Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/140203

Probabilistic Flood Height Estimation of the Limpopo River at the Beitbridge using r-Largest Order Statistics

Robert Kajambeu[1], *Caston Sigauke*[1,*], *Alphonce Bere*[1], *Delson Chikobvu*[2], *Daniel Maposa*[3] *and Murendeni Maurel Nemukula*[4]

[1] Department of Statistics, University of Venda, Private Bag X5050, Thohoyandou, 0950, South Africa
[2] Department of Mathematical Statistics and Actuarial Science, University of the Free State, P. O. Box 339, Bloemfontein 9300, South Africa
[3] Department of Statistics and Operations Research, University of Limpopo, Private Bag X1106, Sovenga, 0727, South Africa
[4] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X54001, Durban, 4000, South Africa

Received: 3 Mar. 2019, Revised: 3 Dec. 2019, Accepted: 8 Dec. 2019
Published online: 1 Mar. 2020

Abstract: The paper presents modelling of uncertainty in extreme return levels of the flood heights of the Limpopo river at the Beitbridge in which the delta and profile likelihood approaches are used in the estimation of the confidence intervals. The modelling approach discussed in this study is a hybrid modelling framework blending a variety of statistical models, techniques and approaches. Monthly flood height data for the years 1992 to 2014 are used. The method is based on a joint generalised extreme value distribution of the r-largest order statistics. The method is more efficient in its use of data than the traditional single maximum observation per block. Estimation of parameters is done using the maximum likelihood method. Using the r-largest order statistics approach, the paper shows that the flood height data can suitably be modelled by the Gumbel class distribution. The 100-year return level is estimated to be 4.981 metres with a confidence interval estimate of (4.886,5.083) using the profile likelihood method. This study is important as it enables accurate estimation of return levels and periods of extreme flood heights. Such analysis helps in risk mitigation, for example, the design of bridges by civil engineers.

Keywords: Generalised extreme value distribution, maximum flood height, r-largest order statistics, Fr*é*chet class distribution, Return period, Return level

1 Introduction

Floods are not an uncommon natural disaster in Southern Africa. In the 1999-2000 rainfall season, Cyclone Eline hit the region and widespread floods devastated large parts of the Limpopo basin (southern and central Mozambique, south eastern Mozambique, parts of South Africa and Botswana) [1]. Some parts of the region also experienced excessively heavy rain episodes in the 2001-2002, 2005-2006 and 2007-2008 rainfall seasons [1]. Also, in March 2019, Malawi, Mozambique and Zimbabwe were hit by Tropical Cyclone IDAI, the devastating effects of which led to these countries declaring states of emergency [2]-[3].

Flooding causes displacement of people, destruction of crops, crop lands and infrastructure, loss of lives and disruption of basic services such as transport, telecommunication and supply of water and electricity [2]-[4]. The disruption of basic services in Mozambique by Tropical Cyclone IDAI plunged South Africa into phase 4 electricity load shedding [3]. It is reported that Tropical Cyclone IDAI affected over 3 million people in Malawi, Mozambique and Zimbabwe leaving to over 839 people dead, over 201,476 people displaced, over 2,347 people injured and over 300 people reported missing [2]. Several cases of cholera and diarrhoea were also reported in the post cyclone period [2]-[3]. In future, Southern Africa is likely to experience more extreme and/or rare weather events such as draughts and floods due to the impact of climate change.

[*] Corresponding author e-mail: csigauke@gmail.com

© 2020 NSP
Natural Sciences Publishing Cor.

The Beitbridge (coordinates 22.2244^{o}S 29.9865^{o}E) across the Limpopo river is an example of infrastructure that has previously been affected by floods and is likely to suffer the same fate in future. In January 2013 both human and vehicle traffic could not cross the Beitbridge because it was flooded [5]. The bridge is the major rail-road gateway into Zimbabwe from South Africa. It is part of a very important trade route which links South Africa with several landlocked countries on the Southern tip of Africa. It is therefore crucial that mitigatory steps are taken in order to lessen the disruptive impact of floods on this important structure.

Flood modelling can be considered as the basis for effective flood mitigation [6]. Extreme value theory (EVT) gives the stochastic framework for such modelling [6]. One of the main outcomes of extreme value analysis is the estimation of return levels for specified periods of time. For example, one could estimate the flood height that occurs on average once every 100 years. Return level information can be used in the design of structures such as bridges, dam walls, sea barriers and nuclear facilities [7]-[9]. Extreme value analysis also aims to identify covariates that drive extremes and to determine if the probability of extremes is increasing over time as a result of climate change.

The most common approach for describing the extreme events of stationary data is the block maximum approach, which models the maxima/minima of a set of adjacent blocks of observations using the generalised extreme value distribution (GEVD). The cornerstone of that approach is the Fisher-Tippett theorem (see [10]) which asserts that the block maxima of a sequence of independently and identically distributed random variables in the limit follows a GEVD. Application of this method is discussed in [11]-[16], among others.

Extremes, as their name suggests, are rare and so in any extreme value analysis, a limited amount of data is available. This makes model estimation difficult [17] with estimates of extreme return levels having large variances [18]. This has necessitated the search for alternatives and improvements to the block maxima approach.

One such alternative is the peak over threshold (POT) approach. This approach involves identification of observations of the time series which are above a predetermined threshold and fitting of the generalised Pareto distribution to those observations. A theorem by [19] states that excesses over a sufficiently high threshold are generalised Pareto distributed. The POT approach is assumed to be more precise than the block maxima method because it utilizes more data and because maxima are not always extremes [6]. This approach is applied in [6], [9], [14], [17], [20], among many others.

Another method which uses more values than the block maxima approach is based on the r-largest observations within a block for small values of r. The pioneer of the use of this method is [21]. Reference [22] presents a detailed discussion of the extension of the EVT to fitting distributions to annual maxima when fitting a distribution to r-largest observations per year and also applies the method to real-life data. Two specification tests which assist in the automation of the process of selecting r are proposed in [23]. The first is a score test for which the p-values are determined through a multiplier procedure. The second test uses the difference in estimated entropy between the GEVD$_r$ and GEVD$_{r-1}$ models, applied to the r-largest order statistics and the $r-1$ largest order statistics, respectively. The later test was used in this work.

Applications of the r-largest order statistics can be found in the papers of [6], [24], [25] and many others. A comparative study by [8] concluded that the r-largest order statistics and POT methods had lower uncertainty on the distribution of parameter and return level estimates compared to the block maxima method.

This paper uses the generalised extreme value distribution (GEVD) based on the r-largest order statistics to model the flood heights of the Limpopo river at the Beitbridge. The focus is on uncertainty quantification of extreme flood heights through the use of the delta and profile likelihood methods in estimating the confidence intervals of extreme floods. The r-largest order statistics approach is used. This approach uses more data compared to using a single maximum in a block and hence is more efficient. To the best of our knowledge, there is no previous study that has focused on modelling the flood height of the Limpopo river at the Beitbridge.

The rest of the paper is organised as follows: Section 2 presents the models. The empirical results are presented in Section 3 and the discussions in Section 4. Section 5 presents some of the contributions of this paper while Section 6 concludes.

2 Methodology

The methodology used in the study is briefly discussed in this section.

2.1 The r-largest order statistics

The use of r-largest order statistics is usually used if there is limited data. This study is motivated by the desire to search for characterisation of extreme value behaviour other than the use of one observation in a block that would enable modelling observations in the upper tails of distributions. Such an approach is more efficient in its use of data.

Let $X_1, X_2, ..., X_n$, be a sequence of independent and identically distributed (i.i.d.) random variables. Define $M_n^{(k)} = k^{th}$ largest of $\{X_1, ..., X_n\}$. If there exists a sequence of constants $\{a_n > 0\}$ and $\{b_n > 0\}$ such that: $P\left\{\frac{M_n^{(r)} - b_n}{a_n} \leq z\right\} \rightarrow G(z)$ as $n \rightarrow \infty$ for some non degenerate distribution G, then, for fixed r, the limiting distribution as $n \rightarrow \infty$ of $\tilde{M}_n^{(r)} = \left(\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(r)} - b_n}{a_n}\right)$ falls within the family having joint probability density function (for $\xi \neq 0$) [26]

$$f\left(x^{(1)}, ..., x^{(r)}\right) = \exp\left\{-\left[1 + \xi\left(\frac{x^{(r)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \times \prod_{k=1}^{r} \frac{1}{\sigma}\left[1 + \xi\left(\frac{x^{(k)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi} - 1}, \tag{1}$$

where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty; x^{(r)} \leq x^{(r-1)} \leq ... \leq x^{(1)}$; and $x^{(k)} : 1 + \xi\left(\frac{x^{(k)} - \mu}{\sigma}\right) > 0$ for $k = 1, 2, ..., r$. For the case $r = 1$, we have the GEVD model. When $\xi \longrightarrow 0$ usually written as $\xi = 0$, the joint density is given as:

$$f\left(x^{(1)}, ..., x^{(r)}\right) = \exp\left\{-\exp\left[-\left(\frac{x^{(r)} - \mu}{\sigma}\right)\right]\right\} \times \prod_{k=1}^{r} \frac{1}{\sigma}\left[-\left(\frac{x^{(k)} - \mu}{\sigma}\right)\right]. \tag{2}$$

Equation 2 reduces to the Gumbel class distribution when $r = 1$.

2.2 Estimation of parameters

The MLE method is used to estimate the parameters of the GEVD$_r$. Due to the fact that the support of the GEVD$_r$ depends on the unknown parameter values, the usual regularity conditions that underline the asymptotic properties of the MLEs are not satisfied. This problem was studied by [27]. In the case $\xi > -0.5$, the usual asymptotic properties of consistency, asymptotic efficiency, and asymptotic normality hold. When these conditions are violated, the Bayesian estimates are then preferred since they do not necessarily depend on these conditions.

Likelihood-based methods of estimating parameters of the EVT models are more reliable compared to other methods for various reasons that include the adaptability to model change, [26]. The likelihood function for the GEVD$_r$ for the case $\xi \neq 0$ is given in equation (3).

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{n}\left(\exp\left\{-\left[1 + \xi\left(\frac{x_i^{(r_i)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \times \prod_{k=1}^{r_i} \frac{1}{\sigma}\left[1 + \xi\left(\frac{x_i^{(k)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi} - 1}\right), \tag{3}$$

for $1 + \xi\left(\frac{x^{(k)} - \mu}{\sigma}\right) > 0$ and $k = 1, 2, ..., r_i; i = 1, ..., n$. Similarly for the case $\xi = 0$ the likelihood is given as:

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{n}\left(\exp\left\{-\exp\left[-\left(\frac{x_i^{(r_i)} - \mu}{\sigma}\right)\right]\right\} \times \prod_{k=1}^{r_i} \frac{1}{\sigma}\left[-\left(\frac{x_i^{(k)} - \mu}{\sigma}\right)\right]\right). \tag{4}$$

2.3 Entropy difference test

One of the tests used to determine a suitable value of r is the entropy difference test. This specification test for the GEVD$_r$ model is based on the difference in entropy for the GEVD$_r$ and GEVD$_{r-1}$ models. The entropy for a continuous random variable with density f is (e.g., [28]).

$$E[-\ln f(x)] = -\int_{-\infty}^{\infty} f(x) \ln f(x) dx. \tag{5}$$

It is an expectation of the negative log likelihood. Assuming $r - 1$ top order statistics fit the GEVD$_{r-1}$, the difference in the log likelihood between GEVD$_{r-1}$ and GEVD$_r$ provides a measure of deviation from the null hypothesis $H_0^{(r)}$, that the specified value of r provides a good fit to the data . Large deviations from the expected difference under $H_0^{(r)}$ suggest that there may be some misspecification of $H_0^{(r)}$. From the log likelihood in equation (4) the difference in log likelihood for the i^{th} block, is given in equation 6.

$$Y_{ir}(\theta) = -\log\sigma - (1 + \xi x_{ir})^{-\frac{1}{\xi}} + (x_{ir-1})^{-\frac{1}{\xi}} - (\frac{1}{\xi} + 1)\log(1 + x_{ir}). \tag{6}$$

2.4 Uncertainty analysis and return level estimation

The quantile functions for the unified GEVD$_r$ are used to estimate high quantiles and predicting the probability of exceedance levels of flood heights. These are given as:

$$\hat{x}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right], & \text{if } \xi \neq 0 \\ \hat{\mu} - \hat{\sigma}\log y_p & \text{if } \xi = 0, \end{cases} \tag{7}$$

where $y_p = -\log(1-p)$. In order to model the uncertainty in the extreme quantile estimates we will use the delta and profile likelihood methods. These are discussed in Sections 2.4.1 and 2.4.2, respectively.

2.4.1 The delta method

Using the delta method the variance of x_p is given as [24]:

$$\mathrm{Var}(\hat{x}_p) \approx \nabla x_p^T V \nabla x_p, \tag{8}$$

where V is the covariance matrix of $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$ and

$$\begin{aligned}\nabla x_p^T &= \left[\frac{\partial x_p}{\partial \mu}, \frac{\partial x_p}{\partial \sigma}, \frac{\partial x_p}{\partial \xi}\right] \\ &= \Bigg[1, -\xi^{-1}\left(1-y_p^{-\xi}\right), \sigma\xi^{-2}\left(1-y_p^{-\xi}\right) \\ &\quad -\sigma\xi^{-1}y_p^{-\xi}\log y_p\Bigg], \end{aligned} \tag{9}$$

which is evaluated at $\left(\hat{\mu}, \hat{\sigma}, \hat{\xi}\right)$. The approximate confidence interval of the flood heights x_p is then given by

$$\left(\hat{x}_p - z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{x}_p)}, \hat{x}_p + z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{x}_p)}\right). \tag{10}$$

2.4.2 The profile likelihood method

The profile likelihood for some parameter θ_i is defined as [26]:

$$\ell(\theta_i) = \max \ell(\theta_i, \theta_{-i}), \tag{11}$$

where θ_{-i} represents components of θ excluding θ_i [26]. To obtain the confidence interval for x_p a re-parametrisation is required in which x_p is one of the parameters in the GEVD$_r$ model, given as follows:

$$\mu = \begin{cases} x_p - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right], & \text{if } \xi \neq 0 \\ x_p - \hat{\sigma}\log y_p & \text{if } \xi = 0, \end{cases} \tag{12}$$

with $y_p = -\log(1-p)$. Now substituting for μ in equation (3) for $\xi \neq 0$ and in equation (4) for $\xi = 0$ results in the log-likelihood function of the GEVD$_r$ with parameters x_p, σ, ξ.

2.5 Forecast verification

Various verification methods including goodness of fit tests for extreme value distributions are discussed in literature. In this paper we are using graphical plots and some verification tests. A detailed discussion of verification statistics is given in [29]. The verification methods used in this study are discussed in the following sections.

2.5.1 Graphical plots

To assess the goodness of fit of a proposed extreme value distribution to the given data, the Quantile-Quantile (QQ) plot is normally used. A good fit to the data is when the QQ-plot follows a 45^0 line.

2.5.2 Verification statistics

Scoring rules are used in this paper to assess the predictive performance of the developed GEVD$_r$ models. The use of the continuous ranked probability score (CRPS) in assessing the predictive performance of the GEVD$_r$ in probabilistic peak wind prediction is proposed in [30]. The authors carried out a comparative analysis with other scoring functions and the CRPS gave the best performance for high quantiles. The expression of the CRPS for the GEVD$_r$ ($\xi \neq 0$) is given as [30]:

$$\begin{aligned} S_{CRP}(F_{GEVD_{\xi\neq 0}}) &= \left[\mu - y_t - \frac{\sigma}{\xi}\right]\left[1 - 2F_{GEVD_{\xi\neq 0}}\right] \\ &\quad -\frac{\sigma}{\xi}\left[2^{\xi}\Gamma(1-\xi) - 2\Gamma_l(1-\xi, -\log F_{GEVD_{\xi\neq 0}})\right], \end{aligned} \tag{13}$$

where Γ represents the gamma function and Γ_l denotes the lower incomplete gamma function. For the case $\xi = 0$ we have

$$\begin{aligned} S_{CRP}(F_{GEVD_{\xi=0}}) &= \mu - y_t + \sigma\left[C - \log 2\right] \\ &\quad -2\sigma Ei(\log F_{GEVD_{\xi=0}}), \end{aligned} \tag{14}$$

where C is the Euler-Mascheroni constant ($C \approx 0.5772$). Usually the CRPS is difficult to calculate for some complex forecast distributions [29]. Alternatives to the use of the CRPS are the logarithmic score (LogS) and Dawid-Sebastiani score (DSS). The LogS is given as:

$$\mathrm{LogS}(y, F) = -\log(f(y)), \tag{15}$$

where y is the density and F is the forecast distribution function. The DSS is given by

$$\mathrm{DSS}(y, F) = \frac{(y - \mu_F)^2}{\sigma_F^2} + 2\log(\sigma_F). \tag{16}$$

For a detailed discussion of these scoring rules see [29].

3 Empirical results

3.1 Exploratory data analysis

In this study, historical monthly flood height data for Limpopo river from Beitbridge are used. The data span over the period 1992 to 2014. Annual maxima from the Beitbridge are derived from the monthly observations. Table 1 shows the descriptive statistics of annual maximum flood heights for the Limpopo river at the Beitbridge. The maximum flood height is 6.71 metres

© 2020 NSP
Natural Sciences Publishing Cor.

Table 1: Table showing descriptive statistics for flood height.

	n	Min	Max	Median	Mean	Std Dev	Skewnes	Kurtosis
Floodheight	23	0.54	6.71	1.64	2.05	1.28	2.18	8.78

which is experienced in the year 2013 and the minimum flood height is 0.54 metres witnessed in the drought year 1992. The mean flood height is 2.05 metres with a standard deviation of 1.28 metres. The distribution of the flood heights is right-skewed as evidenced from the last three panels in Figure 1 and also as shown by the skewness value given in Table 1. A formal stationarity test is carried out for the yearly flood heights using the Kwiatkowski- Phillips-Schmidt-Shin (KPSS) test under the null hypothesis that the data is stationary. The critical values for the 10%, 5%, 2.5% and 1% levels of significance are 0.347, 0.463, 0.574 and 0.739, respectively. Using the maxima data, the KPSS test statistic is 0.2678. Since the test statistic is smaller than the critical values at the given levels of significance we fail to reject the null hypothesis and conclude that the data is stationary.

As shown in Table 2 we fail to reject the null hypothesis that the data follows a GEVD$_r$ for any value of r from 1 to 9. However, attention is limited to $r \leq 6$ order statistics as a result of the reasonable doubt on the validity of the model for all values of $r \geq 7$. For $r \leq 6$, the standard errors of the estimates ($\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$) decrease as the values of r increase, implying an increase in precision of the model. See [18], [26], [30] for details.
The standard errors associated with the shape parameters ξ for various values of r are shown in Figure 2. The variability decreases with an increase in r up to 2, but there is no appreciable change in standard errors for r between 2 and 6. More so, a decreasing trend in standard errors is seen with the modest increase in r. Therefore an optimum choice of r is expected to be between 2 and 6. We choose a fixed value of r to be 6 based on the standard errors in Figure 2 and the plots in Figure 5 in our subsequent analysis.

3.2 MLE results

Table 3 gives the maximum likelihood (ML) estimates, standard errors and the 95% confidence intervals. Using the r-largest order statistics method, the analysis is done for $r = 1$ to $r = 6$. ML estimates of the three GEVD$_r$ parameters and the associated standard errors (SEs) are calculated as shown in Table 3. Four stationary models are considered, M_1 GEVD$_{r=1}$ in which the shape parameter is positive (Fre̒chet class distribution), M_2 GEVD$_{r=1}$ shape parameter is zero (Gumbel class distribution), M_3 GEVD$_{r=6}$ shape parameter is positive and M_4 GEVD$_{r=6}$ shape parameter is zero. Table 4 gives a summary of the AIC values of the four models. We put the models into two groups, i.e. the first group with models M_1 and M_2 and the second group with M_3 and M_4. Likelihood ratio tests are then carried out. The null hypothesis is that the data follows a Gumbel class in each of the two groups. For the first group and second group the $p-$ values are 0.25 and 0.145, respectively. In both cases we fail to reject the null hypothesis and conclude that the Gumbel class provides a better fit of the data from each group. The models M_2 and M_3 (see Table 4) are therefore used in the estimation of high quantiles (extreme return levels).

ML parameter estimates with standard errors in parentheses of r-largest order statistics model fitted to the Limpopo river data at Beitbridge border post with different values of r are given in Table 5 Figure 3 shows the goodness of fit plots. The probability and QQ plots support the model as a good fit.

The corresponding density estimate seems consistent with the histogram of the data. In fact, all four plots do lend support to the fit of the Gumbel model with $r = 1$. The plots in Figure 4 show that the Gumbel model with $r = 6$ also provides a good fit to the data.

3.3 Predictive performance

We also use the CPRS, LogS and the DSS in an attempt to assess the goodness of fit of the models. Using a sample size of 400, the R software is used for the simulations which are based on 1000 Monte Carlo runs. In the simulation study, data is simulated from the parameter estimates of the GEVD$_{r=1}$ and GEVD$_{r=6}$ models. The performance of the estimates are determined through the empirical bias, mean square error (MSE), including the standard error (SE) of the estimates. The coverage probabilities are also calculated and are very close to the specified probabilities which are 0.9 and 0.95 respectively as shown in Table 6. The simulated data was then used to calculate the CPRS, LogS and the DSS, and the results are presented in Table 7. As shown in Table 7 model GEVD$_{r=6}$ has smaller values for all the three skills scores, i.e. CPRS, LogS and the DSS, meaning that it is a better fitting model for probabilistic forecasting of flood heights for the Limpopo river at the Beitbridge border post. The model with $r = 6$ will therefore be used for estimating extreme quantiles (return levels).

© 2020 NSP
Natural Sciences Publishing Cor.

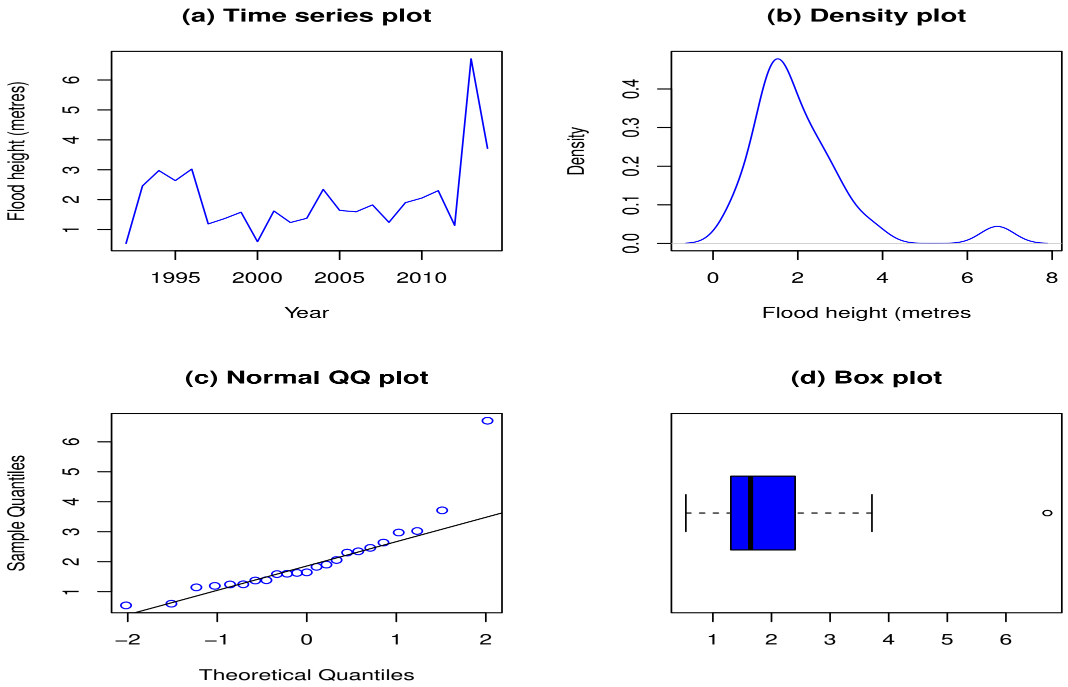

Fig. 1: Flood height for $r = 1$.

Table 2: Entropy difference test for diagnosing generalised extreme value distribution for r-largest order statistics.

r	p	Forwardstop	Strongstop	Statistic	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
2	0.2314	0.9751	0.8499	1.1966	1.7231	0.8619	-0.0186
3	0.8964	1.0641	0.9432	-0.1302	1.6748	0.8431	0.0019
4	0.1277	0.8922	0.8034	1.5232	1.6953	0.7669	0.0332
5	0.1844	1.0181	0.7071	-1.3273	1.6403	0.8089	0.0591
6	0.6320	1.1809	0.7741	0.4790	1.5988	0.8222	0.1212
7	0.5988	1.2263	0.8512	0.5261	1.5619	0.9066	0.2675
8	0.8342	1.3306	1.0684	0.2093	1.9279	2.0890	1.0017
9	0.7610	1.0975	1.3980	0.3042	3.7113	6.8490	1.8391
10	0.5341	0.7638	1.4934	-0.6217	1.5847	1.1578	0.5248

Table 3: Maximum log-likelihood parameter estimates, confidence intervals and standard errors in parentheses of the r-largest order statistics model fitted to the Limpopo river data at Beitbridge border with different values of r.

r	ℓ	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	95%Confidence Interval for ξ
1	32.09	1.49(0.16)	0.72(0.12)	0.14(0.14)	(-0.13;0.13)
2	41.63	1.68(0.16)	0.85(0.09)	-0.01(0.07)	(-0.16;0.13)
3	39.32	1.65(0.15)	0.83(0.09)	0.01(0.07)	(-0.14;0.14)
4	21.98	1.68(0.14)	0.8(0.08)	0.04(0.07)	(-0.13;0.15)
5	7.37	1.63(0.14)	0.81(0.09)	0.04(0.08)	(-0.11;0.20)
6	-22.94	1.58(0.14)	0.82(0.11)	0.12(0.10)	(-0.08;0.32)

© 2020 NSP
Natural Sciences Publishing Cor.

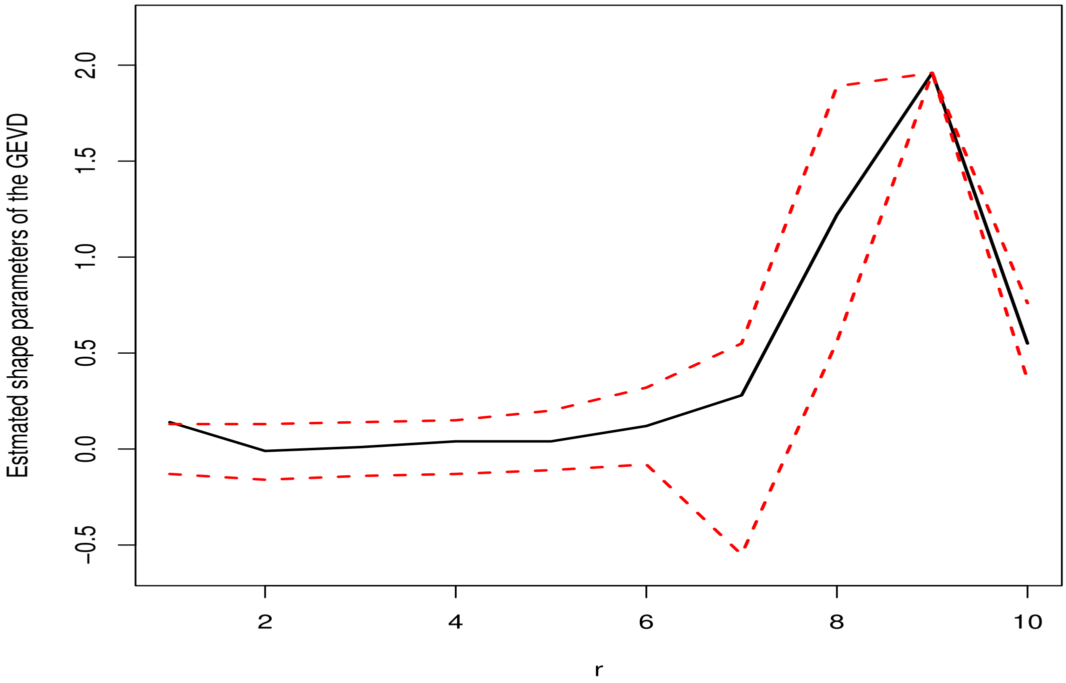

Fig. 2: The GEVD$_r$ shape parameter estimates with 95% confidence interval.

Table 4: Comparative analysis of the proposed models.

	M_1	M_2	M_3	M_4
AIC	68.6561	67.9822	-36.5380	-36.4191
Log Lik	-31.3281	-31.9911	21.2690	20.2096

Table 5: ML parameter estimates with standard errors in parentheses of r-largest order statistics model fitted to the Limpopo river data at Beitbridge border with different values of r.

r	ℓ	$\hat{\mu}$	$\hat{\sigma}$
1	-31.99	1.5531(0.1738)	0.7917(0.1287)
6	21.27	1.6068(0.1263)	0.7383(0.0585)

Table 6: Simulated monthly flood heights parameter estimates, bias, mean square error (MSE), standard error (SE) and coverage probabilities (CP)

	GEVD$_{r=1}$ (Gumbel class)		GEVD$_{r=6}$ (Gumbel class)	
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
Parameter	1.5538	0.7902	1.6077	0.7369
Bias	0.0007	-0.0014	0.0008	-0.0013
MSE	0.0018	0.0010	0.0015	0.0008
SE	0.0416	0.0308	0.0388	0.0288
90% CP	0.8997	0.8944	0.9000	0.9031
95% CP	0.9493	0.9429	0.9496	0.9483

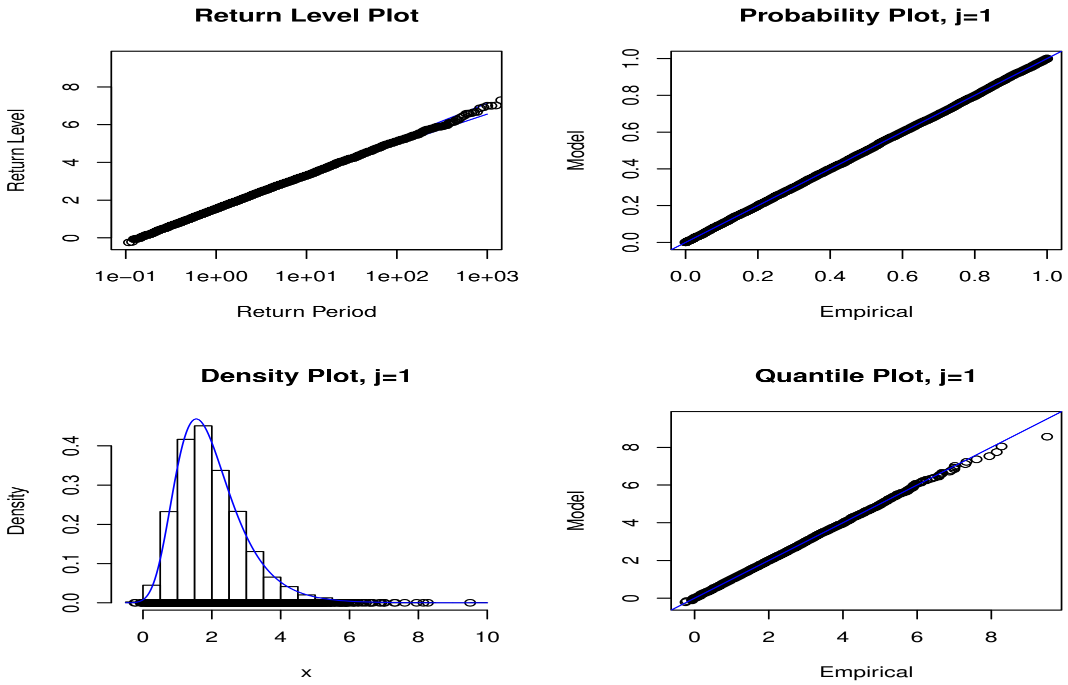

Fig. 3: Diagnostic plots illustrating the fit of the data (Annual flood heights of Limpopo river at Beitbridge border post) to the GEVD$_r$ for r-largest order statistics model with $r = 1$.

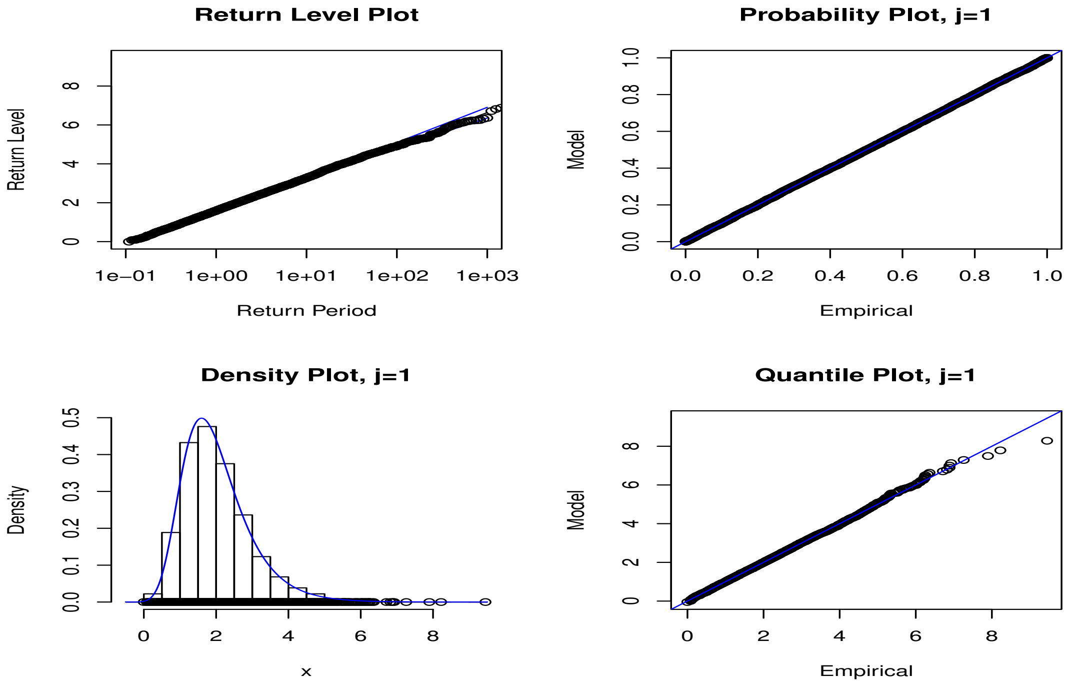

Fig. 4: Diagnostic plots illustrating the fit of the data (Limpopo flood height data at Beitbridge border post) to the GEVD$_r$ for r-largest order statistics model with $r = 6$.

Table 7: Comparison of predictive performance of the models.

Model	CRPS	LogS	DSS
GEVD$_{r=1}$	2.468	4.131	9.481
GEVD$_{r=6}$	2.313	4.077	9.416

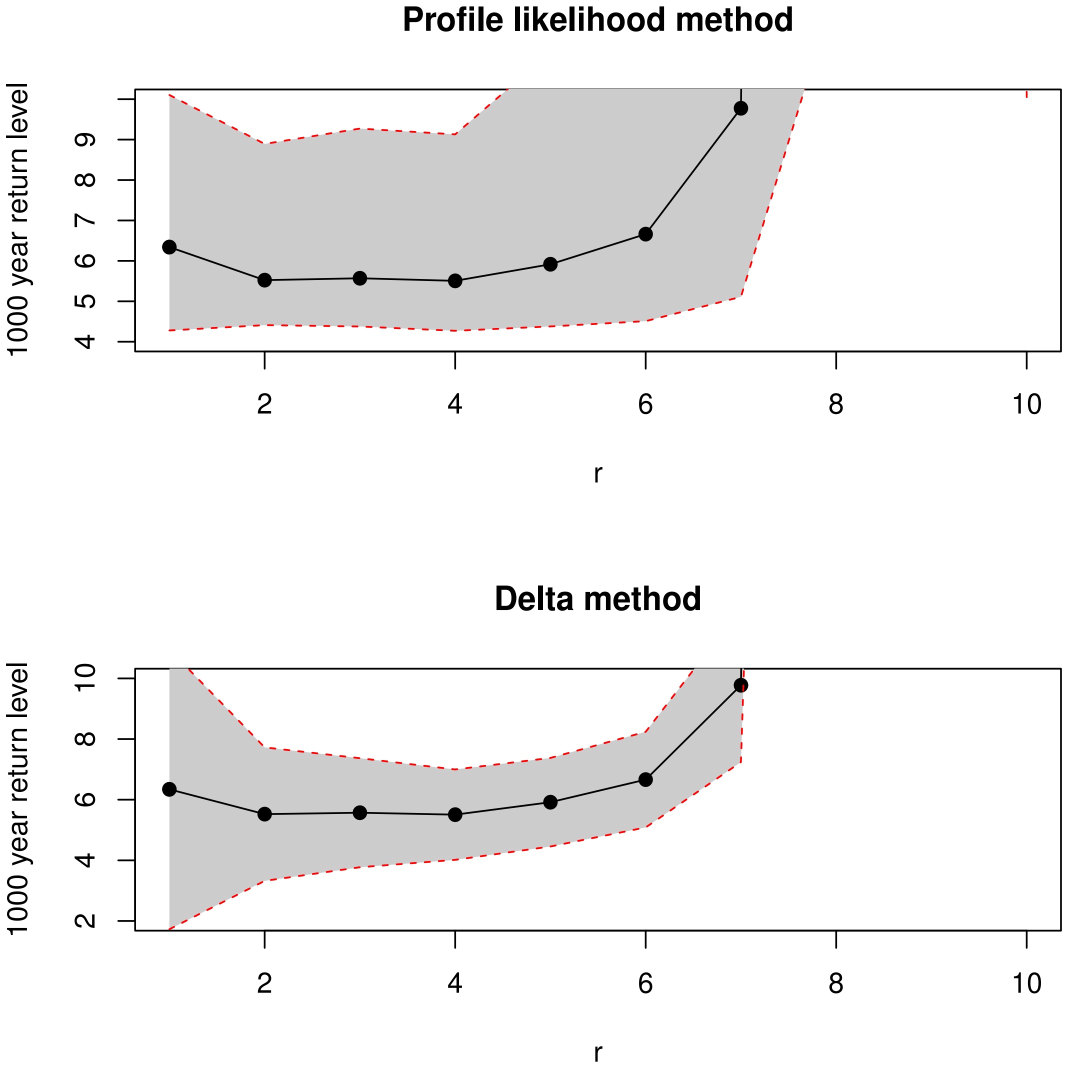

Fig. 5: Return level plot using the profile likelihood and delta methods for determining best value of r.

© 2020 NSP
Natural Sciences Publishing Cor.

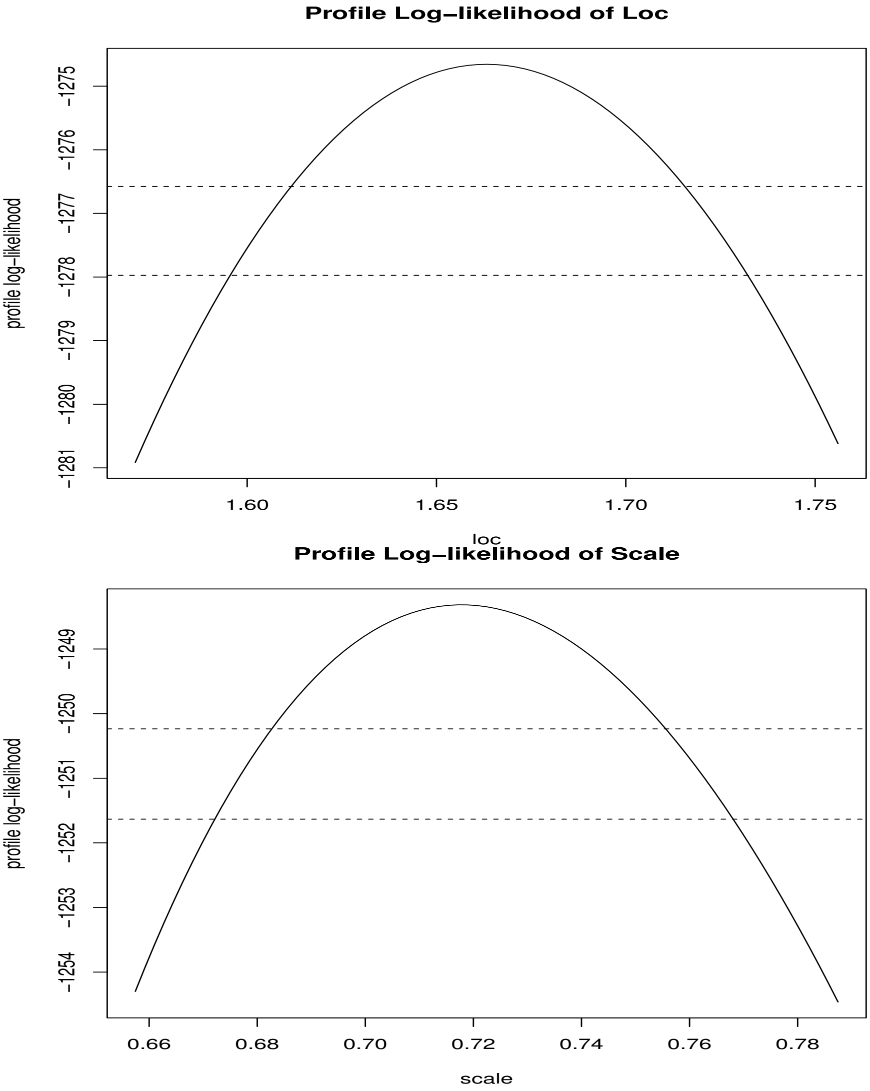

Fig. 6: Profile likelihood for the parameters (GEVD$_{r=6}$ (Gumbel class distribution)) parameters.

© 2020 NSP
Natural Sciences Publishing Cor.

3.4 Uncertainty analysis and extreme quantile estimation of flood heights

Table 8 summarises the return periods together with the 95% prediction intervals using the delta and the profile likelihood methods, respectively. The 100-year return level is estimated to be 4.981 metres with 95% prediction intervals of (4.886, 5.076) and (4.886, 5.083) from the delta and profile likelihood methods, respectively. The interval widths from the profile likelihood method are wider than those from the delta method. This indicates the uncertainty in the return levels from the profile likelihood method is better than that from the delta method. This finding is in agreement with other researchers (see for example [31]-[33] who found the Gumbel distribution being suitable for modelling flood height data. Other researchers (see for example [12]) have found contrasting results. The 100-year return level using the GEVD$_r$ is comparable with the flood height for 2013 when the flow reached 6.707 metres. The results reveal that the year 2013 flood height was well above the 100-year flood height at the site and also lies outside the prediction intervals of both the delta and profile likelihood methods. These findings indicate that the year 2013 flood height was indeed a very rare event and also explains the deleterious impact of the flood event at the site.

4 Return level plots

Figure 5 summarises various return levels with their corresponding return periods using the profile likelihood and delta methods together with 95% confidence intervals. At large return levels, the profile likelihood allows for asymmetric intervals compared to the delta method [24].

The profile likelihood for the GEVD$_{r=6}$ (Gumbel class distribution) for both the location and the scale parameters are given in the appendix, Figure 6.

5 Summary of results and discussion

The highlights of this study are summarised as follows:

1. Initially the best value of r is determined based on the entropy difference test.
2. This is then followed by plotting the standard errors for the estimated shape parameters for each of the models for different values of r from $r = 1$ to $r = 10$.
3. Results from (a) and (b) show that the best value of r is 6.
4. Various models belonging to the Fr*é*chet and Gumbel class distributions are fitted for both $r = 1$ and $r = 6$.
5. Empirical results from (d) show that the Gumbel class fits better than the Fr*é*chet class distribution.
6. One of the contributions of this study is in the use of proper scoring rules, Continuous Rank Probability Score (CPRS), Dawid-Sebastiani Score (DSS) and the Logarithmic Score (LogS). These scores were estimated based on Monte Carlo simulations of GEVD$_{r=1}$ and GEVD$_{r=6}$.
7. Another contribution is in uncertainty modelling of the extreme return levels of the flood heights of the Limpopo river at Beitbridge in which the delta and profile likelihood approaches are used in the estimation of the confidence intervals.
8. The modelling approach discussed in this study is a hybrid modelling framework blending a variety of of statistical models, techniques and approaches which to the best of our knowledge is not discussed in literature in the context of quantifying uncertainty associated with flood heights over bridges.

6 Conclusion

In this paper, we have presented and analysed the flood heights data for the Limpopo river at Beitbridge border post using the generalised extreme value distributions. Modelling of flood heights is quite important in the field of hydrology for decision making. The Gumbel class distribution is found to be the best fit for the data in all the modelling frameworks in this paper. From this work, we can conclude that the distribution of extreme flood heights for the Limpopo river at Beitbridge border post is heavy-tailed and are more likely than predicted with a normal distribution model. Return periods are estimated using the profile likelihood and delta methods. Using the GEVD$_{r=6}$, the 100-year return level is estimated to be 4.981 metres with a 95% confidence interval estimate of (4.886,5.083) based on the profile likelihood method. The delta method gives an estimate of 4.981 metres with a 95% confidence interval estimate of (4.886,5.076). This information could be useful in any future reconstruction of the bridge. The developed models established in this study are consistent with cumulative (or moving sums) annual maximum series flood height and therefore appear reliable to use for flood frequency analysis.

Future research from this paper will involve a probabilistic description and modelling of flood heights. The other area will include a multi-site regional analysis.

Appendix

The following R statistical packages are used in this paper:

1. “eva” developed by [34].
2. “ScoringRules” developed by [35].
3. “urca” developed by [36].

Table 8: Estimating return levels delta and profile 95% CI.

	Delta method	Profile likelihood method
Year	$\hat{x}_p$ (m) (GEVD$_{r=6}$)	$\hat{x}_p$ (m) (GEVD$_{r=6}$)
10	(3.217,3.253,3.288)	(3.218,3.253,3.288)
20	(3.732,3.782,3.831)	(3.734,3.782,3.832)
30	(4.027,4.086,4.145)	(4.028,4.086,4.147)
40	(4.234,4.301,4.368)	(4.235,4.301,4.371)
50	(4.394,4.467,4.540)	(4.395,4.467,4.544)
60	(4.524,4.603,4.681)	(4.525,4.603,4.685)
70	(4.634,4.717,4.800)	(4.634,4.717,4.805)
80	(4.728,4.816,4.904)	(4.729,4.816,4.909)
90	(4.812,4.903,4.995)	(4.812,4.903,5.001)
100	(4.886,4.981,5.076)	(4.886,4.981,5.083)
150	(5.171,5.281,5.391)	(5.171,5.281,5.399)
200	(5.372,5.494,5.614)	(5.372,5.494,5.624)
500	(6.009,6.170,6.331)	(6.008,6.170,6.345)
1000	(6.486,6.682,6.878)	(6.485,6.682,6.895)

References

[1] M.T. Lukamba, Natural disasters in African countries: what can we learn about them? *The Journal for Transdisciplinary Research in Southern Africa*, vol. **6(2)**, 478-495 (2010).

[2] ReliefWeb, SADC Regional Humanitarian Floods Appeal in Response to Tropical Cyclone IDAI [EN/PT]. *ReliefWeb Report from Southern African Development Community*, Published on 11 April 2019.

[3] A. Slabbert, L. Slatter, Eskom blames Cylone IDAI for SA's power outages. *City Press, Johannesburg*, Released: 2019-03-17, 07:13 (2019).

[4] S. Huq, S. Kovats, H. Reid, D. Satterthwaite, *Reducing Risks to Cities from Disasters and Climate Change, Environment and Urbanization* - ENVIRON URBAN; Bartlett, S., Eds.; SAGE publishing: London, UK, 3-15, ISSN: 09562478, (2007).

[5] The Zimbabwe Herald News Paper 21 January (2013).

[6] K. Engeland, H. Hisdal, A. Frigessi, D. Dorah, Practical Extreme Value Modelling of Hydrological Floods and Droughts: A Case Study, *Extremes*, **7(1)**, 5-30 (2004). https://doi.org/10.1007/s10687-004-4727-5.

[7] G.I. Schuëller, *Application of Extreme Values in Structural Engineering, Statistical Extremes and Applications*. NATO ASI Series (Series C: Mathematical and Physical Sciences) **131**; de Oliveira, J.T., Eds.; Springer, Dordrecht:Switzerland, 221-234, (1984).

[8] Y. Hamdi, L. Bardet, C.M. Duluc, V. Rebour, Extreme storm surges: A comparative study of frequency analysis approaches, *Natural Hazards and Earth System Sciences*, **14**, 2053-2067 (2014). https://doi.org/10.5194/nhess-14-2053-2014.

[9] F. Quintero, R. Mantilla, C. Anderson, D. Claman, W. Krajewski, Assessment of Changes in Flood Frequency Due to the Effects of Climate Change: Implications for Engineering Design, *Hydrology*, **5(19)** (2018). https://doi.org/10.3390/hydrology5010019.

[10] R. Fisher, L. Tippett, *Limiting forms of the frequency distribution of the largest or smallest member of a sample*, Mathematical Proceedings of the Cambridge Philosophical Society, **24(2)**, 180-190, (1929). https://doi.org/10.1017/S0305004100015681.

[11] D. Chikobvu, R. Chifurira, Modelling of extreme minimum using generalised extreme value distribution for Zimbabwe, *South African Journal of Science*, 111(9-10), 1-8 (2015). https://dx.doi.org/10.17159/SAJS.2015/20140271

[12] I.S. Kamwi. *Fitting extreme value distributions to the Zambezi River flood water levels recorded at Katima Mulilo in Namibia*. MSc dissertation, University of the Western Cape, South Africa, (2005).

[13] D. Maposa, J. Cochran, M. Lesaoana, C. Sigauke, Estimating high quantiles of extreme flood heights in the lower Limpopo river basin of Mozambique using model based Bayesian approach, *Natural Hazards and Earth System Sciences Discussions*, **2(8)**, 5401-5425 (2014). https://doi.org/10.5194/nhessd-2-5401-2014.

[14] A.H. Tadesse. *Regional Flood Frequency Analysis in Southern Africa*. MSc thesis, University of Oslo, Norway, (2011).

[15] L. Singo, P. Kundu, J. Odiyo, F. Mathivha, T. Nkuna, *Flood frequency analysis of annual maximum stream flows for Luvuvhu river catchment, Limpopo province, South Africa*, Unpublished work, 1-7, (2012).

[16] J.E. Morrison, J.A. Smith, Stochastic modeling of flood peaks using the generalised extreme value distribution, *Water Resources Research*, **38(12)**, 41-1-41-12 (2002). https://doi.org/10.1029/2001WR000502.

[17] L. Kozlowska. *Modelling Extreme Values with reference to River Flooding*. MSc dissertation, City University, London,UK, (2004).

[18] C. Soares, M.G. Scotto, Application of the r-largest-order statistics for long-term predictions of significant wave height, *Coastal Engineering*, **51(5-6)**, 387-394 (2004). https://doi.org/10.1016/j.coastaleng.2004.04.003.

[19] J. Pickands (III), Statistical Inference Using Extreme Order Statistics, *The Annals of Statistics*, **3(1)**, 119-131 (1975). https://doi.org/10.1214/aos/1176343003.

[20] Y. Osman, R. Fealy, J. Sweeney, Modelling extreme temperatures in Ireland under global warming using a hybrid peak-over-threshold and generalised Pareto distribution approach, *International Journal of Global Warming*, **7**, (2015). https://doi.org/10.1504/IJGW.2015.067414.

[21] I. Weissman, Estimation of parameters and large quantiles based on the $k-$largest observations, *Journal of American Statistical Association*, **78(364)**, 812-815 (1978). https://doi.org/10.2307/2286285.

[22] L.R. Smith, Extreme value theory based on the r-largest annual events, *Journal of hydrology*, **86**, 27-43 (1986).

[23] B. Bader, J. Yan, X. Zhang, Automated selection of r for the r-largest order statistics approach is done with adjustment for sequential testing, *Statistics and Computing*, **27(6)**, 1435-1451 (2017). https://doi.org/10.1007/s11222-016-9697-3.

[24] Y. An, M.D. Pandey, The r largest order statistics model for extreme wind speed estimation, *Journal of Wind Engineering and Industrial Aerodynamics*, **95(3)**, 165-182 (2007). https://doi.org/10.1016/j.jweia.2006.05.008.

[25] M.M. Nemukula, C. Sigauke, Modelling average maximum daily temperature using r largest order statistics: An application to South African data, *Jamba: Journal of Disaster Risk Studies*, **10(1)**, a467 (2018). https://doi.org/10.4102/jamba.v10i1.467.

[26] S. Coles, em An introduction to statistical modelling of extreme values, Springer-Verlag, London, (2001).

[27] L.R. Smith, MLE in a class of non regular cases, *Biometrika*, **72(1)**, 67-90 (1985). http://dx.doi.org/10.2307/2336336.

[28] B. Bader, J. Yan, X. Zhang, Automated selection of r for the r-largest order statistics approach is done with adjustment for sequential testing, *Statistics and Computing*, **27(6)**, 1435-1451 (2017). https://doi.org/10.1007/s11222-016-9697-3.

[29] T. Gneiting, M. Katzfuss, Probabilistic forecasting, *The Annual Review of Statistics and its Application*, **1**, 125-151 (2014). https://doi.org/10.1146/annurev-statistics-062713-085831.

[30] P. Friederichs, T.L. Thorarinsdottir, Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction, *Environmetrics*, **23**, 579-594 (2012). https://onlinelibrary.wiley.com/doi/epdf/10.1002/env.2176.

[31] M.J. Mamman, O.Y. Martins, J. Ibrahim, M.I. Shaba, Evaluation of Best-Fit Probability Distribution Models for the Prediction of Inflows of Kainji Reservoir, Niger State, Nigeria, *Air, Soil and Water Research*, **10**, 1-7 (2017). https://doi.org/10.1177/1178622117691034.

[32] N. Bhagat, Flood Frequency Analysis Using Gumbel's Distribution Method: A Case Study of Lower Mahi Basin, India, *Journal of Water Resources and Ocean Science*, **6(4)**, 51-54 (2017). https://doi.org/10.11648/j.wros.20170604.11.

[33] F. Onen, T. Bagatur, Prediction of Flood Frequency Factor for Gumbel Distribution Using Regression and GEP Model, *Arabian Journal for Science and Engineering*, **42**, 3895-3906 (2017). https://doi.org/10.1007/s13369-017-2507-1.

[34] B. Bader, J. Yan, *Extreme value analysis with goodness-of-fit testing*, R package version 0.2.4, (2016).

[35] A. Jordan, F. Krueger, S. Lerch, *Scoring Rules for Parametric and Simulated Distribution Forecasts*, R package version 0.9.5 (2018).

[36] B. Bernhard Pfaff, E. Zivot, M. Stigler, *Unit Root and Cointegration Tests for Time Series Data*, R package version 1.3-0, (2016).

Robert Kajambeu holds an MSc in Statistics from the University of Venda in South Africa. His research interests are in extreme value theory, time series analysis, disaster risk and climate change modelling.

Caston Sigauke holds a PhD in Statistics from the University of the Free State and an MSc in Operations Research from the National University of Science and Technology in Zimbabwe. He is a senior lecturer in the Department of statistics at the University of Venda in South Africa. He is a member of the: International Statistical Institute, International Institute of Forecasters (IIF), Operations Research Society of South Africa (ORSSA) and South African Statistical Association (SASA). He is a Chartered Statistician since 2013 (13ChM012) of the Institute of Certificated and Chartered Statisticians of South Africa (ICCSSA) and currently serves as a Board Member of (ICCSSA). His fields of expertise are forecasting and time series, statistics of extremes, statistical learning and modelling. Caston's research is on probabilistic load and renewable energy (solar and wind) forecasting including the optimization of grid integration of renewable energies.

Alphonce Bere is a senior lecturer in the Department of Statistics at the University of Venda in South Africa. He holds a PhD in Applied Statistics from the University of the Western Cape in South Africa. His research interests are in survival analysis and modelling using copulas.

Delson Chikobvu is a senior lecturer at the University of the Free State in the department of Mathematical Statistics and Actuarial Science, South Africa. He holds a PhD in Mathematical Statistics from the University of the Free State. His research interests are in decision sciences, statistics, econometrics, mathematical finance, statistics of extremes, Bayesian statistics and energy forecasting.

Daniel Maposa is a senior lecturer in the Department of Statistics and Operations Research, School of Mathematical & Computer Sciences, Faculty of Science & Agriculture, University of Limpopo. He holds a PhD degree in Extreme Value Statistics, a Master of Science degree in Operations Research and Statistics, and an Honours degree in Applied Mathematics. He has published more than 21 journal research articles in internationally accredited journals, two book chapters and three conference proceedings. In his research activities, he has been all over the world attending international conferences and presenting his research work in statistics of extremes in countries such as New Zealand, Australia, China, Switzerland, Brazil, Morocco, Botswana and Malaysia. Daniel Maposa is a registered professional natural scientist (Pr.Sci.Nat.) in Statistical Sciences & Mathematical Sciences and is a member of International Statistical Institute (ISI), member of South African Statistical Association (SASA) and member of Operations Research Society of South Africa (ORSSA). He also regularly attends the South African Statistical Association (SASA) conference annually and the ISI World Statistics Congress organised bi-annually.

Murendeni Maurel Nemukula holds a Master of Science (M. Sc) in Mathematical Statistics from University of the Witwatersrand, a Bachelor of Science honours in Statistics from University of Limpopo and a Bachelor of Science in Mathematics and Statistics from University of Venda. He is a full member of the International Statistical Institute (ISI), South African Statistical Association (SASA), and the Operations Research Society of South Africa (ORSSA). He is currently a Statistics lecturer at University of KwaZulu-Natal (UKZN) Westville campus. His area of research is extreme value theory and quantile regression with application in climate change, energy demand, meteorology, disaster risk and finance. He has been lecturing undergraduate Statistics modules since 2012 to date and is currently involved in supervising postgraduate research students. He is currently studying towards a PhD in Statistics.

© 2020 NSP
Natural Sciences Publishing Cor.