Statistics and Its Interface Volume 10 (2017) 131–144

Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalised Pareto distribution

Daniel Maposa*, James J. Cochran, and 'Maseka Lesaoana

In this paper we fit a time-heterogeneous generalised Pareto distribution (GPD) to the flood heights in the lower Limpopo River basin of Mozambique (LLRB). The maximum likelihood method is used for parameter estimation of the nonstationary GPD. We take an in-depth review of the merits of peaks-over-threshold and block maxima. We also show the relationship between generalised extreme value (GEV) distribution and GPD in a mathematical proof and discuss the link between the mathematical proof and the findings. Nonstationary time-dependent GPD models with a trend in the scale parameter are considered in this study. The results show overwhelming evidence in support of the existence of a linear trend in the scale parameter of the GPD models at all the three sites in the LLRB. The time-heterogeneous GPD models developed in this study were found to be statistically worthwhile and provide an improvement in fit over the time-homogeneous GPD models based on the goodness-of-fit tests. This study shows the importance of extending the time-homogeneous GPD models to incorporate climate change factors such as trend in the LLRB. The models developed in this study are expected to be more reliable than their stationary counterparts for planning and decision making processes in Mozambique.

AMS 2000 subject classifications: Primary 62 Statistics; secondary 07 Data Analysis.
Keywords and phrases: Extremes, Peaks-over-threshold, Limpopo River, Generalised Pareto distribution.

1. INTRODUCTION

The presence of long-term trends in extreme events such as annual river flow or precipitation data attributed to climate variability has become an active area of interest for hydrologists and climatologists in the 21st century in order to investigate climate change scenarios and improve research on climate impact on weather extremes [19, 21, 22]. These trends result in nonstationary processes which have the characteristic to systematically change with time [4]. Extreme value theory is the branch of statistics used extensively to study very low or very high values in the tail of some distribution. However, when the process is nonstationary the usual extreme value limit models are not applicable. The usual procedure when dealing with nonstationary processes is to adopt a pragmatic approach of using the standard extreme value models as basic templates that can be enhanced by statistical modelling [4].

It is argued that although the atmosphere-ocean general circulation models (AOGCMs) from an Intergovernmental Panel on Climate Change assessment report (IPCC AR4) were projecting increases in intense precipitation and flooding at a large spatial scale, the models are limited in terms of their ability to quantify extreme events at regional and at-site scales which are crucial for decision making [21]. The use of statistical techniques in river flow or large scale precipitation extremes associated with climate change has been limited [21, and references therein]. According to [22] some researchers have established that global change related extreme events are expected to be on the rise all over Europe although there is no general agreement in concluding that the frequency and magnitude of floods have increased due to climate related changes. The problem in lack of a general agreement is attributed to shortage of instrumental data records in many regions of the continent of Europe. In our view these findings and problems encountered in Europe should draw similarities with climate related extreme events in Africa, particularly in Southern Africa where the study area in this paper is situated although the climatic conditions are very different. In general, several studies in Europe project increases in floods in some regions, for example, in Catalonia the northeast of Spain [22]. In a separate study based on climate simulations, [19] argued that a warmer climate could increase the proportion of floods. It is generally accepted that the expected climatic changes are associated with a higher frequency of occurrence of extreme floods but not associated with a higher intensity of extreme floods [19].

According to [22] with further reference to a review by [10] "there are already a few studies regarding flood magnitude and occurrence changes at river basin level". However scientific literature on these climatic change related ex-

*Corresponding author.

tremes is mainly focused on the continent of Europe, the UK and North America [22, and references therein].

Given that the climatic conditions in these areas (Europe, UK and North America) differ strongly from the current climatic conditions in Southern Africa, and that future projections may also be very different, this study aims to assess and develop climate related models for the future flood trends in the lower Limpopo River basin (LLRB) of Mozambique, an area in Southern Africa that has not been deeply studied. Recently [1] studied the impact of climate change on streamflow in four large African river basins including Limpopo River basin using a geo-scientific model and found that the Limpopo River basin is highly affected by climate variability. Our statistical approach will complement the work by [1] through developing statistical climate change models which will help in decision making for the basin.

According to [7] there are two fundamental approaches widely used in statistics of extremes namely peaks-over-threshold (POT) (or partial duration series) and block maxima (BM) (also called annual maximum series). The BM approach in extreme value theory (EVT) consists of dividing the observation period into blocks (non-overlapping periods of equal size) and restricts attention to the maximum observation in each block (naturally years in the case of floods). The new observations (maxima series) follow approximately the generalised extreme value (GEV) distribution [7]. In the POT approach in EVT, one chooses a reasonably high threshold and selects those of the initial observations that exceed the predetermined threshold. The probability distribution of the new observations (exceedances) follows approximately a generalised Pareto distribution (GPD) [6, 7]. The exact conditions under which the POT statistical method is justified are described by a second order term [6: Section 2.3]. In the case of BM approach it is generally accepted that the maxima follow very well an extreme value distribution and [7] give more theoretical details on the exact conditions of the BM statistical method.

It is argued [5, 7] that the POT method makes better use of the available information since it retains all 'relevant' high observations whereas the BM method on one hand misses some of these high observations and, on the other hand, might retain some lower observations (the latter 'hand' might also be the BM merit over POT as presented by de Haan at the Extreme Value Analysis (EVA2013) conference, 8–12 July 2013, Shanghai, China).

The relevant merits of BM and POT are discussed in detail [7] with references to several papers based on simulated data. Among the merits are that POT estimates are better than BM estimates if the number of exceedances is larger than 1.65 times the number of blocks for the Gumbel family of distributions ($\xi = 0$) when using maximum likelihood (ML) parameter estimators. The POT is as efficient as BM for high quantiles using probability weighted moment (PWM) or L-moment parameter estimators. Provided the number of exceedances is larger than the number of blocks, POT is more preferable for fat-tailed distributions (Fréchet type), whereas BM is more efficient for short-tailed distributions (Weibull type). When using historical data, the gains with the BM are in the range of the gains with the POT method based on ML estimators [7, and references therein]. Based on simulation studies, the POT samples with an average of two or more observations above the threshold per block have more accurate estimates than the corresponding BM estimates and the accuracies become similar and rather good with more than 200 years of historical data [7, and references therein].

All these studies on merits of POT, some with mixed views, show in general that the POT is more efficient than the BM in many ways provided that the number of exceedances is greater than the number of blocks. The two methods have comparable performances when the sample sizes are large. However, the theoretical comparison performed [7] showed that BM is more efficient with lower asymptotic variances of both extreme value index and quantile estimators for BM as compared with POT. The minimal mean square error is also lower for BM under normal circumstances [7].

Our study uses the POT method in order to utilise the richness of information from the big data records contained at the three sites in the LLRB of Mozambique considered in this study. The parameters of the GPD in this paper are estimated by the ML method.

The outline of the rest of the paper is organised in the following structure. Section 2 presents the research methodology, Section 3 presents the results and discussion of the findings, and finally Section 4 gives the concluding remarks.

2. RESEARCH METHODOLOGY

This section presents the study sites and the data used in the study, a brief probability framework of POT approach including the extension of time-homogeneous GPD model to linear trend models. We also prove the link between POT and BM methods through a mathematical proof using the results in literature [4].

2.1 Study sites and data

The data used in this study was obtained from the Mozambique National Directorate of Water (DNA), the authority responsible for water management in Mozambique in the Ministry of Public Works and Housing. The data are hydrometric daily flood heights (in metres) recorded at Chokwe (1951–2010), Combomune (1966–2010) and Sicacate (1952–2010) hydrometric stations for the lower Limpopo River of Mozambique as presented in Figure 1 [14, 15]. The three sites are such that Combomune is located in the upper part of the basin about 162 km from the border with South Africa and Zimbabwe, Chokwe is located in the middle of the basin about 130 km downstream of Combomune and Sicacate is further downstream of Chokwe in the lower part of the basin on the way to the Indian Ocean. The daily flood heights at each site were recorded three times a day,

i.e. morning, afternoon and evening periods. There was a number of missing values in-between the years at each site but this number was counterbalanced by the fact that the data was recorded three times a day making the number of missing values negligible. Further scrutiny on the data reveals that most of the missing values occurred in years of severe droughts or during the winter season which is usually characterised by very low rainfall. Since the interest of this study is in higher values above a certain high threshold, it would mean that these missing values were still likely going to be below the threshold and therefore irrelevant in the study.

2.2 Peaks-over-threshold and generalised Pareto distribution

The approach used in this study is POT. The POT method considers only those of the initial observations that exceed a pre-specified high threshold. In a more formal approach, let $X = X_1, ..., X_n$ be independent and identically distributed (iid) random variables representing flood heights. If F is a distribution function (possibly unknown) of the flood heights X, then the conditional excess $(X - u)$ distribution function is:

$$F_u(y) = P(X - u \le y | X > u) = \frac{P(X - u \le y \text{ and } X > u)}{P(X > u)} = \frac{F(y+u) - F(u)}{1 - F(u)}, \ 0 \le y \le x_F - u, \tag{1}$$

where u is the threshold, $y = x - u$ are the excesses and $x_F < \infty$ is the right endpoint of F [4, 13]. The threshold, u, is selected using the mean residual life plots and threshold choice plots [3, 4, 5]. Based on the [2] and [16] theorems, for a large class of underlying distribution of flood heights F the conditional excess distribution function $F_u(y)$, for large u, is well approximated by a Generalised Pareto Distribution (GPD):

$$G(\sigma, \xi; y) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma} y\right)^{-1/\xi}, & \text{for } \xi \neq 0, \ 0 \le y \le x_F - u, \\ 1 - \exp\left(-\frac{y}{\sigma}\right), & \text{for } \xi = 0, \end{cases} \tag{2}$$

where σ and ξ are, respectively, the shape and scale parameters. These parameters are estimated by the maximum likelihood (ML) method in this study.

Some key points in extreme value theory are that if BM series have a GEV distribution then POT series have an associated GPD. Additionally, the shape parameter ξ in the GPD exactly equals that of the corresponding GEV distribution [4, 9, 13].

We show a mathematical connection between POT and BM methods based on the theoretical results [4]. We start by assuming that the POT model holds, such that values larger than u occur according to a Poisson process with intensity λ (the expected number of events in any time interval of length 1), where this process is independent of the sizes of the exceedances and the sizes of the exceedances are iid and follow a GPD, $G(y) = 1 - (1 + \frac{\xi}{\sigma} y)_+^{-1/\xi}$. Then

$$\begin{aligned} &P(M_T \le u + y) \\ &= \sum_{k=0}^{\infty} P(M_T \le u + y, \text{ there are } k \text{ exceedances in } [0, \mathrm{T}]) \\ &= \sum_{k=0}^{\infty} G(y)^k \frac{(\lambda T)^k}{k!} \exp\{-\lambda T\} \\ &= \sum_{k=0}^{\infty} (1 - (1 + \frac{\xi}{\sigma} y)_+^{-1/\xi})^k \frac{(\lambda T)^k}{k!} \exp\{-\lambda T\} \\ &= \exp\{(1 - (1 + \frac{\xi}{\sigma} y)_+^{-1/\xi}) \lambda T\} \exp\{-\lambda T\} \\ &= \exp\{-(1 + \frac{\xi}{\sigma} y)_+^{-1/\xi} \lambda T\} \\ &= \exp\{-(1 + \xi \frac{y - ((\lambda T)^\xi - 1)\sigma/\xi}{\sigma (\lambda T)^\xi})_+^{-1/\xi}\}, \ \mathrm{T} \text{ is length of observation period.} \end{aligned} \tag{3}$$

In the last expression we note that $((\lambda T)^\xi - 1)\sigma/\xi$ is a constant and also the denominator in the last expression $\sigma(\lambda T)^\xi$ is a constant. If we represent the two constants by μ_0 and σ_0, respectively then Eqn. (3) simplifies to:

$$P(M_T \le \mu + y) = \exp(-(1 + \xi \frac{y - \mu_0}{\sigma_0})^{-1/\xi}$$

which is a GEV distribution.

Thus we have proved the mathematical relationship between GPD and GEV which implies the connection between POT and BM. It will be interesting to see whether the practical results at the three sites in this study will also show some connection.

2.3 Threshold selection

Two threshold selection techniques were used in this study namely mean excess life plot and threshold choice plot. The mean excess life plot is an exploratory technique carried out prior to model selection while threshold choice plot is based on an assessment of the stability of parameter estimates through fitting of models (GPD in this study) across a range of different thresholds [4].

The choice of a threshold is critical to any POT analysis. It is based on a trade-off between bias and variance. Too high a threshold would discard too much data and generate a few exceedances leading to high variance of the estimate of the parameters. On the other hand, too low a threshold would necessitate using data that are no longer considered as being in the tails of the distribution and this will violate the asymptotic basis of the model, thereby leading to an increase in bias [4, 13]. The standard practice is to choose as

low a threshold as possible provided the limit model gives a reasonable approximation. The bias-variance trade-off principle is based on choosing a low enough threshold value to have sufficient data to estimate the parameters σ and ξ, and high enough threshold value for the asymptotic theorem to be considered accurate [4].

Let $x_{(1)} < x_{(2)} < ... < x_{(n_u)}$ be the exceedances $(x_i : x_i > u)$ that are obtained from our sample and define threshold excesses by $y_j = x_{(j)} - u$, for $j = 1, ..., n_u$, then the empirical mean excess life plot is defined by the points:

$$\left\{(u, e_{n_u}(u)) : u < x_{(n_u)}\right\}, \tag{4}$$

where n_u is the number of observations that exceed u, $x_{(n_u)}$ is the largest value of X_i, and $e_{(n_u)} = \frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)} - u)$. The mean residual life plot should be approximately linear above a threshold u_0 at which the GPD provides a valid approximation to the excess distribution [4, 13]. The linearity of the empirical mean excess life plot forms the basis of deciding a threshold [13]. The interpretation of the mean residual life plot in practical situations is not always an easy task [4] and this complexity can be eased by complementing the mean residual life plot with other plots such as the threshold choice, L-moment and dispersion plots [13]. The L-moment and dispersion plots will not be considered in this study.

The threshold choice plot is based on the result that if $X \sim GPD(u_0, \sigma_0, \xi_0)$, then let u_1 be another threshold such that $u_1 > u_0$. Then $X|X > u_1$ is also another GPD with updated parameters $\sigma_1 = \sigma_0 + \xi_0(u_1 - u_0)$ and $\xi_1 = \xi_0$. Threshold choice plots are given by the points defined by:

$$\left\{(u_1, \sigma_*) : u_1 < x_{(n_u)}\right\} \quad \text{and} \quad \left\{(u_1, \xi_1) : u_1 < x_{(n_u)}\right\}, \tag{5}$$

where $\sigma_* = \sigma_1 - \xi_1 u_1$. Thus estimates of σ_* and ξ are constant for all $u_1 > u_0$ if u_0 is a suitable threshold for the excesses to follow a GPD [13]. The estimates of σ_* and ξ will not be exactly constant but approximately due to sampling variability [4].

2.4 Declustering

One of the shortcomings of the POT method as compared to block maxima is that it is prone to producing dependent data particularly when dealing with time series data [8, 20, 23]. Time series data are known to be strongly autocorrelated hence a naïve selection of exceedances above a given threshold may lead to events that are dependent [8, 11, 17, 23]. In order to deal with this problem of clustering of neighbouring events a technique called declustering is used to achieve independence [11, 17]. In [17] a function called *clust* in R package is used to identify exceedances over a fixed threshold while meeting the independence criteria using two arguments: the threshold and a time condition (*tim.cond*). In [17] clusters are identified using the following procedure:

1. The first exceedance initiates the first cluster;
2. The first observation below the threshold ends the cluster unless *tim.cond* does not hold;
3. The next exceedance which holds *tim.cond* initiates a new cluster;
4. The process is iterated as needed.

In all declustering procedures, the main purpose is to identify cluster maxima. Variations usually appear in the choice of the time condition. In [17] two flood events are considered to be independent if they do not lie within a window period of 8 days. In [12] and [23] two flood events are considered to be independent if they are a day (24 hours) apart, that is, the POT values a day prior to and after the peak rainfall event are removed from the data set. For example, if a peak rainfall value (or flood height) in a cluster is selected for 5 December 2015 then rainfall values over the threshold on 4 December 2015 and 6 December 2015 are not considered in the POT data set. In [8] the Ferro-Segers declustering involves the estimation of the extremal index, γ, and the method proposes an automatic selection of the run-length auxiliary parameter, r, used to identify independent clusters. In [8], the exceedance times consist of two groups: one corresponding of inter-cluster times and the other corresponding to intra-cluster times. Based on asymptotic theory [8] postulates that the $1 - \gamma$ proportion of the smallest interexceedance times belong to the intra-cluster times, and the rest belong to the inter-cluster times. Therefore, given m sorted interexceedance times, we can take the $(\lfloor m\gamma \rfloor + 1)^{th}$ interexceedance time as the smallest interexceedance time that separates the clusters. Declustering proceeds with $r = \lfloor m\gamma \rfloor + 1$. In this paper we use the declustering approach based on [8] and programmed in R by [20]. The declustering results for the three sites are presented in Figures 3, 7 and 11 for Chokwe, Combomune and Sicacate, respectively.

2.5 GPD models

Consider the GPD model in Eqn. (2) for $\xi \neq 0$, that is, $G(y) = 1 - (1 + \frac{\xi}{\sigma}y)_+^{-1/\xi}$. Let this model be M_0.

In this present study we also propose one more model M_1. The model M_1 has a linear trend in the scale parameter such that $\log \sigma(t) = \sigma_0 + \sigma_1 \mathrm{t}$ and $\xi(t) = \xi$, and hence model M_1 and its log-likelihood are of the form $G(\sigma(t), \xi; y, t)$ and $l(\sigma_0, \sigma_1, \xi; x, t)$, respectively. In its general form, the nonstationary model, M_1, is given by Eqn. (6):

$$G(\sigma(t), \xi; y, t) = 1 - \left(1 + \frac{\xi y}{\exp(\sigma_0 + \sigma_1 t)}\right)_+^{-1/\xi}, \quad \text{for } \xi \neq 0,\ 0 \leq y \leq x_F - u, \tag{6}$$

2.6 Model choice

An important question to answer is whether the nonstationary model is valid, i.e. is it worthwhile to have the nonstationary model? This is equivalent to testing whether

the nonstationary model provides an improvement in fit over the time-homogeneous (usually simpler) model M_0. The ML estimation of nested models uses a simple procedure called the deviance (D) statistic to compare one model against the other [4]. In this study the time-homogeneous GPD model, M_0, is a special case of the time-dependent model M_1. In general, consider $M_0 \subset M_1$, then we define deviance statistic, D, as in Eqn. (7):

$$\mathrm{D} = 2\left\{l_1\left(M_1\right) - l_0\left(M_0\right)\right\}, \tag{7}$$

where $l_1\left(M_1\right)$ and $l_0\left(M_0\right)$ are the maximised negative log-likelihood (NLLH) values for model M_1 and M_0, respectively [4]. D has a Chi-square, $\chi^2_{k,\alpha}$, asymptotic distribution with k degrees of freedom tested at α (=0.05 or 5%) level of significance, where k is the difference in dimensionality (or difference in number of parameters) of M_1 and M_0. Thus, D is compared to critical values of $\chi^2_{k,\alpha}$ where $D > \chi^2_{k,\alpha}$ suggests that model M_1 explains substantially more of the variability in the data than M_0.

3. RESULTS AND DISCUSSION

This section presents the results of the study as well as discussing the results. The results in this section were obtained using R statistical programming package and R Studio [18]. Results for the time-homogeneous GPD model (M_0) and time-dependent GPD model (M_1) are presented in Tables 1, 2 and 3 for Chokwe, Combomune and Sicacate, respectively. The ML method was used to estimate the parameters of all GPD models.

3.1 Chokwe models

The time series plot for Chokwe shows that, with the exception of a very rare extreme 13 m flood height, the majority of flood heights at Chokwe are below 9 m (Figure 1a). The mean residual plot and the threshold choice plots (Figure 2) were used to come up with a reasonably high threshold of 4.8 m for Chokwe hydrometric station. The threshold of 4.8 m was chosen in order to meet the requirements of the bias-variance threshold trade-off balance such that it is high enough for the asymptotic theorem to be considered accurate and low enough to have sufficient data to estimate the GPD parameters.

The shape parameter ξ is significantly different from zero (p-value < 0.0001) for both the time-homogeneous and nonstationary GPD models (Table 1), suggesting that the distribution of exceedances over the 4.8 m threshold at Chokwe is short-tailed (negative Weibull) and does not come from a Gumbel (exponential) distribution family. The D statistic value for model pair (M_0, M_1) in Table 1 is $2\ (1803.637 - 1794.366) = 18.542$ and the critical value for the pair is $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test from the NLLH values in Table 1 for the test of $\sigma_1 = 0$ is highly significant at 5% level of significance (t-ratio = 3.067, p-value < 0.005). These results show that the nonstationary GPD model, M_1, is both significant and worthwhile over the time-homogeneous GPD model, M_0, in fitting the daily flood heights at Chokwe. These findings suggest that the nonstationary GPD model with a linear trend in the scale parameter provides an improvement in fit to the daily flood heights at Chokwe over the time-homogeneous GPD model since the D statistic value of 18.542 (>3.841) is significantly large. The residual probability plot for the nonstationary GPD model suggests a good fit to the data at Chokwe (Figure 5). On the contrary, however, the residual quantile plot shows a poor fit towards high quantiles indicating that extremely high flood heights at Chokwe may not be adequately modelled by the nonstationary model (Figure 5). The failure of the nonstationary GPD to model extremely high quantiles (flood heights) at Chokwe such as the 13 m flood height of the year 2000 is also highlighted in the time-homogeneous GPD model at the site (Figure 4). The return level plot based on the time-homogeneous GPD model for Chokwe reveals that the 13 m flood height which occurred in the year 2000 has a return period, on average, in excess of 1000 years (Figure 4) which appears to be a ridiculously high return period. In general, the residual diagnostic plots (Figures 4 and 5) suggest that the GPD models, both stationary and nonstationary, may not be quite suitable to model extreme floods at Chokwe implying that an alternative distribution may be necessary [14].

Table 1. Parameter estimates and negative log likelihood (NLLH) of the GPD models for Chokwe (1951–2010)

Model	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\xi}$	**NLLH**
M_0	1.4478677	0	-0.1628891	1803.637
M_1	0.2528215	0.0000059	-0.1940787	1794.366

Despite the shortcomings in the GPD models for Chokwe presented in this study, there is strong evidence that the nonstationary GPD model outperforms the time-homogeneous GPD model. Therefore, based on the results of this study, the proposed model for Chokwe is the nonstationary GPD model with a linear trend in the scale parameter. The general nonstationary GPD model for Chokwe is given in Eqn. (8):

$$G\left(\sigma(t), \xi; y, t\right) = 1 - \left(1 + \frac{-0.1940787 y_i}{\exp\left(0.2528215 + 0.0000059 t_i\right)}\right)^{1/0.1949787}, \quad \text{for } \xi < 0,\ 0 \le y_i < x_F - u, \tag{8}$$

where $y = y_{i,\forall i=1,2,\ldots} = x_i - 4.8$ are the excesses over the 4.8 m threshold, x_i is the daily flood height, and t_i is time such that $t_i = 1, 2, ..., 56058, ...$ where 56058 is the time for the last observed flood height value over the period 19 June 1951 to 31 August 2010. Note that the daily flood height data

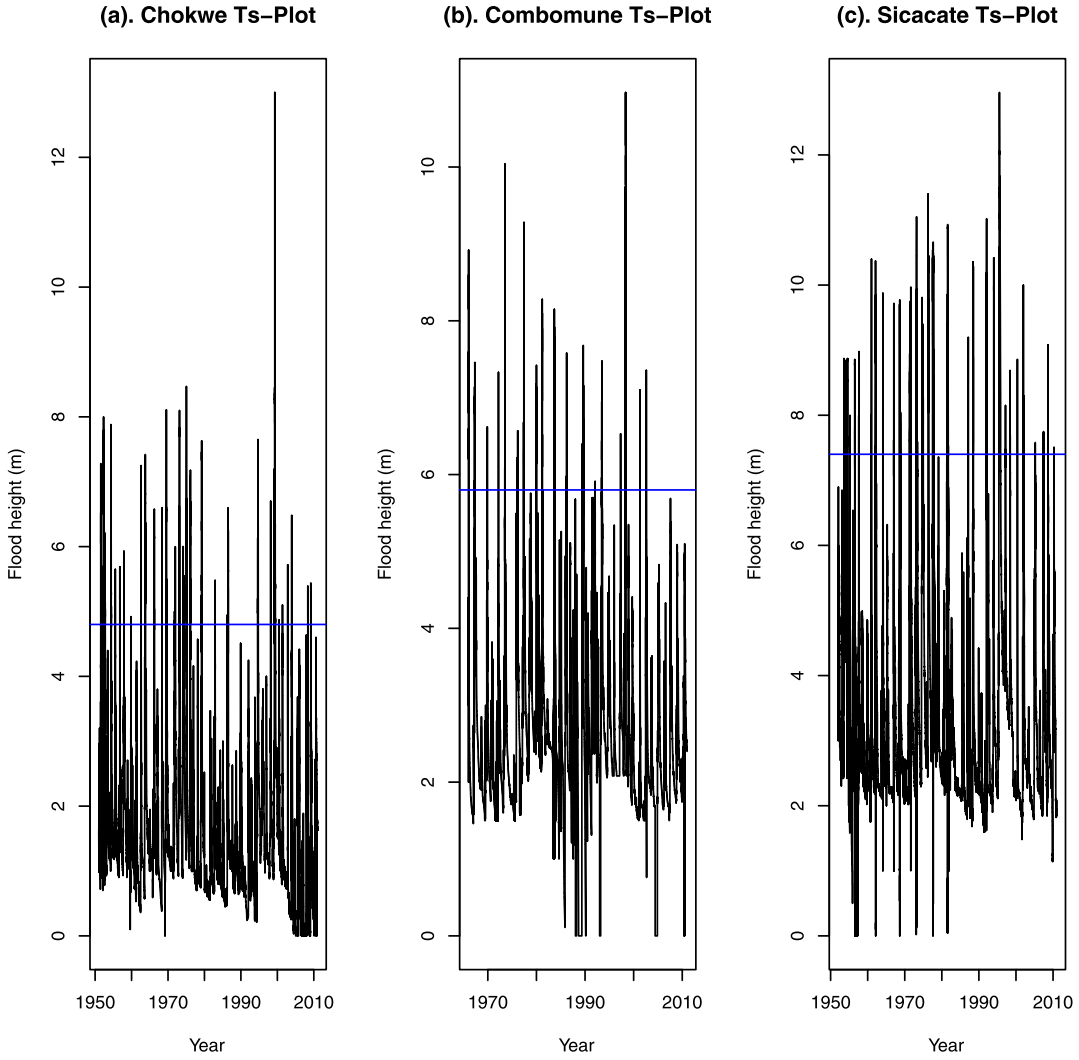

Figure 1. Time series (Ts) plots of the three sites: Panel (a). Chokwe (1951–2010), showing excesses over a 4.8 m threshold; (b). Combomune (1966–2010), showing excesses over a 5.8 m threshold; (c). Sicacate (1952–2010), showing excesses over a 7.4 m threshold.

were recorded three times a day, i.e. morning, afternoon and evening meaning that each day has 3 values of t_i at Chokwe hydrometric station.

3.2 Combomune models

The time series plot for Combomune shows that the majority of flood heights are below 10 metres except for a few rare extreme flood heights of about 11 m (Figure 1b). The mean residual plot and the threshold choice plots (Figure 6) were used to come up with a reasonably high threshold of 5.8 m for Combomune hydrometric station which was chosen in such a way that it is high enough for the asymptotic theorem to be considered accurate and low enough to have sufficient data to estimate the GPD parameters.

The shape parameter ξ is significantly different from zero (p-value < 0.0001) for both the time-homogeneous and non-stationary GPD models (Table 2), suggesting that the distribution of exceedances over the 5.8 m threshold at Combomune is short-tailed (negative Weibull) and does not come from a Gumbel (exponential) distribution family. The D statistic value for the model pair (M_0, M_1) in Table 2 is 33.512 which is too high compared to the critical value of $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test from the NLLH values in Table 2 for the test of $\sigma_1 = 0$ is highly significant at 5% level of significance (t-ratio = 9.481, p-value <

Table 2. Parameter estimates and negative log likelihood (NLLH) of the GPD models for Combomune (1966–2010)

Model	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\xi}$	**NLLH**
M_0	1.3974615	0	-0.1785666	635.826
M_1	0.0482690	0.0000188	-0.2254202	619.070

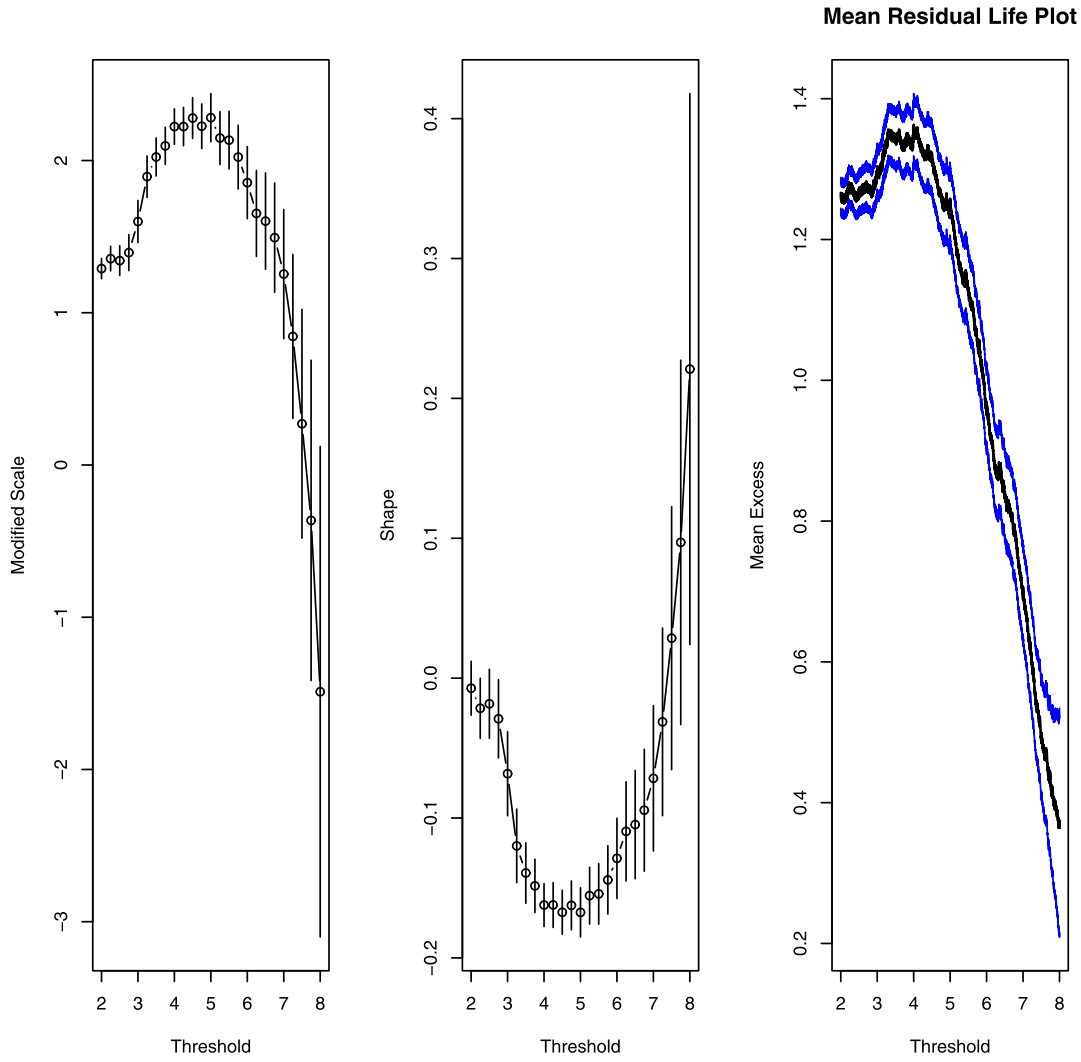

Figure 2. From left to right: Panel (a). First two plots: Threshold choice plots or parameter stability plots, (b). Mean residual life plot for the daily flood height data at Chokwe. Both panels for Chokwe show ML estimates and 95% confidence intervals for transformed parameters in Generalised Pareto model.

0.0001). These results show that the nonstationary GPD model, M_1, is both highly significant and worthwhile over the time-homogeneous GPD model, M_0, in fitting the daily flood heights at Combomune. This suggests that the nonstationary GPD model with a linear trend in scale parameter provides an improvement in fit to the daily flood heights at Combomune over the time-homogeneous GPD model since the D statistic value of 33.512 (>3.841) is significantly large. The residual diagnostic plots for the nonstationary GPD model suggest a good fit to the data (Figure 9). The residual diagnostics for the time-homogeneous model also suggest a good fit to the data (Figure 8). However, results in this study have revealed overwhelming evidence that the nonstationary GPD model outperforms the time-homogeneous GPD model and provides an improvement in fit over the time-homogeneous GPD model.

The proposed model for Combomune based on the findings of this study is the nonstationary GPD model with a linear trend in the scale parameter. The nonstationary GPD model for Combomune is given in Eqn. (9):

(9)

$$G\left(\sigma(t),\xi;y,t\right) = 1-\left(1+\frac{-0.2254202y_i}{\exp\left(0.0482690+0.0000188t_i\right)}\right)^{1/0.2254202},$$
$$\text{for } \xi<0,\ 0\le y_i<x_F-u,$$

where $y=y_{i,\forall i=1,2,\ldots}=x_i-5.8$ are the excesses over the 5.8 m threshold, x_i is the daily flood height, and t_i is time such that $t_i=1,2,...,37907,...$ where 37907 is the time for the last observed flood height value over the period 3 February

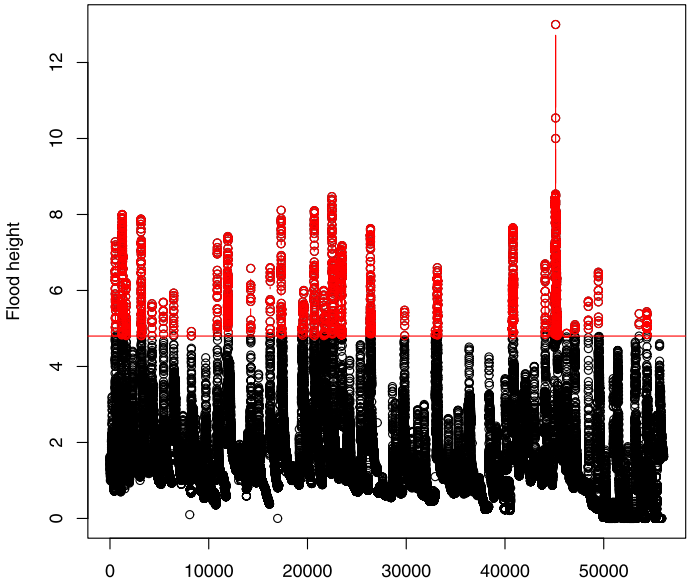

Figure 3. Chokwe declustered flood heights showing cluster maxima above a 4.8 m threshold.

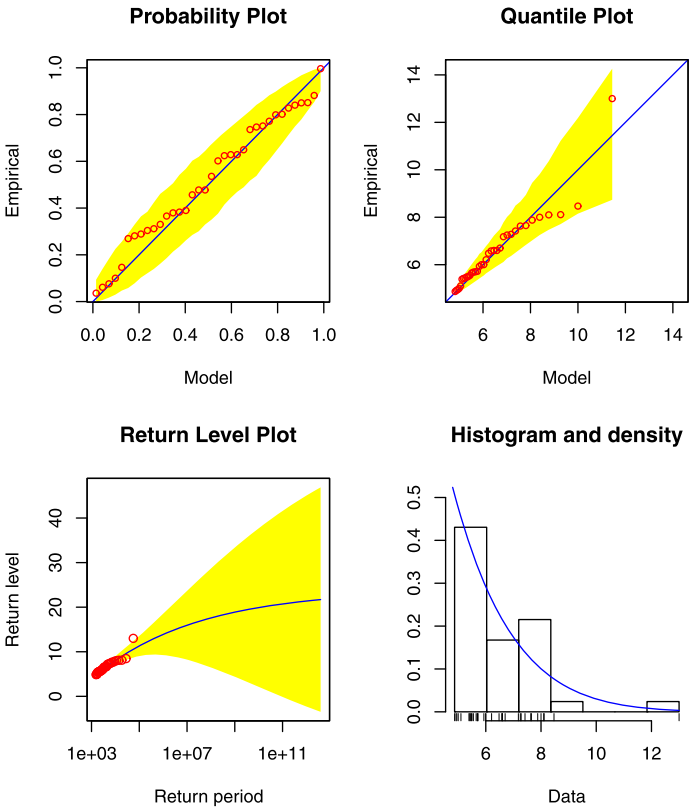

Figure 4. Time-homogeneous GPD diagnostic plots for Chokwe.

1966 to 31 August 2010. Also note that the daily flood height data were recorded three times a day, i.e. morning, afternoon

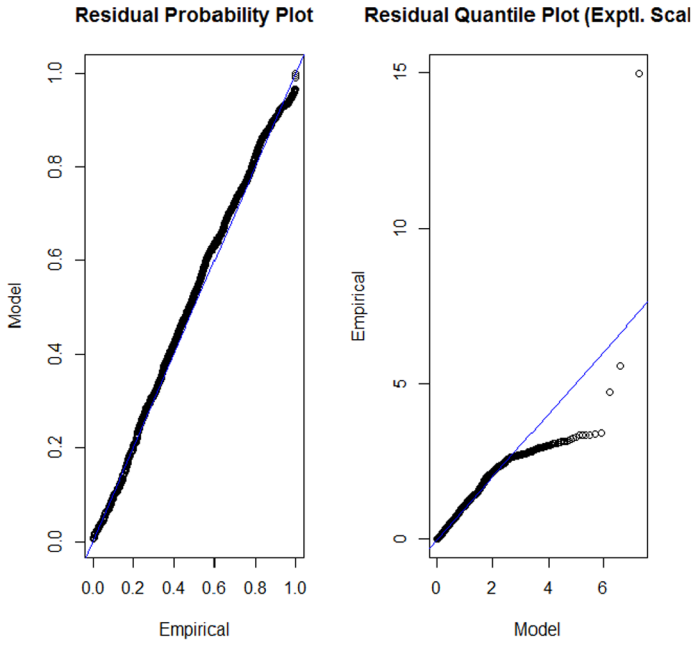

Figure 5. Nonstationary GPD diagnostic plots for Chokwe.

and evening, meaning that each day has 3 values of t_i at Combomune hydrometric station.

3.3 Sicacate models

The time series plot for Sicacate shows that the flood heights are generally high with quite a number of them above 10 m (Figure 1c). The graph also shows one extremely high flood height of about 13 m in magnitude. The mean residual life plot and the threshold choice plots (Figure 10) were used to come up with a reasonably high threshold of 7.4 m for Sicacate hydrometric station which was chosen to meet the bias-variance threshold trade-off balance such that it is high enough for the asymptotic theorem to be considered accurate and low enough to have sufficient data to estimate the GPD parameters.

The shape parameter ξ is significantly different from zero (p-value < 0.0001) for both the time-homogeneous and nonstationary GPD models (Table 3), suggesting that the distribution of exceedances over the 7.4 m threshold at Sicacate is short-tailed (negative Weibull) and does not come from a Gumbel (exponential) distribution family. The model pair (M_0, M_1) from Table 3 has a D statistic value of 360.042 and a critical value of $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test from the NLLH values in Table 3 for the test of $\sigma_1 = 0$ is highly significant at 5% level of significance (t-ratio $= 13.486$, p-value < 0.0001). These results reveal that the nonstationary GPD model, M_1, is both highly significant and worthwhile over the time-homogeneous GPD model, M_0, in fitting the daily flood heights at Sicacate. These findings suggest that the nonstationary GPD model with a linear trend in the scale parameter provides an improvement in fit to the daily flood heights at Sicacate over the

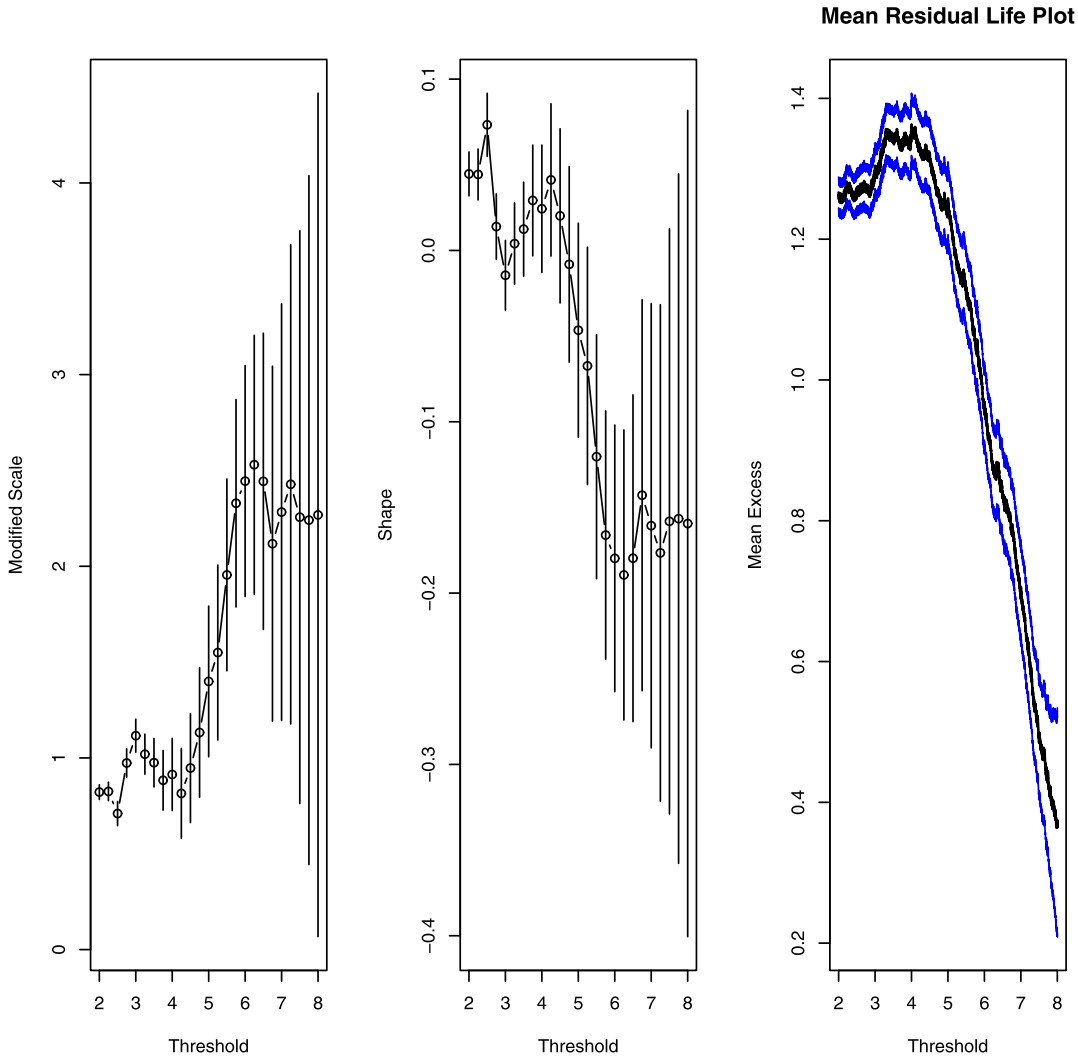

Figure 6. From left to right: Panel (a). First two plots: Threshold choice plots or parameter stability plots, (b). Mean residual life plot for the daily flood height data at Combomune. Both panels for Combomune show ML estimates and 95% confidence intervals for transformed parameters in Generalised Pareto model.

Table 3. Parameter estimates and negative log likelihood (NLLH) of the GPD models for Sicacate (1952–2010)

Model	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\xi}$	**NLLH**
M_0	2.3711619	0	-0.4105294	2702.413
M_1	0.4627278	0.0000261	-0.6105592	2522.392

time-homogeneous GPD model since the D statistic value of 360.042 (>3.841) is significantly large. The residual diagnostic plots for both the time-homogeneous and the nonstationary GPD models suggest a good fit to the data (Figures 12 and 13). It is clear that results of the diagnostic plots from both Figures 12 and 13 suggest a very good fit of the GPD to the data. This may imply that both models can be recommended for modelling the daily flood heights at Sicacate. Nevertheless, overwhelming evidence from the analytical goodness of fit tests suggest that the nonstationary GPD model is more appropriate at the site and is worth proposing because it adds more information in fit over the time-homogeneous model.

The proposed model for Sicacate based on the findings is the nonstationary GPD model with a linear trend in the scale parameter. The nonstationary GPD model for Sicacate is given in Eqn. (10):

(10)

$$G\left(\sigma(t),\xi;y,t\right) = 1-\left(1+\frac{-0.6105592y_i}{\exp\left(0.4627278+0.0000261t_i\right)}\right)^{1/0.6105592},$$
$$\text{for } \xi<0,\ 0\le y_i<x_F-u,$$

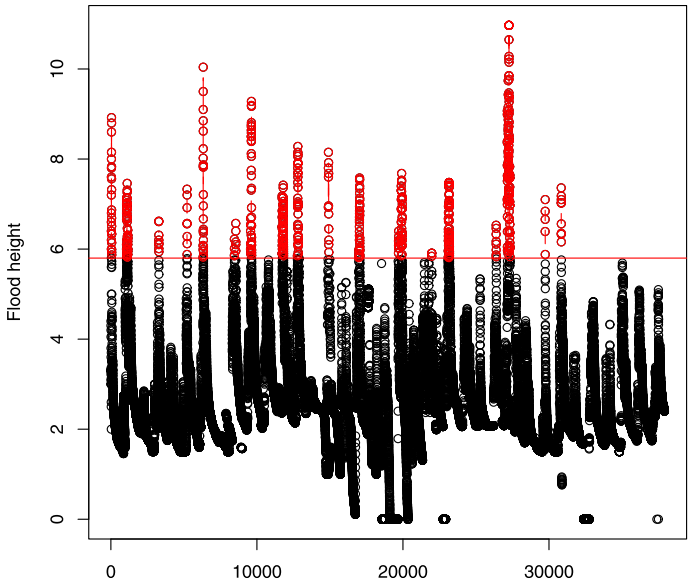

Figure 7. Combomune declustered flood heights showing cluster maxima above a 5.8 m threshold.

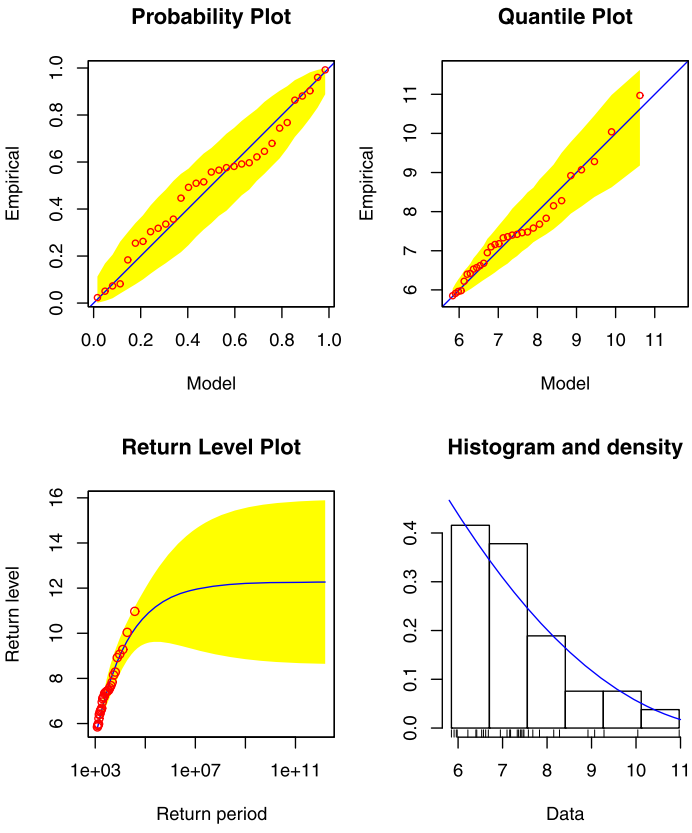

Figure 8. Time-homogeneous GPD diagnostic plots for Combomune.

where $y = y_{i,\forall i=1,2,...} = x_i - 7.4$ are the excesses over the 7.4 m threshold, x_i is the daily flood height, and t_i is time such

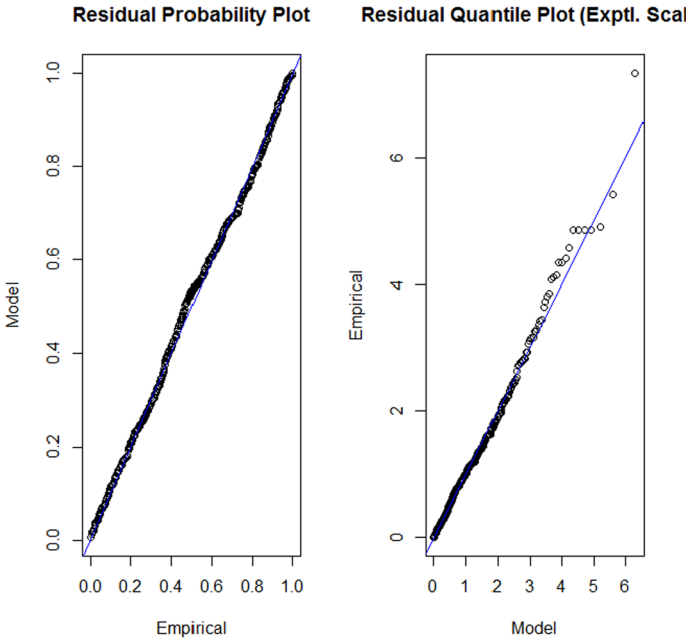

Figure 9. Nonstationary GPD diagnostic plots for Combomune.

that $t_i = 1, 2, ..., 40396, ...$ where 40396 is the time for the last observed flood height value over the period 16 December 1952 to 31 October 2010. Note that the daily flood height data were recorded three times a day, i.e. morning, afternoon and evening, meaning that each day has 3 values of t_i at Sicacate hydrometric station.

3.4 Overall discussion

The data series at all the three sites considered in this study had some missing values in-between the years during the period considered for the study. However, there are two main reasons to be content with the data used to develop the models in this study: (1). In most years where there are missing values they appeared during the winter period when the flood heights are generally low and would have likely missed out on the high threshold even if they were recorded, (2). The fact that the data was recorded three times a day means we have more data than we would have expected if the data was recorded once a day, for instance, the number of exceedances above high thresholds were 1 494, 550 and 1 860 for Chokwe, Combomune and Sicacate, respectively. The percentages of the number of exceedances over a prescribed threshold compared to the total number of observations at each site were 2.7%, 1.5% and 4.6% for Chokwe, Combomune and Sicacate, respectively. According to [7], the POT method is more efficient if the number of exceedances is much larger than the number of blocks. In our case there are 60, 45 and 59 blocks for Chokwe, Combomune and Sicacate which means that the number of exceedances at each site is 24.9, 12.2 and 31.5 times the number of blocks,

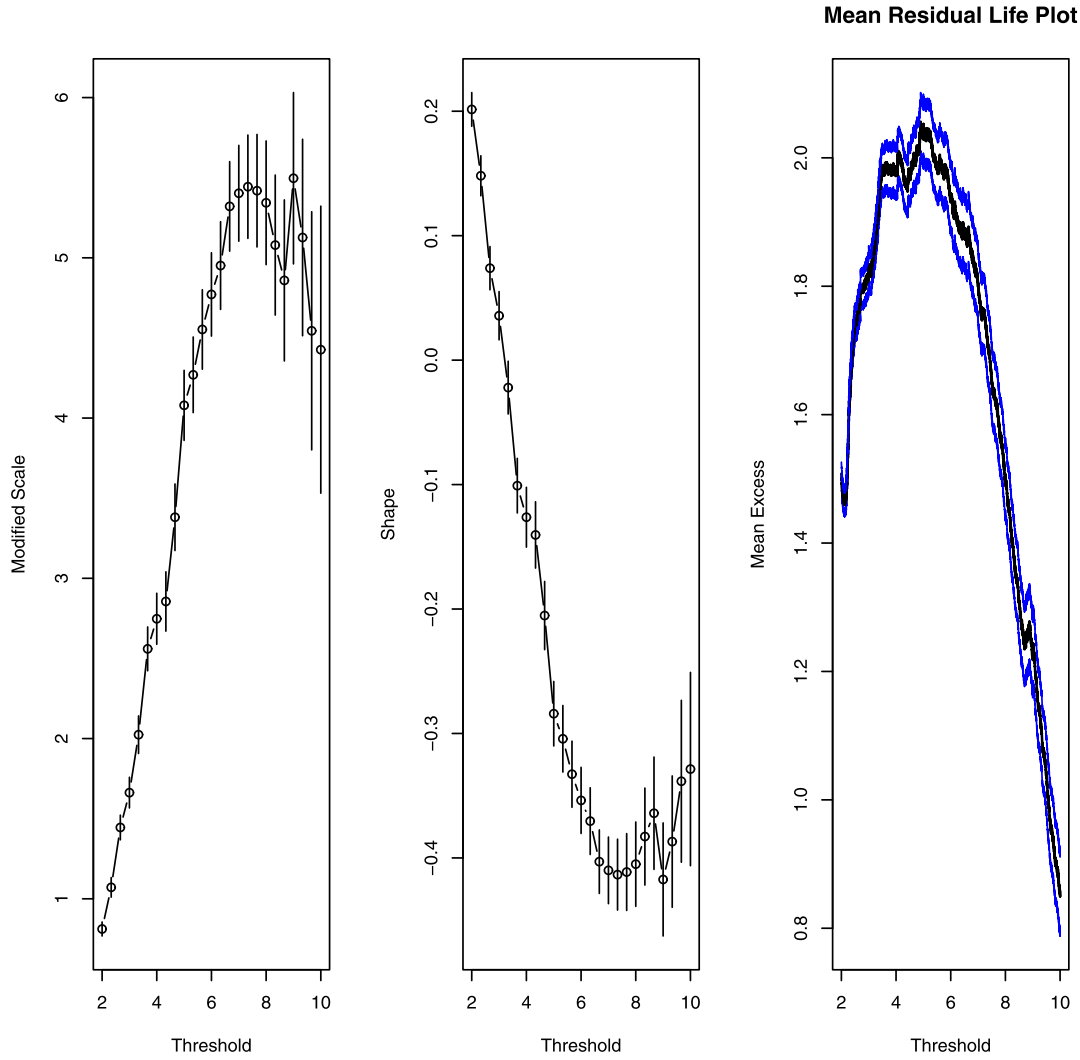

Figure 10. From left to right: Panel (a). First two plots: Threshold choice plots or parameter stability plots, (b). Mean residual life plot for the daily flood height data at Sicacate. Both panels for Sicacate show ML estimates and 95% confidence intervals for transformed parameters in Generalised Pareto model.

respectively. Literature states that the estimation of high quantiles becomes better when the number of exceedances is 1.65 times the number of blocks, particularly when using maximum likelihood estimators for zero shape ($\xi = 0$) parameters [5, 7]. Although this study has revealed that the shape parameter is non-zero ($\xi \neq 0$) at all the three sites, the number of exceedances at all the sites is comparatively higher than 1.65 times the number of blocks. All facts pertaining to the data and the findings in this study suggest that the models developed based on this data can be relied upon.

The findings in this study can be useful to help the lower Limpopo River basin community prepare and protect itself from future disastrous extreme floods. The Limpopo River basin is very important to the economy of Mozambique because it houses the largest irrigation scheme in the country, Chokwe Irrigation Scheme. The basin also forms the backbone of the economy of the country in terms of agriculture as most of the agricultural activities in the country such as rice production are done in the basin mainly in Chokwe district. This implies that a single disastrous extreme flood such as the one that occurred in the basin in the year 2000 may bring the economy of the country to its knees. The major highlights of this paper are in the application of statistics of extremes methods to large volumes of existing data that is untapped in the basin in order to complement the existing methods in the basin used to control and reduce flood disasters.

In a separate study in the basin a GEV distribution estimated by the maximum likelihood method was fitted to block maxima data for the same sites Chokwe, Combomune and Sicacate, and it was found that the distribution of an-

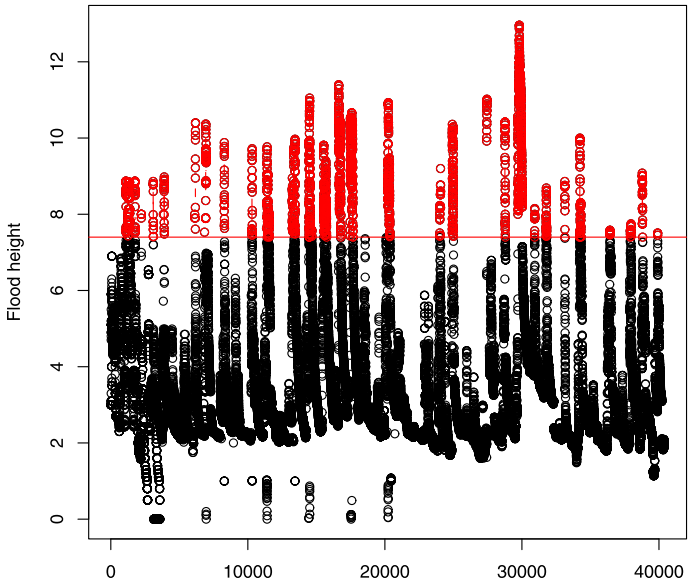

Figure 11. Sicacate declustered flood heights showing cluster maxima above a 7.4 m threshold.

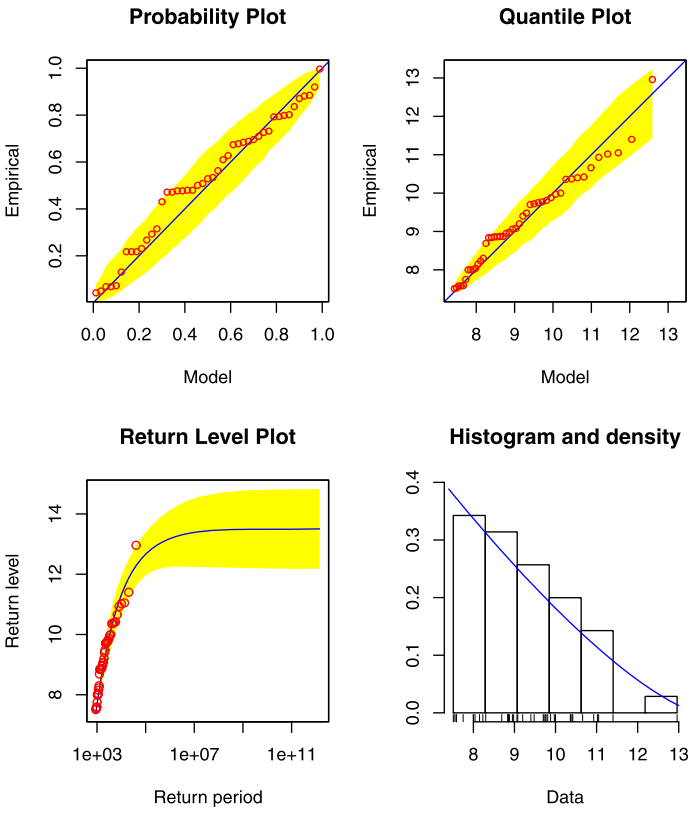

Figure 12. Time-homogeneous GPD diagnostic plots for Sicacate.

nual daily maximum flood heights at Combomune could be modelled by a nonstationary GEV distribution with a linear

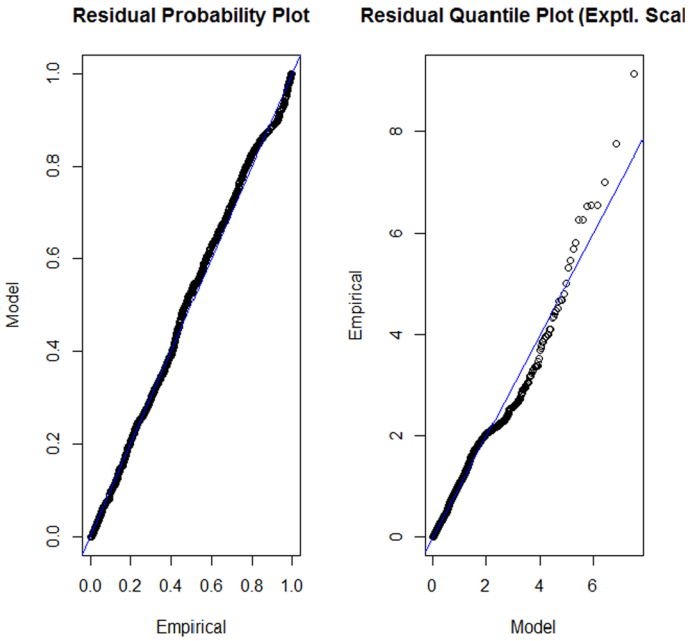

Figure 13. Nonstationary GPD diagnostic plots for Sicacate.

trend in the scale parameter, while for Sicacate the proposed GEV model had a linear trend in both location and scale parameters, whereas no evidence of a linear trend in the GEV model was found at Chokwe (Maposa, Cochran & Lesaoana, unpublished observations). The present study found similar results for Combomune and Sicacate except for the additional presence of a linear trend in the location parameter at Sicacate. The major differences were found at Chokwe where the present study found a linear trend in the scale parameter of the GPD model which is completely different from the time-homogeneous GEV model recommended based on block maxima data.

The findings in this study concur with the results of [1] which found climate variability to have a very big impact on the Limpopo River basin streamflow. This work advances the work of [1] through incorporating the dominant impact of climate variability in the basin into time-heterogeneous GPD extreme value models.

4. CONCLUSION

Statistics of extremes in a changing climate is considered for the lower Limpopo River basin of Mozambique at three sites in an attempt to develop future flood trends for an area that has not been deeply studied in Southern Africa. This is the first time climate change extreme value statistics models are applied to the data in the basin. It is hoped that the findings in this study will contribute towards decision making in the basin and help reduce the impact of floods on humans and properties, as well as reduce the amount of aid money required for disaster recovery and rehabilitation assistance in the basin.

The findings in this study revealed a very strong impact of climate change in the basin which can be modelled by a nonstationary GPD model with a linear trend in the scale parameter. The time-heterogeneous GPD models outperformed the time-homogeneous GPD models at all the three sites suggesting that the nonstationary GPD models are worthwhile and provide an improvement in fit over the time-homogeneous GPD models. This improvement in fit is very important for the planning and policy-making of the government of Mozambique and its partners in the lower Limpopo River basin, where the largest irrigation scheme of the country is situated. The developed time-dependent GPD models would also likely produce more reliable estimates in the frequency of floods since the new models in the basin take into account of the trend in the scale parameter.

Future research will attempt to advance this study to consider Bayesian inference and Markov chain Monte Carlo methods in a changing climate for the lower Limpopo River basin of Mozambique. Covariates in the form of cycles and/or a physical variable such as a dummy variable indicating the occurrence of cyclones in the region will also be considered in future studies involving statistics of extremes in a changing climate.

AUTHORS' CONTRIBUTIONS

D. M. (University of Limpopo) drafted the original manuscript, acquired and analysed the data and made interpretations. J. J. C. (University of Alabama) critically revised the original manuscript and made final approval of the version to be published. M. L. (University of Limpopo) critically revised the original manuscript and made final approval of the manuscript to be published.

ACKNOWLEDGEMENTS

We thank the Mozambique National Directorate of Water (NAM) and Mr. Isac Filimone of NAM, in particular, who provided us with all the necessary data used in this study. We are also indebted to United Nations Office for the Coordination of Humanitarian Affairs-Southern Africa (OCHA) for providing us with weekly update reports of floods in Southern Africa, particularly for the lower Limpopo River basin of Mozambique. We are also greatly indebted to the Department of Science and Technology – National Research Foundation (DST-NRF) Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) of South Africa who provided funds for the postgraduate studies. Lastly, we thank the University of Limpopo for their support in research.

Received 14 April 2015

REFERENCES

[1] Aich, V., Liersch, S., Vetter, T., Huang, S., Tecklenburg, J., Hoffmann, P., Koch, H., Müller, N. and Hattermann, F. F. (2014). Comparing impacts of climate change on streamflow in four large African river basins, *Hydrology and Earth System Sciences*, **18**, 1305–1321, doi:10.5194/hess-18-1305-2014.

[2] Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *Annals of Probability*, **2**, 792–894. MR0359049

[3] Coles, S. and Davison, A. (2008). *Statistical modelling of extreme values. Based on 'An introduction to statistical modelling of extreme values', by Stuart Coles, Springer, 2001.* Copyright 2008. http://stat.epfl.ch (last access: 13 October 2015).

[4] Coles, S. (2001). *An introduction to statistical modelling of extreme values*. Springer-Verlag, London. MR1932132

[5] Cunnane, C. (1973). A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of Hydrology*, **18**, 257–271.

[6] De Haan, L. and Ferreira, A. (2006). *Extreme value theory: An introduction*. Springer, New York. MR2234156

[7] Ferreira, A. and De Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, **43**(1), 276–298, doi:10.1214/14-AOS1280. MR3285607

[8] Ferro, C. A. T. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**, 545–556. MR1983763

[9] Fischer, S. and Schumann, A. (2014). *Comparison between classical annual maxima and peak over threshold approach concerning robustness*. SFB 823 Discussion Paper Nr. 26/2014.

[10] IPCC (2012). *Managing the risks of extreme events and disasters to advance climate change adaption (SREX)*. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change, Cambridge University Press, UK and USA, 582 pp.

[11] Hamidieh, K. (2008). Topics in statistical modelling and estimation of extremes and their dependence. PhD Thesis, University of Michigan, UMI Number 3343082. MR2712736

[12] Jakob, D., Karoly, D. J. and Seed, A. (2011a). Non-stationary in daily and sub-daily intense rainfall – Part 1: Sydney, Australia. *Natural Hazards and Earth System Sciences*, **11**, 2263–2271.

[13] Magadia, J. (2010). *Value-at-Risk modelling via the peaks-over-threshold approach*. Annual BSP-UP Professorial Chair Lectures, 15–17 February 2010, Bangko Sentral ng Pilipinas, Malate, Manila.

[14] Maposa, D., Cochran, J. J. and Lesaoana, M. (2014a). Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme floods in the lower Limpopo River basin, Mozambique. *Journal of Statistics and Management Systems*, **17**(3), 265–283, doi:10.1080/09720510.2014.927602.

[15] Maposa, D., Cochran, J. J., Lesaoana, M. and Sigauke, C. (2014b). Estimating high quantiles of extreme floods in the lower Limpopo River of Mozambique using model based Bayesian approach. *Natural Hazards and Earth System Sciences Discussions*, **2**, 5401–5425, doi:10.5194/nhessd-2-5401-2014.

[16] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, **3**, 119–131. MR0423667

[17] Ribatet, M. A. (2006). *A user's guide to the POT package*. Version 1.0. http://cran-r-project.org/ (last access: 25 November 2015).

[18] R Core Team (2013). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/ (last access: 27 March 2015).

[19] Saidi, H., Ciampittiello, M., Dresti, C. and Ghiglieri, G. (2013). Observed variability and trends in extreme rainfall indices and peaks-over-threshold series. *Hydrology and Earth System Sciences Discussion*, **10**, 6049–6079, doi:10.5194/hessd-10-6049-2013.

[20] Southworth, H. and Heffernan, J. E. (2013). *texmex: Statistical modelling of extreme values*. R package version 2.1.

[21] Towler, E., Rajagopalan, B., Gilleland, E., Summers, R. S. and Yates, D. (2010). Modelling hydrologic and water extremes in a changing climate: A statistical approach based on ex-

treme value theory. *Water Resources Research*, **46**, W11504, doi:10.1029/2009WR008876.

[22] Velasco, M., Versini, P. A., Cabello, A. and Berrera-Escoda, A. (2013). Assessment of flash floods taking into account climate change scenarios in the Llobregat River basin. *Natural Hazards and Earth System Sciences*, **13**, 3145-3156, doi:10.5194/nhess-13-3145–2013.

[23] Yilmaz, A. G., Hossain, I. and Perera, B. J. C. (2014). Effect of climate change and variability on extreme rainfall intensity-frequency-duration relationships: A case study of Melbourne. *Hydrology and Earth System Sciences*, **18**, 4065–4076.

Daniel Maposa
University of Limpopo
Sovenga
South Africa
E-mail address: Daniel.maposa@ul.ac.za
E-mail address: danmaposa@gmail.com

James J. Cochran
University of Alabama
Tuscaloosa, Alabama
United States of America
E-mail address: jcochran@cba.ua.edu

'Maseka Lesaoana
University of Limpopo
Sovenga
South Africa
E-mail address: Lesaoana.maseka@ul.ac.za