An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales

Stelian Curceac[a,*], Peter M. Atkinson[b, c, d, e], Alice Milne[f], Lianhai Wu[a], Paul Harris[a]

[a] Rothamsted Research, Department of Sustainable Agriculture Sciences, North Wyke EX20 2SB, Devon, UK.

[b] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK.

[c] Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

[d] School of Geography, Archaeology and Palaeoecology, Queen's University Belfast, BT7 1NN, Northern Ireland, UK.

[e] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.

[f] Rothamsted Research, Department of Sustainable Agriculture Sciences, Harpenden AL5 2JQ, UK

[*]Corresponding author: stelian.curceac@rothamsted.ac.uk, (+44) 7375596777

Resubmitted to: Journal of Hydrology

21/01/2020

Abstract

This study investigated core components of an extreme value methodology for the estimation of high-flow frequencies from agricultural surface water run-off. The Generalized Pareto distribution (GPD) was used to model excesses in time-series data that resulted from the 'Peaks Over Threshold' (POT) method. First, the performance of eight different GPD parameter estimators was evaluated through a Monte Carlo experiment. Second, building on the estimator comparison, two existing automated GPD threshold selection methods were evaluated against a proposed approach that automates the threshold stability plots. For this second experiment, methods were applied to discharge measured at a highly-instrumented agricultural research facility in the UK. By averaging fine-resolution 15-minute data to hourly, 6-hourly and daily scales, we were also able to determine the effect of scale on threshold selection, as well as the performance of each method. The results demonstrate the advantages of the proposed threshold selection method over two commonly applied methods, while at the same time providing useful insights into the effect of the choice of the scale of measurement on threshold selection. The results can be generalized to similar water monitoring schemes and are important for improved characterizations of flood events and the design of associated disaster management protocols.

Keywords: Generalized Pareto Distribution; Peaks over threshold; Threshold selection; Flood Frequency Analysis; Scale effects; Grassland agriculture.

1. Introduction

The magnitude and frequency of floods is likely to increase as a result of climate change (Bates et al., 2008; Field et al., 2012; Kundzewicz et al., 2007) and this could push ecosystems beyond the threshold of normal disturbance resulting in negative impacts that may be irreversible (e.g. Thibault & Brown, 2008). Floods increase surface run-off, intensify erosion and introduce more soil, organic matter and pollutants into water courses. Floods in areas of steep and unstable slopes increase the possibility of landslides (Clarke & Rendell, 2006). Moreover, increased runoff and flooding generally result in higher sediments and nutrient losses that can lead to soil degradation (Bouraoui et al., 2004). They can have severe impacts on key ecosystem services, such as those of support (e.g. water, nutrient cycling and soil protection), regulation (e.g. climate) and culture (e.g. scenic recreation) (MA, 2005).

Flood Frequency Analysis (FFA) is a classic method to analyze the relationship between flood magnitude and the corresponding frequency of occurrence. Reliable estimation and prediction of high flow quantiles require extrapolation beyond the observed range of events, commonly using parametric probability distributions. There are two main approaches for defining extreme events in stationary time-series. The first is the block (usually annual) maxima (AM) method where the dataset is divided into contiguous blocks of equal size and the maximum values in each segment are considered. According to the Fisher-Tippet theorem (Fisher & Tippett, 1928), these identically, independently distributed (iid) random variables asymptotically follow a Generalized Extreme Value (GEV) distribution (Coles, 2001; Jenkinson, 1955). The second approach is known as the peaks-over threshold (POT) method, which considers the values X that exceed a fixed high threshold u. The distribution function of the excess values $X - u$, conditional on $X > u$, is a Generalized Pareto Distribution (GPD)

(Pickands, 1975). The case study we consider, contains six years of fine resolution (15-minute) flow measurements, which is insufficient for effective fitting of the GEV distribution. Therefore, only the POT method with the GDP was investigated.

The above two families of distributions have fundamental differences, but also theoretical links (see Langousis et al., 2016). The GEV distribution is usually best fitted to annual maxima samples and for this reason long historic records are required. This restriction does not apply to the POT method since it includes all the peaks above a certain threshold allowing for greater flexibility. The threshold must be large enough for the excesses to follow a GPD, but an over-estimated threshold leads to reduced sample size and increases the variance of the estimates. A smaller threshold increases the sample size but also the bias of the estimates as the empirical distribution deviates from a perfect GPD model (Scarrott and MacDonald, 2012). Clearly, GPD threshold selection is of key importance and there is no universally recognized best performing method although various techniques have been proposed (see e.g. Langousis et al. 2016 and Scarrott & MacDonald, 2012). Among them are probabilistic-based techniques (Beirlant et al., 1996, 2006; Choulakian & Stephens, 2001; Deidda & Puliga, 2006; Goegebeur et al., 2008; Hill, 1975), computational approaches (Beirlant et al., 2005; Danielsson et al. 2001; Hall, 1990; Thompson et al., 2009; Zoglat et al., 2014) and mixture models (Behrens et al., 2004; Eastoe & Tawn, 2010; Solari & Losada, 2012). Graphical methods (Das & Ghosh, 2013; Deidda, 2010; Lang et al., 1999; Tanaka & Takara, 2010), such as the Mean Residual Life (MRL) plot (Coles 2001; Beguería, 2005; Davison & Smith, 1990) are used commonly for the selection of an optimal threshold, but have been criticized for the difficulty and subjectivity of their interpretation (Scarrott & MacDonald 2012; Yang et al., 2018). Alternatively, analytical methods have the advantage that they can be automated, and the associated uncertainty can be quantified. Solari et al. (2017) proposed an automated threshold selection method based on AD goodness of fit test. The application of their technique on long records of precipitation and flow resulted in estimated thresholds that were within the stability regions of the shape and modified scale parameters. Durocher et al. (2018) compared several automatic methods and proposed a hybrid one where consistency with shape stability was found for most of the considered sites.

In this study, we propose an empirical automated method for threshold determination, based on threshold stability, which is evaluated against two commonly applied analytical methods, together with eight alternatives for GDP parameter estimation. Furthermore, by averaging the case study's 15-minute flow data to hourly, 6-hourly and daily supports, we determine the effects of temporal measurement scale on threshold selection, as well as the performance of each method.

The remainder of this paper is organized as follows. Section 2 presents the methods for GPD parameter estimation, two analytical threshold selection techniques, this study's proposed automated threshold stability method, and model evaluation diagnostics and indices. Section 3 describes the case study site and flow data, together with the simulation experiment design used to evaluate the performance of the different GDP parameter estimators. Results are presented in Section 4, which includes an investigation of scale effects through a series of flow data integrations. Sections 5 and 6 discuss and conclude the study, respectively.

2. Methodology

The cumulative distribution function (CDF) of the iid excesses over an appropriate threshold u for the GPD is:

$$G(x) = \Pr(X - u < x | X > u) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-\frac{1}{\xi}}, \xi \neq 0 \\ 1 - e^{\left(-\frac{x-u}{\sigma}\right)}, \quad \xi = 0 \end{cases}$$

where x, for this study, is the extreme flow in m^3s^{-1}, u is the location parameter, σ is the scale parameter and ξ is the shape parameter. The value of the shape parameter defines the type of distribution from the GPD family, that is, $\xi = 0$ refers to the exponential distribution, for $\xi > 0$ the corresponding distribution has a heavy upper tail that behaves like a power function with exponent $-1/\xi$ and for $\xi = 1$ the distribution is uniform. The Pareto distribution is obtained when $\xi < 0$.

2.1 GPD parameter estimators

The excesses above a suitable threshold are modelled by the GPD and the parameters of the distribution can be estimated by competing methods, where the Maximum Likelihood estimator (MLE) is the most commonly used (Prescott & Walden, 1980, 1983; Smith, 1985). Hosking and Wallis (1987) showed that MLE provides greater variance and bias for small samples compared to the Probability Weighted Moment (PWM) (Greenwood et al., 1979; Landwehr et al., 1979) and the Method of Moments (MOM) estimators. Coles and Dixon (1999) proposed a modified MLE which contains a penalty function for the shape parameter (i.e. the Maximum Penalized Likelihood estimator (MPLE). Zhang (2007) presented a hybrid Likelihood Moment estimator (LME) which provides feasible estimates and has high asymptotic efficiency. All of these methods are evaluated in this study, together with that suggested by Pickands (1975) and a maximum goodness-of-fit (MGF) estimator (e.g. Luceño, 2006). Estimator performance has been found to depend significantly on sample size and the value of the GPD shape parameter (Ashkar & Tatsambon, 2007; de Zea Bermudez & Kotz, 2010; Hosking & Wallis, 1987), and the choice of the estimator should be made based on the specifics of the situation. The equations for the above estimators can be found in Appendix A: Equations of the estimators.

2.2 Threshold selection methods

The selection of the threshold u is a crucial step in GDP extreme value analysis. On the one hand, a small threshold results in a large sample that makes statistical inference more effective, but can lead to biased estimates due to deviations of the empirical distributions from the GPD model (e.g. Beirlant et al., 2005). On the other hand, when considering large thresholds and consequently small samples, parameter estimates have a smaller expected bias, but a larger variance that can be highly dependent on the estimation method. The two main approaches for threshold selection are graphical methods, such as the MRL plot, and analytical methods that can be automated.

An important assumption for the application of the POT method is that the extracted peaks are independent. A commonly applied method is to use no more than 2-3 peaks per year (Madsen et al., 1997; Todorovic, 1978) but it has been criticised for lack of flexibility. Another solution is to consider a minimum separation interval between successive peaks (Cunnane, 1979; Lang et al., 1999). This minimum separation interval accords to the scale and nature of the measured process, but for daily flow data, an interval of a few days commonly ensures that the peaks are generated from different events (Engeland et al., 2004). The autocorrelation function is a popular choice for the investigation of serial dependence in a time series. However, this approach assumes normally distributed variables, which is not the case for peak discharges, so other independence tests should be implemented (e.g. Ledford and Tawn, 2003; Reiss and Thomas, 2007). In this study, and through prior experimentation, maximum peaks separated by a minimum of three days were considered and their independence was tested using Kendall's τ test (Claps and Laio, 2003; Ferguson et al., 2000).

2.2.1 Graphical methods: MRL plots

The most popular graphical method is the MRL plot (Coles, 2001; Davison & Smith, 1990). If the scaled excesses $X_{u^*} = [X - u^* | X > u^*]$ above a threshold u^* are Generalized Pareto (GP) distributed, then for every $u \geq u^*$, the scaled excesses $X_u = [X - u | X > u]$ are similarly GP distributed with the same shape parameter ξ, a scale parameter $\sigma_u = \sigma_{u^*} + \xi(u - u^*)$ and a mean value:

$$\bar{X}(u) = E[X - u | X > u] = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u^*} + \xi(u - u^*)}{1 - \xi} = Au + B$$

where $A = \xi/(1 - \xi)$ and $B = (\sigma_{u^*} - \xi u^*)/(1 - \xi)$ are the respective slope and intercept of the linear relation. The sample estimates of the mean excesses are then plotted for different values of the threshold and the most appropriate is considered to be the one after which the mean excesses follow a straight line (e.g. Das & Ghosh, 2013).

Another graphical technique is to plot the estimated shape and/or modified scale parameters for different threshold candidates and select the one above which the estimates are constant (Brodin & Rootzén, 2009; Bommier, 2014; Sigauke & Bere, 2017). The main criticism of graphical methods is that the interpretation of the plot can be ambiguous or subjective as it is usually unclear which part of the curve is linear (Scarrott & MacDonald, 2012). In this respect, attempts have been made to automate (Langousis et al., 2016) and estimate the uncertainty (Liang et al., 2019) of the graphical methods.

2.2.2 Analytical methods: Square Error and Normality of Differences

The Square Error (SE) method was developed by Zoglat et al. (2014) following the work of Beirlant et al. (2005), and is implemented as follows. Let $u_1, u_2, \dots, u_n$ be n equally spaced increasing threshold candidates. For each of these thresholds, estimate the scale σ_{u_j} and shape ξ_{u_j} parameters for $j = 1, \dots, n$. Find N_{u_j} the exceedances that correspond to each threshold u_j and simulate m independent samples of size N_{u_j} from the GPD with parameters σ_{u_j} and ξ_{u_j}. For each probability $a \in A = \{0.05, 0.1, \dots, 0.95\}$ and each $i = 1, \dots, m$ calculate the quantiles q^i_{a,u_j} and compute $q^{sim}_{a,u_j} = \frac{1}{m}\sum_{i=1}^{m} q^i_{a,u_j}$. The optimal threshold is the one for which the square error $SE_{u_j} = \sum_{a\in A}\left(q^{sim}_{a,u_j}/q^{obs}_{a,u_j}\right)^2$ between the simulated and the observed quantiles is minimum. The selection of the threshold candidates u_j can be defined by the user or as an automated process. For example, the smallest threshold can be set as zero or the median and the maximum threshold set as a high percentile of the data.

An alternative analytical method for threshold selection was proposed by Thompson et al. (2009). Again, let $u_1, u_2, \dots, u_n$ be n equally spaced increasing threshold candidates. For the excesses above the threshold u_j, $\hat{\sigma}_{u_j}$ and $\hat{\xi}_{u_j}$ are the MLEs of the scale and shape parameters, respectively, for $j = 1, \dots, n$. If $u \le u_{j-1} < u_j$ is an appropriate threshold then according to Coles (2001), $\sigma_{u_{j-1}} = \sigma_u + \xi(u_{j-1} - u)$ and $\sigma_{u_j} = \sigma_u + \xi(u_j - u)$. Consequently, $\sigma_{u_j} - \sigma_{u_{j-1}} = \xi(u_j - u_{j-1})$ and from standard maximum likelihood theory we have that $E[\hat{\sigma}_{u_j}] \approx \sigma_{u_j}$ and $E\left[\hat{\xi}_{u_j}\right] = \xi$ for any j such that $u_j > u$. Respectively, $E\left[\tau_{u_j} - \tau_{u_{j-1}}\right] \approx 0$, $j = 2, \dots, n$ for $\tau_{u_j} = \hat{\sigma}_{u_j} - \hat{\xi}_{u_j} u_j$, $j = 1, \dots, n$. It follows that $\tau_{u_j} - \tau_{u_{j-1}}$ approximately follows a normal distribution. Thompson et al. (2009) suggest Pearson's Chi-square test to examine the null hypothesis of normality. However, this test has been criticised for having inferior power properties (Moore, 1986). For this reason, we also applied the Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov and Shapiro-Francia normality tests (Thode, 2002). Regardless of which of the five normality tests are used, we refer to this method as the 'Normality of Differences' method. According to this approach, a suitable threshold $u \leq u_{j-1} < u_j$ is the one for which all the differences $\tau_{u_j} - \tau_{u_{j-1}}$ are approximately normally distributed. We selected the appropriate threshold as the one for which the *p*-value of $\tau_{u_j} - \tau_{u_{j-1}}, j = 2, \dots, n$ is above 0.05. A smaller threshold would be selected for a smaller *p*-value (e.g. 0.01).

2.2.3 Proposed method based on Threshold Stability

For this study, we propose an automated threshold selection method based on stability plots (Coles, 2001; Scarrott & MacDonald 2012). If the GPD is an appropriate model for the excesses above a threshold u, then for all larger thresholds $u^* > u$ it will also be suitable with the shape parameter being relatively constant. In other words, it is the approximately linear horizontal part on the shape parameters versus thresholds plot. This does not apply for the scale parameter σ_{u^*}, as it changes with the threshold $\sigma_{u^*} = \sigma_u + \xi(u^* - u)$. However, the modified scale parameter $\sigma_1 = \sigma_{u^*} - \xi u$ remains relatively constant. Therefore, we fit a cubic smoothing spline to this plot and calculate the rate of change at each of m consecutive steps. The cubic smoothing spline estimate $\hat{f}$ of a function f in the model $Y_i = f(x_i) + \varepsilon_i$, is defined as the minimizer of $\sum_{i=1}^{n}\left\{Y_i - \hat{f}(x_i)\right\}^2 + \lambda \int \hat{f}''(x)^2 dx$, where λ is the smoothing parameter. The minimum change rate locates the part of the plot where the shape and the modified scale parameters reach a plateau.

A preliminary analysis showed that a smoothing parameter value of λ = 0.4 of the cubic spline function was the most appropriate to avoid both over- and under-fitting. A total of $n = 1000$ threshold candidates were used in each case and a cubic spline was fitted to the corresponding estimated shape and modified scale parameters. The numbers of the consecutive steps for which the minimum change rate was calculated, were $m = 25, 50, 75$ and 100 which corresponds to 2.5%, 5%, 7.5% and 10%, respectively, of the total number of fitted values, that is, the total threshold candidates n.

2.3 Evaluation procedure

Quantile-Quantile (Q-Q) plots are commonly used to investigate the efficiency of the statistical inference of the fitted GPD models. To quantify the difference between the theoretical and empirical quantiles for probabilities $a \in A = \{0.95, 0.951, \ldots, 0.999\}$, various error and agreement diagnostics were calculated. Specifically, we calculated the Mean Square Error (MSE) (e.g. Turan and Yurdusev, 2009), the Normalized Root Mean Square Error (NRMSE) (e.g. Sheta and El-Sherif, 1999) and the Relative Index of Agreement ($RD \in [0,1]$) (Krause et al., 2005; Willmott, 1981). For ideal model performance, both MSE and NRMSE should tend to zero, while RD should tend to unity. The NRMSE was obtained by dividing the root MSE with the difference between minimum and maximum values and, thus, was less sensitive to very large values and provided a more robust diagnostic than MSE.

3. Study site and datasets

3.1 Study site

Flow discharge data come from a single sub-catchment of the North Wyke Farm Platform (NWFP). The NWFP is a farm-scale experiment established in 2011 in the southwest of

England (50°46'10"N, 3°54'05"W) for research into sustainable grassland livestock systems (Orr et al., 2016; Takahashi et al., 2018). The platform is located at an altitude in the range of 120-180 m above sea level. The platform's fields have a declining slope at the west towards the River Taw and to the east, to one of its tributaries, the Cocktree stream. The soil texture consists of a slightly stony clay loam topsoil (approximately 36% clay) above a mottled stony clay (approximately 60% clay). The subsoil is impermeable to water and during rain events most of the excess water moves by surface and sub-surface lateral flow towards the drainage system described below.

Each of the 15 NWFP sub-catchments are hydrologically isolated through a combination of topography and a network of French drains (800 mm deep trenches), which ensure that the total runoff is channeled to instrumented flumes, measuring 15-minute water discharge and water chemistry from October 2012. The discharge from each sub-catchment is measured through a combination of primary and secondary flow devices (Liu et al., 2018). The primary devices are H-type flumes (TRACOM Inc., Georgia, USA) with capacity designed for a 1-in-50 year storm event. The specific design of the H-type flume facilitates the accurate measurement of both low and high flows and is relatively self-cleaning since it allows the ready passage of sediment and particulate matter. A secondary flow measurement device (OTT hydromet, Loveland, CO., USA) is used to measure the water height within the flume and convert it to discharge rate using flume-specific formulas which depend on water height. The flow is generated only from rainfall as the fields are not irrigated. At each sub-catchment, 15-minute precipitation and soil moisture are also monitored. (Figure 1).

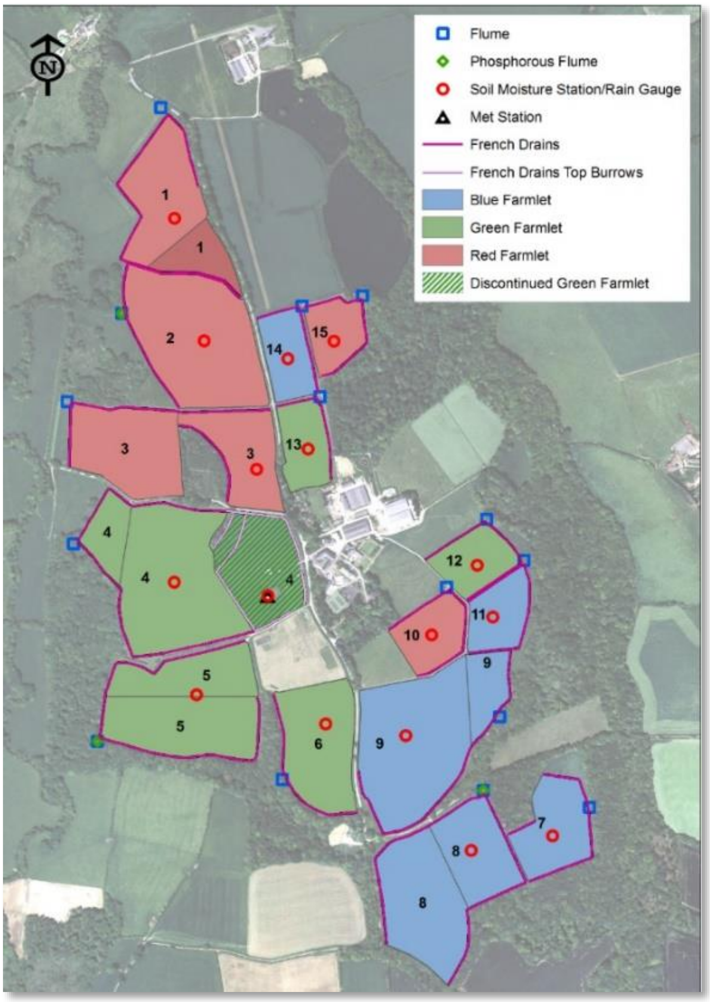


Figure 1: The three farmlets and the 15 sub-catchments of the North Wyke Farm Platform, with: (i) 'blue' farmlet a mixture of white clover and high sugar perennial ryegrass; (ii) 'red' farmlet high sugar perennial ryegrass only and (iii) 'green' farmlet permanent pasture ("business as usual").

3.2 Measured data

For this study, we used the flow discharge measured at sub-catchment 3 of the NWFP, which is part of the 'red' farmlet (Figure 1) and 6.84 ha in size. Given this is a methodological-based study, we chose to use data from this sub-catchment as it has one of the smallest number of missing values (approximately 1%) for the six-year period (2012-2018). Imputation of the missing values was performed using a regularized iterative Principal Components Analysis (PCA impute) model (Josse & Husson, 2013). The largest imputed value was approximately 20 l s^{-1} which is smaller than any threshold suggested (see below) and, therefore, is not considered as a peak flow and does not affect the subsequent analysis. It should be noted that, compared with measurements from many river or stream monitoring systems, the flow data (Figure 2) are highly discontinuous with many zeros, as non-zero measurements occur only after rainfall events.

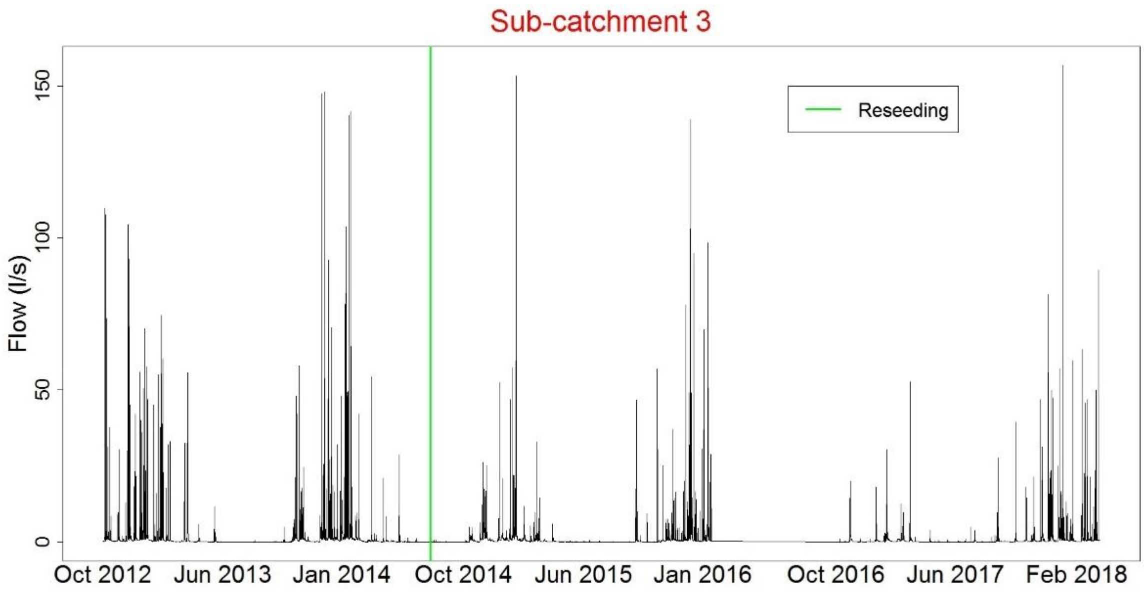


Figure 2: Flow (l s^{-1}) measurements at sub-catchment 3 (2012 to 2018).

3.3 Simulated data

As a precursor to the empirical study, the performance of the eight GDP parameter estimators was assessed through a Monte Carlo experiment. We generated random time-series of different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$) from a GPD distribution with a known shape parameter (ξ = -0.5, -0.25, 0, 0.25 and 0.5). For each combination, 10,000 random samples were generated. The performance of the estimators was evaluated using: (a) bar plots for MSE values and (b) boxplots for estimated ξ. Here the "error" in MSE is the difference between the actual (or known) ξ and that estimated, where MSE incorporates both the variance and the bias of the estimators. Outcomes were used to guide the analyses with the measured NWFP flow data.

4. Results

4.1 Monte Carlo study for Performance of GPD estimators

Our simulated data analysis showed that the performance of the GPD parameter estimators depends on both the sample size n (see performance plots in Figure 3 for a shape parameter of $\xi = 0$ only) and the value of the shape parameter ξ (see supplementary material for performance plots with ξ = -0.5, -0.25, 0.25 and 0.5), which accords with previous studies (e.g. Gharib et al., 2017; Mackay et al., 2011). On viewing all plots, the maximum likelihood (MLE and MPLE) estimators were both negatively biased for small sample sizes for any value of the shape parameter and their performance increased in terms of bias and variance as sample size increased. The MLE outperformed the other estimators for large sample sizes for all values of the shape parameter. The unbiased and biased probability weighted moments, PWMU and PWMB respectively, were consistently the least biased amongst all estimators and provided a small variance, which was less sensitive to sample size compared to the likelihood estimators. According to the MSE, the PWM estimators were most appropriate for small sample sizes and positive shape parameters. The MOM estimator had a similar behavior to the PWMs when $\xi \leq 0$ but had a negative bias for $\xi > 0$ and the bias increased as the value of the shape parameter and the sample size increased. Pickland's estimator ('Pick') and the MGF estimators produced a large variance and the least accurate estimates of the shape parameter, through the whole range of the examined values. LME was among the best performing estimators regarding accuracy and bias, except for the very short tails ($\xi = 0.5$, see supplementary material), when the estimates deviated greatly from the rest of the estimators and the predefined value of the shape parameter. In summary, the MLE/MPLE, PWMU/PWMB and the LME were considered the most unbiased and precise estimators and so we select only from this reduced group of estimators in subsequent analyses using the measured data.

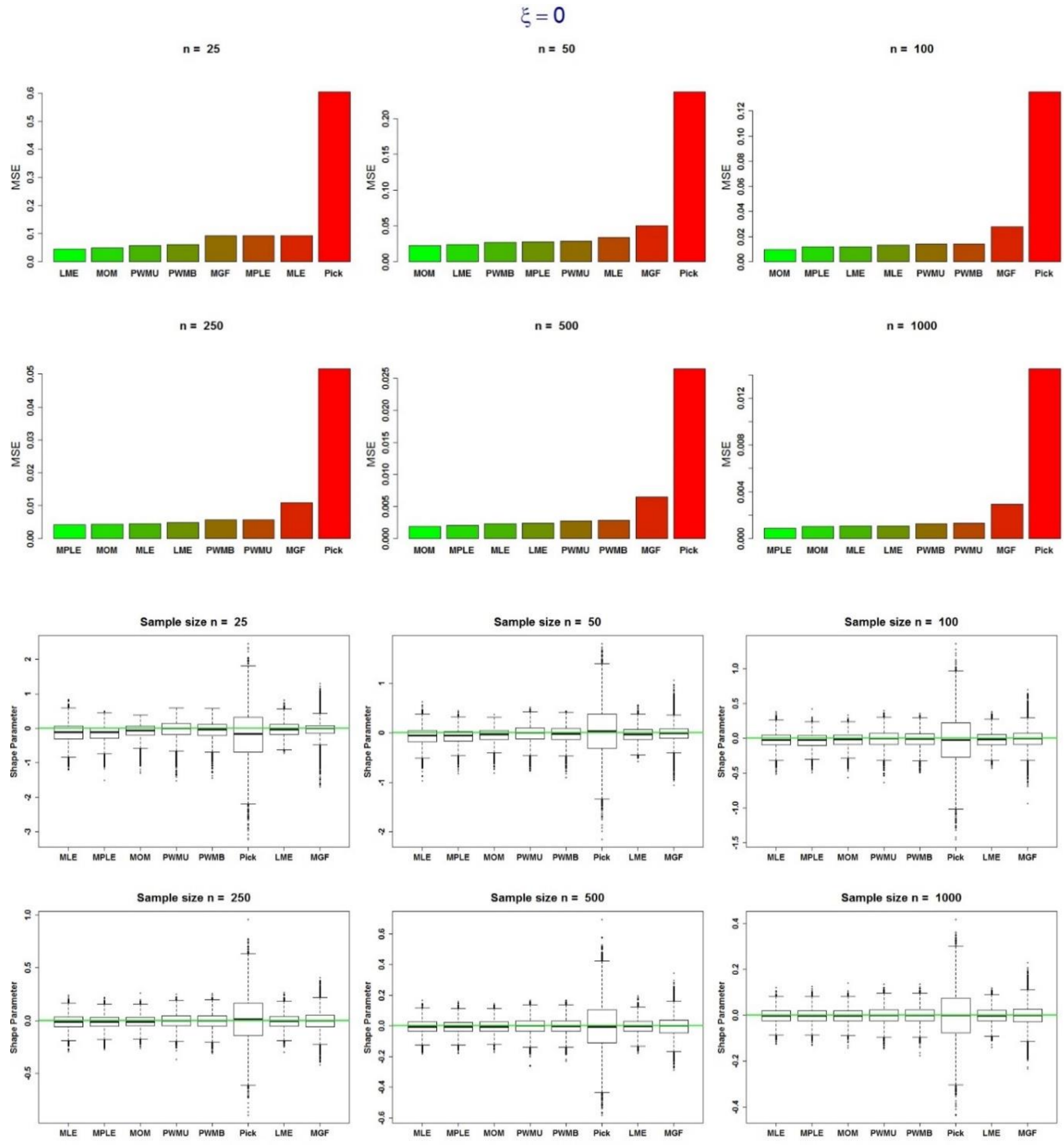


Figure 3: Performance of GPD estimators for shape parameter ξ = 0 and for six different sample sizes ($n = 25,\ 50, 100, 250, 500, 1000$).

4.2 Empirical study for Threshold Selection

4.2.1 Preliminary effects of data aggregation

Initially, the flow (l s^{-1}) time-series of 15-minute resolution was averaged to time-series data of 30 minutes, hourly, 3-hourly, 6-hourly, 12-hourly and daily resolutions. Figure 4 shows the behavior of the MLE-estimated shape parameters for a range of thresholds for the differently aggregated flow data. The range of thresholds was set from the median to the maximum for which daily flow can be fitted efficiently. The shape parameter is in the range of 0.5 to almost 2 for the minimum threshold, has a decreasing trend as the threshold increases and can become negative for the largest thresholds. The similar shape characteristics could be an indication that the shape parameter describes an inherent feature of the process and that changes of scale, which affect the size or variability of the observed values of the process, do not substantially change the shape characteristics of these observations. For the remainder of this study, results from the 30-minute, 3-hourly and 12-hourly aggregations are not reported as retained aggregations (hourly, 6-hourly and daily) communicate all key outcomes adequately.

Kendall's τ test showed that the maximum peaks separated by a minimum of three days were reasonably independent (Figure 5). The statistics τ are large for the lowest thresholds where the peaks are numerous and autocorrelated. With an increasing threshold, the values of the τ decrease rapidly and are below the 95% acceptance limits which supports the null hypothesis of independence of the peaks.

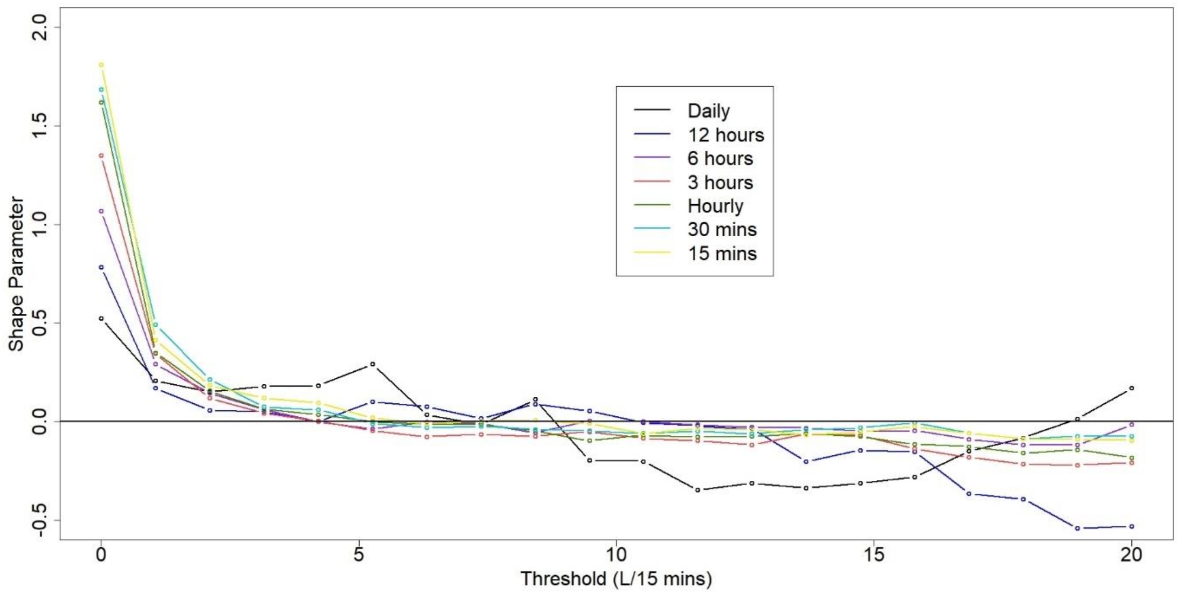


Figure 4: Shape parameter characteristics of measured (15-minute) and a series of averaged (30-minute to daily) flow rates.

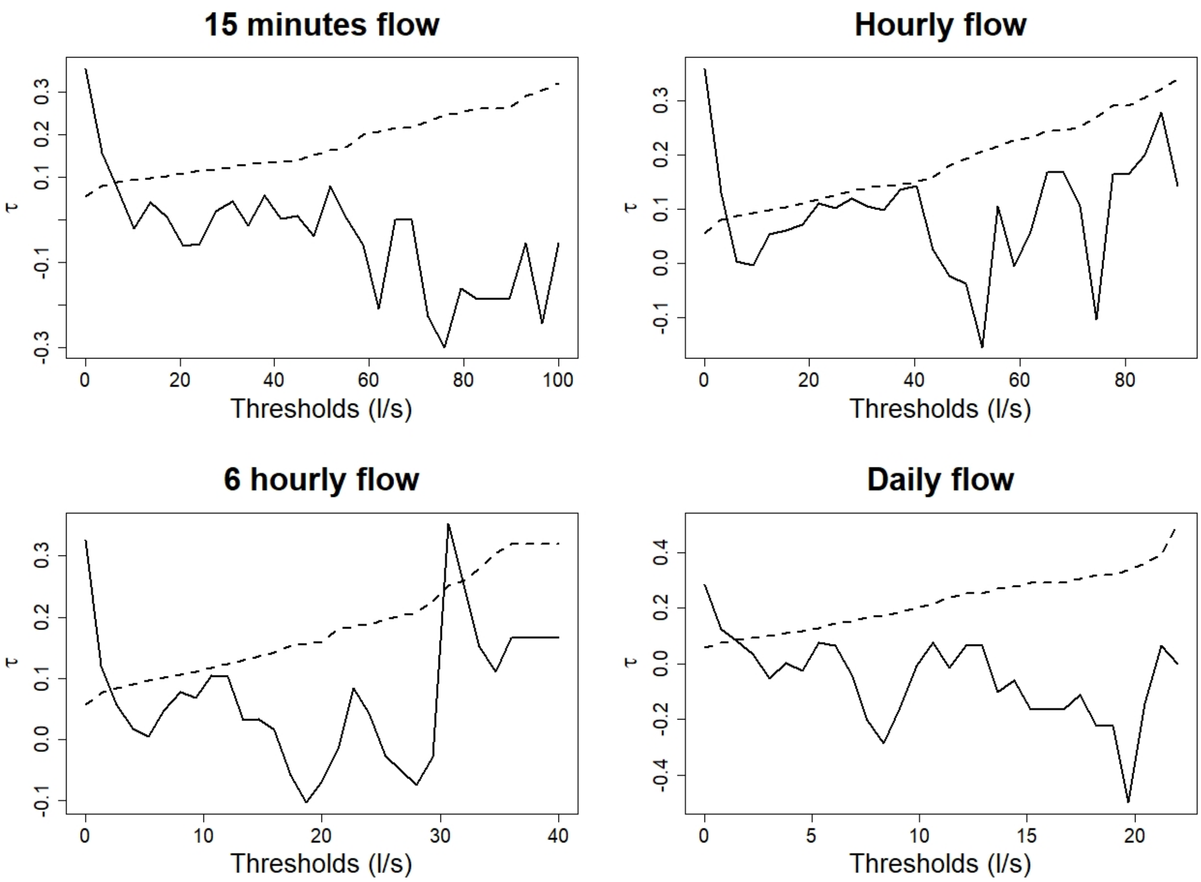


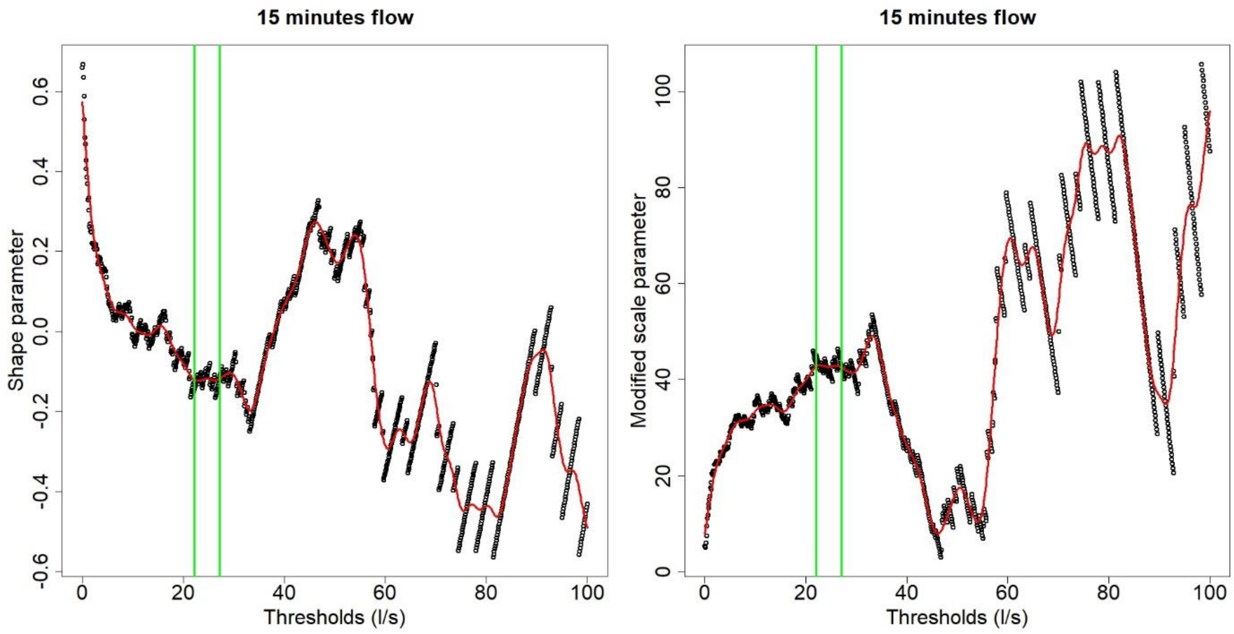
Figure 5: Kendall's test statistic τ (solid lines) along with the 95% acceptance limits of the test (dashed lines).

4.2.2 Automated Threshold Stability plots

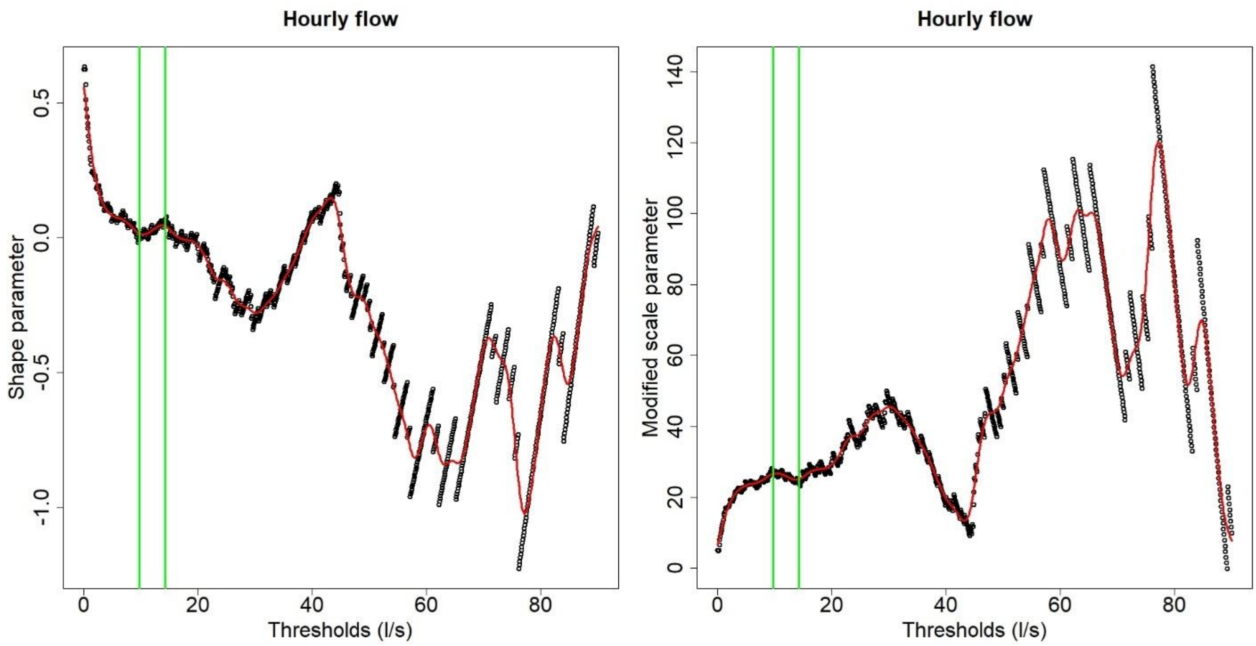
The choice of estimators for the shape and modified scale parameters was guided by the results of the Monte Carlo experiment (Section 4.1). For example, for thresholds $u_j = 1,2, \dots ,5$ of the 15-minute flow data, the number of exceedances was $N_{u_j} > 300$ and the shape parameter ξ_{u_j} between 0.5 and 0.25. For this combination, MLE, MPLE, PWMU, PWMB and LME were the best performing estimators. Thus, for our empirical study, we choose LME due to its consistently precise and unbiased estimates of positive shape parameters for a large sample size. Increasing the thresholds u_j resulted in a reduced sample size ($100 < N_{u_j} < 250$) and negative values of the shape parameter. In this case, we choose MPLE for our empirical work. In all the other cases, the PWMU estimator was preferred as it provided unbiased estimates with small variance.

Stability plots are given in Figure 6 for different flow aggregations, where results reveal our 'Automated Threshold Stability' (ATS) extension to be reasonably robust, since changes in the number of consecutive steps m had a very small impact on the selected threshold and usually resulted in over-lapping regions from which the threshold was considered. The peak flows at 15 minutes and hourly resolution did not provide many regions that could be considered as a plateau, so the number of consecutive steps was set to m = 50 (5% of the total) to also capture the smaller approximately linear horizontal parts. Interestingly, for each aggregation, fitting the same cubic spline functions to both the estimated shape and modified scale parameters, resulted in almost identical suggested thresholds.

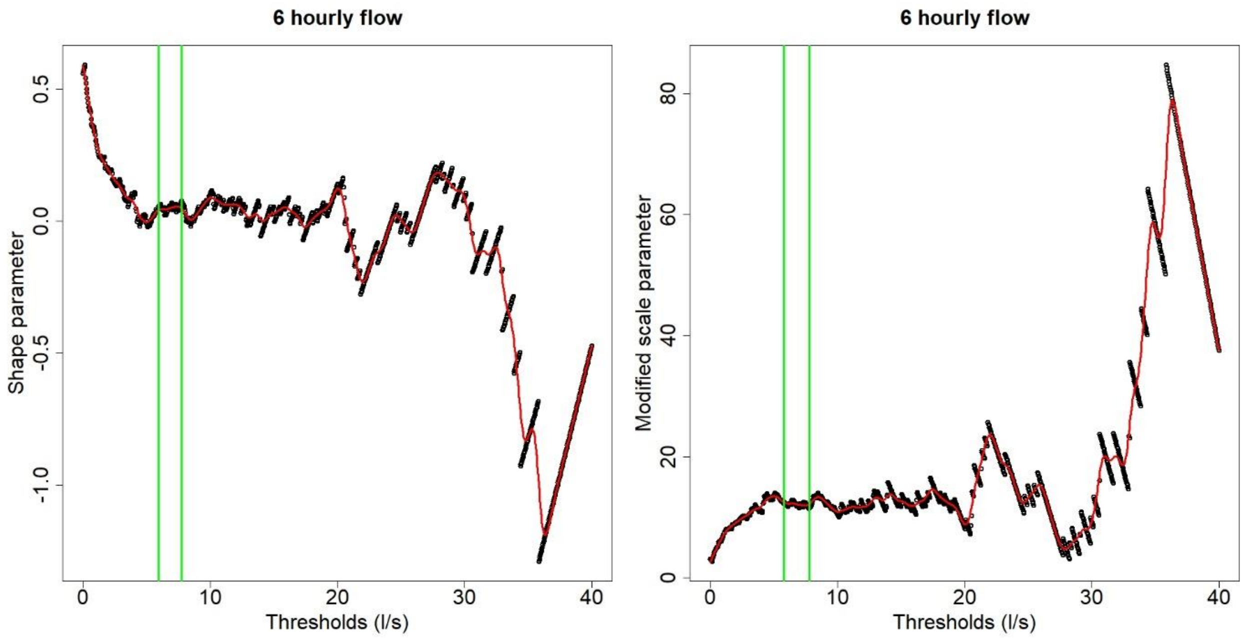
a)


15 minutes flow
Shape parameter
Thresholds (l/s)
15 minutes flow
Modified scale parameter
Thresholds (l/s)

b)


Hourly flow
Shape parameter
Thresholds (l/s)
Hourly flow
Modified scale parameter
Thresholds (l/s)

c)



d)

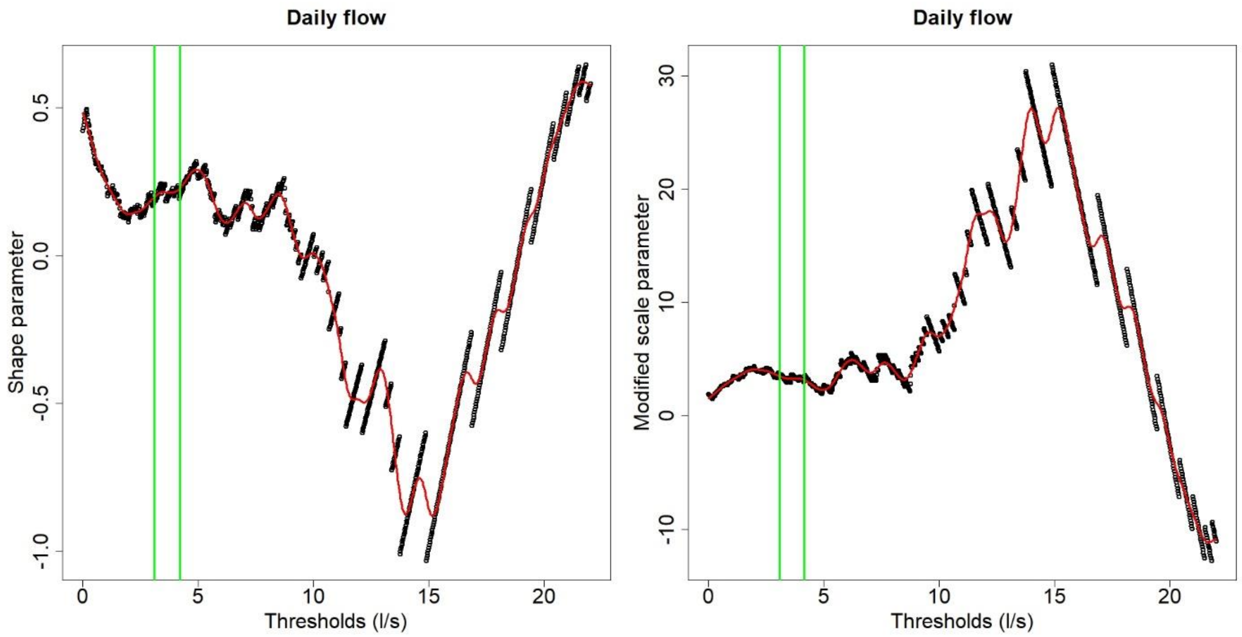


Figure 6: Automated Threshold Stability (ATS) method: Selected threshold (that between the vertical green lines) of a) 15 minutes, b) hourly, c) 6 hourly and d) daily flow based on smoothing splines.

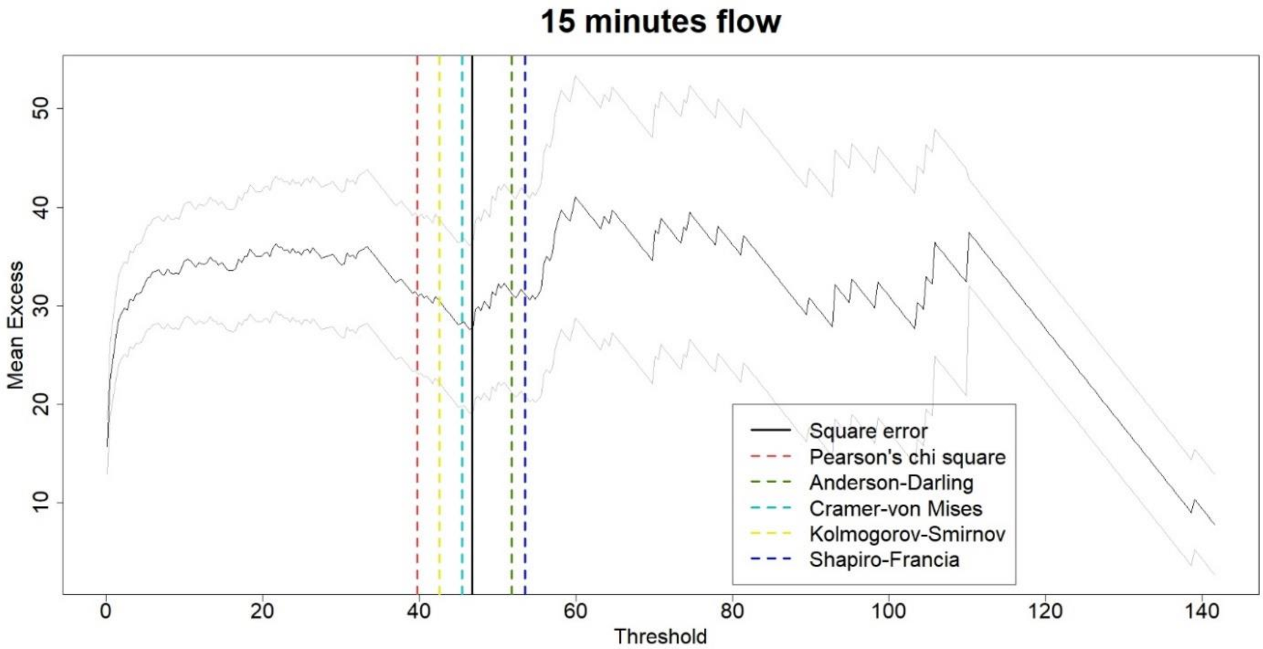
4.2.3 Analytical threshold selection methods: Square Error and Normality of Differences

The choice of GDP estimators for the simulation of the quantiles for the SE method was performed using a similar procedure as described in Section 4.2.2, while the approach based on the Normality of Differences test is based on assumptions of maximum likelihood theory, and consequently the shape parameter was estimated by the MLE. The number n of the considered thresholds u_n plays an important role in the results. Thompson et al. (2009) suggested $n = 100$ and reported that for $n < 100$, less reliable results were obtained. We similarly specified $n = 100$ but also found the thresholds to be over-estimated for $n > 100$.

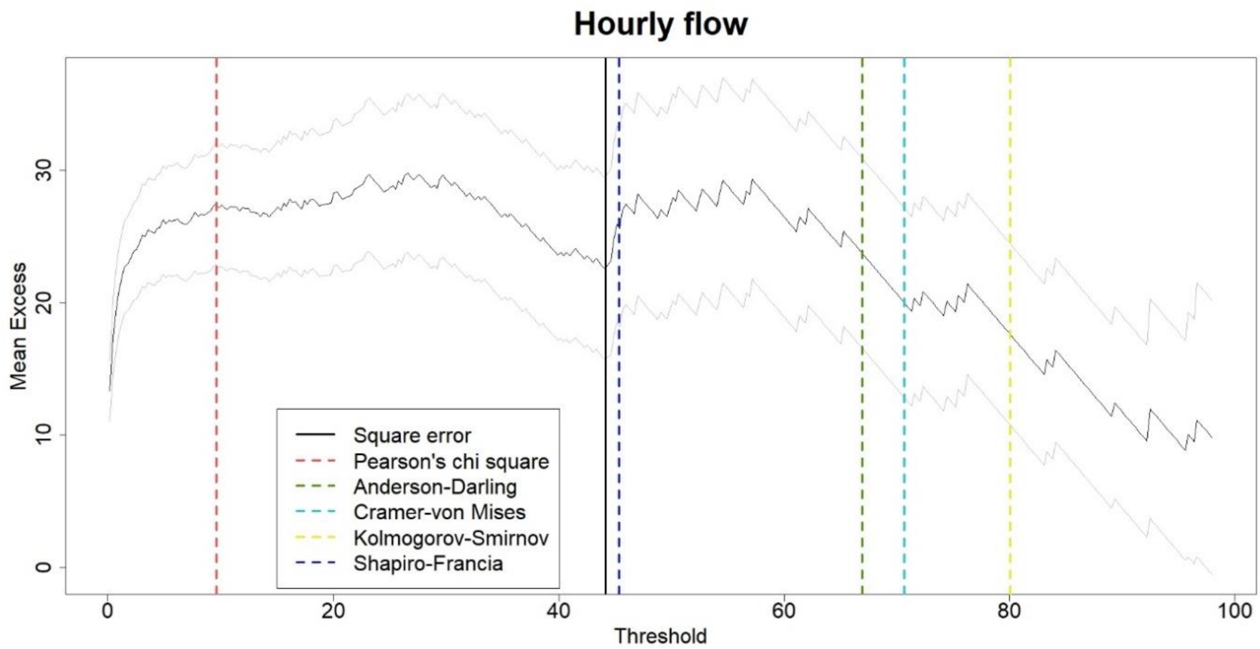
Our results indicated little consistency in the selection of thresholds where a specific part of the MRL plot could be considered approximately linear. The thresholds of the 15-minute peak flow estimated by the SE method and the Normality of Differences tests (Figure 7a) are considerably larger than that based on this study's ATS method (Figure 6a) at around 40 to 50 l/s and 20 to 30 l/s, respectively. Only for the daily flow data (Figure 7d), the threshold estimated by the SE method was smaller than those estimated from the Normality of Differences tests and relatively close to the threshold estimated by ATS (Figure 6d). For hourly flow data (Figure 6b and Figure 7b), ATS and Pearson's chi square test (for Normality of Differences) provided almost identical estimates, while all other methods suggested much larger thresholds. Noticeably, the hourly thresholds estimated by the SE method and the Shapiro-Francia test are very close at 44.68 l/s and 45.33 l/s, respectively (Figure 7b), but result in considerably different shape parameters (Table 1). Figure 6b reveals hourly thresholds to be in the region where the shape characteristics show large fluctuations due to the small sample size that results in an inefficient fit of the GPD and likely spurious estimates of the shape parameter.

The performance of the Normality of Differences method depended greatly on both the given normality test and on data resolution. For the 15-minute flow data, all normality tests provided relatively similar threshold selections (Figure 7a), which was not the case for the hourly and 6-hourly flow data (Figure 7b and Figure 7c). For the daily flow data (Figure 7d), thresholds were estimated too large and consequently result in too few values for efficient statistical inference. In general, the smaller the selected threshold, given that the excesses are satisfactorily modelled by the GPD, the lower the uncertainty and consequently the lower the variance in the parameter estimates due to larger sample sizes.

a)

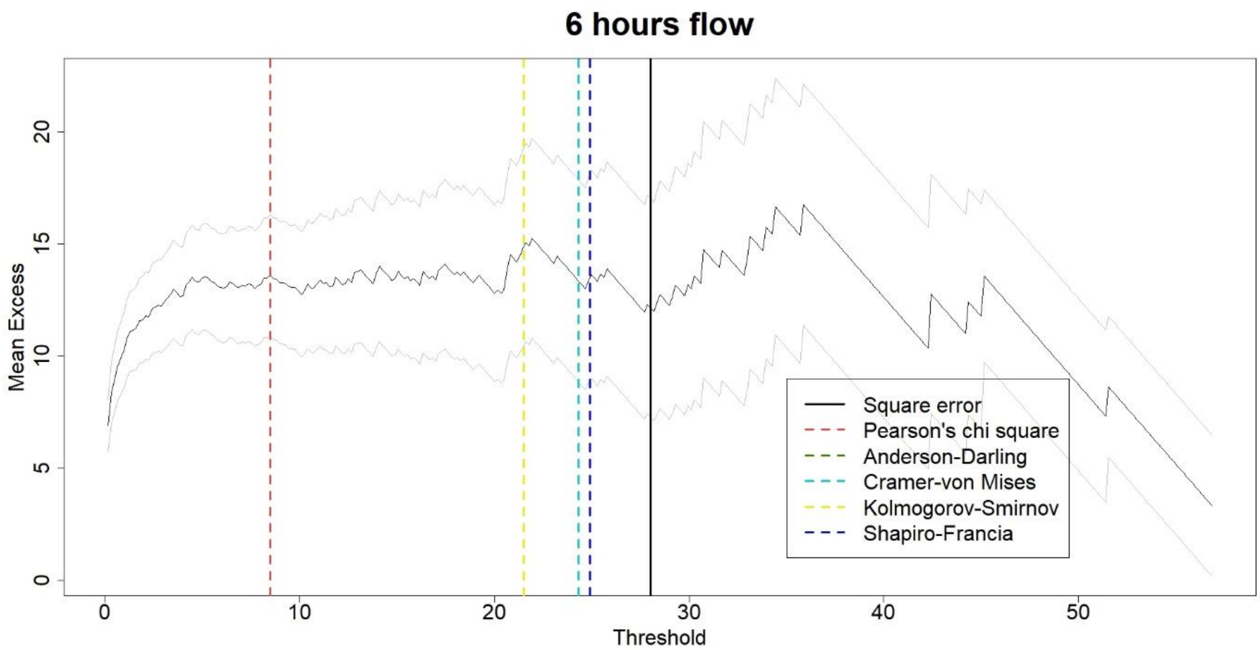


b)


Hourly flow
Mean Excess
Threshold
Square error
Pearson's chi square
Anderson-Darling
Cramer-von Mises
Kolmogorov-Smirnov
Shapiro-Francia
0
20
40
60
80
100
0
10
20
30

411

412 c)


6 hours flow
Mean Excess
Threshold
Square error
Pearson's chi square
Anderson-Darling
Cramer-von Mises
Kolmogorov-Smirnov
Shapiro-Francia
0
10
20
30
40
50
0
5
10
15
20

413

d)

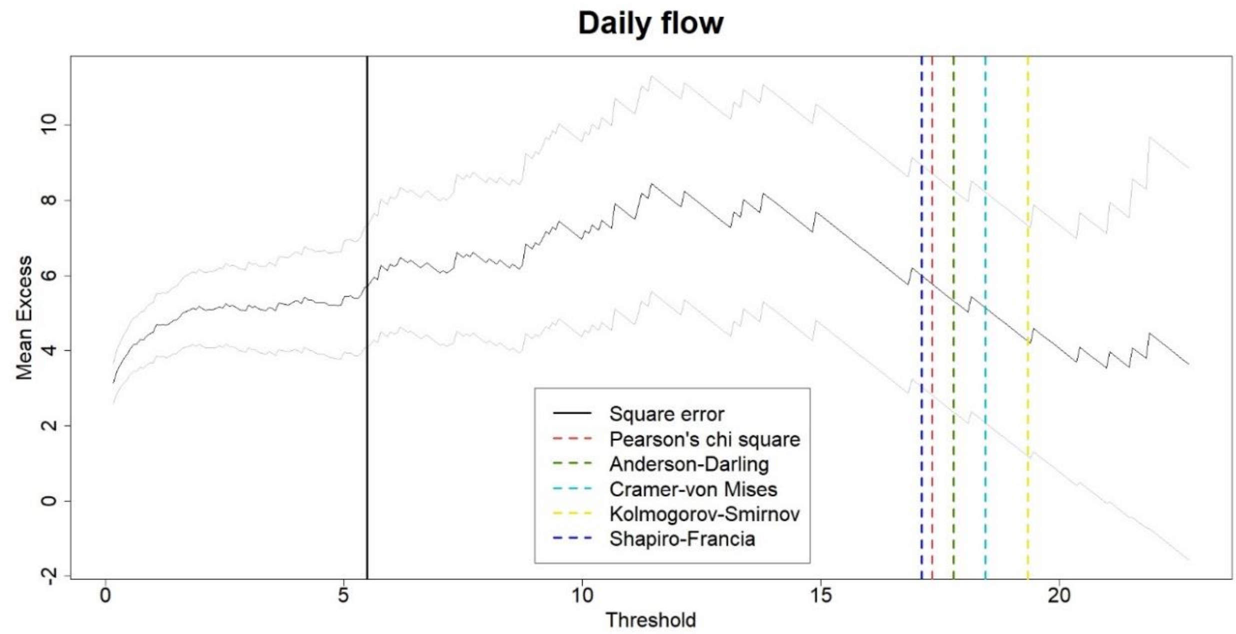


Figure 7: MLR plots: Mean excesses and their 95% confidence intervals plotted against threshold for the a) 15 minutes, b) hourly, c) 6 hourly and d) daily flow data. The threshold selected using the SE method is shown by the vertical solid line and the thresholds selected by the Normality of Differences tests are shown by the dashed vertical lines.

4.2.4 Parameter and fit comparisons

In summary, the estimated shape parameters showed little consistency across the four data resolutions and across the threshold selection techniques investigated (Table 1). The 15-minute extreme flows are characterized by: (i) an exponential tail (Pearson's chi square, Anderson Darling and Kolmogorov-Smirnov tests) as the shape parameter takes values close to zero, (ii) heavy tails (SE method, Shapiro-Francia and Cramer-von Mises tests) and (iii) short tails ($\xi < 0$) (ATS method). ATS and Normality of Differences methods resulted in short tail distributions for both the hourly and 6-hourly flow data, whereas the SE method resulted in a heavier tail, similar to that found across all flow data scales. The ATS and the SE methods provided heavy tails for the daily flow, and the Normality of Differences tests tended to short tails.

Table 1: Estimated thresholds and shape parameters for four flow resolutions and three core threshold selection methods.

		ATS	SE	Normality of Differences tests				
				Pearson's chi square	Anderson-Darling	Cramer-von Mises	Kolmogorov-Smirnov	Shapiro-Francia
15 mins	Threshold	22.2	46.8	39.7	51.8	45.5	42.6	53.5
	Shape Parameter	-0.14	0.33	0.01	0.07	0.26	0.06	0.10
Hourly	Threshold	9.7	44.7	9.6	66.9	70.7	80.1	45.3
	Shape Parameter	-0.09	0.17	-0.09	-0.58	-0.44	-0.48	-0.35
6 hours	Threshold	6.6	28.1	8.5	24.3	24.3	21.5	24.9
	Shape Parameter	-0.01	0.20	-0.05	-0.23	-0.23	-0.34	-0.23
Daily	Threshold	3.1	5.6	17.3	17.8	18.4	19.3	17.1
	Shape Parameter	0.17	0.22	-0.17	-0.10	-0.08	0.10	-0.20

Table 2: MSE between the empirical and theoretical quantiles for different threshold selection methods at four flow resolutions.

MSE	Threshold Stability	SE	Normality of Differences tests				
			Pearson's chi square	Anderson-Darling	Cramer-von Mises	Kolmogorov-Smirnov	Shapiro-Francia
15 mins	252.4	8248.8	123.7	2157.8	6034.9	1242.3	2828.2
Hourly	130.9	2654.1	24.1	14.5	13.6	10.5	28.0
6 hourly	72.1	150.8	61.0	34.0	34.0	12.7	34.8
Daily	38.2	81.9	8.3	10.7	12.6	32.4	7.6

The MSE (Table 2) seems to be an inappropriate diagnostic for deviations between very large theoretical and empirical quantiles as it depends greatly on the shape parameter. Peak flows with very short finite tails will show minimum MSEs, which increase by orders of magnitude as the shape parameter increases. Conversely, the NRMSE does provide a comparative diagnostic since it is normalized by accounting for very large values that are associated with heavy tails. Thus, NRMSE values are reported in Table 3 where compared to the SE and Normality of Differences methods, this study's ATS method gives the smallest NRMSE for flow data of any resolution, except for the Normality of Differences test for the hourly flow.

Table 3: NRMSE between the empirical and theoretical quantiles for different threshold selection methods at four flow resolutions.

NRMSE	ATS	SE	Normality of Differences tests				
			Pearson's chi square	Anderson-Darling	Cramer-von Mises	Kolmogorov-Smirnov	Shapiro-Francia
15 mins	102.6	1017.9	308.0	571.6	866.6	391.4	697.5
Hourly	38.8	244.4	37.7	30.9	29.9	38.2	27.0
6 hourly	51.8	184.2	67.6	87.4	87.4	53.4	88.5
Daily	44.5	69.3	52.6	59.5	72.0	115.3	50.2

The relative index of agreement (Figure 8) is also an efficient measure of proximity between observed and simulated peak flows (Krause et al., 2005). For this diagnostic, the GPD was consistently best fitted to empirical peak flows at all scales when their thresholds were chosen using this study's ATS method. Here, the SE method was the poorest method, especially at the 15-minute data scale. Interestingly, results at the hourly scale behaved very differently to those found at the three other scales. We speculate that this was likely due to the hourly data being at, or close to, the natural water run-off integration rate to the sub-catchment's water flume following a rainfall event (see Discussion).

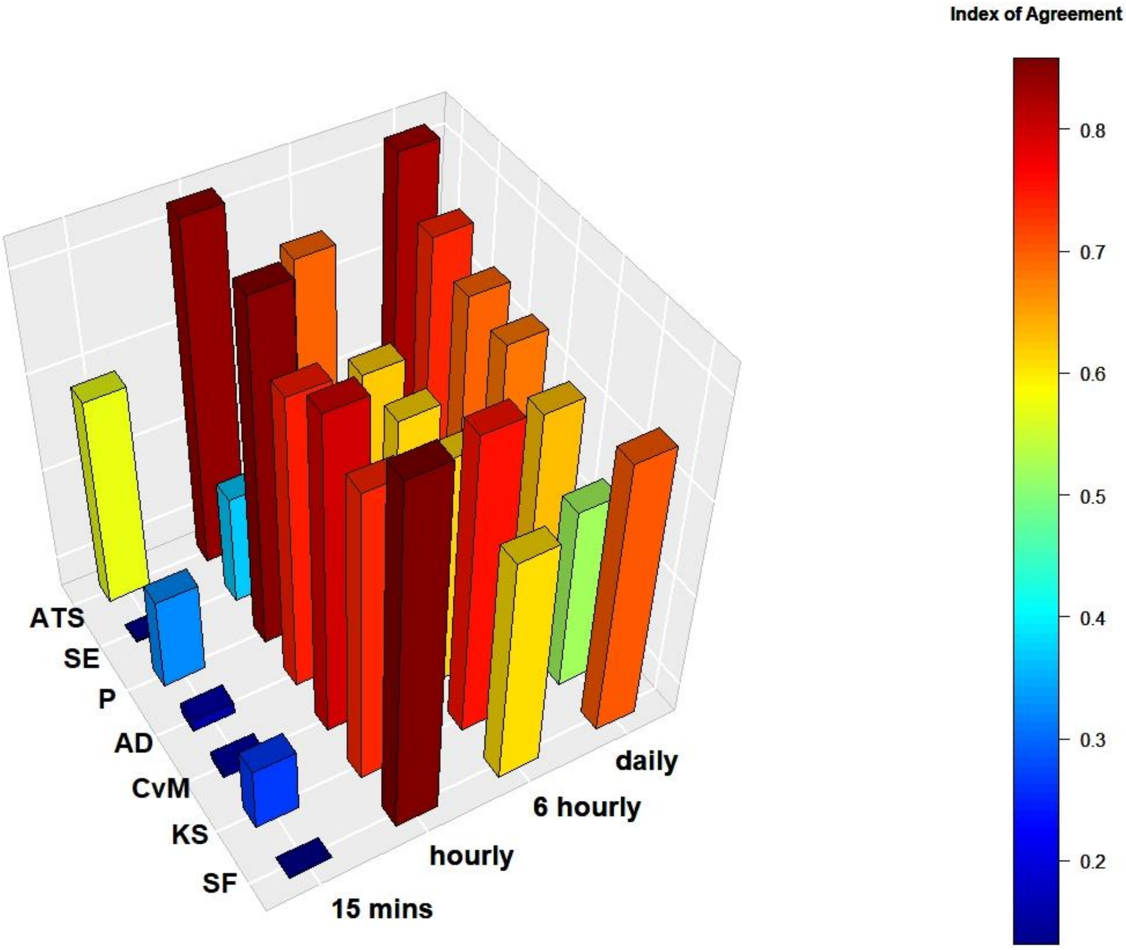


Figure 8: Index of agreement between theoretical and empirical peak flow of different resolutions. The threshold selection methods are Automated Threshold Stability (ATS), Square Error (SE) and the various tests of the Normality of Differences method, the Pearson's chi-square (P), Anderson-Darling (AD), Cramer-von Mises (CvM), Kolmogorov-Smirnov (KS) and Shapiro-Francia (SF).

Figure 9 presents the Q-Q plots of the 15-minute extreme flows for the threshold selection methods that gave the smallest (ATS) and the largest (SE) NRMSE values (Table 3). The Q-Q plots show that an over-estimated threshold results in a sample size that can be too small for efficient statistical inference and results in increased uncertainty. The Q-Q plots also emphasis the superiority of this study's ATS method given its Q-Q plot falls relatively close to the 45^{o} line.

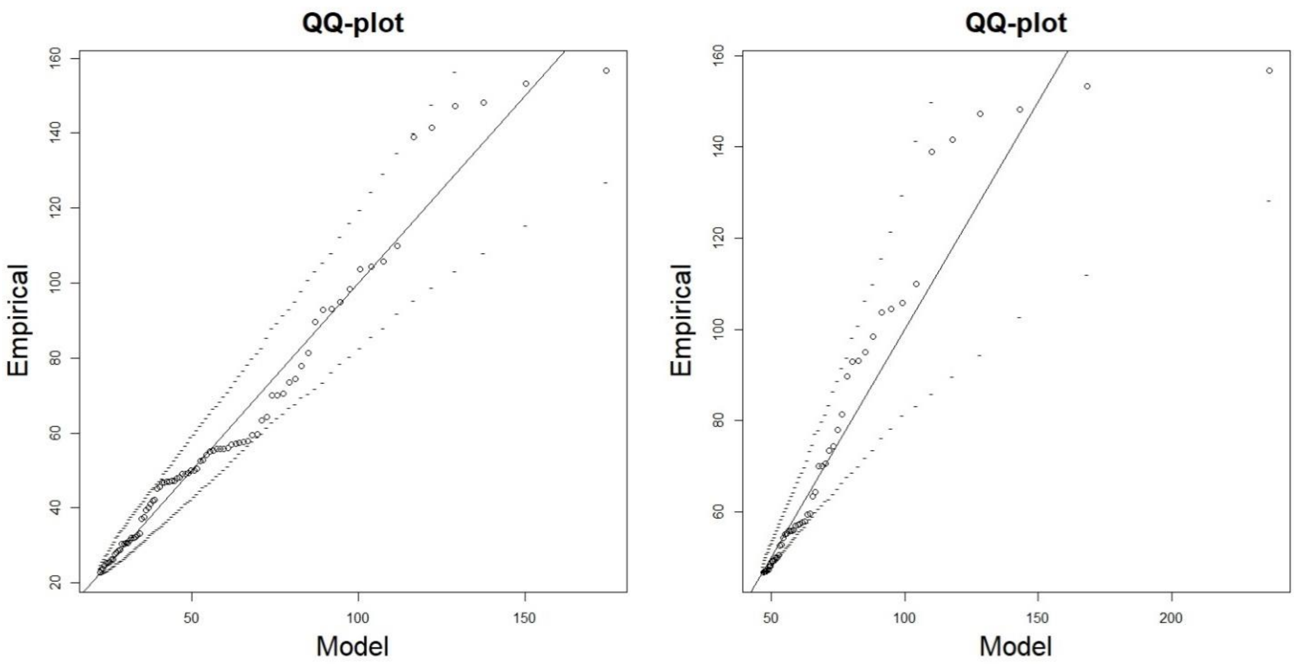


Figure 9: Q-Q plots of the 15-minute peak flows estimated by the ATS (left) and SE (right) methods.

Clear differences in the estimated Return Level / Return Period plots for the ATS and Normality of Difference (Kolmogorov-Smirnov test only) methods (Figure 10) indicate that the combined effects of data scale, the GPD estimator and the threshold selection method - each have a significant impact on the characteristics of the final model that attempts to explain the flow process with the consideration of extremes. This is critically important in cases where reliably informed actions need to be taken or infrastructure needs to be built to mitigate the impacts of future peak flows and likely flood events.

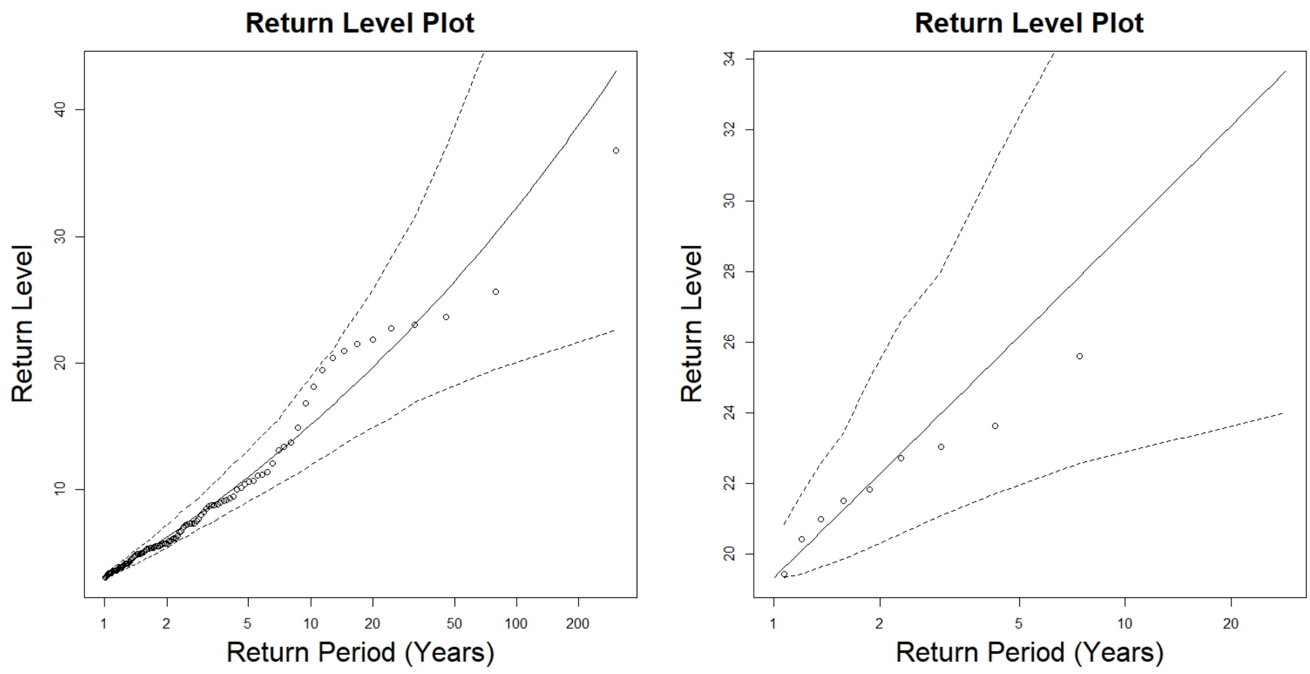


Figure 10: Return level plots of the daily peak flows estimated by the ATS (left) and Normality of Difference Kolmogorov-Smirnov (right) methods.

5. Discussion

In agreement with previous studies (e.g. Bermudez & Kotz, 2010; Engeland et al., 2004), we found that the performance of the GPD parameter estimators examined through a Monte Carlo experiment, depended significantly on the sample size and the value of the shape parameter. The MLE/MPLE, PWMU/PWMB and the LME were consistently the most unbiased and precise estimators and so we chose only from this group in our subsequent analyses. More specifically, for the application of the SE and AST threshold selection methods, a different GPD estimator was used each time according to its strengths. For example, the LME was preferred for positive shape parameters and large sample size.

This study's Automated Threshold Stability (ATS) method was tested against existing SE and Normality of Differences methods. Methods were applied to flow discharge measurements of 15-minute resolution, as well as to the same data aggregated to coarser resolutions of hourly, 6-hourly and daily, to examine scale effects. The Normality of Differences method depended on the normality test applied and resulted in short, exponential and heavy tailed distributions even at the same scale (e.g. shape parameters of $\xi = -0.2$ for the daily flow according to Shapiro-Francia and $\xi = 0.1$ according to the Kolmogorov-Smirnov test). Similar results for the value of the shape parameter were obtained from the ATS method, unlike the SE method which always resulted in positive ξ.

Threshold stability plots were discussed in Scarrott and MacDonald (2012) and Solari and Losada (2012), but these studies did not perform an analytical approximation, as done here with ATS, although Langousis et al. (2016) suggested an automated technique based on the assumption of linearity of the MRL plot and applied it to rainfall data. Our proposed ATS method provided more robust estimates of the threshold compared to: (a) the SE method as it was less sensitive to the resolution of the data and (b) the Normality of Differences method as it was less sensitive to the sample size of the threshold candidates. It also resulted in the smallest errors and the largest agreement indices between the simulated and the empirical quantiles.

Specific to the case study, error and agreement indices indicated that the GPD provided the best fit to the hourly peak flow data relative to 15-minute, 6-hourly and daily peak flow data. For all the applied threshold selection methods, the modelled peak flow at the hourly resolution was consistently the closest to the empirical one, compared to three other scales. These results cannot be attributed to the value of the shape parameter (e.g. short finite tails result in greater agreement between theoretical and empirical quantiles) since the SE method gives a positive ξ. An inspection of the plots and a comparison across various scales does not reveal any pattern that would justify this behavior. A possible explanation could be that the hourly peak flow best captures the signal of the process and integrates more efficiently the way the 6.84 ha sub-catchment (of two pasture fields) transforms intensive rainfall into high discharge flows. It should be noted that the data aggregation was not done at equal intervals. For example, the hourly flow resulted from averaging four 15-minute measurements, whereas the 6-hourly and the daily flow are the averages of 24 and 96 observations, respectively. This does not affect the results but should be borne in mind when interpreting the plots.

An advantage of using fine resolution flow data is that they result in larger sample sizes that can make the statistical inference more efficient even for records of short periods for which a GEV/AM extreme value methodology is not applicable. However, this study showed that for data of the same resolution, the value of the GDP shape parameter varies according to the selected thresholds. This has serious practical implications since the models are commonly extrapolated beyond observed values for forecasting and engineering design purposes to mitigate against future flooding. On one hand, an under-estimated threshold and shape parameter of the extreme flow can result in failure of hydrological infrastructure (e.g. dams, flood protection works) due to higher peak flows than expected. On the other hand, over-estimation of the high flows can lead to over-pricing and mis-use of resources.

6. Conclusions

In this study, we examined the effect of statistical estimators, data resolution, and threshold selection on fitting the Generalized Pareto distribution to peak hydrological flows that resulted from the 'Peaks Over Threshold' method. Through a simulation study, the performance of the estimators depended greatly on the sample size and the shape parameter where the only most accurate and unbiased estimators were used for the selection of thresholds in subsequent empirical evaluations. Here an automated threshold selection method based on the stability of the shape and modified scale parameters was empirically demonstrated to provide more robust estimates compared to two commonly applied alternatives. The proposed method provided the smallest error and the greatest agreement indices between the empirical and theoretical quantiles across all the scales of the case study flow data.

The study results can be generalized to similar water monitoring schemes for improved characterization of likely flood events. However, the study highlights that the combined effect of data scale, threshold selection method and statistical estimator, significantly affects the shape parameter and, as a consequence, the nature of the Generalized Pareto distribution. Such linked effects need to be acknowledged and assessed as they have clear implications for the reliable forecasting of extreme flow events, and the consequences thereof.

Acknowledgements

Rothamsted Research receives grant aided support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom. This research was funded by Rothamsted Research and Lancaster Environment Centre, the BBSRC Institute Strategic Programme (ISP) grant, “Soils to Nutrition” (S2N) grant numbers BBS/E/C/000I0320, BBS/E/C/000I0330 and the BBSRC National Capability grant for the North Wyke Farm Platform grant number BBS/E/C/000J0100.

Declaration of interest

The authors declare no potential conflict of interest associated with this research.

Software and data availability

The statistical software (R Core Team, 2017) and all North Wyke Farm Platform data sets (https://www.rothamsted.ac.uk/north-wyke-farm-platform) are freely available.

References

Ashkar, F. and Tatsambon, C. N. (2007). Revisiting some estimation methods for the generalized Pareto distribution. Journal of Hydrology, 346, 136-143.

Bates, B. C., Kundzewicz, Z. W., Wu, S. and Palutikof, J. P. (2008). Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.

Beguería, S. (2005). Uncertainties in Partial Duration Series Modelling of Extremes Related to the Choice of the Threshold Value. Journal of Hydrology, 303(1), 215-230.

Behrens, C. N., Lopes, H. F. and Gamerman, D. (2004). Bayesian Analysis of Extreme Events with Threshold Estimation. Statistical Modelling, 4(3), 227-244.

Beirlant, J., Dierckx, G. and Guillou, A. (2005). Estimation of the Extreme-Value Index and Generalized Quantile Plots. Bernoulli, 11(6), 949-970.

Beirlant, J., Joossens, E. and Segers, J. (2005). Unbiased tail estimation by an extension of the generalized Pareto distribution. CentER Discussion Paper, Vol. 2005–112, Tilburg: Econometrics.

Beirlant, J., Vynckier, P. and Teugels, J. L. (1996). Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics. Journal of the American Statistical Association, 91(436), 1659-1667.

Beirlant, J., de Wet, T. and Goegebeur, Y. (2006). A Goodness-of-Fit Statistic for Pareto-Type Behaviour. Journal of Computational and Applied Mathematics, Special Issue: Jef Teugels, 186(1), 99-116.

Bommier, E. (2014). Peaks-over-threshold modelling of environmental data (Technical report). U.U.D.M. Project Report, 2014:33.

Bouraoui, F., Grizzetti, B., Granlund, K., Rekolainen, S. and Bidoglio, G. (2004). Impact of Climate Change on the Water Cycle and Nutrient Losses in a Finnish Catchment, Climatic Change, 66(1–2), 109-126.

Brodin, E. and Rootzén, H. (2009). Univariate and Bivariate GPD Methods for Predicting Extreme Wind Storm Losses. Insurance: Mathematics and Economics, 44(3), 345-356.

Choulakian, V. and Stephens, M. A. (2001). Goodness-of-Fit Tests for the Generalized Pareto Distribution. Technometrics, 43(4), 478-484.

Claps, P. and F. Laio. (2003). Can Continuous Streamflow Data Support Flood Frequency Analysis? An Alternative to the Partial Duration Series Approach. Water Resources Research, 39(8).

Clarke, M. L. & Rendell, H. M. (2006). Hindcasting Extreme Events: The Occurrence and Expression of Damaging Floods and Landslides in Southern Italy. Land Degradation & Development, 17(4), 365-380.

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

Coles, S. and Dixon, M. J. (1999). Likelihood-Based Inference for Extreme Value Models, Extremes, 2(1), 5-23.

Cunnane, C. (1979). A Note on the Poisson Assumption in Partial Duration Series Models. Water Resources Research, 15(2), 489-494.

Danielsson, J., de Haan, L., Peng, L. and de Vries, C. G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. Journal of Multivariate Analysis, 76(2), 226-248.

Das, B. and Ghosh, S. (2013). Weak limits for exploratory plots in the analysis of extremes. Bernoulli, 19(1), 308-343

Davison, A. C. and Smith, R. L. (1990). Models for Exceedances over High Thresholds, Journal of the Royal Statistical Society, Series B (Methodological), 52(3), 393-442.

Deidda, R. (2010). A Multiple Threshold Method for Fitting the Generalized Pareto Distribution to Rainfall Time Series. Hydrology and Earth System Sciences, 14(12), 2559-2575.

Deidda, R. and Puliga, M. (2006). Sensitivity of Goodness-of-Fit Statistics to Rainfall Data Rounding Off. Physics and Chemistry of the Earth, 31(18). 1240-1251.

Dekkers, A. L. M. and De Haan, L. (1989). On the Estimation of the Extreme-Value Index and Large Quantile Estimation, The Annals of Statistics, 17(4), 1795-1832.

Durocher, M., Zadeh, S. M., Burn, D. H. and Ashkar, F. (2018). Comparison of Automatic Procedures for Selecting Flood Peaks over Threshold Based on Goodness-of-Fit Tests. Hydrological Processes, 32(18), 2874–2887.

Eastoe, E. F. and Tawn, J. A. (2010). Statistical Models for Overdispersion in the Frequency of Peaks over Threshold Data for a Flow Series. Water Resources Research, 46(2).

Engeland, K., Hisdal, H. and Frigessi, A. (2004). Practical Extreme Value Modelling of Hydrological Floods and Droughts: A Case Study. Extremes, 7(1), 5–30.

Ferguson, T. S., Genest, C. and Hallin, M. (2000). Kendall's Tau for Serial Dependence. Canadian Journal of Statistics, 28(3), 587–604.

Field, C. B., Barros, V., Stocker, T. F. and Dahe, Q. (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change, Cambridge, Cambridge University Press.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proc. Cambridge Philos. Soc., 24(2), 180-190.

Gharib, A., Davies, E. G. R., Goss, G. G. and Faramarzi, M. (2017). Assessment of the Combined Effects of Threshold Selection and Parameter Estimation of Generalized Pareto Distribution with Applications to Flood Frequency Analysis, Water, 9(9), 692.

Goegebeur, Y., Beirlant, J. and de Wet, T. (2008). Linking Pareto-Tail Kernel Goodness-of-Fit Statistics with Tail Index at Optimal Threshold and Second Order Estimation., REVSTAT-Statistical Journal, 6(1), 51-69.

Greenwood, J. A., Landwehr, J. M., Matalas N. C. and Wallis, J. R. (1979). Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressable in Inverse Form. Water Resources Research, 15(5), 1049-1054.

Hall, P. (1990). Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. Journal of Multivariate Analysis, 32(2), 177-203.

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. The Annals of Statistics, 3(5), 1163-1174.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. Technometrics, 29(3), 339-349.

Jenkinson, A. F. (1955). The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. Quarterly Journal of the Royal Meteorological Society, 81(348), 158-171.

Josse, J. and Husson, F. (2013). Handling Missing Values in Exploratory Multivariate Data Analysis Methods. Journal de La Société Française de Statistique, 153(2), 79–99.

Krause, P., Boyle, D. P. and Bäse, F. (2005). Comparison of Different Efficiency Criteria for Hydrological Model Assessment. Advances in Geosciences, 5, 89–97.

Kundzewicz, Z. W., Mata, L. J., Arnell, N. W., Doll, P., Kabat, P., Jimenez, B. et al. (2007). Freshwater Resources and Their Management. In Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by M. L. Parry, O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, 173–210. Cambridge University Press.

Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979). Probability Weighted Moments Compared with Some Traditional Techniques in Estimating Gumbel Parameters and Quantiles. Water Resources Research, 15(5), 1055-1064.

Lang, M., Ouarda, T. B. M. J. and Bobée, B. (1999). Towards Operational Guidelines for Over-Threshold Modeling. Journal of Hydrology, 225(3), 103-117.

Langousis, A., Mamalakis, A., Puliga, M. & Deidda, R. (2016). Threshold Detection for the Generalized Pareto Distribution: Review of Representative Methods and Application to the NOAA NCDC Daily Rainfall Database. Water Resources Research, 52(4), 2659–2681.

Ledford, A. W. and Tawn, J. A. (2003). Diagnostics for Dependence within Time Series Extremes. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 65(2), 521–543.

Liang, B, Shao, Z., Li, H., Shao, M. and Lee, D. (2019). An Automated Threshold Selection Method Based on the Characteristic of Extrapolated Significant Wave Heights. Coastal Engineering, 144, 22-32.

Liu, Y., Li, Y., Harris, P., Cardenas, L. M., Dunn, R. M., Sint, H., Murray, P. J., Lee, M. R. F. and Wu, L. (2018). Modelling Field Scale Spatial Variation in Water Run-off, Soil Moisture, N2O Emissions and Herbage Biomass of a Grazed Pasture Using the SPACSYS Model. Geoderma, 315, 49-58.

Luceño, A. (2006). Fitting the Generalized Pareto Distribution to Data Using Maximum Goodness-of-Fit Estimators. Computational Statistics & Data Analysis, 51(2), 904-917.

Mackay, E. B. L., Challenor, P. G. and Bahaj, A. S. (2011). A Comparison of Estimators for the Generalised Pareto Distribution. Ocean Engineering, 38(11), 1338-1346.

Madsen, H., Rasmussen, P. F. and Rosbjerg, D. (1997). Comparison of Annual Maximum Series and Partial Duration Series Methods for Modeling Extreme Hydrologic Events: 1. At-Site Modeling. Water Resources Research, 33(4), 747–757.

Millenniu Ecosystem Assessment, 2005. Millenniu Ecosystem Assessment (MA), Ecosystems and Human Well-being: Synthesis. Island Press, Washington, DC.

Moore, D. S. (1986). Tests of Chi-Squared Type Goodness of Fit Techniques. Marcel Dekker, New York.

Orr, R. J., Murray, P. J., Eyles, C. J., Blackwell, M. S. A., Cardenas, L. M., Collins, A. L. et al. (2016). The North Wyke Farm Platform: effect of temperate grassland farming systems on soil moisture contents, runoff and associated water quality dynamics, European Journal of Soil Science, 67, 374–385.

de Zea Bermudez, P. and Kotz, S. (2010). Parameter Estimation of the Generalized Pareto Distribution, Part I. Journal of Statistical Planning and Inference, 140(6), 1353-1373.

Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. The Annals of Statistics, 3(1), 119-131.

Prescott, P. and Walden, A. T. (1980). Maximum Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution. Biometrika, 67(3), 723-724.

Prescott, P. and Walden, A. T. (1983). Maximum Likelihood Estimation of the Parameters of the Three-Parameter Generalized Extreme-Value Distribution from Censored Samples. Journal of Statistical Computation and Simulation, 16(3–4), 241-250.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reiss, R. D. and Thomas, M. (2007). Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields. 3rd edition. Birkhäuser Basel.

Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Es-Timation and Uncertainty Quantification. REVSTAT–Statistical Journal, 10(1), 33–60.

Segers, J. (2005). Generalized Pickands Estimators for the Extreme Value Index. Journal of Statistical Planning and Inference, 128(2), 381-396.

Sheta, A. F. and El-Sherif, M. S. (1999). Optimal Prediction of the Nile River Flow Using Neural Networks. International Joint Conference on Neural Networks. Proceedings, 5, 3438-3441.

Sigauke, C. and Bere, A. (2017). Modelling Non-Stationary Time Series Using a Peaks over Threshold Distribution with Time Varying Covariates and Threshold: An Application to Peak Electricity Demand. Energy, 119, 152-166.

Smith, R. L. (1985). Maximum Likelihood Estimation in a Class of Nonregular Cases. Biometrika, 72(1), 67-90.

Solari, S. and Losada, M. A. (2012). A Unified Statistical Model for Hydrological Variables Including the Selection of Threshold for the Peak over Threshold Method. Water Resources Research, 48(10).

Solari, S., Egüen, M., Polo, M. J. and Losada, M. A. (2017). Peaks Over Threshold (POT): A Methodology for Automatic Threshold Estimation Using Goodness of Fit p-Value. Water Resources Research, 53(4), 2833–2849.

Takahashi, T., Harris, P., Blackwell, M. S. A., Cardenas, L. M., Collins, A. L., Dungait, J. A. J. et al. (2018). Roles of Instrumented Farm-Scale Trials in Trade-off Assessments of Pasture-Based Ruminant Production Systems. Animal, 12 (8), 1766-1776.

Tanaka, S. and Takara, K. (2002). A Study on Threshold Selection in POT Analysis of Extreme Floods, Extremes of the Extremes, Extraordinary Floods, IAHS Publ, 271, 299-304.

Thibault, K. M. & Brown, J. H. (2008). Impact of an Extreme Climatic Event on Community Assembly. Proceedings of the National Academy of Sciences, 105(9), 3410-3415.

Thode, H. C. (2002). Testing For Normality. CRC Press.

Thompson, P., Cai, Y., Reeve, D. and Stander, J. (2009). Automated Threshold Selection Methods for Extreme Wave Analysis. Coastal Engineering, 56(10), 1013-1021.

Todorovic, P. (1978). Stochastic Models of Floods. Water Resources Research, 14(2), 345–356.

Turan, M. E., and Yurdusev, M. A. (2009). River Flow Estimation from Upstream Flow Records by Artificial Intelligence Methods. Journal of Hydrology, 369(1), 71-77.

Willmott, C. J. (1981). On the Validation of Models. Physical Geography, 2(2), 184-194.

Yang, X., Zhang, J. and Ren, W. X. (2018). Threshold Selection for Extreme Value Estimation of Vehicle Load Effect on Bridges. International Journal of Distributed Sensor Networks, 14(2).

Yun, S. (2002). On a Generalized Pickands Estimator of the Extreme Value Index. Journal of Statistical Planning and Inference, 102(2), 389-409.

de Zea Bermudez, P. and Kotz, S. (2010). Parameter Estimation of the Generalized Pareto Distribution, Part I. Journal of Statistical Planning and Inference, 140(6), 1353–1373.

Zhang, J. (2007). Likelihood Moment Estimation for the Generalized Pareto Distribution. Australian & New Zealand Journal of Statistics, 49(1), 69-77.

Zoglat, A., EL Adlouni, S., Badaoui, F., Amar A. & Okou, C. G. (2014). Managing Hydrological Risks with Extreme Modeling: Application of Peaks over Threshold Model to the Loukkos Watershed, Morocco, Journal of Hydrologic Engineering, 19(9), 05014010.

Appendix A: Equations of the estimators

The estimators used in this study can be formally defined as follows:

1. MLE method:

$$L = -n\log\sigma + \left(\frac{1}{\xi} - 1\right)\sum_{i=1}^{n}\log\left(1 - \frac{\xi x_i}{\sigma}\right), \quad \xi \neq 0$$

$$L = -n\log\sigma - \frac{1}{\sigma}\sum_{i=1}^{n} x_i\,, \quad \xi = 0$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the order statistics of a random sample $x_1, \ldots, x_n$ from the GPD. The estimated parameters are obtained when the log-likelihood function L is maximized.

2. MPLE method:

$$P(\xi) = \begin{cases} 1 & \xi \leq 0 \\ \exp\{-\lambda\left(\frac{1}{1-\xi} - 1\right)^{a}\} & 0 < \xi < 1 \\ 0 & \xi \geq 1 \end{cases}$$

where a and λ are the penalizing non-negative constants. The corresponding penalized likelihood function is $L_{pen} = L \times P$.

3. LME is a combination of both likelihood and moment estimators and is derived from:

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}(1 - \theta x_i)^{P} - \frac{1}{1-r} = 0, \qquad \theta < x_{(n)}^{-1},$$

where $\theta = \xi/\sigma$ and $P = -\frac{rn}{\sum_{i=1}^{n}\log(1-\theta x_i)}$. The parameter $r < 1, r \neq 0$ must be pre-defined before the estimation and either be set as ξ if there is an initial estimate of it or taken as $r = -1/2$.

4. MOM estimators (Hosking & Wallis, 1987) of the scale σ and shape ξ parameters of the GPD distribution are given by:

$$\hat{\sigma} = \frac{1}{2}\bar{x}\left(\frac{\bar{x}}{s^2} + 1\right), \quad \hat{\xi} = \frac{1}{2}\left(\frac{\bar{x}^2}{s^2} - 1\right)$$

where $\bar{x}$ and s^2 are the sample mean and variance.

5. PWM estimators provide estimates with smaller bias and variance than MLE when the sample size is less than 500 (Hosking & Wallis 1987). The PWM's of the random variable X with a distribution function $G \equiv G(x) = P(X \leq x)$ is defined as:

$$M_{l,j,k} = E\left[X^l F^j (1-F)^k\right] = \int_0^1 [x(F)]^l F^l (1-F)^k dF$$

where l, j and k are real numbers. For $j = k = 0$ and l a nonnegative integer, $M_{l,0,0}$ is the classical moment of order l.

6. The estimator suggested by Pickands (1975) (referred to as 'Pick') is based on the ascending order statistics $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ from an independent sample of size n and is defined as:

$$\hat{\xi}_{n,k}^{Pick} = \frac{1}{log2}\log\left(\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n}}\right), \text{ for } k = 1, \ldots, [n/4]$$

This estimator is largely dependent on k and provides a large asymptotic variance (e.g. (Dekkers & Haan, 1989; Segers, 2005; Yun, 2002).

7. There are many MGF statistics that can be used for GPD parameter estimation, such as Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling (see Luceño, 2006).

List of Tables

List of Figures