MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 3, pp. 819–831 (2025)

An automated threshold selection procedure for generalized Pareto distribution with application to rainfall dataset

Alif F. K.*, Ali N., Safari M. A. M.

Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
**Corresponding author: gs67618@student.upm.edu.my*

(Received 10 February 2025; Revised 30 August 2025; Accepted 31 August 2025)

In hydrological datasets, particularly rainfall, the study of extreme values is crucial. The appropriate analysis of such datasets can provide vital information about the return levels of extreme rainfall, which can play a significant role in disaster prevention. In many situations, the GPD has been a well-respected option for studying extreme data; nonetheless, there are still concerns about the GPD's threshold selection method. The commonly used Mean Residual Life (MRL) plot technique for threshold selection in Generalized Pareto Distribution (GPD) analysis suffers from subjectivity and requires extensive prior knowledge, limiting its reproducibility. This paper introduces a straightforward, computationally inexpensive, and automated procedure for threshold selection. By employing interval-based candidate thresholds and goodness-of-fit (GOF) tests, the proposed method determines the optimal threshold that maximizes the p-value, enhancing objectivity and accuracy. Several combinations of estimation methods and GOF tests were investigated, with the CVM-Lmoment combination emerging as the most robust. Through extensive simulation studies, our approach demonstrated significant improvements in reducing bias and RMSE compared to traditional methods. The application of the proposed methodology to a rainfall dataset from South-West England confirmed its robustness and practical utility, making it a valuable tool for extreme value modeling and disaster management.

Keywords: *generalized Pareto distribution; threshold selection; goodness of fit; L-moments; extreme values; return level; extreme rainfall.*

2010 MSC: 00B20, 62P12, 62-07, 62G32 **DOI:** 10.23939/mmc2025.03.819

1. Introduction

Extreme value is a crucial area of statistics that examines values near the very end of a dataset that may hold essential information. The lowest and greatest values of a given property are called extreme values. For instance, the measure of rainfall in mm in a certain area or the record of a country's or city's annual highest and lowest temperature. Extreme value analysis (EVA) is a statistical method that determines the likelihood that extreme values will occur using measured or observed data and a few key presumptions [1]. There are two common distributions available for dealing with EVA: the Generalized Pareto Distribution (GPD) and the Generalized Extreme Value (GEV).

The analysis of extreme events is crucial in fields such as finance, hydrology, and environmental sciences, where understanding the tail behavior of distributions is essential for assessing risk. The Peak Over Threshold (POT) method is a powerful approach within Extreme Value Theory (EVT) that focuses on modeling these rare events by examining data points that exceed a specified threshold. By concentrating on these exceedances, the POT method effectively captures the most extreme values in a dataset, providing a more accurate representation of the underlying risk associated with rare occurrences [2]. The selection of an appropriate threshold is a critical step in the POT method, as it determines the number of exceedances and influences the stability and reliability of the resulting statistical model [3]. Proper threshold selection ensures that the model captures the extremal behavior while maintaining a sufficient sample size for meaningful analysis.

Once the threshold is established, the exceedances are modeled using the GPD, which was first introduced by John Pickands in 1975 [4]. The GPD is well-suited for this purpose because it is specifically designed to describe the tail behavior of distributions, making it ideal for analyzing extreme values [5]. The GPD is characterized by two key parameters: the shape parameter, which controls the heaviness of the tail, and the scale parameter, which determines the spread of the exceedances above the threshold [2,5,6]. Accurate estimation of these parameters is essential for reliable predictions of extreme events. Several methods, including Maximum Likelihood Estimation (MLE) and Bayesian approaches, have been developed to estimate these parameters, each offering different advantages depending on the context of the application [7].

The combination of the POT method and GPD provides a robust framework for understanding and predicting the occurrence and magnitude of extreme events. This approach has been widely applied in various domains, such as assessing the risk of financial market crashes, forecasting extreme weather conditions, and evaluating the reliability of engineering structures [6]. The ability of the POT-GPD model to accurately characterize the tail behavior of distributions makes it an invaluable tool for risk management and decision making, particularly in scenarios where extreme outcomes can have significant consequences [8].

Subjectivity in determining the threshold point is one of the main problems with GPD [9–15]. The visual method of selecting an appropriate threshold value can result in a number of issues. The proper interpretation of threshold choice plots, such the Mean Residual Life plot (see [6]) for examples), is a prerequisite for achieving a good model fit with these visual approaches. The threshold should be set high enough to allow the GPD to approximate the excesses accurately without causing bias, but not so high that a drop in sample size (the number of exceedances) considerably increases the variance of the estimator [15]. The most common method for threshold selection is still to use graphical methods (see, for example, [6] for a detailed description of these methods and [16] for an application to some of the series used in this work). The inability to quantify threshold selection uncertainty and the inability to estimate the uncertainty of quantiles with extended return periods are two drawbacks of graphic techniques [9].

Recent studies have highlighted the need for more robust and objective threshold selection techniques. For instance, Ref. [15] proposed an automated threshold selection method based on batched return level mapping, demonstrating improved reproducibility but requiring significant computational resources. Reference [9] introduced a goodness of fit (GOF) p-value approach, effectively minimizing bias in threshold selection but facing limitations when applied to datasets with complex structures. Reference [17] evaluated several automated threshold selection methods for hydrological extremes across multiple scales, offering a comprehensive comparison of their performance under varying conditions. Reference [11] investigated threshold selection methods in wave extreme value analysis, emphasizing the necessity of tailoring approaches to specific environmental applications. Reference [13] proposed an automated sampling method for identifying independent and identically distributed samples in long-term wave processes, which was effective in analyzing extreme wave heights caused by tropical and non-tropical cyclones. However, its application in regions with complex weather conditions requires further investigation. Reference [3] noted that extreme and non-extreme events often arise from different physical processes, necessitating separate modeling approaches. Reference [18] employed detrended fluctuation analysis (DFA) to study long-range correlations in wave height series, showing that extreme events minimally affect overall trends, while non-extreme data significantly influence long-term behavior. Furthermore, Ref. [19] demonstrated that the multifractal DFA (MF-DFA) method provided objective thresholds for extreme precipitation, outperforming traditional methods. Reference [11] extended the application of DFA-based approaches to large-scale regions characterized by spatiotemporal heterogeneity, showcasing their potential for robust threshold estimation. These studies collectively highlight the importance of leveraging both statistical properties and physical processes in threshold selection methodologies, while also underlining the ongoing challenges of computational complexity and sensitivity to dataset characteristics.

The novelty of this study lies in its development of an automated threshold selection procedure that eliminates the subjectivity inherent in traditional methods while maintaining computational efficiency. This will remove the requirement for the user of this methodology to take into account human error, subjective measurements, or any form of prior understanding with sophisticated graphical metrics. By focusing on a direct comparison with the widely used MRL plot technique, the proposed approach demonstrates significant improvements in both objectivity and accuracy. Leveraging interval-based candidate thresholds combined with goodness-of-fit tests, this methodology offers a computationally inexpensive and effective alternative for extreme value modeling. The robustness and practicality of this method are validated through extensive simulation studies and its application to a real-world environmental dataset, highlighting its potential to advance the state-of-the-art in extreme value analysis.

For the application of our approach in real-life scenario, we will use a rainfall dataset. The dataset consists of daily rainfall accumulations at a site in south-west England from 1914 to 1962, which contains of 17531 observations [20].

2. Methodology

2.1. Theoretical background

The GPD models consist of two parameters, which are scale (σ), and shape (ξ). The threshold is denoted as u. The cumulative distribution function (CDF) representing the relationship among these parameters is as follows:

$$G(x;\sigma,\xi) = 1 - \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \quad \xi \neq 0,$$
$$G(x;\sigma,\xi) = 1 - \exp\left[-\left(\frac{x-u}{\sigma}\right)\right], \quad \xi = 0,$$

where for $\xi \geqslant 0 : x > u$ and for $\xi < 0 : u \leqslant x < u - \frac{\sigma}{\xi}$. The parameters $\sigma > 0$, $-\infty < \xi < \infty$ and $-\infty < u < \infty$. Here x is the random variable. To estimate σ and ξ parameters, our study will consider two of the very well-known and effective estimation technique, which are Maximum Likelihood Estimation (MLE) and L-moment [2,21]. MLE is a statistical method that uses observed data to calculate a supposed probability distribution's parameters [22]. To do this, one must optimize a likelihood function in order to maximize the likelihood of the observed data under the similar statistical model [22]. Where the likelihood function is maximized in the parameter space is known as the maximum likelihood estimate [22]. Now, let's consider probability distribution function (pdf) of GPD when $\xi \neq 0$,

$$g(x;\sigma,\xi) = \frac{1}{\sigma}\left(1 + \xi\frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}-1},$$

so, the likelihood function will be:

$$L(\sigma,\xi) = \prod_{i=1}^{n} \frac{1}{\sigma}\left(1 + \xi\frac{x_i-u}{\sigma}\right)^{-\frac{1}{\xi}-1}$$

taking the logarithm to obtain the log-likelihood function:

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

Now, let us consider the other case when $\xi = 0$. The pdf of GPD of such case can be represented as:

$$g(x;\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x-u}{\sigma}\right).$$

The likelihood function in this case is:

$$L(\sigma) = \prod_{i=1}^{n}\frac{1}{\sigma}\exp\left(-\frac{x_i-u}{\sigma}\right)$$

and the log-likelihood function is:

$$\log L(\sigma) = -n \log \sigma - \frac{1}{\sigma}\sum_{i=1}^{n}(x_i - u).$$

The likelihood function of the GPD is used to estimate the parameters ξ and σ by maximizing the likelihood (or equivalently, the log-likelihood) function given the observed data.

To illustrate the shape of a probability distribution, statisticians employ a set of statistics known as L-moments [21, 23–25]. Comparable to traditional moments, these are linear combinations of order statistics, or L-statistics. A continuous stochastic variable, such as Y, has a probability distribution whose quantile function (QF) is defined as

$$Q_Y(p) = F_Y^{-1}(p) = \inf\{y \in \mathbb{R} \colon F_Y(y) \geqslant p\}, \quad \text{for } 0 \leqslant p \leqslant 1,$$

where $F_y(Y) = p$ is the cumulative distribution function of Y [26]. A compilation of the L-moment theory was made by [21]. The rth L-moment in terms of the QF is defined by

$$L_r = \int_0^1 Q_y(p)\, P_{r-1}^*(p)\, dp,$$

where

$$P_{r-1}^*(p) = \sum_{k=0}^{r}\left[-1^{r-k}\binom{r}{k}\binom{r+k}{k}p^k\right]$$

is the Legendre polynomial with r^{th} shift and L_r is the r^{th} L-moment. More detailed information is available in [26].

Our methodology uses p-value of the goodness of fit (GOF) test along with the estimation technique to find the best possible threshold from the dataset which defines the extreme values. The statistical incompatibility of the data with the null hypothesis is larger with smaller p-values, if the underlying assumption used to calculate the p-values is correct, according to [27]. With 0 representing complete incompatibility and 1 representing perfect compatibility, this metric may be understood as evaluating how well the model fit the data [28]. Parameters for GPD must be estimated in order to perform GOF testing. The Kolmogorov–Smirnov (KS) and Cramér–von Mises (CVM) tests are the two GOF tests that the study will take into account while choosing the ideal threshold point.

A nonparametric test for determining if two continuous, one-dimensional probability distributions are equal is the KS test, sometimes referred to as the KS test.

$$\mathrm{KS} = \max_{1 \leqslant i \leqslant n}\left(\left|G\big(x_{(i)}; \hat{u}, \hat{\sigma}, \hat{\xi}\big) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - G\big(x_{(i)}; \hat{u}, \hat{\sigma}, \hat{\xi}\big)\right|\right),$$

where n is the sample size, $x_{(i)}$ denotes the i-th order statistic of the sample, meaning the i-th smallest value in the sorted sample data, $G\big(x_{(i)}; \hat{u}, \hat{\sigma}, \hat{\xi}\big)$ is the CDF of the GPD evaluated at $x_{(i)}$, with the parameters $\hat{u}$, $\hat{\sigma}$, and $\hat{\xi}$ being the estimated threshold, scale, and shape parameters, respectively. By using the two-sample KS test or the one-sample KS test, one can ascertain if two samples or one sample came from the same reference probability distribution. The KS statistic quantifies the empirical distribution functions of two samples, or the sample's empirical distribution function and the reference distribution's CDF.

On the other hand, a popular set of goodness of fit statistics for fitting to a continuous distribution is the CVM family [29],

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[G\big(x_{(i)}; \hat{u}, \hat{\sigma}, \hat{\xi}\big) - \frac{2i-1}{2n}\right].$$

The CVM is frequently used to evaluate how well observed sample data fits into a given model [30]. CVM is used to evaluate how well a cumulative distribution function fits to a given empirical distribution function or to compare two empirical distributions. CVM should function in GPD with similar integrity as comparable to the KS test, as [31] has discussed.

2.2. Automated threshold selection method

The interval-based methodology aims to systematically divide the dataset into manageable segments to identify candidate thresholds for extreme value analysis. This approach ensures that threshold selection is both objective and computationally efficient by utilizing equal-sized intervals and statistical criteria for refinement. By focusing on intervals, the methodology reduces subjectivity, leverages inherent dataset structure, and improves the reproducibility of results.

The methodology we propose for the selection of optimum threshold in GPD is as follows:

1. Divide the dataset into N equal interval after sorting it in ascending order. We found that when N is around 150 to 200 the results for the first candidate threshold u_1 (see point 2) seem to converge to a single point.
2. Determine the means $M_1, \ldots, M_N$ for each interval. Choose the means $M_i, \ldots, M_N$ that exceed the dataset's first quantile, Q_1. Find the average from these newly chosen means. This average will be the first candidate threshold, u_1, where $u_1 = \frac{[M_i > Q_1] + \ldots + M_N}{N}$.
   We do this because a rain dataset's first quantile typically has substantially lower values than the dataset's remaining values. Therefore, it increases the computational load of selecting the appropriate candidate thresholds by artificially increasing the number of candidate thresholds.
3. The end point of the threshold is denoted as u_n and u_n is a significantly large value which depends on the dataset, usually around the 5th most largest value as we got good results through our simulation study when u_n is around this region. But it obviously depends on the structure of the dataset.
   The candidate thresholds $u_1, u_2, \ldots, u_n$ are series of values with equal differences between them. Simulation studies are used to determine the value of n, which is chosen to be in the $100-300$ range. In general, better results will come from larger values of n. Our study's findings consistently appear to converge when n is near the 100ı300 range. Anything between 100 and 300 is acceptable because it won't significantly impact the outcome. The study did take into consideration the conclusion of [14], which states that when $n = 100$, better results were obtained.
4. Estimate the parameters of GPD utilizing the candidate thresholds $u_1, u_2, \ldots, u_n$. To estimate the GPD parameters, only the data where $x_i > u_0$ will be used.
5. Apply GOF test to achieve the p-values. The ideal threshold u_0 for the GPD will be chosen as the one that maximized the p-value of the GOF test.

2.3. Simulation study

The objective of the simulation study is to find the best combination between the GOF test and the estimation technique, which best serves our methodology. For the simulation study, the exponential GPD composite model was chosen to generate the dataset. The exponential GPD model that this study is adapting is also supported by research by [32],

$$f(x|\theta) = \begin{cases} \rho\, f_1^*(x|\theta), & \text{if} \quad x \leqslant u_0, \\ (1-\rho)\, f_2^*(x|\theta), & \text{if} \quad x > u_0. \end{cases}$$

where

$$\rho = \frac{\alpha(1 - e^{-\lambda u_0})}{\alpha + e^{-\lambda u_0}}.$$

Here α is the tail index and λ is the rate. The distribution is divided by the model at the threshold value u_0. The pdf of the parent distribution is defined by,

$$f_1^*(x|\theta) = \frac{f_1(x|\theta)}{F_1(x|\theta)},$$

$$f_2^*(x|\theta) = \frac{f_2(x|\theta)}{1 - F_2(x|\theta)}.$$

The data in the lower tail are modeled by the pdf $f_1(x|\theta)$, which is considered to be exponential distribution with λ as the rate. The observations in the higher tail are modeled by the pdf $f_2(x|\theta)$, which is the GPD [33–35]. The ρ the mixing weight [35]. The details about composite models, specially exponential pareto composite model is discussed in detail by [33] and [34].

The Rstudio programming environment was used to carry out the simulation study. Following are the steps for simulation study:

1. Generate random number from exponential GPD composite model with the parametric values as follows: Rate(λ) $= 0.303$ for the exponential portion, Threshold(u) $= 35$, Scale(σ) $= 12.5$, and Shape(ξ) $= 0.010$ for the GPD component. With small, moderate, and big datasets in mind, the sample sizes are considered to be 1000, 10000, and 50000, respectively. The *gendist* package was used to generate random values for the Exponential GPD composite model. The details about the functionality of the *gendist* package is discussed by [35].
2. Apply the automated threshold selection method (see section 2.2) to determine the optimum threshold value in each sample.
3. Repeat the steps 1 and 2 for 1000 times for each sample size and store the results to determine the bias and RMSE.
4. For comparison we apply the classic graphical method (MRL plot) using the same data generating technique and settings in step 1 to determine the optimum threshold.
5. As MRL plot technique is a subjective method, it is very difficult to repeat step 4 a thousand times. Hence for our study, we repeated the step 4 ten times and stored the results to determine the bias and RMSE.

3. Results and discussion

3.1. Selection of optimal pair of GOF and estimator

Selecting the ideal GOF test and the best estimating method are crucial for determining the ideal threshold in GPD. As mentioned in Methodology's section, this study considered two estimating methods in addition to two GOF tests. A total of 4 combinations of estimators and GOF tests have been taken into consideration. At this point, the investigation proceeded to calculate the bias and RMSE for each estimated parameter in each combination. This should give a broad picture of the performance for all sample sizes. It is important to remember that the most exact and consistent outcomes are the ones that this research is most interested in. From Table 1, the most consistent result is obtained by thoroughly examining the CVM's combined results with L-moments. One could argue that it doesn't always provide the most accurate outcome. However, a closer look at the estimated parameter values, bias and RMSE reveals that it is not far off from the true parameters. In terms of bias and RMSE for u and σ, the combination of CVM and L-moment may not yield the greatest results, but for the shape parameter, it has shown to be the most accurate. Considering the bias and RMSE of u, and σ, the results are not entirely unsatisfactory. They are regarded as being at the upper end of the accuracy spectrum. The overview of the simulation study indicates that any combination from KS-MLE or CVM-Lmoment is a good choice when the sample size is small.

As we move on to sample sizes of 10 000 (intermediate datasets), it becomes difficult to identify the perfect combination. As of now, KS and the L-moment appear to be the ideal combination. As with all other combinations in this sample size range, the results produced by combining KS and MLE are not all that different from the results produced by combining KS and L-moment. Using CVM and L-moment together yields results that are closest to the actual threshold of 35, with an estimated threshold of 33.685. Moreover, this combination provides the lowest bias and RMSE for ξ. Consequently, it can be said that for modest datasets, both the KS-Lmoment and CVM-Lmoment are optimal.

Finally, when it comes to large datasets (sample size = 50000), it becomes extremely challenging to choose the optimum combination of GOF test and method of estimation technique. Strictly examining the dataset, one can be persuaded that the combination of CVM and MLE is adequate. According to [36], it is a well-known phenomena that the accuracy level rises as sample size does. As a result, when the sample size is relatively large, these parallels in the results are visible. Looking at the CVM-

Table 1. The table reports the results obtained from the simulation study including the RMSE, bias and average P-value at different sample sizes for KS and CVM for each estimated parameter.

Sample size	GOF	Method of estimation	$E[\hat{u}]$	$E[\hat{\sigma}]$	$E[\hat{\xi}]$	p-value	u		σ		ξ	
							RMSE	Bias	RMSE	Bias	RMSE	Bias
1000	KS	MLE	33.380	13.018	0.002	0.986	1.668	1.620	0.534	−0.518	0.011	0.008
		Lmoment	32.106	13.316	−0.003	0.988	2.920	2.894	0.825	−0.816	0.014	0.013
	CVM	MLE	33.013	13.757	−0.026	0.981	2.028	1.987	1.265	−1.257	0.037	0.036
		Lmoment	33.004	13.078	0.009	0.988	2.035	1.996	0.588	−0.578	0.006	0.001
10000	KS	MLE	34.101	13.151	0.008	0.986	0.985	0.899	0.651	−0.650	0.003	0.002
		Lmoment	34.636	13.206	0.007	0.989	0.540	0.362	0.707	−0.706	0.004	0.003
	CVM	MLE	33.868	13.255	0.002	0.981	1.202	1.132	0.756	−0.755	0.008	0.008
		Lmoment	33.685	13.179	0.008	0.990	1.373	1.315	0.681	−0.680	0.003	0.002
50000	KS	MLE	34.254	13.145	0.010	0.982	0.850	0.745	0.645	−0.645	0.001	0.000
		Lmoment	34.270	13.161	0.009	0.988	0.834	0.730	0.661	−0.661	0.001	0.001
	CVM	MLE	33.635	13.142	0.010	0.980	1.421	1.365	0.642	−0.642	0.001	0.000
		Lmoment	33.975	13.100	0.012	0.989	1.106	1.025	0.600	−0.600	0.002	−0.002

Lmoment result, none of the bias and RMSE are the lowest, with the exception of σ. This is not a huge issue, though, because the predicted parameters from this CVM-Lmoment pair are close enough to the exact parameters to be usable, thus even while it might not be the most accurate, it cannot be stated that it won't function with a larger dataset. For the parameters that this pair estimated, the bias and RMSE are also relatively low.

Generally, if we look at Table 1, as the sample size increases, the results through the table become more consistent. The values of RMSE and bias also reduces for every single combinations of GOF and Estimation technique. Ultimately, the results from the simulation study affirm that moving forward with the CVM-Lmoments combination is a rational and well-supported choice.

3.2. Comparison with MRL plot technique

The MRL plot method is a highly subjective graphical style that requires some prior knowledge to understand. References [2,17] include further information regarding MRL plots. The inability to repeat the MRL plot in a simulation study as much as one would like due to its high time consumption is one of its main drawbacks. However, analyzing a single MRL plot produced from a single set of random variables would be unfair. The data used in this comparison study was produced using the identical settings from the preceding section of the Exponential GPD model. With small, moderate, and big datasets in mind, the sample sizes are 1000, 10000, and 50000, respectively. Ten MRL plots have been produced for each sample size, and the investigation is still ongoing to ascertain the bias and RMSE for $\hat{u}$. This should give a general notion of how the accuracy of the two techniques compares.

Table 2. The table reports the comparison between MRL Plot and proposed Automated Threshold selection approach in terms of threshold, RMSE and bias based on the exact same parameter settings as of Table 1.

Sample size	MRL plot			Automated approach		
	$E[\hat{u}]$	RMSE	Bias	$E[\hat{u}]$	RMSE	Bias
1000	30.250	4.781	4.750	33.004	2.035	1.997
10000	31.200	3.932	3.800	33.685	1.373	1.315
50000	32.000	3.066	3.000	33.975	1.106	1.025

When compared to the graphical method, our automated threshold selection strategy appears to have reduced RMSE and bias values. Therefore, it is evident from both tables that our approach outperformed the MRL plot technique of the POT method.

4. Application to the rainfall data

Upon examining the South–West England daily rainfall dataset and applying the automatic threshold selection technique with the P-value of the GOF test (refer to Automated Threshold Selection Method), we obtain 33.557 mm as the value of $\hat{u}_0$. This cutoff setting is sensible and practical given that the dataset's maximum value is 86.6 mm. Scale and shape are calculated to be 8.461 and 0.169, respectively, using L-moment. In order to address the validity of such out, let us examine the graphs in Figure 1. The predicted model obtained by the automated threshold selection method has a very good fit, as indicated by the Probability plot and the Q-Q plot. Additionally, we contrast our automated threshold selection method with the MRL plot method that Coles et al. (2001) examined. According to [6], 30 mm is the threshold value, shown in Figures 2 and 3. According to [14], an automatic threshold is typically defined as a value over which the plot becomes linear, subject to sampling error.

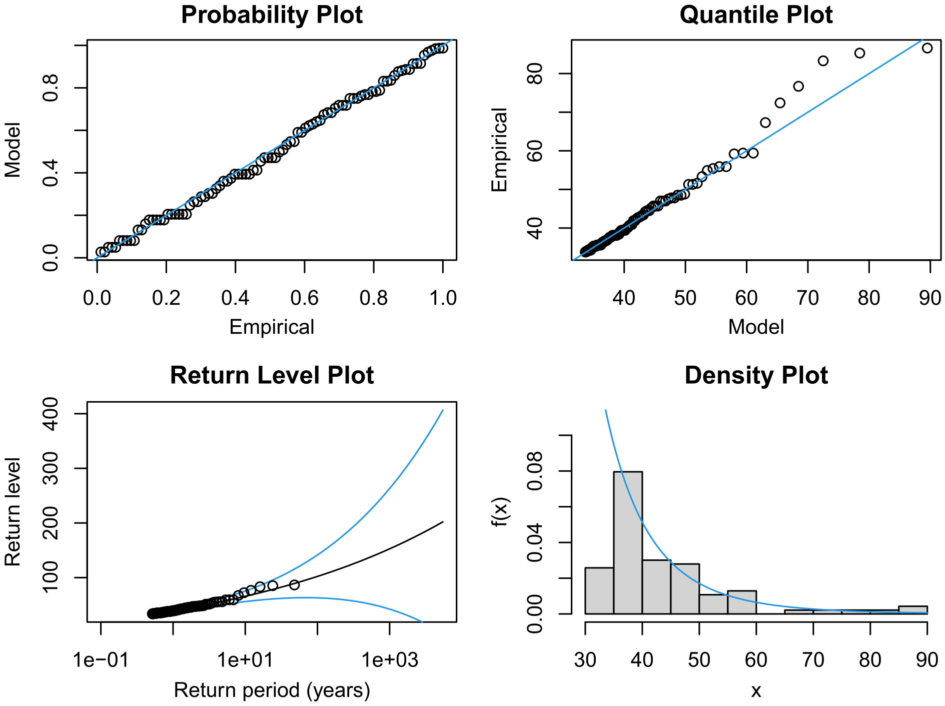

Fig. 1. Plots showing diagnostic information for the fitted GPD model when the threshold is chosen using our automated threshold selection approach on the south–west England rainfall dataset. Return level in the third figure is a reference to rainfall (mm). The rainfall (mm) in the fourth figure is denoted by x, and its probability density is represented by $f(x)$.

According to [6], linearity also happens between 30 and 60 mm, but for 60 mm few data points are above cutoff with sampling error [14]. Hence 30 mm is a reasonable choice. We may also use a similar justification to support our automated threshold selection of 33.557 mm. It is a challenge in the MRL plots. Which particular area on the graph represents the best outcome depends on the researcher. The difficulties in interpreting the MRL plot and its subjective nature are well-illustrated in [14].

Table 3. The estimated parameter of GPD model fitted to rainfall data.

GOF	Method of estimation	$\hat{u}$	$\hat{\sigma}$	$\hat{\xi}$
KS	MLE	35.600	10.608	0.057
	L-moment	36.057	9.065	0.160
CVM	MLE	59.557	57.674	−2.132
	L-moment	33.557	8.461	0.169

Table 3 shows the estimated parameter values using CVM with MLE, which is one thing to note. Based on MLE and CVM, the estimated value of u is about 60 mm. According to [14], the MRL plot for the rainfall dataset in South-West England is linear (with sampling error) over 30 and 60 mm, as we can see in Figure 2. Thus, the combination of CVM and MLE is more oriented in the 60 mm range. In comparison to other combinations, the estimated values of the shape and scale parameters deviate considerably for CVM-MLE and hence may be considered less reliable.

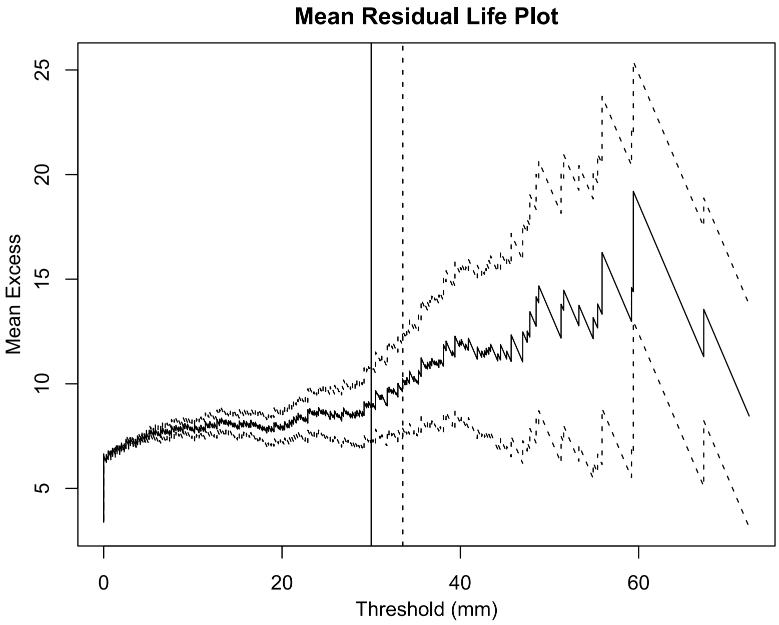

Fig. 2. An illustration of the MRL plot derived from daily rainfall data. The dashed line represents automated threshold selection approach. The solid line represents threshold value using the MRL plot technique.

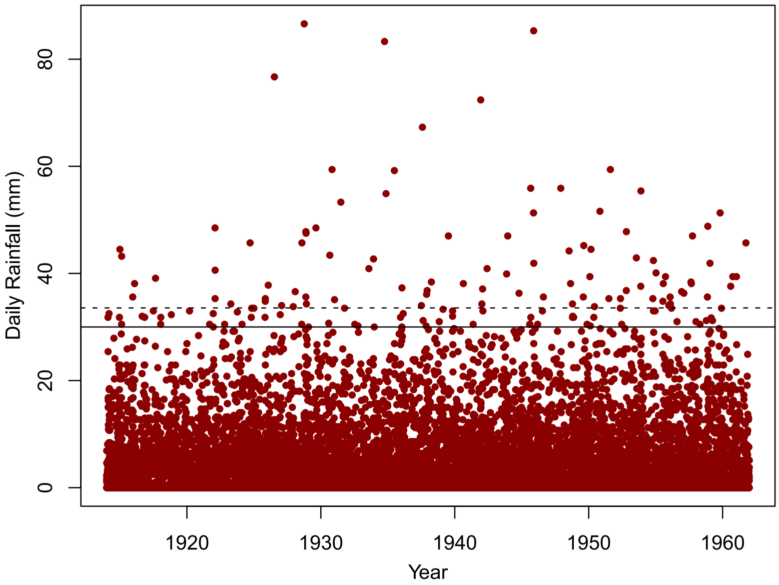

Fig. 3. Daily rainfall dataset in mm plotted versus time in years in a scatter plot. The automated method's threshold selection is represented by the dashed line, and [6] suggested threshold is shown by the solid line.

Table 3 shows that the KS yields consistent estimates for the estimated threshold ($\hat{u}$), ranging from 35.600 to 36.057 mm. However, the CVM method with MLE produces a notably higher value (59.557 mm), which appears to be an outlier. The $\hat{\sigma}$ and $\hat{\xi}$ parameters also show variations across methods, with the CVM-MLE method again diverging significantly from the others. Most methods estimate a positive shape parameter, indicating a heavy-tailed distribution, except for the CVM-MLE method which suggests a bounded distribution.

The 95% confidence intervals presented in Table 4 provide further insight into the reliability of these estimates. Most methods show overlapping confidence intervals for the threshold parameter, ranging from approximately $25-30$ mm at the lower bound to $71-84$ mm at the upper bound. The L-moment method for KS shows the narrowest interval $[30.450, 84.270]$, potentially indicating higher precision, while the CVM method with L-moment shows the widest interval $[26.450, 82.060]$, reflecting greater uncertainty.

Table 4. The 95% confidence interval for all the combinations of methods of estimation and GOF test using bootstrap percentile method.

GOF	Method of estimation	$\hat{u}$	95% confidence interval
KS	MLE	35.600	[27.600, 81.890]
	Lmoment	36.057	[30.450, 84.270]
CVM	MLE	59.557	[25.300, 71.000]
	Lmoment	33.557	[26.450, 82.060]

Given these results, the suggested threshold of 33.557 mm based on the methodology appears reasonable, falling within or close to the lower bounds of most confidence intervals. This aligns with the conclusion that the genuine threshold value should be in the range of 33 mm to 36 mm. However, the wide range of confidence intervals, spanning approximately 50 mm in most cases, highlights significant uncertainty in the threshold estimation. This uncertainty should be carefully considered in any subsequent analysis or decision-making processes related to extreme rainfall events.

In conclusion, while the analysis supports a threshold value in the range of 33 to 36 mm for defining extreme rainfall events, the consistency between KS and MLE methods provides more confidence in their estimates compared to the CVM method. The wide confidence intervals underscore the need for cautious interpretation and the importance of considering upper bounds (up to approximately 84 mm) in risk assessments to account for potential underestimation of extreme events. Future studies might benefit from incorporating additional estimation methods or exploring the sensitivity of results to different threshold choices within the identified range.

The 95% confidence intervals were calculated using the bootstrap percentile method. We used 1000 bootstrap samples for our confidence intervals. Each sample used our automated threshold selection technique to estimate the optimum threshold value. The step by step protocol is discussed in detail by [14]. See [37] for deep understanding of the bootstrap methodology. From Table 4, the combination of CVM-MLE has the lowest uncertainty, but as we can observe in Table 3, the estimated parameters including the threshold value are a bit unrealistic. With the exception of this pair, every other pair's 95% confidence interval range is similar. Moreover, all the values of $\hat{u}$ falls within the 95% confidence interval range, which makes them statistically significant. Therefore, it is reasonable to conclude that while our simulation analysis indicates that the CVM-Lmoment pair should yield the ideal outcomes, other pairs aside from CVM-MLE should not be too far off.

5. Conclusion

This study introduces an automated approach for threshold selection in the GPD framework, addressing critical challenges in extreme value analysis. The methodology, which integrates GOF tests with systematic threshold selection, offers an objective and computationally efficient alternative to traditional methods like the MRL plot. Through this research, significant contributions have been made to enhance the reliability and practicality of extreme value modeling.

The simulation studies conducted in this research highlight the effectiveness of the proposed method in reducing bias and RMSE across different sample sizes. By systematically identifying thresholds that optimize model fit, the methodology ensures accurate parameter estimation and overcomes the subjectivity associated with graphical techniques. This improvement is particularly evident in datasets with varying sample size, where the method consistently outperforms the MRL plot in terms of accuracy and reproducibility.

The practical utility of the approach is demonstrated through its application to South–West England's rainfall dataset. The results showcase the method's robustness in real-world scenarios, providing reliable estimates for critical thresholds and enabling precise modeling of extreme rainfall events. These findings underscore the method's potential in applications requiring accurate risk assessments, such as hydrological studies, disaster management, and climate adaptation planning. By delivering a reliable framework for threshold selection, this research addresses the practical needs of stakeholders in these domains.

Beyond its immediate contributions, this study offers a foundation for further advancements in extreme value analysis. The approach's simplicity and adaptability make it suitable for diverse applications, including financial risk assessment, environmental monitoring, and engineering reliability studies. Future research could explore integrating additional estimation techniques or alternative GOF criteria to enhance model performance further. Moreover, extending the methodology to account for challenges like data heterogeneity and non-stationarity would broaden its applicability to dynamic and complex datasets.

In conclusion, this research significantly advances the field of extreme value theory by providing a structured, objective, and efficient threshold selection methodology. By addressing the limitations of conventional methods, it equips researchers and practitioners with a robust tool for modeling extremes across a range of disciplines. This contribution not only enhances the accuracy and reliability of extreme value models but also paves the way for innovative approaches in risk management and decision-making under uncertainty. The outcomes of this study have the potential to influence future developments in statistical modeling and its application to real-world challenges.

Acknowledgments

As authors we acknowledge the support of Special Graduate Research Allowance Scheme (SGRA) scholarship provided by University Putra Malaysia through the PUTRA GRANT GP/2023/9753100. We also whole heartily thank the referees.

[1] Benstock D., Cegla F. Extreme value analysis (EVA) of inspection data and its uncertainties. NDT & E international. **87**, 68–77 (2017).

[2] Davison A. C., Smith R. L. Models for exceedances over high thresholds. Journal of the Royal Statistical Society Series B: Statistical Methodology. **52** (3), 393–425 (1990).

[3] Scarrott C., MacDonald A. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT – Statistical journal. **10** (1), 33–60 (2012).

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

[5] Embrechts P., Klüppelberg C., Mikosch T. Modelling Extremal Events: for Insurance and Finance. Vol. 33, Springer Science & Business Media (2013).

[6] Coles S., Bawa J., Trenner L., Dorazio P. An introduction to Statistical Modeling of Extreme Values. Vol. 208, Springer (2001).

[7] Beirlant J., Goegebeur Y., Teugels J. L., Segers J. Statistics of Extremes: Theory and Applications. John Wiley & Sons (2006).

[8] McNeil A. J., Frey R. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Journal of Empirical Finance. **7** (3–4), 271–300 (2000).

[9] Solari S., Egüen M., Polo M. J., Losada M. A. Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. Water Resources Research. **53** (4), 2833–2849 (2017).

[10] Wu G., Qiu W. Threshold Selection for POT Framework in the Extreme Vehicle Loads Analysis Based on Multiple Criteria. Shock and Vibration. **2018** (1), 4654659 (2018).

[11] Liu H., Yang F., Wang H. Research on Threshold Selection Method in Wave Extreme Value Analysis. Water. **15** (20), 3648 (2023).

[12] Fukutome S., Liniger M., Süveges M. Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland. Theoretical and Applied Climatology. **120**, 403–416 (2015).

[13] Liang B., Shao Z., Li H., Shao M., Lee D. An automated threshold selection method based on the characteristic of extrapolated significant wave heights. Coastal Engineering. **144**, 22–32 (2019).

[14] Thompson P., Cai Y., Reeve D., Stander J. Automated threshold selection methods for extreme wave analysis. Coastal Engineering. **56** (10), 1013–1021 (2009).

[15] Bader B., Yan J., Zhang X. Automated threshold selection for extreme value analysis via Goodness-of-Fit tests with application to batched return level mapping. Preprint arXiv:1604.02024 (2016).

[16] Solari S., Losada M. A. A unified statistical model for hydrological variables including the selection of threshold for the peak over threshold method. Water Resources Research. **48** (10), W10541 (2012).

[17] Curceac S., Atkinson P. M., Milne A., Wu L., Harris P. An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales. Journal of Hydrology. **585**, 124845 (2020).

[18] Ozger M. Scaling characteristics of ocean wave height time series. Physica A: Statistical Mechanics and its Applications. **390** (6), 981–989 (2011).

[19] Du H., Wu Z., Zong S., Meng X., Wang L. Assessing the characteristics of extreme precipitation over northeast China using the multifractal detrended fluctuation analysis. Journal of Geophysical Research: Atmospheres. **118** (12), 6165–6174 (2013).

[20] Coles S. G., Tawn J. A. Modelling Extremes of the Areal Rainfall Process. Journal of the Royal Statistical Society: Series B: Statistical Methodology. **58** (2), 329–347 (1996).

[21] Hosking J. R. M. L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Journal of the Royal Statistical Society Series B: Statistical Methodology. **52** (1), 105–124 (1990).

[22] Rossi R. J. Mathematical Statistics: An Introduction to Likelihood Based Inference. John Wiley & Sons (2018).

[23] Asquith W. H. Distributional Analysis with L-moment Statistics using the R Environment for Statistical Computing. CreateSpace Scotts Valley, CA, USA (2011).

[24] Hosking J. R. M. Moments or L Moments? An Example Comparing two Measures of Distributional Shape. The American Statistician. **46** (3), 186–189 (1992).

[25] Hosking J. On the characterization of distributions by their L-moments. Journal of Statistical Planning and Inference. **136** (1), 193–198 (2006).

[26] van Staden P. J., Loots M. T. Method of L-moment estimation for the generalized lambda distribution. Proceedings of the Third Annual ASEARC Conference. 1–4 (2009).

[27] Wasserstein R. L., Lazar N. A. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician. **70** (2), 129–133 (2016).

[28] Greenland S., Senn S. J., Rothman K. J., Carlin J. B., Poole C., Goodman S. N., Altman D. G. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. **31** (4), 337–350 (2016).

[29] Choulakian V., Lockhart R. A., Stephens M. A. Cramér – von Mises statistics for discrete distributions. The Canadian Journal of Statistics. **22** (1), 125–137 (1994).

[30] Ahsan-ul-Haq M. A new Cramèr – von Mises Goodness-of-fit test under Uncertainty. Neutrosophic Sets and Systems. **49** (1), 262–268 (2022).

[31] Chu J., Dickin O., Nadarajah S. A review of goodness of fit tests for Pareto distributions. Journal of Computational and Applied Mathematics. **361**, 13–41 (2019).

[32] Martins A. L. A., Liska G. R., Beijo L. A., de Menezes F. S., Cirillo M. Â. Generalized Pareto distribution applied to the analysis of maximum rainfall events in Uruguaiana, RS, Brazil. SN Applied Sciences. **2** (9), 1479 (2020).

[33] Majid M. H. A., Ibrahim K. Composite pareto distributions for modelling household income distribution in Malaysia. Sains Malaysiana. **50** (7), 2047–2058 (2021).

[34] Teodorescu S., Vernic R., et al. Some composite Exponential – Pareto models for actuarial prediction. Romanian Journal of Economic Forecasting. **12** (4), 82–100 (2009).

[35] Abu Bakar S. A., Nadarajah S., ABSL Kamarul Adzhar Z. A., Mohamed I. Gendist: An R Package for Generated Probability Distribution Models. PLOS One. **11** (6), e0156537 (2016).

[36] Ramachandran K. M., Tsokos C. P. Mathematical statistics with applications in R. Academic Press (2020).

[37] Hesterberg T. Bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics. **3** (6), 497–526 (2011).

Автоматизована процедура вибору порогового значення для узагальненого розподілу Парето із застосуванням до набору даних про опади

Аліф Ф. К., Алі Н., Сафарі М. А. М.

Кафедра математики та статистики, Факультет природничих наук,
Університет Путра Малайзія, 43400 UPM Серданг, Селангор, Малайзія

У гідрологічних наборах даних, зокрема про опади, вивчення екстремальних значень має вирішальне значення. Відповідний аналіз таких даних може надати життєво важливу інформацію про рівні повторюваності екстремальних опадів, що відіграє значну роль у запобіганні стихійним лихам. У багатьох випадках узагальнений розподіл Парето (GPD) є надійним методом для вивчення екстремальних даних. Проте, існують певні проблеми з вибором порогового значення для цього розподілу. Широко використовуваний метод побудови графіка середньої залишкової тривалості життя (MRL) для вибору порогового значення в GPD аналізі є суб'єктивним, вимагає значних попередніх знань і обмежує відтворюваність результатів. У цій роботі представлена проста, обчислювально недорога та автоматизована процедура вибору порогового значення. Використовуючи інтервальні порогові значення та критерії згоди (GOF), запропонований метод визначає оптимальне порогове значення, що максимізує p-значення, підвищуючи об'єктивність та точність. Було досліджено кілька комбінацій методів оцінки та критеріїв згоди, серед яких комбінація CVM-Lmoment виявилася найбільш стійкою. Завдяки розширеним симуляційним дослідженням запропонований підхід продемонстрував значні покращення у зменшенні зміщення та середньоквадратичної похибки (RMSE) порівняно з традиційними методами. Застосування запропонованої методології до набору даних про опади з Південно-Західної Англії підтвердило її надійність та практичну цінність, що робить її цінним інструментом для моделювання екстремальних значень та управління стихійними лихами.

Ключові слова: *узагальнений розподіл Парето; вибір порогу; якість наближення; L-моменти; екстремальні значення; рівень повернення; екстремальні опади.*