Malaysian Journal of Mathematical Sciences 19(3): 871–899 (2025)
https://doi.org/10.47836/mjms.19.3.06

Malaysian Journal of Mathematical Sciences

Journal homepage: https://mjms.upm.edu.my

An Assessment on Threshold Selection for Generalized Pareto Distribution using Goodness of Fit

Alif, F. K. [1], Ali, N. * [1], and Safari, M. A. M. [1]

[1]*Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

E-mail: norhaslinda@upm.edu.my
**Corresponding author*

Received: 19 December 2024
Accepted: 10 March 2025

Abstract

In real-world datasets, particularly those related to finance and rainfall, the study of extreme values is essential for understanding the return levels of extreme events and assessing financial risks. Accurate analysis of these extremes can play a crucial role in disaster prevention and risk management. While the generalized Pareto distribution remains a widely used tool for extreme value modeling, its threshold selection method poses challenges, notably the subjectivity of the mean residual life plot. This research presents an automated, step-by-step threshold selection procedure that is computationally efficient and objective. The method evaluates interval-based candidate thresholds and employs goodness-of-fit tests to identify the optimal threshold, maximizing the p-value. Of the various combinations of estimation methods and goodness of fit tests assessed in this study, the Anderson Darling-L-moments and Cramer-von Mises-Lmoments combinations demonstrated superior performance. Simulation studies indicated that our approach offers notable performance improvements compared to widely recognized non-automated method and several existing automated procedures. The proposed method was applied to real-life datasets from both the rainfall and financial domains, confirming its robustness. Additionally, a bootstrap approach was used to quantify the uncertainty of the selected threshold and its impact on return level estimates.

Keywords: extreme values; generalized Pareto distribution; automated; threshold selection; return level; goodness of fit.

1 Introduction

Extreme Value Analysis (EVA) plays a pivotal role in understanding the behavior of rare but consequential events across various fields, including finance, environmental science, and engineering. These extreme values, such as record-breaking rainfall, severe droughts, or significant fluctuations in financial markets, are crucial for informing risk assessments and decision-making processes. Among the statistical methods employed in EVA, the Generalized Pareto Distribution (GPD) is particularly notable for its efficacy in modeling the tail behavior of distributions, thus enhancing our ability to predict and manage extreme events [13, 20].

The urgency of accurately assessing extreme events has become even more pronounced in the face of climate change and economic volatility. The increasing frequency and intensity of extreme events pose significant challenges for risk management and policy development, highlighting the necessity for robust statistical methodologies [8]. The Peak Over Threshold (POT) method, a cornerstone of Extreme Value Theory (EVT), focuses on modeling observations that exceed a defined threshold. This technique effectively captures the essence of extreme behavior, enabling more precise risk estimates [16]. However, the selection of an appropriate threshold remains a significant challenge. This choice is critical, as it can significantly affect both the number of exceedances and the stability of the estimated statistical model [16, 14].

Selecting the right threshold is essential to obtain precise estimates of model parameters and return levels [7]. For the probability distribution model, a lower threshold is probably going to yield more samples. Nevertheless, the sample also contains smaller data, which results in an underestimation of return significant estimates and a decreased representativeness of extreme values. As a result, the distribution estimates are more biased but have less variance. On the other hand, a greater threshold can guarantee that extreme values are representative. The return significant estimates, however, are outside of the appropriate and stable range for return significant estimates as the number of remaining samples declines, increasing uncertainty in the estimates. Therefore, the distribution estimates are less biased but have higher variance [28].

One commonly employed graphical method for threshold selection is the Mean Residual Life (MRL) plot. While this technique provides some insights into threshold behavior, it is fraught with significant drawbacks. The MRL plot relies heavily on subjective interpretation, leading to inconsistencies and potential biases among analysts [28, 30]. To avoid bias, the threshold should be set high enough for the excesses to be adequately approximated by the GPD, but not so high that the estimator's variance is significantly increased as a result of a decrease in sample size (the number of exceedances). Furthermore, the choice of threshold can significantly impact GPD parameter estimates, often resulting in biases that compromise the reliability of risk assessments [7]. Moreover, reliance on graphical methods can obscure the quantification of uncertainty related to threshold selection, complicating the estimation of quantiles for extended return periods [44, 28].

While different approaches to automating threshold selection have been proposed (e.g., Weight based threshold selection of GPD by Dupuis [18], automatic threshold estimation using goodness of fit (GOF) p-value by Solari et al. [44], and threshold selection method based on the distribution of the difference of parameter estimates when the threshold is changed by Thompson et al. [49] among others), the most popular approach remains the use of graphical method. Coles et al. [13] provide a detailed discussion on graphical MRL plot technique, which is widely used despite its subjectivity. This widespread reliance on the MRL plot underscores the challenge posed by complex automated methods, as practitioners often prefer the subjective yet familiar graphical techniques despite their limitations. Recent advances, however, have further underscored the importance of integrating robust statistical procedures into threshold selection. For instance,

Gaigall and Gerstenberg [22] investigate the asymptotic behavior of Cramer-von Mises (CVM) type statistics for excesses over a confidence level and introduce bootstrap techniques to account for parameter uncertainty, thereby reinforcing the need for methods that address estimation variability.

Similarly, Murphy et al. [37] developed an automated threshold selection methodology that directly addresses the bias-variance trade-off, enhancing the reliability of extreme value models. Moreover, Mínguez [35] introduced an automatic threshold selection method based on weighted mean square errors and internally studentized residuals, demonstrating improved precision over conventional techniques in extreme rainfall modeling. Their study reinforces the need for reliable, objective, and statistically sound procedures for determining an appropriate threshold, particularly in hydrological applications. Alaswed [4] also explored graphical diagnostics for threshold selection and evaluated multiple threshold selection plots, showing that threshold choice plots can provide a reliable alternative for identifying stable threshold ranges while minimizing subjectivity.

Additionally, Curceac et al. [15] investigated the role of modified scale parameter estimation in threshold stability analysis, demonstrating how adjusting the scale parameter can enhance the robustness of automated threshold selection in extreme value modeling. These studies collectively highlight the ongoing advancements in threshold selection methodologies, balancing automation with statistical rigor. Additionally, Hambuckers et al. [25] proposed a simultaneous estimation method for tail and threshold parameters in extreme value regression models, offering a more efficient approach to threshold selection over the classical POT method. These recent contributions, together with the continued reliance on graphical methods, highlight that while automated techniques exist, their complexity often limits widespread adoption.

Building upon these advancements, our study introduces an alternative automated threshold selection procedure that enhances robustness while maintaining computational simplicity. Unlike more intricate methodologies, our approach leverages GOF tests to systematically evaluate candidate thresholds, reducing subjectivity while preserving interpretability. By systematically evaluating multiple estimation methods in combination with GOF criteria, our approach aims to provide a more objective and reliable foundation for modeling extreme values, improving accuracy in parameter estimates.

For the application of our approach in real-life scenarios, we will use two datasets. The first dataset consists of daily rainfall accumulations at a site in South West England from 1914 to 1962, containing 17,531 observations [14]. Additionally, we will analyse the daily closing prices of the Dow Jones Index from 1996 to 2000 with 1304 observations to demonstrate the efficacy of our methodology across diverse contexts.

Through this research, we aim to contribute to the growing body of knowledge in extreme value analysis by presenting an innovative, automated, and more reliable method for threshold selection. By enhancing the objectivity of this critical step, we hope to improve both the theoretical understanding and practical management of extreme events, ultimately contributing to more effective risk management strategies in fields affected by rare and extreme occurrences.

2 Methodology

2.1 Theoretical background

The GPD is a key tool for modeling exceedances above a defined threshold u in extreme value analysis [13]. A threshold in extreme value theory refers to a predefined value above which only the data points are considered extreme, and these exceedances are modeled to capture the tail behavior of a distribution. The choice of an appropriate threshold is crucial for ensuring a balance between bias and variance in model estimation [16]. Introduced by Pickands [39], the GPD is parameterized by the scale, σ and shape, ξ parameters. The Cumulative Distribution Function (CDF) representing the relationship among these parameters and the threshold u can be defined as,

$$G(x;\sigma,\xi) = \begin{cases} 1 - \left[1 + \xi\left(\dfrac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\dfrac{x-u}{\sigma}\right), & \text{if } \xi = 0, \end{cases} \tag{1}$$

where, x is independent and identically distributed (iid) random variable, and for $\xi \geq 0 : x > u$ and for $\xi < 0 : u \leq x < u - \dfrac{\sigma}{\xi}$. The parameters $\sigma > 0$, $-\infty < \xi < \infty$ and $-\infty < u < \infty$. The choice of ξ governs the tail behavior: $\xi > 0$ corresponds to heavy tails, $\xi = 0$ to the exponential distribution, and $\xi < 0$ implies an upper bound on the data [20, 16]. To estimate the parameters σ and ξ, our study will utilize three well-established estimation techniques: Maximum Likelihood Estimation (MLE), L-moments, and Maximum Product Spacing (MPS) [16, 10].

Lets consider the probability density function (PDF) of GPD when $\xi \neq 0$, which is,

$$g(x;\sigma,\xi) = \frac{1}{\sigma}\left(1 + \xi\frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}-1}. \tag{2}$$

Taking the logarithm to obtain the log-likelihood function, which can be written as,

$$\log L(\sigma,\xi) = -n\log\sigma - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left(1 + \xi\frac{x_i - u}{\sigma}\right). \tag{3}$$

Next, we examine the scenario where $\xi = 0$. In this case, the PDF of the GPD can be expressed as,

$$g(x;\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x-u}{\sigma}\right), \tag{4}$$

and the log-likelihood function can be written as,

$$\log L(\sigma) = -n\log\sigma - \frac{1}{\sigma}\sum_{i=1}^{n}(x_i - u). \tag{5}$$

The likelihood function of the GPD serves to estimate the parameters ξ and σ through the maximization of the likelihood function, or, alternatively, the log-likelihood function, based on the observed data. To estimate the parameters using MLE in this study we will be utilizing the R packages $extRemes$ [23] and $ismev$ [47].

L-moments provide a robust framework for describing the shape of a probability distribution, functioning as linear combinations of order statistics [5, 27]. L-moments offer greater resistance to outliers, exhibit lower sampling variability, and remain nearly unbiased even with small sample sizes [9]. For a continuous random variable Y, the Quantile Function (QF) that defines its distribution is expressed as $Q_y(p) =$ The value of y such that $F_y(Y) = p$, where $0 \leq p \leq 1$ and $F_y(Y) = p$ is the CDF of Y [50]. Hosking's [26] comprehensive work on L-moments theory provides a foundation for its application. The r^{th} L-moments, expressed in terms of the QF, is defined as,

$$L_r = \int_0^1 Q_y(p) P^*_{r-1}(p) dp, \tag{6}$$

where

$$P^*_{r-1}(p) = \sum_{k=0}^{r} \left[-1^{r-k} \binom{r}{k} \binom{r+k}{k} p^k \right], \tag{7}$$

represents the Legendre polynomial with the r^{th} shift, and L_r denotes the r^{th} L-moment [50]. To estimate the parameters using L-moments in this study we will be utilizing the R package $extRemes$ [23].

MPS is a technique used to estimate parameters in univariate statistical models [10]. This approach involves maximizing the geometric mean of the spacings within the data, defined as the differences in the CDF values at adjacent data points [10]. Given an identically distributed random sample $x_1, \ldots, x_n$ of size n from a univariate distribution with a continuous CDF of $F(x; \theta_0)$, where $\theta_0 \in \Theta$ is the unknown parameter to be estimated. let $x_{(1)}, \ldots, x_{(n)}$ represent the corresponding ordered sample. For convenience, we also define $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$. The spacings are defined as the gaps between the values of the distribution function at adjacent ordered points [40],

$$D_i(\theta) = F(x_{(i)}; \theta) - F(x_{(i-1)}; \theta), \quad i = 1, \ldots, n+1. \tag{8}$$

The maximum spacing estimator $\hat{\theta}$ is obtained by maximizing the logarithm of the geometric mean of the sample spacings, expressed as $\hat{\theta} = \arg\max_{\theta \in \Theta} S_n(\theta)$, where $S_n(\theta) = \dfrac{1}{n+1} \sum_{i=1}^{n+1} \log D_i(\theta)$. By the inequality of arithmetic and geometric means, the function $S_n(\theta)$ is bounded from above by $-\log(n+1)$, ensuring that a maximum exists at least in the supremum sense [10]. To estimate the parameters using MPS in this study, we will be utilizing the R package eva [6].

Our approach utilizes the p-value from GOF tests in conjunction with the estimation technique to identify the optimal threshold in the dataset, which defines the extreme values. Smaller p-values indicate a stronger statistical incompatibility of the data with the null hypothesis, assuming the assumptions for p-value calculations are correct, as noted by Wasserstein and Lazar [51]. This metric, ranging from 0 (indicating total incompatibility) to 1 (indicating perfect compatibility), evaluates the degree to which the model fits the data [24]. To conduct GOF testing, the GPD parameters must first be estimated. In this study, we consider the Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Cramer-von Mises (CVM) tests to determine the most appropriate threshold point.

The KS test measures the largest absolute deviation between the Empirical Distribution Function (EDF) and the theoretical CDF, providing insight into how well the model represents the observed data. It is defined as,

$$KS = \max_{1 \leq i \leq n} \left(\left| G(x_{(i)}; u, \hat{\sigma}, \hat{\xi}) - \frac{i-1}{n} \right|, \left| \frac{i}{n} - G(x_{(i)}; u, \hat{\sigma}, \hat{\xi}) \right| \right), \tag{9}$$

where n is the sample size, $x_{(i)}$ denotes the i^{th} order statistic, and $G(x_{(i)}; u, \hat{\sigma}, \hat{\xi})$ represents the CDF of the GPD evaluated at $x_{(i)}$. Simulation studies by [12] have demonstrated that the KS test performs well for GPD modeling, particularly when identifying discrepancies across the entire distribution. The AD test, unlike KS, places greater emphasis on deviations in the tails of the distribution, making it particularly useful for extreme value analysis. The AD test serves as a statistical procedure designed to assess whether a specific sample of data originates from a specified probability distribution. In its fundamental application, the test operates under the assumption that no parameters are to be estimated within the distribution being evaluated, allowing for a distribution-free framework with corresponding critical values. However, the test is predominantly utilized in scenarios involving a family of distributions, necessitating parameter estimation. In such cases, adjustments must be made to either the test statistic or its critical values to account for this estimation process [46]. The AD test statistic is defined as,

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\log G(x_{(i)}; u, \hat{\sigma}, \hat{\xi}) + \log\left(1 - G(x_{(n+1-i)}; u, \hat{\sigma}, \hat{\xi})\right)\right]. \tag{10}$$

Here, $G(x_{(i)}; u, \hat{\sigma}, \hat{\xi})$ denotes the estimated CDF of the GPD at the i^{th} order statistic. The weighting mechanism of the test enhances sensitivity to extreme deviations, making it one of the most powerful GOF tests for identifying tail discrepancies [31, 45]. Another well-recognized GOF test for continuous distributions is the CVM test, widely employed for assessing the conformity of sample data to a model [11]. The CVM test statistic is given by,

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[G(x_{(i)}; u, \hat{\sigma}, \hat{\xi}) - \frac{2i-1}{2n}\right]^2. \tag{11}$$

This test is particularly useful for evaluating how well an empirical distribution function fits a theoretical CDF [3]. As a statistical measure, it is effective in quantifying discrepancies between observed sample data and the theoretical distribution. For GPD modeling, CVM provides robust performance, comparable to the KS test, as reviewed by [12].

In practice, however, the theoretical null distributions of these GOF test statistics assume that the u and the parameters $\hat{\sigma}$, and $\hat{\xi}$ are known a priori. In our application, these parameters are estimated from the data, and this estimation introduces additional variability that shifts the null distribution of the test statistic away from its standard form [45, 29]. Such shifts can affect the accuracy of the p-values, as the test statistic becomes dependent on the estimation process, resulting in modified critical values compared to the classical theory. This dependence is particularly critical in extreme value analysis, where the tail of the distribution is inherently sparse and sensitive to parameter uncertainty. Although various approaches, such as bootstrap resampling, have been proposed to adjust for this effect, our study focuses on employing the p-values from the GOF tests directly while acknowledging that these p-values are influenced by parameter estimation [17]. By carefully considering these factors, our automated threshold selection procedure is designed to maintain its robustness across datasets of varying sizes and characteristics, thereby ensuring more reliable inference in the modeling of extreme events.

The p-value associated with each GOF test quantifies the probability of observing a test statistic as extreme as, or more extreme than, the one computed from the sample under the null hypothesis [44]. Given that the null distributions of these test statistics are affected by parameter estimation, the calculation of p-values often requires numerical approximations or resampling methods to obtain reliable results [45]. For the KS test, the p-value is computed by comparing the observed KS statistic to the theoretical distribution under the null hypothesis, adjusted for parameter estimation effects [29]. The standard KS test assumes that the parameters are known, but when they are estimated, the null distribution changes, requiring adjusted critical values [34]. In this study, we

utilize the $stats$ package in R, which implements methods to derive p-values from the empirical distribution function [41].

For the AD test, the theoretical null distribution does not have a closed-form solution when parameters are estimated [45]. Instead, the critical values are approximated through numerical integration or simulation-based methods [46]. The dependence of the AD test on estimated parameters has been extensively studied, showing that it requires modifications in its test statistic or critical values for accurate inference [31]. We employ the $goftest$ package in R to compute the p-values for the AD test [21]. Similarly, for the CVM test, the null distribution is obtained from asymptotic approximations, which are modified when parameter estimation is involved [11]. The CVM test is particularly sensitive to deviations across the entire distribution, making it a robust alternative to KS and AD in extreme value settings [12]. The $goftest$ package in R is used to compute p-values for the CVM test, leveraging tabulated critical values for common significance levels [21]. Since these GOF tests are employed for threshold selection in our study, the p-values serve as key indicators of model fit, allowing us to systematically determine the most appropriate threshold for GPD modeling. Given the dependency of p-values on estimated parameters, our approach acknowledges this limitation while utilizing these values as a relative measure for threshold selection across different candidate thresholds [44].

2.2 Threshold selection using *p*-value of the GOF test

This section presents a refined approach for selecting the optimal threshold in GPD modeling that is designed to be robust across datasets of various sizes and characteristics. The methodology systematically partitions the dataset into a fixed number of equal intervals, thereby generating a comprehensive set of candidate thresholds that can be evaluated using GOF tests. For datasets containing a significant number of zero values, the procedure is further adapted by computing the means of each interval and selecting the candidate threshold based on those means exceeding the first quantile of the dataset.

The robustness of this approach is evidenced by its consistent performance across varying sample sizes and different parameter configurations in extensive simulation studies (Section 3.1). Despite the inherent variability in distributional forms and parameter sets, the method consistently yielded accurate and reliable threshold estimates, minimizing bias, Standard Error (SE), and Root Mean Squared Error (RMSE). Its adaptability was further confirmed through the application to real-world datasets with distinct characteristics, including dataset with significant number of zeros, demonstrating its practical applicability and effectiveness across a broad spectrum of scenarios. These findings underscore the method's resilience and suitability for diverse modeling requirements in extreme value analysis.

Following are the steps of our threshold selection approach using p-value of the GOF test:

1. **Initial Data Partitioning:**

   The dataset is first sorted in ascending order and divided into 200 equal intervals. The decision to divide the dataset into $N = 200$ intervals is based on empirical observations, as it provides a suitable level of granularity without excessive fragmentation. While this specific number has not been derived from a formal simulation study, it has consistently demonstrated stable performance in determining the first candidate threshold u_1 across different datasets in our analysis (for simulation study in Tables 1 and 2 as well as real-life datasets in Sectionn 3.4). If the dataset consists of a significant number of zeros (such as in rainfall datasets), the next step is to follow the outlined method for handling such data. However, in

datasets where zeros are not prevalent, the median of the dataset can be selected directly as the candidate threshold u_1, bypassing the more complex step that follows. This adjustment simplifies the process when zeros are not an issue.

2. **Handling Zeros in Data:**

   In cases where the dataset contains significant zero values, the next step involves computing the means $M_1, M_2, \ldots, M_N$ for each of the 200 intervals. From these means $M_1, M_2, \ldots, M_N$, those exceeding the first quantile (Q_1) are selected. The first candidate threshold u_1 is the average of the means which are greater than the Q_1,

$$u_1 = \frac{M_i + M_{i+1} + \ldots + M_N}{C}, \quad \text{where } M_i > Q_1.$$

   Here, the total number of computed means corresponds to the number of intervals into which the dataset was divided. However, only the subset of means that exceed Q_1 contribute to the calculation of u_1, and the count of these selected means is denoted as C. This ensures that the threshold selection process is not biased toward smaller values.

3. **Determining the End Point:**

   The endpoint of the candidate thresholds, denoted as u_n, is usually around the 5th largest observation in the dataset. However, the number of exceedances above the threshold needs to be considered, especially in large datasets. If the dataset has at least 5000 observations, the number of exceedances should not be less than 80. For datasets with 1000 to 5000 observations, a minimum of 20 exceedances is recommended. If the dataset contains fewer than 1000 observations, no specific minimum for exceedances is imposed, allowing for more flexibility in small sample sizes. These criteria ensure that the selected threshold maintains a balance between sample size and statistical reliability which is reflected in comparatively lower uncertainties as presented in Section 3.3 and 3.4, preventing excessive bias from too few exceedances and consistent results which can be observed in Tables 1 and 2.

4. **Selection of Candidate Thresholds:**

   The candidate thresholds $u_1, u_2, \ldots, u_n$ are spaced with equal intervals between them. The total number of candidate thresholds generated in this manner is denoted as n. In other words, n represents the count of thresholds considered for evaluation, spanning from the initial candidate u_1 to the final candidate u_n. Through empirical testing, we have observed that selecting around 300 candidate thresholds provides a robust and reliable range for threshold selection. Although no formal simulation study was conducted to determine this value, it has proven to be a reasonable choice across multiple datasets which can be observed across Tables 1 and 2 as well as in Section 3.4.

5. **Parameter Estimation and p-value Optimization:**

   For each candidate threshold, the GPD parameters are estimated using the data where $x_i > u_0$. After estimating the parameters for each threshold, a GOF test is applied to compute the corresponding p-values. In Section 2.1, we discussed the computation of p-values from various GOF tests. The optimal threshold u_0 is selected as the one that maximizes the p-value of the GOF test, ensuring that the selected threshold provides the best fit for the tail of the distribution.

2.3 Simulation study

The goal of this simulation study is to identify the optimal combination of GOF tests and estimation techniques that enhance the effectiveness and robustness of the proposed threshold selec-

tion methodology. This study also aims to evaluate the accuracy of the method across a range of simulated datasets. For the simulation framework, we have employed an exponential GPD composite model to generate the datasets. This property of having separate models in composite GPD models for observations below and above the threshold value is desirable when the upper tail observations are heavily skewed or believed to be distributed differently compared to the rest of the data [1]. By utilizing this composite model, we can rigorously test the performance of the proposed threshold selection method under controlled conditions, ensuring its applicability across diverse dataset structures. The composite models can generally be defined as,

$$f(x|\theta) = \begin{cases} \rho f_1^*(x|\theta), & \text{if } x \leq u_0, \\ (1-\rho) f_2^*(x|\theta), & \text{if } x > u_0, \end{cases} \tag{12}$$

where

$$\rho = \frac{\alpha(1-e^{-\lambda u_0})}{\alpha + e^{-\lambda u_0}}. \tag{13}$$

The α represent the tail index and λ denotes the rate parameter. The model divides the distribution at the threshold u_0, distinguishing the tail behavior from the bulk of the data. The PDF of the underlying distribution is given by,

$$f_1^*(x|\theta) = \frac{f_1(x|\theta)}{F_1(x|\theta)}, \tag{14}$$

$$f_2^*(x|\theta) = \frac{f_2(x|\theta)}{1 - F_2(x|\theta)}. \tag{15}$$

The data in the lower tail are modeled by the probability density function $f_1(x|\theta)$, assumed to follow an exponential distribution with λ as the rate. Observations in the upper tail are modeled using the probability density function $f_2(x|\theta)$, which corresponds to the GPD [48, 2]. The parameter ρ represents the mixing weight, as discussed by Abu Bakar [2]. Further insights into composite models, especially the exponential-Pareto composite model, are elaborated by Majid and Ibrahim [32] and Teodorescu et al. [48].

The simulation study was conducted within the RStudio environment, leveraging its comprehensive toolset for statistical computing and data analysis [42]. The steps of simulation are as follows:

Step 1: **Dataset generation:** Two sets of parameters will be used to generate datasets from the exponential-GPD composite model:

- **Set 1:** rate, $\lambda = 0.303$ (exponential), threshold, $u = 35$, scale, $\sigma = 12.5$, and shape, $\xi = 0.010$ for the GPD component.
- **Set 2:** rate, $\lambda = 0.250$ (exponential), threshold, $u = 11.4$, scale, $\sigma = 5.0$, and shape, $\xi = -0.050$ for the GPD component.

The parameter sets were considered based on real life datasets. The sample sizes considered will be 500, 1000, and 5000, representing small to large datasets. Random numbers are generated using the *gendist* package, as discussed by Abu Bakar [2], which simulates from a variety of composite models, including the exponential-GPD.

Step 2: **Automated threshold selection:** The automated threshold selection methodology (detailed in Section 2.2) is applied to each generated dataset to determine the optimal threshold for the GPD component.

Step 3: **Repetition for statistical analysis:** Steps 1 and 2 are repeated 1000 times for each combination of sample size and parameter set. The thresholds are stored to compute bias, RMSE and SE, enabling robust performance evaluation.

Step 4: Additional comparative methods: Thompson's [49] and Solari's [44] methods are applied to estimate thresholds. These methods allow quantification of threshold uncertainty along with our automated threshold selection method, unlike the MRL plot which lacks objectivity [44].

Step 5: Threshold uncertainty estimation: To assess the uncertainty in threshold selection, the bootstrap percentile method is employed. This involves generating 5000 bootstrap samples for each dataset using both parameter sets. Thresholds are recalculated for each sample, and confidence intervals are constructed based on the distribution of these thresholds. Thompson's and Solari's methods are used for comparison.

Step 6: Return level uncertainty analysis: A second simulation examines the uncertainty in estimating 50, 100, and 1000-year return levels. For each bootstrap sample, the threshold is recalculated using the core methodology, and return levels are estimated for each return period. Confidence intervals are calculated using the bootstrap percentile method. Both Thompson's and Solari's methods are compared, and return levels are evaluated for both parameter sets using a sample size of 5000 (the exact same datasets as in Step 5).

2.4 Bootstrap percentile method

The bootstrap percentile method is a resampling technique commonly used to estimate confidence intervals [49]. It involves generating multiple bootstrap samples from the original dataset and recalculating the statistic of interest (in this case, thresholds or return levels). Confidence intervals are then constructed from the distribution of these recalculated statistics. As was discussed by Mooney et al. [36] along with Efron and Tibshirani [19], the steps of the bootstrap percentile method are as follows:

Step 1: Generate B bootstrap samples from the original dataset. For our study, we considered $B = 1000$.

Step 2: For each bootstrap sample, calculate the desired statistics which are threshold and return level.

Step 3: Sort the bootstrap estimates.

Step 4: The lower and upper bounds of the confidence interval are taken from the percentiles corresponding to the desired confidence level (e.g., 95% confidence interval).

3 Results and Discussion

3.1 Selection of optimal pair of method of estimation and GOF

Determining the most suitable GOF test alongside the optimal estimation method is paramount to accurately identifying thresholds within the GPD. As outlined in Section 2, this study considers three estimation methods alongside three GOF tests, resulting in nine possible combinations. For each combination, the bias, SE, and RMSE of the estimated parameters are assessed, providing insight into the accuracy and consistency across different sample sizes. The goal is to identify combinations that yield results with both minimal bias and high reliability, aligning with the study's focus on robust and reliable outcomes. As discussed in Section 2.3, there are 2 sets of parameters to be considered for the simulation study on the selection of the optimal pair of the method of estimation and GOF alongside a comparison with the existing method. For clarity, the metrics

that are used to evaluate the performance of the threshold selection procedure in Tables 1 and 2 are discussed below:

- **E[u], E[$\hat{\sigma}$] and E[$\hat{\xi}$]:** These represent the expected (mean) values of the determined threshold u, $\hat{\sigma}$, and $\hat{\xi}$ across all simulation runs.
- **p-value:** The p-value reported in the Tables 1 and 2 is the average p-value obtained from the GOF tests over the simulation runs.
- **Bias:** Bias is defined as the difference between the expected estimate and the true parameter value,

$$\text{Bias} = E[\hat{\theta}] - \theta,$$

  where $\hat{\theta}$ represents the estimated parameter and θ is the corresponding true value.
- **SE (Standard Error):** The standard error quantifies the variability of the estimator and is calculated as,

$$\text{SE} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i - \bar{\theta}\right)^2},$$

  where $\bar{\theta} = \dfrac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$ is the mean estimate across N simulation runs.
- **RMSE (Root Mean Squared Error):** RMSE is defined as,

$$\text{RMSE} = \sqrt{\text{Bias}^2 + \text{SE}^2},$$

  providing an overall measure of estimation accuracy.

Table 1 represents results obtained from the first set of parameters (λ=0.303, u=35.000, σ=12.500 and ξ=0.010) for our automated threshold selection on every possible combination method of estimation and GOF test over different sample sizes. Analysis of the results presented in Table 1 reveals that the most reliable outcomes are observed when considering the combined performance of the CVM with L-moments and the AD with L-moments. Although these combinations do not consistently yield the most precise estimates, a detailed examination of the estimated parameter values, bias, SE, and RMSE indicates that they approximate the true parameters closely.

Specifically, while the CVM-L-moments pairing may not always provide optimal results for the u and σ, it excels in estimating the ξ. Conversely, the AD-L-moments combination demonstrates strong performance across all parameters and sample sizes. Notably, for a sample size of 500, the AD-L-moments combination exhibited a negative bias for the shape parameter; nevertheless, it achieved among the lowest bias (-0.021), RMSE (0.023), and SE (0.010) for this parameter. According to Ramachandran and Tsokos [43], it is a well-known phenomenon that the accuracy level rises as the sample size increases. So as anticipated, the accuracy of all combinations tends to improve with larger sample sizes.

Table 1: The table reports the results obtained from the simulation study for the first set of parameters including the RMSE, bias, SE and average p-value at different sample sizes for KS, AD and CVM for each estimated parameter. In this table, u is the determined threshold.

Sample Size	GOF	Method of estimation	E[u]	E[$\hat{\sigma}$]	E[$\hat{\xi}$]	p-value	u			$\hat{\sigma}$			$\hat{\xi}$		
							RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
500	KS	MLE	36.715	13.281	-0.048	0.989	1.777	1.715	0.465	0.817	0.781	0.240	0.060	-0.058	0.015
		L-moments	33.093	12.727	-0.013	0.991	1.950	-1.906	0.413	0.285	0.228	0.172	0.025	-0.023	0.009
		MPS	39.592	11.172	0.294	0.993	4.614	4.592	0.450	1.355	-1.327	0.274	0.285	0.284	0.025
	AD	MLE	33.883	11.986	0.046	0.987	1.200	-1.116	0.440	0.537	-0.513	0.161	0.038	0.037	0.011
		L-moments	**32.918**	**12.685**	**-0.011**	**0.993**	**2.118**	**-2.081**	**0.393**	**0.257**	**0.185**	**0.178**	**0.023**	**-0.021**	**0.010**
		MPS	36.918	11.479	0.237	0.994	1.968	1.918	0.440	1.056	-1.021	0.273	0.228	0.227	0.024
	CVM	MLE	37.224	12.872	-0.025	0.987	2.270	2.224	0.458	0.434	0.373	0.223	0.039	-0.036	0.016
		L-moments	**32.808**	**12.677**	**0.000**	**0.989**	**2.227**	**-2.191**	**0.399**	**0.244**	**0.177**	**0.168**	**0.013**	**-0.010**	**0.009**
		MPS	38.723	11.106	0.273	0.992	3.751	3.723	0.453	1.412	-1.393	0.228	0.265	0.264	0.024
1000	KS	MLE	31.902	13.086	-0.019	0.984	3.117	-3.097	0.349	0.597	0.586	0.114	0.029	-0.029	0.006
		L-moments	31.361	12.513	0.012	0.990	3.655	-3.638	0.355	0.100	0.013	0.099	0.005	0.003	0.005
		MPS	33.095	11.700	0.117	0.986	1.937	-1.904	0.351	0.806	-0.799	0.110	0.108	0.107	0.007
	AD	MLE	32.453	13.000	-0.016	0.986	2.570	-2.546	0.347	0.516	0.500	0.127	0.026	-0.026	0.006
		L-moments	**31.970**	**12.456**	**0.018**	**0.991**	**3.050**	**-3.029**	**0.350**	**0.115**	**-0.043**	**0.106**	**0.009**	**0.008**	**0.005**
		MPS	33.793	11.601	0.127	0.989	1.258	-1.206	0.356	0.906	-0.898	0.117	0.117	0.117	0.008
	CVM	MLE	33.036	13.110	-0.021	0.980	1.994	-1.963	0.349	0.623	0.610	0.126	0.031	-0.031	0.006
		L-moments	**31.892**	**12.453**	**0.019**	**0.987**	**3.127**	**-3.107**	**0.349**	**0.113**	**-0.046**	**0.103**	**0.010**	**0.009**	**0.005**
		MPS	34.152	11.596	0.131	0.984	0.919	-0.847	0.356	0.910	-0.903	0.114	0.121	0.121	0.008
5000	KS	MLE	33.580	12.895	0.005	0.984	1.466	-1.420	0.366	0.400	0.400	0.050	0.007	-0.006	0.002
		L-moments	33.836	12.741	0.014	0.990	1.219	-1.164	0.363	0.246	0.241	0.050	0.004	0.004	0.003
		MPS	34.550	12.400	0.050	0.984	0.577	-0.449	0.362	0.122	-0.100	0.051	0.040	0.040	0.003
	AD	MLE	33.574	12.883	0.005	0.987	1.473	-1.426	0.369	0.387	0.383	0.055	0.006	-0.005	0.003
		L-moments	**33.747**	**12.680**	**0.018**	**0.991**	**1.306**	**-1.253**	**0.369**	**0.187**	**0.180**	**0.051**	**0.008**	**0.008**	**0.003**
		MPS	34.045	12.331	0.052	0.987	1.025	-0.955	0.371	0.177	-0.170	0.053	0.042	0.042	0.003
	CVM	MLE	33.196	12.843	0.006	0.981	1.841	-1.804	0.371	0.346	0.343	0.051	0.005	-0.004	0.003
		L-moments	**33.342**	**12.768**	**0.014**	**0.988**	**1.700**	**-1.659**	**0.371**	**0.272**	**0.268**	**0.050**	**0.005**	**0.004**	**0.003**
		MPS	34.063	12.400	0.050	0.981	1.008	-0.937	0.371	0.116	0.100	0.051	0.040	0.040	0.003

Further examination of Table 1 illustrates that combinations involving MPS yield highly accurate estimates for u, although they tend to diverge from optimal accuracy for other parameters, particularly ξ. In summary, the bias, SE, and RMSE values for u, σ, and ξ suggest that the AD-L-moments and CVM-L-moments combinations are commendable choices, positioned at the upper echelon of accuracy. The overall findings from this simulation study underscore that either the AD-L-moments or CVM-L-moments combination is a sound selection across varying sample sizes.

The findings derived from the second parameter set ($\lambda = 0.250$, $u = 11.400$, $\sigma = 5.000$, and $\xi = -0.050$), displayed in Table 2, illustrate notable consistency and precision in the AD-L-moments combination, particularly for estimating u and other parameters across different sample sizes. For the smallest sample size of 500, the AD-L-moments combination yielded a bias as low as 0.001 for ξ, along with highly accurate estimates for u and σ, with biases of -0.099 and -0.686, respectively. The parameter estimates across other sample sizes also display strong alignment with the true values, with close examination of biases, SE, and RMSE revealing a minimal deviation. Although the CVM-L-moments combination may not perform optimally for u and σ, it consistently provides accurate estimates for σ and ξ. As expected, all combinations show an improvement in accuracy as the sample size increases. A deeper analysis of Table 2 once again further reveals that combinations involving MPS offer high accuracy for u but consistently lack precision in estimating ξ. Therefore, with respect to bias, SE, and RMSE for u, σ, and ξ, the AD-L-moments and CVM-L-moments combinations emerge as the most favorable choices, occupying the upper range of accuracy. Overall, the simulation study suggests that the AD-L-moments and CVM-L-moments combinations are robust choices across varying sample sizes, underscoring their suitability for precise parameter estimation.

This simulation study systematically evaluates nine combinations of estimation methods and GOF tests to identify the optimal pairing for threshold selection within the GPD. The analysis, conducted across two parameter sets and multiple sample sizes, emphasizes bias, SE, and RMSE as performance indicators. Results underscore the reliability of the AD-L-moments and CVM-L-moments combinations, both of which show high accuracy and consistency, especially for the critical shape parameter, ξ. Despite minor variability, AD-L-moments consistently performs well across all parameters and sample sizes, maintaining minimal bias even at smaller samples, while CVM-L-moments proves particularly robust in larger sample contexts. Conversely, combinations with MPS, though effective for the threshold parameter u, demonstrate lower precision for ξ.

Table 2: The table reports the results obtained from the simulation study for the second set of parameters including the RMSE, bias, SE and average p-value at different sample sizes for KS, AD and CVM for each estimated parameter. In this table, u is the determined threshold.

Sample Size	GOF	Method of estimation	E[u]	E[$\hat{\sigma}$]	E[$\hat{\xi}$]	p-value	u			$\hat{\sigma}$			$\hat{\xi}$		
							RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE
500	KS	MLE	12.547	4.463	-0.106	0.989	1.157	1.147	0.157	0.540	-0.536	0.071	0.058	-0.057	0.015
		L-moments	11.426	4.350	-0.062	0.991	0.142	0.026	0.141	0.652	-0.650	0.056	0.015	-0.012	0.009
		MPS	13.637	3.707	0.238	0.994	2.242	2.237	0.151	1.294	-1.292	0.072	0.289	0.288	0.025
	AD	MLE	11.476	4.193	-0.013	0.986	0.167	0.076	0.148	0.808	-0.806	0.057	0.039	0.037	0.011
		L-moments	**11.301**	**4.314**	**-0.049**	**0.992**	**0.164**	**-0.099**	**0.132**	**0.688**	**-0.686**	**0.056**	**0.009**	**0.001**	**0.009**
		MPS	12.604	3.898	0.177	0.994	1.213	1.204	0.148	1.104	-1.102	0.081	0.228	0.227	0.024
	CVM	MLE	12.638	4.434	-0.096	0.985	1.247	1.238	0.154	0.570	-0.566	0.073	0.048	-0.046	0.016
		L-moments	**11.247**	**4.356**	**-0.049**	**0.989**	**0.203**	**-0.152**	**0.135**	**0.645**	**-0.643**	**0.055**	**0.009**	**0.001**	**0.009**
		MPS	13.304	3.814	0.204	0.991	1.910	1.904	0.150	1.187	-1.185	0.083	0.255	0.254	0.025
1000	KS	MLE	10.990	4.539	-0.082	0.983	0.427	-0.409	0.120	0.462	-0.460	0.039	0.032	-0.032	0.006
		L-moments	10.714	4.347	-0.041	0.990	0.695	-0.685	0.121	0.653	-0.652	0.033	0.009	-0.008	0.005
		MPS	11.423	4.074	0.052	0.986	0.123	0.023	0.121	0.926	-0.925	0.038	0.103	0.102	0.008
	AD	MLE	11.074	4.506	-0.076	0.985	0.346	-0.325	0.118	0.496	-0.494	0.041	0.027	-0.027	0.007
		L-moments	**11.021**	**4.289**	**-0.032**	**0.992**	**0.396**	**-0.378**	**0.119**	**0.711**	**-0.711**	**0.036**	**0.017**	**0.017**	**0.006**
		MPS	11.469	4.044	0.056	0.990	0.140	0.069	0.121	0.956	-0.956	0.040	0.106	0.106	0.008
	CVM	MLE	11.274	4.484	-0.075	0.979	0.173	-0.126	0.119	0.517	-0.516	0.040	0.025	-0.025	0.007
		L-moments	**10.901**	**4.313**	**-0.035**	**0.988**	**0.512**	**-0.498**	**0.118**	**0.687**	**-0.687**	**0.035**	**0.016**	**0.016**	**0.005**
		MPS	11.750	4.013	0.065	0.984	0.370	0.350	0.121	0.987	-0.987	0.040	0.116	0.116	0.008
5000	KS	MLE	11.704	4.422	-0.055	0.981	0.328	0.305	0.123	0.577	-0.577	0.017	0.006	-0.005	0.003
		L-moments	11.799	4.372	-0.044	0.989	0.418	0.399	0.122	0.627	-0.627	0.017	0.006	0.006	0.003
		MPS	11.874	4.255	-0.012	0.983	0.489	0.474	0.123	0.745	-0.745	0.018	0.038	0.038	0.003
	AD	MLE	11.523	4.434	-0.055	0.985	0.174	0.123	0.123	0.565	-0.565	0.018	0.006	-0.005	0.003
		L-moments	**11.526**	**4.361**	**-0.04**	**0.991**	**0.177**	**0.126**	**0.124**	**0.639**	**-0.638**	**0.018**	**0.010**	**0.009**	**0.003**
		MPS	11.660	4.254	-0.01	0.986	0.288	0.260	0.126	0.746	-0.746	0.020	0.040	0.040	0.003
	CVM	MLE	11.531	4.443	-0.057	0.978	0.182	0.131	0.126	0.557	-0.556	0.017	0.007	-0.007	0.003
		L-moments	**11.355**	**4.403**	**-0.044**	**0.988**	**0.134**	**-0.044**	**0.126**	**0.596**	**-0.596**	**0.017**	**0.005**	**0.005**	**0.003**
		MPS	11.721	4.265	-0.013	0.980	0.344	0.321	0.124	0.734	-0.734	0.018	0.037	0.037	0.003

So, one might think about using MPS (particularly with AD) to estimate the threshold value first and then, apply other methods of estimation to estimate σ and ξ, but this will make the core methodology more complex. On the other hand, as an estimation method, MLE also performs well when the sample size is larger. However, as noted by [33], outliers in the data can lead to unreliable sample estimates and MLE results, as seen in the smaller sample sizes in Tables 1 and 2.

So, these findings validate AD-L-moments and CVM-L-moments as robust choices, especially in applications requiring precise modeling of extreme values, and highlight their suitability for diverse sample sizes in GPD threshold selection.

3.2 Comparison with the existing automated threshold selection methods

Over the years, numerous approaches have been developed for automated threshold selection within the GPD. For this study, we have chosen two computationally efficient and effective methods: those proposed by Solari et al. [44] and Thompson et al. [49]. Solari et al. [44] propose an automated threshold selection procedure that relies on the modified version of the Anderson-Darling EDF statistic to identify an optimal threshold. Their method involves selecting candidate thresholds by incrementally examining peaks in the data through a moving window (without stating any exact number of candidate thresholds) and determining the threshold at which the AD statistic meets a predetermined criterion. In their approach, bootstrapping is employed to quantify threshold uncertainty and assess its impact on high return period quantiles. Extensive testing on simulated data and four precipitation and river flow datasets further substantiated the method's robustness. Comprehensive insights into their approach can be found in [44].

In contrast, our method adopts a comprehensive yet operationally simple strategy. We generate a fixed number of candidate thresholds by partitioning the dataset into equal intervals; this extensive candidate threshold set increases the likelihood that the optimal threshold is among those considered, thereby enhancing the robustness of the threshold selection process. For datasets with significant zeros, we further refine the candidate set by computing the interval means and selecting those above the Q_1 of the dataset. Importantly, our method relies on widely used GOF tests (KS, AD, and CVM) and standard parameter estimation techniques (MLE, L-moments, and MPS) rather than complex or modified versions, making it both operationally inexpensive and broadly accessible. In our simulation studies (Tables 1 and 2), we demonstrated that both the CVM-L-moments and AD-L-moments combinations yield consistently robust results across various sample sizes and parameter sets. This multi-criteria, data-driven approach contrasts with Solari et al.'s method, which focuses on a single GOF test and estimation strategy, thereby offering greater flexibility and appeal to users from diverse disciplines. Moreover, keeping the base methodology in mind, our method does provide the grounds for modification by incorporating different types of GOF tests and methods of estimation (supported by simulation study) to be tailored for specific applications and yield favorable results.

On the other hand, Thompson et al. [49] proposed an efficient, automated approach to selecting thresholds for GPD that is both computationally economical and conceptually straightforward. Their method leverages the distributional behavior of parameter estimates across varying thresholds to identify optimal threshold levels. This technique has been applied to rainfall and wave height data to demonstrate its practicality and reliability. To quantify uncertainty in threshold selection, Thompson et al. [49] used the bootstrap method, assessing its impact on return level estimation. They further validated their approach through a simulation study, comparing it to the established JOINSEA software. Additionally, they introduced a method to allow the threshold

choice to adjust dynamically based on covariates, such as the cosine of wave direction, enhancing flexibility and applicability.

For a direct comparison of both [44] and [49] methods with our automated threshold selection approach, Table 3 presents the results obtained by applying the AD-L-moments combination to the dataset from parameter set 1, as specified in Section 2.3. This table provides a comparative analysis of all three threshold selection methods using identical datasets. Across all sample sizes, our approach demonstrates overall superior accuracy and consistency in determining u compared to the methods proposed by Solari et al. and Thompson et al. Notably, there is an exception: Solari's method yields exceptionally low bias and RMSE values for u at a sample size of 1000, with a bias of -0.162. This outcome stands out as the most precise result among the automated methods discussed. However, at sample sizes of 500 and 5000, Solari's approach shows higher bias and RMSE for u, although estimates for the remaining parameters remain accurate. Similarly, Thompson's method also exhibits improved accuracy at a sample size of 1000, with reduced bias and RMSE relative to other sample sizes, yet it still deviates from the actual threshold.

Table 3: Comparison of the three methods in terms of threshold determination for parameter set 1.

Method	**Sample Size**	**E**$[u]$	u		
			RMSE	**Bias**	**SE**
Solari	500	27.102	7.914	-7.898	0.511
	1000	34.837	0.569	-0.162	0.546
	5000	54.314	19.323	19.314	0.564
Thompson	500	20.091	14.909	-14.909	0.050
	1000	27.521	7.479	-7.478	0.120
	5000	44.596	9.599	9.596	0.244
Our	500	32.918	2.118	-2.081	0.393
	1000	31.970	3.050	-3.029	0.350
	5000	33.747	1.306	-1.253	0.369

Table 4: Comparison of the three methods in terms of threshold determination for parameter set 2.

Method	**Sample Size**	**E**$[u]$	u		
			RMSE	**Bias**	**SE**
Solari	500	9.529	1.878	-1.870	0.173
	1000	12.059	0.683	0.659	0.179
	5000	18.093	6.695	6.693	0.173
Thompson	500	6.785	4.614	-4.615	0.018
	1000	9.444	1.956	-1.955	0.042
	5000	14.958	3.559	3.558	0.086
Our	500	11.301	0.164	-0.099	0.132
	1000	11.021	0.396	-0.378	0.119
	5000	11.526	0.177	0.126	0.124

Similarly, Table 4 provides a comparison of all three threshold selection methods for parameter set 2 (see Section 2.3), each applied to identical datasets. Consistently across all sample sizes,

our method demonstrates a notable advantage in consistent accuracy in determining u over the methods by Solari et al. and Thompson et al. Although Solari's technique achieves particularly low bias and RMSE for u at a sample size of 1000, with a bias of 0.659, our approach yields even more precise outcomes, achieving a bias of -0.378 for u. However, at sample sizes of 500 and 5000, Solari's method shows increased bias and RMSE for $\hat{u}$. Similarly, Thompson's approach improves in bias and RMSE at a sample size of 1000 compared to smaller and larger samples, though it remains less accurate than our method relative to the actual threshold.

3.3 Threshold and return level uncertainties

Our research incorporates two distinct real-world datasets: daily rainfall data with over 17,000 observations, representing a large-scale dataset, and the daily closing prices of the Dow Jones index, which has around 1,300 observations, representing a comparatively smaller dataset. The combination of these datasets, each from different domains and with contrasting sizes, serves to demonstrate the versatility and generalization of our method across varied applications. Additionally, these datasets align with the scale of many real-world data sources, where observational counts can range significantly depending on the field. In our simulation study, we preemptively accounted for these variations by setting up three distinct sample sizes: 500, 1000, and 5000, representing small, moderate, and large datasets, respectively.

This approach not only aligns with common scales found in practical datasets but also helps create a simulation framework that can address diverse data sizes across disciplines. The selection of 5000 as the large dataset size was made with the real-life size of the rainfall data in mind, as it captures the large-sample dynamics we aimed to replicate. Simultaneously, 5000 is reasonably close to the 1,300 observation size of the Dow Jones index data, allowing us to analyze uncertainties in a dataset size that bridges both the large-scale data, such as the rainfall dataset, and smaller but substantial datasets like the Dow Jones index. This choice also enhances the practical relevance and methodological rigor of our uncertainty analysis. By adopting 5000 as a representative large sample, we enable meaningful comparisons with existing automated methods while also achieving a level of statistical power and stability essential for reliable uncertainty estimation with the bootstrap percentile method.

To evaluate the uncertainties associated with our approach, we examine the results for both parameter sets 1 and 2, focusing on a sample size of 5000 across each combination of GOF and method of estimation. Based on the simulation findings in Section 3.1, where AD-L-moments and CVM-L-moments emerged as promising combinations, and the comparison analysis in Section 3.2, which further utilized AD-L-moments, this pair will be the primary focus for threshold uncertainty analysis. Notably, we found that the AD-L-moments combination is among those with the lowest uncertainty levels for both threshold and return levels, reflecting its robustness for both parameter sets.

To assess threshold and return level uncertainties across the three automated threshold selection methods, an identical dataset of sample size 5000 from both parameter sets was applied. Table 5 provides a direct comparison of threshold uncertainties for parameter set 1, showing that Thompson's method achieved the narrowest 95% Confidence Interval (CI), while Solari's approach displayed the widest. Although Thompson's method yielded the least threshold uncertainty, it produced a considerable bias in determining u, which is 14.683, in contrast to our method's lower bias of -3.246. Additionally, threshold uncertainty from our approach remains moderate, reinforcing its effectiveness.

Table 5: Comparison of the three methods in terms of threshold uncertainties for parameter set 1. In this table, u is the determined threshold value.

Method	u	CI	Width of CI
Thompson	49.683	[27.210, 52.570]	25.360
Solari	57.766	[27.049, 81.909]	54.860
Our	31.754	[16.544, 50.769]	34.225

Table 6 presents return level estimates for 50, 100, and 1000-year return periods, along with corresponding uncertainties, across the three automated threshold selection methods, using the same dataset as in Table 5 for parameter set 1. Here, the bootstrap percentile approach incorporates the entire threshold selection process to compute 95% CI for each return period. As highlighted by Coles et al. [13], the threshold u is fundamental to return level estimation, and is define as,

$$\hat{Z}_r = \begin{cases} u + \dfrac{\hat{\sigma}}{\hat{\xi}}\left[(rn_y\hat{\beta}_u)^{\hat{\xi}} - 1\right], & \text{if } \xi \neq 0, \\ u + \hat{\sigma}\log(rn_y\hat{\beta}_u), & \text{if } \xi = 0, \end{cases} \tag{16}$$

where β_u is the probability of any value exceeding the threshold point u, $\beta_u = P(X > u)$. The estimated value of $\hat{\beta}_u$ is known as the empirical threshold exceedance. The estimate of the level Z_r is exceedance on average once every r observation is obtained. In other words, $\hat{Z}_r$ is the r observation return level. Giving return levels on an annual scale, however, is frequently more practical. This way, the r year return level represents the level anticipated to be exceeded once every r year. This corresponds to the t-observation return level with $t = rn_y$ if there are n_y observations annually. Hence, it is also important to evaluate the bias in u estimates to enhance forecast reliability.

Table 6: Comparison of the three methods in terms of return level uncertainties for parameter set 1 at 50, 100 and 1000 years return periods.

Method	Return period	Return Level	CI
Thompson	50	114.964	[93.641, 143.362]
	100	121.052	[95.726, 156.509]
	1000	139.334	[100.304, 208.546]
Solari	50	132.141	[94.767, 156.807]
	100	147.122	[98.402, 180.489]
	1000	210.476	[104.406, 331.574]
Our	50	114.891	[95.240, 142.811]
	100	121.162	[97.851, 156.672]
	1000	140.306	[104.151, 206.162]

In Table 6, our method and Thompson's approach show closely aligned return level estimates and CI across the return periods. However, Thompson's method reveals a substantial bias of 14.683 in determining u, in contrast to our method's lower bias of -3.246, indicating a more accurate threshold estimation in our approach. Meanwhile, Solari's method produces higher return level estimates at each period with a broader 95% CI, suggesting increased uncertainty in its predictions.

Table 7: Comparison of the three methods in terms of threshold uncertainties for parameter set 2. In this table, u is the determined threshold value.

Method	u	CI	Width of CI
Thompson	12.804	[7.450, 17.840]	10.390
Solari	19.455	[9.252, 27.338]	18.086
Our	10.912	[5.524, 17.363]	11.839

Table 7 provides a direct comparison of threshold uncertainties among the three methods for parameter set 2, revealing that both Thompson's method and our approach produce a 95% CI of similar width, with our method achieving a slightly more precise determination of u. Solari's method, however, demonstrates lower accuracy in threshold estimation, accompanied by a notably broader 95% CI. Regarding return period estimates and associated uncertainties for parameter set 2, Table 8 shows that Thompson's method yields statistically insignificant return level estimates across all return periods. Conversely, Solari's method provides statistically significant return level estimates, but with a wider 95% CI than those from our approach. Overall, it is safe to say from the entire simulation study that our method consistently provides competitive results in all conditions.

Table 8: Comparison of the three methods in terms of return level uncertainties for parameter set 2 at 50, 100 and 1000 years return periods.

Method	Return period	Return Level	CI
Thompson	50	72.089	[29.440, 42.183]
	100	73.612	[30.042, 45.385]
	1000	77.990	[31.011, 55.828]
Solari	50	39.480	[29.992, 45.504]
	100	43.052	[30.738, 51.129]
	1000	57.148	[32.030, 90.051]
Our	50	34.919	[29.895, 41.619]
	100	36.344	[30.506, 44.684]
	1000	40.357	[31.911, 54.771]

3.4 Practical applications

3.4.1 Daily rainfall in South West England

After analyzing the daily rainfall data for South West England, the automatic threshold selection technique (AD-L-moments) was applied, yielding a threshold, u_0 of 33.668mm, which is a fair determination as the maximum value is 86.6mm in the dataset. The scale and shape parameters, calculated using L-moments, are 8.167 and 0.189, respectively. To assess model fit, we ran a diagnosis with the Probability-Probability (PP) and QQ plots to validate the model's accuracy and they showed a strong alignment with observed data. Additionally, we compared our automated threshold approach with the MRL plot method as discussed by Coles et al. [13], which suggests a threshold of 30mm, as depicted in Figure 1. According to Thompson et al. [49], a threshold is generally defined at a point where the MRL plot exhibits linearity, subject to sampling variability. According to Coles et al. [13], linearity is observable between thresholds of 30 and 60mm, though at 60mm, limited data above this level may introduce sampling errors, as noted by Thompson et

al. [49]. Thus, 30mm emerges as a practical threshold selection. A similar rationale can support our automated threshold choice of 33.668mm. The MRL plot presents challenges in interpretation, with determining the most appropriate threshold often relying on subjective judgment by the researcher. These interpretational complexities and the inherent subjectivity in MRL plot assessments are well-illustrated by Thompson et al. [49]. In addition, we conducted diagnostic assessments using PP and QQ plots for the model proposed by Coles, as well as models developed according to the methods of Thompson and Solari. These diagnostics indicated an acceptable fit across all models, demonstrating alignment comparable to the fit achieved with our proposed method.

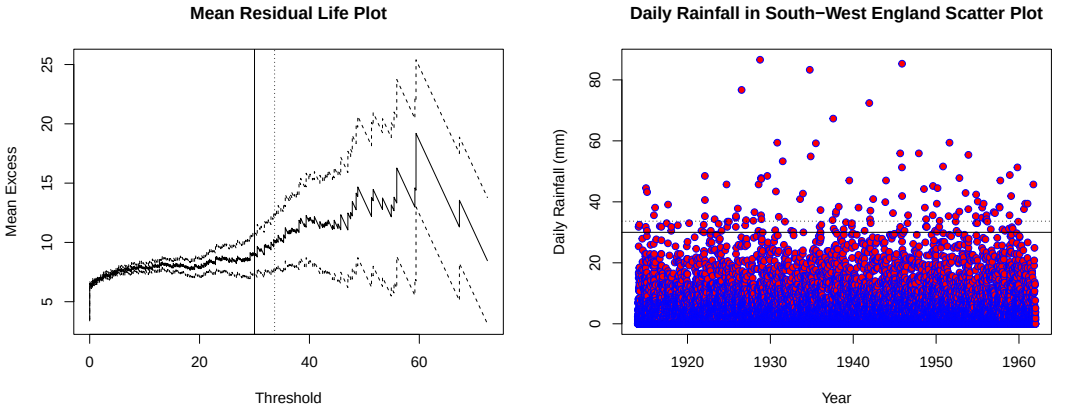

Figure 1: Visualization of the MRL plot and scatter plot based on daily rainfall data from South-West England. In both plots, the dotted line indicates the threshold selected by our method, while the solid line represents the threshold selected by Coles.

However, the automated threshold selection techniques stand out by eliminating the need for user familiarity with complex graphical procedures and interpretations, thereby reducing subjective influence in the threshold selection process. Hence, in the automated procedures, the threshold uncertainties can be quantified, unlike in the MRL plot technique. The 95% CIs for all three automated threshold selection techniques are presented in Table 9.

Table 9: Comparison of the three methods in terms of threshold uncertainties for South-West England daily rainfall dataset. In this table, u is the determined threshold value.

Method	u	**CI**	**Width of CI**
Thompson	23.100	[12.180, 24.100]	11.920
Solari	32.300	[24.700, 59.395]	34.695
Our	33.668	[23.777, 35.037]	11.260

Table 9 reveals that our method consistently yields the narrowest CI, indicating the lowest level of threshold uncertainty, while Solari's method shows the widest interval, suggesting the greatest uncertainty. The results in Table 9 emphasize that our method provides a more precise threshold estimate, reducing the degree of uncertainty compared to other techniques. This highlights its reliability in extreme value modeling, where precise threshold selection is crucial. Although Thompson's method achieves a similar width of uncertainty to our approach, it tends to determine the threshold, u, much lower compared to both the automated threshold selection methods and that of Coles. This discrepancy arises from a constraint within Thompson's method, which stipulates

a minimum of 100 exceedances above the chosen u. This requirement effectively pushes the determined u lower than any estimates generated by the other approaches discussed in this study. In this case, Thompson's method produces a threshold of 23.1mm, yielding 349 exceedances, while our threshold of 33.668mm results in 93 exceedances. While Thompson's method may seem to benefit from a larger dataset, this does not necessarily translate to more accurate parameter estimates.

Table 10: p-value comparison among all four methods using KS, AD and CVM GOF tests.

Method	p**-values**		
	KS	**AD**	**CVM**
Thompson	0.231	0.177	0.268
Solari	0.958	0.963	0.980
Coles	0.887	0.857	0.943
Our	0.966	0.989	0.981

Including a greater number of moderate rainfall values could introduce bias, potentially underestimating the return significant rainfall and compromising the representativeness of extreme conditions, as noted by Liang et al. [28]. To show further evidence, Table 10 provides a direct comparison of how well each threshold selection method aligns with the observed data through the computed p-values while Table 11 shows the respective return level estimates along with 95% CIs. These results allow us to assess which method produces the most statistically reliable model fit for extreme value analysis. The p-values presented in Table 10 are derived from the KS, AD, and CVM GOF tests, which assess the compatibility of the fitted GPD model with the observed data under each threshold selection method. In each approach, the threshold u was determined based on the respective methodology, whether through predefined statistical criteria, optimization of GOF measures, or automated selection techniques. Once u was established, the parameters $\hat{\sigma}$ and $\hat{\xi}$ were estimated, and the empirical CDF of the dataset was compared against the theoretical GPD CDF. The test statistics were then computed and evaluated against their respective null distributions to obtain the corresponding p-values.

Since parameter estimation influences the distribution of these test statistics, adjustments were applied within the respective R packages used for computation. Specifically, the KS test was performed using the $stats$ package, while the AD and CVM tests were conducted using the $goftest$ package. These p-values quantify the extent to which the fitted GPD model aligns with the observed data, with differences across methods reflecting variations in threshold selection and parameter estimation strategies. Table 10 shows that Thompson's method yields lower p-values compared to the other three methods, indicating a lesser degree of model compatibility with the data. As defined by Asserstein and Lazar [51], the p-value is the probability of observing a test statistic as extreme or more extreme than the observed one, assuming the null hypothesis holds. A smaller p-value signifies a stronger deviation from the null hypothesis, or lower model compatibility, given that the underlying assumptions are valid.

Additionally, as outlined in the ASA Statement on Statistical Significance and p-values and reiterated by Greenland et al. [24], p-values serve as a measure of the model's fit, ranging from zero (indicating no compatibility) to one (indicating complete compatibility) with the observed data. From this perspective, the low p-values for Thompson's method suggest it may be less aligned with the data than the other methods, potentially affecting the GPD model's reliability. In contrast, our approach produces the highest p-values as presented in Table 10, indicating a stronger alignment between the model assumptions and the observed data. Table 11 reveals that, as an-

ticipated, Thompson's method consistently yields return level estimates at 50, 100, and 1000-year return periods that are significantly lower than those generated by the other three methods. This trend supports previous findings on the tendency of lower thresholds to yield underestimated return levels, as discussed by Liang et al. [28]. Our automated threshold approach offers return level estimates that align closely with those of both Coles's and Solari's methods, indicating comparable accuracy. It is noteworthy that, given the inherent subjectivity of the MRL plot technique, the delta method described in [38] was applied to calculate return level uncertainties.

Table 11: Comparison of the four methods in terms of return level estimates and uncertainties for daily rainfall dataset of South-West England at 50, 100 and 1000 years return periods.

Method	Return Period	Return Level	CI	Width of CI
Thompson	50	82.432	[68.549, 100.847]	32.298
	100	98.049	[73.413, 116.213]	42.800
	1000	122.672	[90.104, 182.467]	92.363
Solari	50	92.946	[69.837, 110.173]	40.336
	100	107.277	[75.801, 132.212]	56.411
	1000	170.935	[86.837, 257.565]	170.728
Coles	50	92.336	[64.167, 120.507]	56.340
	100	106.342	[65.481, 147.204]	81.723
	1000	168.098	[51.841, 284.354]	232.513
Our	50	91.0153	[70.894, 118.042]	47.148
	100	103.992	[76.502, 147.542]	71.040
	1000	159.100	[93.274, 308.868]	215.594

Most 95% confidence intervals are quite similar across methods, except for the 1000-year return period, where our method, although not the widest, shows a broader uncertainty range. However, due to the extended projection horizon of the 1000-year return level, some variability in uncertainty is not unexpected and should not be a significant cause for concern in practical applications involving rainfall extremes. Table 11 indicates that Thompson's method yields the narrowest return level uncertainty ranges for each return period. However, the return level values themselves raise concerns about underestimation and potential model bias, which could impact the reliability of predictions, especially for extreme values.

3.4.2 Daily closing prices of the Dow Jones index

The Dow Jones index data, as discussed in Section 1, offers an additional case for examining the threshold exceedance model's effectiveness. Due to the evident non-stationarity in the original series $X_1, \ldots, X_n$ and the strong trend component visible in the left panel of Figure 2, a transformation is applied to address this.

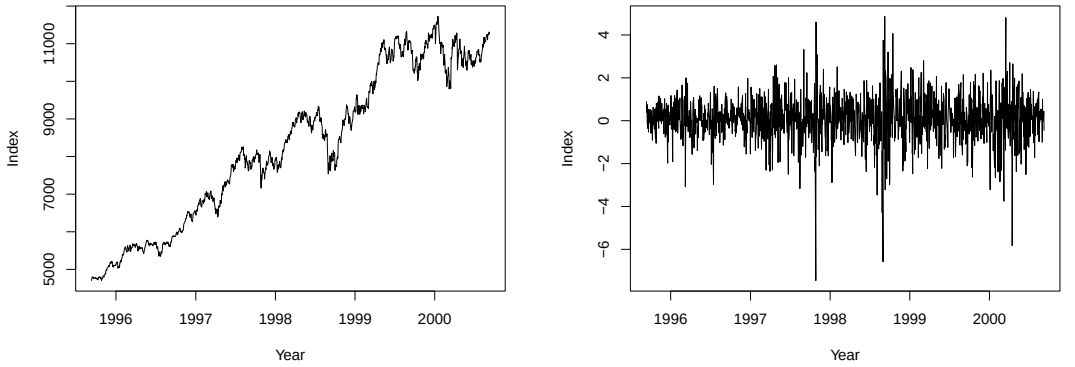

Figure 2: (a) Daily closing price of Dow Jones index. (b) Transformed daily returns of the Dow Jones Index.

Specifically, the data are transformed as $\tilde{X}_i = \log(X_i) - \log(X_{i-1})$, and then rescaled by multiplying by 100 for ease of presentation. This transformation follows the approach suggested by Coles et al. [13]. In Figure 2(b), the transformed series demonstrates a reasonable approximation to stationarity, as further confirmed by the augmented Dickey-Fuller test, which produces a p-value of 0.01, indicating strong evidence of stationarity. Following the analysis of the transformed dataset, our automated threshold selection technique identified an optimal threshold, $u = 0.932$, with corresponding parameter estimates of $\hat{\sigma} = 0.540$ and $\hat{\xi} = 0.084$. We again conducted model diagnostics, including the PP and QQ plots, which confirmed an excellent fit, as the model closely aligns with observed data points. For comparison, we also applied the MRL plot approach outlined in [13], which suggested a threshold of $\hat{u} = 2$, as shown in Figure 3.

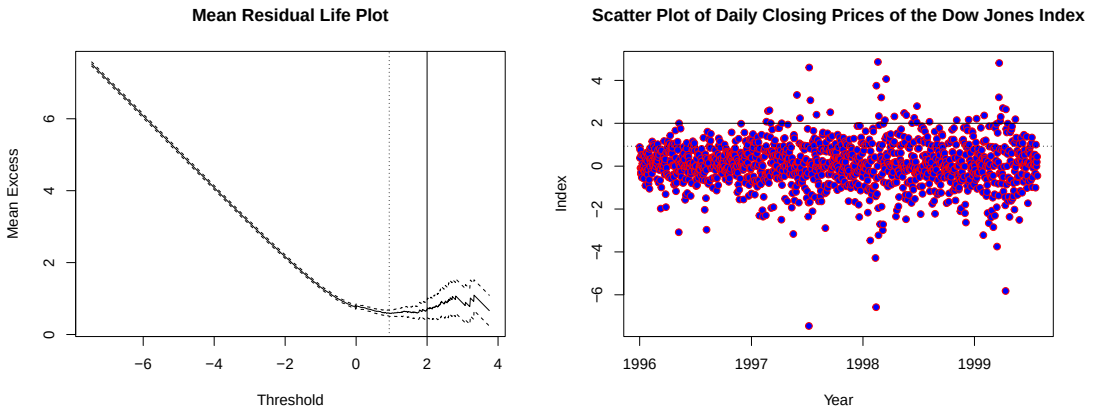

Figure 3: Visualization of the MRL plot and scatter plot based on the transformed daily returns of the Dow Jones index. In both plots, the dotted line indicates the threshold selected by our method, while the solid line represents the threshold selected by Coles.

The MRL plot also clearly shows linearity when the threshold is 0.932. The PP and QQ diagnostic plots for the fitted model suggested by Coles, Solari and Thompson also revealed a strong

fit across all four methods, underscoring minimal practical discrepancies between them based on these diagnostic visuals. This similarity highlights that the methods, despite varying approaches, yield comparable graphical assessments. To deepen this comparison, examining the threshold uncertainty quantifications for each automated method, along with p-values from model fitting across all four techniques, can offer further meaningful insights. Table 12 presents a comparative summary of the threshold uncertainty assessments for the three automated methods, applied to the transformed daily returns of the Dow Jones index.

Table 12: Comparison of the three automated methods in terms of threshold uncertainties for transformed daily returns of the Dow Jones index. In this table, u is the determined threshold value.

Method	u	**CI**	**Width of CI**
Thompson	1.049	[0.821, 1.552]	0.731
Solari	1.693	[0.752, 3.205]	2.453
Our	0.932	[0.713, 2.271]	1.558

In Sections 3.3 and 3.4.1, we explored some limitations inherent to Thompson's method; however, this does not imply unreliability across all datasets. For example, in this case of transformed daily returns of the Dow Jones index dataset, Thompson's method demonstrates strong performance alongside our proposed approach. As shown in Table 12, Thompson's method achieves a low uncertainty level, although not with the lowest u value, as that is produced by our method. Solari's method, on the other hand, yields the broadest 95% CI for u. To evaluate further model fit across all four techniques, we now examine the associated p-values. The p-values for KS, AD and CVM GOF tests for the fitted model using all four methods are given in Table 13.

Table 13: p-value comparison among all four methods using KS, AD and CVM GOF tests.

Method	p**-values**		
	KS	**AD**	**CVM**
Thompson	0.974	0.995	0.985
Solari	0.948	0.990	0.993
Coles	0.885	0.633	0.982
Our	0.999	0.998	0.999

Our fitted model demonstrates the highest p-values, indicating optimal compatibility with the GPD model. Unlike the rainfall dataset, Thompson's method also shows relatively high p-values across all GOF tests, alongside Solari's method. In contrast, Coles's method exhibits lower p-values specifically for the AD test, with a threshold estimate of $u = 2$, notably higher than those produced by automated methods.

Table 14: Comparison of the four methods in terms of return level estimates and uncertainties for transformed daily returns of the Dow Jones index at 50, 100 and 1000 years return periods.

Method	Return Period	Return Level	CI	Width of CI
Thompson	50	7.414	[4.473, 11.791]	7.318
	100	8.296	[4.658, 13.933]	9.275
	1000	11.934	[5.182, 25.225]	20.043
Solari	50	8.243	[4.766, 13.089]	8.323
	100	9.502	[4.909, 16.312]	11.403
	1000	15.586	[4.989, 37.595]	32.606
Coles	50	10.679	[-3.683, 25.040]	28.723
	100	12.975	[-8.107, 34.057]	42.164
	1000	24.910	[-40.826, 90.646]	131.472
Our	50	7.322	[4.756, 15.830]	11.074
	100	8.121	[4.856, 21.300]	16.444
	1000	11.169	[5.070, 58.915]	53.845

A higher threshold can often ensure that extreme values are adequately representative, though the reduced sample size can lead to increased variability in the estimates and uncertainty in predictions for extreme outcomes. Consequently, while the distributional estimates tend to exhibit reduced bias, their variances become larger [28]. Using the delta method to calculate the 95% CI for return levels (in this context, value-at-risk [13]), this increased variance translates into greater uncertainties, as shown in Table 14. Our method achieves the lowest u, which allows for a larger number of exceedances, enhancing the stability of model fitting. While Thompson's method offered a similar advantage in the South West England rainfall dataset (Section 3.4.1), it faced limitations due to concerns over model fit and potential underestimation of return levels. In contrast, in this analysis, our model not only exhibited the highest p-values among the four methods, confirming its strong compatibility with the GPD, but also produced comparable return level estimates. As presented in Table 14, the return level estimates across our method, Thompson's, and Solari's method remain closely aligned across all return periods. However, Coles's method produces substantially higher estimates with considerably wider 95% CIs, as illustrated in Table 14. Among the methods, Thompson's yields the narrowest uncertainties for return level estimates at each return period, underscoring its precision in this specific context.

4 Conclusion

Based on the comprehensive results and analysis in this paper, our proposed automated threshold selection method demonstrates consistent accuracy, reliability, and low uncertainty across diverse datasets, particularly when compared to existing methods by Solari et al. [44], Coles et al. [13] and Thompson et al. [49]. Throughout the simulation study, AD-L-moments and CVM-L-moments combinations emerged as robust choices, yielding minimal bias, RMSE, and SE, which affirms their suitability for GPD modeling under varying conditions. In real-data applications, such as the South West England rainfall and Dow Jones index datasets, our approach effectively addresses limitations in existing threshold selection techniques by minimizing subjective bias and ensuring model compatibility, as evidenced by high p-values and return level estimates that are both logically consistent and competitively robust in comparison to established methods. Unlike

traditional approaches like the MRL plot, our automated method enhances objectivity and accuracy while providing narrower confidence intervals for threshold estimates and return levels.

Furthermore, the comparative studies underscore our method's advantage in reliably identifying the optimal threshold while maintaining high compatibility with the GPD model. Although Solari and Thompson's methods provide accurate estimates for certain datasets, they exhibit increased uncertainties and limitations in determining threshold values in some instances. This distinction highlights our method's ability to balance accuracy and model stability. Therefore, by integrating this automated approach into extreme value analysis, we offer a valuable tool that improves threshold selection processes and extends the reliability of GPD modeling across applications in environmental science, finance, and risk management, advancing the field's capability to handle rare and extreme events effectively. While our proposed method achieves a strong balance of reliability, precision, and compatibility within the GPD framework, opportunities remain for further refinement. We anticipate that with targeted enhancements, future approaches could narrow both threshold and return level uncertainties even further. These advancements may help establish this methodology as a definitive tool for EVA in GPD modeling, paving the way for even greater accuracy and robustness in practical applications.

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

Conflicts of Interest The authors declare no conflict of interest.

References

[1] M. H. Abdul Majid & K. Ibrahim (2021). On Bayesian approach to composite Pareto models. *PLOS One*, *16*(9), Article ID: e0257762. https://doi.org/10.1371/journal.pone.0257762.

[2] S. A. Abu Bakar, S. Nadarajah, Z. A. ABSL Kamarul Adzhar & I. Mohamed (2016). Gendist: An R package for generated probability distribution models. *PLOS One*, *11*(6), Article ID: e0156537. https://doi.org/10.1371/journal.pone.0156537.

[3] M. Ahsan-ul Haq (2022). A new Cramèr–von Mises goodness-of-fit test under uncertainty. *Neutrosophic Sets and Systems*, *49*, 262–268. https://doi.org/10.5281/zenodo.6426399.

[4] H. Alaswed (2024). Graphical diagnostics for threshold selection in fitting the generalized Pareto distribution. *Journal of Pure & Applied Sciences*, *23*(1), 90–95. https://doi.org/10.51984/jopas.v23i1.2997.

[5] W. H. Asquith (2011). *Distributional Analysis with L-moment Statistics Using the R Environment for Statistical Computing*. CreateSpace Scotts Valley, California, USA.

[6] B. Bader & J. Yan. *eva: Extreme Value Analysis with goodness-of-fit Testing, CRAN R Package version 0.2.6* 2020. https://cran.r-project.org/web/packages/eva/eva.pdf.

[7] B. Bader, J. Yan & X. Zhang. Automated threshold selection for extreme value analysis via goodness-of-fit tests with application to Batched return level mapping. arXiv: Statistics 2016. https://doi.org/10.48550/arXiv.1604.02024.

[8] J. Beirlant, Y. Goegebeur, J. Segers & J. L. Teugels (2006). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, Chichester, West Sussex, England.

[9] S. C. Borujeni (2009). *Development of L-moment based models for extreme flood events*. PhD thesis, Universiti Putra Malaysia, Selangor, Malaysia.

[10] R. C. H. Cheng & N. Amin (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of The Royal Statistical Society: Series B (Methodological)*, *45*(3), 394–403. https://doi.org/10.1111/J.2517-6161.1983.TB01268.X.

[11] V. Choulakian, R. A. Lockhart & M. A. Stephens (1994). Cramér-von Mises statistics for discrete distributions. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, *22*(1), 125–137. https://doi.org/10.1007/978-3-642-04898-2_198.

[12] J. Chu, O. Dickin & S. Nadarajah (2019). A review of goodness of fit tests for Pareto distributions. *Journal of Computational and Applied Mathematics*, *361*, 13–41. https://doi.org/10.1016/j.cam.2019.04.018.

[13] S. Coles, J. Bawa, L. Trenner & P. Dorazio (2001). *An Introduction To Statistical Modeling of Extreme Values*. Springer, London. https://doi.org/10.1007/978-1-4471-3675-0.

[14] S. G. Coles & J. A. Tawn (1996). Modelling extremes of the areal rainfall process. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(2), 329–347. https://doi.org/10.1111/j.2517-6161.1996.tb02085.x.

[15] S. Curceac, P. M. Atkinson, A. Milne, L. Wu & P. Harris (2020). An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales. *Journal of Hydrology*, *585*, Article ID: 124845. https://doi.org/10.1016/j.jhydrol.2020.124845.

[16] A. C. Davison & R. L. Smith (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *52*(3), 393–425. https://doi.org/10.1111/j.2517-6161.1990.tb01796.x.

[17] A. C. Davison & D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, New York, USA. https://doi.org/10.1017/CBO9780511802843.

[18] D. J. Dupuis (1999). Exceedances over high thresholds: A guide to threshold selection. *Extremes*, *1*, 251–261. https://doi.org/10.1023/A:1009914915709.

[19] B. Efron & R. J. Tibshirani (1993). *An Introduction to the Bootstrap*, chapter Assessing the Error in Bootstrap Estimates, pp. 271–282. Chapman & Hall/CRC, Boca Raton, Florida. https://doi.org/10.1201/9780429246593.

[20] P. Embrechts, C. Klüppelberg & T. Mikosch (2013). *Modelling Extremal Events: For Insurance and Finance*. Springer Science & Business Media, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33483-2.

[21] J. Faraway, G. Marsaglia, J. Marsaglia & A. Baddeley. *goftest: Classical goodness-of-fit tests for univariate distributions, CRAN R Package version 1.2-3* 2019. https://cran.r-project.org/web/packages/goftest/goftest.pdf.

[22] D. Gaigall & J. Gerstenberg (2023). Cramér-von-Mises tests for the distribution of the excess over a confidence level. *Journal of Nonparametric Statistics*, *35*(3), 529–561. https://doi.org/10.1080/10485252.2023.2173958.

[23] E. Gilleland & R. W. Katz (2016). extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, *72*(8), 1–39. https://doi.org/10.18637/jss.v072.i08.

[24] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman & D. G. Altman (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3.

[25] J. Hambuckers, M. Kratz & A. Usseglio-Carleve. Efficient estimation in extreme value regression models of Hedge fund tail risks. arXiv: Statistics 2023. https://doi.org/10.48550/arXiv.2304.06950.

[26] J. R. Hosking (1990). 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. https://doi.org/10.1111/j.2517-6161.1990.tb01775.x.

[27] J. Hosking (2006). On the characterization of distributions by their L-moments. *Journal of Statistical Planning and Inference*, *136*(1), 193–198. https://doi.org/10.1016/j.jspi.2004.06.004.

[28] B. Liang, Z. Shao, H. Li, M. Shao & D. Lee (2019). An automated threshold selection method based on the characteristic of extrapolated significant wave heights. *Coastal Engineering*, *144*, 22–32. https://doi.org/10.1016/j.coastaleng.2018.12.001.

[29] H. W. Lilliefors (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402. https://doi.org/10.2307/2283970.

[30] B. Liu & M. M. Ananda (2023). A new insight into reliability data modeling with an exponentiated composite Exponential-Pareto model. *Applied Sciences*, *13*(1), Article ID: 645. https://doi.org/10.3390/app13010645.

[31] A. Luceño (2006). Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics & Data Analysis*, *51*(2), 904–917. https://doi.org/10.1016/j.csda.2005.09.011.

[32] M. H. A. Majid & K. Ibrahim (2021). Composite Pareto distributions for modelling household income distribution in Malaysia. *Sains Malaysiana*, *50*(7), 2047–2058. http://doi.org/10.17576/jsm-2021-5007-19.

[33] A. S. M. A. Mamun, A. G. Hussin, Y. Z. Zubairi & S. Rana (2020). A modified maximum likelihood estimator for the parameters of linear structural relationship model. *Malaysian Journal of Mathematical Sciences*, *14*(2), 209–220.

[34] F. J. Massey Jr (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.2307/2280095.

[35] R. Mínguez (2025). Automatic threshold selection for generalized Pareto and Pareto–Poisson distributions in rainfall analysis: A case study using the NOAA NCDC daily rainfall database. *Atmosphere*, *16*(1), Article ID: 61. https://doi.org/10.3390/atmos16010061.

[36] C. Z. Mooney, R. D. Duval & R. Duvall (1993). *Bootstrapping: A Nonparametric Approach To Statistical Inference*. Sage Publications, Newbury Park, California. https://doi.org/10.4135/9781412983532.

[37] C. Murphy, J. A. Tawn & Z. Varty (2024). Automated threshold selection and associated inference uncertainty for univariate extremes. *Technometrics*, *67*(2), 1–10. https://doi.org/10.1080/00401706.2024.2421744.

[38] G. W. Oehlert (1992). A note on the delta method. *The American Statistician*, *46*(1), 27–29. https://doi.org/10.2307/2684406.

[39] J. Pickands III (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, *13*(1), 119–131. https://doi.org/10.1214/aos/1176343003.

[40] R. Pyke (1965). Spacings. *Journal of the Royal Statistical Society: Series B (Methodological)*, *27*(3), 395–436. https://doi.org/10.1111/j.2517-6161.1965.tb00602.x.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria 2013. http://www.R-project.org/.

[42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria 2023. http://www.R-project.org/.

[43] K. M. Ramachandran & C. P. Tsokos (2020). *Mathematical Statistics With Applications In R*. Academic Press, London, United Kingdom.

[44] S. Solari, M. Egüen, M. J. Polo & M. A. Losada (2017). Peaks over threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, *53*(4), 2833–2849. https://doi.org/10.1002/2016WR019426.

[45] M. A. Stephens (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, *69*(347), 730–737. https://doi.org/10.1080/01621459.1974.10480196.

[46] M. A. Stephens (2017). *Goodness-of-Fit-Techniques*, chapter Tests Based on EDF Statistics, pp. 97–194. Routledge, New York. https://doi.org/10.1201/9780203753064.

[47] M. A. Stephenson. *The ismev package, CRAN R Package version 1.42* 2006. https://cran.r-project.org/web/packages/ismev/ismev.pdf.

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

[49] P. Thompson, Y. Cai, D. Reeve & J. Stander (2009). Automated threshold selection methods for extreme wave analysis. *Coastal Engineering*, *56*(10), 1013–1021. https://doi.org/10.1016/j.coastaleng.2009.06.003.

[50] P. J. van Staden & M. Loots (2009). Method of l-moment estimation for the generalized lambda distribution. In *Proceedings of the Third Annual ASEARC Conference*, pp. 7–8. New Castle, Australia.

[51] R. L. Wasserstein & N. A. Lazar (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108.