DATA ANALYTICS AND APPLIED MATHEMATICS
E-ISSN: 2773-4854
VOLUME 7, ISSUE 1, 2026, 50 – 67
DOI: https://doi.org/10.15282/daam.v7i1.13686

RESEARCH ARTICLE

Comparative study of threshold selection methods in the generalised Pareto distribution with application to rainfall datasets

Farabe Khan Alif* and Norhaslinda Ali

Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Malaysia

Abstract - Extreme rainfall events pose significant challenges for flood risk management and infrastructure planning, necessitating robust statistical tools for accurate risk assessment. This study rigorously compares four threshold selection methods for the generalised Pareto Distribution across five distinct rainfall datasets from Southwest England, New Zealand, Bangladesh, Singapore, and the US (Seattle). The methods evaluated include the classical mean residual life plot, a goodness-of-fit p-value-based approach, a parameter stability method, and an automated procedure that combines goodness-of-fit testing with the method of estimation. Return level estimates for 10-, 50-, and 100-year events were estimated, with uncertainties quantified via a bootstrap percentile method. The results reveal that, while each method has its merits, the approach based on goodness-of-fit criteria coupled with the method of estimation generally provides a slight edge. It consistently delivers logically interpretable thresholds that balance bias and variance effectively, particularly in managing datasets with a high prevalence of zero rainfall events. Although alternative methods occasionally yield narrower confidence intervals, they sometimes sacrifice the accurate representation of tail behaviour. Importantly, the study does not dismiss the reliability of other techniques; rather, it underscores that threshold selection is inherently dataset dependent. Overall, this study proposes that while each method offers specific advantages depending on the dataset's characteristics, the approach that integrates goodness-of-fit testing with estimation techniques consistently achieves a favourable balance between simplicity, interpretability, and statistical robustness for threshold selection in generalised Pareto modelling of rainfall extremes. These findings highlight the importance of methodological adaptability and contribute valuable insights toward improving flood risk assessments under diverse climatic conditions.

Article history
Received : 7 January 2026
Revised : 18 February 2026
Accepted : 4 March 2026
Published : 31 March 2026

Keywords
Extreme values
Threshold
Generalised Pareto distribution
Goodness of fit
p-values

1. Introduction

Extreme rainfall events represent an escalating global hazard with devastating social and economic repercussions. Worldwide, such events have resulted in thousands of fatalities and billions of dollars in losses. For instance, monsoon flooding in Pakistan caused nearly 2,000 deaths, while the mid-July 2021 deluge across Belgium and Germany led to over 220 fatalities within just two days [1-2]. In China, Zhengzhou recorded an unprecedented hourly rainfall of 201.9 mm on Jul 20 2021, surpassing its historical maximum of 198.5 mm and causing approximately US$5.8 billion in economic losses and nearly 400 deaths, as reported by the Ministry of Emergency Management. Similarly, extreme rainfall linked to Hurricane Ida in September 2021 severely disrupted vital infrastructure in New York City [3]. These events underscore the critical need for robust statistical methodologies that accurately capture the tail behaviour of rainfall distributions. In this context, Extreme Value Analysis (EVA), particularly methods based on the Generalised Pareto distribution (GPD), has become indispensable for estimating return levels and for guiding effective, data-driven flood risk management strategies [4-5]. The GPD, first introduced by Pickands [6] as a model for threshold exceedances, has become a core element of extreme value theory. It is particularly effective for capturing the tail behaviours of distributions, providing a flexible framework for modelling data that exceeds a high threshold. This threshold, defined as the minimum value beyond which observations are classified as extreme, serves as the basis for the peak over threshold (POT) method, which isolates rare extremes from the main body of the data [7]. However, a major challenge in applying the GPD lies in selecting an appropriate threshold, as this decision directly affects both the number of exceedances and the stability of parameter estimates.

Selecting the optimal threshold is critical in EVA, as it directly influences the accuracy of GPD parameter estimates and subsequent return-level predictions [7]. In practice, a lower threshold tends to yield more exceedances; this may include non-extreme observations, leading to underestimation of true return levels and biased estimates with low variability. Conversely, a higher threshold captures only the most extreme events and better represents the distribution's tail behaviour, but at the expense of a reduced sample size, which increases the variance and uncertainty of the estimates [8]. This trade-off between bias and variance remains one of the most challenging aspects of applying the peak-over-threshold method, prompting the development of various automated procedures to determine the optimal threshold objectively. The mean residual life (MRL) plot is a widely used graphical tool for threshold selection in EVA, offering an intuitive means to assess whether the excesses above a chosen threshold follow the GPD. The method examines the expected excesses over candidate thresholds, where a relatively stable linear trend suggests an appropriate choice [8-9]. Despite its utility, the MRL plot is highly dependent on visual interpretation, introducing subjectivity that can lead to inconsistent threshold selection among analysts [7]. Different approaches to automating threshold selection have been

***CORRESPONDING AUTHOR | Farabe Khan Alif|** ✉ gs67618@student.upm.edu.my

proposed, including weight-based threshold selection by Dupuis [10], goodness-of-fit (GOF) p-value-based threshold estimation by Solari et al. [11], and a method based on changes in parameter estimates by Thompson et al. [12]. However, the graphical MRL plot remains the most widely used approach for threshold selection [11].

Despite the increasing development of automated threshold selection methods, graphical techniques such as the MRL plot remain prevalent owing to their intuitive interpretation and widespread familiarity among practitioners [11]. Nonetheless, recent advances underscore the importance of embedding rigorous statistical principles to enhance the robustness of extreme value models. For instance, Gaigall et al. [13] analysed the asymptotic behaviour of the Cramer–von Mises (CVM) statistic for exceedances and incorporated bootstrap methods to account for parameter uncertainty, highlighting the need to address estimation variability. In parallel, Murphy et al. [14] introduced an automated method specifically designed to manage the bias–variance trade-off, thereby improving the stability of extreme value estimates. Likewise, Minguez [15] developed a weighted mean-square-error approach using internally studentized residuals, achieving higher precision in extreme rainfall modelling than conventional techniques. Complementary to these, Alaswed [16] investigated graphical diagnostics and multiple threshold stability plots to mitigate subjectivity and formalize threshold identification. Within the broader domain of extreme value modelling, Curceac et al. [17] refined scale parameter estimation to improve the resilience of automated threshold selection, while [18] proposed a simultaneous estimation framework for both tail and threshold parameters, offering an efficient alternative to the traditional POT approach. Given the range of existing threshold selection techniques, this study focuses on four principal approaches: the MRL plot shown by Coles et al. [4], the modified GOF-based method by Solari et al. [11], the parameter stability approach by Thompson et al. [12], and our recently proposed automated AD-L method in Alif et al. [19], which integrates the Anderson–Darling (AD) GOF test with L-moment parameter estimation. These methods represent both traditional and advanced perspectives in threshold selection, with Solari et al. [11] and Thompson et al. [12] serving as established benchmarks against which newer, computationally efficient strategies can be evaluated. The AD-L method, in particular, offers a robust yet streamlined framework for determining optimal thresholds, aligning with the study's goal of balancing statistical rigor and simplicity. Since MRL plots remain widely used despite their subjectivity, this research aims to promote a more objective yet practical alternative. Accordingly, the study compares the four techniques for GPD threshold estimation, assesses the suitability of automated approaches like AD-L as reliable substitutes for graphical methods, and evaluates the impact of varying threshold strategies on 10-, 50-, and 100-year return level estimates and their 95% confidence intervals to inform hydrological risk assessment.

2. Methodology

2.1. Theoretical Background

The GPD serves as a fundamental tool in extreme value theory, specifically for modeling exceedances above a selected threshold u [4]. In this context, a threshold defines the lower bound beyond which data points are classified as extreme, allowing for an effective representation of tail behavior. Selecting an appropriate threshold is a critical step, as it directly influences the trade-off between bias and variance in parameter estimation [5]. Setting the threshold too low may introduce deviations from the GPD assumption, whereas an excessively high threshold can significantly reduce the available sample size, increasing estimation variability. Originally introduced by Pickands [6], the GPD is characterized by two key parameters: the scale parameter σ and the shape parameter ξ, as shown in Eq. (1). The cumulative distribution function (CDF) representing the relationship among these parameters and the threshold u can be defined as Eq. (1).

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

where x represents an independent and identically distributed (iid) random variable, where the domain of x depends on the shape parameter ξ. Specifically, for $\xi \geq 0$, the excess $x > u$, while for $\xi < 0$, the constraint $u \leq x < u - \sigma$ ensures a finite upper bound. The model parameters are constrained as follows: $\sigma > 0$, $-\infty < \xi < \infty$, and $-\infty < u < \infty$. The shape parameter ξ plays a crucial role in defining the tail behavior of the distribution: when $\xi > 0$, the distribution exhibits heavy tails; $\xi = 0$ corresponds to an exponential decay; and $\xi < 0$ imposes a finite upper bound on the data [18,20].

For general comparison, we will use L-moments to estimate the GPD model parameters from real-life rainfall datasets. L-moments provide a robust statistical framework for characterizing probability distributions and estimating their parameters. They are a sequence of statistics used to summarize distributional properties and serve as alternatives to conventional moments. Defined as linear combinations of order statistics (L-statistics), L-moments allow for the computation of quantities analogous to standard deviation, skewness, and kurtosis, referred to as L-scale, L-skewness, and L-kurtosis, respectively [21-22]. Standardized L-moments, known as L-moment ratios, provide insight into the shape of a distribution. A theoretical distribution has a set of population L-moments, while sample L-moments are derived from empirical data and serve as estimators of these population characteristics. L-moments are robust to the influence of outliers and remain defined as long as the mean of the distribution is finite [23]. [24] demonstrated that L-moments provide a unified framework for drawing statistical inferences from continuous univariate distributions. This approach is particularly advantageous in scenarios with small sample sizes, where traditional estimation techniques, such as maximum likelihood estimation (MLE) or the method of moments, may fail to perform accurately. The robustness of L-moments to sample variability enhances their applicability in real-world datasets that may include extreme or rare events. These attributes make L-moments highly suitable for modelling and inference involving data characterized by extreme values. Finally, the computational efficiency of L-moments offers an advantage, especially when implementing bootstrapping

techniques for uncertainty analysis, as noted by [11]. For a continuous random variable Y, the quantile function $Q_{Y(p)}$ is defined as the value of y such that the CDF satisfies$F_{Y(Y)} = p$ for $0 \leq p \leq 1$, where $F_{Y(Y)} = p$ is the CDF of Y [25]. Hosking's [21] comprehensive work on L-moment theory provides a foundation for its application. The rth L-moment, expressed in terms of the quantile function, is defined as Eq. (2).

$$L_r = \int_0^1 Q_Y(p)\, P_{r-1}^*(p)\, dp, \tag{2}$$

where in Eq. (2), $P_{r-1}^*(p)$ can be defined as,

$$P_{r-1}^*(p) = \sum_{k=0}^{r} \left[(-1)^{r-k} \binom{r}{k} \binom{r+k}{k} p^k\right] \tag{3}$$

Eq. (2) represents the shifted Legendre polynomial of order $r - 1$, and Lr denotes the rth L-moment. For further details, refer to [48]. To estimate the parameters using L-moments in this study, we will be utilizing the R package extRemes [18].

Next, to compare the compatibility of models generated by different threshold selection methods across datasets, we will use GOF tests and evaluate their p-values. Smaller p-values indicate stronger statistical incompatibility of the data with the null hypothesis, assuming the assumptions for p-value calculations hold, as noted in the literature [26]. This metric, ranging from 0 (indicating total incompatibility) to 1 (indicating perfect compatibility), assesses how well the model fits the data [27]. Before conducting the GOF tests, the GPD parameters must be estimated. In this study, we will employ three GOF tests: the Kolmogorov-Smirnov (KS) test, the Anderson-Darling (AD) test, and the CVM test, to determine the most appropriate threshold selection method.

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 fits the observed data. It is computed as Eq. (4).

$$KS = \left(\left|G\left(x_{(i)}; u, \hat{\sigma}, \hat{\xi}\right) - \frac{i-1}{n}\right|, \left|\frac{i}{n} - G\left(x_{(i)}; u, \hat{\sigma}, \hat{\xi}\right)\right|\right) \tag{4}$$

where in Eq. (4), n is the sample size, $x_{(i)}$ represents the ith order statistic, and $G\left(x_{(i)}; u, \hat{\sigma}, \hat{\xi}\right)$ denotes the estimated CDF of the GPD. Simulation-based evaluations by Chu et al. [7] highlight the effectiveness of the KS test in GPD modelling, particularly in detecting deviations across the entire distribution. Unlike KS, the AD test places greater emphasis on tail discrepancies, making it highly relevant for extreme-value analysis. Eq. (5) represents the test statistics for the AD test used in this study. The test assesses whether a given sample originates from a specified probability distribution. In its simplest form, AD assumes that the distributional parameters are known, enabling the derivation of critical values within a distribution-free framework. However, in practical applications involving parameter estimation, adjustments to the test statistic or its critical values are necessary to account for estimation uncertainty [29]. The AD statistic is given by

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n} \left[(2i-1)\, log\, log\, G\left(x_{(i)}; u, \hat{\sigma}, \hat{\xi}\right) + log\, log\, \left(1 - G\left(x_{(n+1-i)}; u, \hat{\sigma}, \hat{\xi}\right)\right)\right] \tag{5}$$

where $G\left(x_{(i)}; u, \hat{\sigma}, \hat{\xi}\right)$ represents the estimated CDF of the GPD at the ith order statistic. The weighting scheme used in AD enhances its sensitivity to extreme deviations, making it one of the most powerful goodness-of-fit (GOF) tests for detecting deviations in the tails of distributions [30-31].

The CVM test, another widely used GOF measure, assesses the alignment of sample data with a theoretical distribution [32]. Eq. (6) represents the test statistic for the CVM test.

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[G\left(x_{(i)}; u, \hat{\sigma}, \hat{\xi}\right) - \frac{2i-1}{2n}\right]^2 \tag{6}$$

This test is particularly useful in evaluating the goodness-of-fit by quantifying the overall discrepancy between the empirical and theoretical distributions [1]. In the context of GPD modelling, CVM has been shown to perform comparably to KS, offering a robust approach for model validation [28]. The p-value for each GOF test quantifies the probability of obtaining a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis [11]. However, since the null distributions of these test statistics are influenced by parameter estimation, obtaining accurate p-values often necessitates numerical approximations or resampling techniques [31]. For the KS test, p-values are computed by comparing the observed KS statistic to its theoretical distribution under the null hypothesis, accounting for parameter-estimation effects [33]. The standard KS test assumes known parameters, but when parameters are estimated from data, the null distribution shifts, requiring adjusted critical values [34]. In this study, we employ the stats package in R, which provides methods for computing p-values from empirical distribution functions [35]. The AD test, unlike KS, does not have a closed-form theoretical null distribution when parameters are estimated [31]. Instead, critical values are obtained through numerical integration or simulation-based approaches [29]. Prior research has highlighted the AD test's sensitivity to parameter estimation, necessitating modifications to either its test statistic or critical values for valid

inference [30]. We utilize the goftest package in R to compute p-values for the AD test [36]. Similarly, the CVM test relies on asymptotic approximations for its null distribution, which are further adjusted when parameter estimation is involved [32]. The CVM test is particularly robust at detecting deviations across the entire distribution, making it a strong complement to KS and AD in extreme-value settings [28]. The goftest package in R is used to determine p-values for the CVM test, leveraging tabulated critical values at common significance levels [36].

2.2 Return Level Estimation

Accurate estimation of return levels is a crucial component of extreme value analysis, particularly after selecting an appropriate threshold u. As discussed by Coles et al. [4], the threshold plays a vital role in defining return levels. The r-observation return level, denoted by $\widehat{Z_r}$, can be computed as:

$$\widehat{Z_r} = \{u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(rn_y\widehat{\beta_u}\right)^{\hat{\xi}} - 1\right], \ if\ \xi \neq 0\ u + \hat{\sigma}\ log\ log\ \left(rn_y\widehat{\beta_u}\right), \ if\ \xi = 0 \quad (7)$$

In Eq. (7), $\widehat{\beta_u}$ is the empirical probability of exceedance, i.e., the estimated proportion of observations greater than the selected threshold u, and ny represents the average number of observations per year. The formula in Eq. (7) provides the return level $\widehat{Z_r}$, which corresponds to the magnitude expected to be exceeded once every r observations. To make return levels more interpretable in practical terms, especially in hydrological applications, they are often expressed on an annual basis. In that context, the r-year return level refers to an event likely to be exceeded once every r years, which corresponds to rny observations when ny observations are recorded annually. Since return level estimation depends on the threshold, any bias or uncertainty in selecting u may directly influence the final estimates. Therefore, assessing the reliability of both threshold and return-level estimates is essential. This is further addressed in our study through uncertainty quantification using bootstrap-based methods.

2.3 Bootstrap Percentile Method

To assess the uncertainty associated with the threshold and return-level estimates, we use the bootstrap percentile method. This method is widely accepted in the literature due to its intuitive interpretation, computational efficiency, and ease of implementation, making it particularly suitable for large-scale data studies [12]. The bootstrap percentile approach is a resampling-based technique for constructing confidence intervals for estimated statistics. Rather than relying on strict distributional assumptions, this method builds an empirical distribution of the estimator by repeatedly resampling the data. In our context, it enables robust estimation of the 95% confidence intervals for both the threshold values and the corresponding return levels. The core steps of the method, as established in the works of Mooney et al. [37] and Tibshirani et al. [38], are adapted in our study as follows:

i) We generate B = 1000 bootstrap samples from the original rainfall dataset by sampling with replacement.

ii) For each resampled dataset, we re-estimate the threshold and the return level using the same set of methodologies.

iii) The collection of all bootstrap estimates is sorted to form empirical distributions of the statistics.

iv) The lower and upper bounds of the 95% confidence intervals are obtained from the 2.5th and 97.5th percentiles of the sorted estimates, respectively.

This method ensures that both the inherent sampling variability and the influence of threshold selection are properly accounted for, offering a more complete picture of the reliability of our extreme value estimates.

2.4 The AD-L Threshold Selection Method

The AD-L method introduced an automated threshold selection approach for GPD modeling, ensuring robustness across datasets of varying sizes and characteristics. The approach systematically partitions the dataset into 200 equal intervals, a choice empirically found to provide a suitable level of granularity without excessive fragmentation. This method generates a structured set of candidate thresholds, which are then assessed using GOF tests to determine the most suitable threshold for extreme value modelling. The method proceeds through the following steps:

i) Partition the Dataset
   a. Sort the dataset in ascending order.
   b. Divide the sorted data into 200 equal intervals.

ii) Determine the Initial Candidate Threshold u_1
   a. If the dataset contains a significant number of zeros:
      i. Compute the mean of each interval ($M_1, M_2, \dots, M_{200}$).
      ii. Identify the first quantile of the full dataset, denoted Q_1.
      iii. Select the means greater than Q_1 and calculate: $u_1 = \frac{\sum_{M_i > Q_1} M_i}{C}$ where C is the number of means exceeding Q_1.
   b. If zeros are not a concern:
      i. Simply set u1 as the median of the dataset.

iii) Define the Endpoint Threshold u_n

a. Set an approximation around the 5th largest observation.
b. Impose the following minimum exceedance rules to ensure statistical reliability:
   i. At least 80 exceedances if the dataset contains at least 5000 observations.
   ii. At least 20 exceedances if the dataset contains 1000-5000 observations.
   iii. No strict minimum if the dataset has fewer than 1000 observations.

iv) Generate Candidate Thresholds
   a. From u_1 to u_n, generate approximately 300 equally spaced candidate thresholds.

v) Estimate Parameters Using L-moments
   a. For each candidate threshold u_k, extract exceedances above u_k.
   b. Estimate the GPD parameters (σ, ξ) using the L-moments method.

vi) Apply the Anderson-Darling Goodness-of-Fit Test
   a. For each threshold u_k, compute the AD test statistic and its corresponding p-value.

vii) Select the Optimal Threshold u_0
   a. Choose the threshold that maximizes the AD p-value:

$$u_0 = arg\ arg\ p_{AD}\ (u_k)$$

This stepwise framework ensures a reproducible, data-driven process for threshold selection in extreme value modelling. Its effectiveness has been validated through simulation and real-world datasets, offering a reliable alternative to traditional graphical methods.

2.5 Threshold Selection Method

2.5.1By Solari et al. [11]

Solari et al. [11] proposed an automated threshold selection approach based on the AD test statistic, specifically the right-tail-weighted version. This methodology aims to systematically determine the optimal threshold by evaluating the GOF between the empirical and parametric distributions. The process begins by identifying all peaks in the dataset using a moving window, ensuring that the selected peaks are independent. The identified peaks are then sorted to form a candidate threshold series, and for each threshold, the GPD parameters are estimated using L-moments. After sorting the identified peaks, duplicate values are eliminated. Since the methodology of Solari et al. [11] does not specify an optimal number of candidate thresholds, we refer to their simulation study. Based on their approach, we adopt 100 candidate thresholds for our comparative analysis, as this is the maximum number considered in their simulation study. The right-tail weighted AD statistic, denoted as A_R^2 , is computed for each candidate threshold. The selected threshold is the one that minimizes $1 - p$, which is equivalent to minimizing the test statistic value since lower values of A_R^2 correspond to higher p-values, indicating better model fit [11]. Given the lack of explicit tables or formulas for obtaining p-values when parameters are estimated via L-moments, the test statistic values are used directly for threshold selection. Based on simulation studies, 100 candidate thresholds will be considered to ensure a comprehensive evaluation. The right-tail weighted AD test, originally introduced by Sinclair et al. [39], modifies the classical test by placing more weight on the upper tail of the distribution. The test statistic for the right-tail weighted AD test is given by Eq. (8). This modification is particularly relevant for extreme-value modelling, where accurate tail estimation is of primary concern. The test statistic A_R^2 is defined as:

$$A_R^2 = \frac{n}{2} - \sum_{i=1}^{n} \left[\left(2 - \frac{2i-1}{n}\right) log\ log\ (1 - z_i)\ + 2z_i\right], \tag{8}$$

where z_i represents the cumulative distribution function evaluated at the ordered sample values x_i, and n is the sample size. Since the parameters of the fitted distribution are estimated from the sample using L-moments, the distribution of A_R^2 is determined through simulation techniques rather than standard tables. Once the optimal threshold is selected, the GPD parameters are re-estimated using data above this threshold. High-return-period quantiles and their confidence intervals are subsequently obtained using nonparametric bootstrapping. The results from Solari et al. [11] demonstrated that, despite high uncertainty in the threshold selection process, the impact on the confidence intervals of high-return-period quantiles remains minimal, confirming the robustness of the methodology. By incorporating the right-tail-weighted Anderson-Darling statistic and systematically evaluating multiple candidate thresholds, this approach provides an automated, statistically sound means of selecting an appropriate threshold for extreme-value analysis.

2.5.2 By Thompson et al. [12]

Thompson et al. [46] proposed an automated, computationally efficient threshold selection method that relies on differences in parameter estimates across thresholds. The core idea is that, for a suitable threshold u, the difference in the estimated scale parameter between consecutive thresholds should be approximately normally distributed with a mean of zero. Their approach systematically tests for this property to determine the lowest valid threshold. The method begins by selecting a set of n equally spaced candidate thresholds, $u_1, u_2, \ldots, u_n$, where the lowest candidate threshold u_1 is the median of the data, and the highest candidate threshold un is set at the 98th percentile of the data unless fewer than 100 values exceed this percentile, in which case un is set to the 100th largest value. The parameters are estimated using

maximum likelihood estimation for exceedances above each threshold. Let $\widehat{\sigma_{u_j}}$ and $\widehat{\xi_{u_j}}$ be the maximum likelihood estimates of the scale and shape parameters for a given threshold u_j. In MLE, the likelihood function is formulated based on the assumed probability distribution of the data, and parameter estimates are obtained by maximizing it. For the GPD, the PDF when $\xi \neq 0$ is presented in Eq. (9).

$$g(x;\sigma,\xi) = \frac{1}{\sigma}\left(1 + \xi\frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}-1} \tag{9}$$

Taking the logarithm on both sides of Eq. (9) to obtain the log-likelihood function, which is shown in Eq. (10).

$$log\ log\ L\ (\sigma,\xi) = -n\ log\ log\ \sigma\ - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{n}\quad log\ log\ \left(1 + \xi\frac{x_i - u}{\sigma}\right) \tag{10}$$

When $\xi = 0$, the GPD reduces to an exponential distribution with PDF given by Eq. (11).

$$g(x;\sigma) = \frac{1}{\sigma} exp\ exp\ \left(-\frac{x-u}{\sigma}\right), \tag{11}$$

Taking the log on both sides of Eq. (11), the corresponding log-likelihood function is presented in Eq. (12).

$$log\ log\ L\ (\sigma) = -n\ log\ log\ \sigma\ - \frac{1}{\sigma}\sum_{i=1}^{n}\quad (x_i - u) \tag{12}$$

MLE provides parameter estimates $\hat{\sigma}$ and $\hat{\xi}$ by maximizing the log-likelihood function. In this study, these estimates are obtained using the R packages extRemes and ismev [18, 19]. According to [4], if $u \leq u_j - 1 < u_j$, the relationship between consecutive scale parameters is given by Eqs. (13) and (14).

$$\sigma_{u_{j-1}} = \sigma_u + \xi(u_{j-1} - u), \tag{13}$$

$$\sigma_{u_j} = \sigma_u + \xi(u_j - u), \tag{14}$$

Eqs. (13) and (14) imply that the difference in scale estimates between consecutive thresholds satisfies Eq. (15).

$$\sigma_{u_j} - \sigma_{u_{j-1}} = \xi(u_j - u_{j-1}). \tag{15}$$

Since the MLE of $\widehat{\sigma_{u_j}}$ and $\widehat{\xi_{u_j}}$ satisfy $E\left[\widehat{\sigma_{u_j}}\right] \approx \sigma_{u_j}$, $E\left[\widehat{\xi_{u_j}}\right] \approx \xi$, for any suitable threshold u, Thompson et al. [11] defined the transformation shown in Eq. (16).

$$\tau_{u_j} = \widehat{\sigma_{u_j}} - \widehat{\xi_{u_j}}u_j. \tag{16}$$

Taking the first-order difference in Eq. (16), they obtained Eq. (17).

$$\tau_{u_j} - \tau_{u_{j-1}} = \left(\widehat{\sigma_{u_j}} - \widehat{\xi_{u_j}}u_j\right) - \left(\widehat{\sigma_{u_{j-1}}} - \widehat{\xi_{u_{j-1}}}u_{j-1}\right). \tag{17}$$

Under a suitable threshold, the differences in Eq. (17) should follow a normal distribution with mean zero, $E\left[\tau_{u_j} - \tau_{u_{j-1}}\right] \approx 0$. To identify a valid threshold, Thompson et al. [12] applied Pearson's chi-square test for normality to the sequence of differences. The procedure starts with the lowest candidate threshold u_1 and considers all differences $\tau_{u_2} - \tau_{u_1}, \tau_{u_3} - \tau_{u_2}, \ldots, \tau_{u_n} - \tau_{u_{n-1}}$. If the normality test does not reject the null hypothesis, u_1 is selected as the threshold. Otherwise, the next threshold u_2 is considered, removing $\tau_{u_2} - \tau_{u_1}$ from the differences, and the test is repeated. This iterative process continues until the test no longer rejects the null of normality. If no candidate threshold satisfies this condition, the highest candidate threshold is returned with a warning.

Through simulation studies, Thompson et al. [12] found that using a 0.2 significance level in the Pearson test yields stable results across different datasets. Lower significance levels tend to yield lower thresholds, while higher levels yield higher thresholds. The method is computationally efficient and does not require graphical interpretation, making it a practical alternative to traditional methods like the mean residual life plot. The authors implemented their approach in R and validated it using rainfall and wave-height datasets, demonstrating its effectiveness across diverse applications. Additionally, they extended their method to allow the threshold to depend on covariates, such as the cosine of wave direction, further increasing its applicability in environmental studies.

2.5.3 By Coles et al. [8]

The MRL plot, also known as the mean excess plot, is a widely used graphical technique in extreme value analysis to determine an appropriate threshold for the GPD. The core idea behind the MRL plot is that for a suitable threshold u, the excesses above u should exhibit a linear trend, which suggests that the tail of the distribution can be well modelled by the

GPD. The MRL plot is constructed using the mean residual life function, which quantifies the expected excess over a given threshold. Formally, for a random variable X and a threshold u, the mean residual life function is given by:

$$e(u) = E(X > u), \tag{18}$$

In Eq. (18), $E(X > u)$ represents the expected value of the excess $X - u$ given that X exceeds u. In practice, for a dataset $x_1, x_2, \ldots, x_n$, Eq. (19) presents the empirical estimate of the mean residual life function.

$$\hat{e}(u) = \frac{\sum_{i=1}^{n} (x_i - u) I(x_i > u)}{\sum_{i=1}^{n} I(x_i > u)}, \tag{19}$$

In Eq. (19), $I(x_i > u)$ is an indicator function that takes the value 1 if $x_i > u$ and 0 otherwise. This function in Eq. (19) essentially calculates the mean of exceedances over each threshold, which is then plotted against the threshold values.

To select an appropriate threshold, analysts examine the MRL plot to identify a region where it exhibits roughly linear behaviour. A suitable threshold u is one at which the plot stabilizes into a linear trend, suggesting that the excess distribution follows a GPD. If the plot remains nonlinear across all candidate thresholds, this may indicate that the GPD model is not suitable for the dataset or that an alternative threshold selection method is required. An example of an MRL plot is illustrated for the Southeast England daily rainfall dataset in Figure 1. In this case, multiple points exhibit linear behaviours, suggesting potential threshold choices. The solid line at u = 30 represents the threshold choice recommended by Coles et al. [4], while a second linear region appears around u = 60, but selecting such a high threshold would result in too few exceedances to reliably estimate the GPD parameters, as noted by Thompson et al. [12]. The MRL plot remains a valuable tool for threshold selection, providing an intuitive visual approach to identifying a suitable threshold for extreme value modelling. However, its reliance on subjective graphical interpretation may introduce variability in the selection process, making it beneficial to compare its results with automated threshold selection methods.

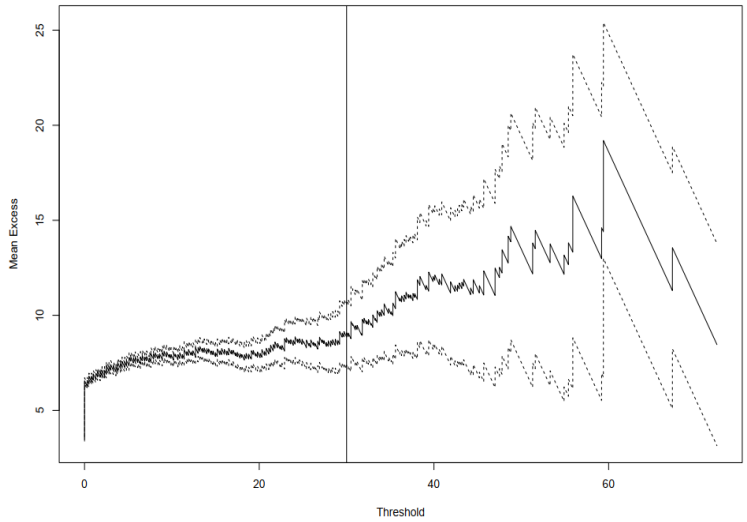

Figure 1. MRL plot illustrating threshold selection for GPD modelling in the Southwest England daily rainfall dataset. The linear behaviour above the selected threshold indicates suitability for GPD application

2.6 Datasets

To comprehensively evaluate the performance of various threshold selection techniques for modeling extreme rainfall, we consider datasets from multiple regions across the globe, namely Southwest England, New Zealand, Bangladesh, Singapore, and Seattle (United States). These locations represent a diverse range of climatic conditions, including temperate maritime, tropical, and subtropical environments, ensuring that the analysis captures a broad spectrum of rainfall patterns. Rainfall datasets are inherently challenging to handle due to their high spatial and temporal variability. The frequency, intensity, and distribution of rainfall events differ significantly across regions, influenced by complex meteorological and topographical factors. By incorporating datasets from distinct geographical locations, this study aims to assess the adaptability of threshold selection methodologies under varied climatic regimes. This approach allows for a rigorous examination of the strengths and limitations of each technique, offering valuable insights into their robustness and reliability when applied to real-world scenarios. Furthermore, including multiple datasets ensures the study does not rely on a single region-specific rainfall characteristic, thereby enhancing the generalizability of the findings. The presence of different data structures, ranging from low-intensity frequent rainfall to extreme, sporadic events, provides an ideal testbed for evaluating threshold selection methods. By subjecting these techniques to datasets with such variability, we can critically assess their effectiveness in capturing extreme values and their potential for broader application in hydrological modelling.

2.6.1 Southwest England daily rainfall dataset

The Southwest England dataset consists of daily rainfall accumulations recorded at a specific location from 1914 to 1962, totaling 17,531 observations [40]. This dataset provides a valuable basis for studying extreme rainfall events and evaluating different threshold selection methodologies. To illustrate the dataset, Figure 2 presents a scatter plot of daily rainfall values (mm), highlighting the distribution and frequency of rainfall events. Additionally, the MRL plot in Figure 1 provides insights into the dataset's extreme-value characteristics. The MRL plot plots the sample mean of excesses over a range of threshold values, along with pointwise confidence intervals, providing a graphical diagnostic tool for selecting an appropriate threshold for extreme value modelling [41]. This technique follows the approach proposed by Coles et al. [4] and is widely used in threshold selection for the GPD.

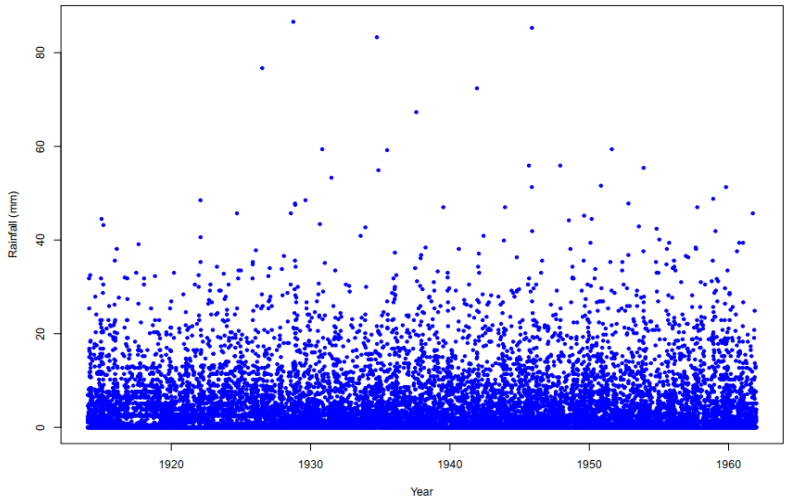

Figure 2. Visualization of the scatter plot representing the characteristics of the daily rainfall dataset from Southwest England

2.6.2 New Zealand daily rainfall dataset

The New Zealand rainfall dataset used in this study comprises daily precipitation records from the Auckland region spanning 1960 to 2019. The data were sourced from the National Institute of Water and Atmospheric Research (NIWA) via the National Climate Database (CLiDB), which compiles high-quality climate records across the country [29]. For our analysis, only the Auckland region was considered, as it provides a continuous, complete time series suitable for extreme-value analysis. Each entry in the dataset corresponds to daily rainfall accumulations measured from 9 am to 9 am the following day. In earlier years, particularly before 1962, missing records were reconstructed using the Virtual Climate Station Network (VCSN) to ensure temporal continuity [42]. After all filtering and pre-processing steps, the dataset contains 21,915 observations.

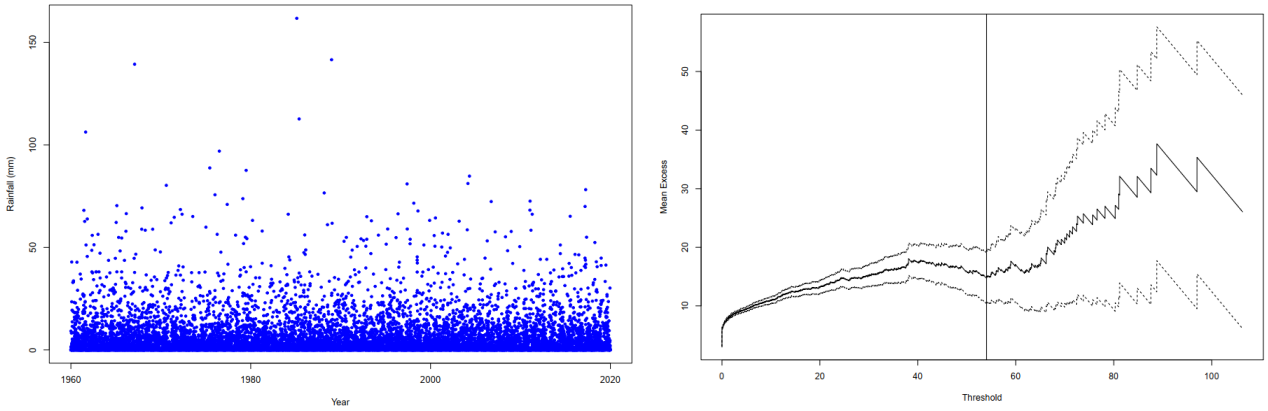

Figure 3. Visualization of the Scatter plot (left panel) and MRL plot (right panel) based on daily rainfall data from Auckland, New Zealand. The solid line in the MRL represents the threshold selected by following the methodology of Coles et al. [4]

To facilitate understanding of the dataset's behaviour, especially from an extreme-value modelling perspective, two key visualizations are presented: a scatter plot of the daily rainfall series and an MRL plot in Figure 3. The threshold choice of 54 mm present in the MRL plot in the right panel of Figure 3 has been chosen by following the methodology instruction from Coles et al. [4] discussed in Section 2.7.

2.6.3 Bangladesh daily rainfall dataset

The Bangladesh rainfall dataset used in this study comprises daily precipitation measurements from 35 strategically distributed weather stations across the country. The data were provided by the Bangladesh Meteorological Department (BMD), which maintains a national network for recording meteorological observations [43]. Although the complete dataset includes rainfall records from multiple cities, this study focuses on data from Dhaka, the capital city, to ensure consistency and maintain a focused regional analysis of rainfall extremes. Each entry in the dataset corresponds to the total daily rainfall recorded over a 24-hour interval. After filtering and pre-processing steps, including the removal of missing values, a total of 23,010 daily rainfall observations from the Dhaka station over the period of 1961 to 2023 were retained for analysis [43]. The continuity and volume of this dataset make it suitable for studying extreme rainfall events and for evaluating the performance of different statistical models in simulating real-world precipitation data. To explore the dataset's statistical properties, with an emphasis on extreme values, two key visualizations are presented in Figure 4. The left panel shows the full scatter plot of the daily rainfall time series, while the right panel presents the MRL plot used to support threshold selection. The threshold of 72 mm indicated in the MRL plot was selected according to the methodology outlined by Coles et al. [4], as discussed in detail in Section 2.7.

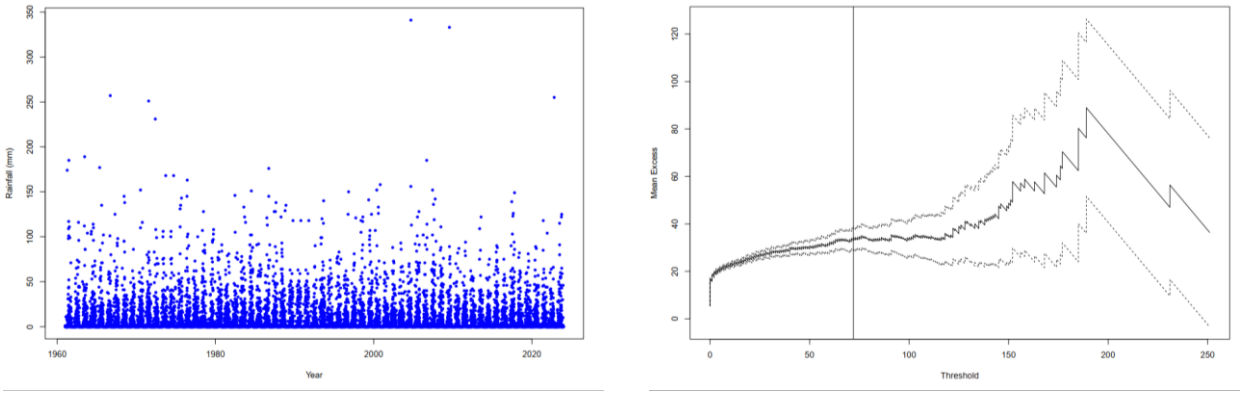

Figure 4. Visualization of the Scatter plot (left panel) and MRL plot (right panel) based on daily rainfall data from Dhaka, Bangladesh. The solid line in the MRL represents the threshold selected by following the methodology of Coles et al. [4]

2.6.4 Singapore daily rainfall dataset

The Singapore rainfall dataset utilized in this study comprises daily precipitation records collected from 2009 to 2017. The data were sourced from the National Environment Agency (NEA) of Singapore and are publicly available via the national open data portal, data.gov.sg. The historical weather records represent officially monitored rainfall amounts and are reported in mm [44].

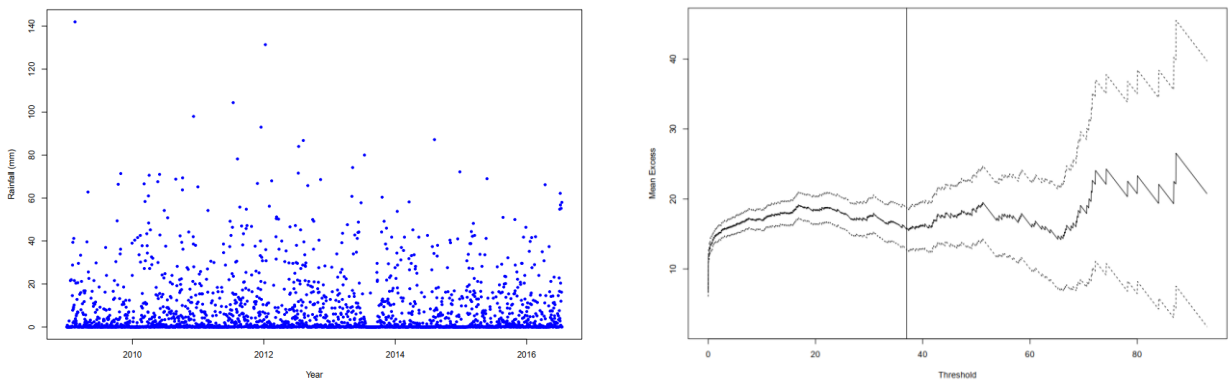

Figure 5. Visualization of the Scatter plot (left panel) and MRL plot (right panel) based on daily rainfall data from Singapore. The solid line in the MRL represents the threshold selected by following the methodology of Coles et al. [4]

For this study, daily rainfall values were extracted and pre-processed to retain only valid entries with measurable rainfall. The dataset, although covering a relatively short time span compared to other regions, provides reliable, high-resolution daily measurements, making it suitable for short-term extreme-value modelling and rainfall distribution analysis specific to equatorial urban settings such as Singapore. To understand the dataset's characteristics in the context of rainfall extremes, Figure 5 presents a time-series scatter plot alongside an MRL plot. The threshold of 37 mm selected in the MRL plot presented by the solid line in Figure 5 (right panel) was determined in accordance with the methodology introduced by Coles et al. [4], as detailed in Section 2.7.

2.6.5 US (Seattle) daily rainfall dataset

The US rainfall dataset used in this study comprises daily precipitation records from Seattle, Washington. This dataset is part of a broader historical climate compilation covering 210 cities across the United States, developed by researchers at Carnegie Mellon University. The original data were sourced from the Global Historical Climatology Network - Daily (GHCN-D) and retrieved through the Applied Climate Information System (ACIS), with the Seattle records assembled using the "ThreadEx" methodology. This approach ensures a coherent and continuous climate record by integrating data from multiple nearby stations [44]. For our research, only daily precipitation data from Seattle were used, covering the period from 1894 to 2023. Each observation represents daily rainfall totals measured in inches, which were subsequently converted into mm to maintain consistency with other datasets in the study. Furthermore, missing entries were systematically removed during pre-processing to ensure the reliability of subsequent statistical modelling. To gain insights into the statistical structure of the Seattle dataset, particularly in the context of extreme value modelling, Figure 6 displays both a scatter plot and an MRL plot. The MRL plot in the right panel illustrates the threshold of 31 mm for extremes, guided by the framework of Coles et al. [4], as detailed in Section 2.7.

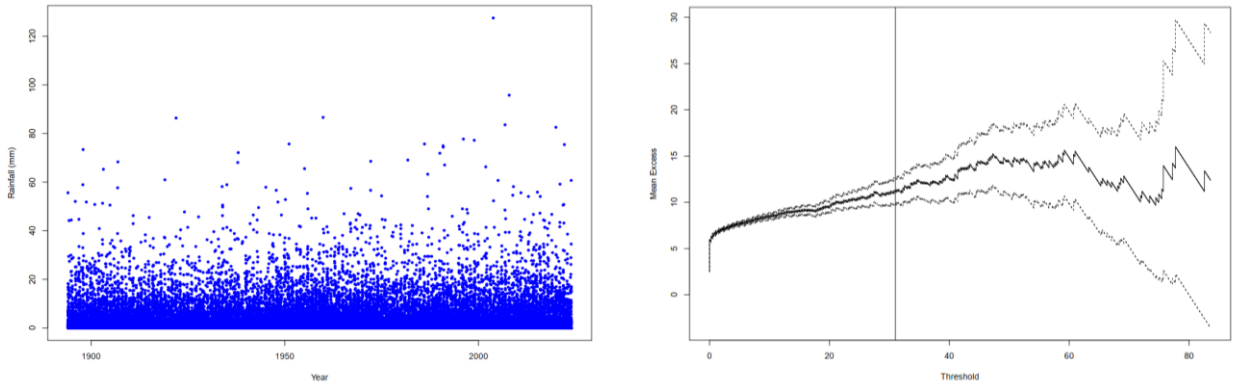

Figure 6. Visualization of the Scatter plot (left panel) and MRL plot (right panel) based on daily rainfall data from Seattle, United States. The solid line in the MRL represents the threshold selected by following the methodology of Coles et al. [4]

3. Results and Discussion

Before evaluating the performance of threshold selection methodologies, it is essential to understand the structure of the datasets utilized in this study. We analysed daily rainfall records from five distinct geographic regions: Southwest England, New Zealand (Auckland), Bangladesh (Dhaka), Singapore, and the US (Seattle). These datasets span a diverse range of climatic and geographical settings, offering a rigorous test bed for assessing the adaptability and robustness of extreme value modelling approaches. A notable feature of daily rainfall data is the frequent occurrence of zero rainfall days, reflecting dry periods. These zero values can influence the dataset's statistical characteristics and have been discussed in the context of extreme-value modelling in the literature [46-47]. While these works focus on specialized modelling frameworks for zero-inflated data, they collectively highlight the importance of understanding how structural zeros may interact with the tails of the distribution in certain modelling contexts.

Table 1. Prevalence of zero rainfall days in the selected datasets

Region	Percentage of Zero
Southwest England	47.025%
New Zealand	50.892%
Bangladesh	67.371%
Singapore	44.843%
US (Seattle)	58.027%

Table 1 presents the proportion of zero rainfall observations in each dataset. The prevalence of zeros varies across regions, from 44.843% in Singapore to 67.371% in Dhaka, underscoring the climatic heterogeneity and rainfall variability across sites. This structural difference is a critical factor in analysing extreme rainfall behaviour, as it shapes the overall distributional form and influences the density of threshold exceedances. It is important to note that in this study, we retain all zero values in the datasets without any modification. We aim to evaluate how threshold selection techniques perform in realistic settings where zero rainfall days naturally coexist with extreme precipitation events. The variation in zero prevalence across datasets, as presented in Table 1, thus provides a valuable opportunity to test the durability and flexibility of each method under diverse data conditions.

Table 2 summarizes the threshold values (u) selected by four different methods, those proposed by Alif et al. [19], Solari et al. [11], Thompson et al. [12], and the classical graphical method of Coles et al. [4], alongside the corresponding parameter estimates for the GPD: the scale parameter ($\hat{\sigma}$) and the shape parameter ($\hat{\xi}$). These results are reported for each

of the five rainfall datasets, providing a comparative view of how each method responds to different data characteristics as presented in Table 2. An evident variation in the selected thresholds across all datasets is evident in Table 2. Notably, the method proposed by Solari et al. [11] consistently yields the highest threshold values, whereas Thompson et al.'s [12] approach tends to yield the lowest. This contrast is particularly striking in the Bangladesh dataset, where Solari et al. [11] recommend a threshold of 114 mm, nearly double the 56 mm suggested by Thompson et al. [12]. This divergence in threshold selection also leads to considerable differences in the estimated scale and shape parameters, with $\hat{\sigma}$ ranging from 25.064 to 33.855 and $\hat{\xi}$ varying between 0.072 and 0.264, as detailed in Table 2. In terms of $\hat{\xi}$, which governs the tail behavior and is thus crucial for estimating return levels, several interesting insights emerge in Table 2. For the New Zealand dataset, Solari et al.'s [11] method yields the highest shape estimate (ξ = 0.390), suggesting a heavier tail and a greater propensity for extreme rainfall events. In contrast, Thompson et al.'s [12] threshold yields a significantly lower $\hat{\xi}$ of 0.115, implying a more moderate tail behavior. This divergence underscores the influence of threshold selection on the interpretation of extreme value characteristics.

Table 2. Comparison of the four methods in terms of determined threshold (u) and estimated scale (σ), and shape (ξ) for five daily rainfall datasets

Dataset	Method	u	$\hat{\sigma}$	$\hat{\xi}$
Southwest England	AD-L	33.668	8.167	0.189
	Solari et al. [11]	32.300	7.878	0.189
	Thompson et al. [12]	23.100	8.685	0.015
	Coles et al. [4]	30.000	7.299	0.197
New Zealand	AD-L	53.070	12.061	0.209
	Solari et al. [11]	62.000	9.789	0.390
	Thompson et al. [12]	26.200	13.061	0.115
	Coles et al. [4]	54.000	11.317	0.250
Bangladesh	AD-L	82.523	30.253	0.089
	Solari et al. [11]	114.000	25.064	0.264
	Thompson et al. [12]	56.000	29.470	0.072
	Coles et al. [4]	72.000	33.855	0.010
Singapore	AD-L	43.092	16.888	0.038
	Solari et al. [11]	72.200	24.332	-0.010
	Thompson et al. [12]	35.816	14.804	0.082
	Coles et al. [4]	37.000	13.958	0.121
US (Seattle)	AD-L	44.913	14.049	0.015
	Solari et al. [11]	71.882	7.326	0.316
	Thompson et al. [12]	21.082	9.347	0.080
	Coles et al. [4]	31.000	9.822	0.133

The Singapore dataset also presents a noteworthy case in Table 2, where Solari et al.'s [11] method yields a slightly negative shape parameter ($\xi = -0.010$). While the magnitude is small, a negative shape parameter suggests the presence of an upper bound on extreme rainfall, which corresponds well with the generally moderate extremes observed in tropical urban areas like Singapore. In contrast, the methods by Coles et al. [4] and Alif et al. [19] yield small but positive values for $\hat{\xi}$, highlighting the sensitivity of the shape parameter to the selected threshold. For the Southwest England and US (Seattle) datasets, the $\hat{\xi}$ across the methods show relatively consistent values, typically ranging from 0.01 to 0.20, as shown in Table 2. The estimates are reasonably aligned, suggesting that for these datasets, the extreme value behavior is less volatile to changes in the threshold selection method, though even here, Thompson et al.'s [12] lower thresholds still lead to the smallest $\hat{\xi}$ values. From a methodological standpoint, Coles et al.'s [4] graphical method tends to yield middle-ground thresholds, producing $\hat{\xi}$ values that are generally close to those from the AD-L method. For example, in Table 2, the US (Seattle) dataset, the shape parameter from Coles et al. [4] is 0.133, compared to 0.015 from AD-L, and 0.080 from Thompson et al. [12]. These slight variations are expected due to the proximity of the threshold values among the methods in this case. It is also important to note the behaviours of the $\hat{\sigma}$ in Table 2. Larger thresholds generally correspond to larger scale values, reflecting the fact that more extreme events are being modelled. However, this pattern is not always linear or consistent. For instance, in the New Zealand dataset, although Solari et al. selected a higher threshold than Alif et al. [19], the corresponding scale parameter is slightly lower (9.789 vs. 12.061), suggesting nuances in dataset-specific variability. Overall, Table 2 illustrates how sensitive GPD parameter estimation is to the choice of threshold. The findings support the view that a single, universal thresholding strategy may not be appropriate, especially for zero-inflated datasets like those used in this study. The variability in the $\hat{\xi}$ has direct consequences for return level estimation and risk quantification, which are explored in the following sections.

A closer look at Table 3 further emphasizes this point, particularly when comparing the thresholds and exceedances selected by the different methods. One of the most striking contrasts arises from Thompson et al.'s method, which includes a constraint requiring at least 100 exceedances above the selected u. This built-in condition systematically pushes the threshold lower than those identified by the other approaches. For instance, in the Southwest England dataset, Thompson et al.'s [12] method yields a u of 23.1 mm with 349 exceedances, while the AD-L method selects a higher threshold of 33.668 mm with 93 exceedances. Although a larger number of exceedances might initially seem advantageous, this does not necessarily improve the accuracy of parameter estimates. A lower threshold can include moderate, non-extreme observations, diluting the tail characteristics that the GPD is intended to capture. This can lead to biased parameter estimates and an underestimation of high return levels, ultimately compromising the reliability of risk assessments. Liang et al. [8] highlight this issue, noting that including non-extreme data can obscure the true extremal behaviours of the distribution. Conversely, selecting a threshold that is too high limits the sample to only the most extreme observations. While this helps better represent the tail, it comes at the cost of fewer data points, as observed in the case of Solari et al. in Table 3, increasing the variance and uncertainty of the estimates, as vividly evident in Table 4. This classic bias–variance trade-off is a core challenge in effectively implementing the POT method.

Table 3. A comparison of the four methods highlights the selected threshold values (u) and the number of exceedances identified, providing an overall perspective on how each method isolates the extreme value portion relative to the net dataset size

Dataset	Size	Missing Values	Net Size	Method	u	Exceedances
South West England	17531	0	17531	AD-L	33.668	93
				Solari et al. [11]	32.300	111
				Thompson et al. [12]	23.100	349
				Coles et al. [4]	30.000	152
New Zealand	21915	0	21915	AD-L	53.070	81
				Solari et al. [11]	62.000	44
				Thompson et al. [12]	26.200	438
				Coles et al. [4]	54.000	77
Bangladesh	23010	0	23010	AD-L	82.523	194
				Solari et al. [11]	114.000	75
				Thompson et al. [12]	56.000	454
				Coles et al. [4]	72.000	257
Singapore	2859	105	2754	AD-L	43.092	82
				Solari et al. [11]	72.200	11
				Thompson et al. [12]	35.816	140
				Coles et al. [4]	37.000	132
US (Seattle)	47481	17	47464	AD-L	44.913	83
				Solari et al. [11]	71.882	15
				Thompson et al. [12]	21.082	915
				Coles et al. [4]	31.000	320

Against this backdrop, the AD-L method's performance stands out. Across all datasets in Table 3, it consistently identifies thresholds that maintain a balance, high enough to focus on genuine extremes, yet not so high as to leave too few exceedances for reliable inference. This careful calibration avoids the pitfalls of both under- and over-thresholding, offering a more stable and context-sensitive approach. The logical consistency of the thresholds selected by the AD-L method and the corresponding exceedance counts reemphasizes the strength of its automated framework. This method adapts well to varying data characteristics without relying on rigid rules or arbitrary cutoffs. To further substantiate these observations, we next evaluate the GOF of each method through model compatibility, as summarized in Table 4. To reinforce the effectiveness of the threshold selection procedures, we assess model compatibility using GOF tests, including the KS, AD, and CVM. Table 4 presents the corresponding p-values for each method across all datasets. As evident from Table 4, the AD-L method consistently achieves the highest or among the highest p-values across all GOF tests and datasets, confirming strong compatibility between the fitted GPD and the empirical exceedances. Notably, Solari et al. [11] also show competitive performance, often matching or slightly outperforming the AD-L method in terms of GOF test p-values. For instance, in the New Zealand dataset, both the AD-L method and Solari et al. [41] yield exceptionally high p-values, with Solari et al. [11] reaching near-perfect fits (KS = 0.999, AD = 0.996, CVM = 0.998). In contrast, despite Thompson et al. [12] method yielding the highest number of exceedances in each dataset, due to its constraint of a minimum of 100 exceedances, its p-values are consistently lower than those of the AD-L method and Solari et al. [11], and in some cases notably poor. For example, in the US (Seattle) dataset, Thompson et al. [46] show a

KS p-value of just 0.201 and an AD p-value of 0.209, suggesting weak model fit. This pattern reinforces the earlier concern that a lower threshold leads to the inclusion of moderate, non-extreme values, which can bias parameter estimates and reduce the compatibility of the fitted model with the observed tail behaviours. Similarly, Coles et al. [4], which relies on the traditional graphical MRL plot method, performs moderately across datasets, often with p-values trailing those of the AD-L method and Solari et al. [11]. Despite sometimes achieving a better fit than Thompson et al. [12], Coles et al. [4] still fall short of the consistency shown by the automated methods, particularly in capturing the underlying extremal behaviours.

Table 4. A comparative analysis of the p-values from the KS, AD, and CVM GOF tests across the four methods

Dataset	Method	p-values		
		KS	AD	CVM
Southwest England	AD-L	0.966	0.989	0.981
	Solari et al. [11]	0.958	0.963	0.983
	Thompson et al. [12]	0.450	0.234	0.345
	Coles et al. [4]	0.850	0.862	0.953
New Zealand	AD-L	0.946	0.900	0.876
	Solari et al. [11]	0.999	0.996	0.998
	Thompson et al. [12]	0.513	0.742	0.755
	Coles et al. [4]	0.869	0.957	0.907
Bangladesh	AD-L	0.961	0.904	0.802
	Solari et al. [11]	0.985	0.914	0.963
	Thompson et al. [12]	0.300	0.357	0.694
	Coles et al. [4]	0.862	0.723	0.874
Singapore	AD-L	0.903	0.946	0.881
	Solari et al. [11]	0.983	0.984	0.979
	Thompson et al. [12]	0.950	0.855	0.840
	Coles et al. [4]	0.920	0.903	0.883
US (Seattle)	AD-L	0.978	0.979	0.973
	Solari et al. [11]	0.902	0.932	0.876
	Thompson et al. [12]	0.201	0.209	0.528
	Coles et al. [4]	0.677	0.762	0.882

3.2.1 Threshold uncertainties

While Solari et al.'s [41] method demonstrates very strong model compatibility, this comes at a cost. As shown in Table 3, Solari et al. [11] often select a much higher threshold, resulting in a very small number of exceedances. This approach ensures that only the most extreme values are modelled, naturally leading to a good fit, but at the cost of greater threshold uncertainty, as shown in Table 5. With fewer exceedances, the resulting confidence intervals become much wider due to higher variance. An examination of Table 5 reveals notable variations in the uncertainty associated with threshold selection across the four methods. The most prominent feature is the considerably wider 95% CIs produced by the Solari et al. [11] method. In several datasets, such as Bangladesh and Singapore, the CI widths exceed 150 and 70, respectively. This reflects substantial uncertainty, likely arising from the method's conservative selection of higher thresholds. While these thresholds may better isolate extreme events, the wide confidence bounds indicate limited precision, potentially complicating model interpretation and inference. In contrast, Thompson et al.'s [12] method consistently yields narrow CIs that compete with the AD-L method, in some cases producing the narrowest CIs with widths as small as 3.01 mm (US) and 5.00 mm (Bangladesh). These tightly bound intervals suggest high precision in threshold estimation. However, this should not be taken at face value as a strength; as noted earlier, the Thompson et al. [12] method tends to select substantially lower thresholds, which may inadequately represent the distribution's tail behaviour. This undermines the utility of its narrower CI, since the underlying threshold choice remains questionable for reliably capturing extreme values. AD-L method, meanwhile, offers balanced performance, balancing extremely competitive model compatibility with reasonably narrow CIs while maintaining logical and defensible threshold values across all datasets.

It is also important to note that Coles et al.'s [4] method, due to its graphical and subjective nature, does not lend itself to quantitative uncertainty estimation. As Solari et al. [11] highlight, this subjectivity limits the ability to rigorously assess the confidence or robustness of thresholds derived through visual inspection. In summary, Table 5 reinforces the reliability of the AD-L method, which provides well-balanced uncertainty measures without sacrificing model performance. While Solari et al.'s [11] method excels at capturing extreme values by setting higher thresholds, it does so at the cost of greater uncertainty. Thompson et al. [12], though seemingly precise, raise questions about the validity of their threshold estimates. These insights further validate the AD-L method as a strong candidate for robust, consistent threshold selection across diverse rainfall datasets.

Table 5. Comparison of the four methods in terms of threshold uncertainties across five daily rainfall datasets. In this table, u is the determined threshold value

Dataset	Method	u	CI	Width of CI
South West England	AD-L	33.668	[23.777, 35.037]	11.26
	Solari et al. [11]	32.300	[24.700, 59.395]	34.695
	Thompson et al. [12]	23.100	[12.180, 24.100]	11.92
New Zealand	AD-L	53.070	[12.471, 54.544]	42.073
	Solari et al. [11]	62.000	[26.902, 88.800]	61.898
	Thompson et al. [12]	26.200	[5.260, 27.340]	22.080
Bangladesh	AD-L	82.523	[56.807, 115.724]	58.917
	Solari et al. [11]	114.000	[72.025, 231.000]	158.975
	Thompson et al. [12]	56.000	[54.000, 59.000]	5.000
Singapore	AD-L	43.092	[13.059, 43.344]	30.285
	Solari et al. [11]	72.200	[22.420, 93.000]	70.580
	Thompson et al. [12]	35.816	[17.600, 42.800]	25.200
US (Seattle)	AD-L	44.913	[22.523, 45.382]	22.859
	Solari et al. [11]	71.882	[32.766, 86.614]	53.848
	Thompson et al. [12]	21.082	[18.510, 21.520]	3.010

3.2.2 Return level uncertainties

In this section, we examine the uncertainty associated with estimated return levels for each method across five distinct rainfall datasets: Southwest England, New Zealand, Bangladesh, Singapore, and US-Seattle. The analysis covers 10-, 50, and 100-year return levels and includes 95% bootstrap CIs and their respective widths. One thing to note is that, due to the subjective nature of the MRL plot approach, return level uncertainties were instead computed using the delta method, as outlined by Oehlert [36]. Now, if we observe the Southwest England dataset in Table 6, the AD-L method and Solari et al. [11] produce comparable return levels with slightly varying CI widths, indicating robust performance. Thompson et al. [12] report consistently narrower intervals, especially at higher return periods, which may again reflect the impact of their lower threshold selection. Coles et al. [4], on the other hand, in Table 6 demonstrate the widest CI at the 100-year return period (81.722), indicative of higher variability or model uncertainty due to their subjective threshold estimation.

Table 6. Comparison of the four methods in terms of return level estimates and uncertainties for the daily rainfall dataset of Southwest England at 10, 50, and 100 years return periods

Method	Return Period	Return Level	CI	Width of CI
AD-L	10	66.145	[56.977, 74.279]	17.302
	50	93.091	[70.292, 112.291]	42.000
	100	107.479	[75.395, 137.314]	61.919
Solari et al. [11]	10	65.945	[56.774, 75.081]	18.307
	50	89.516	[70.324, 110.754]	40.430
	100	101.315	[76.404, 133.638]	57.234
Thompson et al. [12]	10	64.106	[56.450, 71.956]	15.506
	50	83.484	[68.248, 100.847]	35.599
	100	92.702	[73.181, 116.213]	43.032
Coles et al. [4]	10	65.961	[55.667, 76.254]	20.587
	50	92.336	[64.167, 120.506]	56.339
	100	106.342	[65.481, 147.203]	81.722

For the New Zealand dataset in Table 7, all four methods yield fairly similar return levels, but Thompson et al. [12] report the highest returns across all periods. While this could indicate better tail modelling, the associated CIs are quite wide, especially for the 100-year level (110.185). Coles et al. [8] again show the widest CI for the 100-year return level (135.705) in Table 7, which may reflect its subjectivity and reliance on visual thresholding. The AD-L method shows competitive return levels with narrower CIs, supporting its reliability.

In Table 8, Thompson et al. [12] and the AD-L method yield similar return level estimates with moderate Confidence Intervals (CIs) for Bangladesh. In contrast, Solari et al. [11] yield slightly higher 100-year return levels and significantly larger uncertainty (CI width = 216.799), reinforcing its tendency toward conservative but uncertain estimates. Coles et al. [4] produce the narrowest CIs in Table 8 across all return levels, yet their return levels are consistently lower than those

from the other methods, possibly reflecting an underestimation from the graphical approach. For the smallest dataset (Singapore), all methods estimate similar return levels, as shown in Table 9; however, CI widths vary notably.

Table 7. Comparison of the four methods in terms of return level estimates and uncertainties for the daily rainfall dataset of Auckland, New Zealand, at 10, 50, and 100 years return periods

Method	Return Period	Return Level	CI	Width of CI
AD-L	10	94.808	[82.716, 111.814]	29.098
	50	134.622	[98.806, 171.816]	73.010
	100	156.361	[105.461, 210.720]	105.259
Solari et al. [11]	10	93.806	[76.513, 124.271]	47.758
	50	136.711	[98.049, 169.419]	71.370
	100	159.988	[108.441, 209.332]	100.891
Thompson et al. [12]	10	102.540	[86.201, 119.980]	33.779
	50	148.418	[111.038, 191.474]	80.436
	100	173.269	[121.394, 231.579]	110.185
Coles et al. [4]	10	94.411	[80.616, 108.206]	27.590
	50	136.852	[93.296, 180.408]	87.112
	100	161.093	[93.241, 228.946]	135.705

Thompson et al. [12] and Coles et al. [4] produce the widest CIs at the 100-year level, over 220 and 211, respectively, hinting at higher uncertainty in tail estimates. AD-L method and Solari et al. [41] offer slightly tighter bounds while maintaining plausible return levels as shown in Table 9, suggesting balanced modelling. The largest dataset (US (Seattle)) in our study exhibits the lowest return levels and CI widths overall, as shown in Table 10. Thompson et al. [12] deliver the narrowest intervals (as low as 11.506), consistent with its lower threshold strategy. However, the narrowness of CIs does not necessarily imply accuracy, as lower thresholds may not truly reflect tail behaviours. AD-L method and Coles et al. [4] perform competitively, with the AD-L method maintaining a strong balance between precision and realism in extreme value modelling. Also, as shown in Table 10, the return level estimates are relatively similar across the four methods at all given return periods.

Table 8. Comparison of the four methods in terms of return level estimates and uncertainties for the daily rainfall dataset of Dhaka, Bangladesh, at 10, 50, and 100 years return periods

Method	Return Period	Return Level	CI	Width of CI
AD-L	10	203.733	[177.031, 234.054]	57.023
	50	274.716	[216.874, 371.361]	154.487
	100	308.556	[232.076, 459.578]	227.502
Solari et al. [11]	10	200.718	[161.962, 243.265]	81.303
	50	296.795	[217.991, 368.375]	150.384
	100	349.862	[242.201, 459.000]	216.799
Thompson et al. [12]	10	206.010	[180.865, 237.063]	56.198
	50	280.106	[228.406, 349.782]	121.376
	100	315.909	[249.223, 412.045]	162.822
Coles et al. [4]	10	199.910	[176.410, 223.410]	47.000
	50	256.914	[206.338, 307.490]	101.152
	100	281.748	[214.441, 349.055]	134.614

A clear pattern emerging from Tables 6 to 10 is that datasets with fewer observations, such as the Singapore daily rainfall dataset (Table 9), tend to produce noticeably wider 95% CIs. In contrast, larger datasets, such as the US (Seattle) daily rainfall dataset (Table 10), yield much narrower CIs. Across all datasets, the AD-L method demonstrates consistent and stable performance. It provides reasonably narrow CIs across all datasets and return periods while maintaining reliable, defendable return-level estimates. This suggests a strong ability to generalize across diverse climatic patterns without excessive uncertainty. Solari et al. [11] regularly report wide confidence intervals, particularly for the 100-year return levels. This supports the idea that while it isolates extreme events well, it does so at the cost of increased uncertainty, possibly stemming from higher thresholds and fewer exceedances. Thompson et al. [12] often yield the narrowest intervals but frequently report higher return levels due to their lower threshold selection. The artificially reduced variance from using more exceedances may give a false sense of precision, potentially undermining the trustworthiness of its return level projections. Coles et al. [4] show the most variable performance. While it sometimes reports narrow CIs, particularly

in the Bangladesh dataset, it also exhibits very wide CIs in other datasets. Its subjectivity in threshold selection may contribute to these inconsistencies, making it less reliable than automated methods in comparative studies.

Table 9. Comparison of the four methods in terms of return level estimates and uncertainties for the daily rainfall dataset of Singapore at 10, 50, and 100 years return periods

Method	Return Period	Return Level	CI	Width of CI
AD-L	10	129.845	[97.012, 170.929]	73.917
	50	163.401	[107.081, 264.565]	157.484
	100	178.503	[110.564, 317.778]	207.214
Solari et al. [11]	10	130.561	[95.669, 167.474]	71.805
	50	174.646	[105.825, 256.748]	150.923
	100	198.522	[107.989, 317.487]	209.498
Thompson et al. [12]	10	130.186	[96.348, 176.428]	80.080
	50	168.587	[105.073, 273.722]	168.649
	100	187.431	[108.529, 331.864]	223.335
Coles et al. [4]	10	137.153	[96.969, 177.337]	80.368
	50	183.487	[102.024, 264.950]	162.926
	100	206.395	[100.510, 312.280]	211.770

Table 10. Comparison of the four methods in terms of return level estimates and uncertainties for the daily rainfall dataset of Seattle, US, at 10, 50, and 100 years return periods

Method	Return Period	Return Level	CI	Width of CI
AD-L	10	71.324	[64.142, 76.662]	12.520
	50	94.842	[83.069, 110.434]	27.365
	100	105.145	[89.627, 130.441]	40.814
Solari et al. [11]	10	70.580	[61.025, 79.614]	18.589
	50	89.878	[78.087, 107.801]	29.714
	100	98.804	[82.333, 123.861]	41.528
Thompson et al. [12]	10	69.386	[63.735, 75.241]	11.506
	50	94.083	[82.151, 107.770]	25.619
	100	106.123	[90.445, 124.910]	34.465
Coles et al. [4]	10	70.234	[63.761, 76.707]	12.946
	50	97.226	[80.403, 114.049]	33.646
	100	110.753	[86.824, 134.683]	47.859

4. Conclusions

In this study, we conducted a comprehensive comparison of four threshold selection methods for the GPD applied to diverse rainfall datasets from Southwest England, New Zealand, Bangladesh, Singapore, and the US (Seattle). Our investigation reveals that the choice of threshold not only substantially affects GPD parameter estimation but also has significant implications for return-level predictions. The results demonstrate that methods based on automated statistical criteria, particularly the AD-L method, tend to achieve a favourable balance between bias and variance. The AD-L method, along with a simple, computationally favourable method, consistently produces thresholds that are both robust and defensible, yielding return level estimates with moderate uncertainty and strong model compatibility across datasets. In contrast, the Solari approach, while effective at isolating the most extreme events, often yields wider confidence intervals due to its conservative threshold selection. The Thompson method, which systematically selects lower thresholds to secure a high number of exceedances, delivers narrow confidence intervals but may risk underrepresenting tail behaviours. The classical graphical MRL plot method of Coles, though widely used and intuitive, lacks the quantitative rigor required to assess threshold uncertainty objectively.

Overall, our findings clearly demonstrate that the threshold selection procedure proposed by the AD-L method generally exhibits a slight edge across the diverse rainfall datasets examined. AD-L method consistently delivers logically interpretable thresholds that strike an effective balance between bias and variance, yielding an optimal number of exceedances. It is particularly effective in managing the high prevalence of zero rainfall events, a common challenge in realistic datasets, which further enhances the robustness of extreme value estimates. It is important to note, however, that we do not claim that the alternative methods are unreliable or untrustworthy; in some instances, methods other than the AD-L method even produced more precise outcomes or narrower uncertainties. Rather, given all the observations and results presented in our study, the AD-L method appears to have a slight edge due to its consistency, simplicity, and

computational affordability. These insights provide a robust framework for both practitioners and researchers engaged in hydrological risk assessment and flood management. Future research should explore hybrid or adaptive methods to further refine threshold selection and account for evolving climate patterns, thereby enhancing the precision of return-level forecasting in extreme-event modelling.

Acknowledgement
The authors gratefully acknowledge Universiti Putra Malaysia for its support of this work. The authors also express their gratitude to the anonymous referees for their valuable comments and suggestions.

Funding
This study was supported by the Special Graduate Research Scheme (SGRA) provided by Universiti Putra Malaysia through the Putra Grant GP/2023/9753100.

Declaration of Competing Interest
The authors declare no conflict of interest.

CRediT Authorship Contribution Statement
Farabe Khan Alif (Conceptualisation; Methodology; Software; Data curation; Formal analysis; Visualisation; Writing – original draft; Writing – review & editing; Investigation; Validation; Project administration)
Norhaslinda Ali (Supervision; Conceptualisation; Funding acquisition; Methodology; Writing – review & editing)

Availability of the Data and Materials
The data used to support the findings of this study are included within the article.

Ethical Declaration
No artificial intelligence tools were used in the preparation of this manuscript. All content was developed manually by the authors. This study did not involve human participants or animals. Ethical approval was therefore not required.

Generative Artificial Intelligence Declarations
The authors claim that artificially intelligent-assisted technologies, such as generative AI, were not used to generate content, ideas, or theories. We have just utilised AI to enhance readability and refine the language. This was used with extreme human control and oversight. The authors take full responsibility for reviewing and approving the content.

References
[1] Else H. Climate change implicated in Germany's deadly floods. Nature. 2021 Jul 20. Available from: https://www.nature.com/articles/d41586-021-01968-3.
[2] Taye MT, Ntegeka V, Ogiramoi NP, Willems P. Assessment of climate change impact on hydrological extremes in two source regions of the Nile River Basin. Hydrology and Earth System Sciences. 2011;15(1):209–222.
[3] Zhou XH, Zhou A, Shen SL. How to mitigate the impact of climate change on modern cities: lessons from extreme rainfall. Smart Construction and Sustainable Cities. 2023;1(1):7.
[4] Coles S. An introduction to statistical modeling of extreme values. London: Springer; 2001.
[5] Davison AC, Smith RL. Models for exceedances over high thresholds. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1990;52(3):393–425.
[6] Pickands J III. Statistical inference using extreme order statistics. The Annals of Statistics. 1975;3(1):119–131.
[7] 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. arXiv [Preprint]. 2016. Available from: https://arxiv.org/abs/1604.02024.
[8] 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. 2019;144:22–32.
[9] Liu B, Ananda MMA. A new insight into reliability data modeling with an exponentiated composite exponential-Pareto model. Applied Sciences. 2023;13(1):645.
[10] Dupuis DJ. Exceedances over high thresholds: A guide to threshold selection. Extremes. 1999;1:251–261.
[11] Solari S, Eguen M, Polo MJ, Losada MA. Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. Water Resources Research. 2017;53(4):2833–2849.
[12] Thompson P, Cai Y, Reeve D, Stander J. Automated threshold selection methods for extreme wave analysis. Coastal Engineering. 2009;56(10):1013–1021.
[13] Gaigall D, Gerstenberg J. Cramer-von-Mises tests for the distribution of the excess over a confidence level. Journal of Nonparametric Statistics. 2023;35(3):529–561.
[14] Murphy C, Tawn JA, Varty Z. Automated threshold selection and associated inference uncertainty for univariate extremes. Technometrics. 2024;66(3):363-375.
[15] Minguez R. Automatic threshold selection for generalized Pareto and Pareto–Poisson distributions in rainfall analysis: A case study using the NOAA NCDC daily rainfall database. Atmosphere. 2025;16(1):78.

[16] Alaswed H. Graphical diagnostics for threshold selection in fitting the generalized Pareto distribution. Journal of Pure & Applied Sciences. 2024;23(1):90–95.
[17] Curceac S, Atkinson PM, Milne A, Wu L, Harris P. An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales. Journal of Hydrology. 2020;585:124845.
[18] Hambuckers J, Kratz M, Usseglio-Carleve A. Efficient estimation in extreme value regression models of hedge fund tail risks. arXiv [Preprint]. 2023. Available from: https://arxiv.org/abs/2304.06950.
[19] Alif FK, Ali N, Safari MAM. An assessment on threshold selection for the generalized Pareto distribution using goodness of fit. Malaysian Journal of Mathematical Sciences. 2025;19(3):871–899.
[20] Embrechts P, Kluppelberg C, Mikosch T. Modelling extremal events for insurance and finance. British Actuarial Journal. 1999;5(2):465–465.
[21] Hosking JRM. On the characterization of distributions by their L-moments. Journal of Statistical Planning and Inference. 2006;136(1):193–198.
[22] Asquith WH. Distributional analysis with L-moment statistics using the R environment for statistical computing. CreateSpace Scotts Valley, CA, USA; 2011.
[23] Simkova T, Picek J. A comparison of L-, LQ-, TL-moment, and maximum likelihood high quantile estimates of the GPD and GEV distribution. Communications in Statistics-Simulation and Computation. 2017;46(8):5991–6010.
[24] Hosking JRM. L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1990;52(1):105–124.
[25] van Staden PJ, Loots MT. Method of L-moment estimation for the generalized lambda distribution. In: Proceedings of the Third Annual ASEARC Conference. 2009 Dec 1; Newcastle, Australia. pp. 7–8.
[26] Thompson P, Cai Y, Reeve D, Stander J. Automated threshold selection methods for extreme wave analysis. Coastal Engineering. 2009;56(10):1013–1021.
[27] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. 2016;31(4):337–350.
[28] Chu J, Dickin O, Nadarajah S. A review of goodness of fit tests for Pareto distributions. Journal of Computational and Applied Mathematics. 2019;361:13–41.
[29] Stephens MA. Goodness-of-fit techniques. In: D'Agostino RB, Stephens MA Eds. Goodness-of-Fit Techniques. New York: Routledge, 2017.
[30] Luceno A. Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. Computational Statistics & Data Analysis. 2006;51(2):904–917.
[31] Stephens MA. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association. 1974;69(347):730–737.
[32] Choulakian V, Lockhart RA, Stephens MA. Cramer-von Mises statistics for discrete distributions. The Canadian Journal of Statistics. 1994;22(1):125–137.
[33] Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association. 1967;62(318):399–402.
[34] Massey FJ Jr. The Kolmogorov-Smirnov test for goodness-of-fit. Journal of the American Statistical Association. 1951;46(253):68–78.
[35] R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.
[36] Faraway J, Marsaglia G, Marsaglia J, Baddeley A, et al. goftest: Classical Goodness-of-Fit Tests for Univariate Distributions. R package version 1.2-3; 2021.
[37] Mooney CZ, Duval RD. Bootstrapping: A nonparametric approach to statistical inference. Thousand Oaks, CA: SAGE Publications; 1993.
[38] Asquith WH. Distributional analysis with L-moment statistics using the R environment for statistical computing. CreateSpace Scotts Valley, CA, USA; 2011.
[39] Sinclair CD, Spurr BD, Ahmad MI. Modified Anderson-Darling test. Communications in Statistics-Theory and Methods. 1990;19(10):3677–3686.
[40] Coles SG, Tawn JA. Modelling extremes of the areal rainfall process. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(2):329–347.
[41] Niu D, Sayed T, Fu C, Mannering F. A cross-comparison of different extreme value modeling techniques for traffic conflict-based crash risk estimation. Analytic Methods in Accident Research. 2024;44:100352.
[42] Simkova T, Picek J. A comparison of L-, LQ-, TL-moment, and maximum likelihood high quantile estimates of the GPD and GEV distribution. Communications in Statistics-Simulation and Computation. 2017;46(8):5991–6010.
[43] Zubair M, Ishtiaque Mahee MN, Reza KM, Salim MS, Ahmed N. Climate data dynamics: A high-volume real-world structured weather dataset. Data in Brief. 2024;57:111156.
[44] National Environment Agency. Historical daily weather records. Retrieved from https://data.gov.sg/datasets/d_03bb2eb67ad645d0188342fa74ad7066/view; Apr 5 2025.
[45] Lai Y, Dzombak D. Compiled historical daily temperature and precipitation data for selected 210 U.S. cities. Carnegie Mellon University [Dataset]. 2019. doi:10.1184/R1/7891151.v4
[46] Couturier DL, Victoria-Feser MP. Zero-inflated truncated generalized Pareto distribution for the analysis of radio audience data. The Annals of Applied Statistics. 2010;4(4):1824–1846.
[47] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.