Jurnal Teknologi

Full Paper

OUTLIER DETECTION IN RAINFALL DATA USING EXTREME VALUE THEORY

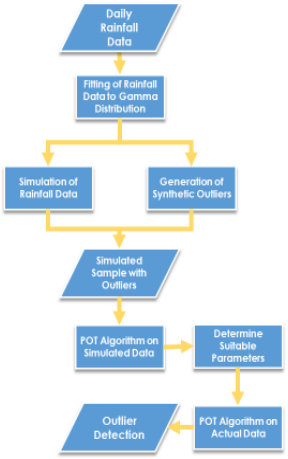
Wan Muhammad Haiqal Shah Mohd Sariff, Sayang Mohd Deni, Nor Azura Md Ghani, Ahmad Zia Ul-Saufie, Noor Fadhilah Ahmad Radi*

School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia

Article history
Received
14 May 2024
Received in revised form
13 November 2024
Accepted
20 January 2025
Published Online
22 August 2025

*Corresponding author
noorfadhilah@uitm.edu.my

Graphical abstract



Abstract

Extreme rainfall modelling has gained increased attention in recent decades due to its importance for spatial analysis and risk assessment. Similar to any statistical analysis, stochastic modelling involving extreme data is susceptible to errors due to presence of outliers. However, the precise definition of outliers and extreme events remains vague despite the extensive research on the topic. The current outlier detection method often assumes that the sample data follows a normal distribution, which is implausible for rainfall data due to its positively-skewed and heavy tail characteristics. In this study, we focus on eliminating the presence of outlier in daily rainfall series while ensuring the preservation of observed extreme events through the implementation of Extreme Value Theory. The contribution of this study is two folds; foremost, the Peaks-Over-Threshold (POT) algorithm is demonstrated for outlier detection in univariate rainfall data. Secondly, the study introduces an algorithm for generating synthetic outlier using Gamma distribution. The algorithm's performance was tested in various settings using simulated rainfall data to evaluate its effectiveness and dependability before applying it to real data. The result indicates that the algorithm successfully identified outliers without affecting the extreme daily precipitation values in the sample dataset. This finding will greatly enhance future research by improving data quality management, hence enabling more precise analysis of extreme rainfall events.

Keywords: Extreme value theory, outlier detection, rainfall series, univariate data, synthetic outlier

Abstrak

Pemodelan siri hujan ekstrem telah mendapat perhatian yang lebih dalam beberapa dekad belakangan ini kerana kepentingannya untuk analisis ruangan dan penilaian risiko. Sama seperti analisis statistik yang lain, pemodelan stokastik yang melibatkan data ekstrem terdedah kepada ralat disebabkan kehadiran data terpencil ('*outlier*'). Walau bagaimanapun, takrifan yang tepat tentang data terpencil dan peristiwa ekstrem masih samar walaupun terdapat penyelidikan yang meluas mengenai topik tersebut. Kaedah pengesanan data terpencil yang sedia ada sering mengandaikan bahawa data sampel mengikut taburan normal, yang mana adalah tidak munasabah untuk data hujan kerana cirinya yang condong positif dan ekor taburan yang berat. Kajian ini memfokuskan kepada penghapusan data terpencil yang wujud dalam siri hujan harian sekali gus memastikan pemeliharaan kejadian ekstrem yang diperhatikan melalui pelaksanaan Teori Nilai Ekstrem ('*Extreme Value Theory*'). Terdapat dua sumbangan daripada kajian ini; pertama, algoritma '*Peaks-Over-Threshold*' digunakan untuk pengesanan data terpencil dalam data hujan univariat. Keduanya, kajian ini memperkenalkan algoritma untuk menghasilkan data terpencil sintetik menggunakan taburan Gamma. Prestasi algoritma telah diuji dalam pelbagai tetapan menggunakan data taburan hujan yang disimulasi untuk menilai keberkesanan dan kebolehpercayaannya sebelum diaplikasikan pada data sebenar. Hasil menunjukkan bahawa algoritma berjaya mengenal pasti data

terpencil tanpa menjejaskan nilai hujan harian ekstrem dalam data sampel. Dapatan ini akan meningkatkan mutu kajian masa hadapan dengan memastikan kualiti data terjaga, justeru membolehkan analisis yang lebih tepat tentang peristiwa hujan ekstrem.

Kata kunci: Teori nilai ekstrem, pengesanan data terpencil, siri hujan, data univariat, data terpencil sintetik

1.0 INTRODUCTION

Global warming has been a prominent concern because of its connection to climate change. The rise in temperature and alterations in atmospheric conditions may lead to changes in rainfall patterns, variability, and distribution. Even a small adjustment in the average and spread could result in significant alterations in the likelihood of extreme weather events [1]. Many researches have been conducted on rainfall modelling for different reasons, such as drought and flood assessments [2,3], extreme event predictions [4], and water management [5]. By utilizing stochastic methods to create a rainfall model, it is feasible to capture the statistical characteristics of the observed distribution, enabling the simulation of many potential rainfall patterns. In recent decades, researchers have turned their attention to modelling extreme rainfall series for spatial analysis and risk assessment using return periods [6-9]. Modelling extreme occurrences involves statistical analysis, which is susceptible to errors caused by outliers in the dataset [10]. Liao *et al.* (2023) considers outlier detection essential before doing any statistical analysis to provide a dependable outcome and hence, precise decision-making [11]. The definition on the concept of outliers remains broad despite the numerous existing literatures, and there is no singular or unambiguous separation between outliers and extreme events. Outliers are events that deviate greatly from the population distribution or may originate from a completely distinct population distribution [12,13]. On that basis, the extreme events can be defined as observations that occurs on the lower probability region of the population distribution.

Numerous approaches to identifying outliers have been explored in the past; nevertheless, discussions regarding the preservation of extreme observations within rainfall series remain scarce. According to a prior definition provided by Kim *et al.* (2024), an outlier in the context of rainfall data is a substantially big precipitation observation that deviates from the typical pattern or represents an abrupt change in both place and time [14]. On another note, Zakaria *et al.* (2017) defines extreme rainfall events as precipitation amount observed that is more than the normal occurrence [15]. While the two prior definitions provided are somewhat similar to a certain extent, it can be highlighted that upper outliers usually deviates more significantly compared to extreme events, however there is no precise way on quantifying the said deviation. Various outlier detection techniques have been discussed such as the the Grubb's test [12,16,17], the Chi-square test [16] and the Z-score method [12], however, a common limitation shared by these methods, and most outlier detection techniques alike, is that all of the methods works under the assumption that the data is sampled from a normally distributed population [14]. Studies by Radi *et al.* (2017) and Suhaila (2023) refutes this assumption as they suggest that rainfall series often exhibit a right-skewed distribution, making them better suited for representation by positively skewed distributions like gamma rather than the symmetrical Gaussian (normal) distribution [7,18]. Liao *et al.* (2023) in their study introduced an outlier detection technique specifically tailored for gamma distributed sample, whereby the computed test statistics will be compared to a critical value which can be obtained via Monte-Carlo method, however, the concern of misclassifying an observed extreme event as outliers were not addressed [11].

A method that requires no assumption on the underlying distribution of the sample data has also been discussed before, namely the box-plot method which estimates the outlier threshold value to be 1.5 to 3 box length from the upper or lower edge of the box [12,19], however, a recent study done by Bhattacharya *et al.* (2023) shows how this method would work exceptionally well when the underlying distribution is normal but it tends severely overestimates the number of outliers that exists in the dataset as the tail heaviness of the underlying distribution increases [20]. In other words, the box-plot method is more likely to misidentify an extreme event as an outlier should it be applied to rainfall series which is known for its right-skewed tendency. This highlights another concern that, while the methods mentioned above produces reliable result in their respective studies, they do not ensure that the supposedly flagged 'outliers' are not extreme events of the population. Considering how outliers can be values that are abnormally large, it may exhibit behavior similar to extreme events, in the sense that it is usually isolated due to these events having lower probability of occurring.

As such, the Extreme Value Theory (EVT) is an ideal basis to be implemented for outlier detection process while preserving the actual extreme events observed in the sample considering how the foundation of this theory itself is based on the behavior of extreme events irrespective of their population distribution. The EVT dictates the law of extreme events and Siffer *et al.* (2017) provided an analogous explanation of how, even when the independent variables have different underlying distributions, the extreme events extracted from two different independent variables will most likely follow a similar distribution [21]. This would enable a more versatile and extensive implementation of the outlier detection method, irrespective of the distribution of the sample dataset.

The application of EVT for outlier detection has been reported previously by Gbenro (2020) and Bhattacharya *et al.* (2023) where both studies found that the respective proposed technique performs effectively regardless of whether the sample distribution is Gaussian or not, aligning with the postulated property that extreme events distribution is independent of its parent distribution. However, both of these methods work by testing whether the largest observation is an outlier, thus, the process needs to be repeated iteratively depending on the number of outliers. In this study, we are more interested in a simpler approach introduced by Siffer *et al.* (2017) where the Generalised Pareto Distribution (GPD) was implemented for outlier detection in streaming dataset with relatively large sample sizes [21]. This method is preferable for detecting possible outliers in daily rainfall series under a long period of observations as it is tailored to deal with large sample size. The use of EVT makes it capable of accommodating to the positively skewed tendency exhibited by rainfall data and most importantly, will preserve the actual extreme events that are observed in the sample data. Hence, the study aims to explore its suitability for actual implementation on univariate rainfall data.

However, in this study, we propose a slight modification whereby we introduce the empirical quantile threshold, which classifies any observation exceeding this threshold as extreme events in the algorithm, to be a non-constant parameter. As opposed to the original implementation, this study works with data in a large batch, rather than continuously streaming data. Hence, the adjustments are made to allow for a reasonable amount of observations to be fitted to the GPD and ultimately produce a more accurate estimation of outlier threshold. Additionally, the Monte-Carlo method proposed by Liao *et al.* (2023) can serve as a basis for generating synthetic outlier detection in gamma distribution [11]. This is crucial in designing our simulation study phase, where the generated outliers will act as the "ground truth" to assess and validate the performance of our proposed model. This study extends the seminal work of Liao *et al.* (2023) and Siffer *et al.* (2017) by incorporating their research findings into the analysis of outlier in rainfall data [11,21]. Drawing inspirations from past approaches, we design this study to address the difficulty of outlier detection in sample data that are generally right-skewed or exhibits heavy tail behaviour while maintaining the actual extreme events that exist in the dataset. The study focuses on how the proposed algorithm performs under diverse settings, such as sample size, empirical quantile threshold and risk variable, which allow us to evaluate the algorithm's robustness across a wide range of scenario. This work is critical because it not only broadens the relevance of existing research, but also provides practical answer for better quality control on rainfall series, benefitting subsequent studies on extreme events for a more accurate findings that can be extended to the area of agriculture, water resource management and risk mitigations.

2.0 METHODOLOGY

This study will utilize the Peaks-Over-Threshold (POT) approach for outlier detection, which was initially developed by Siffer *et al.* (2017) for analysing continuously streaming datasets that updates over time [21]. Contrarily, this study focuses on identifying outliers in a large batch of daily rainfall data collected over 40 to 50 years, thus, due to the fundamental differences in the dataset, it is important to assess the appropriateness of using the POT method. This study will be divided into two parts where the first part involves a simulation study where synthetic rainfall data will be generated from the gamma distribution. Additionally, a method inspired by Liao *et al.* (2023) will be proposed to generate outliers [11], which will then be included in the dataset as "labelled data", or data that can be explicitly labelled as outlier for the purpose of validation. We will proceed with applying the outlier detection algorithm to the real rainfall data in the second phase of the study. Figure 1 below provides a visual summary of the research workflow.

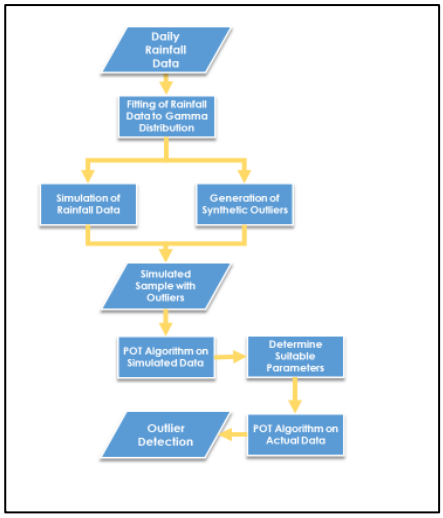


Figure 1 Research workflow

2.1 Study Area and Data

The state of Pahang covers 35 960 km^2 of land area, making it the largest state in Peninsular Malaysia with tropical rainforests occupying two thirds of the area [22]. While the state generally receives an increased amount of rainfall during the transition period of Southwest Monsoon (SWM) to Northeast Monsoon (NEM), and during NEM itself [23], it was also previously suggested that the area of Pahang (inland region) receives less precipitation as compared to other regions due to the mountain range that shelters the area during the NEM season [18].

For this study, we will be working with daily rainfall series recorded at three different stations located across the state; Batu Embun (***A*** :102.3500°E, 3.9667 °N), Kuantan (***B*** :103.2167°E, 3.7667 °N) and Temerloh (***C*** :102.3833°E, 3.4667 °N), all provided by the Malaysian Meteorological Department. Figure 2 shows the location of the selected rainfall stations in the state of Pahang. The observation period of the rainfall series is 50 years (1973 – 2022) for Kuantan and 40 years (1983 - 2022) for Batu Embun and Temerloh, and it is decided based on the longest availability of observations. Only 0.005% of missing data was observed in Kuantan which was easily replaced using the inverse distance weighing method [28] with the information from the neighboring stations while Batu Embun and Temerloh have complete observations in the whole period. The descriptive statistics of the rainfall series recorded the selected stations can be seen in Table 1. Aligning with the previous suggestion of the right skewed tendency of rainfall data, the skewness measure at all three stations shows a positive value as expected.

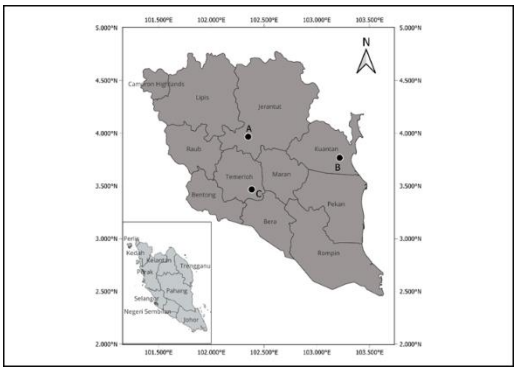


Figure 2 The location of the selected rainfall stations in Pahang, (**A**) Batu Embun, (**B**) Kuantan and (**C**) Temerloh

Temerloh station receives the least average daily rainfall of 5.2976 mm, followed by Batu Embun with 5.9379 mm and Kuantan with the highest amount at 8.1057 mm, with the coefficient of variation also following this order of increment. In particular, we would like to emphasize on the maximum daily rainfall observed at the three stations, where Batu Embun and Temerloh with records of 175.6 mm and 200.1 mm respectively, but the maximum value recorded at Kuantan can be considered to be very large. For insight, Ghani *et al.* (2016) stated that the monthly average precipitation received in the city of Kuantan is approximately 438 mm, but there is an observation in the dataset with a large value of 527.5 mm recorded in a single day. Therefore, in this study we will be implementing the proposed outlier detection algorithm on the dataset observed at all three stations to test for any outlier existence for maintaining data quality.

2.2 Simulation Study

In this phase of the study, the suitability of the Peaks-Over-Threshold method, which was developed on the fundamentals of EVT and utilises the GPD, will be examined using synthetic data simulated from the gamma distribution. A number of past literatures have suggested that the gamma distribution is adequate for modelling precipitation due to its heavy tail property that accommodates the positive skewness exhibited by rainfall data [8,9,18]. Working with simulated data also allows us to introduce labelled data for further analysis and error measure to validate the appropriateness of the POT.

2.2.1 Simulation of Rainfall Data and Outlier from Gamma Distribution

The gamma distribution is a skewed distribution commonly utilized in different applications, particularly in rainfall modelling and generating synthetic data [25]. Husak *et al.* (2008) state that this distribution is favoured because it can depict a variety of distributions using only two parameters; shape and scale [26], while Soleh *et al.* (2016) further emphasised that as the shape parameter value approaches zero, the distribution will be more positively skewed, and the distribution will achieve symmetricity as the shape parameter approaches infinity [27]. The distribution is bounded on the left at zero and positively skewed, reflecting the empirical rainfall data where negative precipitation values are impossible, and the likelihood of exceptionally heavy rainfall events is not zero. The gamma distribution is commonly used for modelling various patterns of rainfall due to its favourable characteristics. Radi *et al.* (2017) stated that a random variable X of a gamma distribution, denoted X~G(α, β)

Table 1 Descriptive statistics of daily rainfall series

Stations	Elevation (m a.s.l)	Average (mm)	Maximum (mm)	Coefficient of Variation	Skewness
Batu Embun	59.5	5.9379	175.6	186.2469	3.9492
Kuantan	15.2	8.1057	527.5	477.6170	6.8616
Temerloh	39.1	5.2976	200.1	169.2746	4.5422

with the probability distribution function (PDF) given by [9]:

$$f_X(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \quad (1)$$

for $x > 0$ and $\alpha, \beta > 0$ where α, and β are the shape and scale parameters, respectively with the mean of the distribution given as $\alpha\beta$ and variance given as $\alpha\beta^2$. The Maximum Likelihood Estimator (MLE) is a common method utilized for parameter estimation of the PDF [4]. Based on Equation (1) above, the likelihood function of the Gamma distribution can be given by:

$$\begin{aligned} L^* &= \ln \prod_{i=1}^{n} f(x_i;\alpha,\beta) \\ &= (\alpha-1)\sum_{i=1}^{n}\ln x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i \\ &\quad - n\ln\Gamma(\alpha) - n\alpha\ln\beta \end{aligned} \quad (2)$$

To assess the appropriateness of the shape and scale parameters, α and β which were estimated via MLE, we will deploy several goodness-of-fit tests which are specifically designed to measure the agreement between empirical distribution and the theoretical probability distribution [28]. Let $x_1 < x_2 < \ldots < x_n$ be the sample data with size n, sorted in increasing manner. The population is then defined by a continuous cumulative distribution function $F(x)$ and $F_0(x_i, \hat{\theta})$ be the theoretical distribution with the set of estimated parameters, $\hat{\theta}$. The test statistics will be computed based on three different empirical distribution function test statistics, namely the Kolmogorov-Smirnov (K-S) test, Cramer-von Mises (CvM) test and Anderson-Darling (A-D) test. All of these tests measure the distance between a continuous distribution function and the empirical distribution function, with the null hypothesis, H_0 stating that the empirical distribution is equal to the theoretical distribution and the alternative hypothesis, H_1 stating otherwise. Song & Singh (2010) suggests that the test statistics can be represented by [28]:

(1) Kolmogorov-Smirnov test statistics, D_n

$$\begin{aligned} D_n &= \max_{1 \le i \le n}\left(\hat{\delta}_i\right), \\ \hat{\delta}_i &= \max\left[\frac{i}{n} - F_0(x_i;\hat{\theta}), F_0(x_i;\hat{\theta}) - \frac{i-1}{n}\right] \end{aligned} \quad (3)$$

(2) Cramer-von Mises test statistics, W_n^2

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F_0(x_i;\hat{\theta}) - \frac{2i-1}{2n}\right]^2 \quad (4)$$

(3) Anderson-Darling test statistics, A_n^2

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left\{(2i-1)\ln\left[\begin{matrix} F_0(x_i;\hat{\theta}) + \\ \ln\left(1 - F_0(x_i;\hat{\theta})\right)\end{matrix}\right]\right\} \quad (5)$$

Ultimately, the POT algorithm will be implemented to suggest the existence of potential outliers in the dataset. However, working in the context of rainfall data we have no access to any "ground truth", or in other words even if an outlier exists in the dataset, the observation will not be explicitly labelled as such. Therefore, we have no mean to gauge the efficacy of the POT algorithm should we directly implement it with the rainfall data. A recent study by Liao *et al.* (2023) which primarily discusses on detecting outliers within the gamma distribution provides a solution to this limitation [11]. The idea is to first, fit the sample to the gamma distribution, estimate the shape and scale parameters by utilising the MLE [4], calculate the outlier threshold value based on the method introduced in [11], and lastly generate random values exceeding this threshold to be declared as synthetic outliers. Here, we choose the uniform distribution to produce the random values for the synthetic outliers as to align with prior definition of outliers coming from a different population distribution. We will then generate a synthetic rainfall series using the estimated parameters of the gamma distribution, and concatenate the simulated outliers into it to validate the performance of the POT algorithm. The outlier simulation process can be summarized in the following steps:

Generating Synthetic Outliers in Gamma Distribution

1. Fit the observed rainfall series to the gamma distribution and estimate the shape, α and scale, β parameters using MLE method.
2. Simulate n = 5000 gamma observations with the estimated α and β values.
3. Let $x_1 < x_2 < \ldots < x_n$ be the simulated observations, sorted increasingly and $\bar{x}$ be the sample mean. Compute the test statistics T_k, given by:
$$T_k = \frac{x_n + x_{n-1}}{\bar{x}}$$
4. Simulate 5000 instances of T_k by repeating step 2 and step 3.
5. Rank the simulated T_k instances.
6. For 95% confidence level, the critical value, Tc will be denoted as the 95th empirical quantile of the simulated T_k instances.
7. Compute the outlier threshold, T_H as follows:
$$T_H = \frac{T_C \times \bar{x}}{2}$$
8. Generate random number from the uniform distribution that exceeds T_H to be introduced as synthetic outlier for gamma distribution.

2.2.2 *POT Algorithm for Outlier Detection in Daily Rainfall Series*

Generally, past studies have demonstrated several ways of extracting and modelling of extreme events using EVT. The first method is on block maxima, in which the maximum observation will be extracted from each block, for example the maximum annual daily precipitation, or the maximum observed daily rainfall in the block of one year. Extreme events identified through this method are modeled using the GEV distribution. However, it has been noted that this

approach reduces the sample size and may result in the loss of extreme observations, particularly when multiple maxima occur within a single block [29]. The second approach, which is also the foundation of this study, is the peaks-over-threshold technique, where all observations exceeding the specified threshold will be extracted as extreme events, and these events can be quantified through GPD.

The Peaks-Over-Threshold (POT) algorithm introduced by Siffer *et al.* (2017) was designed for outlier detection in streaming dataset, in which the sequence of data will be continuously generated over time [21]. Siffer *et al.* (2017) highlighted that one of the main drawbacks in current outlier detection technique is due to the need of an a priori assumption on the sample distribution, and they overcame this limitation by implementing the EVT as the foundation of their anomaly detection approach [21]. The POT algorithm revolves mainly around the second theorem of EVT, specifically the Pickands- Balkema-de Haan theorem, as shown below [21]:

$$F_t(x) = \mathbb{P}(X - t > x \mid X > t) \sim \left(1 + \frac{\gamma x}{\sigma(t)}\right)^{-\frac{1}{\gamma}}, \; t \to \infty \qquad (6)$$

where F is the cumulative distribution function, X is a random variable and threshold, t. Equation (6) above shows that the excess over a certain quantile threshold t, also denoted as $X - t$, will likely follows the Generalized Pareto Distribution (GPD) with parameters shape, γ and scale, σ. The general idea of this algorithm is to set a small "risk" parameter, q, which is associated with the probability of observing an event where $X > z_q$ where z_q is the outlier threshold, shown as in the equation [21]:

$$\mathbb{P}(X > z_q) < q \qquad (7)$$

and the computation of z_q is based on the estimated parameter of the fitted GPD parameters $\hat{\sigma}$ and $\hat{\gamma}$ [21],

$$z_q \simeq t + \frac{\hat{\sigma}}{\hat{\gamma}}\left(\left(\frac{qn}{N_t}\right)^{-\hat{\gamma}} - 1\right) \qquad (8)$$

where t is the empirical quantile threshold, n is the number of observations and N_t is the number of observed "peaks" or observations that exceeds t. The general procedure of this algorithm can be summarized in a few steps as follows:

Peaks-over-Threshold (POT) Algorithm
1. Set initial empirical quantile threshold, t.
2. Retrieve all "peaks" or excesses over t, given as $X_i > t$ for i = 1, 2, …, n.
3. Fitting the set of peaks to the Generalized Pareto Distribution.
4. Parameter estimation, $\hat{\sigma}$ and $\hat{\gamma}$ of GPD via maximum likelihood estimates.
5. Set "risk" parameter, q to a suitable value, generally between 10^{-3} and 10^{-5}.
6. Computation of outlier threshold z_q.

A more comprehensive discussions and explanations on the algorithm development is available in the work of Siffer *et al.* (2017) [21]. Considering how we will be implementing this algorithm on rainfall data, it is crucial to first assess its performance prior to actual application. To achieve that, we will be observing how the algorithm perform under differing conditions, primarily on the sample size, n, empirical quantile threshold, t and risk, q.

2.2 Implementation on Actual Rainfall Dataset

After examining the appropriateness of the POT approach through the simulation study, we will then apply the said method on the actual observed dataset on all stations, Batu Embun, Kuantan and Temerloh for the whole daily rainfall series during the longest available duration. The threshold value, z_q and number of detected outliers will be of utmost importance to be observed in this phase of the study.

3.0 RESULTS AND DISCUSSION

In this section, we will be reviewing the properties of the POT algorithm, its drawbacks and strengths under differing conditions and assess its applicability on actual precipitation data for outlier detection. Foremost, we will discuss on the results obtained during the simulation study phase where the algorithm's performance is tested against synthetic data generated from the gamma distribution. Following that, the actual implementation will then be carried out based on the preliminary findings obtained during the simulation study.

The simulation study begins with the fitting of daily rainfall dataset to the gamma distribution. We will then simulate synthetic rainfall series based on the estimated parameters of the distribution. For that purpose, the Batu Embun station was chosen, where the wet days (observed daily rainfall amount is greater than zero) occurrence in the month of December through the span of 40 years was modelled using the gamma distribution. In fact, there was no particular reason on choosing which station to be modelled considering how the main purpose is to only generate synthetic data that mirrors the behaviour of actual daily rainfall data. The period for modelling the data was selected based on the heightened rainfall during the NEM season, which extends through December. To avoid any potential trends, the data will be modelled specifically for this month. The parameters are estimated using MLE method with the shape, α = 0.45813 and scale, β = 27.09699 with standard errors of 0.01803 and 1.72854, respectively. Based on the three goodness-of-fit tests carried out which are the Kolmogorov-Smirnov test (K-S), the Anderson-Darling test (A-D) and the Cramer-von-Mises test (CvM), as shown in Table 2, the estimated parameters of the gamma model are appropriate to represent the observed daily rainfall dataset at 95% confidence level since the p-value > 0.05.

Table 2 Goodness of fit results for estimated gamma model

Goodness of Fit	*p*-value
K-S	0.05021
A-D	0.10680
CvM	0.56150

Following that, we apply the synthetic outlier generation procedure from gamma distribution discussed in the previous section using the estimated α and β values. The resulting computed gamma outlier threshold is T_H = 237.7205. For further inspection on the impact of outliers in the dataset, n = 5000 - k gamma instances were generated and we randomly introduce a value that exceeds T_H as outliers. The process is iterated for a total of 10 times and in each iteration k number of outliers will be introduced in the sample where k = 1, 2, …, 10. The synthetic sample data was then refitted to the gamma model to observe any notable changes on the estimated parameters α and β, and also the effect of outliers in the sample mean and variance, given by αβ and $\alpha\beta^2$, respectively. As seen in Table 3, as the number of outliers in the sample grows, the estimated α shows a decreasing trend while the estimated β increases. More importantly, the existence of the synthetic outliers introduced also shifted the sample mean and variance at an increasing trend, and at k = 10 or 0.2% of the total sample size the absolute error observed was 5.63% for sample mean and 14.74% for sample variance. Specifically, in the context of this study, considering how all of the outliers introduced are upper outliers, the estimated gamma model will now assign more probability towards the more extreme values in the tail region of the distribution; except these points are not actually extreme events but rather anomalies that does not come from the same population distribution. Figure 3 illustrates this argument in the form of the probability distribution function of the fitted gamma model for the dataset with 0, 5 and 10 outliers introduced, respectively. These findings are parallel to the suggestion made by Liu *et al.* (2020) where a relatively small amount of outlier presence is capable of distorting both the sample mean and standard deviation [30]. Hence, based on the results obtained it can be justified that the synthetical generation of outliers for gamma distribution was achieved.

The simulation study then transitions into evaluating the POT algorithm's performance by utilizing the simulated synthetic data in various scenarios. Foremost, we wish to monitor the behavior of z_q with differing number of observations, n and risk parameter, q. This is important as in the original work by Siffer *et al.* (2017), the POT algorithm is run repetitively every time a new entry flagged as a "peak" is detected to update the outlier threshold based on current available information [21].

In other words, the z_q value in the original implementation is ever-changing and once the gathered sample size of "peaks" is large enough, the z_q value will approach a single limit. Contrarily, for the implementation of POT algorithm in a batch of univariate rainfall data, it is crucial to first determine the suitable value for the parameters and monitor how the threshold value would behave under differing conditions. Siffer *et al.* (2017) mentioned that the suitable risk parameter, q is set between 10^{-3} and 10^{-5} for better precision while the sample size experimented was up to n = 15 000, while the value of empirical quantile, t can be set to a value "high enough" and was set to 98% (0.98) in practice [21].

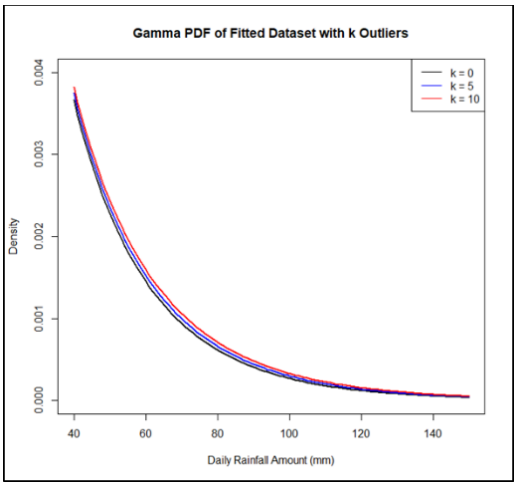


Figure 3 Portion of the gamma PDF fitted to different samples with k outliers

Figure 4 shows the value of z_q when n is set from 500 to 10,000 observations, and when q is set to 0.0001, 0.0005 and 0.001, respectively at t = 0.98. It can be seen that the identical pattern of fluctuation is present for all cases of q, and the fluctuation is not as excessive for n ≥ 1000 as compared to the large difference from n = 500 to n = 1000. A smaller value of q translates to a smaller probability of observing an event where $X > z_q$, hence the outlier threshold value would be inversely proportionate. We shall expect larger threshold value for smaller value of risk parameter, as portrayed in Figure 4. Both findings are in line as the suggestion made by Siffer *et al.* (2017) in their work [21].

Next, we would like to inspect the effect of the empirical threshold quantile, t on the efficacy of the POT algorithm. According to Siffer *et al.* (2017), the relevance of the GPD fit becomes greater with increasing t value, however, if set too high, the number of instances in the set of "peaks" will be reduced and the GPD model will lose more accuracy [21]. To further study on this, we ran the algorithm at a fixed risk parameter, q = 0.0001 and n = 5000 observations at differing level of t = 0.90, t = 0.95 and t = 0.98 and k .

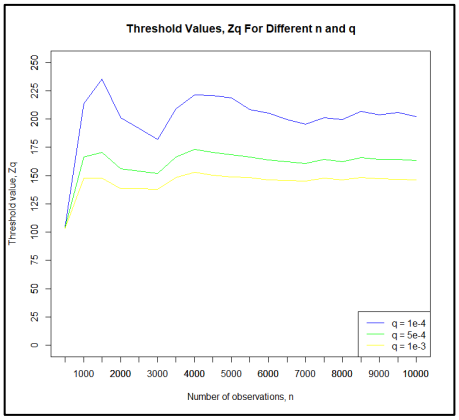


Figure 4 Threshold value, Z_q at differing sample size, n and risk, q

The red line in Figure 5 below shows the synthetic gamma outlier threshold computed at 5000 simulations. A desirable result would be when the z_q values are visually closer to the red line, as this would mean that the outlier threshold is closely estimated. For $k < 3$ outliers introduced, the POT algorithm estimates close z_q values at all t = 0.90, 0.95 and 0.98. Starting from $k \geq 3$, the algorithm shows similar increasing pattern of z_q estimation for t = 0.90 and 0.95, except the recorded outlier threshold is greater for the latter, however, when t = 0.98, the estimated z_q starts to become extremely high to the point that it is no longer reliable, especially with k = 10 outliers introduced to the sample.

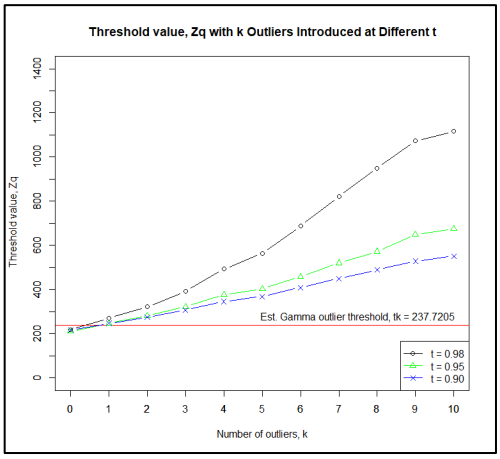


Figure 5 Threshold value, Z_q at differing empirical quantile, t

In this case, the t parameter is what defines the "peaks" or the "extremes" in the dataset. Smaller value of t means that the threshold for identifying "peaks" is lower, and more observations will be defined as such, which can be represented by the red and blue line, respectively in Figure 6. For $k \geq 3$, the algorithm performed terribly at t = 0.98 where the "peaks" defined are too scarce and the outlier now have a more dominant effect in the GPD fitting process prior to the z_q estimation. When the effect of outliers becomes more evident on the sample, the parameters shape, γ and scale, σ were not properly estimated, leading to the severe overestimation of the outlier threshold, z_q. In this situation, the extracted "peaks" are considered as extreme events and it is evident how the introduced synthetic outliers affect the parameter estimation of the GPD, resulting in an unreliable fit. This further justifies the significance of segregating the outliers from extreme observations prior to any statistical analysis to ensure a more accurate result. Based on the result, it can be inferred that the algorithm will perform generally better when the quantile threshold is set to t = 0.90.

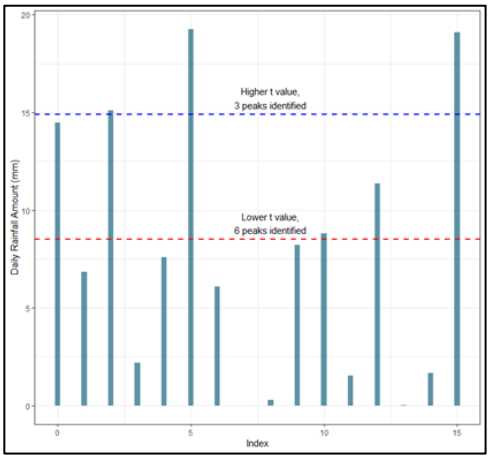
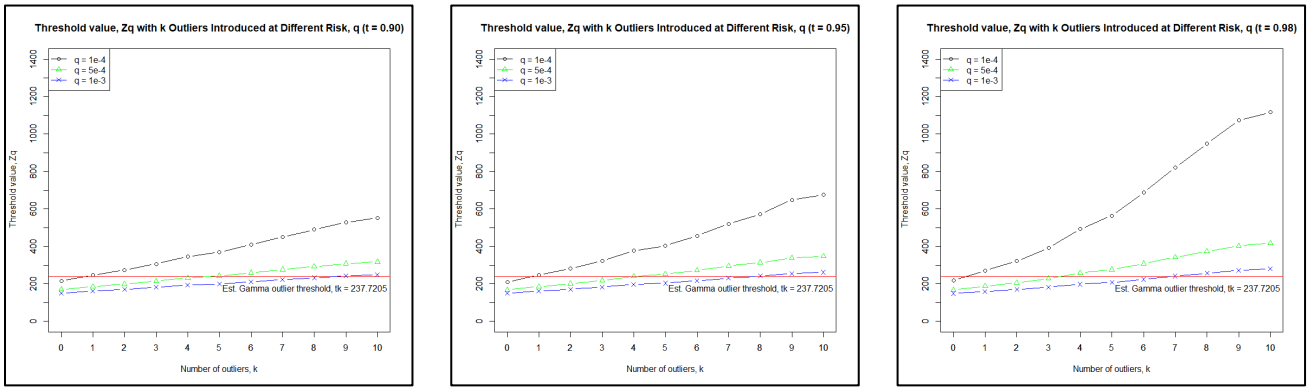


Figure 6 Visualisation of peak identification based on t value

For a more comprehensive insight on the algorithm's efficacy and performance, we also examined the threshold value, z_q and the number of outliers detected when the algorithm is applied to the synthetic dataset with k outliers introduced under different combination of risk, q and empirical quantile, t to look for the optimal set of parameters. Based on the visual representation as shown below, the POT algorithm performs best at all cases of q when t = 0.90 in Figure 7 (a) considering how all the lines are closer to the gamma outlier threshold in contrast to when t is set to 0.95 and 0.98 as in Figure 7 (b) and (c) respectively. Table 4 reflects this finding in terms of the number of detected outliers, where it can be seen that as the value of t increases, the algorithm's accuracy deteriorates for all cases of q. For instance, when t = 0.90, the algorithm with q = 0.0001 can still accurately detect outliers up until k = 3, but when t is set to 0.98, the algorithm fails to detect any outliers for $k \geq 3$. The same behavior is observed for q = 0.0005 and q = 0.001, where the number of detected outliers decreases as t increases, especially at $k \geq 5$. Based on this result, it can be suggested that t = 0.90 is a good baseline to run the algorithm.

Table 3 Effect of outliers on sample property

Sample outliers, *k*	Shape, α	Scale, β	Sample Mean, αβ	Absolute Error (relative to *k* = 0), %	Sample Variance, $\alpha\beta^2$	Absolute Error (relative to *k* = 0), %
0	0.45318	28.48566	12.90924	-	367.72812	-
1	0.45220	28.68209	12.97015	0.47	372.01115	1.16
2	0.45117	28.89645	13.03732	0.99	376.73228	2.45
3	0.44980	29.15213	13.11269	1.58	382.26273	3.95
4	0.44817	29.44916	13.19830	2.24	388.67889	5.70
5	0.44724	29.63864	13.25562	2.68	392.87860	6.84
6	0.44567	29.93277	13.34002	3.34	399.30385	8.59
7	0.44402	30.24839	13.43103	4.04	406.26698	10.48
8	0.44266	30.52181	13.51065	4.66	412.36934	12.14
9	0.44141	30.78944	13.59087	5.28	418.45538	13.79
10	0.44067	30.94251	13.63555	5.63	421.91796	14.74



(a) (b) (c)

Figure 7 Threshold value, Z_q tested under different risk, *q* and empirical quantile, *t* with (a) at *t* = 0.90, (b) at *t* = 0.95 and (c) at *t* = 0.98

Table 4 Number of Detected Outliers by POT Algorithm

Sample outliers, *k*	Number of Detected Outliers								
	t = 0.90			t = 0.95			t = 0.98		
	q = 0.0001	q = 0.0005	q = 0.001	q = 0.0001	q = 0.0005	q = 0.001	q = 0.0001	q = 0.0005	q = 0.001
0	0	2	4	1	2	5	0	2	5
1	1	3	4	1	3	4	1	3	4
2	2	4	4	2	4	4	1	4	4
3	3	4	5	2	3	5	0	3	5
4	2	4	6	2	4	6	0	4	6
5	2	5	7	1	5	7	0	5	7
6	2	6	7	2	6	6	0	4	6
7	3	7	7	0	6	7	0	4	7
8	0	7	8	0	6	8	0	5	8
9	0	7	9	0	6	9	0	4	9
10	0	7	10	0	6	9	0	3	9

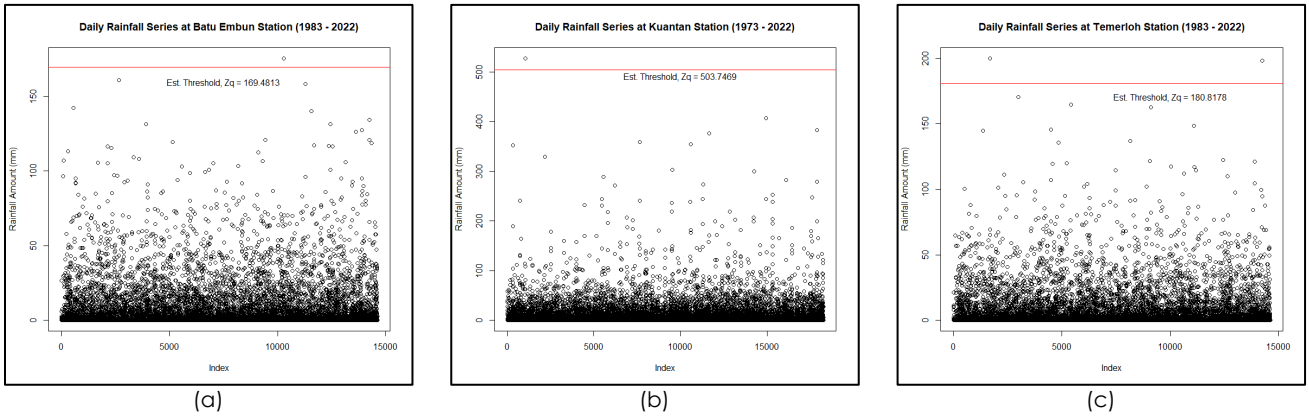


Figure 8 Scatter plot of daily rainfall series at the stations (a) Batu Embun, (b) Kuantan and (c) Temerloh

Focusing on t = 0.90, the findings in Table 4 also suggests that the risk parameter, q shall be set bigger when there are more suspicious observations in the sample. This premise was drawn from the fact that the algorithm accurately detects the outliers at q = 0.0001 for $k \leq 3$, q = 0.0005 for $4 \leq k \leq 7$ and q = 0.001 for $k \geq 7$. This finding is mainly attributed to Equation 7, where it was shown how the value q is associated with the probability of observing an event larger than the threshold value. Consequently, as the number of outliers increases, the aforementioned probability also increases, and to satisfy the relation in Equation 7, q shall also be increased. However, the limitation is that, if q is set too high, then there will be an increase in false positive rate where the actual extreme event is flagged as outlier, and contrarily, if q is set too low, then actual outliers may not be detected. From this finding, it can also be inferred that q = 0.0001is a proper choice when there are not many suspicious observations in the sample.

Following that, the method will be applied on the observed daily rainfall series in three different stations located in Pahang; namely the station of Kuantan, Batu Embun and Temerloh. In order to gain some preliminary insight on the dataset, the rainfall series with sample size, n = 14 262 (Batu Embun and Temerloh) and n = 18 610 (Kuantan) was scattered on a plot. As can be seen in Figure 7 (a) and (c), there are no obvious rainfall occurrence that deviates too far from the main group, but in (b), the Kuantan station has one suspicious observation that exceeds 500 mm while the next largest observation is approximately at 400 mm. Considering the small number of suspected outliers, we decided to run the POT algorithm at t = 0.90 and q = 0.0001 for all three stations. The result shows that there are outliers in all three stations, as portrayed in Table 5. A red line representing the computed threshold was added to the scatter plot to further visualise the result obtained from the outlier detection method, isolating observations with high likelihood of being outliers. Most importantly, the POT algorithm is able to isolate these potential outliers while still preserving the extreme events. As can be seen in Figure 8, the isolated observations all shows significant deviation from other events observed in the stations. Concurrently, extreme observations which occurs less frequently and of higher values, is well preserved in the region below the illustrated outlier threshold. In other words, the removal of these flagged observation will not adversely affect, yet improve subsequent statistical analysis on extreme rainfall events.

Table 5 Goodness of fit results for estimated gamma model

Station	Estimated Outlier Threshold, z_q	Detected Outliers
Batu Embun	169.4813 mm	1
Kuantan	503.7469 mm	2
Temerloh	180.8178 mm	1

4.0 CONCLUSION

In this study, we have demonstrated the method of generating synthetic outliers for gamma distribution and explored on the applicability of EVT in detecting outliers in univariate rainfall series via the application of the POT method, which was originally designed for outlier detection in streaming dataset. The findings suggest that overall, the empirical quantile, t should be set at 90% (0.90) for best result from the algorithm. However, the limitation to be highlighted is that the study has yet to address on how to optimally determine the suitable value of q to be used. From simulation study, it was shown how the risk parameter, q works differently under different influence of outliers, where if it is set too high under no presence of outlier, the algorithm will flag the actual extreme event as outliers, and vice versa in the case where it is set too low under the presence of many outliers.

In the implementation phase, we overcame this problem by gaining some insight based on the scatter plot of the rainfall series, where an abnormally large event was observed in Kuantan station. Due to the low number of suspected observations, the risk parameter was set to q = 0.0001 as suggested based on the simulation study. The result flags the largest observations in Batu Embun and Temerloh station as outliers, and

flagged two largest rainfall events in the Kuantan station. Future studies are recommended to further improve the algorithm's performance by exploring methods to optimize the value of *q* and extend the methodology for outlier detection in multivariate data. To conclude, the study was able to produce satisfactory result for outlier detection in extreme rainfall data and hopefully the findings of this study will be able to improve the data quality in future studies for an improved statistical analysis of extreme rainfall events.

Acknowledgement

Authors acknowledge the Ministry of Higher Education (MOHE) for funding under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2022/STG06/UITM/02/5). Authors would also like to extend our gratitude to the Malaysian Meteorological Department for the daily rainfall record provided which made this study possible.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Mayowa, O. O., Pour, S. H., Shahid, S., Mohsenipour, M., Harun, S. Bin, Heryansyah, A., & Ismail, T. 2015. Trends in Rainfall and Rainfall-related Extremes in the East Coast of Peninsular Malaysia. *Journal of Earth System Science*. 124(8): 1609–1622. Doi: https://doi.org/10.1007/s12040-015-0639-9.

[2] Bin Luhaim, Z., Tan, M. L., Tangang, F., Zulkafli, Z., Chun, K. P., Yusop, Z., & Yaseen, Z. M. 2021. Drought Variability and Characteristics in the Muda River Basin of Malaysia from 1985 to 2019. *Atmosphere*. 12(9): 1–19. Doi: https://doi.org/10.3390/atmos12091210.

[3] Latif, S., & Mustafa, F. 2020. Parametric Vine Copula Construction for Flood Analysis for Kelantan River Basin in Malaysia. *Civil Engineering Journal (Iran)*. 6(8): 1470–1491. Doi: https://doi.org/10.28991/cej-2020-03091561.

[4] Nashwan, M. S., Ismail, T., & Ahmed, K. 2019. Non-stationary Analysis of Extreme Rainfall in Peninsular Malaysia. *Journal of Sustainability Science and Management*. 14(3): 17–34.

[5] Hao, Z., & Singh, V. P. 2013. Modeling Multisite Streamflow Dependence with Maximum Entropy Copula. *Water Resources Research*. 49(10): 7139–7143. Doi: https://doi.org/10.1002/wrcr.20523.

[6] Ma, J., Cui, B., Hao, X., He, P., Liu, L., & Song, Z. 2022. Analysis of Hydrologic Drought Frequency using Multivariate Copulas in Shaying River Basin. *Water (Switzerland)*. 14(8): 1–18. Doi: https://doi.org/10.3390/w14081306.

[7] Radi, N. F. A., Zakaria, R., & Satari, S. Z. 2017. Generating Monthly Rainfall Amount using Multivariate Skew-t Copula. *Journal of Physics: Conference Series*. *890*(1). Doi: https://doi.org/10.1088/1742-6596/890/1/012133.

[8] Win, N. L., & Win, K. M. 2014. The Probability Distributions of Daily Rainfall for Kuantan River Basin in Malaysia. *International Journal of Science and Research*. 3(8): 977–983.

[9] Radi, N. F. A., Zakaria, R., Piantadosi, J., Boland, J., Wan Zin, W. Z., & Azman, M. A. zuhri. 2017. Generating Synthetic Rainfall Total Using Multivariate Skew-t and Checkerboard Copula of Maximum Entropy. *Water Resources Management*. 31(5): 1729–1744. Doi: https://doi.org/10.1007/s11269-017-1597-6.

[10] Marik, R. 2018. Thresholding using Extreme Value Theory Threshold Models. *Proceedings of the 2018 18th International Conference on Mechatronics - Mechatronika, ME 2018*. 1–8.

[11] Liao, X., Wang, T., & Zou, G. 2023a. A Method for Detecting Outliers from the gamma Distribution. *Axioms*. 12(2). Doi: https://doi.org/10.3390/axioms12020107.

[12] Asikoglu, O. L. 2017. Outlier Detection in Extreme Value Series. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*. 4(5): 2458–9403. www.jmest.org.

[13] Gbenro, N. 2020. *Using Extreme Value Theory to Test for Outliers*. https://ssrn.com/abstract=3516056.

[14] Kim, Y., Kim, D., Park, J., & Jun, C. 2024. An Effective Algorithm of Outlier Correction in Space-time Radar Rainfall Data based on the Iterative Localized Analysis. *IEEE Transactions on Geoscience and Remote Sensing*. 1–1. Doi: https://doi.org/10.1109/tgrs.2024.3366400.

[15] Zakaria, R., Ahmad Radi, N. F., & Satari, S. Z. 2017. Extraction Method of Extreme Rainfall Data. *Journal of Physics: Conference Series*. 890(1). Doi: https://doi.org/10.1088/1742-6596/890/1/012154.

[16] Mallick, J., Talukdar, S., Alsubih, M., Salam, R., Ahmed, M., Kahla, N. Ben, & Shamimuzzaman, M. 2021. Analysing the Trend of Rainfall in Asir Region of Saudi Arabia using the Family of Mann-Kendall Tests, Innovative Trend Analysis, and Detrended Fluctuation Analysis. *Theoretical and Applied Climatology*. 143(1–2): 823–841. Doi: https://doi.org/10.1007/s00704-020-03448-1.

[17] Mahajan, M., Kumar, S., Pant, B., & Khan, R. 2021. Improving Accuracy of Air Pollution Prediction by Two Step Outlier Detection. *Proceedings of the 2021 1st International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies, ICAECT 2021*. DOI: https://doi.org/10.1109/ICAECT49130.2021.9392404.

[18] Suhaila, J. 2023. Tweedie Models for Malaysia Rainfall Simulations with Seasonal Variabilities. *Journal of Water and Climate Change*. 14(10): 3648–3670. Doi: https://doi.org/10.2166/wcc.2023.275.

[19] Walfish, S. 2006. A Review of Statistical Outlier Methods. *Pharmaceutical Technology*. 30(11): 82–86.

[20] Bhattacharya, S., Kamper, F., & Beirlant, J. 2023. Outlier Detection based on Extreme Value Theory and Applications. *Scandinavian Journal of Statistics*. 50(3): 1466–1502. Doi: https://doi.org/10.1111/sjos.12665.

[21] Siffer, A., Fouque, P. A., Termier, A., & Largouet, C. 2017. Anomaly Detection in Streams with Extreme Value Theory. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F1296*. 1067–1075. Doi: https://doi.org/10.1145/3097983.3098144.

[22] Ghani, N. A. A. abdul, Mohamad, N. A., & Hui, T. W. 2016. Rainfall Analysis to Determine the Potential of Rainwater Harvesting Site in Kuantan, Pahang. *Journal of Engineering and Applied Sciences*. 11

[23] Wong, C. L., Venneker, R., Uhlenbrook, S., Jamil, a. B. M., & Zhou, Y. 2009. Variability of Rainfall in Peninsular Malaysia. *Hydrology and Earth System Sciences Discussions*. 6(4): 5471–5503. Doi: https://doi.org/10.5194/hessd-6-5471-2009.

[24] Lebay, M., & Le, M. 2020. Techniques of Filling Missing Values of Daily and Monthly Rain Fall Data: A Review. *SF Journal of Environmental and Earth Science*. 3(1): 1036. https://scienceforecastoa.com/

[25] Zakaria, R., Boland, J. W., & Moslim, N. H. 2013. Comparison of Sum of Two Correlated Gamma Variables for Alouini's Model and McKay Distribution. *Proceedings - 20th International Congress on Modelling and Simulation*, MODSIM 2013, December, 408–414. Doi: https://doi.org/10.36334/modsim.2013.a9.zakaria.

[26] Husak, G. J., Michaelsen, J., & Funk, C. 2008. Use of the Gamma Distribution to Represent Monthly Rainfall in Africa

for Drought Monitoring Applications. *International Journal of Climatology*. 2029(March 2008). 2011–2029. Doi: https://doi.org/10.1002/joc.

[27] Soleh, A. M., Wigena, A. H., Djuraidah, A., & Saefuddin, A. 2016. gamma Distribution Linear Modeling with Statistical Downscaling to Predict Extreme Monthly Rainfall in Indramayu. *International Conference on Mathematics, Statistics, and Their Applications (ICMSA)*. Doi: https://doi.org/10.1109/ICMSA.2016.7954325.

[28] Song, S., & Singh, V. P. 2010. Meta-elliptical Copulas for Drought Frequency Analysis of Periodic Hydrologic Data. Stochastic Environmental Research and Risk Assessment, 24(3): 425–444. DOI: https://doi.org/10.1007/s00477-009-0331-1.

[29] Boluwade, A., Sheridan, P., & Farooque, A. A. 2024. Spatial Modeling of Extreme Temperature in the Canadian Prairies using Max-Stable Processes. *Results in Engineering*. 101879. Doi: https://doi.org/10.1016/j.rineng.2024.101879.

[30] Tingfeng Liu, Hui Gao, & Jianjun Wu. 2020. Review of Outlier Detection Algorithms Based on Grain Storage Temperature Data. *International Conference on Artificial Intelligence and Computer Applications (ICAICA)*.