Sains Malaysiana 50(7)(2021): 1843-1856
*http://doi.org/10.17576/jsm-2021-5007-02*

# A Comparative Flood Frequency Analysis of High-Flow between Annual Maximum and Partial Duration Series at Sungai Langat Basin

(Suatu Perbandingan Analisis Kekerapan Banjir Aliran Tinggi antara Siri Maksimum Tahunan dan Siri Tempoh Separa di Lembangan Sungai Langat)

FIRDAUS MOHAMAD HAMZAH*, HAZRINA TAJUDIN & OTHMAN JAAFAR

## ABSTRACT

*Flood frequency analysis should consider small and frequent floods. Despite the complexities in partial duration series implementation, it can give a better flood estimation in a way that it does not exclude any significant high flow events, even if it is not the highest event of the year. This study employs the streamflow data recorded at Kajang station, Sungai Langat, Malaysia over a 36-year period spanning from 1978 to 2013. The paper attempts to conduct flood frequency analysis using two approaches, annual maximum and partial duration series. The optimal threshold value is selected to be 48.7 m$^3$/s, where the dispersion index stabilizes at around 1, DI = 1. The results have shown that generalized extreme value (GEV) distribution describes the annual maximum data while the lognormal (LN3) and generalized Pareto (GPA) distribution is chosen as the best fit distribution at Kajang station for a partial duration series. There is a slight difference between estimated streamflow magnitude when using GPA and LN3 for selected return periods, while a considerable difference was observed when using annual maximum at a higher return period. As a conclusion, PDS gives more relevant magnitude estimation rather than AMS. Flood frequency plays an important role in understanding the nature and magnitude of high flow, which in turn can assist relevant agencies in the design of hydrological structures and reduce flood impacts.*

*Keywords: Flood frequency analysis; generalized extreme value; generalized Pareto; Sungai Langat; three-parameter lognormal*

## ABSTRAK

*Analisis kekerapan banjir harus mempertimbangkan kejadian banjir dengan magnitud kecil dan kerap. Walaupun terdapat kerumitan dalam pelaksanaan data siri separa, ia memiliki kemampuan untuk memberikan anggaran banjir yang lebih baik, dengan tidak mengecualikan kejadian aliran tinggi yang signifikan, walaupun ia bukan peristiwa tertinggi tahun ini. Kajian ini menggunakan data aliran yang direkodkan di stesen Kajang, Sungai Langat, Malaysia dalam jangka masa 36 tahun yang merangkumi tahun 1978 hingga 2013. Objektif utama kajian ini adalah Nilai ambang optimum dipilih menjadi 48.7 m$^3$/s dengan indeks penyebaran stabil pada sekitar 1, DI = 1. Taburan nilai ekstrim teritlak (GEV) menerangkan data maksimum tahunan sementara taburan lognormal dan Pareto teritlak dipilih sebagai taburan yang paling sesuai di stesen Kajang untuk data siri separa. Terdapat sedikit perbezaan antara magnitud aliran dengan menggunakan taburan Pareto teritlak dan lognormal untuk tempoh pulangan yang dipilih. Manakala, perbezaan yang cukup besar dapat dilihat apabila menggunakan data tahunan maksima terutamanya pada tempoh pulangan yang lebih tinggi. Secara kesimpulan, PDS memberikan anggaran magnitud yang lebih relevan berbanding AMS. Kekerapan banjir memainkan peranan penting dalam memahami sifat dan besarnya aliran tinggi, yang seterusnya dapat membantu agensi yang berkaitan dalam merancang struktur hidrologi dan mengurangkan kesan kejadian banjir.*

*Kata kunci: Analisis kekerapan banjir; lognormal tiga-parameter; nilai ekstrim teritlak; Pareto teritlak; Sungai Langat*

## INTRODUCTION

Flood frequency analysis is important in various area including projects management, areal and water resource planning (Engeland et al. 2018) such as in the design of important infrastructures and hydrological planning. According to hydrological perspective, flood frequency analysis is an important tool used to estimate future flood events based on the historical data of streamflow events. The short period of time series data available for the presents study is the main challenge in making the computations for hydrological analysis (Gado & Nguyen 2016). Result of the analysis is presented in terms of frequency and magnitude of flood events (Keast & Ellison 2013). There are two main methods for computing flood frequency analysis, graphical, and analytical. The graphical method is represented by Gringorten Plotting Position which is essentially a plot magnitude of selected return period on a probability paper (Makkonen 2006). The analytical method involves the implementation of a statistical distribution of the data series which is used to compute the average recurrence interval (known as return period) of discharge magnitude. The data used to carry out frequency analysis must be independent and identically distributed (Malamud & Turcotte 2006). Additionally, the data must come from the same population and must not show any seasonal pattern in the time series (Tallaksen & Hewa 2008).

Most research consider extreme flood events instead of medium and frequent floods (Cheong & Gabda 2018; Jiang & Kang 2019; Keast & Ellison 2013; Madsen et al. 1997). This problem can be solved by taking into consideration the partial duration series in preference to the annual maximum series data (Claps & Laio 2003). There are two factors which cause difficulty in implementing PDS which are determination of peak independence and threshold level. Selecting a higher threshold value in a series would results in a smaller number of events being included in the series and this leads to a loss of valuable information. It often increases the likelihood of independence of each peak. On the other hand, selecting a low threshold value would result in a bigger number of events being selected for this study. This would provide a more reliable parameter estimation whilst also increasing the likelihood of dependence of the series, is sometimes known as peak over threshold value (POT). PDS does not exclude any significant high flow events, even if it is not the highest event of the year. As such, this method ensures a better representation of the sampling procedure of extreme values.

Failure to allocate appropriate FFA can lead to rupture in hydrological design. Overspill of water could damage the engineering properties of the material structure, thus affects the structure itself. For example, construction of a main road usually needs to last for 50 to 100 years, while streamflow magnitude with less than 20 years recurrence interval is used for drainage design. Basically, the application of FFA is closely related to design life and failure probability of the hydrological design (Leščešen & Dolinaj 2019). Over-design of a structure incurs more cost, while under design hydrological structure increase the structure maintenance. Thus, it is important to generate an accurate streamflow magnitude for each recurrence interval.

Malaysia being a rapidly growing country, needs to have a proper planning to prevent natural disaster from repeating to minimize property damage and environmental impact (Syed Hussain & Ismail 2013). Sungai Langat was chosen for the analysis due to rapid growth in Selangor area. There is vast construction development, includes housing, shopping complex, and commercial building due to increase in the number of populations and demand. According to the statistics recorded by the Department of Statistics Malaysia, the total number of population in Selangor reached from 6.38 million in 2017 to 6.5 million in 2019, with population density of 819/km$^2$ (Department of Statistics Malaysia 2019). Rapid industrialisation and urbanization have led to deforestation and uncontrolled land use and, in many areas, this has altered the relationship between rainfall and flooding events (Tekolla 2010). Previous research has shown that the intensity and frequency of extreme rainfall events are on the increase, thus, creating a non-stationary component. This is essentially the consequence of climate change (Agilan & Umamahesh 2017). There are two main factors which leads to the flooding occurrence in Selangor area which are improper drainage system along with heavy and continuous precipitation (Franchini et al. 2005). In addition, due to heavy precipitation, the water release from dam to prevent cracking on the dam is one of the factors causing the flooding in Langat area. This is usually called flash flooding, which happen due to lack of improper management of the drain and sewer, especially during heavy or continuous rainfall, resulted in excess runoff which eventually exceeds river capacity. As a result, many people were affected by the incidence especially those who stay in the lowland area. Hence, this study aims to conduct flood frequency analysis near Kajang station to determine the magnitude at selected return period by implementing appropriate statistical distribution. Next, it aims to compare between the usage of annual maximum series and partial duration series. This study is important to develop an effective planning and management of flood mitigation.

## STUDY AREA

The present study focuses on Sungai Langat in the State of Selangor. The catchment area for Sungai Langat is approximately 2350 km$^2$. It is located at latitude 2° 40’M 152” N to 3° 16’M 15” and longitudes 101° 19’M 20”E to 102° 1’M 10”E. The main river course of Sungai Langat is about 141 km and is mostly located about 40 km east of Kuala Lumpur. The principal rivers of Sungai Langat are Sungai Semenyih, Sungai Lui and Sungai Beranang. The four streamflow stations along Sungai Langat are located in Pekan Dengkil, Kajang, Sungai Semenyih and Sungai Lui. This study focuses on the Sungai Langat-Kajang sub basin, which has an area of approximately 389.4 km$^2$. Figure 1 shows the streamflow stations along Sungai Langat, namely Sg. Lui, Kajang, Rinching and Dengkil stations.

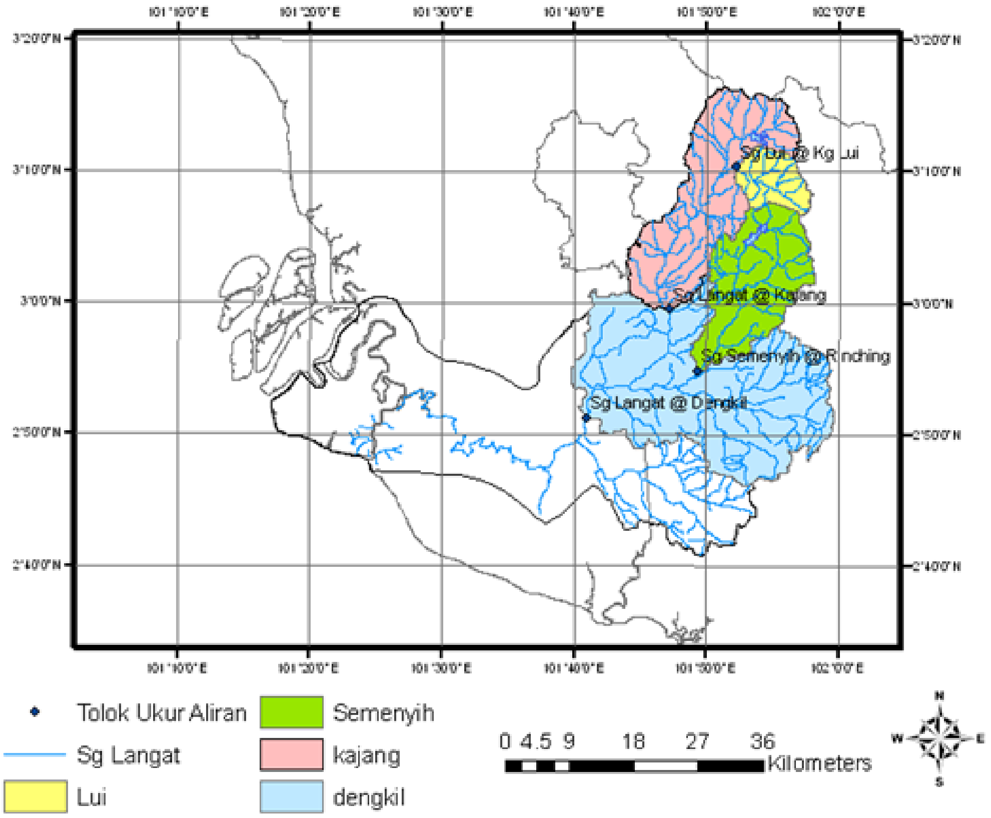

FIGURE 1. Streamflow station at Sungai Langat

## Methods

This section comprises several parts, i.e., data collection, distribution fitting and parameter estimation, goodness of fit testing, and return period.

### DATA COLLECTION

The site chosen for this study is the Kajang station in Sungai Langat sub basin. The station's ID is 2917401 and it is located at latitude 02° 59' 34" and longitude 101° 47' 13" (Figure 1). The duration of the streamflow data is approximately 36 years of daily data spanning from 1978 to 2013. The data is measured in cubic meter per second (m$^3$/s). There is no missing data in the dataset provided by the Department of Irrigation and Drainage (DID) Malaysia.

### ANNUAL MAXIMUM SERIES

Annual maximum series can be extracted from highest flow each year. Since the period of record is between 1978 and 2013, hence there are 36 data extracted from the daily series.

### PARTIAL DURATION SERIES

Partial duration series (PDS) is extracted from the daily streamflow based on certain condition. There are two factors affecting the appropriate selection of PDS namely threshold selection and independence criteria. Based on the traditional approach, 5% of the total maximum value represents the extreme maximum of the hydrological series. After determining the least optimal threshold value, it should satisfy certain statistical rules, the dispersion index which stabilize at around 1 and average number of occurrences per year to be at least 1.68. Then, the independence criteria should satisfy the following (Bezak et al. 2014):

$$\theta < 5 + \log (A) \qquad (1)$$

where $\theta$ represents number consecutive days which considered as independent series and $A$ represent the total area of the basin.

### DISTRIBUTION FITTING AND PARAMETER ESTIMATES

After obtaining the data series, it is important to determine the proper statistical distribution function that is able to describe the time series data. There may be several distributions that fit the data well and it would be difficult to determine the best model amongst these distributions. Some of the statistical distributions which can be implemented are Gumbel, generalized extreme value (GEV), generalized logistics (GLO), generalized Pareto (GPD), and lognormal (LN) distributions. GEV distribution is frequently used for AMS data (Jiang & Kang 2019). Generalized Pareto distribution, also known as GPD, was introduced by Pikands in 1975 (Mierlus-Mazilu 2010). It is often implemented in peak over threshold data set (Gharib et al. 2017; Mierlus-Mazilu 2010). Table 1 shows the probability density function and cumulative distribution functions.

TABLE 1. Probability density function and cumulative distribution function for each distribution

| Distribution | Probability density function (pdf) | Cumulative distribution function (cdf) | Quantile function |
|---|---|---|---|
| Gumbel | $f(x) = \frac{1}{\sigma} e^{(-z-e^{-z})}$<br>where $z = \frac{x-\mu}{\sigma}$ | $F(x) = 1 - e^{-\left(\frac{x-\zeta}{\beta}\right)^{\delta}}$ | $x(F) = \zeta + \beta(-\log(1-F))^{\frac{1}{\delta}}$ |
| Generalized extreme value | $f(x) = \frac{1}{\sigma} e^{(-(1+kz)^{-\frac{1}{k}})(1+kz)^{1-\frac{1}{k}}}$ | $F(x) = e^{(-(1+kz)^{-\frac{1}{k}})}$<br>where<br>$y = -k^{-1} log(1 - \frac{k(x-\xi)}{\alpha})$ | $x(F) = \xi + \frac{\alpha}{k}\{1 - (-\log(F))^{k}\}$ |
| Lognormal (3P) | $f(x) = \frac{e^{(-\frac{1}{2}(\frac{\ln(x-\gamma)-\mu}{\sigma})^{2})}}{(x-\gamma)\sigma\sqrt{2\pi}}$ | $F(x) = \phi(y)$<br>where<br>$y = \frac{(\log(x-\zeta)-\mu)}{\sigma}$ | $x(F) = e^{\sigma\phi^{-1}(F)+\mu} + \gamma$ |
| Generalized logistics | $f(x) = \frac{(1+kz)^{-1-1/k}}{\sigma(1+(1+kz)^{-\frac{1}{k}})^{2}}$ | $F(x) = \frac{1}{1+(1+kz)^{-1/k}}$ | $x(F) = \mu + \frac{\sigma}{k}((y^{-1}-1)^{-k} - 1)$ |
| Generalized Pareto | $f(x) = \frac{1}{\sigma}(1 + k(\frac{x-\mu}{\sigma})^{-1-\frac{1}{k}}$ | $F(x) = 1 - (1 + k(\frac{x-\mu}{\sigma})^{-\frac{1}{k}})$ | $x(F) = \xi + \frac{\alpha}{k}\{1 - (1-F)^{k}\}$ |

$\phi^{-1}$ ($F$) is the inverse of probability distribution function for normal distribution.

where $f(x)$ is probability density function; $F(x)$ is cumulative distribution function; $x$ is the data series; $z$ represents standard value of normal distribution and $k$, $\sigma$, $\mu$ represent shape, scale and location parameters of the distribution (Chang et al. 2016).

Several methods, including method of moments (MOM), maximum likelihood function (MLE), and L-moment method (ML), can be used to estimate parameters. Evaluation method performance is dependent on sample size and skewness of the data. MLE can give the best parameter value compared to other methods. It maximizes the likelihood or joint probability of occurrence of the observed sample. However, MLE is not suitable for implementation in a small sample size. In MOM, the estimators of the population moments must be equal to the sample moments. MOM is best implemented when moments are available. ML is a particular linear combination of probability weighted moments (PWMs) which gives simple interpretations of the location, shape and dispersion of a sample data. Unlike other product moment estimators, ML is not affected by sample variability. MLE estimator does not exist for a shape parameter of −1. L-moment has theoretical advantages over conventional moments in that it is able to characterize a broader range of distribution and is less affected by bias (Bílková 2014; Schlögl & Laaha 2017). Hydrological parameters usually contain outliers. ML is said to be robust and is not significantly affected by sampling variability (Srinivasa Murthy et al. 2017). Unlike other methods of parameter estimation, the L-moment method is not biased (Alahmadi et al. 2014; Ummi Nadiah et al. 2013).

Among all the distributions, only Gumbel is represented by two parameters. The location parameter of a distribution indicates where the distribution lies along the x-axis (horizontal axis). The scale parameter of a distribution determines the degree of spread in a distribution. The shape parameter of a distribution allows the distribution to take different shapes. The threshold parameter of a distribution represents the minimum value of the distribution along the x-axis (Scarrott & Macdonald 2012).

MOM and ML have the same moments, which can be defined as follows (Khan et al. 2017):

$$l_1 = \beta_0 \tag{2}$$

$$l_2 = 2\beta_1 - \beta_0 \tag{3}$$

$$l_3 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{4}$$

$$l_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \tag{5}$$

where $\beta r$ ($r$ = 0,1,2,3) represents probability weighted moments such that:

$$\beta_r = n^{-1} \sum_{i=r+1}^{n} \binom{j-1}{r} \binom{n-1}{r}^{-1} * X(j,n)\,, r = 0, n-1 \tag{6}$$

Each moment represents location, dispersion, symmetry, and peakedness of a data series. Based on the calculated moments, ML ratios can be established using the following calculation for coefficient of variation (CoV), skewness, and kurtosis.

$$\tau_2 = \frac{l_2}{l_1} \tag{7}$$

$$\tau_3 = \frac{l_3}{l_2} \tag{8}$$

$$\tau_4 = \frac{l_4}{l_2} \tag{9}$$

## GOODNESS FOR FIT TESTING

The Kolmogorov-Smirnov and Anderson-Darling statistical tests were implemented to evaluate the performance of the distribution. The test statistics used to assess the goodness of fit are as follow:

$H_0$: The data follows specific distribution
$H_A$: The data does not follow a specific distribution

Evaluation of the tests is based on the calculated p-value. $H_0$ is rejected if the calculated p-value is less than 0.05 ($p<0.05$). Rejection of $H_0$ means the data is not explained by the specific distribution. The Kolmogorov Smirnov and Anderson Darling tests can be used only for continuous data. AD is essentially an updated version of the KS test which considers the tail of the tested distribution; this is in contrast to KS which is known to be distribution free. It involves the computation of a critical value of the test at which AD will give more accurate judgement than KS since the mathematical computation involves a cumulative distribution function of the tested distribution. In general, goodness of fit testing compares the difference between theoretical and empirical distribution functions (Singla et al. 2016).

KS test is computed based on an empirical distribution function (ECDF) and theoretical distribution function (Ghasemi & Zahediasl 2012). The critical value is computed as follow:

$$F_n(X_i) = \frac{N(i)}{n} \quad (10)$$

$$F(x) = \int_a^x f(y,\theta)dy \quad (11)$$

$$D_n = \sup_{1\le i\le n}|F(x_i) - F_n(x_i)| \quad (12)$$

where $N(i)$ is the number of points that is less than $X_i$ and $f(y,\theta)$ is the probability density function. The test statistic, $D_n$, is rejected if it exceeds the tabulated critical value or when the p-value is lower than the significance level. KS test is also suitable for a small sample size. However, to utilize this test, the location, shape, and scale parameters have to be specified since they cannot be estimated directly from the data.

AD test is an upgraded version of the KS test which considers the tail of a distribution. This test can overcome the limitation of KS although it can only be used for certain distributions. AD is more sensitive towards the tail of the distribution. The test statistics can be mathematically written as follow:

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(lnF(X_i) + \ln(1 - F(X_{n-i+1}))) \quad (13)$$

where $n$ is sample size; $F(x)$ is the cumulative distribution function for the tested distribution; and the $i$th sample is calculated after sorting the data in ascending order. The p-value is the probability that the data in a sample is random, and the p-value is dependent on the statistical value obtained from the above equation. The value of maximum difference between the two is a measure of the difference between the calculated and the observed data (Garba et al. 2013). AD test is a better version of goodness of fit test than the K-S test (Özonur et al. 2013).

### RETURN PERIOD

Return period, also known as recurrence interval, is an estimation of the likelihood of the occurrence of an event. Mathematically, the magnitude at selected return period can be calculated using the inverse of the cumulative distribution function (also known as quantile function). The formula for each distribution is presented in Table 1.

## RESULTS

This section presents the descriptive analysis, distribution fitting and parameter estimates, goodness of fit testing and computation of magnitude for a selected return period. All topics comprise two parts, namely the computed results for the annual maximum and the partial duration series.

### DESCRIPTIVE ANALYSIS

Hydrological data such as rainfall and streamflow are always skewed to the right. There is no negative value in the data structure. It is either zero, which indicates that there is no event, or greater than zero, which indicates any possible hydrologic event. The duration of the recorded data is crucial in determining the skewness of data distribution. The larger the number of recorded data, the higher probability of observing infrequent events of high magnitude; therefore, the data will be more skewed and the analysis more accurate (Schlögl & Laaha 2017; van Westen & Jetten 2015).

### DAILY STREAMFLOW DATA

Figure 2 presents the daily plot for Kajang station for the period from 1978 to 2013. There is a total of 13149 data for the 36-year period. The descriptive statistics of the daily data is shown in Table 2.

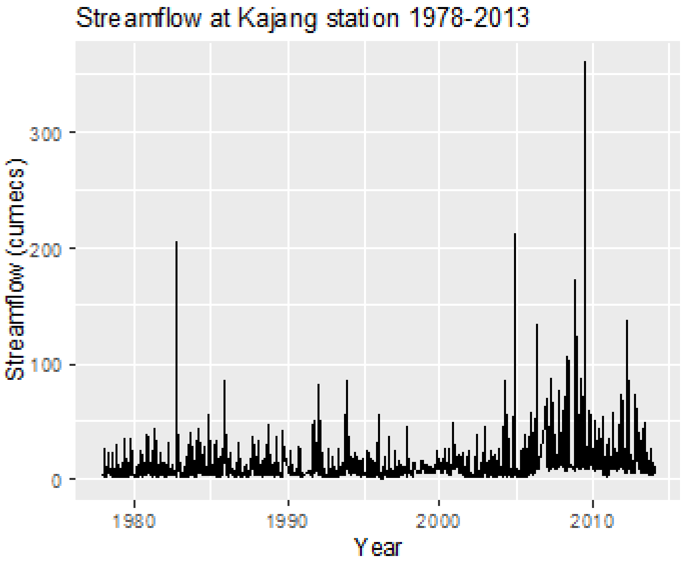

FIGURE 2. Daily plot at Kajang station between 1978-2013

### ANNUAL MAXIMUM MODEL

Annual maximum series consist of the highest maximum data for each year between 1978 and 2013. The descriptive statistics of AMS is summarized in Table 2. Figure 4 is a histogram and density plot for the annual maximum data at Kajang station. Only one maximum event of 360.8 m$^3$/s was recorded during this period, as can be seen in Figure 3.

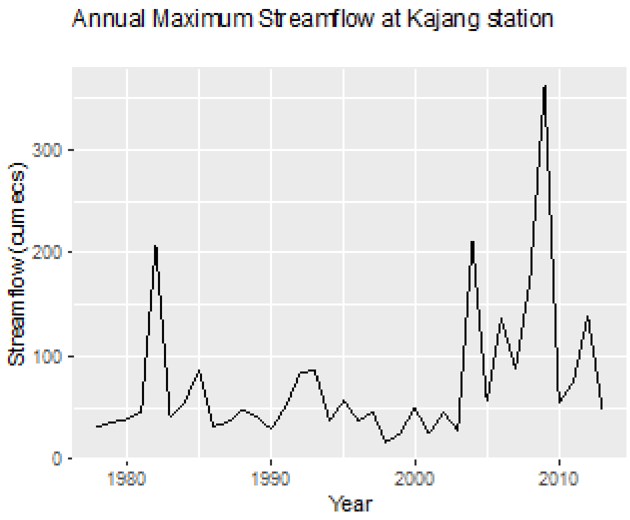

FIGURE 3. Annual maximum flow at Kajang station

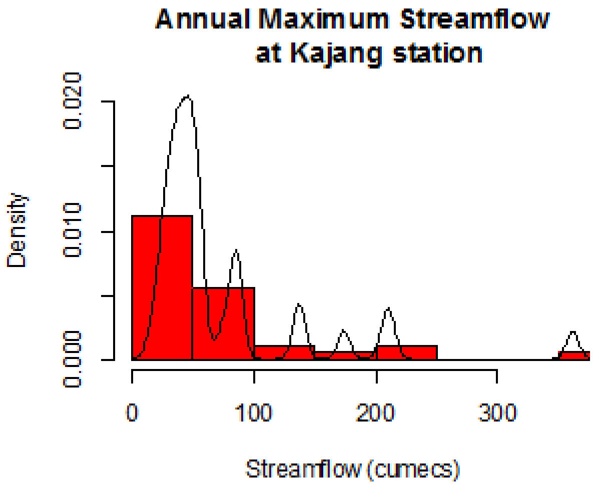

FIGURE 4. Histogram and density plot of annual maximum at Kajang station

TABLE 2. Descriptive analysis of annual maximum streamflow at Kajang station

| Description | Daily streamflow data | Annual maximum series |
|---|---|---|
| Minimum streamflow (m$^3$/s) | 0.1 | 16.55 |
| Maximum streamflow (m$^3$/s) | 360.8 | 360.8 |
| Average streamflow (m$^3$/s) | 9.81 | 73.53 |
| Standard deviation (m$^3$/s) | 11.61 | 69.48 |
| Skewness | 7.15 | 2.4 |
| Kurtosis | 114.33 | 6.15 |

### PARTIAL DURATION SERIES MODEL

The threshold value selected in this study is based on percentile, where data above 90% in flow duration curves (FDC) is selected for the analysis. The values exceeding several percentages were tested and shown in the first row of Table 3 which presents the characteristics of PDS samples for the period from 1978 and 2013.

TABLE 3. Characteristics of PDS samples at Kajang station

| Percentile (%) | 90.0 | 91.5 | 93.0 | 94.5 | 96.0 | 97.5 | 98.0 | 98.5 |
|---|---|---|---|---|---|---|---|---|
| Threshold (m$^3$/s) | 18.6 | 20.2 | 22.6 | 25.6 | 30.6 | 39.2 | 43.6 | 48.7 |
| Sample size | 380 | 337 | 291 | 249 | 187 | 116 | 91 | 66 |
| Rejected Peaks | 935 | 781 | 629 | 474 | 339 | 213 | 172 | 131 |
| $\lambda$ | 10.56 | 9.36 | 8.08 | 6.92 | 5.19 | 3.22 | 2.53 | 1.83 |
| Mean (m$^3$/s) | 36.2 | 38.93 | 42.02 | 45.94 | 52.27 | 63.79 | 70.03 | 79.12 |
| Standard Deviation (m$^3$/s) | 28.55 | 29.72 | 30.97 | 32.35 | 35.18 | 40.84 | 44.14 | 48.91 |
| Skewness | 5.87 | 5.65 | 5.51 | 5.33 | 5.02 | 4.37 | 4.03 | 3.58 |
| Kurtosis | 51.13 | 46.90 | 43.73 | 40.2 | 34.28 | 24.65 | 20.4 | 15.53 |

Note: $\lambda$ = Average number of peaks per year

The sample size in Table 3 represents the number of events which exceeds the selected threshold after considering the independence of the events. The rejected peaks are the number of events excluded from the study based on independence criteria stated in the methods section. As the threshold value gets larger, the number of samples being considered decreases. The number of extreme maxima is smaller at higher threshold level. The magnitude of discharge selected is the highest peak in each cluster. Based on the dispersion index plot, the optimal threshold should be selected when the plot stabilizes at around 1. Figure 5 shows the threshold value against the dispersion index based on the assumption of the Poisson process. The dispersion index plot in Figure 5 shows that the index begins to stabilize as the threshold approaches 50 m$^3$/s. Hence, a threshold of 48.7 m$^3$/s is selected for this study. The density plot for each tested threshold is shown in Figure 6.

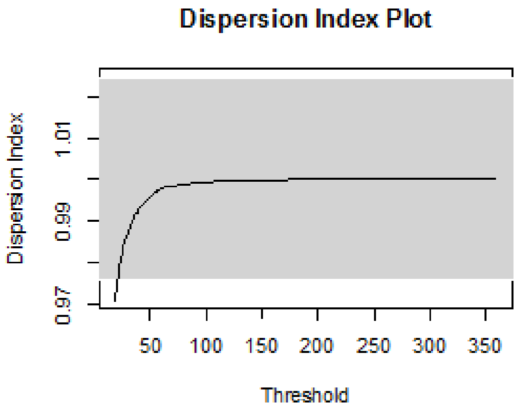

FIGURE 5. Dispersion Index plot

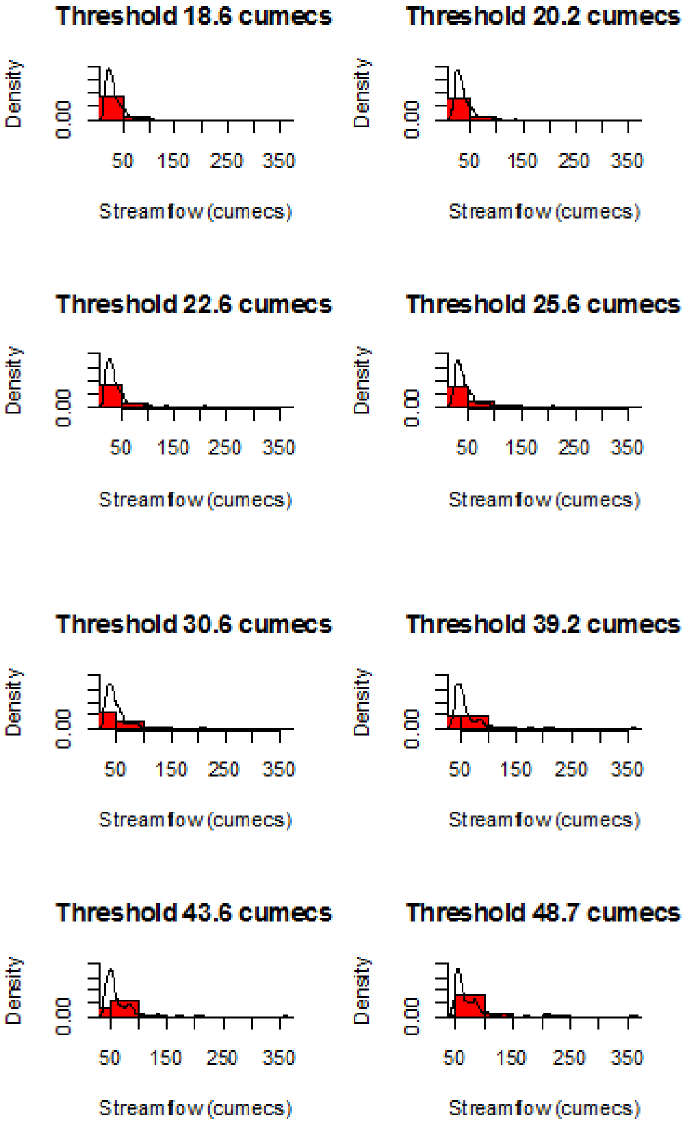

FIGURE 6. Histogram and density plot for partial duration series

### DISTRIBUTION FITTING AND PARAMETER ESTIMATES

ML is suitable for estimating the parameters for a hydrological data series. Table 4 presents the estimation of parameters for the annual maximum series and partial duration series at 48.7 m$^3$/s.

TABLE 4. Parameter estimate for annual maximum series using L-moment method

| Distribution | Parameters | | | |
|---|---|---|---|---|
| Lognormal (3 parameter) | AMS | ζ= 20.08 | μ= 3.33 | σ= 1.14 |
| | PDS | ζ= 47.81 | μ= 2.68 | σ= 1.23 |
| Generalized Logistics | AMS | ξ = 50.46 | α= 18.99 | κ = -0.52 |
| | PDS | ξ = 63.97 | α= 10.92 | κ = -0.55 |
| Generalized extreme value | AMS | ξ= 41.11 | α= 22.09 | κ= -0.48 |
| | PDS | ξ= 58.71 | α= 12.41 | κ= -0.52 |
| Generalized Pareto | AMS | ξ= 22.94 | α= 32.15 | κ= -0.36 |
| | PDS | ξ= 48.73 | α= 17.39 | κ= -0.48 |
| Gumbel | AMS | ξ= 47.77 | α=44.63 | - |
| | PDS | ξ= 63.03 | α=27.88 | - |

### GOODNESS OF FIT TESTING

Anderson Darling test considers the performance of each tested distribution in contrast to the Kolmogorov Smirnov test which is known to be distribution free.

### ANNUAL MAXIMUM SERIES

The result presented in Table 5 shows that lognormal (3 Parameters) and generalized Pareto gives small p-value, showing that the annual maximum data are not from these distributions. The calculated p-value shows that generalized extreme value is most likely able to explain the data set. Therefore, the quantile function of GEV distribution will be used to calculate the magnitude of streamflow at selected return period.

### PARTIAL DURATION SERIES

Two distributions at the 48.7 m$^3$/s threshold value are suitable for representing the data set. The two distributions are lognormal (3 parameters) and generalized Pareto distribution, and both distributions have a p-value greater than 0.05. Thus, the quantile function of these two distributions will be used to calculate the magnitude of streamflow at selected return period.

TABLE 5. Goodness of fit testing for annual maximum series

| Tests | Annual maximum | Partial duration series | Annual maximum | Partial duration series |
|---|---|---|---|---|
| | Kolmogorov-Smirnov | | Anderson Darling | |
| Lognormal (3 parameter) | 0.6724 | 0.0086 | <0.05 | 0.3371 |
| Generalized logistics | <0.05 | 0.0002 | <0.05 | <0.05 |
| Generalized extreme value | 0.4549 | 0.0003 | 0.5027 | 0.0205 |
| Generalized Pareto | 0.6136 | 0.0060 | <0.05 | 0.2444 |
| Gumbel | 0.0385 | 0.0095 | 0.0076 | <0.05 |

### RETURN PERIOD

A flood frequency analysis is carried out to evaluate future occurrence of flood with certain magnitude at return period of 5, 10, 20, 50 and 100 years. The magnitude value for each return period is given in Table 6. The calculation of magnitude at selected return period for AMS employs the quantile function of GEV while, PDS utilises the LN (3) and GP distributions.

TABLE 6. Streamflow magnitude at selected return period

| Return period (years) | Streamflow magnitude (m$^3$/s) | | |
|---|---|---|---|
| | Annual maximum series | Partial duration series | |
| | | Lognormal (3 Parameters) | Generalized Pareto |
| 5 | 136.1 | 89.0 | 89.0 |
| 10 | 197.1 | 115.9 | 116.9 |
| 20 | 280.4 | 150.6 | 154.5 |
| 50 | 441.1 | 212.2 | 224.7 |
| 100 | 618.4 | 274.5 | 299.5 |

The calculated return period which considers annual maximum series gives a higher estimated magnitude compared to that of partial duration series. There is slight difference between the estimated streamflow value for the 5- and 10-year return period, and the difference increases as longer periods of between 20 and 100 years return period are used. However, the estimated flow determined using the partial duration series for both distributions shows a small difference for a short period although the difference is rather significant at about 25 m$^3$/s for a 100-year return period.

### DISCUSSION

The utilisation of PDS data in preference to AMS allows for much more data to be included in the analysis. For instance, for the data from a 36-year period, 66 data are considered at a threshold value of 48.7 m$^3$/s instead of only 36 data. All peaks above the threshold value are selected and filtered to ensure that only peaks which satisfies the independence criteria are used in this study. To fulfil the independence criteria of a data series, the occurrence of two consecutive discharge peaks must exceed an interval of 8 days. The average number of peaks used to select optimum threshold cannot be easily determined (Pham et al. 2014). Additionally, the average number of occurrences must satisfy $\lambda > 1.65$ where $\lambda = 1.83$. The number of occurrence per year follows a Poisson distribution such that threshold values of less than 30.6 m$^3$/s are rejected. This is due to the fact that the distribution is almost symmetrical if $\lambda > 5$ and becomes normally distributed, thus, violating the assumption of the Poisson process (Cunnane 1979).

The selection of a proper threshold value for extracting PDS is crucial. According to Beguería (2005), the selected threshold value will determine the lower part of the distribution where any small change in the threshold value will modify the lower tail. Thus, minor changes on the left tail of the distribution leads to a significant difference in the estimated parameters. Moreover, compared to AMS, PDS is more suitable when estimating the high frequency of a small magnitude flood. The estimation of small magnitude flood is suitable for a return period of less than 10 years (Keast & Ellison 2013). PDS series is the better option when the chosen value is close to the actual flood discharge (Karim et al. 2017).

Skewness is a measure of symmetry of a distribution, while kurtosis measures the combined probability in the two tails (tail-heaviness). Generally, normal distribution has a skewness of 0 and a kurtosis of 3. It describes the relative size of the left and right tails. A kurtosis greater than 3 indicates that the data set have heavier tails and that there are more data in the tail part of the distribution. Hydrological data such as rainfall and streamflow are always skewed to the right. There is no negative value in the data structure. It is either zero, indicating that there is no event, or greater than zero, indicating that there is a possible hydrologic event. The skewness of data distribution is also influenced by the duration of recorded data. A large amount of recorded data means a higher probability of observing infrequent events with high magnitude; as a result, the data will be more skewed and the analysis will be more accurate (Schlögl & Laaha 2017; van Westen & Jetten 2015).

Based on the descriptive analysis presented in the results section, all kurtosis has a value greater than 3, which implies that all data set have heavier tails and that the tail of the distribution contains more data. A skewness greater than one (>1) also implies that the models are highly skewed. The result of the partial duration series shows a lower kurtosis value, indicating that the tail of the data set becomes lighter with higher threshold value. The shape parameter is explained by the skewness and kurtosis of the data series. Location parameter is known to shift the distribution while the scale parameter determines whether the distribution shrinks or stretches.

The Anderson-Darling (A-D) test is an improvement of Cramer-von Mises statistics. It gives more weight to the tail distribution in contrast to Kolmogorov-Smirnov test. Hence, it is very useful for detecting outliers in a data series (Anderson & Darling 1954). A small value of the A-D test implies the suitability of the distribution for the available data set. Previous research shows that the distribution tested with AD statistics value above 1.038 will not be considered while distributions with AD statistics value between 0.474 and 1.038 may be considered for the analysis (Pettitt & Stephens 1977). AD test can be used for a small sample size and in most hydrologic conditions where the data is highly asymmetric.

Based on the technical information obtain from DID, the danger water level recorded at Kajang station is at 26.3 m, where the streamflow reading is around 157 m$^3$/s. According to flood frequency analysis (FFA), at return period of 20-years, there are 5% chance of flooding occurring in any year. Seven flooding occurrences throughout 36 years of data record, where the extreme flood happens on 2009 with magnitude of 360.8 m$^3$/s, followed by 225.9 m$^3$/s which is recorded on the next day. Sungai Langat around Kajang station has experienced change in river profile for the past year. Based on the rating curve obtained, there is difference between water level against streamflow during 1978 to 2002, and from 2003 onwards. This is probably due to expansion of river area or increase in river sedimentation from active construction around Kajang station.

Malaysia has tropical and equatorial climate, which explains high temperature, humidity, and heavy rainfall throughout the year. Malaysia is among country experiencing no severe natural disasters like typhoons and earthquake, however, this country often faced with floods issue. It is caused by several factors such as high intensity or duration of precipitation, seasonal monsoon and improper drainage system is some places. The rainfall distribution is affected by the increase in temperature, which directly affect the evapotranspiration and air moisture. Seasonal monsoon in Malaysia is divided into two which are Northeast monsoon and Southwest monsoon. The first brings more precipitation intensity in east coast of Malaysia, Sabah and Sarawak, while usually less rainfall at other location in Malaysia, known as wet season in Malaysia. The latter is known as dry season, but higher precipitation amount at Kuala Lumpur, Penang, and Langkawi.

Rapid industrialization and urbanization have led to change in land use in most areas in Malaysia. For example, in Kajang, new area such as Kajang 2 and Kajang Utama has been upgraded to residential area. Population growth in Kajang in line with the increase in the amount of infrastructure and facilities that need to be provided. Many new infrastructures have been built around the area. Other than that, public transportation, such as MRT has been added as accommodation for Kajang citizen. However, due to improper maintenance on the drainage system, flash floods in certain area had become severe problem. For example, flash flood in front of the new MRT stations, Jalan Reko had caused traffic problem since vehicles cannot passed through the road. This incident happens frequently especially during heavy rainfall in short period or sometimes during continuous light precipitation. Therefore, in line with the increase in facilities and population, effective management in terms of flood management and proper infrastructure development can help reduce the impact of disasters in Kajang area.

## Conclusion

This article presents the analysis of maximum streamflow by considering peak over threshold series instead of annual maximum. The proper threshold value must be established in order to extract partial duration series data. This study selected a threshold value of 49.5 m$^3$/s since it satisfies the average number of event per year suggested by previous research with a lambda of approximately 1.8. The number of events taken into consideration by using the proposed threshold is 63 data instead of 35 data if the annual maximum was considered in the computation. In order to ensure that the independence criteria is satisfied, the recurrence interval between each peak should be more than 8 days between one event and the next. The distribution with the best fit for the data is lognormal with three parameter distributions. The L-moment method is the most suitable for estimating the parameter of the distribution since it is able to deal with outliers. According to the analysis, it can be discovered that population growth along with climate change could leads to more flood events in the future. Among the approach that can be taken is to update the flood frequency analysis from time to time in order to reduce the impact caused by flooding. The analysis is important especially for hydrological planning, construction of infrastructure and urban planning. It is recommended to consider multivariate distribution for the computation of flood frequency analysis, where each different characteristic of overflow or flood can be captured. Flood frequency analysis can help relevant government agencies prepare for flooding events which may occur in the near future.

## ACKNOWLEDGEMENTS

The support provided by the Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia communities in this study is highly appreciated. The authors are grateful for the research fund provided by the Universiti Kebangsaan Malaysia and Ministry of Higher Education, Malaysia via GUP-2020-013 and FRGS/1/2016/STG06/UKM/02/1 grants, respectively.

## REFERENCES

Agilan, V. & Umamahesh, N.V. 2017. Non-stationary rainfall intensity-duration-frequency relationship: A comparison between annual maximum and partial duration series. *Water Resources Management* 31(6): 1825-1841.

Alahmadi, F.S., Abd Rahman, N. & Abdulrazzak, M. 2014. Evaluation of the best fit distribution for partial duration series of daily rainfall in Madinah, Western Saudi Arabia. *Proceedings of the International Association of Hydrological Sciences 364*. Göttingen: Copernicus Publications. pp. 159-163.

Anderson, T.W. & Darling, D.A. 1954. A test of goodness of fit. *Journal of the American Statistical Association* 49(268): 765-769.

Beguería, S. 2005. Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value. *Journal of Hydrology* 303(1-4): 215-230.

Bezak, N., Brilly, M. & Šraj, M. 2014. Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis. *Hydrological Sciences Journal* 59(5): 959-977.

Bílková, D. 2014. Alternative tools of statistical analysis: L-moments and TL-moments of probability distributions. *Pure and Applied Mathematics Journal* 3(2): 14-25.

Chang, K.B., Lai, S.H. & Othman, F. 2016. Comparison of annual maximum and partial duration series for derivation of rainfall intensity-duration-frequency relationships in Peninsular Malaysia. *Journal of Hydrologic Engineering* 21(1): 1-11.

Cheong, R.Y. & Gabda, D. 2018. Frequency analysis of annual maximum river flow by generalized extreme value distribution with Bayesian MCMC. *Journal of Computer Science & Computational Mathematics* 8(4): 77-81.

Claps, P. & Laio, F. 2003. Can continuous streamflow data support flood frequency analysis? An alternative to the partial duration series approach. *Water Resources Research* 39(8): 1-11.

Cunnane, C. 1979. A note on the Poisson assumption in partial duration series models. *Water Resources Research* 15(2): 489-494.

Department of Statistics Malaysia. 2019. Poket stats negeri – Selangor. https://www.dosm.gov.my/v1/index.php?r=column/cone&menu_id=dFc3aExhVktPbUpoZys1dWoyUWFPQT09.

Engeland, K., Wilson, D., Borsányi, P., Roald, L. & Holmqvist, E. 2018. Use of historical data in flood frequency analysis: A case study for four catchments in Norway. *Hydrology Research* 49(2): 466-486.

Franchini, M., Galeati, G. & Lolli, M. 2005. Analytical derivation of the flood frequency curve through partial duration series analysis and a probabilistic representation of the runoff coefficient. *Journal of Hydrology* 303(1-4): 1-15.

Gado, T. & Nguyen, V.T.V. 2016. Regional estimation of floods for ungauged sites using partial duration series and scaling approach. *Journal of Hydrologic Engineering* 21(12): 1-12.

Garba, H., Ismail, A. & Tsoho, U. 2013. Fitting probability distribution functions to discharge variability of Kaduna River. *International Journal of Modern Engineering Research* 3(5): 2848-2852.

Gharib, A., Davies, E.G.R., Goss, G.G. & Faramarzi, M. 2017. Assessment of the combined effects of threshold selection and parameter estimation of generalized Pareto distribution with applications to flood frequency analysis. *Water* 9(9): 692-708.

Ghasemi, A. & Zahediasl, S. 2012. Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism* 10(2): 486-489.

Jiang, S. & Kang, L. 2019. Flood frequency analysis for annual maximum streamflow using a non-stationary GEV model. *E3S Web of Conferences: International Symposium on Architecture Research Frontiers and Ecological Environment 79, 03022*. United Kingdom: EDP Sciences. pp. 1-5.

Karim, F., Hasan, M. & Marvanek, S. 2017. Evaluating annual maximum and partial duration series for estimating frequency of small magnitude floods. *Water* 9(7): 481-497.

Keast, D. & Ellison, J. 2013. Magnitude frequency analysis of small floods using the annual and partial series. *Water* 5(4): 1816-1829.

Khan, S.A., Hussain, I., Hussain, T., Faisal, F., Muhammad, Y.S. & Shoukry, A.M. 2017. Regional frequency analysis of extremes precipitation using L-moments and partial L-moments. *Advances in Meteorology* 2017: Article ID. 6954902.

Leščešen, I. & Dolinaj, D. 2019. Regional flood frequency analysis of the Pannonian Basin. *Water* 11(193): 1-15.

Madsen, H., Rasmussen, P.F. & Rosbjerg, D. 1997. Comparison of annual maximum series and partial duration series methods for modelling extreme hydrologic events. *Water Resources Research* 33(4): 747-757.

Makkonen, L. 2006. Plotting positions in extreme value analysis. *Journal of Applied Meteorology and Climatology* 45(2): 334-340.

Malamud, B.D. & Turcotte, D.L. 2006. The applicability of power-law frequency statistics to floods. *Journal of Hydrology* 322(1-4): 168-180.

Mierlus-Mazilu, I. 2010. On generalized Pareto distributions. *Romanian Journal of Economic Forecasting* 13(1): 107-117.

Özonur, D., Gökpinar, E., Gökpinar, F., Bayrak, H. & Gül, H.H. 2013. Comparison of the goodness of fit tests for the geometric distribution. *Gazi University Journal of Science* 26(3): 369-375.

Pettitt, A.N. & Stephens, M.A. 1977. The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics* 19(2): 205-210.

Pham, H.X., Shamseldin, A.Y. & Melville, B.W. 2014. Statistical properties of partial duration series and its implication on regional frequency analysis. *Journal of Hydrologic Engineering* 19(7): 1471-1480.

Scarrott, C. & Macdonald, A. 2012. A review of extreme value threshold estimation and uncertainty quantification. *Revstat Statistical Journal* 10(1): 33-60.

Schlögl, M. & Laaha, G. 2017. Extreme weather exposure identification for road networks - A comparative assessment of statistical methods. *Natural Hazards and Earth System Sciences* 17(4): 515-531.

Singla, N., Jain, K. & Sharma, S.K. 2016. Goodness of fit tests and power comparisons for weighted gamma distribution. *Revstat Statistical Journal* 14(1): 29-48.

Srinivasa Murthy, D., Aruna Jyothy, S. & Mallikarjuna, P. 2017. Probability distributions of annual maximum daily streamflows using L-moments - A case study. *International Journal of Civil Engineering and Technology* 8(6): 290-302.

Syed Hussain, T.P.R. & Ismail, H. 2013. Flood frequency analysis of Kelantan River basin, Malaysia. *World Applied Sciences Journal* 28(12): 1989-1995.

Tallaksen, L. & Hewa, G. 2008. Extreme value analysis. In *Manual on Low-flow Estimation and Prediction*, edited by Gustard, A. & Demuth, S. Geneva, Switzerland: World Meteorological Organization. pp. 57-70.

Tekolla, A.W. 2010. Rainfall and flood frequency analysis in Pahang River basin, Malaysia. Master of Science Thesis. Lund University, Lund, Sweden (Unpublished).

Ummi Nadiah, A., Shabri, A. & Zakaria, Z.A. 2013. An analysis of annual maximum streamflows in Terengganu, Malaysia using TL-moments approach. *Theoretical and Applied Climatology* 111(3-4): 649-663.

van Westen, C.J. & Jetten, V. 2015. Magnitude-frequency analysis. In *Caribbean Handbook on Risk Information Management*, edited by van Westen, C.J., Jetten, V., Sliuzas, R., Brussel, M., Alkema, D., Van den Bout, B. & Hazarika, M. Washington D.C., United States: World Bank.

Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: fir@ukm.edu.my

Received: 14 August 2020
Accepted: 30 November 2020