Parameter Estimation of Extreme Rainfall Distribution in Johor using Bayesian Markov Chain Monte Carlo

N Nazmi¹, H F S Saipol¹, F Yusof², S A Mazlan¹, M A A Rahman¹, N A Nordin¹, N Johari¹ and S A A Aziz¹

¹ Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
² Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

*Corresponding author e-mail: nurhazimah@utm.my

Abstract. Heavy rainfall and the associated floods occur frequently in the Malaysia and have caused huge economic losses as well as massive impact on agriculture and people. As a consequence, it is necessary to understand the distribution of extreme rainfall in order to improve the managements in a country. Thus, the aim of this paper is to determine the best method to estimate parameters of Generalized Extreme Value (GEV) distribution that represent the annual maximum series (AMS) data of daily rainfall by using method of moments (MOM), maximum likelihood estimators (MLE) and Bayesian Markov Chain Monte Carlo (MCMC). The daily precipitation rainfall amount of 12 rain gauge stations in Johor from year 1975 to 2008 were used and the AMS data of each year were fitted with GEV distribution. Based on goodness-of-fit tests, namely Relative Root Mean Square Error (RRMSE) and Relative Absolute Square Error (RASE), the performances of three parameters of GEV distribution estimated by MOM, MLE and Bayesian MCMC were compared for each station. The results indicated that Bayesian MCMC method was performed better than MOM and MLE method in estimating the parameters of GEV distribution.

1. Introduction

As Malaysia is located in a tropical climate zone, extreme rainfall is estimated to take place regularly every year resulting from local tropical wet season. Extreme rainfall event is often related with the changes of climate, which may be followed by a series of natural disasters like flash flood and landslides. The rapid changes in the climate has steadily risen the number of extreme floods in Malaysia, particularly in Johor. Johor experienced a tropical rainforest climate with monsoon rains from November to February from the South China Sea. The mean annual rainfall is 1778 mm with average temperatures ranging between 25.5°C and 27.8°C. The worst history of flooding was recorded in 1967 as it effects 250,000 people in urban and rural areas of West Malaysia [1]. Then, continuous heavy downpour had occurred in Johor on December 2006, that led to the 2006/2007 Malaysia floods with water levels as high as 10 feet (3.0 m) above ground level in Muar, Kota Tinggi and Segamat. In 2014, the flood disaster hit several states includes Kelantan, Perlis, Kedah and Johor as well as Sarawak and Sabah can be classified as worst as the floods in 1967 as 200,000 people were affected with 21 people were killed [2]. Unpredictable extreme rainfall events phenomenon that increase in frequency lately has brought damages costing millions of Malaysian ringgit.
Extreme rainfalls data need to be modelled by suitable statistical distributions while providing the best inferences for the patterns of extreme rainfall. There are many distributions that can be used in extreme rainfall analysis, such as Generalized Extreme Value (GEV) distribution, Generalized Pareto (GP) distribution, Generalized Logistic (GLO), Exponential and Gamma distribution. In an overview on rainfall modelling in Malaysia, Shamsudin et al. [3] found GP distribution was the most suitable distribution of rainfall intensity using hourly data than Exponential, Beta and Gamma distributions. Other than GP distributions, GEV, GLO and Wakeby distributions can represent the extreme rainfall [4,5]. Nevertheless, GEV distribution was the most appropriate distribution to describe the extreme rainfall compared with Gamma, Generalized Normal (GNO), GP, Gumbel, log Pearson Type III (PE3), Pearson Type III and Wakeby distributions [6]. Similarly, Ibrahim et al. [7] and Ismail et al. [8] showed that GEV distribution was best fit distribution than other distributions.

For extreme distributions, three parameters namely shape, scale and location can be estimated by using several methods. Adetan et al. [9] used Method of Moments (MOM) and Maximum Likelihood Estimator (MLE) to analyse the lognormal raindrop size distribution. For distribution with multiple parameters, MOM was usually more tractable than MLE. This is supported by Hosking et al. [10] as parameter estimation via MOM method produce better estimation than MLE. Nonetheless, Coles et al. [11] stated that parameter estimation via MLE method was the best method as its all-around utility and adaptability to model change. This statement had proven with the performance of MLE in estimate the parameter for GEV and GP distributions [12,13]. Even so, the method incapable to estimate the parameters for small sample [14]. Eli et al. [15] improved the performance of parameter estimation by adopting Bayesian MCMC approach. Bayesian approach capable to measure the uncertainties in extreme rainfall event and does not rely on asymptotic theory as MLE method [16,17]. Therefore, the aim of this study is to evaluate the performance of MOM, MLE and Bayesian MCMC in estimating the three parameters of GEV distributions. The annual maximum series (AMS) of daily precipitation rainfalls in Johor will be fitted with GEV distribution. Then, the performance of parameters estimated by MOM, MLE and Bayesian MCMC is determined by conducting goodness-of-fit tests.

2. Methodology

2.1 Data sources

A datasets of the daily precipitation rainfalls amount of 12 rain gauge stations in Johor for 34 years (1975 to 2008) were obtained from Department of Irrigation and Drainage Malaysia. Table 1 shows the details locations of each station.

| Station | Location                  | Latitude    | Longitude   |
|---------|---------------------------|-------------|-------------|
| S01     | Ladang Getah Kukup Pontian| 1.3167° N   | 103.4500° E |
| S02     | Ladang Benut Renggam      | 1.8831° N   | 103.3946° E |
| S03     | Stor JPS JB               | 1.2634° N   | 103.7547° E |
| S04     | Pintu Kawalan Tampok Batu Pahat | 1.6499° N | 103.2000° E |
| S05     | Senai                     | 1.6203° N   | 103.6563° E |
| S06     | Sek Men Bkt Besar         | 1.7835° N   | 103.6819° E |
| S07     | Sek Men Inggeris Batu Pahat | 1.8500° N | 102.9333° E |
| S08     | Pintu Kawalan Sembong     | 2.0000° N   | 193.1667° E |
| S09     | Pintu Kawalan Separap     | 1.9167° N   | 102.8667° E |
| S10     | Kluang                    | 2.0336° N   | 103.3194° E |
| S11     | Tangkak                   | 2.2667° N   | 102.5500° E |
The data selection for extreme rainfall used in this study was annual maximum series (AMS). AMS methods involve selection of the highest value of rainfall observed each year. Figure 1 shows the scatter plot / time series plot of annual maximum amount of daily rainfall 12 stations in Johor for 34 years. From the figure, the highest annual maximum rainfall in Johor was 433.4 mm which occurred in year 2004 and year 2002 was the lowest annual maximum rainfall with 117.4 mm. The average annual maximum rainfall in Johor was 113.0478 mm.

![Scatter plot / time series plot of annual maximum amount of daily rainfall.](image)

2.2 Probability Distribution

By the extreme value theorem, the GEV distribution is the limit distribution of properly normalized maxima of a sequence of independent and identically distributed random variables. Being that, the GEV distribution is used as an approximation to model the maxima of long (finite) sequences of random variables.

Let $X_1, \ldots, X_n$ denoted the independent annual maximum rainfall distribution, the probability density function of GEV is

$$f(k, \xi, \alpha)(x) = \frac{1}{\alpha} \left[ 1 + k \left( \frac{x - \xi}{\alpha} \right) \right]^{-1+k} \exp \left\{ - \left[ 1 + k \left( \frac{x - \xi}{\alpha} \right) \right]^{-1+k} \right\}$$

where $\xi$ is the location, $\alpha$ is the scale and $k$ is the shape parameters with parameter space $-\infty < \xi < \infty$, $\alpha > 0$ and $-\infty < k < \infty$, respectively.

2.3 Parameter Estimation Method

2.3.1 Method of Moment (MOM). Three parameters of GEV distribution were calculated by using equations (2), (3), and (4) by substituted the value of sample mean, standard deviation and skewness.

$$k = \left\{ \begin{array}{ll}
0.0087\gamma^3 + 0.0582\gamma^2 - 0.32\gamma + 0.2778, & -0.7 \leq \gamma \leq 1.15 \\
-0.31158\left[1 - \exp\left(-0.4556(\gamma - 0.97134)\right)\right], & 1.15 \leq \gamma
\end{array} \right.$$
The variances of the proposal distribution, controlling the mixing properties of the output chain and the convergence rate to its stationary distribution was set as 0.01 for $\hat{\alpha}$, 0.001 for $\alpha$ and 0.02 for $k$.

$$\frac{\alpha}{\sigma} = -0.1429k^3 - 0.7631k^2 + 1.0145k + 0.7795, \quad -0.5 \leq k \leq 0.5$$ (3)

$$\frac{\varepsilon - \mu}{\sigma} = \begin{cases} 0.514075k_{1.33199} - 0.44901, & 0.01 \leq k \leq 0.5 \\
19.357k^4 + 13.749k^3 + 0.5212k - 0.4427, & -0.5 \leq k \leq 0.01 
\end{cases}$$ (4)

where $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\gamma}$ are the sample mean, standard deviation and skewness respectively.

2.3.2 Maximum Likelihood Estimators (MLE). The MLE of $k$, $\sigma$, and $\mu$ after differentiating with respect to each parameter of distribution as follows;

$$\frac{\partial l}{\partial k} = -\frac{1}{k^2} \sum_{i=1}^{n} \left\{ \ln(Z_i) \left[ 1-k-(Z_i)^\frac{\iota}{\kappa} \right] + \frac{1-k-(Z_i)^\frac{\iota}{\kappa}}{Z_i} k \left( \frac{X_i - \mu}{\sigma} \right) \right\} = 0$$ (5)

$$\frac{\partial l}{\partial \alpha} = -\frac{1}{\sigma} \sum_{i=1}^{n} \left[ \frac{1-k-(Z_i)^\frac{\iota}{\kappa}}{Z_i} \right] = 0$$ (6)

$$\frac{\partial l}{\partial \varepsilon} = -\frac{n}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^{n} \left\{ \frac{1-k-(Z_i)^\frac{\iota}{\kappa}}{Z_i} k \left( \frac{X_i - \mu}{\sigma} \right) \right\} = 0$$ (7)

The MLEs ($\hat{\epsilon}, \hat{\alpha}, \hat{k}$) were obtained by maximising equation (5), (6) and (7) with respect to the vector ($\epsilon, \alpha, k$) for $i = 1, \ldots, n$ using numerical techniques. In this study, three parameters of GEV distribution had es using the ismev package within the free statistical environment R software.

2.3.3 Bayesian Markov Chain Monte Carlo (MCMC). Given a distribution of interest, $\pi$, a reversible Markov chain, which has this distribution as its stationary distribution can be constructed. Simulating from such a Markov chain will result in values from the distribution of interest. The procedure was to construct a transition kernel $p(\theta, \phi)$ such that the equilibrium distribution of the chain was $\pi$. This transition kernel was made up of two elements which was an arbitrary transition kernel $q(\theta | \phi)$, also known as the proposal distribution and an acceptance probability $a(\theta | \phi)$.

$$a(\theta | \phi) = \min \left\{ 1, \frac{\pi(\phi)q(\phi | \theta)}{\pi(\theta)q(\theta | \phi)} \right\}$$ (8)

The particular type of MCMC method used in this study was based on simulation of a random walk chain. The proposed value $\phi$ at point $j$ is $\phi = \theta^{(j-1)} + w_j$. The $w_j$ were identically and independent distributed random variables and have density $f(\cdot)$. Supposing $f(\cdot)$ was easy to simulate from, an innovation, $w_j$, can be simulated. The candidate point was then set to $\phi = \theta^{(j-1)} + w_j$ and the transition kernel had been given by $q(\theta | \phi) = f(\phi - 0)$. This was then used to calculate the acceptance probability. The variance of the innovation affects the acceptance probability. If the variance was too low most proposals would be accepted, resulting in very slow convergence, and if it was too high very few would be accepted and the moves in the chain will often be large. The calculations had been done by using evdbayes package within the free statistical environment R software. The estimated value of $\hat{\varepsilon}, \hat{\alpha}$, and $\hat{k}$ by using MOM method were used as the initial value in evdbayes packages for MCMC simulation.
1000 iterations of the posterior density with the Metropolis-Hastings algorithm had been determined. If the convergence of the posterior density not converged to the limiting distribution, tune the variances of the proposal distributions in order to have reasonably good acceptance rates were between 10% to 40%. Else, modify the initial values until achieve the good acceptance rates. For example, for station S01, the simulated valued for three parameters of GEV were found to converge to the limiting distribution as shown in Figure 2, suggesting that no obvious tendencies and periodicities. Thus, the variances for non-informative priors had been chosen to be large enough in order to create flat priors.

![Figure 2: Graph plot (a) location, (b) scale and (c) shape parameters of GEV parameters for 1000 iterations](image)

2.4 Goodness-of-fit tests

The performance between MOM, MLE and Bayesian MCMC to estimate the GEV distribution are analysed based on Relative root mean squared error (RRMSE) and relative absolute square error (RASE). These formulas are used to measure the discrepancy between observed and estimated values under distributions given as:

\[
RRMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_{i,n} - \hat{Q}(F_i)}{Y_{i,n}} \right)^2}
\]

(9)

\[
RASE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_{i,n} - \hat{Q}(F_i)}{Y_{i,n}} \right|
\]

(10)
where \( n \) is the random samples of size, \( Y_i \) is the observed values and \( \hat{Q}(F_i) \) is estimated values for the \( i \)th order statistics of a random sample of size \( n \). The smallest values of RRMSE and RASE indicate the best method to fit the distribution.

3. Results and Discussion

3.1 Parameters Estimation

The MOM, MLE and Bayesian MCMC method were used to compare the performance in estimate the parameters of GEV distribution. For MOM method, three parameters of GEV distribution that were calculated by using equations (2), (3) and (4) by substituted the value of sample mean, standard deviation and skewness. Meanwhile, the three parameters of MLE and Bayesian MCMC were obtained by using the ismev package and evdbayes package of R software, respectively. Table 2 shows the parameters estimated value of \( \varepsilon \), \( \alpha \) and \( k \) representing the location, shape and scale parameters of GEV distribution for each method, respectively.

| Station | \( \varepsilon \)  
|---------|------------------|------------------|------------------|
|         | \( \alpha \)  
|---------|------------------|------------------|------------------|
|         | \( k \)  
|---------|------------------|------------------|------------------|
| S01     | 80.2404          |
|         | 14.6287          |
|         | -0.0040          |
| S02     | 81.8856          |
|         | 25.2609          |
|         | 0.0973           |
| S03     | 96.1237          |
|         | 25.7437          |
|         | 0.1991           |
| S04     | 93.0223          |
|         | 31.5881          |
|         | 0.0852           |
| S05     | 94.0156          |
|         | 27.7122          |
|         | 0.4023           |
| S06     | 91.7740          |
|         | 23.0340          |
|         | 0.3636           |
| S07     | 84.0615          |
|         | 30.508           |
|         | 0.0852           |
| S08     | 81.8914          |
|         | 26.7675          |
|         | 0.1815           |
| S09     | 73.4046          |
|         | 23.4974          |
|         | -0.1966          |
| S10     | 92.1264          |
|         | 27.4194          |
|         | 0.4284           |
| S11     | 79.9307          |
|         | 14.5155          |
|         | 0.0061           |
| S12     | 149.8015         |
|         | 51.8051          |
|         | 0.2095           |

For MOM method, Station S09 had the lowest location parameter value, 73.4046 and Station S12 had the highest with 149.8015. In addition, the value of scale parameter was between 14.5155 to 51.8051. Station S01 had the minimum value of scale parameter while Station S12 had the maximum value of scale parameter. As the value of location parameter bigger, the value of scale parameter also bigger.

For MLE method, the value of location parameter was between 79.566 to 156.05. The minimum value of \( \varepsilon \), 79.566 was in Station S11, while the maximum value of \( \varepsilon \), 156.05 was in Station S12. Station S01 had the lowest value of \( \alpha \), 13.909 and the station S12 had the highest value of \( \alpha \), 62.981. The value of shape parameter, \( k \) mostly was less than zero. Station S04, S07, S09 and S12 were the only stations with non-negative value of \( k \).

This is contrary to the findings of Bayesian MCMC method, the lowest value of location parameter was Station S11 and the highest was Station S12. For scale parameter, Station S11 had the minimum value with 14.2616374 and Station S12 had the maximum value with 57.6173102. The value of shape parameter was in range 0.003766 to 0.1404307.
3.2 Goodness-of-fit tests
The performance between MOM, MLE and Bayesian MCMC methods in estimating the three parameters of GEV distribution were evaluated. Two goodness-of-fit tests, RRMSE and RASE were used in this study and the results were shown in Table 3 and 4, respectively.

Table 3. RRMSE for each station of MOM, MLE and Bayesian MCMC techniques

| Method of estimation | Station | MOM  | MLE  | Bayesian |
|----------------------|---------|------|------|----------|
|                      | S01     | 0.256902 | 0.229621 | 0.246127 |
|                      | S02     | 0.501242 | 0.551212 | 0.535344 |
|                      | S03     | 0.306121 | 0.210266 | 0.200859 |
|                      | S04     | 0.46232 | 0.41378 | 0.427525 |
|                      | S05     | 0.465458 | 0.460676 | 0.460631 |
|                      | S06     | 0.358596 | 0.375547 | 0.349043 |
|                      | S07     | 0.373299 | 0.343539 | 0.319048 |
|                      | S08     | 0.453333 | 0.451499 | 0.446852 |
|                      | S09     | 0.287355 | 0.296446 | 0.280579 |
|                      | S10     | 0.440113 | 0.438457 | 0.439653 |
|                      | S11     | 0.419095 | 0.418318 | 0.419139 |
|                      | S12     | 0.895192 | 0.946235 | 0.921786 |

Referring to Table 3, MOM method gives the smallest value of RRMSE for Station S02 with 0.501242 while MLE method give the smallest value of RRMSE for Station S01, S04, S10 and S11 with 0.229621, 0.41378, 0.438457 and 0.418318 respectively. It can be seen that Bayesian MCMC method had the smallest value of RRMSE for station S03, S05, S06, S07, S08, S09 and S12 with 0.200859, 0.460631, 0.349043, 0.319048, 0.446852, 0.280579 and 0.921786 respectively.

Table 4. RASE of each station for MOM, MLE and Bayesian MCMC techniques

| Method of estimation | Station | MOM  | MLE  | Bayesian |
|----------------------|---------|------|------|----------|
|                      | S01     | 0.192852 | 0.17016 | 0.182527 |
|                      | S02     | 0.387559 | 0.479713 | 0.424989 |
|                      | S03     | 0.459326 | 0.413477 | 0.336432 |
|                      | S04     | 0.349278 | 0.331559 | 0.336325 |
|                      | S05     | 0.398912 | 0.396061 | 0.343142 |
|                      | S06     | 0.347624 | 0.420437 | 0.300725 |
|                      | S07     | 0.287221 | 0.367409 | 0.215896 |
|                      | S08     | 0.363576 | 0.363189 | 0.287255 |
|                      | S09     | 0.337847 | 0.347842 | 0.316543 |
|                      | S10     | 0.382431 | 0.377048 | 0.37692 |
|                      | S11     | 0.301373 | 0.301143 | 0.309687 |
|                      | S12     | 0.919324 | 0.943251 | 0.893338 |

Based on Table 4, Station S01, S04, S10 and S11 had the smallest value of RASE with 0.17016, 0.336432, 0.300725 and 0.309687 respectively for MLE method. MOM method only had the smallest value of RASE for station S02 while Bayesian MCMC method had the smallest value of RASE for the rest of stations.
From Table 3 and 4, this study recommended MLE method than MOM to estimate the parameters of GEV distribution as MLE method gained a small difference between the observed and estimated. This result was contrast with Adetan et al. [9] and Hosking [10] which declared that MOM method give the best fit compared with the MLE method. Van et. al [18] stated that MOM was a very simple approach but may result in increased sampling errors due to the squaring of observations but MLE method was the best method due to of its all-around utility and adaptability to model change [11] Coles (2001). Yet, Bayesian MCMC method was the best method due to its small value of RRMSE and RASE followed by MLE and MOM. This result in line with the study conducted by Eli et al. [15] and prove that Bayesian MCMC method does not rely on asymptotic theory and could measure the uncertainties in extreme rainfall event [16,17].

4. Conclusion
The best method of parameter estimation for GEV distribution had been explored in this study. The yearly AMS of rainfall data in Johor for 34 years was used to fit GEV distribution. Three parameters namely location, shape and scale of GEV distribution were estimated using MOM, MLE and Bayesian MCMC method. Based on goodness-of-fit tests, MLE method gained lower value of RRMSE and RASE than MOM method. It can be concluding that MLE method fairly better agreement with the measured data than MOM method. However, Bayesian MCMC method produced small value of RRMSE and RASE than MLE method. The result revealed Bayesian MCMC gained the smallest value of RRMSE and RASE for seven stations followed by MLE (four stations) and MOM (one station). Although extreme data are limited in nature, Bayesian inferences have the ability to incorporate other source of information via prior distribution. In addition, Bayesian MCMC method offers a way of dealing with information conceptually different from all other statistical methods which it provides a method in which observations are used to update estimates of the unknown parameters of a statistical model. Another approach which would benefit from future research is applying peak over thresholds series (POT) approach (also known as partial duration series, PDS) instead of AMS approach or change the posterior distribution for Bayesian MCMC for $\epsilon, \alpha$ and $k$.

5. References
[1] Ishak N S, Mohamed Dali A and Abdul Razak M R 2014. Perspektif: Jurnal Sains Sosial Dan Kemanusiaan, 3 54
[2] Buslima F S, Omar R C, Jamaluddin T A, and Taha H 2018. International Journal of Engineering and Technology(UAE), 7 760
[3] Shamsudin S and Aris A 2010 American Journal of Environmental Sciences 6 238
[4] Hassan Z and Shamsudin S and Harun S 2015 Jurnal Teknologi (Sciences and Engineering). 75 3058
[5] Shabri A B, Zalina M D, and Ariff N M 2011 Theor. Appl. Climatol. 104 561
[6] Zalina M D, Desa M N, Nguyen V T V and Hashim M K 2002 Water Science and Technology 45 63
[7] Ibrahim I 2004 Identification of a probability distribution for extreme rainfall series in East Malaysia. Thesis. Universiti Teknologi Tun Hussein Onn
[8] Ismail A Zuhdi, Yusop Z and Yusof Z 2015 J. Teknologi 74 123
[9] Adetan O, and Afullo T J 2012 Proceedings of Southern Africa Telecommunications Networks Applications Conference 6 1
[10] Hosking JRM 1990 Journal of Royal Statistical Society 52 105
[11] Coles S 2001 An Introduction to Statistical Modeling of Extreme Values Springer
[12] Onwuegbuche F, Kenyatta A, Affognon S B, Enock, E, and Akinade M 2019 International Journal of Statistics and Probability. 8 85
[13] Nashwan M, Ismail T, and Ahmed K. 2019 Journal of Sustainability Science and Management. 14 17
[14] Rao A R and Hamed H K 2000 Flood Frequency Analysis Boca Raton
[15] Eli A, Shaffie M and Zin W Z W 2012 Sains Malaysia 41 1403
[16] Lazoglou G, and Anagnostopoulou C. 2017 Proceedings. 1 110
[17] Jung I S and Yongku K 2016 Communications in Statistics - Theory and Methods 3 1
[18] Van d M S, Steven D and Pretorius M 2013 Bayesian Extreme Value Analysis of Stock Exchange Data. Department of Mathematical Statistics and Actuarial Science, University of the Free State.

Acknowledgments
The work presented in this study is funded by Universiti Teknologi Malaysia under Encouragement Research university grant. VOTE NO: 18J26.