Frequency Analysis of Annual Maximum Flood for Segamat River

Noor Suraya Romali[1, 2, *] and *Zulkifli* Yusop[1]

[1]Faculty of Civil Engineering, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
[2]Faculty of Civil Engineering and Earth Resources, Universiti Malaysia Pahang, Gambang, Malaysia

Abstract. Several major floods had occurred in the last few decades in Segamat, causing extensive damage to properties and harm local community. For the purpose of flood risk management, this study estimated the average recurrence interval (ARI) and peak flows associated with the ARI based on the distributions of annual peak flow. The flood frequency analysis was performed for flood series data of Segamat River, at Sg. Segamat gauging station (Site 2528414) for the years 1960 – 2011. Five distribution models, namely Generalized Pareto, Generalized Extreme Value, Log-Pearson 3, Log-Normal (3P) and Weibull (3P) were tested for the 52 years flood series data. The goodness of fit test (GOF) of Kolmogorov-Smirnov (K-S) was used to evaluate and estimate the best-fitted distribution. The results obtained using Generalized Pareto distribution provided the best fit, followed by Generalized Extreme Value, Log-Pearson 3, Log-Normal (3P), and the least for Weibull (3P). The estimated peak flows for Segamat River for 50, 100 and 200 ARIs are 1362.2 m^3/s, 1914 m^3/s and 2642 m^3/s respectively. Results can be useful as a reference for further/future flood risk assessment works in the study area.

1 Introduction

Nowadays, a higher frequency of extreme rainfall is expected to occur more frequently due to the climate change phenomenon [1]. Flood had causes tremendous damages to properties [2] and may lead to the loss of human life [3]. In Malaysia, flood occur annually and affected an approximate area of 29,720 km^2, involving more than 4.915 million people and causing up to RM 915 million damage yearly [4]. Efforts have been made by researchers and local authorities to reduce the risk and mitigate the impact of flooding. Flood modelling has been used in flood mitigation to estimate floods associated with return periods of interest, which is called design flood. Design flood is essential in the flood plain management, development and planning controls, and in the design of hydraulic structures [5]. In Malaysia, a 100 year ARI has been used as a practice for designing hydraulic structures. However, recently this standard has been extended to 200 years return period [6].

* Corresponding author: suraya@ump.edu.my

Flood frequency analysis is the most direct method for determining design flood [5]. The purpose of flood frequency analysis is to estimate the return period associated with a given flood magnitude. It shows the relationship between the magnitude of an event and the frequency with which that event is exceeded [7]. Furthermore, the catchment characteristics, water availability and possible extreme hydrological conditions like floods and droughts at various locations of any river system may be illustrated through the flood frequency analysis [8]. The flood frequency analysis primarily uses observed annual maximum flood data at a gauging station to estimate flood magnitude [9]. A long period of recorded flood data is required for this purpose, and a statistical distribution method is needed [6].

Numerous probability distribution models have been used in flood frequency studies, such as log-Pearson 3 [3, 5], Generalized Extreme Value [2, 5], Generalized Pareto [8], log-normal (3P) [8] and Weibull [7]. The selection of appropriate probability distribution and associated parameter procedure is important in flood frequency analysis to avoid under- or over-estimation of design floods [5]. Hence, this paper is aimed at determining the most appropriate probability distribution model that could provide the hydrological frequency i.e. ARI and peak flows of the study area.

2 Model formulations

2.1 Study area and data

Segamat River is located in the southern part of Peninsular Malaysia at 102° 49" East and 2° 30.5' North, with a length of 23 km. The average width of Segamat River is 40 m and is 14 m above sea level. About 70% of the Segamat river watershed is classified as hilly with elevation up to 1000 meters above the mean sea level (msl), and the rest (30%) is undulating with little swamp. Segamat River is a tributary of Sungai Muar that flows through the Segamat town. The data used in the flood frequency analysis were 52 annual maximum flows of Sg. Segamat gauging station for the years 1960 until 2011. These data were provided by the Department of Irrigation and Drainage of Malaysia (DID). The location of Sg. Segamat gauging station (Site 2528414) is shown in Fig. 1.

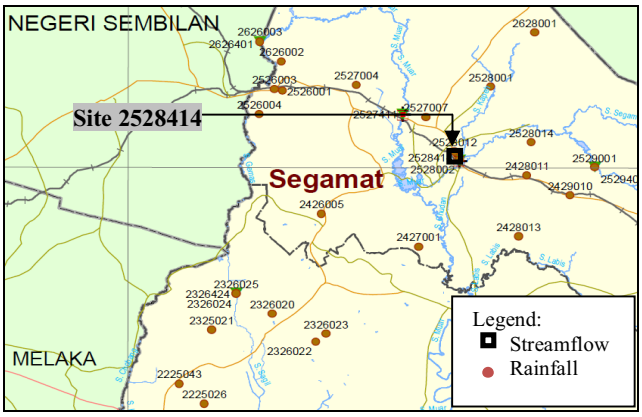


Fig. 1. The location of Sg. Segamat gauging station (Site 2528414)[10].

2.2 Flood Frequency Models

The purpose of flood frequency analysis is to extract information from a flow record to estimate the relationship between flows and return periods. Three different models i.e. annual maximum series (AM) model, partial duration series (PD) or peaks over the threshold (POT) model, and time series (TS) model could be considered for this purpose [11]. A numbers of probability distribution such as Generalized Extreme Value (GEV), Log-Pearson, Log-Normal, Gumbell, Weibull, and Generalized Pareto had been utilised in flood frequency studies worldwide. In order to determine whether the distribution model could fit the data properly, Goodness-of-Fit test such as Kolmogorov-Smirinov, Anderson-Darling, and Chi squared tests can be used [12].

Singo et al. [13] had adopted AM model in their study. 50 years annual maximum flow data from 8 stations were used to analyse flood frequencies in the Luvuvhu River Catchment in Limpopo province, South Africa. The result showed that Gumbel and Log Pearson type III distributions provided the best fit in the extreme value analysis. Rahman et al. [5] used a large annual maximum flood data set to select best probability distributions for at-site flood frequency analysis in Australia. They identified Log Pearson type III, GEV and Generalized Pareto as the top-three best fitted distributions. A frequency analysis study using POT flood data was conducted by Guru and Jha [8]. Comparison was made with another analysis using AM flood data where Generalized Pareto and Log Normal (3P) showed the best result for AM and PO flood data series respectively.

Study by Mohd Daud et al. [14] found that GEV was the most suitable distribution for annual maximum rainfall in Peninsular Malaysia. The analysis was done using annual maximum rainfall series for several time resolutions obtained from 17 recording rain gauges that are located all over the peninsular. Meanwhile, GEV and Generalized Logistic distribution are identified as the best fitted distribution for frequency analysis using annual flood data from more than 23 gauged river basin in Sarawak, Malaysia [15].

3 Methodology

Fig. 2 shows the general methodology adopted in this study. The first stage is the estimation of annual maximum stream flow that based from the flow historical data for certain years [6]. Then 52 selected flow data from the year 1960 until 2011 were analyzed using EasyFit Software to determine the distribution models that can best fit the data. EasyFit Software is a data analyzer and simulation software which is capable to fit and simulate statistical distributions with sample data, choose the best model, and then use the obtained result of analysis to provide better decisions [11].

3.1 Probability distributions, parameter estimation methods and Goodness of fit (GOF)

In this study, the annual maximum series (AM) model was adopted where only the peak flow in each water year is considered. Five different probability distributions i.e. Generalized Pareto, Generalized Extreme Value (GEV), Log-Pearson 3, Log-Normal (3P) and Weibull (3P) are considered for comparison. The selections of the distribution models are based on the previous studies where most of these have been used and recommended in various countries.

In EasyFit software, different parameter estimation methods are used for different probability distributions. Table 1 list the method uses for the five selected probability distributions. Method of L-moments is used for Generalized Pareto and GEV. Whereas, maximum likelihood method is used for Log-Normal (3P) and Weibull (3P), and method of

moments is used for Log-Pearson 3. The most commonly adopted GOF tests are Kolmogorov-Smirinov (KS), Anderson-Darling, and Chi squared test. However, KS test is found to be the most used GOF test [12]. Hence KS test is applied in this study to determine whether the distribution is fitted to the data or not. K-S at 5% level of significant ($p<0.05$) was used to define the best fit ranking [6].

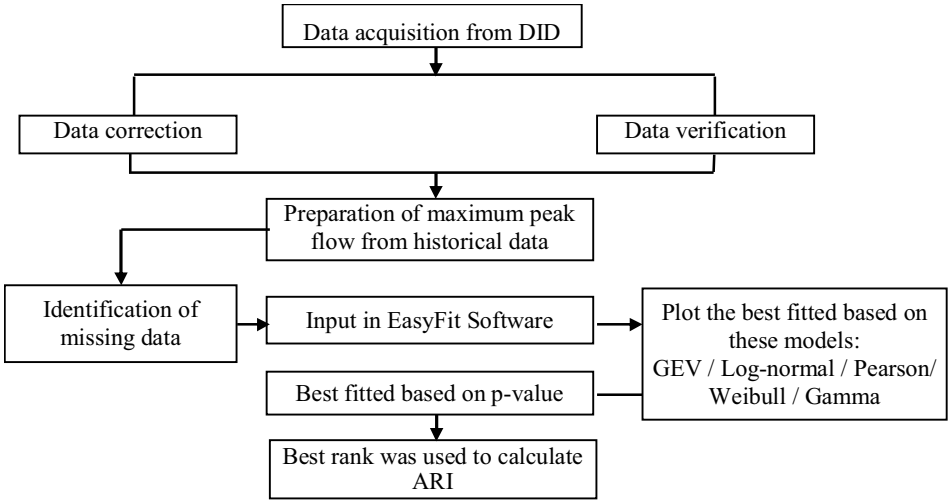


Fig. 2. The general methodology for flood frequency analysis and the determination of Average Recurrence Interval (ARI)

Table 1. Parameter estimation methods applied in this study.

Probability distribution	Parameter estimation method
Generalized Extreme Value, Generalized Pareto	Method of L-moments
Log-Pearson 3	Method of moments
Lognormal (3P), Weibull (3P)	Maximum likelihood method

3.2 Quantile estimation of Generalized Pareto

After the parameters of a distribution are estimated, quantile estimates (X_T) which correspond to different return periods may be computed [11]. For the case of Generalized Pareto distribution, the distribution function $F = F(x)$ is given by Equation (1) [11]:

$$F(x) = 1 - \left[1 - \frac{k}{\alpha}(x - \varepsilon)\right]^{1/k} \quad (1)$$

Using the inverse form of Equation (1), $x = x(F)$ and $F = 1 - (1/T)$, the T-year quantile (X_T) for Generalized Pareto distribution is given by Equation (2):

$$X_T = \varepsilon + \frac{\alpha}{k}\left[1 - T^{-k}\right] \quad (2)$$

4 Results and discussion

The annual flood variation for the respective years is shown in Fig. 3. The highest flow was recorded in 1983 which is 1615.5 m^3/s, while the minimum flow of 3.3 m^3/s was recorded

in the year 1989. The average flow for the 52 years was 234.37 m^3/s. Four major flood events, labelled as 1, 2, 3 and 4 in Fig. 3 occurred in year 1969, 1979, 1983 and 2007 with peak flow more than 1000 m^3/s. A large flood had occurred in Segamat for years 2007 and 2011, which had caused tremendous damages and disruptions to local communities.

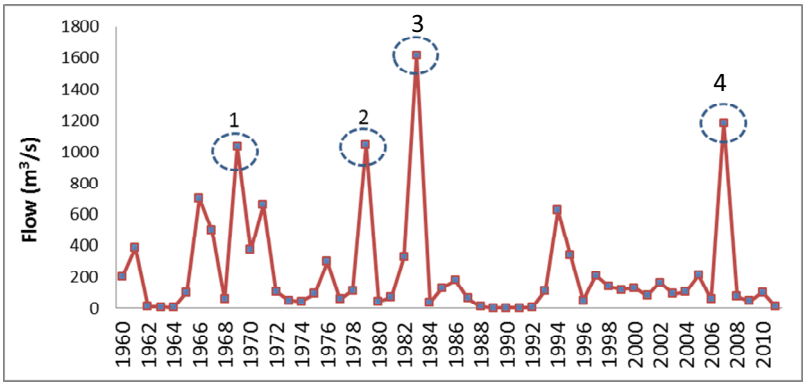


Fig. 3. The annual peak flow at Sg. Segamat gauging station from July 1960 to June 2011

4.1 Goodness of fit test result

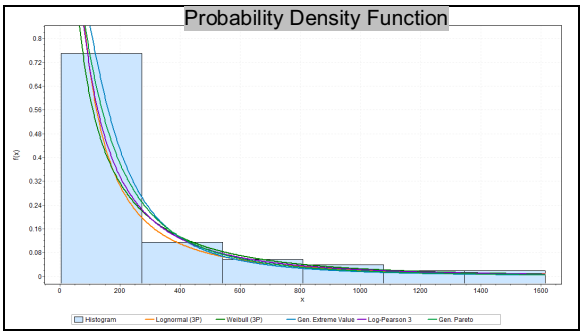
The best parameter estimates generated from the EasyFit Software for the five distribution models are displayed in Table 2. Parameters (α, k) represent shape parameters, while (σ, β) and (μ, γ) representing the continuous scale parameters and continuous location parameters respectively. Table 3 ranks the performance of various cumulative density based on the K-S GOF tests. Generalized Pareto shows the best performance, followed by Generalized Extreme Value, Log-Pearson 3, Log-Normal (3P), and the least for Weibull (3P). The ranking is based on the p-value. A p-value closer to one indicates a better-fit distribution. The highest p-value is 0.83277 for the Generalized Pareto and the lowest is 0.29474 for Weibull (3P).

Table 2. Fitting results for probability distribution of annual flood

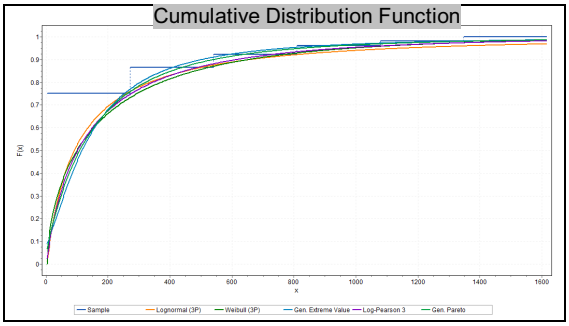
	Distribution	Parameters
1	Generalized Extreme Value	k=0.50376 σ=101.87 μ=75.491
2	Generalized Pareto	k=0.3988 σ=145.27 μ=-7.2636
3	Log-Pearson 3	α=23.428 β=-0.31913 γ=11.998
4	Lognormal (3P)	σ=1.5441 μ=4.512 γ=0.24503
5	Weibull (3P)	α=0.62873 β=172.5 γ=3.3

Table 3. Fitting results for probability distribution of annual flood

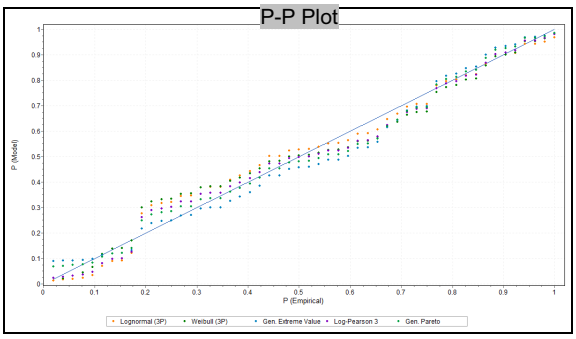
Distribution	Kolmogorov Smirnov	
	P	Rank
Generalized Pareto	0.83277	1
Generalized Extreme Value	0.67816	2
Log-Pearson 3	0.67300	3
Log-Normal (3P)	0.45227	4
Weibull (3P)	0.29474	5



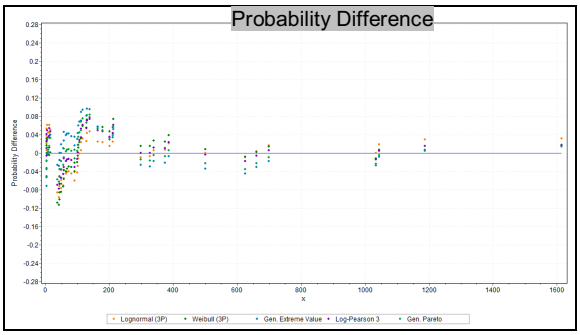
(a)



(b)



(c)



(d)

Fig. 4. a) Probability density function, b) Cumulative distribution function c) Probability-probability plot and d) Probability difference plot for the five distributions.

The Probability Density Functions (PDFs) for the five distribution models; Generalized Pareto (green), Generalized Extreme Value (blue), Log-Pearson 3 (purple), Log-Normal (3P) (orange) and Weibull (3P) (dark green) are shown in Fig. 4a. The Cumulative Distribution Function (CDF) in Fig. 4b shows the non-exceedance probability for a given magnitude. The P-P plot (Fig. 4c) is a graph of the empirical CDF values against the theoretical CDF values. The distribution that has the most number of points close to the line represents the best fitted distribution model. Through all the patterns shown in Fig. 4a to 4d, the best fitted is Generalized Pareto (GP) distribution. The second most chosen best fitting distribution is Generalized Extreme Value (GEV), followed by Log-Pearson 3, Log-Normal (3P), and the least for Weibull (3P). These result is differ from the finding by Mohd Daud et al. [14], which stated that GEV is the most suitable distribution for annual maximum rainfall in Peninsular Malaysia. However, it is in agreement with those obtained by Dan'azumi et al. [16] where in their study, GP is found to be the most suitable distribution for modelling the hourly rainfall intensity in Peninsular Malaysia. Furthermore, this findings are in accord with another study by Wan Zin et al. [17], indicating that GP is the most frequently selected fitting distribution of annual maximum rainfall in Peninsular Malaysia based on LQ-moment methods, together with Generalized Logistic distribution.

4.2 Peak Flow Estimation

Peak flows corresponding to return periods of 2, 5, 10, 50, 100 and 200 years were estimated using the best-fitted distribution models i.e. Generalized Pareto, as shown in Table 4. The estimated flows for 50, 100 and 200 years ARI are 1362.2 m^3/s, 1914 m^3/s and 2642.0 m^3/s respectively.

Table 4. Fitting results for probability distribution of annual flood

Return period, T (years)	Probability, P (%)	Flood discharge, Q (m^3/s)
2	50	108.7
5	20	320.6
10	10	540.9
25	4	943.5
50	2	1362.2
100	1	1914.2
200	0.5	2642.0

5 Conclusion

A total of five probability distributions namely Generalized Pareto, Generalized Extreme Value, Log-Pearson 3, Log-Normal (3P) and Weibull (3P) were tested using 52 annual flow data of Sg. Segamat gauging station to identify the best distribution model that fit the annual flood of Segamat River. This study found that Generalized Pareto distribution provided the best-fit, followed by Generalized Extreme Value, Log-Pearson 3, Log-Normal (3P), and the least for Weibull (3P). Peak flow of Segamat River for 50, 100 and 200 years ARI are estimated as 1362.2 m^3/s, 1914 m^3/s and 2642.0 m^3/s respectively. This information is useful for the flood risk management where the ARI and estimated flow values may be used to generate future flood risk mapping.

The authors would like to acknowledge Universiti Malaysia Pahang and Universiti Teknologi Malaysia for the financial support through research grant Institut Inovasi Strategik Johor (IISJ) vot

15H00 and Department of Irrigation and Drainage (DID) Malaysia for providing data and relevant information.

References

[1] M. Ren, B. Wang, Q. Liang and G. Fu, Classified real-time flood forecasting by coupling fuzzy clustering and neural network. Int. J. of Sediment Research, **25**(2), 134–148 (2010)

[2] L. Chang, C. Lin and M. Su, Application of geographic weighted regression to establish flood-damage functions reflecting spatial variation. Water SA, **34**(2)**,** 209–216 (2008)

[3] S. N. Jonkman, Global perspectives on loss of human life caused by floods. Natural Hazards, **34**(2), 151–175 (2005)

[4] DID Malaysia, *DID Manual: Flood management*, **1**(2009)

[5] A. S. Rahman, A. Rahman, M. A. Zaman, K. Haddad, A. Ahsan and M. Imteaz, A study on selection of probability distributions for at-site flood frequency analysis in Australia. Natural Hazards, **69**(3), 1803–1813(2013)

[6] A. Z. Ismail, Z. Yusop and Z. Yusof, Comparison of flood distribution models for Johor River basin. Jurnal Teknologi (Science and Engineering), **72**(1), 1–6 (2015)

[7] O.S. Selaman, S. Said and F. J. Putuhena, Flood frequency analysis for Sarawak using Weibull, Gringorten and L-Moments Formula. The Intitutions of Engineers, Malaysia, **68** (1)**,** 43–52 (2007)

[8] N. Guru and R. Jha, Flood frequency analysis of Tel Basin of Mahanadi river system , India using annual maximum and POT flood data. Aquatic Procedia ,**4**, 427–434, (2015)

[9] B. P. Parida, R. K. Kachroo and D. B. Shrestha, Regional flood frequency analysis of Mahi-Sabarmati Basin (Subzone 3-a) using Index Flood Procedure with L-Moments. Water Resources Management, **12**(1), (1998)

[10] Department of Irrigation and Drainage of Malaysia (2016) ,Location of Sg. Segamat Gauging Station, Retrieve on September 26, 2016 from http://h2o.water.gov.my/

[11] A. Ramachandra Rao, K. H. Hamed. *Flood Frequency analysis*, CRC Press,(2000)

[12] H. Mehrannia and A. Pakgohar, Using Easy Fit Software For Goodness-Of-Fit Test and Data Generation, IJMA, **5**(1), 118–124, (2014)

[13] L. R. Singo, P. M. Kundu, J. O. Odiyo, F. I. Mathivha and T. R. Nkuna, Flood frequency analysis of annual maximum stream flows for Luvuvhu River Catchment , Limpopo Province , South Africa. In *16th SANCIAHS Hydrology Symposium*, 1-3 October 2012 at the University of Pretoria, South Africa, (2012)

[14] Z. Mohd Daud, A. H. Mohd Kassim, M. N. Mohd Desa and V. T. V. Nguyen, Statistical analysis of at-site extreme rainfall processes in Peninsular Malaysia. FRIEND 2002-Regional Hydrology: Bridging the Gap between Research and Practice, *(Proceedings of the Fourth International FRIEND Conference held at Cape Town, South Africa),* IAHS Publication, **No. 274**, 61–68, (2002)

[15] Y. H. Lim and L.M. Lye, Regional flood estimation for ungauged basins in Sarawak, Malaysia, Hydrological Sciences J., **48** (1), 79–94, (2003)

[16] S. Dan'azumi, S. Shamsudin, A. Azmi, Modeling the distribution of rainfall intensity using hourly data, American J. of Envi. Sciences, **6**(3), 238-243, (2010)

[17] W. Z. Wan Zin, A. A. Jemain and K. Ibrahim, The best fitting distribution of annual

maximum rainfall in Peninsular Malaysia based on methods of L-moment and LQ-moment, Theory Appl Climatol, **96**:. 337-334, (2009)