Research Journal of Applied Sciences, Engineering and Technology 6(4): 733-738, 2013
DOI:10.19026/rjaset.6.4189
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 22, 2012 Accepted: November 23, 2012 Published: June 20, 2013

**Research Article**

# Comparison of Distribution Models for Peakflow, Flood Volume and Flood Duration

[1]Mohsen Salarpour, [2]Zulkifli Yusop and [3]Fadhilah Yusof
[1]Faculty of Civil Engineering, Universiti Teknologi Malaysia,
[2]Water Research Alliance, Universiti Teknologi Malaysia,
[3]Faculty of Science, Department of Mathematics, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

**Abstract:** Besides peakflow, a flood event is also characterized by other possibly mutually correlated variables. This study was aimed at exploring the statistical distribution of peakflow, flood duration and flood volume for Johor River in south of Peninsular Malaysia. Hourly data were recorded for 45 years from the Rantau Panjang gauging station. The annual peakflow was selected from the maximum flow in each water year (July-June). Five probability distributions, namely Gamma, Generalized Pareto, Beta, Pearson and Generalized Extreme Value (GEV) were used to model the distribution of peakflow events. Anderson-Darling and Chi-squared goodness-of-fit tests were used to evaluate the best fit. Goodness-of-fit tests at 5% level of significance indicate that all the models can be used to model the distribution of peakflow, flood duration and flood volume. However, Generalized Pareto distribution was found to be the most suitable model when tested with the Anderson-Darling-Smirnov test and the Chi-squared test suggested that Generalized Extreme Value was the best for peakflow. The result of this research can be used to improve flood frequency analysis.

**Keywords:** Flood frequency characteristic, goodness-of-fit test, probability distribution

## INTRODUCTION

Flood disasters caused by monsoonal storms in Malaysia can pose disastrous impact on the country's economy and social life of the population. According to the MNRE (Ministry of Natural Resources and Environment) (2007), the total flood affected area in Malaysia was 29,799 km$^2$ or about 9% of the total area of the country. The total number of people living in the flood prone areas was estimated to be 4.819 million, which was about 22% of the total population as of year 2000 and the estimated total annual average flood damage was RM 915 million (MNRE, 2007).

Results gained from flood distribution studies are important for a country's water resources planning in terms of assessing decision making processes in planning and management strategies. Garde and Kothyari (1990), Gunasekara and Cunnane (1992), Haktanir (1992), Bobee *et al*. (1993), Haktanir and Horlacher (1993), Vogel *et al*. (1993), Mutua (1994), Bobee and Rasmussen (1995), Mitosek and Strupczewski (2004) and Mitosek *et al*. (2002) used statistical distributions to model the long term flood characteristics. This study, on the other hand, was aimed at analyzing the statistical distribution of flood variables, notably peakflow, duration and volume of storm runoff that may be mutually correlated, as pointed out by Laio *et al*. (2009).

The estimation of extreme rainfalls or flood peak discharges in engineering practices relies on statistical analysis of maximum precipitation or stream flow records that uses available sample data to calculate the selected frequency distribution parameters. The fitted distribution is then utilized to estimate event magnitudes pertaining to return periods unequal to recorded events (Laio *et al*., 2009). For hydraulic design, it is important to obtain accurate estimations of extreme rainfall to alleviate possible damages.

Normally, selection of statistical distributions for any flood frequency analysis is done through statistical tests or by using graphical methods (Bobee *et al*., 1993). Cunnane (1989) summarized different distributions and parameter estimation procedures that were tested and recommended for different regions. Commonly used distributions for annual flood series modeling include Extreme Value type 1 (EV1), General Extreme Value (GEV), Extreme Value type 2 (EV2), two components Extreme Value, Normal, Log Normal (LN), Pearson type 3 (P3), Log Pearson type 3 (LP3), Gamma, Exponential, Weibull, Generalized Pareto and Wake by Cunnane (1989) and Bobee *et al*. (1993). Cunnane (1989) also revealed through a survey conducted on 54 agencies in 28 countries that EV1,

**Corresponding Author:** Zulkifli Yusop, Water Research Alliance, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

EV2, LN, P3, GEV and LP3 distributions had been preferred in ten, three, eight, seven, two and seven countries respectively (Hadda and Rahman, 2011).

In this study, five probability distributions were considered as potential candidates. These were Beta, Generalized Pareto, Gamma, Pearson and Generalized Extreme Value (GEV). The reason for selecting these distributions for analysis is that they are commonly used in flood frequency studies (Chowdhury *et al*., 1991; Vogel and McMartin, 1991; Takara and Stedinger, 1994; Zalina *et al*., 2002).

## MATERIALS AND METHODS

**Data collection and study area:** Discharge and rainfall data recorded at hourly intervals was obtained from the Department of Irrigation and Drainage, Malaysia. Discharge data at the Rantau Panjang gauging station (01° 46’ 50’’N and 103° 44’ 45’’E) was used in this analysis. The data covered 45 years. Missing records were removed. Figure 1 shows the map of Peninsular Malaysia and the location of the flow gauging station.

In this study, flood duration begins from the start of hydrograph rise and ends when the falling limb intercepts an extended line with a slope of 0.0055 L/s/ha/h as suggested by Hewlett and Hibbert (1967) and Yusop *et al*. (2006). The flood volume includes both base flow and storm flow, as shown in Fig. 2

**Modeling the peakflow, flood duration and flood volume:** Generalized Pareto, Pearson, Exponential, Beta and GEV were used to model the distribution of the flood variables. The Cumulative Distribution Function (CDF) was determined using the equation:

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad (1)$$

The theoretical CDF is displayed as a continuous curve. The empirical CDF is denoted by:

$$F_n(x) = \frac{1}{n}[Number\ of\ observatio\ ns \le x] \quad (2)$$

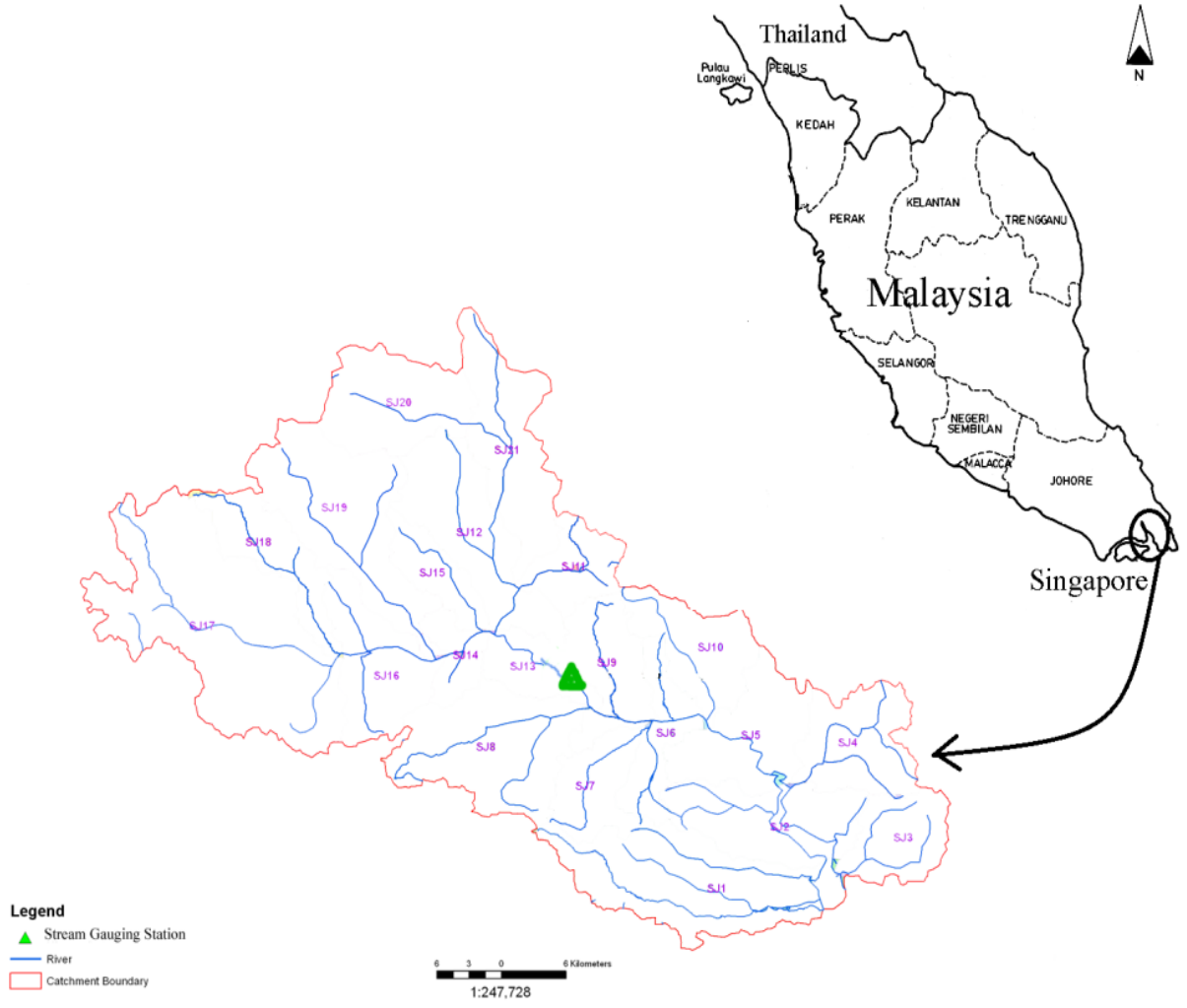


Fig. 1: Map of Johor River, south of Malaysia

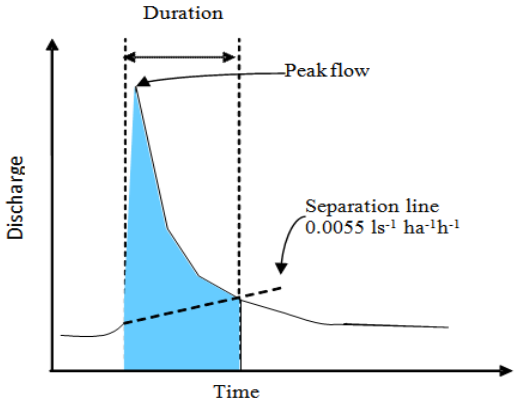


Fig. 2: Method for defining flood duration

where,

x = The random variable representing the hourly rainfall intensity

The Probability Density Function (PDF) is the probability that the variate has the value x:

$$\int_a^b f(x)\,dx = P(a \le X \le b) \tag{3}$$

For discrete distributions, the empirical (sample) PDF is displayed as vertical lines representing the probability mass at each integer X:

$$f(x) = P(X = x) \tag{4}$$

The empirical PDF is shown as a histogram with equal-width vertical bars (bins). Each bin represents the number of sample data that fall into the corresponding interval divided by the total number of data points. Theoretically, the PDF is in the form of a continuous curve appropriately scaled to the number of intervals.

The Probability Density Functions (PDF) and Cumulative Distribution Function (CDF) for the five models are given as follows:

**Generalized Pareto distribution:** The Generalized Pareto distribution with continuous shape parameter (K), continuous scale parameter (σ>0) and continuous location parameter (μ) have PDF and CDF given by:

$$F(x) = \begin{cases} \frac{1}{\sigma}\left(1+\kappa\frac{(x-\mu)}{\sigma}\right)^{-1-1/k} & k \neq 0 \\ \frac{1}{\sigma}\exp\left(-\frac{(x-\mu)}{\sigma}\right) & k = 0 \end{cases} \tag{5}$$

and

$$F(x) = \begin{cases} 1 - \left(1 + k\frac{x-\mu}{\sigma}^{-\frac{1}{k}}\right) & k \neq 0 \\ 1 - exp\left(-\frac{x-\mu}{\sigma}\right) & k = 0 \end{cases} \tag{6}$$

where,

$$\mu \le x \prec +\infty \qquad for\ \ \kappa \ge 0$$

$$\mu \le x \le \mu - \sigma/\kappa \qquad for\ \ \kappa \prec 0$$

**Pearson distribution:** The Pearson distribution with continuous shape parameter (α>0), continuous scale parameter (β>0) and continuous location parameter (γ) have PDF and CDF given by:

$$f(x) = \frac{\exp(-\beta/(x-\gamma))}{\beta\Gamma(\alpha)((x-\gamma)/\beta)^{\alpha+1}} \tag{7}$$

$$F(x) = 1 - \frac{\Gamma_{\beta/(x-\gamma)}(\alpha)}{\Gamma(\alpha)} \tag{8}$$

where,

$$\gamma \le x \prec +\infty$$

**Gamma distribution:** The Gamma distribution with continuous shape parameter (α), continuous scale parameter (β) and continuous location parameter (γ) have PDF and CDF given by:

$$f(x) = \frac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\,\Gamma(\alpha)}\exp(-(x-\gamma)/\beta) \tag{9}$$

$$F(x) = \frac{\Gamma_{\frac{x-\gamma}{\beta}}(\alpha)}{\Gamma(\alpha)} \tag{10}$$

where,

$$\gamma \le x \prec +\infty$$

**Beta distribution:** The Beta distribution with continuous scale parameter ($\alpha_1$>0), continuous shape parameter ($\alpha_2$>0) and continuous location parameter (a<b) have PDF and CDF given by:

$$f(x) = \frac{1}{\beta(\alpha_1,\alpha_2)}\frac{(x-a)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{(b-a)^{\alpha_1+\alpha_2-1}} \tag{11}$$

$$F(x) = I_z(\alpha_1,\alpha_2) \tag{12}$$

$$z \equiv \frac{x-a}{b-a}$$

$I_z$ = The Regularized Incomplete Beta Function

where,

$$a \le x \le b$$

**Generalized Extreme Value (GEV):** The general extreme value I with continuous shape parameter (K), continuous scale parameter (σ and continuous location parameter (μ) have PDF and CDF given by:

$$f(x)=\begin{cases}\frac{1}{\sigma}\exp(-(1+kz)^{-1/k})(1+kz)^{-1-1/k} & k\neq 0\\ \frac{1}{\sigma}\exp(-z-\exp(-z)) & k=0\end{cases} \quad (13)$$

$$F(x)=\begin{cases}\exp(-(1+kz)^{-1/k} & k\neq 0\\ \exp(-\exp(-z)) & k=0\end{cases} \quad (14)$$

where,

$$z\equiv\frac{x-\mu}{\sigma}$$

$$1+k\frac{(x-\mu)}{\sigma}\rangle 0 \quad for \quad k\neq 0$$
$$-\infty\langle x\langle+\infty \quad for \quad k=0$$

**Goodness-of-fit tests:** The Goodness-of-Fit (GOF) tests measure the compatibility of a random sample with a theoretical probability distribution function. In other words, these tests show how well a selected distribution fits the data. Two goodness-of-fit tests were conducted at 5% level of significance. Note that *X* denotes the random variable and *n* is the sample size. The mathematical explanation of two goodness-of-fit tests is as follows:

**Anderson-Darling (A-D) test:** This statistical test is used to find out if a given sample belongs to a specific probability distribution. The test assumes that there are no parameters to be estimated in a distribution under scrutiny, which means that the test and its critical value sets are distribution-free. This test is more often used to test a family of distributions where the parameters in the family need to be estimated; this has to be noted in adjusting the test-statistics or its critical values. The test statistic ($A^2$) is given as:

$$A^2=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1).[\ln F(X_i)+\ln(1-F(X_{n-i+1}))] \quad (15)$$

**Chi-Squared (C-S) test:** This test is a statistical hypothesis test to simply compare how well the theoretical distribution fits the empirical distribution PDF. The Chi-squared test statistic is given by:

$$X^2=\sum\frac{(O_i-E_j)^2}{E_i} \quad (16)$$

where,
$O_i$ = The observed frequency for bin i
$E_i$ = The expected frequency for bin I and is given by:

$$E_1=f(X_2)-F(X_1) \quad (17)$$

where,
$X_1$ & $X_2$ : The lower and upper limits for bin *i*

**The Cumulative Distribution Function (CDF):** The cumulative distribution function is the probability that the variate takes on a value less than or equal to *x*:

$$F(x)=P(X\leq x) \quad (18)$$

For continuous distributions, the CDF is expressed as:

$$F(x)=\int_{-\infty}^{x}f(t)\,dt \quad (19)$$

so the theoretical CDF is displayed as a continuous curve. The empirical CDF is denoted by:

$$F_n(x)=\frac{1}{n}[Number\ of\ observatio\ n\leq x] \quad (20)$$

## RESULTS AND DISCUSSION

The averages of peakflow, flood duration and flood volume at the study site were 248 $m^3$/sec, 349 h and 104 mm, respectively and the corresponding standard deviations were 163 $m^3$/sec, 125 h and 49 mm. Table 1 presents the fitting result parameters for various distributions of flood variables. In this table the amount of continues shape parameter (α, K), continues scale parameter (σ, β) and continues location parameter (μ, γ)

Table 1: Fitting result parameters for various distributions of flood variables

| | Flood variables | | |
|---|---|---|---|
| Distributions | Peakflow (P) | Duration (D) | Volume (v) |
| Beta | $\alpha_1$ = 0.54000 | $\alpha_1$ = 1.1200 | $\alpha_1$ = 1.350 |
| | $\alpha_2$ = 1.78000 | $\alpha_2$ = 1.3700 | $\alpha_2$ = 2.010 |
| Gen. Pareto | κ = -0.40000 | κ = -0.8200 | κ = -0.560 |
| | σ = 184.48000 | σ = 373.0200 | σ = 111.330 |
| | μ = 70.68000 | μ = 144.3400 | μ = 33.480 |
| Gamma | α = 0.88255 | α = 6.0071 | α = 6.810 |
| | β = 177.13000 | β = 52.7900 | β = 18.840 |
| | γ = 76.89900 | γ = 32.2000 | γ = 23.540 |
| Pearson | α = 2.81000 | α = 77.7800 | α = 228.241 |
| | β = 454.29000 | β = 2067.2000 | β = 6867.600 |
| | γ = 8.97000 | γ = -739.5200 | γ = -47.280 |
| Gen. Extreme | κ = 0.21558 | κ = -0.20041 | κ = -0.074 |
| Value (GEV) | σ = 98.51500 | σ = 122.4500 | σ = 42.820 |
| | μ = 164.97000 | μ = 144.3400 | μ = 83.017 |

Based on the Anderson-darling test, the generalized extreme value distribution is the best fitted to flood volume and duration and Gen. Pareto distribution is the best for peakflow

Table 2: Goodness-of-fit test ranking for various distributions of flood variables

| | Goodness-of fit tests | | | | | |
|---|---|---|---|---|---|---|
| | Anderson-darling | | | Chi-squared | | |
| Distributions | Peakflow (P) | Duration (D) | Volume (v) | Peakflow (P) | Duration (D) | Volume (v) |
| Beta | 5 | 4 | 4 | 5 | 1 | 2 |
| Gen. Pareto | 1 | 5 | 5 | 3 | 5 | 5 |
| Gamma | 4 | 3 | 2 | 2 | 4 | 3 |
| Pearson | 2 | 2 | 3 | 4 | 3 | 4 |
| Gen. Extreme Value (GEV) | 3 | 1 | 1 | 1 | 2 | 1 |

Ranking is in the order of 1, 2, 3, 4 and 5; 1 is the best ranking and 5 the worst ranking, so the generalized extreme value distribution is the best fitted to flood volume and duration and Gen. Pareto distribution is the best for peakflow based on the Anderson-darling test

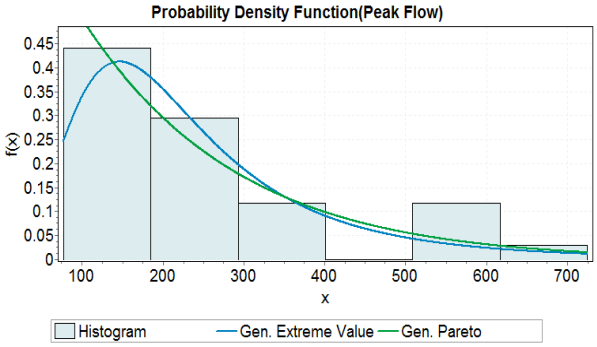


Fig. 3: Generalized extreme value distribution and generalized Pareto distribution fitted to the peakflow

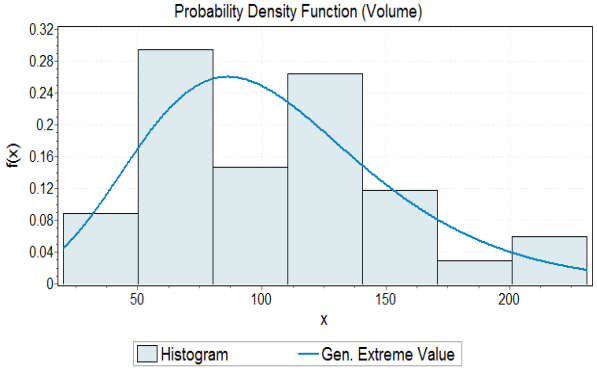


Fig. 4: Generalized extreme value distribution fitted to the volume of peakflow

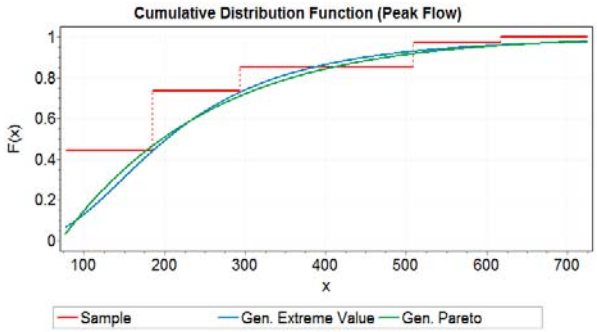


Fig. 5: Comparison between generalized extreme value and generalized Pareto distributions in the cumulative distribution function of peakflow

are valid. The goodness-of-fit test ranking results are shown in Table 2. Based on the Anderson goodness-of-fit test method, it was found that the Generalized Pareto was the best distribution to fit the peakflow and Generalized Extreme Value was the best for flood duration and volume. However, when the Chi-Squared test was used, GEV became more favorable for fitting peakflow and flood volume and Beta was the best distribution for the flood duration. Figure 3 presents the PDF for the GEV and GP distributions fitted to the peak flow. Since the goodness-of-fit test statistics indicate the distance between the observed data and the fitted distributions, it is obvious that the distribution with the lowest statistic value is the best fitting model. Based on this fact, the statistics from the Anderson goodness-of-fit test using GP for peak flow, GEV for flood duration and flood volume were 0.1551, 0.3547 and 0.2376, respectively. Also for the Chi-Squared test, the statistics for GEV for peakflow and flood volume and Beta for flood duration were 0.0571; 0.15231 and 1.2941, respectively. Figure 4 presents the PDF of GEV distribution fitted to the flood volume whereas Fig. 5 compares the CDF of peakflow between GEV and GP distributions. The CDF graph is useful to precisely determine how well the distributions can fit the observed data. The results showed that the GP distribution was more significant for peakflow compared to GEV. Previous studies on flood variables mostly focused on peak flow. GEV distribution had been found to be the best distribution to fit peakflow data over several stations in Malaysia (Ahmad *et al*., 2011; Ashkar and Mahdi, 2006). Also, Suhaila and Jemain (2007, 2008) found that GEV distribution was the best for fitting daily rainfall throughout Peninsular Malaysia.

## CONCLUSION

Flow data were used to analyze the statistical distribution of the peakflow, duration and volume of annual flood for Johor River at Rantau Panjang gauging station. Five probability distributions, namely Beta, Generalized Pareto, Gamma, Pearson and Generalized Extreme Value were tested. Based on the Anderson-Darling test, the Generalized Extreme Value distribution was found to be the most suitable for modeling the flood volume and duration and General Pareto was the most fitted to peakflow. Meanwhile,

based on the Chi-squared test, the Generalized Extreme Value distribution was the most suitable for modeling the flood volume and peakflow and Beta was the most fitted to the duration. Goodness-of-fit tests at 5% level of significance indicate that all the models can be used to model the distribution of peakflow, flood duration and flood volume. For further study it is recommended to evaluate the performance of other distributions such as Log Normal, Log Pearson Type 3 and Normal. In addition, different goodness-of-fit tests such as Anderson-Darling and Kolmogorov-Smirnov can be attempted.

## ACKNOWLEDGMENT

We are grateful to the Department of Irrigation and Drainage (DID) Malaysia for providing the data. Support from the Research Management Center (RMC) of Universiti Teknologi Malaysia is greatly acknowledged. This study was also partly supported by the Ministry of Higher Education, Malaysia and the Japanese Society for the Promotion of Science (JSPS) under the Asian Core Program.

## REFERENCES

Ahmad, U.N., A. Shabri and Z.A. Zakaria, 2011. Flood frequency analysis of annual maximum stream flows using L-moments and TL-moments approach. Appl. Math. Sci., 5(5): 243-253.

Ashkar, F. and S. Mahdi, 2006. Fitting the log-logistic distribution by generalized moments. J. Hydrol., 328: 694-703.

Bobee, B. and P.F. Rasmussen, 1995. Recent advances in flood frequency analysis. Rev. Geophy., 33(S2): 1111-1116.

Bobee, B., G. Cavidas, F. Ashkar, J. Bernier and P. Rasmussen, 1993. Towards a systematic approach to comparing distributions used in flood frequency analysis. J. Hydrol., 142: 121-136.

Chowdhury, J.U., J.R. Stedinger and L.H. Lu, 1991. Goodness-of-fit test for regional GEV flood distribution. Water Resour., 27(77): 1765-1776.

Cunnane, C., 1989. Statistical Distributions for Flood Frequency Analysis. Secretariat of the World Meteorological Organization, Geneva, pp: 64.

Garde, R.J. and U.C. Kothyari, 1990. Flood estimation in Indian catchments. J. Hydrol., 113: 135-146.

Gunasekara, T.A.G. and C. Cunnane, 1992. Split sampling technique for selecting a flood frequency analysis procedure. J. Hydrol., 130: 189-200.

Hadda, K. and A. Rahman, 2011. Selection of the best fit flood frequency distribution and parameter estimation procedure: A case study for Tasmania in Australia. Stoch. Env. Res. Risk Assess., 25(3): 415-428.

Haktanir, T., 1992. Comparison of various flood frequency distributions using annual flood peaks data of rivers in Anatolia. J. Hydrol., 136: 1-31.

Haktanir, T. and H.B. Horlacher, 1993. Evaluation of various distributions for flood frequency analysis. Hydrol. Sci. J. Des. Sci. Hydrol., 2(1-2): 15-32.

Hewlett, J.D. and A.R. Hibbert, 1967. Factors Affecting the Response of Small Watersheds to Precipitation in Humid Areas. In: Sopper, W.E. and H.W. Lull (Eds.), Proceedings of the International Symposium on Forest Hydrology. Pergamon, Oxford, pp: 275-290.

Laio, F., G.D. Baldassarre and A. Montanari, 2009. Model selection techniques for the frequency analysis of hydrological extremes. Water Resour. Res., 45: W07416.

Mitosek, H.T. and W.G. Strupczewski, 2004. Simulation Results of Discrimination Procedures. Retrieved from: http://www.igf.edu.pl/.

Mitosek, H.T., W.G. Strupczewski and V.P. Singh, 2002. Toward an objective choice of an annual flood peak distribution. Proceeding of the 5th International Conference on Hydro-science and-engineering, Published on CR ROM: Advances in Hydro-Science and Engineering, Warsaw.

MNRE, 2007. Flood and Drought Management in Malaysia. Ministry of Natural Resoirces and Environment.

Mutua, F.M., 1994. The use of the akaike information criterion in the identification of an optimum flood frequency model. Hydrol. Sci. J. Des. Sci. Hydrol., 39(3): 235-244.

Suhaila, J. and A.A. Jemain, 2007. Fitting daily rainfall amount in peninsular Malaysia using several types of exponential distributions. J. Appl. Sci. Res., 3(10): 1027-1036.

Suhaila, J. and A.A. Jemain, 2008. Fitting the statistical distribution for daily rainfall in peninsular Malaysia based on AIC criterion. J. Appl. Sci. Res., 4(12): 1846-1857.

Takara, K.T. and J.R. Stedinger, 1994. Recent Japanese contributions to frequency analysis and quantile and quantile lower bound estimators. Stochast. Statist. Meth. Hydrol. Env. Eng., 1: 217-234.

Vogel, R.M. and D.E. McMartin, 1991. Probability plot goodness-of-fit and skewness estimation procedures for the Pearson Type 3 distribution. Wat. Resour. Res., 27(2): 3149-3158.

Vogel, R.M., W.O. Thomas and T.A. McMahon, 1993. Flood-flow frequency model selection in southeastern United States. J. Water Resour. Plann. Manag., 119(3): 353-366.

Yusop, Z., I. Douglas and A.R. Nik, 2006. Export of dissolved and undissolved nutrients from forested catchments in Peninsular Malaysia. Forest Ecol. Manag., 224: 26-44.

Zalina, M.D., M.N. Desa, V.T. Nguyen and A.H. Kassim, 2002. Selecting a probability distribution for extreme rainfall series in Malaysia. Water Sci. Technol., 45(2): 63-68.