Journal of Computer Science & Computational Mathematics, Volume 9, Issue 1, March 2019
DOI: 10.20967/jcscm.2019.01.002

Modelling Annual Maximum River Flows with Generalized Extreme Value Distribution

R Y Cheong*, Darmesah Gabda

Department of Mathematics with Economics, Faculty of Science and Natural Resources, Universiti Malaysia Sabah
*Corresponding author email: riying93@hotmail.com

***Abstract*:** A good understanding of probability distribution of annual maximum river flow is believed to improve water resources planning and design. Based on the annual maximum river flow record over 20-48 years at 9 individual river sites in Sabah, the data set are fitted into generalized extreme value (GEV) distribution with maximum likelihood estimator. Both stationary and non-stationary models are considered. Likelihood ratio test shows that most of the river flows are stationary. Over a homogeneous region, a parent distribution with common shape parameter is found well describing the behaviour of selected annual maximum river flow. Hence, 10- and 100-year return levels are estimated using the single model.

***Keywords*:** generalized extreme distribution, maximum likelihood estimator, annual maximum river flow.

1. Introduction

The modelling of annual maxima river flows has been the popular topic for long time. In Sabah, river flows are important for survival and country economic development. The population growth is putting pressure on state's water resources. Poor water resources planning and management might bring to flood or drought due to climate change and even low quality of drinking water. In short, improper control of river flow will bring severe destroy to crops, economic loss and casualty. Hence, identify the behaviour of extreme river flows, mainly in long term trends, is of essential. A fairly accurate estimation of extreme flows with given return period will improve in decision making so that to avoid waste of investment or severe damage and sacrifice of life [1],[2].

In previous researches, annual maximum river flow was used as an indicator for flood trends analysis. It is believed that a good understanding of river flows pattern may be useful in reducing the negative impacts as mentioned above. The first step in modelling extreme value is analysing the data in the form of cumulative distribution and determine the best fitting distribution function. In annual maxima analysis, generalized extreme value (GEV) distribution is always suggested due to the advantage in allowing uncertainty to be considered, in particular scale parameter. Hence, a more robust prediction can be obtained from GEV distribution [3]-[5]. In Malaysia, 3-p log-normal distribution and generalized Pareto distribution (GPD) are suggested in analysing annual maximum river flows at Johor [6],[7]. In this study, GEV distribution is employed.

Suitable parameter estimation could reduce bias and uncertainty in estimates. There are several frequentist methods suggested in cooperating with GEV distributions such as maximum likelihood estimation (MLE), probability weighted moment (PWM) and L-moment. In hydrological events, MLE is often chosen which shows less bias and provides more consistent approach to parameter estimate [8]-[10].

Recently, environmental scientists and statisticians start focus on non-stationary probability distribution in flood frequency analysis. Environmental process could exhibit trend in time [11]. Climate variability and anthropogenic activities such as deforestation as well as land misuse are found affecting on the behaviour of extreme river flows [12]. In the study of [13], stationary model is found underestimate in flood quantile relative. [14] who test the hypothesis of stationarity in estimating flood events, conclude that a linear trend in location parameter is important. The similar finding is supported by [15] in the study of annual maximum stream flow analysis in Canada.

The objective of this study is to describe the behaviour of selected extreme river flows in Sabah by using a parent probability distribution. In particular, the annual maximum series data of river flow from nine sites with small sample size are fitted into GEV distribution with MLE as parameter estimation. Both stationary and non-stationary cases are studied. Location parameter in non-stationary model is accordance with time dependent. Likelihood ratio test is conducted to compare both models. Next, we examine the possible common GEV parameters between sites. Lastly, we obtain the return level estimate from chosen model.

2. Research Methodology

Extreme value theory (EVT) provides analogues of the central limit theorem for the extreme values in a sample, which normally situated in the tail distribution.

2.1 Generalized extreme value distribution

EVT focuses on the statistical behaviour of $M_n = \max\{X_1,\ldots,X_n\}$ where $X_1,\ldots,X_n$ is a sequence of i.i.d. EVT states that, if there exits of normalising constant $\{a_n > 0\}$ and $\{b_n\}$, G is non-degenerate distribution function such that

$$\Pr\left\{\frac{(M_n - b_n)}{a_n} \le z\right\} \to G(z), n \to \infty \tag{1}$$

where G is a non-degenerate distribution function, then G belongs to one of the families of GEV distribution. The cumulative distribution function (cdf) of GEV distribution is denoted as follow:

$$G_{\zeta,\mu,\sigma}(x)=\exp\left\{-\left[1+\zeta\frac{(x-\mu)}{\sigma}\right]^{-\frac{1}{\zeta}}\right\};1+\zeta\frac{(x-\mu)}{\sigma}>0 \quad (2)$$

Location (μ), scale (σ) and shape (ζ) are the parameters in GEV distribution which combines Fréchet distribution $(\zeta>0)$, Gumbel distribution $(\zeta=0)$ and Weibull distribution $(\zeta<0)$.In EVT, GEV distribution is used to model the tail behaviour of a distribution where the shape parameter plays the role [16].

2.2 Maximum likelihood estimation

A likelihood function can be formed when the observations are known which gives the probability of observed data. The joint likelihood function of the sample follows from probability distribution of (2) as

$$L(\theta\mid x)=\frac{1}{\sigma^n}\prod_{i=1}^{n}\left[1+\zeta\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\zeta}-1}\times\exp\left[-\sum_{i=1}^{n}\left[1+\zeta\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\zeta}}\right] \quad (3)$$

MLE can be described by log-likelihood as

$$\ell(\theta\mid x)=\log L(\theta\mid x)=\sum_{i=1}^{n}\log g(x_i;\theta);\ g(x_i;\theta)=\frac{\partial G(x)}{\partial x} \quad (4)$$

The parameters are estimated by solving the partial derivatives of the log-likelihood function and equate them to zero with Newton-Raphson algorithm [17].

2.3 Model Comparison

Likelihood ratio test is an efficient method to compare nested models for covariance. Suppose that Model 1 is a reduced model with 3-parameter and Model 2 is a full model with 4-parameter.

$$M_1: X\sim GEV(\mu,\sigma,\zeta)$$
$$M_2: X\sim GEV(\mu(t)=\beta_0+\beta_1 t,\sigma,\zeta) \quad (5)$$

The considered hypothesis testing is given by

$$H_0:\beta_1=0$$
$$H_a:\beta_1\neq 0 \quad (6)$$

Let L_1 and L_2 be the maximum likelihood of Model 1 and Model 2 respectively. The likelihood ratio test is given by

$$\gamma=-2\ln\frac{L_1}{L_2} \quad (7)$$

and distributed as $(1-\alpha)$ quantile chi-square distribution. The degree of freedom k is corresponding to the difference number of parameters between the two models. If $\gamma<\chi^2_{k,1-\alpha}$, H_0 is not rejected.

3. Return level estimate

A return period is an estimate of the likelihood of an event. Return level z_p is expected to be exceeded on average once every $\frac{1}{p}$ periods, where p is the probability of the extreme event $(0<p<1)$. By inverting the cdf in (2), return level estimates are obtained by

$$z_p=\left\{\mu-\frac{\sigma}{\zeta}\left[1-\{-\log(1-p)\}^{\zeta}\right]\right\} \quad (8)$$

4. Results and Discussions

4.1 Descriptive analysis

In this study, nine river flows with small sample size $(n<50)$ are selected for the annual maximum river flow analysis. The secondary data are obtained from Hydrology and Survey Division under Department of Irrigation and Drainage, Sabah. The data are collected as daily mean of 24-hour periods beginning at 8.00am every day. The observations are measured in m^3s^{-1}. Table 1 summarizes the information of the selected river flows in this study. The range of annual maxima observations used from 20 to 48 years with average 40 years.

Table 1. Information of selected river flows

Site	Stations	No. of years	Period	Max. observation
1	Sungai Apin-apin at Waterworks	20	1996-2015	50.41
2	Sungai Baiayo at Bandukan	21	1993-2013	39.94
3	Sungai Padas at JPS Beaufort	35	1981-2015	1506.30
4	Sungai Sook at Biah	47	1969-2015	313.99
5	Sungai Wariu at Bridge No.2	47	1969-2015	524.90
6	Sungai Kadamaian at Tamu Darat	47	1969-2015	490.20
7	Sungai Papar at Kaiduan	48	1969-2016	468.86
8	Sungai Papar at Kogopon	48	1969-2016	970.30
9	Sungai Pegalan at Ansip	48	1969-2016	688.63

4.2 Stationarity of data

Annual maximum series of river flow are fitted into GEV distribution. The results are analysed according to individual location. Two models as demonstrate in (5) are built. In stationary model, probability weighted moment is employed as the initial value of MLE. The intercept parameters of MLE in non-stationary model are PWM estimates and the parameter based on covariate is initially set as zero. Covariate t in non-stationary model represents number of year observations. The GEV parameter estimates for both cases are shown in Table 2 and Table 3.

Next, likelihood ratio test is employed to compare both models. In the presence case, the degree of freedom equals to 1. This suggesting that, at significance level of 5%, the critical value is $\chi^2_{1,0.95}=3.8415$. Likelihood ratio test at each site is summarized at Table 4. Since the value of likelihood ratio test at site 5 and site 6 are larger than critical value, null hypothesis is rejected. Hence, we do not have enough evidence to prove there is trend exiting at each sites except site 5 and site 6.

Table 2. GEV parameters (stationary model)

Site	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\zeta}$
1	21.33	10.61	-0.09
2	21.22	5.06	0.08
3	822.51	218.11	-0.18
4	127.78	61.16	-0.19
5	113.78	58.54	0.09
6	204.29	80.58	0.08
7	120.18	54.78	0.003
8	256.72	119.89	0.02
9	222.31	111.94	-0.02

Table 3. GEV parameters (non-stationary model)

Site	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	$\hat{\zeta}$
1	23.01	-0.09	10.40	-0.06
2	19.96	0.05	4.96	0.10
3	735.46	2.34	204.00	-0.11
4	87.77	1.97	28.69	0.14
5	140.98	-0.59	52.96	0.17
6	144.53	1.24	71.34	0.17
7	141.86	-0.45	52.04	0.02
8	303.51	-1.08	108.69	0.10
9	206.51	0.38	114.26	-0.05

Table 4. Likelihood ratio test

Site	Likelihood Ratio Test	H_0
1	0.19	Not rejected
2	0.25	Not rejected
3	1.55	Not rejected
4	0.86	Not rejected
5	**4.78**	**Rejected**
6	**6.92**	**Rejected**
7	3.10	Not rejected
8	3.29	Not rejected
9	0.29	Not rejected

4.3 Model with common GEV parameters

Over a homogeneous region, it is reasonable to assume that the individual sites follow the same distribution type with common shape parameter but different scale parameter [18]. In this study, a single model with common shape parameter that links all nine models together is build. This single model is expected to describe the probability distribution of the selected river flows within a homogeneous region.

The model fitting technique in previous section is applied here as well. The initial value for shape parameter is the average value of 9 independent shape parameter estimates from Table 2, which is $\hat{\zeta} = -0.02$. Since most of the cases are stationary, therefore only stationary model is considered. The two models consider in this section as stated below.

$$M_1 : X \sim GEV(\mu_j, \sigma_j, \zeta);\ j = 1,2,\ldots,9$$
$$M_{.2} : X \sim GEV(\mu_j, \sigma_j, \zeta_j);\ j = 1,2,\ldots,9 \quad (9)$$

Hypothesis testing is given by

$$H_0 : \zeta_j = \zeta$$
$$H_a : \text{at least one } \zeta_j \text{ is different from other} \quad (10)$$

Since the difference in number of parameters between both models is 8, hence at 5% significance level, the critical value is $\chi^2_{8,0.95} = 15.5073$. The likelihood ratio test is 7.8289 which is smaller than the critical value. Thereby, H_0 is not rejected. There is no enough evidence to prove that there is at least one shape parameter is different from other. The single model with common shape parameter is able to describe the selected river sites. The respective parameter estimates as shown in Table 5.

Table 5. GEV parameter estimates with $\hat{\zeta} = -0.02$

Site	$\hat{\mu}_j$	$\hat{\sigma}_j$
1	20.83	10.32
2	24.47	5.24
3	809.36	215.52
4	123.48	59.77
5	116.99	60.61
6	209.07	84.69
7	120.19	54.77
8	259.06	120.37
9	220.99	111.37

4.4 Return level estimate

Application of GEV model fitting is to predict the return level that the annual maximum river flows exceeding the maximum observations in Table 1. MLE estimators in Section3.3 are substituted with p=0.1 and p=0.01 into (8) to estimate 10- and 100-year of return level at each site. Return level estimates are shown in Table 6.

Table 6. 10-,100-year return level estimates (m^3)

Site	p=0.1	p=0.01
1	44.00	68.08
2	33.23	45.44
3	1293.17	1795.85
4	257.66	397.08
5	253.06	**394.43**
6	399.12	596.65
7	243.25	**370.89**
8	529.29	**810.05**
9	471.00	730.75

As comparing to maximum observations in Table 1, most of the maximum river flows are expected to be exceeded on average once in every 100-year except site 5, site 7 and site 8.

5. Conclusion

In this study, nine annual maximum river flows in Sabah are fitted into GEV distribution. Both stationary and non-stationary models are considered. The parameters are estimated by employing MLE with PWM as initial value. Likelihood ratio test suggests that most of the river flows are stationary except site 5 and site 6. A model with common shape parameter is found suitable to describe all the annual maximum river flows in this study. With this single model, most of the maximum river flows are expected to exceed, on average 100-year, except site 5, site 7 and site 8. In conclusion, modelling annual maximum river flow using GEV distribution with common shape parameter seems reasonable. For future

study, non-stationary case might take into consider when modelling the data set with common shape parameter. Spatial modelling as studied in [19] can be applied to improve the analysis as well. As shown in [20], Bayesian approach will be better parameter estimation as compared to traditional frequentist methods especially in uncertainty analysis.

Acknowledgement

The authors sincerely thank to reviewer for helpful comments. This research is supported by a grant from UMS Great 2017 (GUG0136-1/2017).

References

[1] B. Saghafian, S. Golian, A. Ghasemi, "Flood frequency analysis based on simulated peak discharges", *Natural Hazards*, vol. 71, no. 1, pp. 403-717, 2014.

[2] M. Ellouze, H. Abida, "Regional flood frequency analysis in Tunisia: identification of regional distributions", *Water Resources Management*, vol. 22, no. 8, pp. 943-957, 2008.

[3] S. Coles, L. R. Pericchi, S. Sisson, "A fully probabilistic approach to extreme rainfall modelling", *Journal of Hydrology*, vol. 273, no. 1-4, pp. 35-50, 2003.

[4] E. Chung, S. U. Kim, "Bayesian rainfall frequency analysis with extreme value using the informative prior distribution", *KSCE Journal of Civil Engineering*, vol. 17, no. 6, pp. 1502-1514, 2013.

[5] S. U. Kim, G. Kim, W. M. Jeong, K. Jun, "Uncertainty analysis on extreme value analysis of significant wave height at eastern coast of Korea", *Applied Ocean Research*, vol. 41, pp.19-27, 2013.

[6] S. R. Noor, Z. Yusop, "Frequency analysis of annual maximum flood for Segamat river", in *MATEC Web of Conferences*, Malaysia, vol. 103, pp. 1-9, 2017.

[7] A. M. J. Nur, S. Ani, "Estimating distribution parameters of annual maximum streamflows in Johor, Malaysia using TL-moments approach", *Theoretical and applied climatology*, vol. 127, no. 1-2, pp. 213-227, 2017.

[8] W. Zhang, Y. Cao, Y. Zhu, Y. Wu, X. Ji, Y. He, Y. Xu, W. Wang, "Flood frequency analysis for alterations of extreme maximum water levels in the Pearl River Delta", *Ocean Engineering*, vol. 129, pp. 117-132, 2017.

[9] O. Zisheng, Y. Xiangqun, "Extreme value flood frequency analysis at water Gauging station near the Dongting lake", in *2011 International Symposium on Water Resource and Environmental Protection*, China, pp. 592-294, 2011.

[10] S. Nadarajah, "Extreme value models with application to drought data", *AStA*, vol. 90, no. 3, pp. 403-418, 2006.

[11] S. Coles, *An introduction to Statistical Modelling of Extreme Values*, Springer-Verlag, London, 2001.

[12] W. H. J. Toonen, "Flood frequency analysis and discussion of non-stationarity of the Lower Rhine flooding regime (AD 1350-2011): Using discharge data, water level measurements, and historical records", *Journal of Hydrology*, vol. 528, pp. 490-502, 2015.

[13] M. Šraj, A. Viglione, J. Parajka, G. Blöschl, "The influence of non-stationarity in extreme hydrological events on flood frequency estimation", *Journal of Hydrology and Hydromechanics*, vol. 64, no. 4, pp. 426-437, 2016.

[14] J. Houkpè, B. Diekkrüger, D. F. Badou, A. A. Afouda, "Non-stationary flood frequency analysis in the Ouémé River Basin, Benin Republic", *Hydrology*, vol. 2, no. 4, pp. 210-229, 2015.

[15] X. Tan, T. Y. Gan, "Nonstationary analysis of annual maximum streamflow of Canada", *Journal of Climate*, vol. 28, pp. 1788-1805, 2015.

[16] S. Kotz, S. Nadarajah, *Extreme Value Distributions, Theory and Applications*, Imperial College Press, London, 2000.

[17] S. Yoon, W. C. Cho, J. H. Heo, "A full Bayesian approach to generalized maximum likelihood estimation of generalized extreme value distribution", *Stoch Environ Res Risk Assess*, vol. 24, no. 5, pp. 761-770, 2010.

[18] M. Naghettini, E. J. d. A. Pinto, "Regional frequency analysis of hydrologic variables", In *Fundamentals of Statistical Hydrology*, Springer International Publishing, Switzerland, pp. 441-495, 2017.

[19] D. Gabda, R. Towe, J. Wadsworth, J, Tawn, "Discussion of 'Statistical Modeling of Spatial Extremes' by A. C. Davison, S. A. Padoan and M. Ribatet", *Statistical Science*, vol. 27, no. 2, pp. 189-192, 2012.

[20] R. Y. Cheong, D. Gabda, "Modelling maximum river flow by using Bayesian Markov Chain Monte Carlo", *Journal of Physics: Conference Series*, vol. 890, pp. 1-7, 2017.