Journal of Computer Science & Computational Mathematics, Volume 8, Issue 4, December 2018
DOI: 10.20967/jcscm.2018.04.004

Frequency Analysis of Annual Maximum River Flow by Generalized Extreme Value Distribution with Bayesian MCMC

R Y Cheong[1], Darmesah Gabda[2*]

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

***Abstract*:** The aim of this paper is to fit 9 annual maximum river flows in Sabah for a period record of over 20-48 years into the generalized extreme value (GEV) distribution. Bayesian Markov Chain Monte Carlo is employed as the parameter estimation which is believed to provide a more robust inference through prior and posterior distribution. In this study, scale parameter is being associated with the linear trend function. Based on the 95% credible interval in this study, the results suggest that the additional covariate to the model has no impact at most of the river sites. Hence, return level with 10- and 100- year for each river sites have been obtained by using a simple model which is urged in substituting complex models such as logistic model.

***Keywords*:** Bayesian Markov Chain Monte Carlo, annual maximum river flows.

1. Introduction

Sabah is the northern part of Borneo region. The annual rainfall received in Sabah is between 2500mm to 3500mm. Most of the residents rely on the river water for survival. The state plays an important role in planning the water management in order to avoid economic loss caused by droughts or floods, casualty and damage of infrastructures. There are two types of floods that occurs in Sabah, which is flash flood and monsoon flood. Deforestation and land misuse increases the flood risk in Sabah.

A suitable probability distribution in annual maximum analysis is able to reduce the negative impacts. A 3-parameter log-normal distribution and a generalized Pareto distribution (GPD) are suggested in analysing the annual maximum river flow at Johor [1, 2]. A full three parameters GEV distribution is recommended than the reduced distribution. This is because GEV distribution could allow uncertainty to be considered especially in scale parameter which then provides a more reliable estimation [3, 4].

GEV distribution with maximum likelihood method (MLE) is widely applied in hydrological events due to being less bias and provides a more consistent approach to parameter estimation. However, there are weaknesses of MLE shown in return level estimates [5, 6]. An overlooked or ignorance of the model uncertainty leads to an underestimation of the probability of extreme events [3]. From the point of view of uncertainty analysis, Bayesian Markov Chain Monte Carlo (MCMC) with Metropolis-Hastings algorithm approach is found to perform better than MLE when dealing with small sample size. Bayesian MCMC method is able to reduce the range of uncertainty [7, 8]. Also, Bayesian MCMC approach could give a better prediction due to the allowance of estimation uncertainty [3] and a more complete inference through posterior distribution [9]. A better performance in root mean square error, relative absolute square error and probability plot correlation coefficient in Bayesian MCMC than MLE is presented in [10].

There are various algorithms studied and suggested for MCMC method such as Metropolis-Hasting algorithm, Gibbs sampling, Hamiltonian Monte Carlo (HMC) and Riemann manifold HMC (RMHMC). Metropolis-Hastings algorithm may yield the best match between the observed and modelled parameters [11]. On the other hand, Gibbs sampler needs more work than MH algorithm, particularly in point evaluations of the posterior density [12]. Following from these reasons, MH algorithm is applied in this study.

The assumption of independent and identically distributed data sets in extreme event seems challengeable nowadays. The presence of trend is violated under such assumptions. River management, global warming, land use and nature protections are the leading causes of human-induced non-stationarity in flood frequency analysis [13]. A suitable covariate will improve the model fitting and reduces the modelling uncertainty [14]. Generally, location parameter and scale parameter are assumed to be time dependent or other covariate dependent; shape parameter is fixed and constant because it is difficult to reliably estimate the value [15]. The presence of trends in the data may influence the design values estimation [16]. However, argument about the data used in non-stationary model may bias results since the past condition and present situation may not be similar [13].

The objective of this study is to compare the stationary and non-stationary models in modelling annual maximum river flows in Sabah. In particular, the annual maximum series data of river flow from nine sites are fitted into the GEV distribution with Bayesian MCMC approach as the parameter estimation. Here, MLE is treated as the initial value. Scale parameter in non-stationary model is accordance with the linear time dependent. 95% credible interval is computed for every parameter estimates to examine the impact of additional covariate. Lastly, we obtain the return level estimation from a suitable model.

2. Research Methodology

Most of the traditional statistical analyses study the body of the underlying distribution. However, extreme value theory (EVT) focuses on the tail distribution over a certain period of time [17].

2.1 Generalized extreme value distribution

According to EVT, the sequence of random variables is said to be independent and identically distributed (i.i.d.) if each of the random variable has the same probability distribution as the others and all of them are mutually independent [18]. 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

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

then, G belongs to one of the families of GEV distribution which have cumulative distribution function as below

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

Here,μ, σ andζ are the location, scale and shape parameters in GEV distribution, respectively. (2) is the generalized distribution of Weibull distribution$(\zeta < 0)$, Gumbel distribution$(\zeta = 0)$ and Fréchet distribution$(\zeta > 0)$[15].

2.2Bayesian parameter estimation

All GEV parameters $\theta = (\mu, \sigma, \zeta)$ are treated as random variables in Bayesian approach but other frequentist methods treat only one shape parameter [19, 20]. Based on the central idea of Bayes', Bayesian inference is given by

$$\pi(\theta \mid x) = \frac{L(\theta \mid x) \cdot \pi(\theta)}{\int_{\kappa} L(\theta \mid x) \cdot \pi(\theta) \partial \theta} \tag{3}$$

where x is the given observation

$L(\theta \mid x)$ is the likelihood function

κ is the parameter space of θ and

$\pi(\theta)$ denotes the normal prior distribution.

Likelihood of GEV distribution which follows from (2) under Bayesian estimation is given by

$$L(\theta; x) = f(x \mid \theta) = \prod_{i=1}^{n} f(x_i; \theta) \cdot \tag{4}$$

If we treat the denominator in (3) as a normalizing constant, then

$$\pi(\theta \mid x) \propto L(\theta \mid x) \cdot \pi(\theta). \tag{5}$$

A conclusion can be drawn from the credible interval as "The probability that θ lies in the credible interval given the observed data is at$(1 - \varepsilon)$." Advantages of prior belief is when scarcity of data occurs, performance in uncertainty analysis and accuracy of prediction are known as the benefits of Bayesian approach [4, 8, 21].

2.2.1 Markov Chain Monte Carlo

The complexity in computing the posterior distribution in (3) may be solved by using Markov Chain Monte Carlo (MCMC). MCMC is applied to estimate the probability of a function by using samples generated directly from posterior [15, 22]. Also, MCMC approach provide chains that are irreducible and aperiodic which then produces a unique stationary distribution, known as the basis of MCMC [23]. The asymptotic converge of MCMC methods suggest that the posterior distribution may be simply estimated from the samples generated provided that simulation is long enough. The burn-in period, which is the initial period of the chains, will be discarded because it is not initialized from posterior distribution. MCMC method is efficient in handling high dimensional distributions. Following from there, Markov chain is said to be reversible if detailed balance is fulfilled [24, 25].

2.2.2 Metropolis-Hastings algorithm

Metropolis-Hastings algorithm, a form of generalized rejection sampling, is used to overcome the high dimensional distributions in MCMC. Arbitrary probability rule$q(\cdot \mid \theta_i)$ generates the candidate random variable θ^* forθ_{i+1}. The movement of Markov chains is determined by a specified acceptance probability where

$$\alpha_i = \min \left\{ 1, \frac{\pi(\theta^*) q(\theta_i \mid \theta^*)}{\pi(\theta_i) q(\theta^* \mid \theta_i)} \right\}. \tag{6}$$

If probability equals toα_i, then the proposed value is accepted, otherwise, Markov chains will remain at current status θ_i. Tuning parameter θ_i is normally decided by user and might affect the performance of the sampler [26]. The steps involved are summarized as follows [27].

```
1.Initialize $\theta_0$
2.In $i$ iteration
    a. Draw a candidate $\theta^*$ from proposal distribution $q(\theta^* \mid \theta_i)$.
    b.Calculate acceptance probability.
    c. Draw $u \sim Uniform\ (0,1)$
        if $\alpha_i < u$,
        set $\theta_{i+1} = \theta^*$
        else $\theta_{i+1} = \theta_i$
3. Increment $i$ and return to step 2.
```

2.3Return level estimates

Application of model fitting is to predict the return level. Let p be the probability of the extreme events, the events are expected to exceed a return level of z_p, on average once every $\frac{1}{p}$ $(0 < p < 1)$ time periods. Return level is obtained by inverting cdf of GEV distribution in (2) given by

$$z_p = \left\{\mu - \frac{\sigma}{\zeta}\left[1 - \{-\log(1-p)\}^{\zeta}\right]\right\} \quad (7)$$

3.Results and Discussions

3.1Descriptive analysis

In this study, nine river flows with similar geographical factor are selected for the annual maximum river flow analysis. The record period is over 20 to 48 years. 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 8.00 am every day and measured in m^3s^{-1}. Table 1 summarizes the information of the selected river flows in this study. The average maximum observation is 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

3.2Model fitting

Annual maximum data series of the river flows are fitted into the GEV distribution. Based on the result in [28], Bayesian MCMC is suggested as the parameter estimation in modelling the annual maximum river flow data due to the advantage in reducing the parameter uncertainty. In stationary case, the GEV parameters are estimated by using the maximum likelihood estimators from previous work as the initial values in respective sites. The employment of the result as starting values is to make sure the chains converge efficiently, hence, avoid burn-in value [29].

In pass study, the attention is paid on location with time dependent. However, time is usually observed simultaneously in location and scale parameters [30]. In this study, the non-stationary model is built by scale parameter in accordance with the linear trend function. In order to make sure the scale parameter is always positive, application of $\ln \sigma(t) = \beta_0 + \beta_1 t$ is then needed [30-32]. The starting values in full model are initialized by using MLE method. In a nutshell, the two models are considered in this study as stated below.

$M_1 : X \sim GEV\ (\mu, \sigma, \zeta)$

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

In Model 2, location and shape parameters are constant, where t is measured in yearly unit. GEV parameters estimate for non-stationary model are shown in Table 2.

95% credible interval with respect to each parameter is shown in parenthesis. In both cases, 50000 iterations and uninformative prior distribution with large variance for each parameter $\theta \sim N(0, 1000^2)$ are used.

From the table we can observe that, the 95% credible intervals of scale parameters with time-varying function from all river sites contain zero except Site 3, Site 8 and Site 9. This suggests that there are not enough evidence to prove that there exists a linear trend in the scale parameter. In other words, the additional new parameter has no impact at most of the river sites except Site 3, Site 8 and Site 9. However, the width of credible intervals of the three sites are narrow and almost near to zero at Site 8 and Site 9. There is only a little evidence to prove the existing of trend in the corresponding parameter. Hence, we can conclude that a simple model is fair enough to describe the data sets in this study.

3.3Return level estimate

Return level is employed to predict the probability of river flows exceeding the maximum observations shown in Table 1. A 10- and 100-year return level estimate is obtained by substituting p=0.1 and p=0.01 into (7). Since the posterior distributions are right skewed, posterior medians are considered instead of posterior mean. Return level estimates with 95% credible interval for each site are shown in Table 3.

Upon comparing to the maximum observations in Table 1, only Site 3 is expected to exceed the maximum observations on average once in every 10-year. The maximum river flows of Site 1, Site 2, Site 3 and Site 6 are expected to exceed on average once in every 100-year.

4. Conclusions

In this study, GEV distribution is fitted into nine annual maximum river flows in Sabah. Both stationary and non-stationary cases are discussed. Based on previous simulation results and the advantages of Bayesian MCMC approach, Bayesian MCMC approach is hence chosen to estimate the parameters. MLE parameter estimates in earlier works are treated as starting values for Bayesian MCMC approach. Note that, this technique is consistent with [28]. 95% credible interval suggests that there is only little evidence to support the existence of time-varying function in the scale parameter at Site 3, Site 8 and Site 9. Hence, extra complexity to the model is unnecessary. A stationary model is found suitable for probability distribution in describing the behaviour of the selected river flows. Hence, return level estimates are obtained by using a simpler model. Most of the river flows are not expected to exceed, on average 10-year, except Site 3. Moreover, Site 1, Site 2, Site 3 and Site 6 are expected to exceed on an average once in every 100-year. In conclusion, modelling annual maximum river flow in Sabah using GEV distribution with Bayesian MCMC method as parameter estimation seems reasonable. For future study, a model which treats both location and scale parameters in accordance with linear trend functions may be studied. The trend parameter estimates can be improved by using sandwich estimator as discussed in [33].

Table 2. GEV parameters estimates for Model 2

Site	$\hat{\mu}$ (95% credible interval)	$\hat{\beta}_0$ (95% credible interval)	$\hat{\beta}_1$ (95% credible interval)	$\hat{\zeta}$ (95% credible interval)
1	20.88	9.68	0.25	0.05
	(15.03,27.50)	(0.05,19.54)	(-0.34,1.00)	(-0.49,0.77)
2	21.56	7.48	-0.08	0.08
	(18.95,24.12)	(0.00,0.16)	(-0.66,0.87)	(-0.37,0.60)
3	863.30	392.05	-8.96	-0.20
	(788.60,925.37)	(256.78,552.59)	**(-14.48,-2.98)**	(-0.43,0.10)
4	126.80	45.86	0.66	-0.07
	(108.12,146.54)	(25.04,78.53)	(-0.33,1.56)	(-0.31,0.20)
5	114.84	35.85	0.94	0.26
	(98.23,132.58)	(17.20,62.54)	(-0.12,1.74)	(0.03,0.53)
6	204.35	86.26	0.06	0.11
	(175.97,236.53)	(51.01,136.64)	(-1.58,1.55)	(-0.21,0.51)
7	120.38	50.46	0.30	0.06
	(103.18,138.19)	(29.87,78.70)	(-0.55,1.06)	(-0.10,0.27)
8	255.60	57.14	2.35	0.16
	(225.11,288.17)	(29.45,97.84)	**(0.98,3.86)**	(-0.02,0.39)
9	206.08	52.11	2.24	0.16
	(177.59,239.54)	(20.55,96.85)	**(0.77,3.77)**	(-0.07,0.46)

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

Site	p=0.1	p=0.01
1	42.26	**62.89**
	(15.84,69.12)	(17.08,242.15)
2	40.25	**58.67**
	(21.43,65.86)	(21.64,196.50)
3	**1563.74**	**1995.96**
	(1317.53,1967.97)	(1628.86,3270.46)
4	219.54	300.84
	(177.36,273.42)	(234.99,433.99)
5	219.92	409.85
	(162.37,305.66)	(265.21,792.89)
6	411.04	**666.69**
	(312.95,673.01)	(431.29,2071.86)
7	238.86	377.90
	(191.98,304.14)	(288.31,577.25)
8	404.99	626.34
	(321.19,533.17)	(440.46,1055.66)
9	342.32	542.94
	(246.81,476.88)	(346.81,1016.70)

Acknowledgement

Sincerely thanks to the reviewers in polishing this paper. This research is supported by the grant from UMS Great 2017.

References

[1] S. R. Noor, Y. Zulkifli, “Frequency analysis of annual maximum flood for Segamat river”, in *International Symposium on Civil and Environmental Engineering 2016 (ISCEE 2016)*, Melaka, Malaysia, vol. 103, pp. 1-9, 2017.

[2] A. M. J. Nur, S. Ani, “Estimating distribution parameters of annual maximum streamflows in Johor, Malaysia using TL-moments approach”, *Theor Appl Climatol*, vol. 127, no. 1-2, pp. 213-227, 2017.

[3] S. Coles, L. R. Pericchi, S. Sisson, “A fully probabilistic approach to extreme rainfall modelling”, *J. Hydrol.,* 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 J. Civil Eng,* vol. 17, no. 6, pp. 1502-1514, 2013.

[5] W. Zhang, W. 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 Eng*, vol. 129, pp. 117-132, 2017.

[6] Z. Ouyang, X. Yang, "Extreme value flood frequency analysis at water Gauging station near the Dongting lake", in *2011 International Symposium on Water Resource and Environmental Protection*, pp. 592-294, 2011.

[7] K. S. Lee, S. U. Kim, "Identification of uncertainty in low flow frequency analysis using Bayesian MCMC method", *Hydrological Processes*, vol. 22, no. 12, pp. 1949-1964, 2008.

[8] 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.

[9] A. T. Silva, M. M. Portela, M. Naghettini, W. Fernandes, "A Bayesian peaks-over-threshold analysis of floods in the Itajaí-açu River under stationarity and nonstationarity", *Stoch Environ Res Risk Assess,* vol. 31, no. 1, pp. 185-204, 2017.

[10] A. Eli, M. Shaffie, W. Z. Wan Zin, "Preliminary study on Bayesian extreme rainfall analysis: a case study of Alor Setar, Kedah, Malaysia", *Sains Malaysiana*, vol. 41, no. 11, pp. 1403-1410, 2012.

[11] P. K. Panday, C. A. Williams, K. E. Frey, M. E. Brown, "Application and evaluation of a snowmelt runoff model in the Tamor River basin, Eastern Himalaya using a Markov Chain Monte Carlo (MCMC) data assimilation approach", *Hydrological Processes,* vol. 28, no. 21, pp. 5337-5353, 2014.

[12] M. A. El-Sayed, M. M. Mohie El-Din, S. Danial, F. H. Riad, "Algorithms of credible intervals from generalized extreme value distribution based on record data", *Int. J. Stat and Applications 2017,* vol. 7, no. 4, pp. 215-221, 2017.

[13] 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", *J. Hydrology*, vol. 528, pp. 490-502, 2015.

[14] P. Jonathan, K. Ewans, "Statistical modelling of extreme ocean environments for marine design: A review", *Ocean Eng,* vol. 62, pp. 91-109, 2013.

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

[16] J. M. Cunderlik, D. H. Burn, "Non-stationary pooled flood frequency analysis", *J. Hydrology,* vol. 276, no. 1-4, pp. 210-223, 2003.

[17] R. Minkah, "An application of extreme value theory to the management of a hydroelectric dam", *SpringerPlus,* vol. 5, no. 96, pp. 1-12, 2016.

[18] J. Blanchet, C. Marty, M. Lehning, "Extreme value statistics of snowfall in the Swiss Alpine region", *Water Resources Research,* vol. 45, no. 5, pp. 1-12, 2009.

[19] H. A. Saadi, F. Ykhlef, A. Guessoum, "MCMC for parameters estimation by Bayesian approach", in *Eighth International Multi-Conference on Systems, Signals & Devices,* pp. 1-6, 2016.

[20] S. Y. 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.

[21] M. H. Chen, Q. M. Shao, J. G. Ibrahim, *Monte Carlo Methods in Bayesian Computation*, Springer-Verlag, New York, 2000.

[22] W. A. Link, R. J. Barker, *Bayesian Inference with Ecological Applications*, Elsevier, London, 2009.

[23] S. P. Brooks, "Markov Chain Monte Carlo Method and its application", *J. Royal Statistical Society. Series D (The Statistician),* vol. 47, no. 1, pp. 69-100, 1998.

[24] J. S. Rosenthal, "Markov Chain Monte Carlo algorithms: Theory and Practice", in *Monte Carlo and Quasi-Monte Carlo Methods 2008, Springer*, Berlin, Heidelberg, pp. 157-169, 2009.

[25] G. O. Roberts, J. S. Rosenthal, "General state space Markov chains and MCMC algorithms", *Probability Surveys,* vol. 1, pp. 20-71, 2004.

[26] D. van Ravenzwaaij, P. Cassey, S. D. Brown, "A simple introduction to Markov Chain Monte-Carlo sampling", *Psychonomic Bulletin & Review*, vol. 25, no. 1, pp. 143-154, 2016.

[27] C. Andrieu, N. De Freitas, A. Doucet, M. I. Jordan, "An introduction to MCMC for machine learning", *Machine Learning,* vol. 50, no. 1-25, pp. 5-43, 2003.

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

[29] M. O. Isikwue, S. B. Onoja, D. S. Naakaa, "Classical and Bayesian Markov Chain Monte Carlo (MCMC) modeling of extreme rainfall (1979-2014) in Makurdi, Nigeria", *International Journal of Water Resources and Environmental Engineering*, vol. 7, no. 9, pp. 123-131, 2015.

[30] S. El Adlouni, T. B. M. J. Ouarda, X. Zhang, R.Roy, B. Bobée, "Generalized maximum likelihood estimators for the nonstationary generalized extreme value model", *Water Resources Research,* vol. 43, no. 3, pp. 1-13, 2007.

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

[32] J. D. Salas, J. Obeysekera, "Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events", *J. Hydrol Eng,* vol. 19, no. 3, pp. 554-568, 2014.

[33] D. Gabd, J. Tawn, "Inference for an extreme value model accounting for inter-site dependence", *AIP Conf. Proc*, vol. 1830, no. 1, pp. 1-8, 2017.