Modelling maximum river flow by using Bayesian Markov Chain Monte Carlo

R Y Cheong and D Gabda

Department of Mathematics with Economics, Faculty of Science and Natural Resources
University Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia
E-mail: cheongry30@gmail.com

Abstract. Analysis of flood trends is vital since flooding threatens human living in terms of financial, environment and security. The data of annual maximum river flows in Sabah were fitted into generalized extreme value (GEV) distribution. Maximum likelihood estimator (MLE) raised naturally when working with GEV distribution. However, previous researches showed that MLE provide unstable results especially in small sample size. In this study, we used different Bayesian Markov Chain Monte Carlo (MCMC) based on Metropolis-Hastings algorithm to estimate GEV parameters. Bayesian MCMC method is a statistical inference which studies the parameter estimation by using posterior distribution based on Bayes’ theorem. Metropolis-Hastings algorithm is used to overcome the high dimensional state space faced in Monte Carlo method. This approach also considers more uncertainty in parameter estimation which then presents a better prediction on maximum river flow in Sabah.

1. Introduction

Extreme value theory (EVT) is a branch of statistic dealing with statistical techniques for modelling and estimation of rare events [1]. In general, EVT study the tail estimation instead of the body of the underlying distribution. EVT has been widely applied in financial markets, survival analysis, risk management and geophysical variables [2, 3, 4]. For geophysical processes, EVT is applied for better natural disasters’ preparedness, prevention and mitigation. For example in wind speed analysis [5, 6], modelling and predicting earthquake magnitudes [7], extreme temperature analysis [8, 9, 10] and extreme snowfall analysis [11].

Flooding was the costliest natural disaster for four consecutive years [12]. An accurate estimation of extreme flows with given return period is important. An overestimation of flood magnitude may lead to waste of investment; however, an underestimation of flood potential will lead to severe damages and casualty [13, 14]. In previous researches, the annual maximum river flow was used as an indicator for the analysis of flood trends. Therefore a good understanding of the probability distributions of river flows may be useful for water resource planning and management. However, there is no standardised technique in studying this. The objective of this study is to apply generalized extreme value (GEV) distribution in modelling maximum river flow in Sabah by using Bayesian Markov Chain Monte Carlo based on Metropolis-Hastings algorithm.

2. Research methodology

Extreme value theory focuses on the statistical behaviour of $M_n = \max\{X_1, ..., X_n\}$ where $X_1, ..., X_n$ is a sequence of independent and identically distributed variables [15]. GEV distribution having non-degenerate distribution function that fulfill
\[
\Pr \left( \frac{M_n - b_n}{a_n} \leq z \right) \to G(z) \text{ as } n \to \infty
\] (1)

where \( \{a_n > 0\} \) and \( \{b_n\} \) as normalising constant. The cumulative distribution function (cdf) of GEV distribution is denoted as followed:

\[
G_{\xi, \mu, \sigma}(x) = \exp \left\{ \left[ 1 + \frac{(x - \mu)}{\sigma} \right]^{\frac{1}{\tau}} \right\} \text{ with } 1 + \frac{(x - \mu)}{\sigma} > 0
\] (2)

Location \((\mu)\) scale \((\sigma)\) and shape \((\xi)\) are the parameters in GEV distribution. Equation (2) is a generalized distribution of Fréchet \((\xi < 0)\), Gumbel \((\xi = 0)\) and Weibull distribution \((\xi > 0)\)[16].

In this study, the data were divided into annually block as stated in [17] that a block of \( n \) is usually a year in hydrology events. Hence, \( M_n \) corresponds to the annual maxima.

2.1. Maximum likelihood estimation (MLE)

Maximum likelihood method is based on maximizing the likelihood of the observed sample; hence adopt to the model [18]. Logarithm of a function achieves maximum value at the same point; thereby log-likelihood tends to be much simpler than likelihood by solving the system of equation (3).

\[
\frac{\partial \ell(\theta \mid x)}{\partial \theta_j} = 0, \ j = 1,2,\ldots,k
\] (3)

where \( k \) is the dimension of the vector \( \theta \) using R software [19]. The corresponding log-likelihood is

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

When the sample size grows to infinity, the MLE is said to be a consistent estimator and the variance goes to zero. The asymptotic theory allows MLE to be normally distributed as the sample size increase. In [1] and [20], MLE was chosen due to the stable performance in large sample size \((n > 50)\).

2.2. Bayesian estimation

Bayesian approach improves the estimation accuracy by assuming the parameter \( \theta = (\mu, \sigma, \xi) \) as random variable [21]. Bayesian inference, based on central idea of Bayes’ theorem, which as follows,

\[
\pi(\theta \mid x) = \frac{L(\theta \mid x) \cdot \pi(\theta)}{\int_{\omega} L(\theta \mid x) \cdot \pi(\theta) d\theta}
\] (5)

where \( x \) is the given observations, \( L(\theta \mid x) \) denote likelihood function, \( \pi(\theta) \) is the normal prior distribution, \( \omega \) denote the parameter space. Bayesian inference is applied to calculate and display marginal posterior densities which provide complete information about parameters of interest [22]. The Bayesian inference is hence leading to the likelihood function \( L(\theta \mid x) \).

\[
L(\theta; x) = f(x \mid \theta) = \prod_{i=1}^{n} f(x_i; \theta)
\] (6)
The denominator in equation (5) is treated as a normalizing constant so that the posterior distribution integrated to one. Thereby, result \( \pi(\theta | x) \propto L(\theta | x) \cdot \pi(\theta) \). Once the posterior distribution is concerned, the normalizing constant is ignored.

Bayesian inference is desirable in extreme value due to ability of owing to scarcity of data through a prior distribution and does not adapt to the asymptotic theory. Markov Chain Monte Carlo is employed to solve the complex computational of posterior distribution in equation (5).

2.2.1. Markov Chain Monte Carlo (MCMC). MCMC is used to examine probability distribution [23] by using samples generated from posterior directly [13]. MCMC method provides the chain that satisfies ergodicity theorem. This implies that the chain is irreducible and aperiodic that produces a unique stationary distribution [24]. Based on the asymptotic converge of MCMC methods, for sufficient long simulation, the posterior distribution can be estimated easily by using the samples generated [25]. Since the initial sample is not from posterior distribution, the respective initial period is called as burn-in period which are usually been discarded. Besides, if the Markov chain transition distribution fulfils the detailed balance, then is considered as reversible Markov chain [26].

2.2.2. Metropolis-Hastings algorithm. Metropolis-Hastings algorithm is a form of generalized rejection sampling. The proposed value, \( \theta^* \) for \( \theta_{i+1} \) is generated from arbitrary probability rule \( q(\cdot | \theta_i) \). The Markov chain moves to \( \theta^* \) with a specified acceptance probability [15]. Specifically, let

\[
\alpha_i = \min \left(1, \frac{\pi(\theta^*) q(\theta_i | \theta^*)}{\pi(\theta_i) q(\theta^* | \theta_i)} \right)
\]

where \( q(\theta^* | \theta_i) \) is denoted as proposal distribution and \( \theta_i \) is the tuning parameter that determined by user [27] which will influence the performance of sampler. The candidate is accepted if the probability is equal to \( \alpha_i \), otherwise the Markov chain remains as current state \( \theta_i \). The steps involved can be illustrated into the following algorithm [28]:

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

2.3. Return level estimate

Return level \( Z_p \) expected to be exceeded on average once every \( \frac{1}{p} \) \( (0 < p < 1) \) period, where \( p \) is the probability of the extreme event. Equation 8 shows the return level estimate of GEV distribution.

\[
Z_p = \left\{ \mu - \frac{\alpha}{\xi} \left[ 1 - \left( \ln (1 - p) \right)^\xi \right] \right\}
\]
3. Results and discussion

3.1. Descriptive analysis
In this study, the secondary data were obtained from Hydrology and Survey Division under Department of Irrigation and Drainage, Sabah. 18 stations with small sample size were chosen (n < 50). The data were collected as daily mean of 24-hour periods beginning at 8.00 am each day. Table 1 shows the number of annual maxima, the period and the maximum observation for each site.

Table 1. The number of years, period and maximum value for each site.

| Site          | Station         | No. of years | Period          | Maximum observation |
|---------------|-----------------|--------------|-----------------|---------------------|
| 1             | Sg. Sapulut     | 27           | 1990-2016       | 874.70              |
| 2             | Sg. Bongon      | 29           | 1988-2016       | 595.28              |
| 3             | Sg. Tungad      | 30           | 1986-2015       | 903.25              |
| 4             | Sg. Kinabatangan at Pagar | 31 | 1986-2016 | 2213.90 |
| 5             | Sg. Sugut       | 32           | 1984-2015       | 4528.00             |
| 6             | Sg. Mengalong   | 33           | 1984-2016       | 453.33              |
| 7             | Sg. Padas at JPS Beaufort | 35 | 1981-2015 | 1506.30 |
| 8             | Sg. Kinabatangan at Balat | 38 | 1978-2015 | 3506.80 |
| 9             | Sg. Bengkoka    | 45           | 1972-2016       | 2016.40             |
| 10            | Sg. Kuamut      | 47           | 1969-2015       | 3142.90             |
| 11            | Sg. Milian      | 47           | 1969-2015       | 2396.90             |
| 12            | Sg. Sook        | 47           | 1969-2015       | 313.99              |
| 13            | Sg. Wariu       | 47           | 1969-2015       | 524.90              |
| 14            | Sg. Labuk       | 48           | 1969-2016       | 2901.30             |
| 15            | Sg. Padas at Kemabong | 48 | 1969-2016 | 1644.30 |
| 16            | Sg. Papar at Kaiduan | 48 | 1969-2016 | 468.86 |
| 17            | Sg. Papar at Kogopon | 48 | 1969-2016 | 970.30 |
| 18            | Sg. Pegalan     | 48           | 1969-2016       | 688.63              |

3.2. Simulation
We performed a simulation study before real data analysis. The extreme event simulated from the GEV distribution \( \mu = 0, \sigma = 0, \zeta = 0 \). Then the simulated data were fitted into GEV distribution and the parameters were estimated by MLE and Bayesian MCMC approach. Bayesian MCMC method used uninformative normal prior distribution, \( N \sim (0,1000^2) \) for all parameters. The results of MLE and Bayesian estimates are shown in table 2. The respective trace plot and density plot are shown in figure 1. Both methods gave estimates that similar to actual parameters hence verified our R code. We then continue the analysis in modelling the annual maximum river flow in Sabah.

![Figure 1](image1.png)

Figure 1. (a), (b) and (c) show the trace plot for \( \mu, \sigma \) and \( \xi \) respectively using Bayesian MCMC approach. 50000 iterations were carried out and first 5000 iterations were discarded where the chains start to converge. Hence, (d), (e) and (f) show the density plot for \( \mu, \sigma \) and \( \xi \) respectively after 15 subsample feed.
3.3. Parameter estimation

In real data analysis, the starting value of MLE was getting by employing probability weighted moment. The result of MLE was then treated as the starting value in Bayesian MCMC method. This is to avoid burn-in period in Bayesian MCMC. Table 3 shows the results of both parameter estimations.

| Site | GEV parameter | Maximum likelihood estimation (s.e) | Bayesian MCMC (credible interval) |
|------|---------------|-------------------------------------|----------------------------------|
|      | μ (s.e)       | σ (s.e)                             | μ (95% credible interval)        |
|      | ξ (s.e)       |                                     | σ (95% credible interval)        |
|      |               |                                     | ξ (95% credible interval)        |
| 1    | 452.76 (141.28) | -0.15                              | 450.58 (113.16, 212.56)         |
|      | (30.47)       | (0.14)                             | (10.39, 0.18)                   |
| 2    | 112.62 (80.01)  | -0.32                              | 113.97 (88.45)                  |
|      | (17.23)       | (0.18)                             | (0.34)                          |
| 3    | 468.93 (271.73)| -0.56                              | 456.79 (278.53)                 |
|      | (56.80)       | (0.20)                             | (0.47)                          |
| 4    | 1268.6 (363.73)| -0.27                              | 1254.36 (388.20)               |
|      | (73.52)       | (0.13)                             | (0.28)                          |
| 5    | 674.70 (403.87)| 0.23                               | 670.32 (431.38)                 |
|      | (82.18)       | (0.16)                             | (0.29)                          |
| 6    | 212.40 (101.32)| -0.34                              | 209.70 (107.92)                 |
|      | (19.36)       | (0.11)                             | (0.31)                          |
| 7    | 822.51 (218.11)| -0.18                              | 818.79 (231.66)                 |
|      | (40.72)       | (0.11)                             | (0.15)                          |
| 8    | 1258.31 (354.35)| 0.11                             | 1253.24 (377.17)                |
|      | (64.13)       | (0.11)                             | (0.14)                          |
| 9    | 281.52 (224.06)| 0.32                               | 277.09 (232.72)                 |
|      | (41.54)       | (0.19)                             | (0.37)                          |
| 10   | 872.18 (354.48)| 0.36                               | 864.74 (367.85)                 |
|      | (62.46)       | (0.17)                             | (0.40)                          |
| 11   | 1859.87 (285.69)| -0.047                           | 1855.06 (299.94)                |
|      | (45.58)       | (0.074)                            | (0.022)                         |
| 12   | 1277.86 (61.16)| -0.19                              | 128.34 (63.61)                  |
|      | (9.86)        | (0.098)                            | (0.17)                          |
| 13   | 113.78 (58.54)| 0.090                              | 114.42 (61.82)                  |
|      | (9.27)        | (0.078)                            | (0.10)                          |
| 14   | 918.55 (484.13)| -0.022                            | 924.22 (510.63)                 |
|      | (78.00)       | (0.10)                             | (0.017)                         |
| 15   | 651.34 (256.44)| -0.063                            | 653.55 (270.75)                 |
|      | (40.63)       | (0.083)                            | (0.057)                         |
| 16   | 120.18 (54.78)| 0.0027                             | 119.97 (57.48)                  |
|      | (8.54)        | (0.065)                            | (0.018)                         |
| 17   | 256.72 (119.89)| 0.020                             | 256.48 (126.04)                 |
|      | (18.65)       | (0.067)                            | (0.031)                         |
| 18   | 222.31 (111.94)| -0.0178                           | 222.17 (117.68)                 |
|      | (17.96)       | (0.097)                            | (0.0073)                        |

Table 2. Result of simulation with set seed =100.

Table 3. The maximum likelihood estimates (standard error) and posterior means for GEV parameters.
By observing the lower and upper quantile of Bayesian MCMC, the maximum likelihood estimator was lie between the 95% credible intervals. This implies that both methods produced consistent estimates. Besides, Bayesian MCMC method produced the parameter distribution which allows probability statements to be made including summaries of the uncertainty of the parameters. Allowance for parameter uncertainty can improve the return level estimate [29]. Hence, Bayesian MCMC was chosen in return level estimation.

3.4. Return level estimates
In this study, the return levels for each site were obtained from Bayesian MCMC approach. The Bayesian MCMC estimators were substituted with $p=0.1, 0.01$ to estimate the 10 and 100-year of return levels for each site. Due to the right skew posterior distribution, the posterior medians were concerned with 95% credible interval. These estimates are shown in table 4.

Table 4. Estimates for 10,100-year return levels (m$^3$) at each site using Bayesian MCMC.

| Site | Return period (years) | Site | Return period (years) |
|------|----------------------|------|----------------------|
|      | 10                   | 10   | 100                  |
| 1    | 743.08               | (656.73,909.60) | 2147.61               | (1688.52,3395.58) | 5354.69               | (3004.80,18888.14) |
| 2    | 396.11               | (281.31,741.88) | 1003.21               | (816.97,1558.84) | 2123.4               | (529.42,4352.13) |
| 3    | 838.31               | (757.66,9988.08) | 1925.9               | (953.57,1470.95) | 346.23               | (1341.16,1761.81) |
| 4    | 1925.9               | (2306.68,3395.58) | 2123.4               | (529.42,4352.13) | 346.23               | (1341.16,1761.81) |
| 5    | 1980.81              | (1754.80,2231.74) | 346.23               | (2085.33,3321.95) | 346.23               | (1341.16,1761.81) |
| 6    | 382.84               | (1485.34,3366.62) | 346.23               | (2085.33,3321.95) | 346.23               | (1341.16,1761.81) |
| 7    | 1253.49              | (342.76,444.067)  | 346.23               | (2085.33,3321.95) | 346.23               | (1341.16,1761.81) |
| 8    | 2222.5               | (1141.95,1442.59) | 346.23               | (2085.33,3321.95) | 346.23               | (1341.16,1761.81) |
| 9    | 1060.94              | (1924.76,2842.77) | 346.23               | (2085.33,3321.95) | 346.23               | (1341.16,1761.81) |
|      | (785.59,1806.80)     | (2722.51,6670.04) | 346.23               | (2085.33,3321.95) | 346.23               | (1341.16,1761.81) |

As comparing the maximum river flow in Table 1 and return level estimates, the river flow which exceeded the maximum river flow of the observation period started after $p = 0.01$ for most of the cases except Site 5, Site 11, Site 13, Site 16 and Site 17.

4. Conclusion
In this study, we fit GEV distribution to model the annual maximum river flow in Sabah. We used an alternative method of Bayesian MCMC based on Metropolis-Hastings algorithm to estimate the parameters. This method produces the distribution of the parameter. For application, we could estimate for 10 and 100-year return level for each site. This will help in water and flood management for any return period. We will study non-stationary model to the annual maximum river flow in Sabah in the future. Logistic model is another suggested model in modelling maximum river flow.

References
[1] Minkah R 2016 SpringerPlus 5 1-12
[2] Gilli M and Këllezi E 2006 Computational Economics 27 207-228
[3] Adesina O, Ismail I and Oladeji T 2016 the J. of Risk Management and Insurance 20 40-51
[4] Lanzarone, Pasquali S, Gilioli G and Marchesini E 2017 J. Mathematical Biology 1-21
[5] Soukissian T H and Tsalis 2015 *Natural Hazards* **78** 1777-1809
[6] Rajabi M R and Modarres R 2008 *J. Wind Engineering and Industrial Aerodynamic* **96** 78-82
[7] Pisarenko V F, Sornette A, Sornette D and Rodkin M V 2014 *Pure Appl. Geophys.* **171** 1599-1624
[8] Hasan H, Ahmad Radi N F and Kassim S 2012a *Proc. World Congress on Engineering 2012* vol 1 (London, U.K.) pp 3-8
[9] Hasan H, Salam N and Kassim S 2012b 2012 (ICSSBE) doi:10.1109/ICSSBE.2012.6396634
[10] Lee J 2017 *Asia-Pac J. Atmos. Sci.* **53** 31-41
[11] Blanchet J, Marty C and Lehning M 2009 *Water Resources Research* **45** 1-12
[12] Aon Benfield 2017 *Annual Global Climate and Catastrophe Report* (London: Aon Benfield)
[13] Saghañian B, Goliand S and Ghasemi A 2014 *Natural Hazards* **71** 403-417
[14] Ellouze M and Abida H 2008 *Water Resources Management* **22** 943-957
[15] Coles S 2001 *An Introduction to Statistical Modelling of Extreme Values* (London: Springer-Verlag)
[16] Kotz S and Nadarajah S 2000 *Extreme Value Distribution, Theory and Applications* (London: Imperial College Press)
[17] Maposa D, Cochran J J, Lesaoana M and Sigauke C 2014 *Nat. Hazards Earth Syst. Sci. Discuss* **2** 5401-25
[18] Castillo E, Hadi A S, Balakrishan N and Sarabia J M 2005 *Extreme Value and Related Models with Applications in Engineering and Science* (Canada: John Wiley & sons)
[19] Yoon S Y, Cho W C and Heo J H 2010 *Stoch Environ Res Risk Assess* **24** 761-770
[20] Machado M J, Botero B A, López J, Francés F, Diez-Herrero A and Benito G 2015 *Nat. Hazards Earth Syst. Sci.* **14** 1543-51
[21] Saadi H A, Ykhlef F and Guessoum A 2011 8th *Int. Multi-Conf on Systems, Signals & Devices*
[22] Chen M H, Shao Q M and Ibrahim J G *Monte Carlo Methods in Bayesian Computation* (New York: Springer-Verlag)
[23] Link W A and Barker R J 2009 *Bayesian Inference with Ecological Applications* (London: Elsevier)
[24] Brooks S P 1998 *J. Royal Statistical Society Series D (The Statistician)* **47** 69-100
[25] Rosenthal J S 2009 *Monte Carlo and Quasi-Monte Carlo Methods 2008* ed L’Ecuyer P and Owen A B (Heidelberg: Springer) pp 157-169
[26] Roberts G O and Rosenthal J S 2004 *Probability Surveys* **1** 20-71
[27] van Ravenzwaaij D, Cassee P and Brown S D 2016 *Psychon Bull Rev* **1-12
[28] Andrieu C, de Freitas N, Doucet A and Jordan M I 2003 *Machine Learning* **50** 5-43
[29] Coles S, Pericchi L R and Sisson S 2003 *J. Hydrology*