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

R Y Cheong¹, Darmesah Gabda²*
¹Universiti Malaysia Sabah, Faculty of Science and Natural Resources, Department of Mathematics with Economics
²Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Department of Mathematics with Economics
*Corresponding author email: riving93@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 \( \{ a_n, b_n \} \) and \( \{ \xi_n \} \), \( G \) is non-degenerate distribution function such that

\[
\Pr \left\{ \left( \frac{M_n - b_n}{a_n} \right) \leq z \right\} \to G(z), n \to \infty
\]

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

\[
G_{\xi, \mu, \sigma}(x) = \exp \left\{ -\left[ \left( 1 + \frac{x - \mu}{\sigma} \right)^{\frac{1}{\xi}} \right] \right\}, \quad \text{for } \xi \neq 0
\]

\[
\left\{ \begin{array}{ll}
1 & \text{if } \xi = 0, \\
\left( 1 + \frac{x - \mu}{\sigma} \right)^{\frac{1}{\xi}} & \text{if } \xi > 0,
\end{array} \right.
\]

Here, \( \xi, \mu, \sigma \) and \( \xi \) are the location, scale and shape parameters in GEV distribution, respectively. (2) is the generalized distribution of Weibull distribution (\( \xi > 0 \)), Gumbel distribution (\( \xi = 0 \)) and Fréchet distribution (\( \xi < 0 \)) [15].

2.2 Bayesian parameter estimation

All GEV parameters \( \theta = \{ \mu, \sigma, \xi \} \) 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

\[
\pi(\theta | x) \propto L(\theta | x) \cdot \pi(\theta)
\]

(3)

where \( x \) is the given observation

\( L(\theta | x) \) is the likelihood function

\( \kappa \) is the parameter space of \( \theta \) and

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

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

\[
L(\theta; x) = f(x | \theta) \cdot \prod_{i=1}^{n} f(x_i; \theta).
\]

(4)

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

\[
\pi(\theta | x) \propto L(\theta | x) \cdot \pi(\theta).
\]

(5)

A conclusion can be drawn from the credible interval as “The probability that \( \theta \) lies in the credible interval given the observed data is at(1 - \( c \)).” 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 \( q(\theta' | \theta) \) generates the candidate random variable \( \theta' \) for \( \theta_i \). The movement of Markov chains is determined by a specified acceptance probability where

\[
a_i = \min \left\{ 1, \frac{\pi(\theta') q(\theta | \theta')}{\pi(\theta) q(\theta' | \theta)} \right\}.
\]

(6)

If probability equals to \( a_i \), then the proposed value is accepted, otherwise, Markov chains will remain at current status \( \theta \). Tuning parameter \( \theta_i \) 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 \( q(\theta' | \theta) \).

   b. Calculate acceptance probability.

   c. Draw \( u \sim \text{Uniform} \ (0,1) \)

      \( \text{if} \ a_i < u, \)

      set \( \theta_{i+1} = \theta' \)

      else \( \theta_{i+1} = \theta_i \)

3. Increment \( i \) and return to step 2.

2.3 Return level estimates

Application of model fitting is to predict the return level. Let \( \rho \) 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

\[
z_p = \left\{ \mu - \frac{\sigma}{\xi} \left[ -\log (1 - p) \right]^{\frac{1}{\xi}} \right\}
\]

(7)

3. Results and Discussions

3.1 Descriptive 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 \( \text{m}^3\text{.s}^{-1} \). Table 1 summarizes the information of the selected river flows in this study. The average maximum observation is 40 years.

<table>
<thead>
<tr>
<th>Site</th>
<th>Stations</th>
<th>No. of years</th>
<th>Period</th>
<th>Max. observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sungai Apin-</td>
<td>20</td>
<td>1996-2015</td>
<td>50.41</td>
</tr>
<tr>
<td></td>
<td>Waterworks Sungai</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Baiay at Bandukan</td>
<td>21</td>
<td>1993-2013</td>
<td>39.94</td>
</tr>
<tr>
<td></td>
<td>Sungai Padas at JPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Beaufort Sungai Sook</td>
<td>35</td>
<td>1981-2015</td>
<td>1506.30</td>
</tr>
<tr>
<td></td>
<td>at Biah Sungai Wariu</td>
<td>47</td>
<td>1969-2015</td>
<td>313.99</td>
</tr>
<tr>
<td></td>
<td>at Bridge No.2 Sungai</td>
<td>47</td>
<td>1969-2015</td>
<td>524.90</td>
</tr>
<tr>
<td>6</td>
<td>Kadamaian at Tamu Darat</td>
<td>47</td>
<td>1969-2015</td>
<td>490.20</td>
</tr>
<tr>
<td></td>
<td>Sungai Papar at Kaiduan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sungai Papar at Kogopon</td>
<td>48</td>
<td>1969-2016</td>
<td>468.86</td>
</tr>
<tr>
<td></td>
<td>Sungai at Ansip</td>
<td>48</td>
<td>1969-2016</td>
<td>688.63</td>
</tr>
</tbody>
</table>

3.2 Model 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.

\[ M_1 : X \sim \text{GEV} \left( \mu, \sigma, \xi \right) \]
\[ M_2 : X \sim \text{GEV} \left( \mu, \ln \sigma(t) = \beta_0 + \beta_1 t, \xi \right) \]

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 \( \theta \sim N(0,100^2) \) 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.

3.3 Return 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**

<table>
<thead>
<tr>
<th>Site</th>
<th>$\hat{\mu}$ (95% credible interval)</th>
<th>$\hat{\beta}_x$ (95% credible interval)</th>
<th>$\hat{\beta}_s$ (95% credible interval)</th>
<th>$\hat{\xi}$ (95% credible interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.88 (15.03,27.50)</td>
<td>9.68 (0.05,19.54)</td>
<td>0.25 (-0.34,1.00)</td>
<td>0.05 (-0.49,0.77)</td>
</tr>
<tr>
<td>2</td>
<td>21.56 (18.95,24.12)</td>
<td>7.48 (0.00,1.16)</td>
<td>-0.08 (-0.66,0.87)</td>
<td>0.08 (-0.37,0.60)</td>
</tr>
<tr>
<td>3</td>
<td>863.30 (788.60,925.37)</td>
<td>392.05 (256.78,552.59)</td>
<td>-8.96 (-14.48,-2.98)</td>
<td>-0.20 (-0.43,0.10)</td>
</tr>
<tr>
<td>4</td>
<td>126.80 (108.12,146.54)</td>
<td>45.86 (25.04,78.53)</td>
<td>-0.33 (1.56)</td>
<td>-0.31 (0.20)</td>
</tr>
<tr>
<td>5</td>
<td>114.84 (98.23,132.58)</td>
<td>35.85 (17.20,62.54)</td>
<td>0.94 (-1.12,1.74)</td>
<td>0.26 (0.03,0.53)</td>
</tr>
<tr>
<td>6</td>
<td>204.35 (175.97,236.53)</td>
<td>86.26 (51.01,136.64)</td>
<td>-1.58 (1.58,1.55)</td>
<td>-0.21 (0.51)</td>
</tr>
<tr>
<td>7</td>
<td>120.38 (103.18,138.19)</td>
<td>50.46 (29.87,78.70)</td>
<td>0.30 (-0.55,1.06)</td>
<td>0.06 (-0.10,0.27)</td>
</tr>
<tr>
<td>8</td>
<td>255.60 (225.11,288.17)</td>
<td>57.14 (29.45,97.84)</td>
<td>2.35</td>
<td>0.16</td>
</tr>
<tr>
<td>9</td>
<td>206.08 (177.59,239.54)</td>
<td>52.11 (20.55,96.85)</td>
<td>(0.98,3.86)</td>
<td>(0.02,0.39)</td>
</tr>
</tbody>
</table>

**Table 3. 10-, 100-year return level estimates (m³)**

<table>
<thead>
<tr>
<th>Site</th>
<th>p=0.1</th>
<th>p=0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42.26 (15.84,69.12)</td>
<td>62.89 (17.08,242.15)</td>
</tr>
<tr>
<td>2</td>
<td>40.25</td>
<td>58.67 (21.43,65.86)</td>
</tr>
<tr>
<td>3</td>
<td>1563.74 (1317.53,1967.97)</td>
<td>1995.96 (1628.86,3270.46)</td>
</tr>
<tr>
<td>4</td>
<td>219.54 (177.36,273.42)</td>
<td>300.84 (234.99,433.99)</td>
</tr>
<tr>
<td>5</td>
<td>219.92 (162.37,305.66)</td>
<td>409.85 (265.21,792.89)</td>
</tr>
<tr>
<td>6</td>
<td>411.04 (312.95,673.01)</td>
<td>666.69 (431.29,2071.86)</td>
</tr>
<tr>
<td>7</td>
<td>238.86 (191.98,304.14)</td>
<td>377.90 (288.31,577.25)</td>
</tr>
<tr>
<td>8</td>
<td>404.99 (321.19,533.17)</td>
<td>62.34 (440.46,1055.66)</td>
</tr>
<tr>
<td>9</td>
<td>342.32 (246.81,476.88)</td>
<td>542.94 (346.81,1016.70)</td>
</tr>
</tbody>
</table>

**Acknowledgement**

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

**References**


