Bayesian inference for extreme value flood frequency analysis in Bangladesh using Hamiltonian Monte Carlo techniques

Md Ashraful Alam[1*], *Craig* Farnham[1], and *Kazuo* Emura[1]

[1]Department of Housing and Environmental Design, Graduate School of Human Life Science, Osaka City University, Osaka, Japan

Abstract. In Bangladesh, major floods are frequent due to its unique geographic location. About one-fourth to one-third of the country is inundated by overflowing rivers during the monsoon season almost every year. Calculating the risk level of river discharge is important for making plans to protect the ecosystem and increasing crop and fish production. In recent years, several Bayesian Markov chain Monte Carlo (MCMC) methods have been proposed in extreme value analysis (EVA) for assessing the flood risk in a certain location. The Hamiltonian Monte Carlo (HMC) method was employed to obtain the approximations to the posterior marginal distribution of the Generalized Extreme Value (GEV) model by using annual maximum discharges in two major river basins in Bangladesh. The discharge records of the two largest branches of the Ganges-Brahmaputra-Meghna river system in Bangladesh for the past 42 years were analysed. To estimate flood risk, a return level with 95% confidence intervals (CI) has also been calculated. Results show that, the shape parameter of each station was greater than zero, which shows that heavy-tailed Fréchet cases. One station, Bahadurabad, at Brahmaputra river basin estimated 141,387 $m^3 \cdot s^{-1}$ with a 95% CI range of [112,636, 170,138] for 100-year return level and the 1000-year return level was 195,018 $m^3 \cdot s^{-1}$ with a 95% CI of [122493, 267544]. The other station, Hardinge Bridge, at Ganges basin estimated 124,134 $m^3 \cdot s^{-1}$ with a 95% CI of [108,726, 139,543] for 100-year return level and the 1000-year return level was 170,537 $m^3 \cdot s^{-1}$ with a 95% CI of [133,784, 207,289]. As Bangladesh is a flood prone country, the approach of Bayesian with HMC in EVA can help policy-makers to plan initiatives that could result in preventing damage to both lives and assets.

1 Introduction

In Bangladesh, major floods are frequent, due to its unique geographic location. About one-quarter to one-third of the country is inundated by overflowing rivers during the monsoon season almost every year. Bangladesh is in the active delta of three major rivers, the

* Corresponding author: alam13ocu@gmail.com

Ganges, Brahmaputra, and Meghna (GBM). The intensity and time duration of floods in Bangladesh typically depends on the GBM river system.

Extreme value analysis (EVA) is a branch of probability theory which deals with the stochastic behavior of extreme values of a set of random variables. Typically, EVA is applied for describing a rare event. The fundamental concepts, techniques, and guidelines of the application of the theory of EVA were detailed by Coles [1].

Several techniques have been recommended for parameter estimation in extreme value models. The likelihood-based technique is attractive, but the difficulty is the regularity conditions that are required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid. Subsequently, the Bayesian technique has become popular in recent years [1]. The popularity can be clarified by the rediscovery of the utility of Markov chain Monte Carlo (MCMC) algorithms in the early 1990s. The output result of a Bayesian analysis provides a more complete inference than the corresponding maximum likelihood analysis.

Besides the general algorithms of MCMC to estimate the parameters in GEV distributions, Hamiltonian Monte Carlo (HMC) algorithms are more efficient. Also, the HMC parameter estimation is relatively robust and much faster. In extreme value analysis with GEV models, avoidance of random-walk behavior is one of the major advantages of HMC [2]. The main aim of this article is to calculate a Bayesian regional model for the annual maximum discharge of two major rivers in Bangladesh, as well as flood risk estimation and extrapolation of the return period for the two rivers. For the block maxima or annual maximum discharge, the GEV distribution is used. For describing the Bayesian, the Hamiltonian Monte Carlo is used as the MCMC algorithm.

2 Data and study area

Table 1. Major characteristics of the Ganges, Brahmaputra and Meghna (GBM) river basins.

Item		**Ganges**	**Brahmaputra**	**Meghna**
Basin area (km^2)		907,000	583,000	65,000
River length (km)		2000	1800	946
Total number of dams		75	6	
Elevation (m above sea level (asl))	Area below 500 m asl:	72%	20%	75%
	Area above 3000 m asl:	11%	60%	0%
	Lowest	530	3430	2
	Highest	70,868	102,535	19,900
	Average	11,300	20,000	4,600
Land use (% area)	Agriculture	68%	19%	27%
	Forest	11%	31%	54%

In the GBM basin, most of the rivers originate from the Himalayas, north of the country, and flow through the country to the Bay of Bengal, south of Bangladesh. The basin is situated within four different countries: China (50.5%), India (33.6%), Bangladesh (8.1%)

and Bhutan (7.8%). Major characteristics of the GBM river basin are shown in Table 1 (for more details see [3] and references therein).

Bangladesh is situated between latitudes 20°30′ N and 26°45′ N, and longitude 88°0′ E and 92°45′ E (Figure 1). The total area is 147,570 km^2. Most of the country (79%) is a floodplain. There are some hilly areas (12%), which are situated in the southeast and northeast regions. The rest of the area (9%) is occupied by four uplifted blocks, which are in the northwest and central parts of the country. The climate of this area is the tropical monsoon type, with a hot summer monsoon and dry season in the winter. During the summer monsoon time, from June to October, extreme seasonal rainfall occurs. Also, in the Himalayas, the storage of water in the form of snow and ice (in glaciers) provides a large water reservoir that regulates annual water distribution.

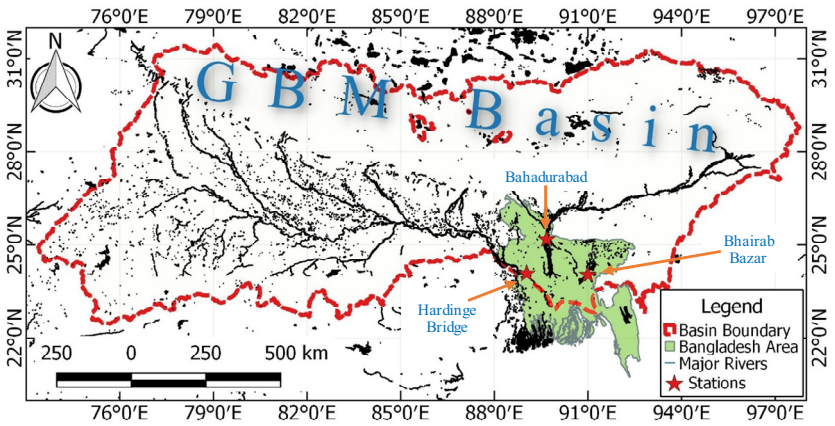


Fig. 1. The GBM basin with three outlets: Hardinge Bridge in the Ganges, Bahadurabad in the Brahmaputra and Bhairab Bazar in the Meghna river basins.

Streamflow data were collected from the Hydrology Division of the Bangladesh Water Development Board (BWDB). Taking the 3 stations closest to the outlets of the 3 rivers of the GBM system should represent the discharge of the GBM system as a whole. Station Hardinge Bridge is in the Ganges river. Station Bahadurabad is in the Brahmaputra river, and station Bhairab Bazar is in the Meghna.

In this study, 42 years of data are used. The Bhairab Bazar station on the Meghna has a large amount of missing data over the period (over 40%) and could not be useful for this analysis. Thus, only the 2 stations (Hardinge Bridge and Bahadurabad) are used in this analysis. The stations' basic information is presented in Table 2.

Table 2. The stations' basic information and description of the data set.

Station Name	River	Drainage area (km^2)	Elevation (m asl)	Observed data	Missing values (%)
Hardinge Bridge	Ganges	907,000	10	1976–2017	5
Bahadurabad	Brahmaputra	583,000	22	1976–2017	3
Bhairab Bazar	Meghna	65,000	8	1976–2017	40

3 Materials and methods

The GEV distribution is a standard tool for modeling flood peaks by using the annual maximum series (AMS). In this work, a Bayesian approach using the Hamiltonian Monte Carlo (HMC) and Metropolis Hasting (MH) algorithms was applied to estimate the parameters in the GEV model.

The GEV distribution comprises into a single form all three Extreme Value (EV) distributions: Gumbel (EV-I, $\xi = 0$), Fréchet (EV-II, $\xi > 0$), and Weibull (EV-III, $\xi < 0$) [1]. MCMC techniques describe a method of simulating from complex distribution by simulating from Markov chains which have the target distributions as their stationary distributions. There are many MCMC techniques, including HMC, which was used in this work. HMC was initially proposed by Duane [4] for simulating molecular dynamics under the name of Hybrid Monte Carlo. For an up to date review about the theoretical and practical aspects on HMC, the reader is referred to Neal [2]. The HMC method is used in the context of GEV models in this work. To create the other benchmark to assess the accuracy of the MCMC results, another MCMC method, the MH algorithm [5,6], was used to estimate the parameters of the distribution models.

Estimates of return levels of the annual maxima are of particular interest in hydrologic extremes, as they give an estimate of the level the process is expected to exceed once, on average, in a given number of years. For the GEV distribution model, the return levels [1] with 95% confidence intervals (CI) were also calculated and presented.

4 Results and discussion

Simulation-based techniques, MCMC such as HMC and MH, have provided a way for analyzing the Bayesian methods of extreme value data for calculating the risk level of flood frequency analysis in the major rivers in Bangladesh.

4.1 Statistical qualities of the sample data

The characteristics of the AM samples and best-fit distribution results of each station are shown in Table 3. The Hardinge Bridge station is highly positive-skewed, and the Bahadurabad station is more symmetric.

In a time series analysis, checking stationarity and homogeneity is essential. Both parametric and non-parametric methods are used to detect the statistical significance of monotonic trends or non-stationarity. The Von Neumann (VN) ratio test [7]

and standard normal homogeneity (SNH) test [8] were used to check the homogeneity of the sample data set at the 5% significance level. The SNH test and VN ratio test passed as homogeneous for both data series. The Augmented Dickey-Fuller (ADF) test [9], a type of statistical test called a unit root test, was used to check the stationarity of sample data. The time series of this sample data pass as stationary. As the analyzed sample data were homogeneous and stationary, flood frequency analysis can be performed.

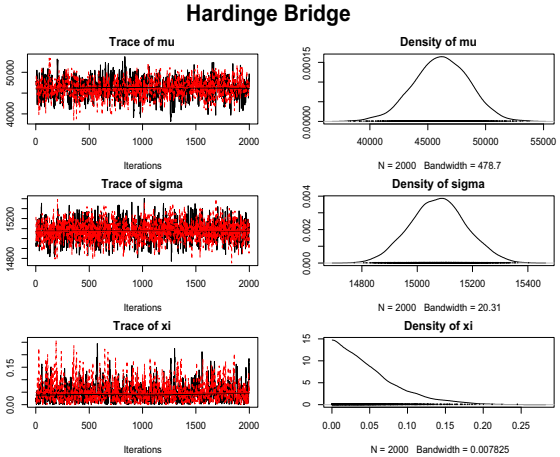
Also, the goodness-of-fit test statistic results of the commonly used frequency distributions [10] in hydrology were also used, and the results are presented in Table 3. The GEV distribution yielded the best-fit for Hardinge Bridge station. For Bahadurabad station, the GEV distribution was ranked third according to the test statistic results of the goodness-of-fit test of the K-S and A-D. But the difference of the test statistic result of the GEV with the first ranked distribution was comparatively smaller.

Table 3. Characteristics of the AM sample data of each station.

Characteristics	Hardinge Bridge Station	Bahadurabad Station
Sample size	42	42
Mean ($m^3 \cdot s^{-1}$)	52,517	66,490
Standard deviation ($m^3 \cdot s^{-1}$)	15,079	14,655
Skewness	1.22	0.52
SNH test	2.67 (Pass)	5.07 (Pass)
VN ratio test	1.94 (Pass)	1.50 (Pass)
ADF test	−5.028 (Pass)	−3.648 (Pass)
First three best-fit distributions by K-S test, with test statistic result in brackets from top to bottom.	GEV (0.067)	LPT3 (0.085)
	Gumbel (0.068)	PT3 (0.088)
	PT3 (0.071)	GEV (0.089)
First three best-fit distributions by A-D test, with test statistic result in brackets from top to bottom.	GEV (0.484)	LPT3 (0.397)
	LPT3 (0.522)	PT3 (0.403)
	Gumbel (0.568)	GEV (0.41)

4.2 Parameter estimation of the GEV by the bayesian MCMC method

Defining a prior distribution is important in Bayesian analysis. The prior distribution assumed was a trivariate normal on (μ, log(σ), ξ) with mean vector zero and diagonal and diagonal variance-covariance matrices. Fig. 2 and Fig. 3 shows the trace plots and posterior densities of the GEV parameters by using HMC and MH respectively.



(a)

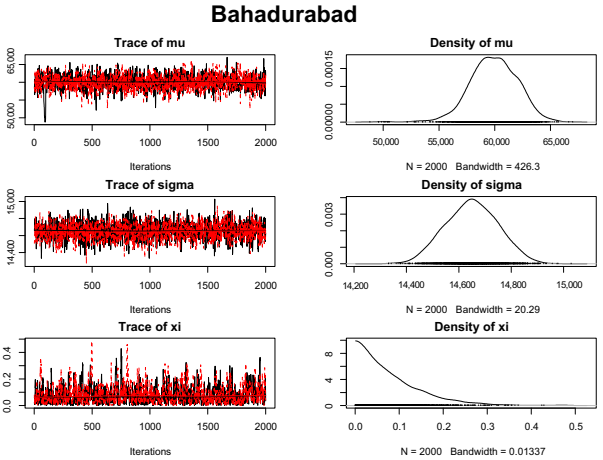


Fig. 2. Trace plots and posterior densities of the GEV parameters using HMC algorithm. The upper panel (**a**) shows the Hardinge Bridge station and the lower panel (**b**) shows the Bahadurabad station.

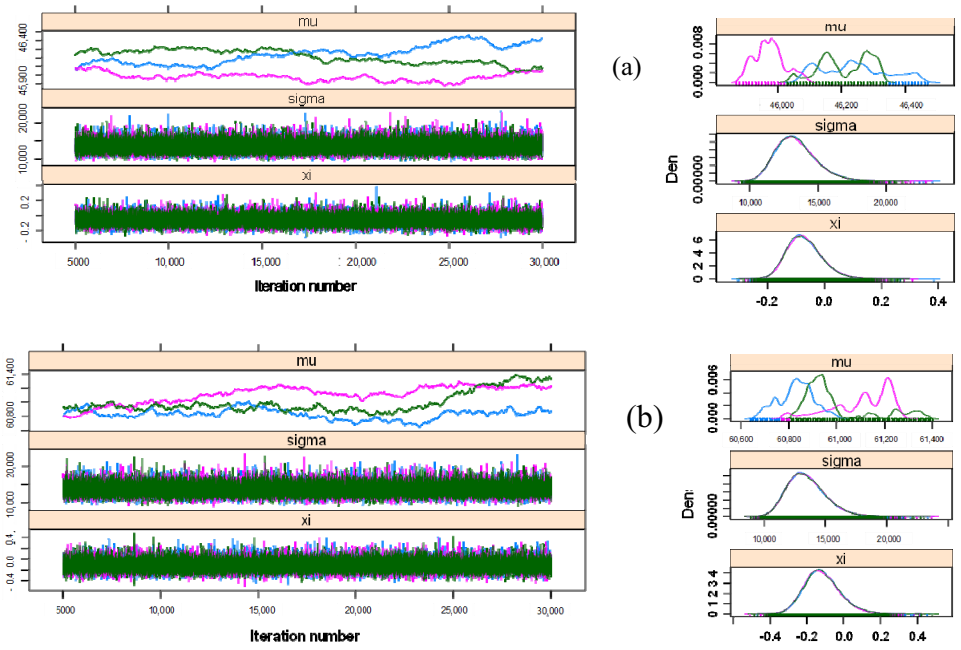


Fig. 3. Trace plots and posterior densities of the GEV parameters using MH algorithm. The upper panel (**a**) shows the Hardinge Bridge station and the lower panel (**b**) shows the Bahadurabad station.

Typically, two properties are desired in the trace plots: stationarity and good mixing. Stationarity means the path is staying within the posterior distribution. The second property, "good mixing" of a chain means each successive sample within each parameter is not highly correlated with the sample before it. Visually, a rapid zig-zag motion of each path is seen. In Figure 2, on the left side, experimental trace plots of the parameters have a stable stationarity characteristic. In the experimental trace plots, the second characteristic

also shows very well. In Fig. 3, the parameter μ in each station was not converged well. Therefore, the estimated parameters and simulated values of the HMC were considered for the further analysis in this work.

The posterior means of the parameters with 95% confidence intervals for each station are given in Table 4. Also, the effective number of samples indicated as "n_eff," and healthy chain status indicated as "Rhat" are presented in the same table.

Table 4. Posteriors means with 95% credible intervals for the GEV parameters.

Station name	Parameters	Mean (95% credible intervals; min, max)		Convergence diagnostics for HMC method	
		By HMC Method	By MH Method	n_eff	Rhat
Hardinge Bridge station	Mu	46,157 (41,507, 50,807)	45,948 (45,920, 45,980)	716	1.00
	Sigma	15,076 (14,878, 15,274)	13,353 (10,860, 16,620)	872	1.00
	xi	0.05 (−0.03, 0.13)	−0.068 (−0.177, 0.073)	740	1.00
Bahadurabad station	Mu	59,924 (55,752, 64,095)	60,971 (60,890, 61,010)	853	1.00
	Sigma	14,650 (14,453, 14,847)	13,461 (10,860, 16,940)	1449	1.00
	xi	0.08 (−0.06, 0.22)	−0.12 (−0.29, 0.1)	559	1.00

When the n_eff, is much lower than the actual number of iterations (3000) minus warmup (1000) in the chains (2 chains), it means the chains are inefficient, but possibly still valid. Warmup samples are used to adjust sampling, and so are not part of the target posterior distribution. When the Gelman-Rubin convergence diagnostic, Rhat is above 1, it indicates that the chain has not yet converged. Thus the result based on the n_eff should not be trusted. In a healthy set of chains, Rhat should approach 1.00.

4.3 Diagnostic plots and return level

The various diagnostic plots for checking the accuracy of the GEV model fitted to the stream flow data of the two main rivers are shown in Figure 4. The practical application part of the EVA is the calculation of the return period analysis, which yields risk estimations for the event. In this figure, the return level plot shows the discharge (m3•s−1) heights of the maximum 1000-year return period with 95% CI.

In Fig. 4, both the probability plot and the quantile plot show the validity of the fitted model. Each set of the plotted points is near-linear. The corresponding density plot in Fig. 4 also seems consistent with the histogram of the data. In summary, all three diagnostic plots support the fitted GEV model.

The estimated 100-year return level at Bahadurabad was 141,387 $m^3.s{-}1$ with 95% CI was [112,636, 170,138]. The 1000-year return level at the same station was 195,018 $m^3.s{-}1$, with a 95% CI range of [122,493, 267,544]. For Hardinge Bridge, the estimated 100-year return level was 124,134 $m^3.s{-}1$ with 95% CI was [108,726, 139,543]. For the 1000-year return level at Hardinge Bridge was 170,537 $m^3.s{-}1$ with a 95% CI range of [133,784, 207,289]. For expressing the uncertainty level, the estimation of CI is important in risk analysis as well as design purposes.

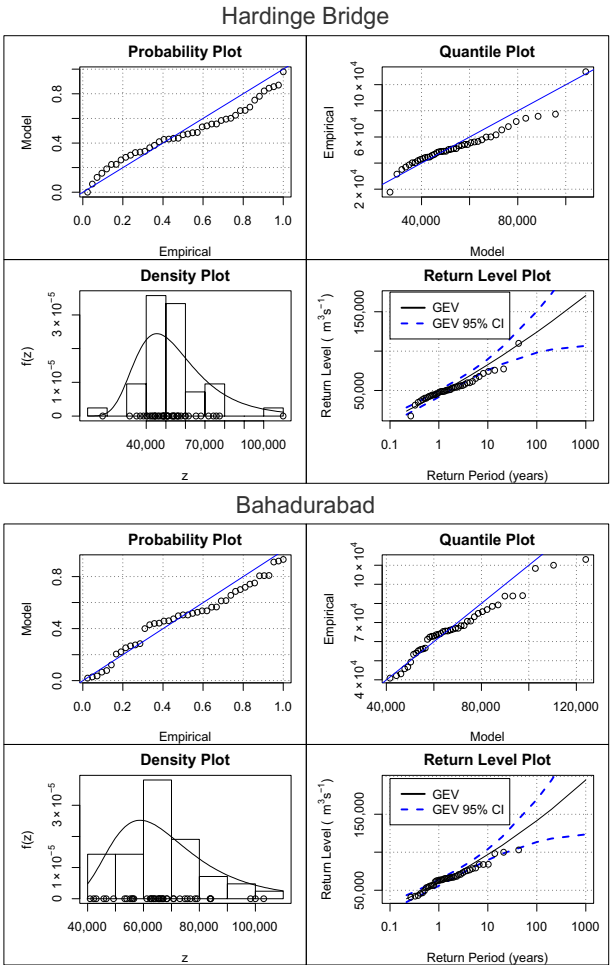


Fig. 4. Diagnostic plots for the GEV fit and return level plot. The upper panel (**a**) shows the Hardinge Bridge station and the lower panel (**b**) shows the Bahadurabad station.

5 Conclusions

This analysis has provided a demonstration of the Bayesian approach to modeling the extreme discharge of two major rivers in Bangladesh. Block maxima data, each from 42 years of data, was used in a GEV distribution model. The sample data passed as homogeneous and stationarity. For solving the Bayesian approach, a simulation-based technique (MCMC) was selected. As a more efficient, robust and much faster MCMC algorithm, the Hamiltonian Monte Carlo (HMC) algorithm was well converged than the Metropolis-Hasting (MH) algorithm. The obtaining parameters using HMC algorithm in the Bayesian analysis was used for estimating return levels of flood discharge in two major rivers. As Bangladesh is a flood-prone country, this study can help policy-makers to plan initiatives that could result in preventing damage to both lives and assets.

References

1. S. Coles, J. Bawa, L. Trenner, P. Dorazio, *An introduction to statistical modeling of extreme values* (Springer, London, 2001)
2. R.M. Neal, *Handbook of Markov Chain Monte Carlo* (Chapman & Hall/CRC, London, 2011)
3. M. Masood, P.-F. Yeh, N. Hanasaki, K. Takeuchi, Hydrol. Earth Syst. Sci. **19** (2015)
4. S. Duane, A.D. Kennedy, B.J. Pendleton, D. Roweth, Phys. Lett. B **195** (1987)
5. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. **21** (1953)
6. W.K. Hastings, Biometrika **57** (1970)
7. J. Von Neumann, Ann. Math. Stat. **12** (1941)
8. H. Alexandersson, Int. J. Climatol. **6** (1986)
9. S.E. Said, D.A. Dickey, Biometrika **71** (1984)
10. M.A. Alam, K. Emura, C. Farnham, J. Yuan, Climate **6**, 9 (2018)