Frequency Analysis and Uncertainty Assessment of Annual Maximum Flood Series Using Bayesian MCMC Method

Yunbiao Wu[1, 2] and Lianqing Xue[1, 2, 3]

[1]College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China

[2]Hohai University Wentian College, Maanshan 243031, China

[3]College of Water Conservancy and Architectural, Shehezi University, Shehezi 832003, China

Corresponding author: wyb_0018@163.com

Abstract. The calculation of flood quantiles and its uncertainty estimation are important subjects of hydraulic engineering planning and water resources management. In this study, the Bayesian theory is used to implement frequency analysis and uncertainty assessment of annual maximum flood series, the Generalized Extreme Value (GEV) distribution is considered as the flood frequency distribution line type, and the Markov chain Monte Carlo (MCMC) method based on Metropolis-Hastings algorithm is used to evaluate the GEV distribution parameters, then the posterior distributions of flood flow quantiles are used to calculate the point estimations and interval estimations of flood design values under different return periods. The results show that the fitting effect of the Bayesian MCMC method is the same as the maximum likelihood estimation (MLE), but the Bayesian MCMC more superior when the uncertainties were considered. Compared with the traditional methods of flood frequency analysis, the proposed Bayesian MCMC method provides not only the design flood estimated values, but also the confidence intervals of the estimated values. In addition, the lengths between upper confidence limits and estimated values are greater than the lower confidence limits and estimated values, this asymmetry is more realistic than the traditional methods such as the delta method, thus improve the reliability of flood frequency analysis.

1. Introduction

Over the past few decades, floods have been seen as one of the most commonly and largely distributed natural disasters in the world. Flood disaster has become one of the obstacles that affected the sustainable development of economy and society, so the flood frequency analysis has gained more and more attention in hydrology [1-5].

In recent years, the Bayesian theory has been gradually introduced into hydrological frequency analysis. One of the most attractive advantages of the Bayesian approach is that it couples prior information with sample information to provide a theoretically consistent framework for integrating systematic flow records with regional and hydrologic information within a unit framework [4]. Another reason for using Bayesian approach in flood frequency analysis is the superiority in assessing the uncertainty of quantile estimations [5-6].

In this article, the GEV distribution is considered as the flood frequency distribution line type, and the Bayesian Markov chain Monte Carlo (MCMC) method based on Metropolis-Hastings algorithm

was used to evaluate the GEV distribution parameters, then the posterior distributions of GEV distribution parameters was used to calculate design flood values, the point estimations and interval estimations of flood design values are deduced in the end to quantitative assessment on uncertainties. As an example, the annual maximum flood series of four hydrological stations in Dongting Lake basin, China, were analyzed to validate the proposed approach.

2. Materials and methods

2.1. Study area and Data sources

Dongting Lake basin is located in the middle and lower reaches of the Yangtze river region, with a total area of approximately 260000 km^2. Runoff in this basin high inter-annual variability, leading to floods and droughts occur frequently. The economic loss of agricultural directly caused by floods is 1.858×10^9 yuan [7]. The annual maximum flood peak flows of 4 hydrological stations in this basin are used. The information of the 4 hydrological stations is showed in Table 1. Data were obtained from the Hydrology and Water Resources Survey Bureau of Hunan Province.

Table 1. Information on the hydrological stations considered in this study

River	Station	Time interval	Record length (year)	Mean ($\boldsymbol{m^3/s}$)
Lishui	Shimen	1951-2014	64	7074.34
Yuanjiang	Taoyuan	1953-2014	62	16007.26
Zishui	Taojiang	1951-2014	64	5530
Xiangjiang	Xiangtan	1951-2012	62	12775.32

2.2. Methods

2.2.1. Bayes' theorem. Assume data $x = (x_1, \cdots, x_n)$ to be realizations of a random variable, whose density falls within a parametric family $\mathcal{F} = \{f(x;\theta): \theta \in \Theta\}$, where the parameter θ is an unknown constant. The Bayes' theorem is as follows:

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta} \tag{1}$$

where $\pi(\theta)$ and $f(\theta|x)$ are the prior and posterior probability density of parameter θ separately. $f(x|\theta)$ is the likelihood function of samples x. Θ is the parameter space of parameter θ. Any statistical inference about the parameter θ only based on this posterior distribution, and the asymptotic normality of the maximum likelihood estimate is no need for getting the asymptotic distribution of the parameter estimator $\widehat{\theta}$. The mean of the posterior distribution is used as point estimates of θ, i.e. $\hat{\theta} = E(\theta|x) = \int_\Theta \theta f(\theta|x)d\theta$, and the certain probability interval (confidence interval) of the posterior distribution as interval estimate of θ.

2.2.2. Flood design values calculation. If z denotes future flood design values, then the predictive density of z can be expressed as:

$$f(z|x) = \int_\Theta f(z|\theta)f(\theta|x)d\theta \tag{2}$$

By solving the following equation:

$$Pr(Z \le z|x) = 1 - \frac{1}{m} \tag{3}$$

It can give an analog of the m-year return level (i.e. $1 - \frac{1}{m}$ quantile) that incorporates uncertainty due to model estimation.

2.2.3. MCMC method. To avoid calculating the integral in the posterior distribution, the MCMC

method can be used to generate samples from the posterior distribution. Various algorithms have been suggested to apply the MCMC method according to types of chains, among which the Metropolis-Hastings algorithm have been most widely used [5-6].

The efficiency of the MCMC algorithm can be checked by acceptance rate. The acceptance rate r is defined as $r = {}^{n_a}/_{n}$, where n_a is the number of times that the proposal value θ^* is accepted, and n is the total number of iterations. If r is between 0.2 and 0.5, then the Markov chain is regarded as convergent [6].

2.2.4. Generalized extreme value distribution. The GEV distribution incorporates the Gumbel's type I ($\xi = 0$), Fréchet's type II ($\xi < 0$), and Weibull's type III ($\xi > 0$) extreme value distributions. The cumulative distribution function of GEV distribution as follows [8]:

$$F(x) = \begin{cases} exp\left\{-\left[1+\xi\frac{(x-\mu)}{\sigma}\right]^{-1/\xi}\right\}, & \xi \neq 0 \\ exp\left\{-exp\left[-\frac{(x-\mu)}{\sigma}\right]\right\}, & \xi = 0 \end{cases} \quad (6)$$

where μ, σ, ξ are the location, scale and shape parameter of GEV distribution, respectively, and $\mu \in R, \sigma > 0, \xi \in R, 1+\xi\frac{(x-\mu)}{\sigma} > 0$.

3. Results

3.1. GEV parameter estimation

For convenience, the unit of annual maximum flood peak flow in the following sections is unified to $10^3\ m^3/s$. Because there is no information on prior knowledge about the parameters, the independent zero-mean normal prior distributions on $\mu, \Phi = \log\sigma, \xi$ with variances $\nu_\mu = \nu_\Phi = 10^4$, $\nu_\xi = 10^2$ were adopted. To make the Markov chain rapidly convergence, the MLE of μ, σ, ξ were assigned to their initial values. The proposed distribution of three parameters μ, Φ, ξ is the random walk on the respective axes, i.e. $\mu^* = \mu + \varepsilon_\mu$, $\Phi^* = \Phi + \varepsilon_\Phi$, $\xi^* = \xi + \varepsilon_\xi$, where ε_μ, ε_Φ, ε_ξ are normally distributed variables, with zero-means and variances ω_μ, ω_Φ, ω_ξ. After a little trial-and-error, the parameters ω_μ, ω_Φ, ω_ξ of the four stations were selected, and all the acceptance rates r of the proposal values $\theta^* = (\mu^*, \Phi^*, \xi^*)$ were between 0.2 and 0.5.

Figure 1 shows the MCMC sampling values produced by the 10,000 iterations of the GEV model parameters μ, σ, ξ from the posterior distribution at Shimen station, where $\sigma = e^\Phi$, and the maximum likelihood estimates $(\mu,\sigma,\xi) = (5.47, 2.56, 0.05)$ were used as the sampling initial values. Figure 1 indicates that the maximum likelihood estimates are reasonable as the initial values of the sampling, and it makes the Markov chains converge rapidly. To ensure the stationarity of the sequences, all the burn-in periods were selected as 500 though all the chains converge near the initial values. In this study, the mean of the MCMC simulated samples are used as the parameter estimates after deleting the first 500 burn-in simulations.

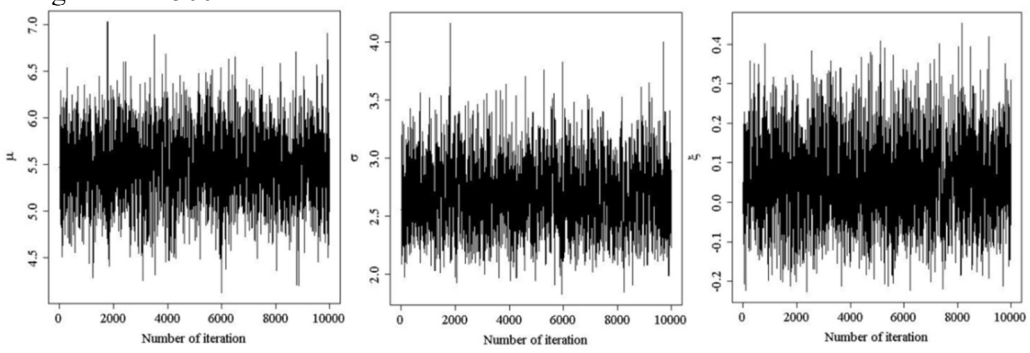

Figure 1. Bayes MCMC simulation for parameters of GEV model at Shimen station

Table 2 shows the parameter estimate results of the GEV distribution parameters based on the Bayesian theory. The confidence level of the parameter confidence interval is 95%. Compared with the traditional parameter estimation methods, the Bayesian method not only gives an estimate of the parameter, but also gives the parameter confidence interval, which represents the uncertainty of parameter estimation.

Table 2. Estimation of parameters of GEV distribution based on Bayesian MCMC

Station	Parameter	Mean	2.5%	97.5%
Shimen	μ	5.478	4.737	6.218
	σ	2.671	2.148	3.285
	ξ	0.053	-0.137	0.280
Taoyuan	μ	14.043	12.689	15.354
	σ	4.843	3.939	5.954
	ξ	-0.196	-0.369	0.019
Taojiang	μ	4.455	3.967	4.977
	σ	1.855	1.510	2.278
	ξ	0.037	-0.117	0.249
Xiangtan	μ	11.326	10.276	12.406
	σ	3.755	3.054	4.650
	ξ	-0.223	-0.444	0.030

3.2. Goodness-of-fit test

Three goodness-of-fit tests were employed: (1) quantile plot; (2) root-mean-square error (RMSE); (3) Kolmogorov-Smirnov (KS) statistic [8].

Figure 2 shows the quantile plots for the GEV models based on Bayesian estimation at the 4 selected stations. From Figure 2, it is seen that the points are sufficiently close to the unit diagonals to lend support to the fitted models.

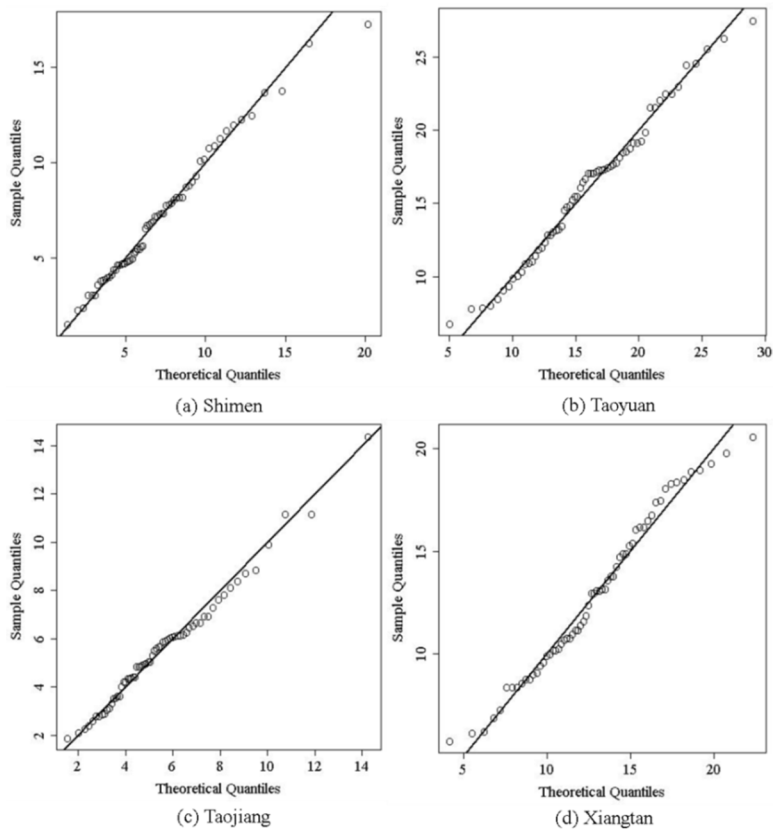

Figure 2. Quantile plots for the GEV models at the 4 selected stations

Table 3 shows the RMSE and KS test goodness-of-fit statistics at the 4 selected stations. From Table 3, it is found that all the KS statistics are between 0.07 and 0.09, and less than the critical values. All the GEV models based on Bayesian MCMC estimation and MLE have passed the KS test, which indicates that both the two estimation methods are applicable to GEV parameter estimation. By contrast the RMSE of the two parameter estimation methods at each station, it is seen that both the two methods almost have the same fitting effects.

Table 3. Results of goodness-of-fit test

Station	Statistics	Bayesian MCMC	MLE
Shimen	KS statistics	0.0755	0.0765
	RMSE	0.4677	0.4026
Taoyuan	KS statistics	0.0841	0.0871
	RMSE	0.5650	0.5516
Taojiang	KS statistics	0.0724	0.0723
	RMSE	0.2654	0.2508
Xiangtan	KS statistics	0.0797	0.0827
	RMSE	0.5284	0.5338

3.3. Return level estimation

The return level of the return period $T = {}^{1}/_{p}$ can be calculated as:

$$x_{1-p} = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-log(1-p)\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma log\{-log(1-p)\}, & \xi = 0 \end{cases} \quad (7)$$

applying this transformation to each of the vectors (μ_i, σ_i, ξ_i) leads to a sample from the corresponding posterior distribution of the $T = 1/p$ year return level, i.e. design flood peak flow.

Table 4 shows that the values of flood design estimated by Bayesian method under different typical return periods at each station. It can be seen that the values of flood design estimated by Bayesian method under different typical return periods are less than the means of the corresponding 95% confidence intervals, and the confidence intervals are not symmetric about the design values. The lengths between upper confidence limits and estimated values are greater than the lower confidence limits and estimated values. This is because the Bayesian approach estimation is based on the true posterior quantile distribution, thus capturing the skewness of their posterior distribution. The asymmetry of the confidence intervals is more realistic than the traditional methods such as the delta method. In addition, it also can be seen from Table 4 that larger return periods have the larger design values, and the width of the corresponding confidence intervals are wider, which indicates that the uncertainty increase with return period.

4. Conclusions

This study shows a Bayesian approach employed for flood frequency analysis using the GEV distribution as the line type. The method was implemented by the MCMC based on Metropolis-Hastings algorithm. The posterior distributions of the flood peak flow were used to calculate design flood values and estimate confidence intervals of design flood values. The results showed that the proposed Bayesian MCMC method provided not only the design flood estimated values, but also the confidence intervals of the estimated values. As the Bayesian approach estimation is based on the true posterior quantile distribution, the confidence intervals of the estimated values are asymmetrical: the lengths between upper confidence limits and estimated values are greater than the lower confidence limits and estimated values. The asymmetry is more realistic than the traditional methods such as the delta method. Therefore, the Bayesian approach can be applied effectively to flood frequency analysis, and its results can improve the reliability of flood frequency analysis.

Table 4. Different return level estimates of annual maximum flood peak flow

Station	Return periods (year)	2.5%	Estimated values	97.5%
Shimen	10	10.261	11.905	14.390
	25	12.315	14.942	19.731
	50	13.674	17.364	24.658
	100	14.929	19.997	30.685
Taoyuan	10	21.121	22.880	25.256
	25	23.436	25.645	29.489
	50	24.714	27.431	32.887
	100	25.718	29.018	36.375
Taojiang	10	7.758	8.830	10.331
	25	9.209	10.831	13.644
	50	10.190	12.398	16.739
	100	11.099	14.035	20.401
Xiangtan	10	16.688	17.997	19.832
	25	18.348	20.015	23.337
	50	19.211	21.304	25.998
	100	19.816	22.442	28.806

Acknowledgments
This study is supported by Program for Outstanding Young Talents in Colleges and Universities of Anhui Province (No. gxyq2018143). The authors are very grateful to the supporting sponsored by the National Scientific Foundation of China (NSFC) (No.51779074, No.41371052). Ministry of Water Resources' special funds for scientific research on public cause (201501059) and State's Key Project of Research and Development Plan (2017YFC0404304) and Jiangsu water conservancy science and technology project (2017027).

References
[1] Kuczera G 1999 Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference *Water Resources Research* **35(5)** pp 1551-1557.
[2] Reis D S and Stedinger J R 2005 Bayesian MCMC flood frequency analysis with historical information *Journal of Hydrology* **313(1-2)** pp 97-116.
[3] Lu F and Yan D 2013 Bayesian MCMC flood frequency analysis based on generalized extreme value distribution and Metropolis-Hastings algorithm *Journal of Hydraulic Engineering* **44(8)** pp 942-949.
[4] Liang Z, Chang W and Li B 2011 Bayesian flood frequency analysis in the light of model and parameter uncertainties *Stochastic Environmental Research and Risk Assessment* **26(5)** pp 721-730.
[5] Reis D S and Stedinger J R 2005 Bayesian MCMC flood frequency analysis with historical information *Journal of Hydrology* **313(1-2)** pp 97-116.
[6] Liu Y, Lu M, Huo X, Hao Y, Gao H, Liu Y, Fan Y, Cui Y and Metivier F 2016 A Bayesian analysis of Generalized Pareto Distribution of runoff minima *Hydrological Processes* **30(3)** pp 424-432.
[7] Li J, Zheng Y, Gao C and Yang Y 2000 A discussion on geographical regularity of flood and drought in Hunan Province *Journal of Natural Disasters* **9(4)** pp 115-120.
[8] Coles S 2001 *An Introduction to Statistical Modeling of Extreme Values* (London: Springer).