Probable Maximum Flood (PMF) Estimation Using a Bayesian Model (Case Study: Latian Dam)

Ahmad Agheshlouei1, Mojtaba Shourian

1. Faculty of Civil, Water and Environmental Engineering, Shahid Beheshti University

Email: m_shourian@sbu.ac.ir

Abstract

Probable Maximum Flood (PMF) is one of the important factors in dam design. In addition, it is the base for sizing the spillway of a dam to prevent overtopping during a flood. PMF is the response of a catchment to the Probable Maximum Precipitation (PMP). Among the different methods, the data-driven models try to estimate the unobserved events only based on a series of observed flows. Bayesian models are used effectively to predict flood discharge. They help hydrologists and other responsible people to make appropriate decisions to save lives and properties in downstream of watersheds. Although, the frequency analysis is popular yet, Bayesian model uses knowledge of an expert together with observation data to produce a more precise estimation. In this study, the hydrometric station located near Latian dam is selected. Its historical report is available for more than 70 years. Two scenarios namely, Annual maximum flows and Peak floods are modeled using WinBUGS software, which estimates the posterior distribution function, using MCMC method, Metropolis-Hastings algorithm with Gibbs sampling. An upper bounded LN4 is chosen as prior function.

Keywords: Probable Maximum Flood (PMF), Bayesian Analysis, Latian Dam.

1. INTRODUCTION

During recent decade, due to climate change, the frequency and severity of floods have been increasing dramatically. It is essential for a water resource planner to know the magnitude of the maximum flood. Knowledge of PMF is required for dams design, because it is a critical input parameter for designing the reservoir of a Dam, and the capacity of the Spillway. Furthermore, it is used for design of a flood control system around cities or industrial facilities (e.g., Oil and Gas plants), since industrial plants are usually located at the foot of a mountain.

PMF maybe the most used criterion in the United States and other countries to design major hydraulic flood defense structures, such as large dam spillways (ICOLD 1987; FEMA 2004; USNRC 1977). [1,2,3]

WMO (1986) defined PMF as the response of a catchment to PMP, probable maximum precipitation that is meteorologically possible to occur over that catchment, during a given season of the year [4]. Usually, an appropriate rainfall-runoff simulation model performs the PMP-PMF transformation [5]. Dawdy and Lettenmaier (1987) stated, since the PMF is astatistic, as other statistic must be estimated, regarded the PMF as “quasi-deterministic”. They declared that its exceedance probabilities is small but in fact, it is not zero [6]. Although engineers accepted this “quasi-deterministic” PMF-like floods as a safety standard for high-hazard dams, the probability of extreme floods should be estimated to incorporate it into quantitative risk assessment studies. This study is needed for both new and existing structures (Stedinger et al. 1989; Dawdy and Lettenmaier 1987; Dubler and Grigg 1996; USBR 2002).[7,6,8,9]

Nathan and Weinmann (2001) have categorized floods [10], based on their return periods, Figure 1. Floods with return period of more than 2000 years (Pr ≤ 2 × 10⁻³) are considered outside the extrapolation range because of very large amount of uncertainty leve. However, PMF or some percent of PMF is used for flood control system, as well as for the detention and retention dam so check the Dam failure for both new and existing ones, and also the spillway’s sizing. Therefore, instead of frequency analysis, Bayesian model is used as an alternative solution.

Fernandez (2010) studied American river basin [5], near the Folsom Reservoir, in California, USA. He combined conventional analysis with the logical context in a Bayesian model to reach a more reliable estimation of PMF. He used two types of non-systematic data: (a) the historical floods so-called paleofloods which has been recorded by human, (b) the remaining physical, geological, and/or botanical evidences that show the risen water level in some pre-historical time.

Mauro Naghettini [11] explained a comprehensive discussion of this method. Anyhow, in this case the non-systematic data are not available in this study.
Bayesian models are established based on three parts: first priories including the background knowledge of researcher, second, likelihood function which is obtained using observed floods and third, marginal function. A LN4 upper bond function is used as a prior function. The estimation of its parameters is explained later.

**Figure 1. Flood categories (adapted from Nathan and Weinmann 2001)**

2. **METHODOLOGY**

Some important assumptions had to be made; (a) despite the fact that flood is non-stationary, the series assumed as stationary homogeneity; (b) non-systematic data is not available for this study; (c) the upper bond LN4 is the best fitted one, considering the available data; (d) the short data (only 70 years) is used to model an event with T>1000 years. These assumptions caused an uncertainty to the results. Some of them can be resolved to some extent in further studies but some is inevitable because of the nature of random process.

Flood estimation is not deterministic and cannot produce any firm result. Only the statistical analysis, either frequency analysis or Bayesian, can be used to estimate flood confidential or credential interval. The result is not unique and it is strongly dependent on availability and accuracy of data. Any bias in input data will result in more uncertainty in output estimation.

In a brief and short form, the Bayesian Model can be explained as Equation 1.

\[
\pi(\theta|x,\Omega) = \frac{f(x|\theta,\Omega)\pi(\theta|\Omega)}{f(x|\Omega)} \text{Posterior PDF} = \frac{\text{Likelyhood Function} \times \text{Prior PDF}}{\text{Marginal Probability}}
\]  

(1)

Where, \( \theta \), \( x \) and \( \Omega \) are the parameter, observation, and prior data about the parameter \( \theta \) respectively. Further information can be found in chapter 11 of Fundamentals of Statistical Hydrology book, by Ma Mauro Naghettini[11]. In fact, this equation for hydrological event and in particular for flood cannot be solved analytically in most cases. So, numerical estimation is needed. One of the common solutions is Markov chain Monte Carlo (MCMC). To solving the MCMC of Bayesian model, using Metropolis-Hastings algorithm within Gibbs sampling from probability function is recommended as an effective and reliable method [12].

The input data then are used to develop a Bayesian model in WinBUGS (Bayesian inference Using Gibbs Sampling). It is developed in the MRC Biostatistics Unit, Cambridge University asan open source software. WinBUGS is used for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods [13].

The upper bound Log Normal Distribution with four parameters (LN4) is used for prior distribution. The transformation of Equation2 is used to normalize X to Y. Where X, and NEX are the possible PMF in the catchment area and the number of “Annual flood-peaks with exact values” respectively.

Transformation: \( Y = \ln \left( \frac{X-\epsilon}{\alpha-\epsilon} \right), Y \sim N(\mu_Y, \sigma_Y) \)  

(2)

The Probable Distribution Function (PDF) and Cumulative Distribution Function (CDF) of LN4 are formulated in Equations 3 and 4 respectively:

PFD: \( f_x(x|\theta) = \frac{1}{x(\alpha-x)\sigma_Y \sqrt{2\pi}} e^{-\frac{1}{2\sigma_Y^2} \left[ \ln \left( \frac{x-\epsilon}{\alpha-\epsilon} \right) - \mu_Y \right]^2 }, 0 \leq x \leq \alpha \)  

(3)
CDF: \( F_x(\theta) = \Phi\left(\frac{1}{\sigma_y} \ln\left(\frac{x - \mu_\theta}{\sigma_\theta}\right) - \frac{\mu_\theta}{\sigma_\theta}\right) \) \hspace{1cm} (4)

Where, \( \Phi \) is the notion for CDF of Normal Distribution. In addition, \( \alpha \in \mathcal{R}_+ \) is the upper bond of LN4 and analogously \( \epsilon \in \mathcal{R}_+ \) is the lower bond of LN4.

There are some hyper-parameters: \( \rho_{\alpha}, \beta_{\alpha}, \rho_{\sigma}, \beta_{\sigma}, \mu_\theta, \sigma_\theta \). Takara and Tosa (1999) reported that there would be some effects of a zero lower bound \( \epsilon=0 \) on the LN4 goodness-of-fit and upper tail behavior[14].

\[ P(\alpha \leq PMF_{local} | \rho_{\alpha}, \beta_{\sigma}) = 0.5 \] \hspace{1cm} (5)

To assess the accuracy of the posterior estimations, two criteria is used. The first criterion is the Coefficient of Variation (CV) which can be calculated by dividing variance by mean. The second one is MC_error, which means that the Monte Carlo error for each parameter that is an estimation of the difference between the mean of the sampled values (posterior mean) and the true posterior mean using the metropolis algorithm. The poorer the convergence, the higher the Monte Carlo error [15].

3. **CASE STUDY**

Latian Dam is located 25 km away from Tehran, the capital of Iran, constructed in 1968. It plays a very important role for providing potable water for the capital city. The main water resource for the dam is Jajrood River. This river springs from Alborz Mountain and flows to the Namak Lake in the center of Iran. In this catchment, there are 16 meteorological stations. Latian Dam has two hydropower plants; the capacity of each one is about 22.5 MW. Watershed code in “wrs.wrm.ir” [16] system is 41-36 and its area is 987 Km². Annual input rainfall is about 334 MM³, and output is about 307 MM³, from which 171 MM³ of it is consumed for potable water of Tehran. The reservoir capacity is 95 MM³, and the height of the dam crest from foundation is 107 m[17].

The hydrometric station (Code 41-119) near the dam, is located at E=51°41′00″, N=35°47′00″, Z=1560 m. The measured discharge means about 2.87 m³/s, with the mean of annual maximum flows of 2.71. This is shown in Table 1, which clearly indicates a bias in the input data.

![Figure 2. Site Location of Latian Dam](image-url)
The data used in this study belongs to Latian Dam; located on Jajrood river basin. Due to 70 years long record, it forms a suitable likelihood function. “Iran Water Resource Management Co” has provided this record [16], from 1946 to 2016, at Latian Hydrometric station coded 41-119, as shown in Figure 3.

![Figure 3. The record of discharge at hydrometric Station](image3.png)

The classical definition of flood is that when the discharge fills the whole stream and starts to flow on the floodplain. However, in this study, for practical engineering reason, the return period of 25 years is selected as threshold. First, all discharge history record is given to a Matlab code to sort the time history based on the magnitude. The frequency analysis is done by assigning the relevant rank as Weibullrankings to each discharge. Considering 25-year return period (T) threshold, or Exceedance probabilities Pr=0.04, resulted in flood with 30 m$^3$/s flow rate as shown in Figure 3.

During 70 years, 1006 days have been recorded with flood (i.e. $Q>30$m$^3$/s), amongst those 692 days with return period larger than 25 years and less than 80 years. Meanwhile, there is only 314 observations with $T>80$ years or $Pr<0.0125$. Due to rare observations, the flood should be estimated using extreme method of statistical interferences.

![Figure 4. Frequency Analysis of the Discharge](image4.png)

![Figure 5. Flow-Duration-Curve (FDC)](image5.png)

Accordingly, the Flow-Duration-Curve (FDC) presented in Fig 5, the daily peak flow of each flood is extracted and saved as “Peak Latian Dam” as the first scenario for the next step; including 139 items, see Fig 6.

![Figure 6. Site Location of Latian Dam](image6.png)
The second scenario includes the set annual maximum discharge in hydrometric station, called “Annual Latian Dam”. It comprises of 70 items, some of them are less than the threshold as illustrated in Figure 7.

![Figure 7. Annual maximum Discharge](image)

### 4. BAYESIAN MODEL

Two selected scenarios (Peak and Annal) of Latian Dam are compared to the American river basin at the Folsom dam site done by Fernandes (2010) did [5]. The input data of Folsom Dam is reported by Ma Mauro Naghettini [11]. Refer to Table 1; it is obvious that CV and variance of Peak series are smaller than Annual’s on, but the skew of Peak is greater than Annual’s.

#### Table 1 - Input Data (X) Comparison

<table>
<thead>
<tr>
<th>Dam</th>
<th>No of Observation</th>
<th>Max Q</th>
<th>Mean Q</th>
<th>Min Q</th>
<th>Variance</th>
<th>CV (=) Var./Mean</th>
<th>Skew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folsom [5]</td>
<td>52</td>
<td>8,438</td>
<td>2,259,423</td>
<td>280</td>
<td>4,386,216</td>
<td>1,941</td>
<td>1.431</td>
</tr>
<tr>
<td>Latian (Annual)</td>
<td>70</td>
<td>187</td>
<td>53.887</td>
<td>2.71</td>
<td>1,433</td>
<td>26</td>
<td>1.078</td>
</tr>
<tr>
<td>Latian (Peak)</td>
<td>139</td>
<td>187</td>
<td>51.241</td>
<td>30.2</td>
<td>733</td>
<td>14</td>
<td>2.059</td>
</tr>
</tbody>
</table>

The transformation of Equation 2 normalize X input data to Y, Table 2. This transformation helps to reduce the CV and variance of data.

#### Table 2 - Transformed Data (Y) Comparison

<table>
<thead>
<tr>
<th>Dam</th>
<th>(\alpha) (PMF)</th>
<th>No</th>
<th>Max</th>
<th>Mean</th>
<th>Min</th>
<th>Variance</th>
<th>CV (=) Var./Mean</th>
<th>Skew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folsom [3]</td>
<td>25,655</td>
<td>52</td>
<td>-0.713</td>
<td>-2.725</td>
<td>-4.507</td>
<td>0.984</td>
<td>-0.361</td>
<td>0.263</td>
</tr>
<tr>
<td>Latian (Annual)</td>
<td>3,000</td>
<td>70</td>
<td>-2.711</td>
<td>-4.288</td>
<td>-7.009</td>
<td>0.737</td>
<td>-0.172</td>
<td>-0.833</td>
</tr>
<tr>
<td>Latian (Peak)</td>
<td>3,000</td>
<td>139</td>
<td>-2.711</td>
<td>-4.153</td>
<td>-4.588</td>
<td>0.183</td>
<td>-0.044</td>
<td>1.134</td>
</tr>
</tbody>
</table>

The available knowledge and prior data about the scenarios are summarized in the Table 3. It is entered in Bayesian model.

#### Table 3- LN4 model as prior distribution for Bayesian Analysis

<table>
<thead>
<tr>
<th>parameter</th>
<th>Distribution of parameter</th>
<th>Description</th>
<th>Selected priors</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha)</td>
<td>(\alpha\sim\text{Gamma}(\rho_\alpha, \beta_\alpha))</td>
<td>Upper bound of the LN4 distribution</td>
<td>Table 4</td>
<td>Marginal Distribution</td>
</tr>
<tr>
<td>(\varepsilon)</td>
<td>-</td>
<td>Lower bound of the LN4 distribution</td>
<td>Zero</td>
<td>[6]</td>
</tr>
<tr>
<td>(\sigma_Y)</td>
<td>(\sigma_Y\sim\text{Gamma}(\rho_{\sigma_Y}, \beta_{\sigma_Y}))</td>
<td>Scale parameter of the LN4 distribution</td>
<td>(\sigma_Y\sim\text{Gamma}(1.0, \frac{1}{10^7}))</td>
<td>non-informative prior distributions</td>
</tr>
<tr>
<td>(\mu_Y)</td>
<td>(\mu_Y\sim\text{Normal}(\frac{1}{\sigma^2}, \frac{1}{\sigma^2}))</td>
<td>Location parameter of the LN4 distribution</td>
<td>(\mu_Y\sim\text{Normal}(1.0, \frac{1}{10^7}))</td>
<td>non-informative prior distributions</td>
</tr>
</tbody>
</table>
The method of moments yields \( \hat{\mu} = CV^{-2} \). In this study, \( CV \) is discrete with values of 0.1, 0.2, 0.3, 0.4 and 0.5; \( \rho_a \) and \( \beta_a \) are calculated accordingly. The existing spillways for the Dam are sized to the capacity of 1930 m³/s [18]. So, the initial value for PMF is considered to be 3000 m³/s (about 50% more than the existing PMF). Table 4 demonstrates results for discrete value of \( CV \) using Equation 5.

<table>
<thead>
<tr>
<th>CV_a</th>
<th>Pr.</th>
<th>( \rho_a )</th>
<th>( \beta_a )</th>
<th>( 1/\beta_a )</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.50</td>
<td>400</td>
<td>7.506</td>
<td>1.332E-01</td>
<td>3.003</td>
<td>3.000</td>
</tr>
<tr>
<td>0.10</td>
<td>0.50</td>
<td>100.00</td>
<td>30.100</td>
<td>3.322E-02</td>
<td>3.010</td>
<td>3.000</td>
</tr>
<tr>
<td>0.15</td>
<td>0.50</td>
<td>44.444</td>
<td>68.009</td>
<td>1.47039E-02</td>
<td>3.023</td>
<td>3.000</td>
</tr>
<tr>
<td>0.20</td>
<td>0.50</td>
<td>25.00</td>
<td>121.618</td>
<td>8.222E-03</td>
<td>3.040</td>
<td>3.000</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>16</td>
<td>191.474</td>
<td>5.22264E-03</td>
<td>3.023</td>
<td>3.000</td>
</tr>
<tr>
<td>0.30</td>
<td>0.50</td>
<td>11.11</td>
<td>278.303</td>
<td>3.593E-03</td>
<td>3.092</td>
<td>3.000</td>
</tr>
<tr>
<td>0.50</td>
<td>0.50</td>
<td>4.00</td>
<td>816.980</td>
<td>1.224E-03</td>
<td>3.268</td>
<td>3.000</td>
</tr>
<tr>
<td>0.70</td>
<td>0.50</td>
<td>2.04</td>
<td>1,745.307</td>
<td>5.730E-04</td>
<td>3.562</td>
<td>2.998</td>
</tr>
<tr>
<td>0.90</td>
<td>0.50</td>
<td>1.23</td>
<td>3,254.522</td>
<td>3.073E-04</td>
<td>4.018</td>
<td>2.989</td>
</tr>
</tbody>
</table>

Where, for the gamma distribution, mean is \( \rho_a \times \beta_a \) and median is about Mean \( \times (3 \rho_a - 0.8) / (3 \rho_a + 0.2) \).

5. RESULTS

Two scenarios (Annual and Peak) have been studied. Based on the results, it is proven that Latian Peak scenario is not sensitive to the changes of \( \sigma_Y \), \( \mu_Y \); the set of \((10^{-8}, 10^{-6})\) or \((10^{-7}, 10^{-6})\) or \((10^{-7}, 10^{-5})\). However, in the second scenario, Latian Annual shows a little improvement in this order: of \((10^{-8}, 10^{-6})\) or \((10^{-7}, 10^{-6})\) or \((10^{-7}, 10^{-5})\). Therefore, the distribution of \( \sigma_Y \), \( \mu_Y \) is selected as \((10^{-7}, 10^{-5})\) to run the final models.

Both scenarios are shown compared in Figure 8 (a,b). The MC_Error for both scenarios are similar and the lower prior CV gives the lower Error. Nonetheless, the Peak Flow scenario is more accurate than the Annual. Analogously, in both scenarios, smaller prior CV results in smaller posterior CV.

Running the models leads to the posterior joint distribution of the LN4 parameters: \( \alpha, \sigma_Y \) and \( \mu_Y \). Lunn et al. (2000) stated that WinBUGS or OpenBUGS makes practical MCMC methods and Metropolis–Hastings algorithm generate the samples from the posterior joint distribution of the LN4 parameters [19]. In this study, WinBUGS sampled 1,000,000 times from the posterior joint distribution of parameters, lagged by 10 samples to avoid correlation. First 10,000 samples as ‘burn in’ were discarded and the 99,000 samples left were used to build the posterior data set.

The shape of posterior probability distribution function varies according to different prior CVs. As shown in Figure 9, the shape of Annual scenario changes from sharp Gamma type to Normal, and to shifted bell shape.
Similarly, the shape of Peak Flow scenario is gradually changing from sharp Gamma to bell type, but it has a smoother distribution as shown in Figure 10.

According to Figure 11, having a suitable estimation of $\alpha$ and a lower amount of CV, the results from two scenarios are very close. However, for higher CV values, peak flow scenario ends up with a larger amount of flow. Conversely, the annual maximum scenario tends to lower amount of flow.

It can be concluded that the annual maximum series underestimates the magnitude of PMF, yet the peak flow overestimates the magnitude of PMF, Table 5.

Furthermore, without non-systematic data, the estimation of Bayesian Model will lead to the same amount of PMF in both scenarios. In the future, with non-systematic data, Bayesian Model can be run and the outcomes can be compared. According to research conducted by Fernandes [5], this can improve the results.

### Table 5- The result of estimation $\alpha$, prior CV=0.1

<table>
<thead>
<tr>
<th></th>
<th>scenario</th>
<th>mean</th>
<th>sd</th>
<th>MC_error</th>
<th>val5.0pc</th>
<th>median</th>
<th>val95.0pc</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Peak</td>
<td>3025</td>
<td>299.3</td>
<td>0.9041</td>
<td>2550</td>
<td>3015</td>
<td>3535</td>
<td>10000</td>
<td>99000</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>2999</td>
<td>301.7</td>
<td>0.9569</td>
<td>2519</td>
<td>2989</td>
<td>3513</td>
<td>10000</td>
<td>99000</td>
</tr>
<tr>
<td>$\mu_Y$</td>
<td>Peak</td>
<td>-4.156</td>
<td>0.1073</td>
<td>0.0003234</td>
<td>-4.33</td>
<td>-4.157</td>
<td>-3.977</td>
<td>10000</td>
<td>99000</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>-4.282</td>
<td>0.147</td>
<td>0.0004661</td>
<td>-4.522</td>
<td>-4.283</td>
<td>-4.039</td>
<td>10000</td>
<td>99000</td>
</tr>
<tr>
<td>$\sigma_Y$</td>
<td>Peak</td>
<td>0.4312</td>
<td>0.0263</td>
<td>0.00009196</td>
<td>0.3905</td>
<td>0.4298</td>
<td>0.4765</td>
<td>10000</td>
<td>99000</td>
</tr>
<tr>
<td></td>
<td>Annual</td>
<td>0.8747</td>
<td>0.07661</td>
<td>0.0002189</td>
<td>0.7584</td>
<td>0.8693</td>
<td>1.009</td>
<td>10000</td>
<td>99000</td>
</tr>
</tbody>
</table>
CONCLUSIONS

This study shows that the estimated PMF by annual maximum flow underestimates the PMF, although the Peak flow overestimates the PMF. Bayesian Model can reduce the difference between the results of the two scenarios by deploying an effective Likelihood function. The upper bond parameter of LN4 distribution function estimates the PMF properly with a CV of less than 0.1. OpenBUGS software generates the sampling by Gibbs method and estimates the posterior function by Metropolis–Hastings algorithm.

To improve the estimation of PMF, the nonsystematic floods should be added. The nonsystematic floods can be extracted from the responsible sources as the needed input into the Bayesian models.

REFERENCES

13. WinBUGS official site: https://www.mrc-bsu.cam.ac.uk/software/bugs/
15. The tutorial, manuals of software: OpenBUGS, http://www.openbugs.net/Manuals/Tutorial.html