

Article

Generalized Beta Distribution of the Second Kind for Flood Frequency Analysis

Lu Chen ^{1,*} and Vijay P. Singh ²

¹ College of Hydropower & Information Engineering, Huazhong University of Science & Technology, Wuhan 430074, China

² Department of Biological and Agricultural Engineering & Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843, USA; vsingh@tamu.edu

* Correspondence: chen_lu@hust.edu.cn; Tel.: +86-027-8754-3992

Academic Editor: Kevin H. Knuth

Received: 18 April 2017; Accepted: 26 May 2017; Published: 12 June 2017

Abstract: Estimation of flood magnitude for a given recurrence interval T (T -year flood) at a specific location is needed for design of hydraulic and civil infrastructure facilities. A key step in the estimation or flood frequency analysis (FFA) is the selection of a suitable distribution. More than one distribution is often found to be adequate for FFA on a given watershed and choosing the best one is often less than objective. In this study, the generalized beta distribution of the second kind (GB2) was introduced for FFA. The principle of maximum entropy (POME) method was proposed to estimate the GB2 parameters. The performance of GB2 distribution was evaluated using flood data from gauging stations on the Colorado River, USA. Frequency estimates from the GB2 distribution were also compared with those of commonly used distributions. Also, the evolution of frequency distribution along the stream from upstream to downstream was investigated. It concludes that the GB2 is appealing for FFA, since it has four parameters and includes some well-known distributions. Results of case study demonstrate that the parameters estimated by POME method are found reasonable. According to the RMSD and AIC values, the performance of the GB2 distribution is better than that of the widely used distributions in hydrology. When using different distributions for FFA, significant different design flood values are obtained. For a given return period, the design flood value of the downstream gauging stations is larger than that of the upstream gauging station. In addition, there is an evolution of distribution. Along the Yampa River, the distribution for FFA changes from the four-parameter GB2 distribution to the three-parameter Burr XII distribution.

Keywords: entropy theory; principle of maximum entropy (POME); GB2 distribution; flood frequency analysis

1. Introduction

Estimation of flood magnitude for a given recurrence interval T (T -year flood) at a given location is essential for the design of hydraulic and civil infrastructure facilities, such as dams, spillways, levees, urban drainage, culverts, road embankments, and parking lots. A key step in flood frequency estimation or analysis (FFA) is the selection of a suitable frequency distribution [1]. Commonly used distributions for flood frequency analysis include Gumbel, gamma, generalized extreme value (GEV), Pearson type III (P-III), log-Pearson type III (LP-III), Weibull, and log-normal (LN). Some of these distributions have been adopted in different countries. For example, the P-III distribution has been adopted in China and Australia as a standard method for hydrologic frequency analysis [2–4]. The LP-III distribution has been adopted in the United States and the GEV distribution in Europe.

Mielke and Johnson investigated the use of two special cases of the generalized beta distribution of the second kind, namely gamma and log normal distributions, for flood frequency analysis [5].

Wilks investigated the performance of eight three-parameter probability distributions for precipitation extremes using annual and partial duration data from stations in the northeastern and southeastern United States [6]. He found that the beta- κ distribution best described the extreme right tail of annual extreme series, and the beta- P distribution was best for the partial duration data.

Recently, some generalized frequency distributions have been used for hydrologic frequency analysis. For example, Perreault et al. presented a family of distributions, named Halphen distributions, for frequency analysis of hydrometeorological extremes [7]. Papalexiou and Koutsoyiannis used the generalized gamma distribution and generalized beta distribution of the second kind (GB2) for rainfall frequency analysis across the world and showed that these distributions were appropriate for worldwide rainfall data [8]. The greatest advantage of these generalized distributions is that they provide sufficient flexibility to fit a large variety of data sets, which facilitates the selection and comparison of different distributions. For instance, the GB2 distribution includes the exponential, Weibull, and gamma distributions as special cases. Since the GB2 distribution has four parameters, logically it should perform better than 3-parameter distributions, such as GEV, P-III, LP-III or LN-III. Papalexiou and Koutsoyiannis concluded that the GB2 distribution was a suitable model for rainfall frequency analysis because of its ability to describe both J-shaped and bell-shaped data [8]. The other advantages of the GB2 distribution can be summarized as: (1) the GB2 distribution can model positive or negative skewness which is an advantage over distributions, such as lognormal, with only positive skew; (2) it can jointly estimate both location and shape parameters, while many other distributions, such as exponential, logistic, normal, etc., usually focus on location only; and (3) it can better capture the long right or left tail. Because of these advantages, the GB2 distribution was employed in this study.

The second step in flood frequency analysis is to estimate parameters of the selected distribution. There are several standard parameter estimation methods, such as moments, maximum likelihood, L-moments, probability weighted moments, and least square. Among these methods, the maximum likelihood (ML) and L-moment methods are widely used in hydrology. In addition, the principle of maximum entropy (POME) has been applied to parameter estimation [9,10]. Singh and Guo indicated that POME method was comparable to ML and L-moment methods, and for certain situations, POME method was superior to these two methods [11]. Therefore, the POME method was considered in this study for parameter estimation.

Another aspect of FFA that is of interest is how the flood frequency distribution evolves from upstream to downstream along a river. The drainage area along the river increases from upstream to downstream. It is interesting to investigate if the same frequency distribution applies at all gauging stations along the stream.

The objective of this study therefore is to employ the GB2 distribution for flood frequency analysis (FFA). The specific objectives are to: (1) estimate the GB2 distribution parameters using the principle of maximum entropy; (2) evaluate the performance of the GB2 distribution and compare it with commonly used distributions in hydrology; (3) select the best distribution; and (4) discuss the evolution of frequency distribution and its parameters along the river.

2. GB2 Distribution

The generalized beta distribution of the second kind, denoted as GB2, is a four-parameter distribution and can be expressed as:

$$f(x) = \frac{r_3}{\beta B(r_1, r_2)} \left(\frac{x}{\beta}\right)^{r_1 r_3 - 1} \left(1 + \left(\frac{x}{\beta}\right)^{r_3}\right)^{-(r_2 + r_1)} \quad (1)$$

where $B(\cdot)$ is the beta function; β is the scale parameter, $\beta > 0$; and $r_1 > 0$, $r_2 > 0$, and $r_3 > 0$ are the shape parameters. Parameter r_3 represents the overall shape; parameter r_1 governs the left tail; parameter r_2 controls the right tail; and β is a scale parameter and depends on the unit of measurement. These parameters allow the distribution to be able to fit data having very different histogram shapes. It can simulate both the J-shaped and bell-shaped distributions. Parameters r_1 and

r_2 together determine the skewness of the distribution. The general shapes of GB2 probability density distribution were shown in Figure 1.

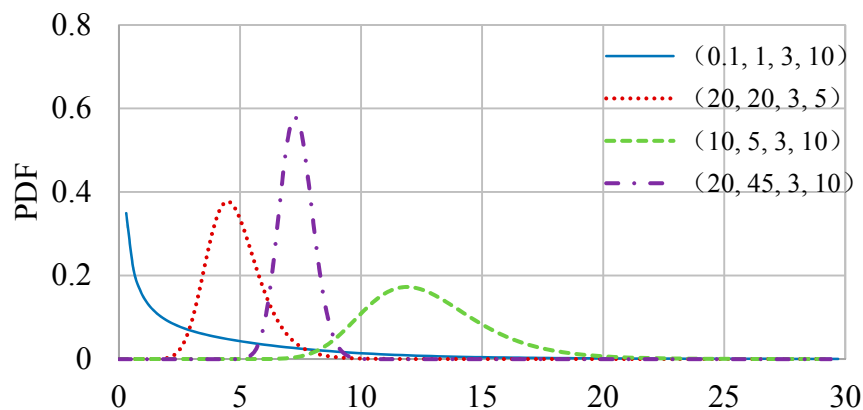


Figure 1. Shapes of PDF of GB2 distribution.

When analyzing extreme rainfall, Papalexiou and Koutsoyiannis showed that the GB2 distribution is a very flexible four-parameter distribution [8]. By fixing certain parameters, the GB2 distribution can yield some well-known distributions, such as the beta distribution of the second kind (B2), the Burr type XII, generalized gamma (GG), and so on. These distributions can be treated as special or limiting cases of the GB2 distribution, as shown in Figure 2. Some of these special cases have been applied in hydrological frequency analysis. For example, Shao et al. employed the Burr type XII distribution for flood frequency analysis [2].

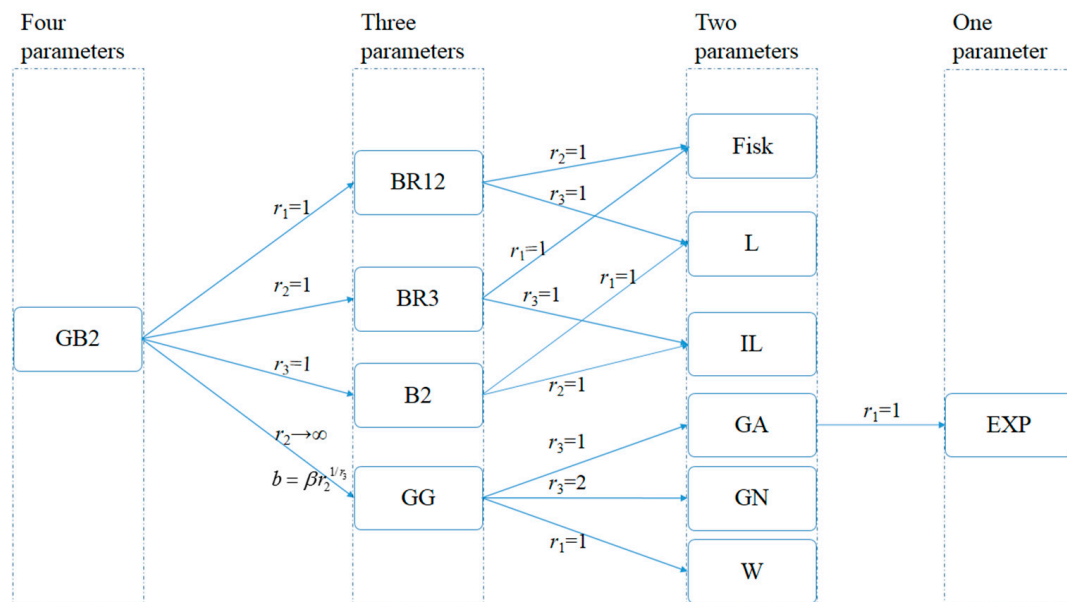


Figure 2. The GB2 distribution and its special cases (where BR12 means the Burr XII distribution; BR3 means the Burr III distribution; B2 means the beta distribution of second kind; Fisk means log-logistic distribution; L means the Lomax distribution; IL means inverse Lomax distribution; GA distribution means the gamma distribution; GN means the generalized normal distribution; W means the Weibull distribution and EXP means the exponential distribution).

3. Estimation of Parameters of GB2 Distribution by POME Method

The GB2 distribution parameters were determined using the principle of maximum entropy (POME). The POME method involves the following steps: (1) specification of constraints; (2) maximization of entropy using the method of Lagrange multipliers; (3) derivation of the relation between Lagrange multipliers and constraints; (4) derivation of the relation between Lagrange multipliers and distribution parameters; and (5) derivation of the relation between distribution parameters and constraints. These steps are discussed in Appendix A. Here only steps (1) and (5) are outlined.

Flood discharge is considered as a random variable X , which ranges from 0 to infinite. Its probability distribution function (PDF) and cumulative distribution function (CDF) are denoted as $f(x)$ and $F(x)$ respectively, where x is a specific value of X . Since constraints encode the information that can be given for the random variable, following Singh (1998), the constraints for the GB2 distribution can be expressed as:

$$\int_0^{\infty} f(x) dx = 1 \quad (2a)$$

$$\int_0^{\infty} f(x) \ln x dx = E(\ln x) \quad (2b)$$

$$\int_0^{\infty} f(x) \ln\left(1 + \left(\frac{x}{\beta}\right)^{r_3}\right) dx = E\left(\ln\left(1 + \left(\frac{x}{\beta}\right)^{r_3}\right)\right) \quad (2c)$$

The first constraint is the total probability law, the second constraint is the mean of log values or the geometric mean, and the third constraint is the mean of log of scaled values raised to a power and then shifted by unity.

Following the derivation in Appendix A, the relation between parameters and constraints can be expressed as:

$$\begin{aligned} -\ln \beta - \frac{1}{r_3} \varphi(r_1) + \frac{1}{r_3} \varphi(r_2) &= -E(\ln x) \\ \beta^{r_3} \varphi(r_2) - \beta^{r_3} \varphi(r_1 + r_2) &= -E(\beta^{r_3} \ln(1 + (\frac{x}{\beta})^{r_3})) \\ -\ln \beta + \frac{1}{r_3^2} \varphi'(r_1) + \frac{1}{r_3^2} \varphi'(r_2) &= \text{var}(\ln x) \\ \varphi'(r_2) - \varphi'(r_1 + r_2) &= \text{var}(\ln(1 + (\frac{x}{\beta})^{r_3})) \end{aligned} \quad (3)$$

where $\varphi(\cdot)$ is the digamma function; and $\varphi'(\cdot)$ is the trigamma function. Detailed information for deriving these relationships can be found in Appendix A.

4. Flood Frequency Analysis

For FFA, three problems were addressed. First, the GB2 distribution was tested using observed flood data, and was compared with commonly used distributions in hydrology. Second, a method for selecting the best distribution was discussed. Third, flood frequency analysis was carried out at several gauging stations from upstream to downstream, and the evolution of frequency distribution along the stream was investigated.

4.1. Flood Data

Flood data from eight gauging stations on the Colorado River and its tributaries, as shown in Figure 3, were considered to test the performance of the GB2 distribution and discuss the evolution of frequency distribution along the river. The Colorado River is the principal river of the Southwestern United States and northwest Mexico. It rises in the central Rocky Mountains, flows generally southwest across the Colorado Plateau and through the Grand Canyon. The basin boundary consists of mountains that are 13,000 to 14,000 feet (3962.4 m to 4267.2 m) high in Wyoming, Colorado, and Utah; and the

boundary drops to elevations of less than 1000 feet (304.8 m) at Hoover Dam. The northern part of the river basin in Colorado and Wyoming is a mountainous plateau that ranges from 5000 to 8000 feet (1524 m to 2438 m) in elevation, which encompasses deep canyons, rolling valleys, and intersecting mountain ranges. The central and southern portions of the basin in eastern Utah, northwestern New Mexico, and northern Arizona consist of rugged mountain ranges interspersed with rolling plateaus and broad valleys. In general, the mountains in the southern part of the basin are much lower than those in the northern part. Of the eight gauging stations considered in this study, gauging stations or sites 1, 2 and 3 are on the Yampa River which is a secondary tributary of the Colorado River. Sites 4, 5, 6, 7 and 8 are on the mainstream of the Colorado River. Site 8 is near the location of the Hoover Dam. The data of these gauging stations is directly downloaded from USGS (United States Geological Survey) website. The characteristics of flow data of these gauging stations, including length of the data, mean, standard deviation, skewness, and kurtosis, were calculated, as shown in Table 1. Since there is a dam, named Glenn Canyon, regulating the river flow past Lees Ferry (shown in Figure 3), the characteristics of the flow at the Hoover dam (site 8) are quite different from those at sites 4, 5, 6 and 7 upstream. It can be seen from Table 1 that for sites 1 to 7 the mean values increase from upstream to downstream, as more rainfall or water flows into the river. Since the standard deviation is related to the flood magnitude, it also increases with the mean value. For site 8, considering the impact of reservoir operation, some streamflow was stored in the reservoir, which leads that the streamflow at site 8 is reduced. The skewness is positive for all gauging stations, indicating that the right tail is longer or fatter than the left side and the mass of distribution is concentrated on the left side. Kurtosis is a measure of the peakedness of the probability distribution. The skewness and kurtosis values in the mainstream are generally lower than those in the tributaries.

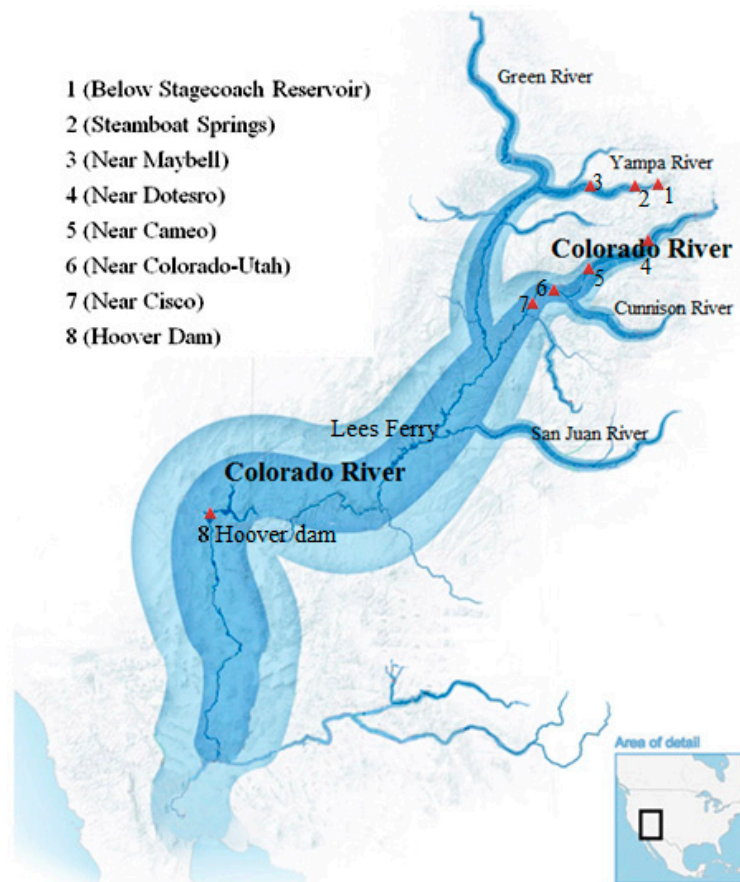


Figure 3. Locations of gauging stations on the Colorado River.

Table 1. Characteristics of the gauging stations used in the study.

River	No.	Gaging Station	Drainage Area (Square Miles)	Length of Data	Mean Value (ft ³ /s)	Standard Deviation	Skewness	Kurtosis
Yampa River	1	Below Stagecoach Reservoir	228	1957–2014	315	189	0.91	3.49
	2	Steamboat Springs	567	1904–2013	3630	1115	0.26	2.94
	3	Near Maybell	3383	1904–2013	10,419	3657	0.90	4.88
Colorado River	4	Near Dotsero	4390	1941–2013	9870	4450	0.39	2.59
	5	Near Cameo	7986	1934–2013	19,049	7687	0.26	2.68
	6	Near Colorado-Utah	17,847	1951–2013	26,714	13,936	0.84	3.53
	7	Near Cisco	24,100	1884–2013	34,329	16,520	0.36	2.31
	8	Hoover Dam	171,700	1934–2013	26,131	6831	1.37	5.83

4.2. Performance Measures

For evaluating the performance of the GB2 distribution, two measures were employed: (1) the root mean square deviation (RMSD); and (2) the Akaike information criterion (AIC). These methods assess the fitted distribution at a site by summarizing the deviations between observed discharges and computed discharges.

A frequently used method for assessing the goodness-of-fit of a function is the RMSD [12]. This method was used by NERC (1975) for ranking candidate distributions [13]. RMSD can be expressed as:

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{Q_{the}(i) - Q_{emp}(i)}{Q_{emp}(i)} \right)^2} \quad (4)$$

where n is the sample size; Q_{the} is the computed discharge at the i^{th} plotting position. Q_{emp} denotes the observed i^{th} smallest discharge. The value of RMSD is from 0 to 1. The smaller is, the better the distribution fits.

AIC is a measure of the relative quality of statistical models for a given set of data. It also includes a penalty that is an increasing function of the number of estimated parameters. The AIC value was calculated as [14]:

$$AIC = n(\ln(MSE)) + 2K \quad (5)$$

where K is the number of parameters of the distribution, and MSE was calculated by

$$MSE = \frac{1}{n} \sum_{i=1}^n (Q_{the}(i) - Q_{emp}(i))^2 \quad (6)$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

4.3. Evaluation of GB2 Distribution

Annual maximum flood peak data from four gauging stations, namely sites 2, 6, 7 and 8 in Figure 3, were selected. The empirical frequencies were calculated first. The purpose of defining the empirical distribution is to compare it with selected theoretical distributions in order to verify whether they fit sample data.

Many plotting positions are proposed, most of which can be expressed in general form:

$$P_i = \frac{i - a}{n + 1 - 2a} \quad (7)$$

where a is a constant having values from 0 to 0.5 in different formula, 0.5 for Hazen's formula, 0.3 for Chegadayev's formula, zero for Weibull's formula, 3/8 for Blom's formula, 1/3 for Tukey's formula, and 0.44 for Gringorten's formula.

Among these formulars, Gringorten’s formular is recognized by lots of researchers, especially for GEV, gumbel, exponential, Generalized pareto distributions which have been widely used for flood frequency analysis [15–20]. The Gringorten formula is also used for GB2 distribution. For normal, generalized normal and Gamma distributions, the Blom’s formula is recommended [21,22]. For Pearson type 3 and log Pearson type 3 distributions, Weibull’s formula is recommended [18,21]. The GB2 distribution was employed to fit the annual maximum (AM) series of the four sites. The distribution parameters were estimated using Equation (3) and given in Table 2. The fitted GB2 distributions and empirical frequency of each AM series are shown in Figure 4. In the left of Figure 4, the line represents the fitted distribution and circle the empirical frequencies of observations. Results show that the marginal distributions fit the empirical data well. Histograms of AM flood peak series fitted by the GB2 distribution for the gauging stations on the Colorado River are shown in the right section of Figure 4. It also indicates that the GB2 distribution can successfully be fitted to empirical histograms.

Several distributions, including normal, exponential, gamma, Gumbel, generalized normal, pearson type III, log Pearson type III, generalized Pareto, and generalized extreme-value that are commonly used in hydrology, were fitted to the AM series at this site. The L-moment method was used to estimate the parameters of these distributions.

Table 2. Parameters of the GB2 distribution for the gauging stations along the Colorado River.

Number	Location	r_1	r_2	r_3	β
4	Near Dotsero	1.58	60.30	1.75	85.11
5	Near Cameo	1.12	77.57	2.53	112.93
6	Near Colorado-Utah	3.94	83.08	0.94	69.05
7	Near Cisco	2.73	76.82	1.07	80.90
8	Hoover Dam	10.59	434.72	1.31	43.62

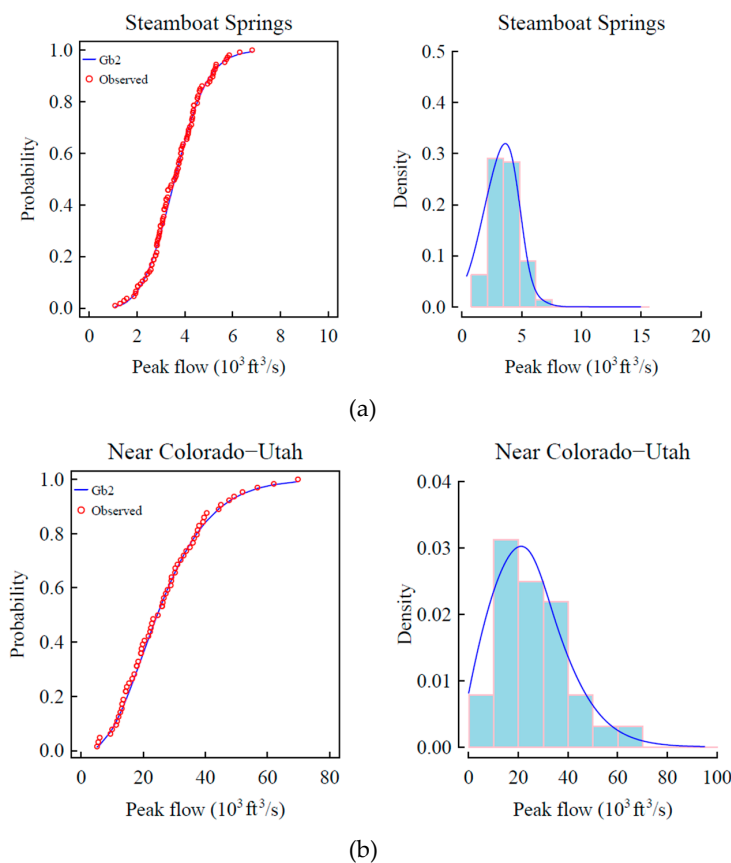


Figure 4. Cont.

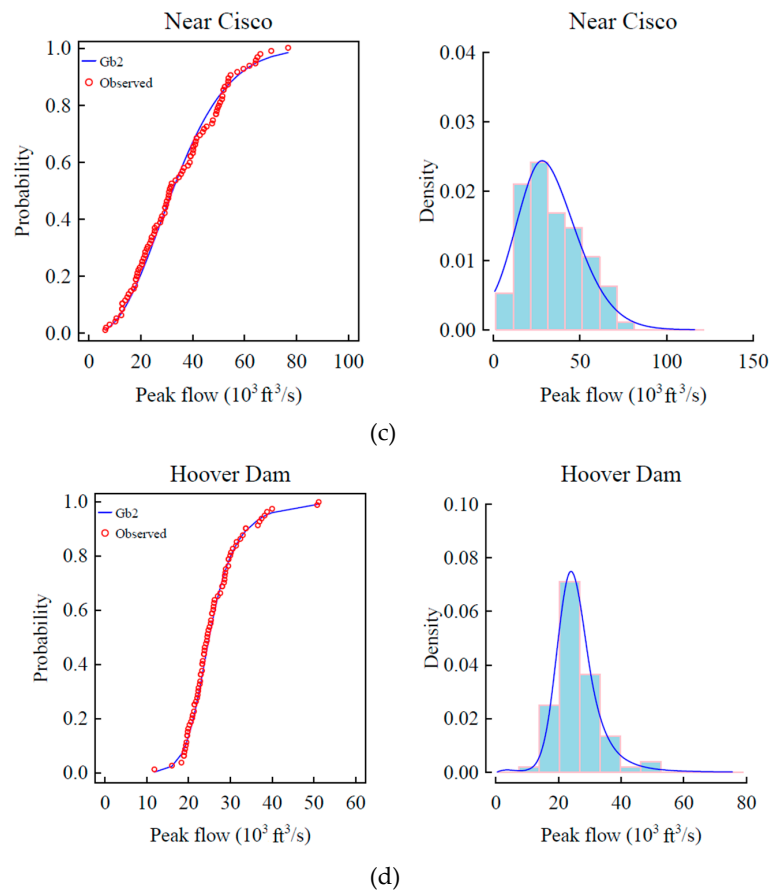


Figure 4. Marginal distributions and histograms of AM flood peak series fitted by the GB2 distribution for the gauging stations on the Colorado River. (a) Steamboat springs; (b) Near Colorado-Utah; (c) Near Cisco; (d) Hoover Dam.

Singh and Guo compared the POME method with the L-moment method, and indicated that the two methods are comparable [11,23,24]. Therefore no matter what method is used, it has little influence on the value of the T-year design discharge. The Kolmogorov-Smirnov test was used here to compare a sample with a reference probability distribution. The p -value was calculated and given in Table 3 as well. The higher or more close to 1 the p -value is the more similar the theoretical and empirical distributions are. It is indicated from Table 3 that the p -value of GB2 distribution is 1 or close to 1, which demonstrates that the GB2 distribution fit the data better. Table 3 also listed the RMSD and AIC values computed for the fitted GB2 distribution using Equations (4)–(7). The smaller the RMSD and AIC values are, the better the distribution fits. For the site streamboat springs, the GB2 and generalized normal distributions have the smallest RMSD values, which is equal to 0.025. For the site Near Cisco, the GB2 has the smallest RMSE values, which is equal to 0.061. For the site Near Colorado-Utah, the GB2 and gamma distributions have the smallest RMSE value. For the site Hoover dam, the GB2 distribution has the smallest RMSE value. Since the GB2 distribution have more parameters, the AIC values of GB2 distribution are larger than those of generalized normal, Gamma and GEV distributions. Thus, generally GB2 distribution gives a getter fit.

Table 3. RMSE and AIC values of different distributions.

Number	Distribution	Steamboat Springs			Near Cisco			Near Colorada-Utah			Hoover Dam		
		<i>p</i> -Value	RMSE	AIC	<i>p</i> -Value	RMSE	AIC	<i>p</i> -Value	RMSE	AIC	<i>p</i> -Value	RMSE	AIC
1	GB2	0.976	0.025	924.1	0.991	0.061	1384.1	1	0.047	852.7	1	0.036	1098.8
2	Normal	0.926	0.043	992.9	0.787	0.194	1502	0.839	0.221	1031	0.436	0.081	1236.2
3	Exponential	0.409	0.122	1306.8	0.336	0.171	1669.6	0.839	0.145	1036.4	0.919	0.055	1152.6
4	Gamma	1	0.045	1005.1	0.959	0.064	1512.8	1	0.047	842.5	0.692	0.057	1192.9
5	Gumbel	0.976	0.066	1143.5	0.959	0.088	1546.2	1	0.107	869.1	0.978	0.039	1122.3
6	Generalized normal	0.844	0.025	922.9	0.991	0.137	1455.5	1	0.083	852.8	0.978	0.039	1106.7
7	Pearson type III	0.976	0.035	953.2	0.991	0.1	1425	1	0.054	895.6	1	0.058	1146.8
8	Log Pearson type III	0.976	0.034	951.3	0.991	0.106	1431.5	1	0.054	893.1	1	0.052	1133
9	Generalized Pareto	0.976	0.078	1158	0.991	0.062	1386.1	1	0.09	960.2	1	0.054	1169
10	GEV	0.976	0.027	929.6	0.991	0.138	1450.1	1	0.128	865.5	1	0.036	1096.8

In order to compare the POME with the current used method, the maximum likelihood (ML) method was also employed for the parameter estimation of GB2 distribution. Taking the site Near Colorada-Utah for an example, the estimated parameters by POME and ML method are given in Table 4. The p -value, RMSE and AIC values are also given in Table 4. It is indicated that the parameters obtained by the two method are more or less the same. And the RMSE and AIC values based on the POME method are smaller.

Table 4. Parameters estimated by POME and ML methods for site Near Colorada-Utah.

Methods	r_1	r_2	r_3	β	p -Value	RMSE	AIC
POME	2.14	24.78	1.40	157.18	1	0.0169	−357.76
ML	2.26	30.85	1.35	158.55	1	0.0170	−357.80

4.4. Flood Frequency Analysis

The Hoover dam is a multi-purpose dam, serving the needs of flood control, irrigation, water supply, and hydropower generation. Therefore, it was desired to determine the most appropriate distribution for FFA at the dam site. The T-year design flood at Hoover dam was calculated using each distribution, as given in Table 5, and it can be seen that different distributions yielded significantly different values. For example, the 1000-year design flood values calculated by the GB2 and gamma distributions were 76,702 and 50,485 ft³/s, respectively. The RMSD and AIC values for GB2 distribution (Gamma distribution) were 0.036 (0.057) and 1098.8 (1192.9), respectively, which indicates that the performance of GB2 distribution is much better than that of the gamma distribution. It concludes that if the gamma distribution were used, the design flood would be underestimated and potential flood risk would be higher.

Table 5. Comparison of T-year design flood discharges (10³ ft³/s) calculated by different distributions for the Hoover dam site.

Number	Return Period	1000	500	100	50	10
1	GB2	76.702	67.914	51.198	34.138	30.125
2	Normal	45.800	44.451	40.938	34.288	31.488
3	Exponential	68.561	63.583	52.024	35.486	30.508
4	Gamma	50.485	48.424	43.314	34.613	31.320
5	Gumbel	58.926	55.332	46.973	34.799	30.912
6	Generalized normal	50.513	49.271	45.325	35.732	31.485
7	Pearson type III	60.025	56.451	47.985	35.145	30.926
8	Log Pearson type III	69.568	64.494	52.713	35.639	31.858
9	Generalized Pareto	64.809	59.870	49.084	34.893	30.695
10	GEV	57.809	54.766	47.324	35.270	31.072

4.5. Change in Flood Frequency Distribution with Change in Drainage Area

The GB2 distribution was applied for FFA along the main stem of the Colorado River. Four gauging stations (sites 4, 5, 6 and 7) from upstream to downstream were used, as shown in Figure 3 and Table 6. These gauging stations were selected, because all these stations are on the mainstream and no dam has been built on this reach. The drainage area and statistical characteristics (including mean, skewness and kurtosis of the annual maximum data) of these stations were calculated, as given in Table 1. The T-year design flood of these gauging stations was calculated, as shown in Figure 5, in which the x -axis represents the return periods and the y -axis represents the design flood values. Figure 5 shows that for a given return period, the design flood value of the downstream gauging stations is larger than that of the upstream gauging stations. The increasing rates of drainage area and T-year design flood values between the adjacent gauging stations were computed, as given in Table 6, which indicates that the percentage increase of the drainage area was nearly the same as

that of the design flood values. For instance, with the increase of drainage area up to 45% from the gauging station near Dotsero to that near Cameo, the flood value increased by 43% on average. It is also seen that from upstream to downstream, when the drainage area increased by 45%, 55% and 26%, the flood value increased by 43%, 42%, and 16%, respectively. It seems that in a mountainous watershed, the upstream the reach is, the greater the impact the drainage area has on flood. This may be because that the runoff coefficient is generally larger in the steep area.

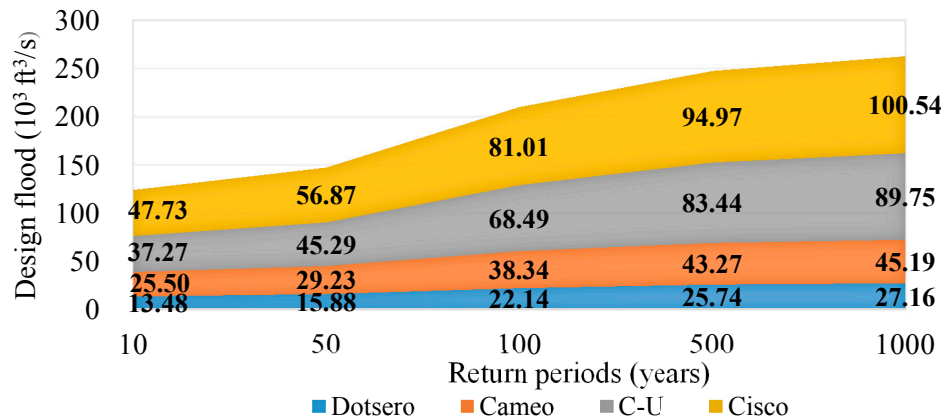


Figure 5. Flood values along the mainstream of the upper Colorado River.

Table 6. Statistical characteristics of the four gauging stations, the increasing rate of drainage area and flood discharge between adjacent gauging stations.

Number	Locations	Drainage Area (Square Miles)	Increase in Drainage Area (%)	Increase in in Flood Value (%)					
				1000	500	100	50	10	Mean
4	Near Dotsero	11370	45	40	41	42	47	46	43
5	Near Cameo	20683	55	50	48	44	32	35	42
6	Near Colorado-Utah	46228	26	11	12	15	22	20	16
7	Near Cisco	62419							

4.6. Evolution of Frequency Distribution along Stream

In order to determine the evolution of frequency distribution and its parameters along the river, data from the Yampa River were applied, because this river is taken as one of the west’s last wild rivers and has only a few small dams and diversions. The Yampa River with a length of 402 km, located in northwestern Colorado, is a tributary of Green River and a secondary tributary of the Colorado River. Data from three gauging stations along this river, designated as sites 1, 2 and 3 in Figure 6, were used. The GB2 distribution was used to fit the AM series of each of the three gauging stations, as shown in Table 7. It can be seen that shape parameters r_1 and r_2 decreased along the river. The value of r_1 became close to be 1. When r_1 equals 1, the GB2 distribution becomes the Burr XII distribution [25]. This distribution has been shown to reasonably fit the income distribution data [20,26,27] and has recently been used in hydrology [2,28]. The PDF of Burr XII distribution can be written as:

$$f(x) = \frac{r_3}{bB(1, r_2)} \left(\frac{x}{b}\right)^{1 \times r_3 - 1} \left(1 + \left(\frac{x}{b}\right)^{r_3}\right)^{-(r_2 + 1)} = \frac{r_3 r_2}{b} \left(\frac{x}{b}\right)^{r_3 - 1} \left(1 + \left(\frac{x}{b}\right)^{r_3}\right)^{-(r_2 + 1)} \quad (8)$$

where b is the scale parameter. The Burr XII distribution was also used to fit the data at the gauging station near Maybell of Yampa River. The estimated parameters of Burr XII distribution were: $r_2 = 1.94$, $r_3 = 4.19$, and $b = 12.33$. The fitting results of the GB2 and Burr distributions for the gauging station near Maybell are shown in Figure 7. For the gauging station near Maybell, parameters of the GB2 distribution estimated by POME method are nearly as the same as the parameters of the Burr XII distribution estimated by MLE method. Thus, Burr XII distribution instead of GB2 distribution can be

used for FFA at that station. In other words, the distribution for FFA changes from the four-parameter GB2 distribution to the three-parameter Burr XII distribution along the Yampa River. There is an evolution of distribution along this river. From Equation (1), the value of scale parameter β increases with the mean value, because more water flows into the stream. Parameters r_1 and r_2 govern the left and right tails, respectively. The smaller the value of r_1 , the fatter the left tail is; and the smaller the value of r_2 , the fatter the right tail is. It can be seen from Table 7 that both r_1 and r_2 decrease along the stream, which demonstrates that both the left and right tails become fatter, and the PDF values become larger in these areas and lower in the central area.

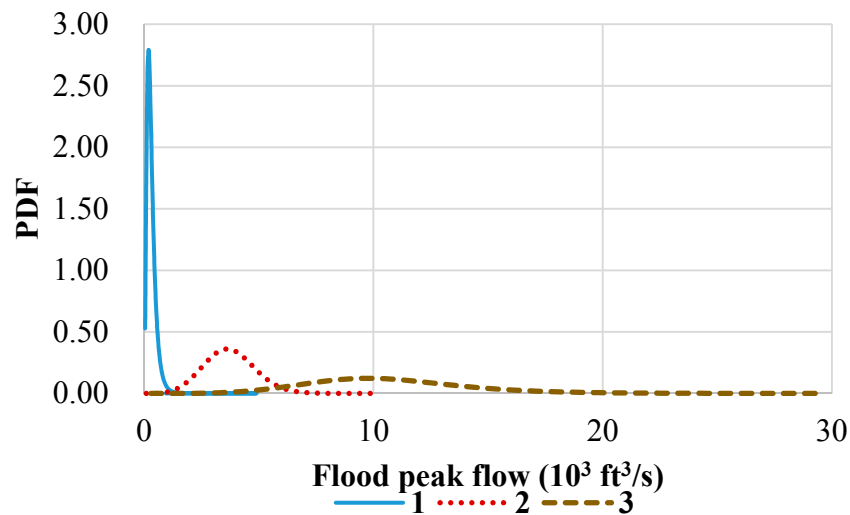


Figure 6. Evaluations of PDF of sites along the Yampa River.

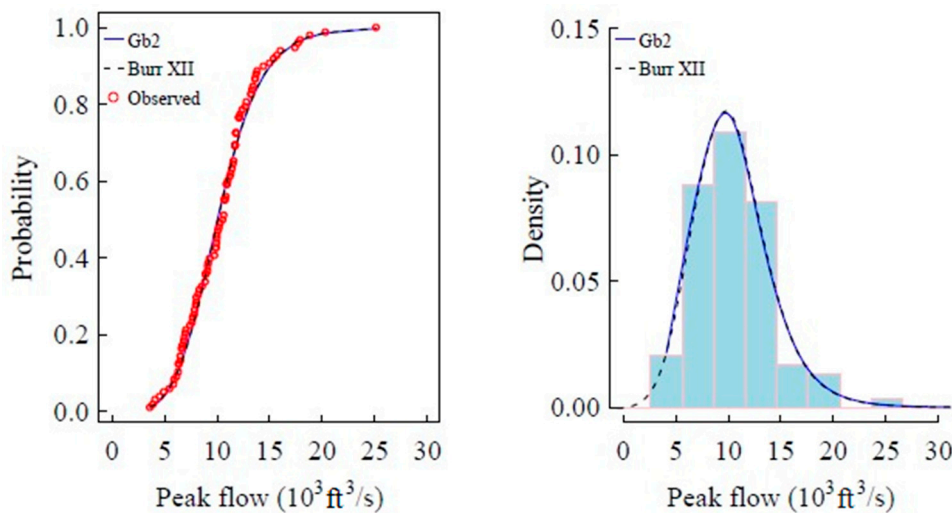


Figure 7. Marginal distribution and histograms of AM flood peak series fitted by the GB2 and Burr XII distributions for the gauging station near Maybell on the Yampa River.

Table 7. Parameters of the GB2 distribution for four gauging stations along the Yampa River.

Number	Location	r_1	r_2	r_3	β
1	Below stagecoach Reservoir	17.44	15.25	0.55	2.10
2	Steamboat springs	1.20	5.49	3.59	5.81
3	Near Maybell	1.14	2.07	3.92	12.11

5. Conclusions

The GB2 provides sufficient flexibility to fit a large variety of data sets. Papalexiou and Koutsoyiannis introduced this distribution in hydrology and used it for rainfall frequency analysis [8]. In this study, the generalized beta distribution of the second kind (GB2) is introduced for FFA for the first time. The POME method was proposed to estimate the parameters of GB2 distribution. Equations of POME method was deduced by ourselves and given in Appendix A. The Colorado River basin was selected as a case study to test the performance of GB2 distribution. Frequency estimates from the GB2 distribution were also compared with those of commonly used distributions in hydrology. In addition, some characteristics of FFA in mountainous areas are discussed. The conclusions can be summarized as follows:

- (1) Results demonstrate that the GB2 is appealing for FFA, since it has four parameters which allows the distribution to be able to fit data having very different histogram shapes, such as the J-shaped and bell-shaped distributions. And by fixing certain parameters, the GB2 distribution can yield some well-known distributions, such as the beta distribution of the second kind (B2), the Burr type XII, generalized gamma (GG), and so on.
- (2) The parameters estimated by POME method are found reasonable. Both the marginal distributions and histograms indicates that the GB2 distribution can successfully be fitted to empirical values using the POME method.
- (3) The performance of the GB2 distribution is better than that of the widely used distributions in hydrology. For the site streamboat springs, the GB2 and generalized normal distributions have the smallest RMSD values. For the site Near Cisco, the GB2 has the smallest RMSE values. For the site Near Colorado-Utah, the GB2 and gamma distributions have the smallest RMSE value. For the site Hoover dam, the GB2 distribution has the smallest RMSE value. Since the GB2 distribution have more parameters, the AIC values of GB2 distribution are larger than those of generalized normal, Gamma and GEV distributions. Thus, generally GB2 distribution gives a better fit.
- (4) When using different distributions for FFA, significant different design flood values are obtained. It concludes that if the wrong distribution were used, the design flood would be underestimated and potential flood risk would be higher.
- (5) The design flood value increase with the drainage area. For a given return period, the design flood value of the downstream gauging stations is larger than that of the upstream gauging stations. In this study, the percentage increase of the drainage area was nearly the same as that of the design flood values. It seems that in a mountainous watershed, the upstream the reach is, the greater the impact the drainage area has on flood. This may be because that the runoff coefficient is generally larger in the steep area.
- (6) There is an evolution of distribution along this river. Along the Yampa River, the distribution for FFA changes from the four-parameter GB2 distribution to the three-parameter Burr XII distribution. And both r_1 and r_2 decrease along the stream, which demonstrates that both the left and right tails become fatter, and the PDF values become larger in these areas and lower in the central area, which means that when the drainage area become larger, the flood magnitudes has a more significant variation.

Acknowledgments: The project was financially supported by the National Natural Science Foundation of China (51679094, 51509273, 91547208 and 41401018), Fundamental Research Funds for the Central Universities (2017KFYXJJ194, 2016YXZD048).

Author Contributions: Vijay P. Singh conceived and designed the experiments; Lu Chen performed the experiments and analyzed the data; Lu Chen wrote the draft of the paper and Vijay P. Singh revised it. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Estimation of Parameters of GB2 Distribution

The GB2 distribution parameters can be estimated by maximizing the Shannon entropy $H(X)$ which, for a random variable X , can be expressed as:

$$H(X) = - \int_0^{\infty} f(x) \log f(x) dx \quad (\text{A1})$$

where $f(x)$ is the probability density function (PDF). The principle of maximum entropy (POME) indicates that the most appropriate PDF is the one that maximizes the value of entropy, given available data and a set of known constraints [29].

Specification of Constraints: Following Singh, the constraints for the GB2 distribution can be expressed as

$$\int_0^{\infty} f(x) dx = 1 \quad (\text{A2a})$$

$$\int_0^{\infty} f(x) \ln x dx = E(\ln x) \quad (\text{A2b})$$

$$\int_0^{\infty} f(x) \ln(1 + (\frac{x}{\beta})^{r_3}) dx = E(\ln(1 + (\frac{x}{\beta})^{r_3})) \quad (\text{A2c})$$

Method of Lagrange Multipliers for Maximizing Entropy: In the search for an appropriate probability distribution for a given random variable, entropy should be maximized. In other words, the best fitted distribution is the one with the highest entropy. The method of Lagrange multipliers was used to obtain the appropriate probability distribution with the maximum entropy. Finally, the form of this distribution is given as:

$$f(x) = \exp(-\lambda_0 - \lambda_1 \ln(x) - \lambda'_2 \ln(1 + (\frac{x}{\beta})^{r_3})) \quad (\text{A3a})$$

in which λ_0 , λ_1 , and λ'_2 are the Lagrange multipliers. Let $p = \beta^{-r_3}$. Then, Equation (A3a) can be written as

$$f(x) = \exp(-\lambda_0 - \lambda_1 \ln(x) - \lambda'_2 \ln(1 + px^{r_3})) \quad (\text{A3b})$$

Let $\lambda'_2 = \frac{\lambda_2}{p}$ and $q = r_3$. Papalexiou and Koutsoyiannis defined the entropy-based PDF as:

$$f(x) = \exp(-\lambda_0 - \lambda_1 \ln(x) - \frac{\lambda_2}{p} \ln(1 + px^q)) \quad (\text{A4})$$

Substitution of Equation (A4) in Equation (A2a) yields:

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} \exp(-\lambda_0 - \lambda_1 \ln(x) - \frac{\lambda_2}{p} \ln(1 + px^q)) dx = 1 \quad (\text{A5})$$

From Equation (A5):

$$\begin{aligned} \exp(\lambda_0) &= \int_0^{\infty} \exp(-\lambda_1 \ln x - \lambda_2 \ln(1 + px^q)/p) dx \\ &= \int_0^{\infty} \exp(-\lambda_1 \ln x) \exp(-\frac{\lambda_2}{p} \ln(1 + px^q)) dx \\ &= \int_0^{\infty} x^{(-\lambda_1)} (1 + px^q)^{(-\frac{\lambda_2}{p})} dx \end{aligned} \quad (\text{A6})$$

Let $t = px^q$. Then $x = (\frac{t}{p})^{\frac{1}{q}}$, and $dx = \frac{1}{pq}(\frac{t}{p})^{\frac{1}{q}-1} dt$. Thus, Equation (A6) can be expressed as:

$$\begin{aligned} \exp(\lambda_0) &= \int_0^\infty x^{(-\lambda_1)}(1 + px^q)^{(-\frac{\lambda_2}{p})} dx \\ &= \int_0^\infty (\frac{t}{p})^{-\frac{\lambda_1}{q}} (1 + t)^{-\frac{\lambda_2}{p}} \frac{1}{pq} (\frac{t}{p})^{\frac{1}{q}-1} dt \\ &= \int_0^\infty \frac{1}{q} p^{\frac{\lambda_1-1}{q}} t^{-\frac{\lambda_1}{q}} (1 + t)^{-\frac{\lambda_2}{p}} t^{\frac{1}{q}-1} dt \end{aligned} \tag{A7}$$

Let $y = \frac{t}{1+t}$. Then $t = \frac{y}{1-y}$, and $dt = \frac{1}{(1-y)^2} dy$.
 Since $y(0) = 0$ and $y(\infty) = 1$, $y \in [0, 1]$.

$$\begin{aligned} \exp(\lambda_0) &= \int_0^1 \frac{1}{q} p^{\frac{\lambda_1-1}{q}} (\frac{y}{1-y})^{-\frac{\lambda_1}{q}} (1 + \frac{y}{1-y})^{-\frac{\lambda_2}{p}} (\frac{y}{1-y})^{\frac{1}{q}-1} \frac{1}{(1-y)^2} dy \\ &= \int_0^1 \frac{1}{q} p^{\frac{\lambda_1-1}{q}} (\frac{y}{1-y})^{-\frac{\lambda_1+1}{q}} (1 + \frac{y}{1-y})^{-\frac{\lambda_2}{p}} \frac{1}{(1-y)^2} dy \\ &= \int_0^1 \frac{1}{q} p^{\frac{\lambda_1-1}{q}} (\frac{y}{1-y})^{-\frac{\lambda_1+1}{q}} (\frac{1}{1-y})^{-\frac{\lambda_2}{p}+2} dy \\ &= \int_0^1 \frac{1}{q} p^{\frac{\lambda_1-1}{q}} (y)^{\frac{1-\lambda_1}{q}-1} (1-y)^{-\frac{1-\lambda_1}{q} + \frac{\lambda_2}{p}-1} dy \\ &= \frac{1}{q} p^{\frac{\lambda_1-1}{q}} B(\frac{1-\lambda_1}{q}, -\frac{1-\lambda_1}{q} + \frac{\lambda_2}{p}) \end{aligned} \tag{A8}$$

The Lagrange multiplier λ_0 can be calculated from Equation (A8) as:

$$\lambda_0 = -\ln q + \frac{\lambda_1 - 1}{q} \ln(p) + \ln \Gamma(\frac{1 - \lambda_1}{q}) + \ln \Gamma(-\frac{1 - \lambda_1}{q} + \frac{\lambda_2}{p}) - \ln \Gamma(\frac{\lambda_2}{p}) \tag{A9}$$

From Equation (A4), the other equation for calculating λ_0 can be defined as:

$$\lambda_0 = \ln(\int_0^\infty \exp(-\lambda_1 \ln x - \frac{\lambda_2}{p} \ln(1 + px^q)) dx) \tag{A10}$$

Relation between Lagrange multipliers and constraints: Defining $a' = \frac{1-\lambda_1}{q}$ and $b' = -\frac{1-\lambda_1}{q} + \frac{\lambda_2}{p}$, differentiate Equation (A9) with respect to λ_1 and λ_2 :

$$\begin{aligned} \frac{\partial \lambda_0}{\partial \lambda_1} &= \frac{\ln p}{q} + \frac{\partial \ln \Gamma(a')}{\partial a'} \frac{\partial a'}{\partial \lambda_1} + \frac{\partial \ln \Gamma(b')}{\partial (b')} \frac{\partial b'}{\partial \lambda_1} - \frac{\partial \ln \Gamma(a+b')}{\partial (a+b')} \frac{\partial (a+b')}{\partial \lambda_1} \\ &= \frac{\ln p}{q} - \frac{1}{q} \varphi(a') + \frac{1}{q} \varphi(b') \end{aligned} \tag{A11a}$$

$$\begin{aligned} \frac{\partial \lambda_0}{\partial \lambda_2} &= \frac{\partial \ln \Gamma(b')}{\partial (b')} \frac{\partial b'}{\partial \lambda_2} - \frac{\partial \ln \Gamma(a+b')}{\partial (a+b')} \frac{\partial (a+b')}{\partial \lambda_2} \\ &= \frac{1}{p} \varphi(b') - \frac{1}{p} \varphi(a' + b') \end{aligned} \tag{A11b}$$

where $\varphi(\cdot)$ is a digamma function. Differentiate Equation (A10) with respect to λ_1 and λ_2 :

$$\frac{\partial \lambda_0}{\partial \lambda_1} = \frac{\int_0^\infty \ln x \exp(-\lambda_1 \ln x - \frac{\lambda_2}{p} \ln(1 + px^q)) dx}{\int_0^\infty \exp(-\lambda_1 \ln x - \frac{\lambda_2}{p} \ln(1 + px^q)) dx} = -E(\ln x) \tag{A12a}$$

$$\frac{\partial \lambda_0}{\partial \lambda_2} = \frac{\int_0^{\infty} x^q \exp(-\lambda_1 \ln x - \frac{\lambda_2}{p} \ln(1 + px^q)) dx}{\int_0^{\infty} \exp(-\lambda_1 \ln x - \frac{\lambda_2}{p} \ln(1 + px^q)) dx} = -E\left(\frac{\ln(1 + px^q)}{p}\right) \quad (\text{A12b})$$

Based on Equations (A11) and (A12), the relation between Lagrange multipliers and constraints can be expressed as:

$$\frac{\ln p}{q} - \frac{1}{q} \varphi(a) + \frac{1}{q} \varphi(b) = -E(\ln x) \quad (\text{A13a})$$

$$\frac{1}{p} \varphi(b) - \frac{1}{p} \varphi(a + b) = -E\left(\frac{\ln(1 + px^q)}{p}\right) \quad (\text{A13b})$$

Since there are four parameters, Equations (A13a) and (A13b) are not sufficient for calculating parameters, and two additional equations are needed that are given as:

$$\frac{\partial^2 \lambda_0}{\partial^2 \lambda_1} = \frac{1}{q^2} \varphi'(a') + \frac{1}{q^2} \varphi'(b') = \text{var}(\ln x) \quad (\text{A14a})$$

$$\frac{\partial^2 \lambda_0}{\partial^2 \lambda_2} = \varphi'(r_2) - \varphi'(r_1 + r_2) = \text{var}(\ln(1 + (\frac{x}{\beta})^q)) \quad (\text{A14b})$$

Relation between Lagrange multipliers and parameters: Substituting Equation (A8) in Equation (A4), it is known that:

$$f(x) = \frac{1}{\frac{1}{q} p^{\frac{\lambda_1-1}{q}} B(\frac{1-\lambda_1}{q}, -\frac{1-\lambda_1}{q} + \frac{\lambda_2}{p})} x^{-\lambda_1} (1 + px^q)^{-\frac{\lambda_2}{p}} \quad (\text{A15})$$

Equation (A15) is the GB2 distribution. Comparing Equation (1) with Equation (A15), the following equations can be obtained:

$$\begin{aligned} \lambda_1 &= 1 - r_1 q \\ \lambda_2 &= p(r_2 + \frac{1-\lambda_1}{q}) \\ p &= (\frac{1}{\beta})^{r_3} \\ q &= r_3 \end{aligned} \quad (\text{A16})$$

Relation between parameters and constraints: Based on the relation between parameters and constraints, and parameters and Lagrange multipliers, the relation between parameters and constraints can be expressed as:

$$\begin{aligned} -\ln \beta - \frac{1}{r_3} \varphi(r_1) + \frac{1}{r_3} \varphi(r_2) &= -E(\ln x) \\ \beta^{r_3} \varphi(r_2) - \beta^{r_3} \varphi(r_1 + r_2) &= -E(\beta^{r_3} \ln(1 + (\frac{x}{\beta})^{r_3})) \\ -\ln \beta + \frac{1}{r_3^2} \varphi'(r_1) + \frac{1}{r_3^2} \varphi'(r_2) &= \text{var}(\ln x) \\ \varphi'(r_2) - \varphi'(r_1 + r_2) &= \text{var}(\ln(1 + (\frac{x}{\beta})^{r_3})) \end{aligned} \quad (\text{A17})$$

References

1. Beven, K.J.; Hornberger, G.M. Assessing the effect of spatial pattern of precipitation in modeling streamflow hydrographs. *J. Am. Water Resour. Assoc.* **1982**, *18*, 823–829. [\[CrossRef\]](#)
2. Shao, Q.; Wong, H.; Xia, J.; Ip, W. Models for extremes using the extended three-parameter Burr XII system with application to flood frequency analysis. *Hydrol. Sci. J.* **2004**, *49*, 685–702. [\[CrossRef\]](#)
3. Chen, L.; Guo, S.L.; Yan, B.W.; Liu, P.; Fang, B. A new seasonal design flood method based on bivariate joint distribution of flood magnitude and date of occurrence. *Hydrol. Sci. J.* **2010**, *55*, 1264–1280. [\[CrossRef\]](#)
4. Chen, L.; Singh, V.P.; Guo, S.; Hao, Z.; Li, T. Flood coincidence risk analysis using multivariate copula functions. *J. Hydrol. Eng.* **2012**, *17*, 742–755. [\[CrossRef\]](#)

5. Mielke, P.W., Jr.; Johnson, E.S. Some generalized beta distributions of the second kind having desirable application features in hydrology and meteorology. *Water Resour. Res.* **1974**, *10*, 223–226. [[CrossRef](#)]
6. Wilks, D.S. Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series. *Water Resour. Res.* **1993**, *29*, 3543–3549. [[CrossRef](#)]
7. Perreault, L.; Bobée, B.; Rasmussen, P. Halphen distribution system. I: Mathematical and statistical properties. *J. Hydrol. Eng.* **1999**, *4*, 189–199. [[CrossRef](#)]
8. Papalexiou, S.M.; Koutsoyiannis, D. Entropy based derivation of probability distributions: A case study to daily rainfall. *Adv. Water Resour.* **2012**, *45*, 51–57. [[CrossRef](#)]
9. Singh, V.P. *Entropy Based Parameter Estimation in Hydrology*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1998.
10. Singh, V.P. *Entropy-Based Parameter Estimation Hydrology*; Springer: Dordrecht, The Netherlands, 1998.
11. Singh, V.P.; Guo, H. Parameter estimation for 3-parameter generalized Pareto distribution by the principle of maximum entropy (POME). *Hydrol. Sci. J.* **1995**, *40*, 165–181. [[CrossRef](#)]
12. Karim, A.; Chowdhury, J.U. A comparison of four distributions used in flood frequency analysis in Bangladesh. *Hydrol. Sci. J.* **1995**, *40*, 55–66. [[CrossRef](#)]
13. Natural Environment Research Council. *Flood Studies Report*; Natural Environment Research Council: London, UK, 1975; Volumes 1–5.
14. Zhang, L.; Singh, V.P. Bivariate flood frequency analysis using the copula method. *J. Hydrol. Eng.* **2006**, *11*, 150–164. [[CrossRef](#)]
15. Ross, R. Graphical method for plotting and evaluating weibull distribution data. In Proceedings of the 4th International Conference on Properties and Application of Dielectric Materials, Brisbane, Australia, 3–8 July 1994; pp. 250–253.
16. Cunnane, C. Unbiased plotting positions—A review. *J. Hydrol.* **1978**, *37*, 205–222. [[CrossRef](#)]
17. Makkonen, L. Notes and correspondence plotting positions in extreme value analysis. *J. Appl. Meteorol. Clim.* **2006**, *45*, 334–340. [[CrossRef](#)]
18. Shabri, A. A Comparison of plotting formulas for the pearson type III distribution. *J. Technol.* **2002**, *36*, 61–74. [[CrossRef](#)]
19. Gringorten, I.I. A plotting rule for extreme probability paper. *J. Geophys. Res.* **1963**, *68*, 813–814. [[CrossRef](#)]
20. Dagum, C. A New Model of Personal Income Distribution: Specification and Estimation. In *Modeling Income Distributions and Lorenz Curves*; Springer: New York, NY, USA, 2008; pp. 3–25.
21. Mehdi, F.; Mehdi, J. Determination of plotting position formula for the normal, log-normal, pearson(III), log-pearson(III) and gumble distribution hypotheses using the probability plot correlation coefficient test. *World Appl. Sci. J.* **2011**, *15*, 1181–1185.
22. Kim, S.; Shin, H.; Kim, T.; Taesoon, K.; Heo, J. Derivation of the probability plot correlation coefficient test statistics for the generalized logistic distribution. In Proceedings of the International Workshop Advances in Statistical Hydrology, Taormina, Italy, 23–25 May 2010; pp. 1–8.
23. Singh, V.P.; Guo, H. Parameter estimation for 2-parameter log-logistic distribution (LLD2) by maximum entropy. *Civ. Eng. Syst.* **1995**, *12*, 343–357. [[CrossRef](#)]
24. Singh, V.P.; Guo, H. Parameter estimations for 2-parameter Pareto distribution by pome. *Stoch. Hydrol. Hydraul.* **1980**, *9*, 81–93.
25. Burr, I.W. Cumulative Frequency Functions. *Ann. Math. Stat.* **1942**, *13*, 215–232. [[CrossRef](#)]
26. Kleiber, C.; Kotz, S. *Statistical Size Distributions in Economics and Actuarial Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2003.
27. Singh, S.K.; Maddala, G.S. A function for size distribution of incomes. *Econometrica* **1976**, *44*, 963–970. [[CrossRef](#)]
28. Hao, Z.; Singh, V.P. Entropy-based parameter estimation for extended Burr XII distribution. *Stoch. Environ. Res. Risk Assess.* **2008**, *23*, 1113–1122. [[CrossRef](#)]
29. Singh, V.P. Hydrologic synthesis using entropy theory: Review. *J. Hydrol. Eng.* **2011**, *16*, 421–433. [[CrossRef](#)]

