water

MDPI

Article

BLP3-SP: A Bayesian Log-Pearson Type III Model with Spatial Priors for Reducing Uncertainty in Flood Frequency Analyses

Dan Tian and Lei Wang *

Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA; dtian2@lsu.edu

* Correspondence: leiwang@lsu.edu; Tel.: +1-225-578-8876

Abstract: Gauge stations have uneven lengths of discharge records owing to the historical hydrologic data collection efforts. For watersheds with limited water data length, the flood frequency model, such as the Log-Pearson Type III, will have large uncertainties. To improve the flood frequency prediction for these watersheds, we propose a Bayesian Log-Pearson Type III model with spatial priors (BLP3-SP), which uses a spatial regression model to estimate the prior distribution of the parameters from nearby stations with longer data records and environmental factors. A Markov chain Monte Carlo (MCMC) algorithm is used to estimate the posterior distribution and associated flood quantiles. The method is validated using a case study watershed with 15 streamflow gauge stations located in the San Jacinto River Basin in Texas, US. The result shows that the BLP3-SP outperforms other choices of the priors for the Bayesian Log-Pearson Type III model by significantly reducing the uncertainty in the flood frequency estimation for the station with short data length. The results have confirmed that the spatial prior knowledge can improve the Bayesian inference of the Log-Pearson Type III flood frequency model for watersheds with short gauge period.

Keywords: BLP3-SP; flood frequency analyses; Log-Pearson Type III distribution; Bayesian; spatial prior; uncertainty

check for updates

Citation: Tian, D.; Wang, L. BLP3-SP: A Bayesian Log-Pearson Type III Model with Spatial Priors for Reducing Uncertainty in Flood Frequency Analyses. *Water* **2022**, *14*, 909. https://doi.org/10.3390/w14060909

Academic Editors: Stefano Morelli, Veronica Pazzi and Mirko Francioni

Received: 28 January 2022
Accepted: 9 March 2022
Published: 14 March 2022

1. Introduction

A design flood is a hypothetical peak discharge graph representation of previous knowledge of precipitation frequency in an area, which is commonly used to evaluate the construction of dams, bridges, canals, and flood damage desistance systems. Flood records do not fit any specific known statistical distributions. Nevertheless, to make the determination of flood frequency trackable, it is convenient to select a reasonable distribution. Bulletin 17C recommends the Log-Pearson Type III (LP3) distribution for design-flood prediction in the United States [1]. Several algorithms can be used to estimate the LP3 distribution parameters. The methods of moments and the maximum likelihood are the most commonly used methods in flood frequency analysis [2]. The limited length of gauged data is one of the major sources of the uncertainties of the predicted design floods. For example, the 100-year flood is an international default design flood. The longer the gauge records, the more accurately predicted design flood. However, most areas are ungauged or recently gauged, leading to large uncertainty in flood frequency models. Spatial information expansion (SIE) is a technique used to employ the knowledge learned from nearby sites or sites from similar environments to substitute space from time [2–5], in order to improve the accuracy of the flood frequency estimate at the site of interest. The assumption is that the hydrological regime of nearby watersheds is similar, therefore resulting in similar flood frequency distribution.

Meanwhile, Bayesian methods have also been applied to flood frequency analysis using instrumental data when it is possible to use conjugate priors or semi-conjugate priors [2,6–9]. The Markov chain Monte Carlo (MCMC) algorithm has been used to estimate the parameters of the Bayesian inference if conjugate or semi-conjugate priors are absent [10,11]. Several flood frequency studies have applied Bayesian approaches with priors obtained from regional information [7,12–16]. Merz et al. and Viglione et al. made the spatial expansion in flood frequency hydrology with a geostatistical regionalization method called top-kriging [17–19]. Nguyen et al. took advantage of the index flood principle, assuming that the average annual peak discharges are scaled to the drainage area in a statistically homogeneous region [20]. Lima et al. applied a hierarchical Bayesian GEV model to improve the estimation of local and regional flood quantiles, which assumed that both the location and scale parameters for all sites were identical except a scale factor based on the watershed area [21]. These studies considered either spatial proximity or catchment attributes for the spatial extension. However, Merz and Blöschl compared four flood regionalization methods and concluded that spatial proximity, together with catchment attributes, outperformed spatial proximity only and then catchment attributes only [22].

This paper proposes a Bayesian Log-Pearson Type III model with spatial priors (BLP3-SP) that considers both spatial proximity and catchment attributes as the prior information to reduce the uncertainty in estimated flood frequency. The hyperparameters of the prior distribution is calculated from regional sites with longer systematic data series than the target site, using the spatial lagged model and the spatial error model. The research question is whether the BLP3-SP model can produce accurate flood prediction without using long-time series observation data. In the following sections, the question is answered by analyzing the data of the 15 streamflow gauge stations located in the San Jacinto River Basin in Texas, US.

2. Methods

To improve the parameter estimation for the LP3 distribution, we incorporate spatial information as the prior of a Bayesian inference framework (Figure 1). The prior distributions are estimated from the parameters of nearby sites using spatial regression models. The posterior distribution is inferred by an ensemble MCMC algorithm as well a Metropolis and Metropolis–Hastings algorithm within a Gibbs sampler. The estimations and uncertainties of the parameters and quantiles are calculated by sampling directly from the posterior distribution.

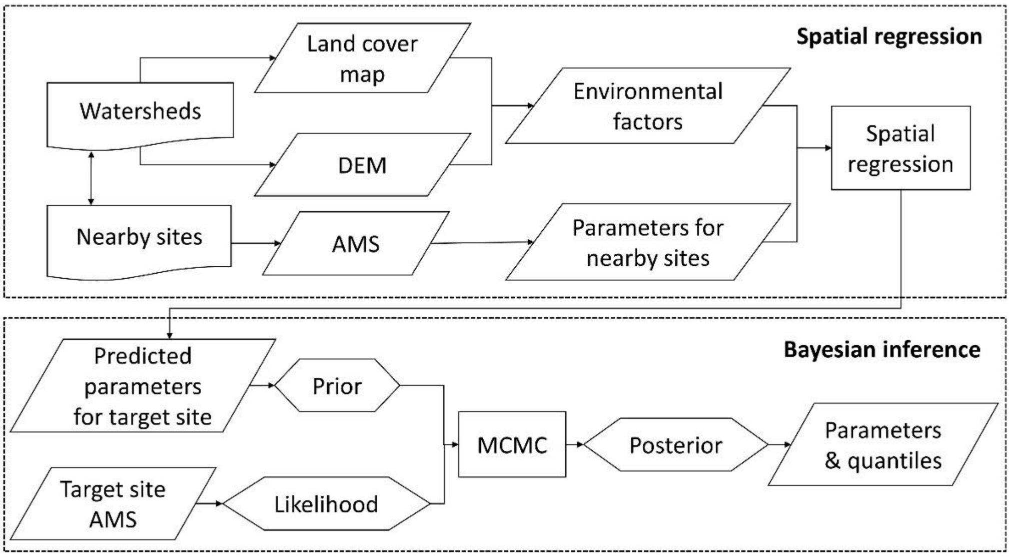

Figure 1. Flowchart of the BLP3-SP processing.

2.1. Log-Pearson Type III Distribution

The Log-Pearson Type III (LP3) distribution is recommended by the United States Water Resources Committee for the flood frequency estimation [1,2,23,24]. When the flood peak discharge time series {Q_1, Q_2, . . . , Q_N} are distributed as a Log-Pearson Type III distribution, $X = \log(Q)$ distributes as a Pearson Type III distribution, with a probability density function (pdf):

$$f_X(x) = \frac{|\beta|}{\Gamma(\alpha)}[\beta(x-\tau)]^{\alpha-1}e^{-\beta(x-\tau)} \tag{1}$$

where α, β, and τ are the shape, scale, and location parameters, respectively; and $\Gamma(\alpha)$ is the gamma function.

Another parameterizing of LP3 distribution is usually used to calculate the pth quantile, which is based on the mean (μ), standard deviation (σ), and skewness (γ) [2]. First, the Pearson variate X (log Q) is transferred to the standard normal variable z for modest skews γ by applying the Wilson–Hilferty transformation [25]:

$$f_X(x) = \phi(z)\frac{dz}{dX} = \frac{\phi(z)}{\sigma[\frac{\gamma}{2}\left(\frac{x-\mu)}{\sigma}\right)+1]^{\frac{2}{3}}} \tag{2}$$

where $\phi(z)$ is the standard normal probability density function for z. The cumulative distribution function (cdf) is:

$$F_X(x) = \int_0^x f_X(t)dt \tag{3}$$

Meanwhile, the pth quantile can be calculated as:

$$x_p = \mu + \sigma K_p(\gamma) \tag{4}$$

where $K_p(\gamma)$ is the pth quantile of the LP3 distribution with mean 0, standard deviation 1, and skewness γ, named as the frequency factor. It can also be approximated by the Wilson–Hilferty transformation for $|\gamma| < 2$ [25]:

$$K_p(\gamma) = \frac{2}{\gamma}(1 + \frac{\gamma z_p}{6} - \frac{\gamma^2}{36})^3 - \frac{2}{\gamma} \tag{5}$$

where z_p is the pth quantile of the standard normal distribution.

2.2. Bayesian Theorem for LP3 Distribution

According to the Bayes theorem, the probability of parameter θ given the observed dataset $X = \{x_1, x_2, x_3, \ldots, x_s\}$ (posterior) is proportional to the product of the probability of θ (prior) and the probability of X given θ (likelihood). Assuming the independence between the observations, the posterior can be calculated as below:

$$p(\theta|X) \propto p(\theta)l(X|\theta) = \prod_{i=1}^{s} p(\theta) \times f_X(x_i) \tag{6}$$

where $p(\theta|X)$ is the posterior distribution, $p(\theta)$ is the prior distribution, $l(X|\theta)$ is the likelihood, and $f_X(\,)$ is the pdf for X. In this study, θ comprises mean μ, standard deviation σ, and skewness γ in Equation (2).

2.3. Prior Distribution

The posterior belief of the parameter's distribution is based on a prior belief. In this study, we assume normal distributions for the mean μ and skewness γ, while a log-normal distribution for the standard deviation σ is based on the previously suggested distributions [2,18,26].

$$\mu \sim N\left(\mu_\mu, \sigma_\mu^2\right) \tag{7}$$

$$\log(\sigma) \sim N\left(\mu_{\log(\sigma)}, \sigma^2_{\log(\sigma)}\right) \tag{8}$$

$$\gamma \sim N\left(\mu_\gamma, \sigma_\gamma^2\right) \tag{9}$$

where $\left\{\mu_\mu, \sigma_\mu^2, \mu_{\log(\sigma)}, \sigma_{\log(\sigma)}^2, \mu_\gamma, \sigma_\gamma^2\right\}$ are the are the hyperparameters for the prior distributions.

The prior distribution of the main model is calculated from the data of nearby stations using spatial regression models. A spatial regression model takes both the catchment characteristics and the spatial proximity into consideration at the same time. It deals with the spatial autocorrelation in two ways: the spatial lagged model (SLM) and the spatial error model (SEM). SLM assumes that the magnitudes of the dependent variable depend on the magnitude of its neighbors [27], which is expressed as follow:

$$y = \rho W y + X\beta + \varepsilon \tag{10}$$

where y is a vector of the variable of interest (the flood distribution parameters in this study), r is the spatial coefficient, W is the spatial weight that defines the strength of the spatial autoregressive process, X is a matric of the catchment characteristics, β is a vector of regression coefficients, and ε is a vector of uncorrelated error assumed to be of normal distribution with zero mean and constant variance.

SEM handles the spatial dependencies among the error term after applying the ordinary least squares (OLS) model to spatial variables, which is given in the following equations:

$$y = X\beta + v \tag{11}$$

$$v = \lambda W v + \varepsilon \tag{12}$$

where v is a vector of error with spatial dependencies and λ is the spatial error coefficient.

Five independent variables—size, elevation, vegetation cover, imperviousness, and slope—were considered; however, not every variable was used to estimate all the three parameters. Before we performed the spatial regression, a multiple linear regression was used to select the important variables for each parameter using the stepwise selection method based on the Akaike information criterion (AIC). We tested both SLM and SEM with several types of weights in this study and selected the ones with the smallest spatial coefficient p-value for each parameter.

A non-informative prior, namely the prior with minimal effect on the posterior distribution compared to the experiment, was also applied to full-length data series to generate a baseline flood frequency estimation based only on the information from data records [28]. Specifically, the non-informative priors for μ, log (σ), and γ are set to a mean of 0 and a variance of 10,000.

2.4. Parameter Estimation

To estimate the parameters and flood quantiles from the posterior distribution, a Markov chain Monte Carlo (MCMC) algorithm was used in this study. MCMC is a type of algorithm for sampling from probability distributions, which formulates a Markov chain that has the desired distribution as its equilibrium distribution [18,29,30]. A Markov chain is a sequence of random variables $\theta^{(1)}$, $\theta^{(2)}$, ... , for which, for any t, the distribution of $\theta^{(t)}$given all previous θ' depends only on the most recent value, $\theta^{(t-1)}$,

$$p\left(\theta^{(t)} \middle| \theta^{(1)}, \ldots, \theta^{(t-1)}\right) = p(\theta^{(t)}|\theta^{(t-1)}) \tag{13}$$

Based on drawing values of θ from approximate distributions and then comparing the probability of proposed location and current location to accept or reject the drawing, the chain with a large number of steps was treated as a sample of the desired distribution.

We applied the Gibbs sampler to sample the three parameters one by one within each iteration [31,32]. Since the proposal distributions for μ and γ are symmetric, we used the Metropolis algorithm to simulate them [33]. For σ, we applied the Metropolis–Hastings algorithm because its proposal distribution is not symmetric, which will be discussed in the next section [34].

2.5. Proposal Distribution

A proper proposal distribution J_θ is key to effective implementation of the Metropolis and Metropolis–Hastings algorithms. Based on the study of Reis and Steginger [2], we generated the proposed values of the three parameters independently based only on their values at the previous step.

The proposal distribution for the mean μ is a normal with mean $\mu^{(t-1)}$ and variance $\sigma^{2(t-1)}/s$,

$$\mu^* \sim N\left(\mu^{(t-1)}, \frac{\sigma^{2(t-1)}}{s}\right) \tag{14}$$

The proposal distribution for σ is a gamma distribution with mean $\sigma^{(t-1)}$ and variance modeled as a function of $\sigma^{(t-1)}$ and $\gamma^{(t-1)}$ [35],

$$\sigma^* \sim \gamma(a,b) \tag{15}$$

$$a = \frac{\sigma^{2(t-1)}}{Var\left(\sigma^{(t-1)}\right)}\ ,\ b = \frac{Var\left(\sigma^{(t-1)}\right)}{\sigma^{(t-1)}} \tag{16}$$

$$Var\left(\sigma^{(t-1)}\right) = \frac{\sigma^{2(t-1)}\left(1 + 0.75\gamma^{2(t-1)}\right)}{2s} \tag{17}$$

The proposal distribution for γ is a normal distribution with mean $\gamma^{(t-1)}$ and variance modeled as a function of $\gamma^{(t-1)}$ and s [1],

$$\gamma^* \sim N\left[\gamma^{(t-1)}, Var(\gamma)\right] \tag{18}$$

$$Var(\gamma) = [1 + \frac{6}{s}]^{2}\, 10^{-b\log\left(\frac{s}{10}\right)} \tag{19}$$

$$a = \begin{cases} -0.33 + 0.08\left|\gamma^{(t-1)}\right| & if\left|\gamma^{(t-1)}\right| < 0.90 \\ -0.52 + 0.30\left|\gamma^{(t-1)}\right| & if\left|\gamma^{(t-1)}\right| > 0.90 \end{cases} \tag{20}$$

$$b = \begin{cases} 0.94 - 0.26\left|\gamma^{(t-1)}\right| & if\left|\gamma^{(t-1)}\right| < 1.50 \\ 0.55 & if\left|\gamma^{(t-1)}\right| > 1.50 \end{cases} \tag{21}$$

After sampling the parameters, we estimated the marginal density distributions, computed means and standard errors, and estimated credible intervals of the parameters and some desired quantiles.

3. Case Study Area and Data

3.1. Study Area and Gauge Station Data

We applied the proposed model to a series of annual peak discharges for 15 streamflow gauges (Table 1) located in the hydrologic accounting unit 120,401, San Jacinto, which covers the San Jacinto River Basin above Galveston Bay, Texas (Figure 2). This area is to the northwest of the city of Houston, with a total area of 10,308 km^2. Flood frequency can be estimated using the annual maximum series (AMS) or partial duration series (PDS). The AMS consists of records of the annual peak discharge, while the PDS is based on all floods exceeding a predefined base line [1]. If minor floods are considered (AEP > 0.10), PDS is more appropriate than AMS. However, for floods with an annual exceedance probability (AEP) less than 0.10, there is no significant difference between the AEP estimation using

AMS or PDS [36]. Meanwhile, due to its wide availability and longer data length, AMS has also been used in many studies [16,20,37]. Therefore, AMS was used in this study.

Table 1. Summary of the 15 watersheds with more than 50 records.

Site ID	Site No.	Latitude	Longitude	Watershed Area (km^2)	Series Length (year)
1	08075780	29.95 N	95.52 W	18.76	55
2	08074150	29.85 N	95.49 W	15.90	53
3	08068000	30.24 N	95.46 W	2158.28	85
4	08075400	29.62 N	95.45 W	48.36	55
5	08068500	30.11 N	95.44 W	1052.12	82
6	08069000	30.04 N	95.43 W	737.34	77
7	08075900	29.96 N	95.42 W	86.42	54
8	08074500	29.78 N	95.40 W	227.43	85
9	08076500	29.86 N	95.33 W	69.14	67
10	08076000	29.92 N	95.31 W	166.20	67
11	08070500	30.26 N	95.30 W	271.47	76
12	08075500	29.67 N	95.29 W	150.76	67
13	08075770	29.79 N	95.27 W	47.79	56
14	08071000	30.23 N	95.17 W	307.53	56
15	08070000	30.34 N	95.10 W	859.51	81

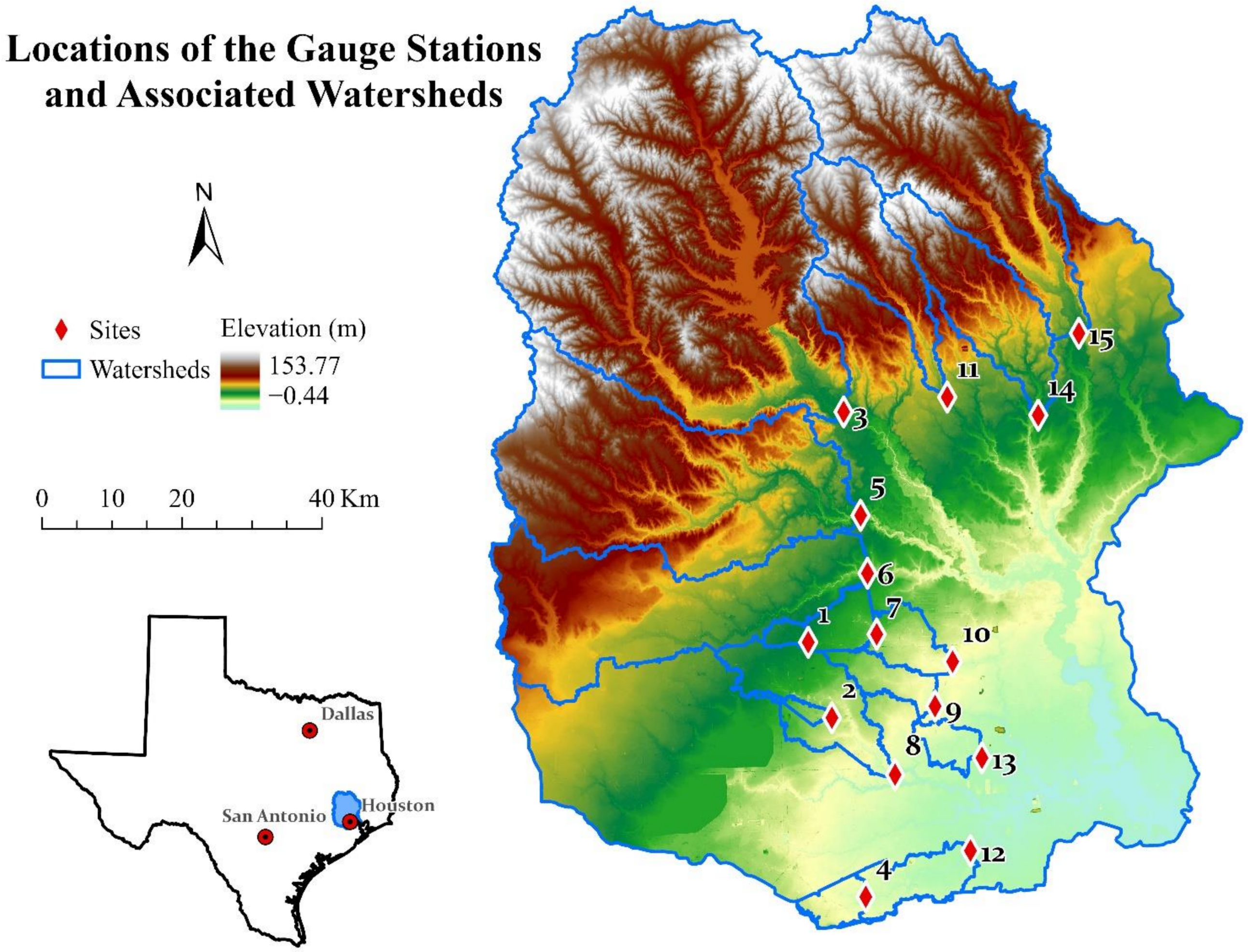

Figure 2. Study area, locations of the 15 gauge stations, and associated watersheds.

The annual peak discharge time series data were obtained from the USGS National Water Information System. Among the 95 sites in the San Jacinto accounting unit, there are 76 sites with records longer than 5 years. Using the $^1/_3$ arc-second seamless DEM dataset of the 3D Elevation Program, we generated 76 watersheds from the gauge stations. 29 of the 76 generated watersheds had areas different to the drainage area for the same site in the USGS National Water Information System; thus, they were removed from the dataset, which left 47 stations. 15 of the 47 stations have more than 50 years of data (Figure 3). The site number 08074150 (ID: 2) was used for testing and validation, revealing 53 peak discharge records since 1964 with missing data for 1987, 1988, and 1989. A baseline Bayesian model was built using the full length of the records from this site with non-informative prior. The baseline model is used for evaluating the BLP3-SP models calibrated using 10, 20, and 30 years of records with spatial prior computed from other 14 stations.

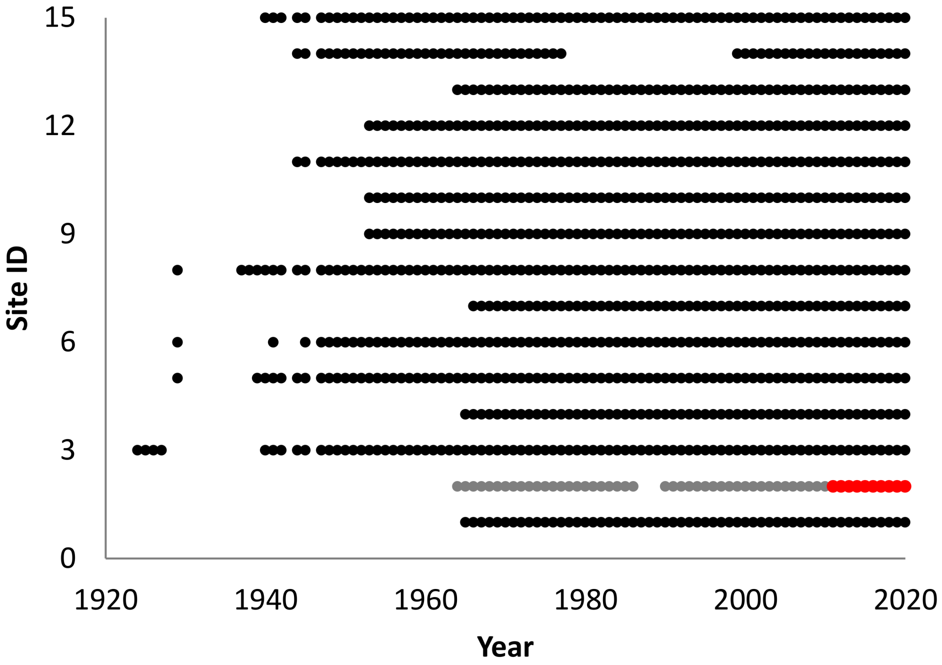

Figure 3. Data availability of the entire set of streamflow gauges used in this study (the gray row is the site for testing and validation and the red dots are the 10-year time series used in the model).

3.2. Spatial Data for Prior Estimation

To use spatial regression, LP3 parameters for the nearby stations were estimated from the gauge station data. The independent environmental factors include the area, elevation, slope, tree canopy cover, and the urban impervious surface for each watershed. The USGS National Water Information System provides the watershed area associated with each gauge station. The elevation and slope were obtained from the $^1/_3$ arc-second seamless DEM, with a spatial resolution of ~10 m. The tree canopy and urban imperviousness were downloaded from the National Land Cover Database (NLCD) with a spatial resolution of 30 m. The mean of the factors and the local LP3 parameters for each watershed are summarized in Table 2.

Table 2. Summary of the variables for the 15 watersheds.

ID	Area (km^2)	Elevation (m)	Slope (%)	Tree Canopy (%)	Imperviousness (%)
1	18.76	37.64	24.22	8.79	39.94
2	15.90	28.61	24.37	3.36	52.30
3	2158.28	87.00	59.96	52.58	2.18
4	48.36	19.90	27.95	5.00	30.34
5	1052.12	69.91	40.04	47.49	5.88
6	737.34	54.79	20.71	11.92	9.41
7	86.42	33.75	23.95	13.54	34.05
8	227.43	29.25	27.29	6.29	44.21
9	69.14	24.25	18.84	9.90	35.42
10	166.20	29.42	25.31	12.73	33.59
11	271.47	88.01	56.58	52.83	1.94
12	150.76	16.02	29.34	6.83	28.67
13	47.79	14.90	19.38	4.66	51.31
14	307.53	69.16	46.66	69.56	1.04
15	859.51	88.35	61.48	71.02	0.60

4. Results

4.1. Estimated Prior Information from Spatial Regression

The Bayesian prior distributions of the three LP3 parameters were estimated using spatial regression models. Before running the spatial regression models, we applied multiple linear regression and selected the important independent variables for each LP3 parameter by the information index, AIC. Based on the results of the multiple regression models, we selected *Area* as the independent variable for μ, *Tree canopy* for σ, and *Area* and *Elevation* for γ.

Both the spatial error model and the spatial lag model were tested in this study with eight weight types, which are (1) first-order Queen, (2) second-order Queen, (3) 4-NN (make symmetric), (4) distance band (max–min distance), (5) distance (15,240 m), (6) distance (60,960 m), (7) distance (45,720 m), and (8) triangular kernel with 3-NN adaptive bandwidth. The p-value of each model is summarized in Table 3. Three models have a p-value less than 0.05 for estimating μ, which are (1) SEM with first-order Queen weight (1st Queen SEM), (2) SEM with triangular kernel with 3-NN adaptive bandwidth (triangular kernel SEM), and (3) SLM with 4-NN (4-NN SLM). There are two potential models for estimating σ, which are (1) SEM with second-order Queen weight (2nd Queen SEM) and (2) SLM with a triangular kernel with a 3-NN adaptive bandwidth (triangular kernel SLM). Only one model has a p-value less than 0.05 for the estimation of γ, which is SLM with a second-order Queen weight (2nd Queen SLM).

Table 3. Summary of the p-values for each spatial regression model (the significant ones are bolded).

Weight Type	μ		σ		γ	
	SEM	SLM	SEM	SLM	SEM	SLM
First-order Queen weight	**0.024**	0.646	0.110	0.563	0.974	0.555
Second-order Queen weight (including the lower order)	0.618	0.497	**0.008**	0.646	0.187	**0.039**
4-NN	0.126	**0.023**	0.237	0.244	0.057	0.280
Distance band (Max-Min distance)	0.187	0.117	0.058	0.073	0.706	0.645
Distance band (15,240 m)	0.359	0.669	0.775	0.908	0.994	0.085
Distance band (60,960 m)	0.389	0.273	0.799	0.144	0.526	0.589
Distance band (45,720 m)	0.137	0.068	0.536	0.142	0.196	0.223
Triangular kernel with 3-NN adaptive bandwidth	**0.001**	0.428	0.071	**0.001**	0.385	0.239

The R^2 and standard deviation for each potential model are summarized in Table 4. For the models to estimate μ and σ, the one with the greatest R^2 and the smallest standard deviation was selected. Therefore, the model for estimating the prior μ distribution is the

4-NN SLM, the model for estimating the prior σ distribution is the 2nd Queen SEM, and the model for estimating the prior γ distribution is the 2nd Queen SLM.

Table 4. Summary of the R^2 and standard deviation of models with significant spatial coefficient (p-value < 0.05).

	Model	**R^2**	**STD**
μ	1st Queen SEM	56.18%	0.20
	4-NN SLM	60.57%	0.18
	Triangular kernel SEM	33.20%	0.30
σ	2nd Queen SEM	78.03%	0.01
	Triangular kernel SLM	47.69%	0.03
γ	2nd Queen SLM	59.02%	0.06

These spatial regression models provide the mean and variance for the distributions of μ, σ, and γ, which represent the prior information used in the Bayesian model. The values are summarized in Table 5.

Table 5. Mean and variance for the LP3 parameters estimated from the spatial regression models.

	μ	σ	γ
mean	7.8957	0.6993	−0.3969
variance	0.1778	0.0143	0.0635

4.2. Posterior Distribution and Flood Quantiles

The prior information obtained in the previous section was applied to the test gauge site using only the last 10 years of records. For comparison, a baseline model with no prior information applied to the Bayes inference was trained with 54 years of data. The posterior means and variances of the three parameters are listed in Table 6.

Table 6. Means and variances of the posterior distributions.

	μ		σ		γ	
	Mean	**Variance**	**Mean**	**Variance**	**Mean**	**Variance**
Spatial regression prior with 10-year data	7.5338	0.0278	0.5264	0.0183	−0.3133	0.2516
Non-informative prior with 54-year data	7.0798	0.0093	0.6624	0.0098	−0.2349	0.3011

Figure 4 shows the means and 95% confidence limits of the predicted design floods from the three scenarios: the Bayesian LP3 model calibrated with 54-year data and non-informative prior, the Bayesian LP3 model calibrated with 10-year data and non-informative prior, and the BPL-SP model calibrated with 10-year data and spatial prior. The means and the lower boundaries of the three scenarios are similar. The upper boundaries, however, show large differences. The baseline model returned the lowest uncertainty for large flood magnitudes. The non-informative prior model with only 10-year records has the largest uncertainty. By using the spatial prior, the uncertainty of the large floods is reduced to the level similar with the baseline model. The test confirms that the Bayesian estimation can use the prior knowledge learned from the nearby stations and environment factors to reduce the uncertainty caused by short data length.

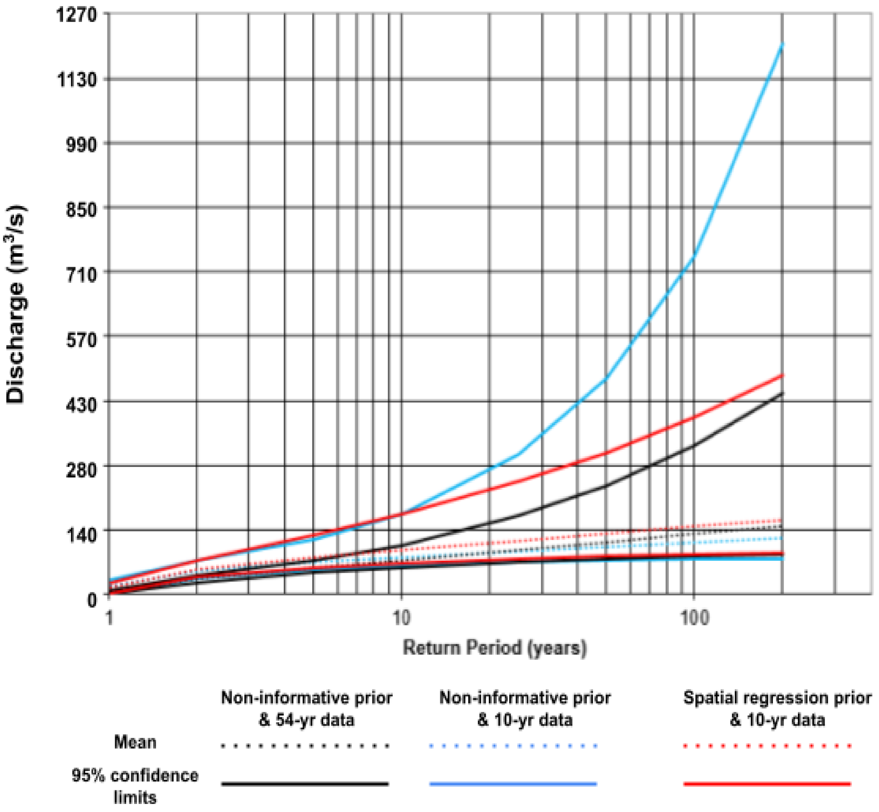

Figure 4. The means and 95% confidence limits for the three scenarios: non-informative prior and 54-year data, non-informative prior and the last 10-year data, and spatial regression prior and the last 10-year data.

Table 7 displays the discharge for certain design floods with a 95% confidence interval and a reduction in confidence intervals for each design flood. With an increase in the return period, the reduction in the confidence interval is more drastic. For floods with a return period more than 50 years, the prior knowledge from spatial regression could reduce almost half of the uncertainty.

Table 7. Estimation of the discharge (m^3/s) for certain design floods with 95% confidence interval.

Return Period	10 Year	25 Year	50 Year	100 Year	200 Year
Non-info prior and 10-year data	82.9 (63.3–175.6)	96.1 (71.8–307.0)	106.1 (75.9–475.4)	115.5 (78.1–744.3)	125.2 (79.2–1027.6)
Spatial regression and 10-year data	97.5 (70.0–176.2)	118.9 (79.3–248.8)	134.5 (84.1–312.0)	149.6 (87.8–390.1)	164.7 (90.6–482.6)
Reduction in confidence interval	5.35%	27.93%	42.95%	54.62%	65.26%

5. Discussion

5.1. Compared with Other Spatial Prior Methods

Spatial regression considers both the catchment characteristics and the spatial proximity at the same time. To demonstrate the superiority of spatial regression, we compared it with two other types of spatial priors: mean prior and areal interpolation prior. The first method uses the arithmetic mean and variance calculated from the nearby site records [38], and the other uses the areal interpolation technique that is similar to the top-kriging algorithm [19].

The priors and associated posteriors are summarized in Table 8. Compared with areal interpolation prior, the posterior generated by mean prior is similar with the one generated by spatial regression. Figure 5 shows that the mean prior can also reduce the uncertainty in the quantile estimation, but much less than the spatial regression prior. However, the areal interpolation prior generates a larger confidence interval compared to the non-informative prior result, even with small return periods. It shows that spatial interpolation may not be applicable to watersheds because of the hierarchical structure of the watershed system. Overall, within the prior types tested in this study, spatial regression provides the best results.

Table 8. Different prior types and associated posterior distributions.

		μ		σ		γ	
		Mean	**Variance**	**Mean**	**Variance**	**Mean**	**Variance**
Mean Prior	Prior	8.2205	0.4291	0.9052	0.0690	−0.2822	0.1627
	Posterior	7.5374	0.0309	0.4848	0.0374	−0.2190	0.3650
Areal Interpolation Prior	Prior	9.1225	0.2963	0.8840	0.0136	−0.8983	0.1506
	Posterior	7.6258	0.0670	0.7340	0.0220	−0.6832	0.8131
Spatial Regression Prior	Prior	7.8957	0.1778	0.6993	0.0143	−0.3969	0.0635
	Posterior	7.5338	0.0278	0.5264	0.0183	−0.3133	0.2516

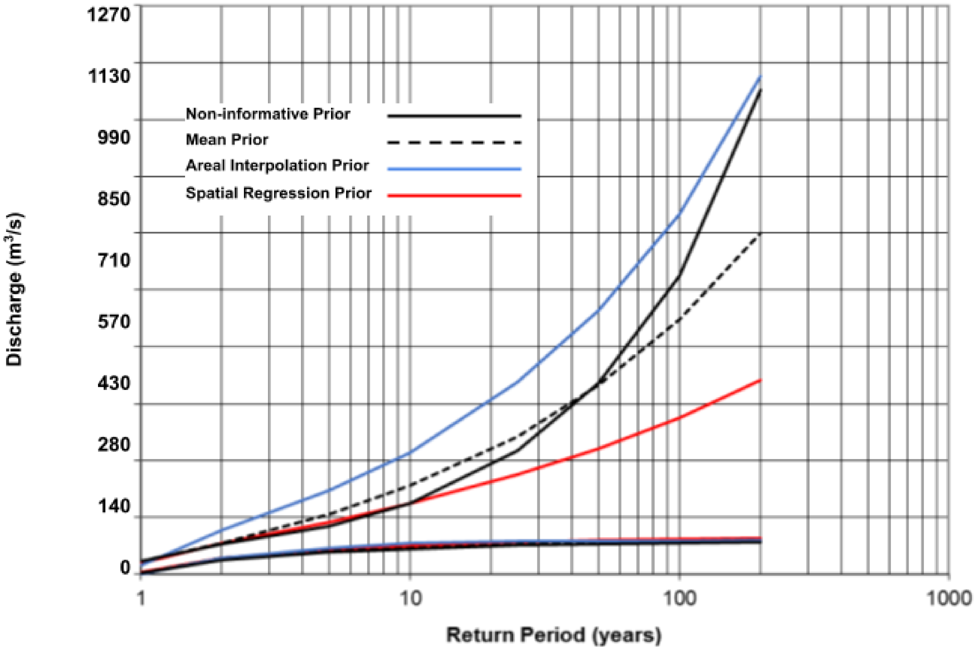

Figure 5. The 95% confidence interval for four scenarios: non-informative prior, mean prior, areal interpolation priors, and spatial regression prior.

5.2. *Effects of Length of Observations*

The results have shown that the BLP-SP algorithm can largely reduce the uncertainty of the flood frequency analysis based on 10-year observations. To further evaluate the improvement of the BLP-SP algorithm with other data lengths, we tested two more scenarios with 20 and 30 years of data length. Figure 6 shows that the 95% confidence interval generated using the last 30 years of systematic records without prior information is similar with the one generated using the entire 54-year records.

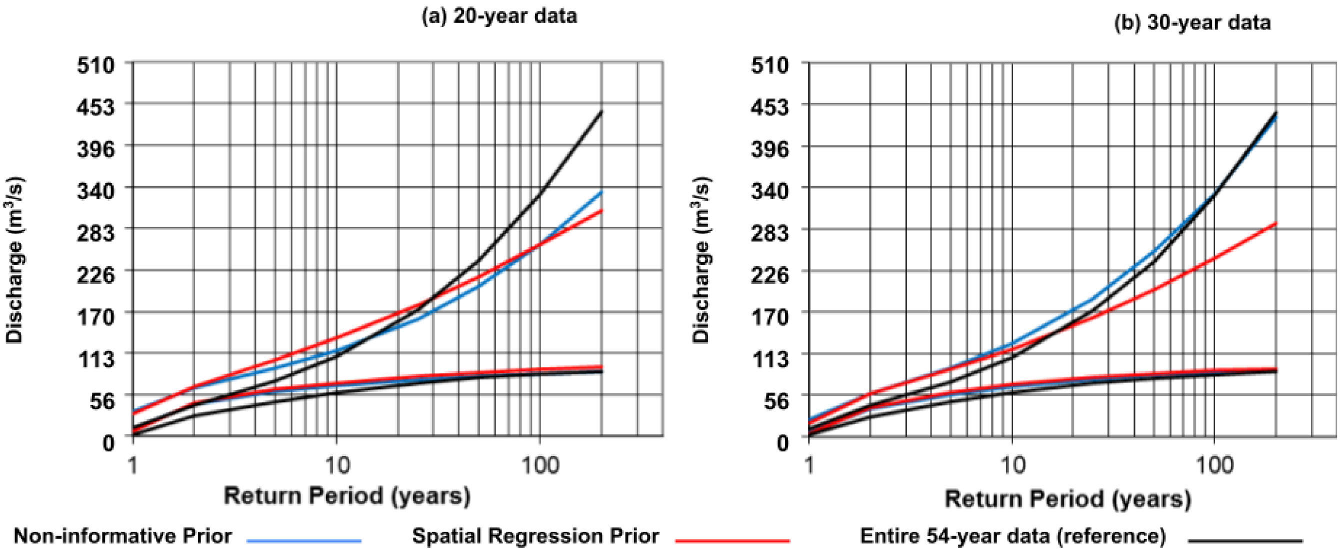

Figure 6. Applying spatial regression prior to (**a**) 20-year and (**b**) 30-year data series.

In Figure 6, both the 20-year model and the 30-year model have similar upper bounds. It indicates that with sufficient data length (e.g., 20 years), the spatial regression based model would produce consistent prediction regardless the data length. Even that the 30-year model with a non-informative prior can achieve a comparable prediction interval as the 54-year baseline model, introducing the spatial prior has increased it prediction accuracy.

The 20-year scenario shows an interesting outcome: the non-informative Bayesian inference has a smaller confidence interval than the model using the spatial prior. In fact, the small confidence interval of the non-informative model might be biased because theoretically longer data should produce smaller confidence interval, not the other way. One of the assumptions for flood frequency analyses is that the records are independent of each other. We suspect that there were strong temporal autocorrelation and seasonal trends in the records over the last 20 years introduced by multidecadal trends or wet and dry cycles [1]. Our model using the spatial prior in the Bayes inference has corrected the bias.

With the spatial regression prior information, the confidence interval of the 30-year result decreased significantly. For example, for the 100-year floods, the confidence interval decreased by 36.88%. In this way, by using both 20- and 30-year records, the results with the prior spatial regression knowledge are more realistic than those using information from the systematic data only. In other words, with the help of the spatial regression prior, the Bayesian estimation can generate a comparable prediction result as the baseline model even the data length is much shorter.

6. Conclusions

Our data analysis confirmed that with only 10 years of records, the flood prediction model would have a much larger uncertainty than the baseline model using 54 years of records. Therefore, we proposed the new model BLP3-SP that can incorporate the prior information from other nearby watersheds with long data series using a spatial regression model. With the spatial prior information, the BLP3-SP model can predict future floods with a similar mean and confidence interval as the baseline model. Specifically, the BLP3-SP model can reduce half of the uncertainty in the predicted discharge rate of a 100-year flood using only 10 years of records. In addition, spatial regression prior can reduce the bias caused by seasonal trends and generate a more accurate representation of the future flood probability.

We also evaluated three spatial models to generate the prior distributions: spatial regression, arithmetic mean, and areal interpolation. The spatial regression model outperformed the other two because the model considered both spatial contiguity and local environmental characteristics. The areal interpolation model did not work at all in our case study. This result indicates that the Log Pearson Type III distribution parameters have some spatial contiguity and are associated with local environmental characteristics.

Overall, BLP3-SP is a useful and robust algorithm for decreasing the uncertainty in the flood frequency estimation, especially for the sites with a short systematic data series. This method can be applied to areas with an uneven length of discharge gauge records to improve the accuracy of predicted flood quantiles.

Author Contributions: Conceptualization, D.T. and L.W.; methodology, D.T. and L.W.; software, D.T.; validation, D.T.; formal analysis, D.T.; writing—original draft preparation, D.T.; writing—review and editing, L.W.; visualization, D.T.; supervision, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study did not report any data.

Acknowledgments: The authors thank the anonymous reviewers for their constructive comments to improve the manuscript and the Open Access Author Fund from Louisiana State University Library.

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

References

1. England, J.F.; Cohn, T.A.; Faber, B.A.; Stedinger, J.R.; Thomas, W.O.; Veilleux, A.G.; Kiang, J.E.; Mason, R.R. *Guidelines for Determining Flood Flow Frequency—Bulletin 17C*; No. 4-B5; US Geological Survey: Reston, VA, USA, 2019.
2. Reis, D.S.; Stedinger, J.R. Bayesian MCMC Flood Frequency Analysis with Historical Information. *J. Hydrol.* **2005**, *313*, 97–116. [CrossRef]
3. Stedinger, J.R.; Lu, L.-H. Appraisal of Regional and Index Flood Quantile Estimators. *Stoch. Hydrol. Hydraul.* **1995**, *9*, 49–75. [CrossRef]
4. Hosking, J.R.M.; Wallis, J.R. *Regional Frequency Analysis: An Approach Based on L-Moments*; Cambridge University Press: Cambridge, UK, 2005; ISBN 9780521019408.
5. Kimber, A. National Research Council Estimating Probabilities of Extreme Floods. *J. Am. Stat. Assoc.* **1989**, *84*, 627. [CrossRef]
6. Vicens, G.J.; Rodriguez-Iturbe, I.; Schaake, J.C. A Bayesian Framework for the Use of Regional Information in Hydrology. *Water Resour. Res.* **1975**, *11*, 405–414. [CrossRef]
7. Wood, E.F.; Rodríguez-Iturbe, I. Bayesian Inference and Decision Making for Extreme Hydrologic Events. *Water Resour. Res.* **1975**, *11*, 533–542. [CrossRef]
8. Stedinger, J.R. Design Events with Specified Flood Risk. *Water Resour. Res.* **1983**, *19*, 511–522. [CrossRef]
9. Parent, E.; Bernier, J. Bayesian POT Modeling for Historical Data. *J. Hydrol.* **2003**, *274*, 95–108. [CrossRef]
10. Coles, S.; Pericchi, L. Anticipating Catastrophes through Extreme Value Modelling. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2003**, *52*, 405–416. [CrossRef]
11. Coles, S.; Pericchi, L.R.; Sisson, S. A Fully Probabilistic Approach to Extreme Rainfall Modeling. *J. Hydrol.* **2003**, *273*, 35–50. [CrossRef]
12. Kuczera, G. Combining Site-Specific and Regional Information: An Empirical Bayes Approach. *Water Resour. Res.* **1982**, *18*, 306–314. [CrossRef]
13. Madsen, H.; Rosbjerg, D. Generalized Least Squares and Empirical Bayes Estimation in Regional Partial Duration Series Index-Flood Modeling. *Water Resour. Res.* **1997**, *33*, 771–781. [CrossRef]
14. Seidou, O.; Ouarda, T.B.M.; Barbet, M.; Bruneau, P.; Bobée, B. A Parametric Bayesian Combination of Local and Regional Information in Flood Frequency Analysis. *Water Resour. Res.* **2006**, *42*, 11. [CrossRef]
15. Micevski, T.; Kuczera, G. Combining Site and Regional Flood Information Using a Bayesian Monte Carlo Approach. *Water Resour. Res.* **2009**, *45*, 4. [CrossRef]
16. Gaume, E.; Gaál, L.; Viglione, A.; Szolgay, J.; Kohnová, S.; Blöschl, G. Bayesian MCMC Approach to Regional Flood Frequency Analyses Involving Extraordinary Flood Events at Ungauged Sites. *J. Hydrol.* **2010**, *394*, 101–117. [CrossRef]
17. Merz, R.; Blöschl, G.; Humer, G. National Flood Discharge Mapping in Austria. *Nat. Hazards* **2008**, *46*, 53–72. [CrossRef]
18. Viglione, A.; Merz, R.; Salinas, J.L.; Blöschl, G. Flood Frequency Hydrology: 3. A Bayesian Analysis. *Water Resour. Res.* **2013**, *49*, 675–692. [CrossRef]
19. Skøien, J.O.; Merz, R.; Blöschl, G. Top-Kriging-Geostatistics on Stream Networks. *Hydrol. Earth Syst. Sci.* **2006**, *10*, 277–287. [CrossRef]

20. Nguyen, C.C.; Gaume, E.; Payrastre, O. Regional Flood Frequency Analyses Involving Extraordinary Flood Events at Ungauged Sites: Further Developments and Validations. *J. Hydrol.* **2014**, *508*, 385–396. [CrossRef]
21. Lima, C.H.R.; Lall, U.; Troy, T.; Devineni, N. A Hierarchical Bayesian GEV Model for Improving Local and Regional Flood Quantile Estimates. *J. Hydrol.* **2016**, *541*, 816–823. [CrossRef]
22. Merz, R.; Blöschl, G. Flood Frequency Regionalisation—spatial Proximity vs. Catchment Attributes. *J. Hydrol.* **2005**, *302*, 283–306. [CrossRef]
23. Bobée, B.; Ashkar, F. *The Gamma Family and Derived Distributions Applied in Hydrology*; Water Resouces Publications: Littleton, CO, USA, 1991; ISBN 9780918334688.
24. Vogel, R.W.; McMartin, D.E. Probability Plot Goodness-of-Fit and Skewness Estimation Procedures for the Pearson Type 3 Distribution. *Water Resour. Res.* **1991**, *27*, 3149–3158. [CrossRef]
25. Kirby, W. Computer-Oriented Wilson-Hilferty Transformation That Preserves the First Three Moments and the Lower Bound of the Pearson Type 3 Distribution. *Water Resour. Res.* **1972**, *8*, 1251–1254. [CrossRef]
26. Mehmood, A.; Jia, S.; Mahmood, R.; Yan, J.; Ahsan, M. Non-Stationary Bayesian Modeling of Annual Maximum Floods in a Changing Environment and Implications for Flood Management in the Kabul River Basin, Pakistan. *Water* **2019**, *11*, 1246. [CrossRef]
27. Ahn, K.-H.; Palmer, R. Regional Flood Frequency Analysis Using Spatial Proximity and Basin Characteristics: Quantile Regression vs. Parameter Regression Technique. *J. Hydrol.* **2016**, *540*, 515–526. [CrossRef]
28. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2009; ISBN 9780470317716.
29. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*; Springer Science & Business Media: Berlin, Germany, 2013; ISBN 9781475730715.
30. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2013; ISBN 9781439840955.
31. Geman, S.; Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721–741. [CrossRef]
32. Casella, G.; George, E.I. Explaining the Gibbs Sampler. *Am. Stat.* **1992**, *46*, 167.
33. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
34. Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]
35. Stedinger, J.R.; Tasker, G.D. Regional Hydrologic Analysis, 2, Model-Error Estimators, Estimation of Sigma and Log-Pearson Type 3 Distributions. *Water Resour. Res.* **1986**, *22*, 1487–1499. [CrossRef]
36. Langbein, W.B. Annual Floods and the Partial-Duration Flood Series. *Trans. Am. Geophys. Union* **1949**, *30*, 879. [CrossRef]
37. Parkes, B.; Demeritt, D. Defining the Hundred Year Flood: A Bayesian Approach for Using Historic Data to Reduce Uncertainty in Flood Frequency Estimates. *J. Hydrol.* **2016**, *540*, 1189–1208. [CrossRef]
38. Gotvald, A.J.; Barth, N.A.; Veilleux, A.G.; Parrett, C. *Methods for Determining Magnitude and Frequency of Floods in California, Based on Data through Water Year 2006*; Scientific Investigations Report; US Geological Survey: Reston, VA, USA, 2012.