atmosphere

MDPI

Article

Bayesian Analysis of Spatial Model for Frequency of Tornadoes

Haitao Zheng [1], Yi Zhang [1], Qiaoju Chen [1], Qingshan Yang [2], Guoqing Huang [2,*], Dahai Wang [3] and Ruili Liu [2]

1 Department of Statistics, Southwest Jiaotong University, Chengdu 610031, China
2 School of Civil Engineering, Chongqing University, Chongqing 400044, China
3 School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430062, China
* Correspondence: ghuang1001@gmail.com

Abstract: Frequency analysis of tornadoes is very important for risk analysis and disaster control. In this paper, the annual frequency of tornadoes that occurred in the United States from 1967 to 2016 is analyzed. The simple analysis shows that frequencies of tornadoes of different sites are spatially correlated and over-dispersed. To explain the two characteristics of the data, the Bayesian hierarchical model is proposed. For comparison purposes, the Bayesian model with negative binomial distribution, Poisson distribution, Polya distribution, and first-order, non-negative, integer-valued autoregressive model with Bell innovations(BL-INAR(1)) are considered to fit the frequency of tornado. The distribution parameters of all sites are assumed to be spatially correlated, and the corresponding Bayesian hierarchical models were established. MCMC (Markov Chain Monte Carlo) method is applied to parameter estimations and relative statistical inference. By comparison of the analysis results, the negative binomial distribution is recommended to analyze the overdispersion and spatial correlation among the sites of the data. The comparison between the simulated frequencies based on the proposed model and the actual frequencies also verifies that the proposed method is a better model for the data.

Keywords: negative binomial distribution; Poisson distribution; spatial correlation; Bayesian hierarchical model; MCMC

check for updates

Citation: Zheng, H.; Zhang, Y.; Chen, Q.; Yang, Q.; Huang, G.; Wang, D.; Liu, R. Bayesian Analysis of Spatial Model for Frequency of Tornadoes. *Atmosphere* **2023**, *14*, 472. https://doi.org/10.3390/atmos14030472

Academic Editor: Jimy Dudhia

Received: 18 November 2022
Revised: 3 February 2023
Accepted: 17 February 2023
Published: 27 February 2023

1. Introduction

A tornado is a small-scale vortex flow with strong destructive force. The United States is the country with the most tornadoes in the world. According to the records of the National Oceanic and Atmospheric Administration (NOAA) (spc2016) [1], a total of 62208 tornadoes occurred from 1950 to 2016. Almost every year, tornadoes cause damage to property and the economy and also result in deaths and injuries. Tornadoes are a major concern in the United States. In order to estimate tornado damages, risk analysis is needed [2].

The risk analysis of tornadoes requires building an appropriate model. In general, the characteristics of tornadoes can be described by physical parameters such as path length and width, moving direction, moving speed, and wind speed model, as well as statistical parameters such as annual frequency and seasonal characteristics. Banik [3] used the frequency, intensity, path length, and width of tornadoes to establish a tornado risk assessment model in southern Ontario, Canada. Shen [4] and Coleman [5] established a tornado risk assessment model in the United States by using the frequency of tornadoes. Tamura [6] established a tornado risk analysis model for Japanese tornado data according to the frequency, length, width, and moving direction of tornadoes. In the above risk analysis, frequency analysis is an important part of tornado analysis. Generally, meteorological variables such as tornadoes have many uncertainties. The Bayesian hierarchical model is often used in the study of the probability model of meteorological variables [7]. For example, Cheng [8] established the probability model by using a Bayesian hierarchical model for the frequency of tornadoes in Canada. Sang [9] also adopted the spatial model for building extreme temperature values with space-time attributes. Potvin [10] used Bayesian hierarchical modeling to estimate tornado reporting rates and expected tornado counts in the central United States from 1975 to 2016. To improve estimates of tornado frequency in the central United States, Potvin [11] developed a sophisticated statistical model that accounts for these population-dependent tornadoes. Moore [12] analyzed the spatial distribution characteristics of tornadoes in the United States; they [13] looked at the relationship between tornado activity in the United States and the El Nio/Southern Oscillation in all four seasons and across multiple regions. Coffer [14] used Random Forest classification to predict tornadoes. Allen [15] employed Kernel Density Estimation for spatial pattern analysis and space-time cubes to visualize the spatiotemporal frequency of tornadoes and potential trends. Gensini [16] analyzed the spatial trends in the United States' tornado frequency. They focused on an environmental covariate approach to examine potential changes in United States tornado frequency. Cao [17] developed a statistical framework to quantitatively explain two-way interconnections between long-term climate trends and internal variabilities. Some of the studies might consider the spatial correlation of tornadoes [18–21] but have not considered the characteristics of dispersion. The aforementioned literature used Poisson distribution to fit the data but did not check if it was appropriate for the data. Recently, Huang [22] has proposed a new first-order, non-negative, integer-valued autoregressive model with Bell innovations(BL-INAR(1)) to analyze the count data, and the model is suitable for counts exhibiting overdispersion. The data is non-negative and integer-valued Poisson distribution is usually used for count data analysis and sometimes generalized linear model since the data is a non-negative integer-value. The count data with repeated measures may have a dispersion issue, which is often ignored in engineering practice.

The simple analysis of the annual frequency of tornadoes in various regions of the United States from 1967 to 2016 shows that the spatial correlation and overdispersion phenomenon exists. To account for those phenomena, the Bayesian method is applied to analyze the annual frequency of tornadoes with three distributions, i.e., Poisson distribution, negative binomial distribution and Polya distribution. The parameters of each distribution are assumed to vary with the grids to account for the inhomogeneity in the dataset. Spatial correlation is assumed by the a priori distribution of spatial model parameters. MCMC (Markov Chain Monte Carlo) method is used for posterior parameter estimation. Finally, the simulation analysis is carried out by using the established spatial model. The numerical characteristics and spatial correlation of the simulated tornado frequency are compared with the actual data to show the effectiveness of the proposed method.

We organize the rest of this article as follows. In Section 2, we briefly introduce the Bayesian hierarchical model and the three discrete distributions, including its definition and some properties. Section 3 introduces the method of data processing and analysis, and explains why we adopt the method in this paper. In Section 4, we use the MCMC algorithm to estimate the parameters of the model and then compare the results with other models to show the superior performance of the proposed model, and the simulation analysis also shows the practicability of the model. The paper concludes in Section 5.

2. Methodology

Here we attempt to establish a probability model of the annual frequency of tornadoes by using the Bayesian inference technique. Firstly, the annual frequency of tornadoes in each region is assumed to follow a discrete distribution, and the discrete distribution is decomposed into a series of conditional probability models to obtain a Bayesian hierarchical model. X, Y, and Z are random variables and their joint probability density function can be expressed as $P(X,Y,Z) = P(Z|X,Y)P(X|Y)P(Y)$.

Hypothesis: D, observation data set; Θ, parameter set; Φ, hyperparameter set.

The Bayesian hierarchical model simplifies complex problems into three main parts:

Sampling distribution: $P(D|\Theta, \Phi)$

Priors for Θ: $P(\Theta|\Phi)$

Priors for Φ: $P(\Phi)$

The first part is mainly concerned with the sampling distribution function or likelihood function of the observed values under the condition that the relevant parameters of the observed data are given. The second part is concerned with the prior distribution, which is the probability distribution of the parameters of the data model in part one under the condition of hyperparameters. The last part is the prior distribution of the hyperparameters, also known as the super prior distribution. In application, these parts can be subdivided into some sub-stages to construct a multi-stage hierarchical model, which makes the model more flexible.

The three parts are introduced to study what distribution the process and parameters of interest follow given the observed data, that is, the posterior of the parameters and processes. The expression is given by Bayes' rule:

$$P(\Theta, \Phi|D) \propto P(D|\Theta, \Phi)P(\Theta|\Phi)P(\Phi) \tag{1}$$

In Bayesian inference, assuming that the parameters of the model are not fixed values, but random variables, the appropriate distribution of parameters can be assigned according to experience and prior knowledge, that is, super prior distribution. For the selection of prior distribution, the conjugate distribution of the likelihood function is generally selected to facilitate iterative operation, or without prior information, the uninformative prior is directly selected.

In this paper, under the condition that the annual frequency data of tornadoes in each region are known, the distribution functions of frequency and parameters are assumed to construct a Bayesian hierarchical model. Finally, the posterior parameters of the annual frequency distribution are estimated by MCMC sampling.

2.1. Sampling Distributions

The data set $\mathbf{X}_{s\times t} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_t) = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,t} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,t} \\ & \vdots & & \\ x_{s,1} & x_{s,2} & \cdots & x_{s,t} \end{pmatrix}$ is given, where $x_{k,i}$, $1 \le k \le s, 1 \le i \le t$ is the frequency of tornadoes in the k-th area in the i-th year. Here, it is assumed that the distribution of tornado frequency in different years is the same in the same year.

(1) Suppose that the annual frequency $\boldsymbol{X}_i = (x_{1,i},\ x_{2,i},\ \cdots\ x_{s,i})^T$ of tornadoes in each region in the i-th year follows a Poisson distribution under the condition of the expected occurrence frequency $\boldsymbol{\lambda} = (\lambda_1,\ \lambda_2, \cdots, \lambda_s)^T$:

The probability density function of the Poisson distribution is:

$$\boldsymbol{X_i}|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \boldsymbol{Poisson}(\boldsymbol{\lambda}) \tag{2}$$

The probability density function of the Poisson distribution is:

$$f(\boldsymbol{X_i}; \boldsymbol{\lambda}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{X}_i|\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\boldsymbol{\lambda}^{\boldsymbol{X_i}} \boldsymbol{e}^{-\boldsymbol{\lambda}}}{X_i!} \tag{3}$$

and the mean and variance of the Poisson distribution are equal.

(2) Suppose that $\boldsymbol{X}_i = (x_{1,i},\ x_{2,i},\ \cdots\ x_{s,i})^T$ follows a negative binomial distribution with parameters $\boldsymbol{r} = (r_1, r_2, \cdots, r_s)^T$ and $\boldsymbol{p} = (p_1, p_2, \cdots, p_s)^T$:

$$\boldsymbol{X_i}|\boldsymbol{r}, \boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim NB(\boldsymbol{r}, \boldsymbol{p}) \tag{4}$$

The probability density function of negative binomial distribution, expectation, and variance are respectively:

$$f(\boldsymbol{X_i}; \boldsymbol{r}, \boldsymbol{p}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{X_i}|\boldsymbol{r}, \boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \binom{\boldsymbol{X_i} + \boldsymbol{r} - \boldsymbol{1}}{\boldsymbol{X_i}} \boldsymbol{p}^{\boldsymbol{k_i}} (1 - \boldsymbol{p})^{\boldsymbol{r}} \tag{5}$$

$$E(\boldsymbol{X_i}|\boldsymbol{r}, \boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\boldsymbol{r}(1 - \boldsymbol{p})}{\boldsymbol{p}} \tag{6}$$

$$Var(\boldsymbol{X_i}|\boldsymbol{r}, \boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\boldsymbol{r}(1 - \boldsymbol{p})}{\boldsymbol{p}^2} \tag{7}$$

Equations (6) and (7) show that the mean value of the negative binomial distribution is smaller than the variance.

(3) Suppose that $\boldsymbol{X_i} = (x_{1,i},\ x_{2,i},\ \cdots\ x_{s,i})^T$ follows a Polya distribution with parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_s)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_s)^T$:

$$\boldsymbol{X_i}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim Polya(\boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{8}$$

The probability density function of polya [20] distribution, expectation, and variance, respectively:

$$f(\boldsymbol{X_i}; \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{X_i}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma(\boldsymbol{k_i} + \boldsymbol{\alpha})\boldsymbol{\beta}^{\boldsymbol{k_i}}}{\Gamma(\boldsymbol{k_i} + 1)\boldsymbol{\Gamma}(\boldsymbol{\alpha})(1 + \boldsymbol{\beta})^{\boldsymbol{\alpha} + \boldsymbol{k_i}}} \tag{9}$$

$$E(\boldsymbol{X_i}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\alpha\beta} \tag{10}$$

$$Var(\boldsymbol{X_i}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\alpha\beta}(1 + \boldsymbol{\beta}) \tag{11}$$

2.2. Priors of Parameters

(1) Suppose that the parameter $\boldsymbol{\lambda}$ of the Poisson distribution follows the lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\boldsymbol{\lambda}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{12}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_s)^T$. Covariance matrix $\boldsymbol{\Sigma}$ is a function of spatial distance, which can be expressed as:

$$\boldsymbol{\Sigma} \propto e^{-\frac{d}{\rho}} \tag{13}$$

The off-diagonal elements of the covariance matrix are not zero, indicating that there is the correlation between the elements of a random variable $\boldsymbol{\lambda}$. Where d is the spatial distance matrix, taking the Euclidean distance, we have $d_{m,n} = \sqrt{(s_{m1} - s_{n1})^2 + (s_{m2} - s_{n2})^2}$, where s_{m1}, s_{n1} represents the latitude of m and n regions, and s_{m2}, s_{n2} represents the corresponding longitude.

(2) The value of parameter $\boldsymbol{r}$ of the negative binomial distribution is non-negative. Assume that uninformative prior is a lognormal distribution, that is:

$$\boldsymbol{r} \sim LN(0, 1) \tag{14}$$

The value of parameter $\boldsymbol{p}$ ranges from 0 to 1. In order to consider the spatial correlation, $\frac{\boldsymbol{p}}{\boldsymbol{1-p}}$ is assumed to follow a lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e.,

$$\frac{\boldsymbol{p}}{\boldsymbol{1 - p}}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{15}$$

The assumption of covariance matrix $\boldsymbol{\Sigma}$ is the same as that of Poisson distribution.

(3) Suppose that the parameter $\boldsymbol{\beta}$ of the Polya distribution follows an uninformative prior of the uniform distribution,

$$\boldsymbol{\beta} \sim U(a,b) \tag{16}$$

where a and b are non-negative integers.

Considering the spatial correlation, assume that the parameter $\boldsymbol{\alpha}$ follows a lognormal distribution with hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{\alpha}|\boldsymbol{\mu},\boldsymbol{\Sigma} \sim LN(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{17}$$

The assumption of covariance matrix $\boldsymbol{\Sigma}$ is the same as that of Poisson distribution.

2.3. Priors of Hyperparameters

Assume that the elements $\mu_j (1 \le j \le s)$ of the mean $\boldsymbol{\mu}$ of lognormal distribution are independent and identically distributed, and its prior distribution is a normal distribution

$$\mu_j \sim N\left(\mu_0, \sigma_0^2\right) \tag{18}$$

where μ_0 and σ_0 are constants.

Suppose that the hyperparameter ρ follows the uninformative prior distribution of $N(\mu_d, \sigma_d)$ in all three distributions, where $\mu_d = \overline{d} = \frac{\sum_{m=1}^{s}\sum_{n=1}^{s} d_{m,n}}{s^2}, \sigma_d = \frac{\overline{d}}{k}$, and k is a constant.

3. Data Source and Data Pre-Analysis

3.1. Data Sources

The data comes from the National Oceanic and Atmospheric Administration (NOAA) [1]. The tornado frequency data of the United States from 1967 to 2016 is used in the analysis. The tornado frequency here refers to the number of recorded tornadoes, not the actual number of occurrences. The data also contains the time, latitude and longitude of the tornado occurrence.

In the data, the latitude and longitude of tornadoes in the United States ranged from $-$W81° to $-$W158° and N21.1° to N60°. The smaller the grid is, the more grids with 0 tornadoes will be covered by grids of equal size. However, areas that have never experienced tornadoes in 50 years are meaningless for practical analysis. Some simple analysis with different sizes of grids has been done. Too many grids with zero counts or grids with large counts will complicate the analysis. To balance the number of grids and the annual frequency of tornadoes in each grid, the longitude interval selected in this paper is 6.7°, and the latitude interval is 3°, resulting in a 15 × 15 grid, shown as a black frame line in Figure 1. The annual frequency of tornadoes in each grid is counted, and the data set $\mathbf{X}_{225\times50}$ is obtained. The subscript of the data set represents the number of grids and the number of years, respectively, and the grid coordinate $S(Latitude, Longitude)$ takes its geometric center coordinate. Excluding the grids whose annual average frequency is less than 1, the remaining 52 grids are shown in the white box in Figure 1, and the data set $\mathbf{X}_{52\times50}$ is obtained. The following analysis is based on this data set.

3.2. Descriptive Analysis

3.2.1. Spatial Correlation Analysis

Each row of the data set $\mathbf{X}_{s\times t}$ represents the annual tornado frequency of a grid, and the grid coordinates are $S(Latitude, Longitude)$. Sorting each grid according to latitude, and one row of the sorted dataset $\mathbf{X}_{s\times t}$ still represents the annual tornado frequency of a grid from 1967 to 2016. Calculate the correlation coefficient between rows, that is, the correlation coefficient of the annual frequency series of tornadoes in each grid, or the spatial correlation coefficient of tornadoes, and obtain the correlation coefficient matrix (a) and correlation coefficient histogram (b) in Figure 2. The obvious correlation is shown in Figure 2a. It can also be seen from Figure 2b that the correlation coefficient between some grids is more than 0.5, and the correlation coefficient of some grids is less than −0.2, indicating that there is a non-negligible spatial correlation between the annual frequency of tornadoes.

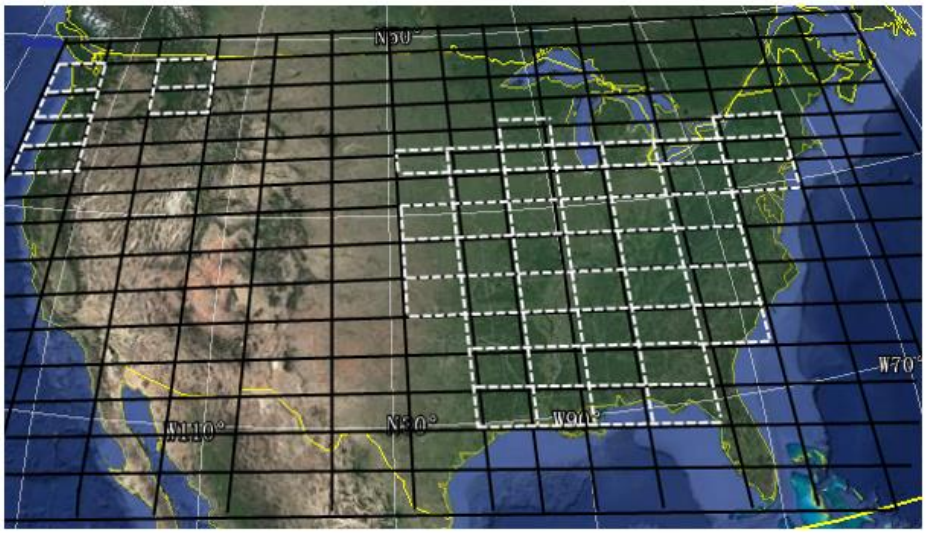
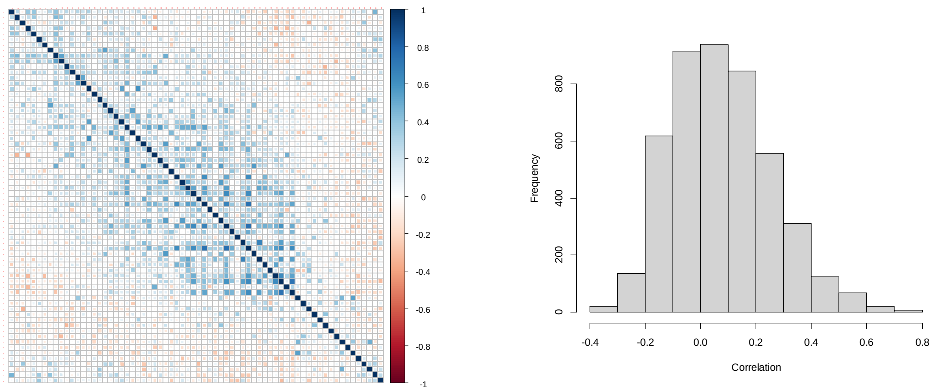


Figure 1. Grid division of tornadoes in the United States from 1967 to 2016.



(**a**) Correlation coefficient matrix (**b**) Histogram of annual tornado frequency

Figure 2. Correlation coefficient matrix and histogram of annual tornado frequency in different grids.

3.2.2. Divergence Analysis

Poisson distribution is commonly used for count data. It has an important property that the mean and variance are equal. If the sample variance of the count data and estimated variance based on the assumed distribution are not equal, it is called uneven dispersion, including two situations: the estimated variance is less than the sample variance, which is called over-dispersion, and vice versa, it is called under-dispersion. For Poisson distribution, the dispersion exists when the sample mean and sample variance are not equal. The data set $\mathbf{X}_{52\times50}$ is analyzed by the divergence. The mean and variance of each grid are shown in Figure 3. Table 1 shows the mean and variance data of six randomly selected grids. It can be seen that the variance in the given grid is significantly larger than the mean, which indicates that the data of $\mathbf{X}_{52\times50}$ has obvious over-discrete characteristics.

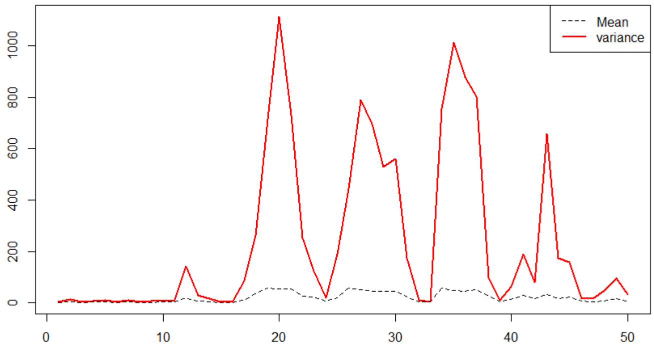


Figure 3. Mean and variance of tornados in every grid.

Table 1. Mean and variance of the partial grid of annual tornado frequency.

Grids \ Statistics	G_1	G_2	G_3	G_4	G_5	G_6
Mean	1.94	3.30	2.46	18.88	7.46	25.44
Var	4.75	13.81	7.06	143.53	30.21	99.23

For the data with overdispersion, Poisson distribution may lead to large errors in data fitting. Therefore, it is necessary to consider a more accurate distribution. For count data, some suitable distributions have been proposed, such as the Poisson distribution based on the Quasi Likelihood and random effect model [23–25]. For the tornado counting data in this paper, we consider the Poisson distribution, negative binomial distribution, and Polya distribution. Note that Poisson distribution is just a special case of negative binomial distribution. By randomizing the distribution parameters, we establish a Bayesian hierarchical model on the premise of considering the spatial correlation of parameters. Then we use the MCMC method to estimate the posterior parameters and compare them with the actual data to evaluate which distribution is more suitable and whether the proposed Bayesian hierarchical model is effective.

4. Analysis Results

4.1. Estimation Algorithm

After establishing the Bayesian hierarchical model, a series of parameters of posterior distribution can be sampled by the MCMC method. The MH (metropolis Hastings) algorithm in the MCMC method is adopted in this paper. The basic steps of this algorithm are as follows:

(1) Select the appropriate proposed distribution $g(.|X_t)$, and f is the objective function.

(2) Generate X_0 from the proposed distribution g as the initial value of the sampling sequence.

(3) Repeat the following steps until the sampling sequence converges to a stable state according to some criteria.

(3.1) Generate candidate values Y from $g(.|X_t)$.

(3.2) Generate a random number u from the uniform distribution $U(0,1)$.

(3.3) If $u < \frac{f(Y)g(X_t|Y)}{f(X_t)g(Y|X_t)}$, accept Y, $X_{t+1} = Y$, and the sampling sequence is updated in step $t+1$, otherwise $X_{t+1} = X_t$, the sampling sequence has not been updated in step $t+1$.

(3.4) Increase the value of t.

In the Bayesian hierarchical model in this paper, the likelihood function $P(D|\Theta,\Phi)$ is the target distribution, and the super prior distribution $P(\Theta|\Phi)P(\Phi)$ is the proposed distribution. On the premise of known observation data, the estimated values of parameters and super parameters are obtained through multiple iterations; that is, the estimated values of parameters λ_k, r_k, p_k, α_k, $\beta_k (1 \leq k \leq s)$.

4.2. Comparison between Simulation Frequency and Actual Frequency

According to the Bayesian hierarchical model established above, the posterior parameter estimates of three distributions $\hat{\lambda}_k, \hat{r}_k, \hat{p}_k, \hat{\alpha}_k, \hat{\beta}_k\ (1 \leq k \leq s)$ are obtained by the MH algorithm. The parameters of Polya's interpretation model in Section 2.2 are a = 0, b = 10; the values of μ_0 and σ_0 in the parameter model in Section 2.3 are $\mu_0 = 0.1$ and $\sigma_0 = 1.5$ in the Poisson distribution; In the negative binomial distribution, $\mu_0 = 0$, $\sigma_0 = 10$; In Polya distribution, $\mu_0 = 0$, $\sigma_0 = 1$; k= 100 in Poisson distribution and negative binomial distribution, k = 20 in Polya distribution. It can be seen from the data model in Section 2.1 that if the parameters of the probability distribution are known, the mean and variance of the distribution function can be estimated from the parameters. Therefore, after the posterior parameter estimates are obtained, the mean and variance corresponding to each grid distribution function are calculated and then compared with the mean and variance calculated from the actual data.

From dataset $\mathbf{X}_{52\times 50}$, eight grids were randomly selected. The mean and variance of the tornado frequency series of each grid are shown in Table 2. At the same time, the values of mean and variance obtained from the estimation of posterior parameters in the corresponding grids under several distributions and BL-INAR(1) model are given. The results are also shown in Table 2.

Table 2. The mean and variance of tornado frequency of partial grids under different distributions.

Girds	Statistic	Actual Value	Poisson Distribution	Negative Binomial Distribution	Polya Distribution	BL-INAR(1)
G_1	Mean	2.46	2.29	2.34	6.12	2.49
	Variance	7.07		8.05	12.86	4.76
G_2	mean	2.96	2.72	2.93	2.56	3.01
	Variance	11.18		12.55	4.87	4.88
G_3	Mean	3.30	3.00	3.31	4.06	3.29
	Variance	13.81		12.99	6.33	6.46
G_4	Mean	4.22	3.56	4.47	6.25	4.32
	Variance	16.66		13.75	11.04	8.66
G_5	Mean	11.68	10.48	12.92	5.62	11.60
	Variance	87.04		56.85	10.29	30.83
G_6	Mean	15.22	13.83	15.14	8.35	15.29
	Variance	174.62		297.51	14.22	36.86
G_7	Mean	21.36	19.40	24.80	11.83	21.66
	Variance	173.58		227.09	19.98	58.11
G_8	Mean	49.22	52.33	50.52	66.58	49.56
	Variance	1011.97		1502.40	182.07	155.77

The results in Table 2 show that under the three distribution assumptions and BL-INAR(1) model, the mean value calculated by the posterior parameters is relatively close to the actual mean value, except for Polya. It was found that the mean values of the negative binomial distribution and BL-INAR(1) model were close to the actual value. The mean of the Poisson distribution was not bad, but the Polya distribution was not suitable for the data. The variance range of the actual data in each grid was 1.24–1111.56, which fluctuated greatly. It was seen that the variance estimated by the posterior parameters of the negative binomial distribution was closer to the real variance, and better described the volatility and overdispersion of the data, only slightly overestimated. BL-INAR(1) model can closely estimate the mean values and explain part of the overdispersion of the data, but its variances were much smaller than the actual data. Note that the BL-INAR(1) model did not consider the spatial correlation among the sites. It means that the BL-INAR(1) model is a potential method if it can be modified by different distributions of spatial structure in the parameters. By comprehensive comparison analysis, the Bayesian hierarchical model, considering spatial correlation, was effective, and the model with negative binomial distribution was the best, which had more consistent results with the actual data.

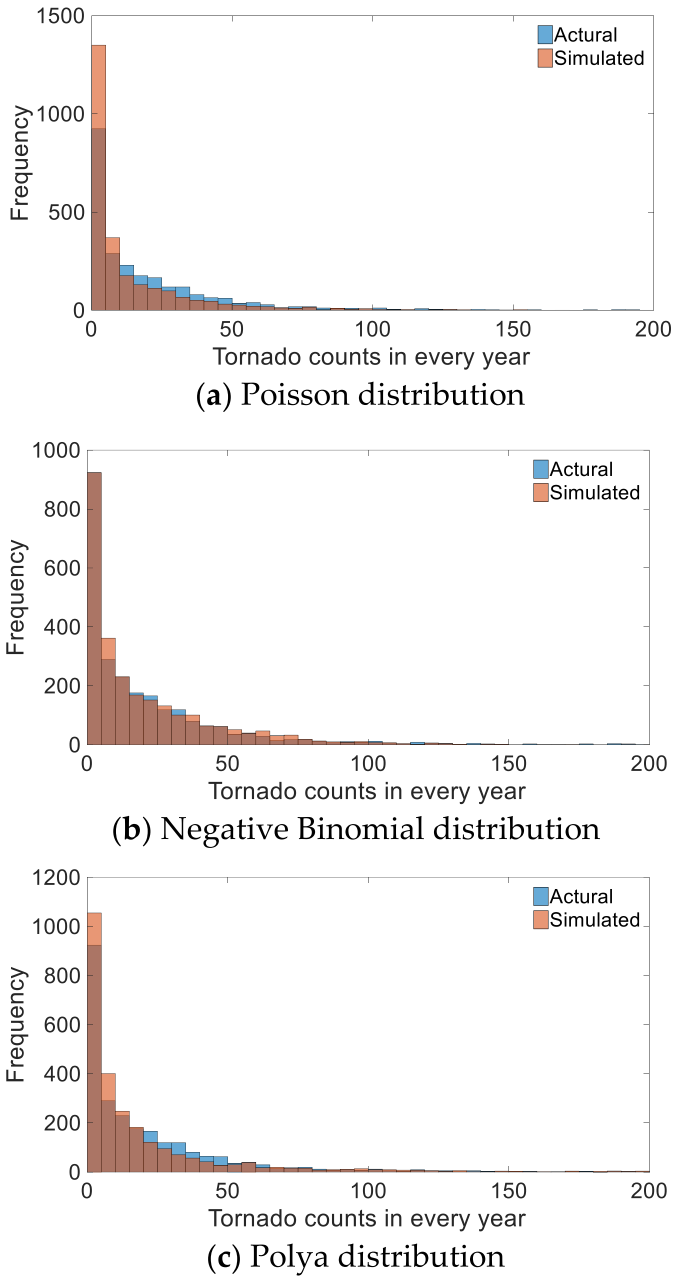
After obtaining the estimates of the posterior parameters of the three distributions, the established models are used to simulate the annual tornado frequency of each grid. We used a histogram to compare the distributions of simulated data and real data. The simulation study did not consider the BL-INAR(1) model since it can not simulate the spatial correlation among the sites. It can be seen from Figure 4 that the simulated data of Poisson distribution and negative binomial distribution are closer to the actual data. More details from Figure 4 are as follows:

(1) The zero frequency of Poisson distribution simulation data is close to that of the real data. The number of grids falling into the first interval of Poisson distribution is higher than that of the real data in the 9–12 interval, and slightly lower than that in the last few intervals.
(2) For the simulated data with a negative binomial distribution, the number of grids in the first two intervals is slightly higher than the actual value, the number of grids in the 5–15 intervals is slightly lower than the real number, and the number of grids in the subsequent intervals is higher than the real number. The simulated values of the negative binomial distribution are more dispersed and can better reflect the characteristics of the original numbers.
(3) The grid number of simulated data of the Polya distribution is concentrated in the first interval, and the grid number of subsequent intervals is basically lower than that of real data. The simulation value of tornado occurrence times of Polya distribution in each grid is obviously small, and its effect is the worst compared with the first two distributions.

It can be seen that the distribution of the simulated frequencies of tornadoes with a negative binomial model is quite close to the distribution of the real value. Such a result reflects the superiority of the proposed method.

4.3. Analysis of Posterior Parameter Estimates

The spatial correlation of the original data is established by the prior distribution of the parameters in the model. Therefore, the spatial correlation of posterior parameters should be similar to the spatial correlation of actual data. The sample of parameters of Poisson distribution, negative binomial distribution, and Polya distribution is obtained through MCMC, the spatial correlation coefficients of the stabilized parameters $\boldsymbol{\lambda}$, $\log\frac{\boldsymbol{p}}{1-\boldsymbol{p}}$ and $\boldsymbol{\alpha}$ are calculated, respectively. The spatial correlation coefficient matrix of the parameters adopts the same eigenvector angular ordering method as the spatial correlation coefficient matrix of the original data, and the correlation coefficient matrix shown in Figure 5 is obtained. The results show that the posterior parameters have an obvious spatial correlation. The correlation coefficient matrix of Poisson distribution parameters is shown in Figure 5a. It can be seen that the Poisson distribution also reflects the spatial correlation, but its characteristics are significantly different from those in Figure 2a. The correlation coefficient matrix of the negative binomial distribution parameters is shown in Figure 5b, and its correlation matrix is more similar to the correlation matrix of the original data (Figure 2). The correlation coefficient matrix (c) of Polya distribution parameters is similar to that of the Poisson distribution and does not really reflect the spatial correlation of the original data. The correlations of the posterior parameters of the three distributions are different since they are related to their distribution characteristics and parameter assumptions. In addition, the posterior parameters of different distributions retain an obvious correlation. Thus, it is feasible to set the superparameters of the model as a function of spatial distance so as to consider the spatial correlation of tornado occurrence frequency.

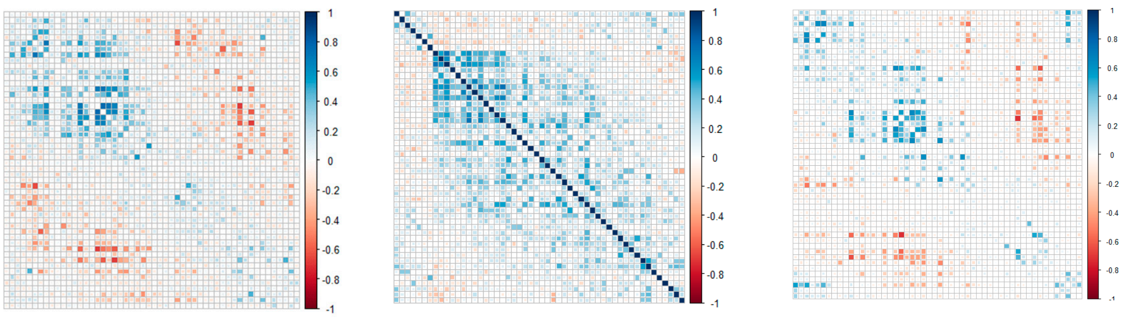


(**a**) Poisson distribution

(**b**) Negative Binomial distribution

(**c**) Polya distribution

Figure 4. Histogram of simulated and real values for three different distributions.



(**a**) Poisson distribution (**b**) Negative binomial distribution (**c**) Polya distribution

Figure 5. Correlation coefficient matrix of posterior parameters of three different distributions.

5. Conclusions

In this paper, three distributions, including Poisson distribution, negative binomial distribution, and Polya distribution, and BL-INAR(1) models are used to fit the annual tornado frequency data. Considering the spatial correlation of the data, probability models of the annual tornado frequency in each region is established based on the Bayesian hierarchical framework. The means and variances of the data were compared with those of the posterior analysis. The established models were also used to simulate the frequencies of tornadoes and the corresponding spatial relationship. The results are summarized as follows:

(1) The posterior analysis shows that estimated means and variances of negative binomial for each grid are closer to those of actual data and provide a better explanation of overdispersion shown in the data. These statistics are better than those by existing distributions, such as the Poisson distribution.
(2) The distributions of the simulated frequency based on a negative binomial is close to the distribution of the actual data, meaning the negative binomial model is more suitable for the data.
(3) The analysis of raw data reveals the spatial correlation of data. The proposed spatial correlation analysis based on negative binomial distribution is consistent with the results of actual data.

Overall, the proposed method with a negative binomial model can better describe the spatial correlation and overdispersion of tornado frequencies. That has a certain guiding significance for the disaster reduction and prevention of tornadoes. However, it should be noted that the occurrence of tornadoes may be related to time and many meteorological factors, such as temperature, air pressure, vertical wind shear, and even illumination and rainfall, etc. The observed data may also be affected by non-meteorological factors such as monitoring equipment, stations, period, and so on. Therefore, data collection of tornadoes with more variates is needed for more accurate analysis. Furthermore, the advanced models can be developed, such as BL-INAR(1) model with different distributions and spatially correlated parameters, semiparametric models with consideration of overdispersion and spatial structure, etc.

Author Contributions: H.Z.: Methodology, Validation, Writing—original draft; Y.Z.: Writing—review & editing; Q.C.: Writing—review & editing; Q.Y.: Writing—review & editing; G.H.: Conceptualization, Supervision, Writing—review & editing; D.W.: Formal analysis; R.L.: Formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are grateful for the financial support provided by the National Natu-ral Science Foundation of China (Grant No. 52178456), the Fundamental Research Funds for the Central Universities (SWJTU, 2682021ZTPY078) and 111 Project (Grant No. B18062).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. NOAA. Severe Weather Database Files (1950–2016). National Oceanic and Atomspheric Administration. Washington, USA. 2017. Available online: http://www.spc.noaa.gov/wcm/#data (accessed on 7 December 2017).
2. Changnon, S.A. Tornado losses in the United States. *Nat. Hazards Rev.* **2009**, *10*, 145–150. [CrossRef]
3. Banik, S.S.; Hong, H.P.; Kopp, G.A. Tornado hazard assessment for southern Ontario. *Can. J. Civ. Eng.* **2007**, *34*, 830–842. [CrossRef]
4. Shen, G.; Hwang, S.N. A spatial risk analysis of tornado-induced human injuries and fatalities in the USA. *Nat. Hazards* **2015**, *77*, 1223–1242. [CrossRef]
5. Coleman, T.A.; Dixon, P.G. An objective analysis of tornado risk in the United States. *Weather Forecast.* **2014**, *29*, 366–376. [CrossRef]

6. Tamura, Y.; Matsui, M.; Kawana, S. Characteristics and risk analysis of tornadoes in Japan. In Proceedings of the 14th international Conference on Wind Engineering, Porto Alegre, Brazil, 21–26 June 2015.
7. Wikle, C.K. Hierarchical models in environmental science. *Int. Stat. Rev.* **2003**, *71*, 181–199. [CrossRef]
8. Cheng, V.Y.S.; Arhonditsis, G.B.; Sills, D.M.L.; Auld, H.; Shephard, M.W.; Gough, W.A.; Klaassen, J. Probability of tornado occurrence across Canada. *J. Clim.* **2013**, *26*, 9415–9428. [CrossRef]
9. Sang, H.; Gelfand, A.E. Hierarchical modeling for extreme values observed over space and time. *Environ. Ecol. Stat.* **2009**, *16*, 407–426. [CrossRef]
10. Potvin, C.K.; Broyles, C.; Skinner, P.S.; Brooks, H.E.; Rasmussen, E. A Bayesian hierarchical modeling framework for correcting reporting bias in the US tornado database. *Weather Forecast.* **2019**, *34*, 15–30. [CrossRef]
11. Potvin, C.K.; Broyles, C.; Skinner, P.S.; Brooks, H.E. Improving estimates of US tornado frequency by accounting for unreported and underrated tornadoes. *J. Appl. Meteorol. Climatol.* **2022**, *61*, 909–930. [CrossRef]
12. Moore, T.W.; DeBoer, T.A. A review and analysis of possible changes to the climatology of tornadoes in the United States. *Prog. Phys. Geogr. Earth Environ.* **2019**, *43*, 365–390. [CrossRef]
13. Moore, T.W. Seasonal frequency and spatial distribution of tornadoes in the United States and their relationship to the El Niño/Southern Oscillation. *Ann. Am. Assoc. Geogr.* **2019**, *109*, 1033–1051. [CrossRef]
14. Coffer, B.; Kubacki, M.; Wen, Y.; Zhang, T.; Barajas, C.A.; Gobbert, M.L. Machine Learning with Feature Importance Analysis for Tornado Prediction from Environmental Sounding Data. *PAMM* **2021**, *20*, e202000112. [CrossRef]
15. Allen, M.J.; Allen, T.R.; Davis, C.; McLeod, G. Exploring Spatial Patterns of Virginia Tornadoes Using Kernel Density and Space-Time Cube Analysis (1960–2019). *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 310. [CrossRef]
16. Gensini, V.A.; Brooks, H.E. Spatial trends in United States tornado frequency. *npj Clim. Atmos. Sci.* **2018**, *1*, 38. [CrossRef]
17. Cao, Z.; Cai, H. Trend Analysis of US Tornado Activity Frequency. *Atmosphere* **2022**, *13*, 498. [CrossRef]
18. Fuentes, M.; Henry, J.; Reich, B. Nonparametric spatial models for extremes: Application to extreme temperature data. *Extremes* **2013**, *16*, 75–101. [CrossRef] [PubMed]
19. Twisdale, L.A.; Dunn, W.L. Probabilistic analysis of tornado wind risks. *J. Struct. Eng.* **1983**, *109*, 468–488. [CrossRef]
20. Ling, Q. Bayesian Spatial-Temporal Models for Areal Count Data. Ph.D. Thesis, Emory University, Atlanta, GA, USA, 2014.
21. Wang, X.; Chen, M.H.; Kuo, R.C.; Dey, D.K. Bayessian spatial-temporal modeling of ecological zero-inflated count data. *Stat. Sin.* **2005**, *25*, 189.
22. Huang, J.; Zhu, F. A new first-order integer-valued autoregressive model with Bell innovations. *Entropy* **2021**, *23*, 713. [CrossRef]
23. Payne, E.H.; Hardin, J.W.; Egede, L.E.; Ramakrishnan, V.; Selassie, A.; Gebregziabher, M. Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling. *Stat. Methods Med. Res.* **2015**, *26*, 1802–1823. [CrossRef]
24. Wedderburn, R.W.M. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **1974**, *61*, 439–447.
25. Bennetts, W.R.E. Analysis of frequency count data using the negative binomial distribution. *Ecology* **1996**, *77*, 2549–2557.