Population, Resources & Environmental Economics (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/pree.2023.040113
ISSN 2616-2253 Vol. 4 Num. 1

# *A Copula Model-Based Joint Probability Analysis of Losses in Torrential Rain and Flood Disasters*

**Chenchen Yang**[1,a,*]

[1]*Tianjin University of Commerce, Tianjin, 300134, China*
[a]*cqyangcc@163.com*
[*]*Corresponding author*

***Keywords:*** Torrential rain and flood disasters, direct economic losses, the death toll, copula function, joint probability distribution

***Abstract:*** This paper analyzes the data of torrential rain and flood disasters in China in recent years, and finds that there is an obvious correlation between the direct economic losses and the death toll caused by the disasters. Firstly, this paper selects six common distributions to fit the marginal distributions of the direct economic losses and the death toll, and determines their optimal types of marginal distributions using the K-S test and the AIC criterion, which are the normal distribution and the lognormal distribution respectively. Next, the optimal copula describing their correlation is the Gumbel copula using the minimum AIC and BIC criteria. Finally we model the joint probability distribution of direct economic losses and the death toll to analyze the joint probability. The study shows that direct economic losses and the death toll are positively correlated.

## 1. Introduction

China is located in the East Asian monsoon region and the spatial and temporal distribution of precipitation is uneven, which leads to frequent torrential rain in China. Due to the complexity of the terrain in China, it is easy to cause the flood disaster [1]. The annual flood season of torrential rain in China is concentrated from May to August. And torrential rain and flood disasters are mainly characterized by the following features: high intensity, wide range, long duration, and heavy losses [2]. Persistent heavy rainfall may lead to flooding and urban waterlogging, which can lead to disasters such as landslides and mudslides. These disasters can have serious adverse impacts on the safety of people's lives, the food security of farmers, and the ecological security of nature, and they can still affect economic and social development after a disaster. Torrential rain and flood disasters are one of the natural disasters that have the most serious socio-economic impacts on our country. From 2003 to 2019, the average annual direct economic loss caused by flood disasters in China is about 137,958 million RMB, and the average annual death toll is about 1,082. The affected areas are mostly in the southern region. Figure 1 shows the regional distribution of the frequency of major torrential rain and flood disasters from 2003 to 2019, and it can be seen that the frequency of disasters is higher in the southern and central region.

In this paper, the copula theory is applied to the correlation analysis of direct economic losses and deaths in torrential rain and flood disasters. Firstly, this paper uses different distributions to fit

the actual data, and the probability density function of the direct economic loss and the total number of deaths is obtained after the fitting goodness and accuracy tests. Then we create joint probability distributions of direct economic losses and the death toll based on Archimedean copula functions. The density plots of different Archimedean copula are compared with the binary histogram of the data and the values of AIC and BIC are calculated for each Archimedean copula function. The optimal copula function is selected as the joint probability model according to the AIC or BIC minimization criterion, and the Spearman and Kendall rank correlation coefficients are calculated for the optimal copula. The coefficients quantify the relationship between direct economic losses and the total death toll.

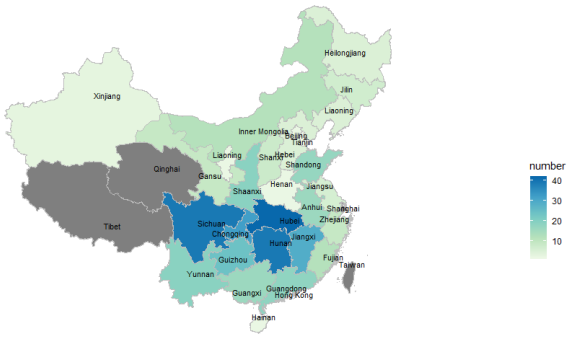


Figure 1: Distribution of major torrential rain and flood disasters

## 2. Data sources and processing

### 2.1. Data description

In this paper, the torrential rain and flood disasters events in China's mainland region from 2003 to 2019 are selected as research samples, and the data are obtained from the China Meteorological Disaster Yearbook published by the National Climate Center [3]. Specifically, this paper collects data related to 411 major torrential rain and flood disasters events from 2003 to 2019, including the time and location of each disaster event, the number of people affected, the total number of death toll (including the number of missing people), and the direct economic loss. Since this paper needs to consider the dependence of direct economic losses and the total death toll, we only consider torrential rain and flood disasters that cause one death at least. So the data used to build the model has 339 events. Among these disaster events, the lowest direct economic loss was 40 million yuan, the highest was 48.24 billion yuan. In one disaster event, there is up to 315 deaths.

### 2.2. Processing of data

In order to facilitate the comparison of direct economic losses in different years, which are caused by torrential rain and flood disasters. Therefore we need to revise the raw data to remove the effect of price level. The revised methodology is to calculate the inflation rate for each year relative to the base year and extrapolate the loss at the same consumer price level [4]. In this paper, we set 2019 prices as the base period prices. First, from 2003 to 2019, the variable-price GDP in each year is divided by constant-price GDP to obtain the inflation rate for each year, where the GDP data is obtained from the RESSET database. Second, the direct economic loss for all years is the inflation rate for each year multiplied by the direct economic loss.

Since the value of the adjusted direct economic loss still fluctuates quite a bit, we do logarization

on it [5]. It is to take the logarithm of the direct economic loss, at the same time this processing will not change the nature of the original data. The kurtosis of logarithmic direct economic losses caused by torrential rain and flood disasters is 3.06589, which is larger than 3, indicating that the loss data is a spike pattern. It has a skewness of -0.01020656, which is smaller than 0. Its data distribution is left skewed and trailing on the left. The total death toll has a kurtosis of 34.77891, which is larger than 3, indicating that the death data is also in a spiky pattern and is more spiky than the logarithmic direct economic losses. It has a skewness of 4.783579, which is larger than 0, and its data distribution is right skewed and trailing off to the right. In summary, the kurtosis of the loss and the death toll are both greater than 3 and both skewed data, indicating that they are both sharp peaks with thick tails.

## 3. Copula function and research methodology

### 3.1. Correlation Analysis

The correlation of different variables can be measured and assessed using a number of different metrics that can help us understand and quantify the relationship of variables. Studying such relationships is helpful in data analysis and inference. Pearson linear correlation coefficient, Kendall rank correlation coefficient and Spearman rank correlation coefficient are commonly used correlation coefficients. Pearson correlation coefficient can only be used to describe the linear correlation between the variables, while the Kendall rank correlation coefficient and Spearman rank correlation coefficient are the indicators of the consistency of the degree of change between the variables. Therefore, Kendall rank correlation coefficient and Spearman rank correlation coefficient are selected to study the relationship between direct economic loss and the death toll. Their expressions are shown in Table 1.

Table 1: The formula for the correlation coefficient

| Type of correlation coefficient | Formula |
|---|---|
| Pearson | $r_p(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2}$ |
| Kendall | $r_K(X_1, X_2) = P[(X_1 - X_1')(X_2 - X_2') > 0] - P[(X_1 - X_1')(X_2 - X_2') < 0]$ |
| Spearman | $r_s(X_1, X_2) = r_P(F_1(X_1), F_2(X_2))$ |

where $Cov(X_1, X_2)$ denotes the covariance of $X_1$ and $X_2$, and $\sigma_1$ and $\sigma_2$ are the standard deviations of $X_1$ and $X_2$ respectively. $(X_1, X_2)$ and $(X_1', X_2')$ are independently and identically distributed. $F_i$ is the variable consisting of the rankings of the continuous random variable $X_i$ $(i = 1, 2)$.

### 3.2. Copula function

According to Sklar theory, copula is a connection function that connects the probability distributions of multiple random variables and can be used to characterize the dependence between multiple dependent variables [6]. The cumulative distribution function (CDF) of the random variables $(X_1, X_2, \cdots, X_n)$ is $F(x_1, x_2, \cdots, x_n) = C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n))$, where $C$ denotes the copula function and $F_i$ denotes the marginal distribution function of $X_i$.

The advantage of the Copula function is that the marginal distribution and the correlation structure can be treated separately. A high-dimensional joint probability distribution is generated by connecting the marginal distributions of each univariate. Common copula functions include Gaussian copula, Archimedean copula, and t copula. In this paper, we select the Archimedean copula, which is often used in several fields. We use this type of copula to characterize the dependence of direct economic losses and death toll of torrential rain and flood disasters. Specifically, we select three copula functions in the Archimedean copula family: the Frank copula, the Gumbel copula, and the Clayton copula, whose expressions are shown in Table 2. Their distributional characteristics are different: the Frank copula presents symmetry in the upper and lower tails, while the other copulas present the opposite characteristics. The Gumbel copula is more sensitive to the thick-tailed properties of the upper tail and is more suitable for constructing joint probabilities where the thick-tailed features of the upper tail are distinct and the thick-tailed features of the lower tail are not. The thick-tailed properties of the Clayton copula are the opposite of the Gumbel copula [7].

Table 2: Formula for the two-dimensional Archimedean copula function

| Type of Copula Function | Formula |
|---|---|
| Frank copula | $C_\alpha(u_1,u_2)=-\frac{1}{\alpha}\ln\left(1+\frac{\left(e^{-\alpha u_1}-1\right)\left(e^{-\alpha u_2}-1\right)}{\left(e^{-\alpha}-1\right)}\right),\alpha\neq 0$ |
| Gumbel copula | $C_\alpha(u_1,u_2)=\exp\left\{-\left[\left(-\ln u_1\right)^\alpha+\left(-\ln u_2\right)^\alpha\right]^{\frac{1}{\alpha}}\right\},\alpha..1$ |
| Clayton copula | $C_\alpha(u_1,u_2)=\left(u_1^{-\alpha}+u_2^{-\alpha}-1\right)^{-\frac{1}{\alpha}},\alpha..0$ |

where $u_i=F_{Xi}(X_i),i=1,2$ is the marginal distribution function and $\alpha$ is the copula parameter.

### 3.3. Research methodology

In this paper, the inference function for margins (IFM) is used to fit the parameters of the copula function [8]. The IFM is divided into two steps: in the first step, the parameters of the marginal probability distribution function are estimated by the great likelihood method; in the second step, the parameters of the copula function are estimated by the great likelihood method.

(1) The first step of IFM is to choose the optimal marginal probability distribution function. In this paper, the first step is to carry out the K-S (Kolmogorov-Smirnov) test with a confidence level of 5%. The original hypothesis of this test is that the samples come from a specific distribution, and the alternative hypothesis is that the samples do not come from a specific distribution. For the distributions that passed the test, we used the AIC and BIC criteria to further determine the optimal fitting distribution.

(2) The second step of the IFM is to select the optimal copula function, and the AIC and BIC criteria are also selected to determine the optimal copula function.

## 4. Analysis of results

### 4.1. Determination of the marginal distribution function

#### 4.1.1. Frequency histogram

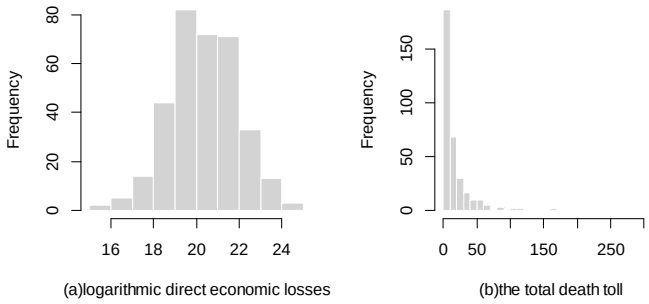


Figure 2: Frequency histograms of logarithmic direct economic losses and the total death toll

As shown in Figure 2(a), the frequency distribution of logarithmic direct economic losses has symmetrical characteristics near the center point, and the common symmetrical distributions are Normal, Cauchy and Logistic. Therefore, the above three distributions are used to fit the data of logarithmic direct economic loss. As seen in Figure 2(b), the distribution of the total death toll is characterized by asymmetry and right skewed. Therefore three asymmetric and heavy-tailed distributions which are Lognormal, Weibull and Generalized Extreme Value are used to fit the data of the total death toll.

#### 4.1.2. Tests of the effectiveness of fitting marginal distribution functions

Table 3: K-S test for logarithmic direct economic losses

| Type of distribution | Normal | Cauchy | Logistic |
|---|---|---|---|
| p-value | 0.8481 | 0.03565 | 0.5523 |

Table 4: K-S test for the total death toll

| Type of distribution | Lognormal | Pareto | Generalized Extreme Value |
|---|---|---|---|
| p-value | 0.1823 | 0.01696 | 0.05243 |

Based on the copula model, we fit the marginal probability distributions of logarithmic direct economic losses and total death toll, and the results are shown in Table 3 and Table 4. According to the K-S test, it is found that the p-value of both Normal and Logistic distributions are greater than 0.05 at a significance level of 5%. Therefore, the original hypothesis is accepted: logarithmic direct economic losses follow Normal and Logistic distributions. Next, we used them to fit logarithmic direct economic losses. Similarly, the p-values of both Lognormal and Generalized Extreme Value distributions are greater than 0.05. Therefore, the original hypothesis that the total death toll follows the Lognormal and Generalized Extreme Value distributions is accepted. Next, we used them to fit the total death toll. Finally, the AIC and BIC values of the fitted distributions of the two variables are calculated separately and their optimal marginal distributions are selected.

Table 5: Goodness of fit of marginal distribution of logarithmic direct economic losses

| | Normal | Logistic |
|---|---|---|
| AIC | 1278.626 | 1284.151 |
| BIC | 1286.278 | 1291.803 |

Table 6: Goodness of fit of marginal distribution of the total death toll

| | Lognormal | Generalized Extreme Value |
|---|---|---|
| AIC | 2577.363 | 2598.183 |
| BIC | 2585.015 | 2609.661 |

Based on the minimum AIC and BIC criteria, and combining with Table 5, it can be seen that for logarithmic direct economic losses, the optimal distribution is the Normal distribution, whose fitted parameter has a mean of 20.3438827 and a standard deviation of 1.58572248. From Table 6, it can be seen that for the total death toll, the optimal distribution is the Lognormal distribution, whose fitted meanlog is 2.14757514, sdlog is 1.25735062.

## 4.2. Determination of the Copula function

In this paper, the commonly used Archimedean copula function is chosen to portray the dependence structure of direct economic losses and the total death toll of torrential rain and flood disasters in China. Firstly, the unknown parameters of the copula model are estimated. Secondly, the AIC and BIC values of Clayton copula, Gumbel copula, and Frank copula are compared, and the copula corresponding to the minimum AIC and BIC is selected as the optimal copula fitting the joint probability distribution. Finally we calculate the rank correlation coefficient of the optimal copula [9].

### 4.2.1. Estimation of parameters

The linear correlation parameters of the three copula function models were estimated through R [10]. The results are as follows shown in Table 7:

Table 7: Linear correlation parameters of Copula

| | Frank copula | Gumbel copula | Clayton copula |
|---|---|---|---|
| Linear correlation parameter | 2.843 | 1.388 | 0.9038 |

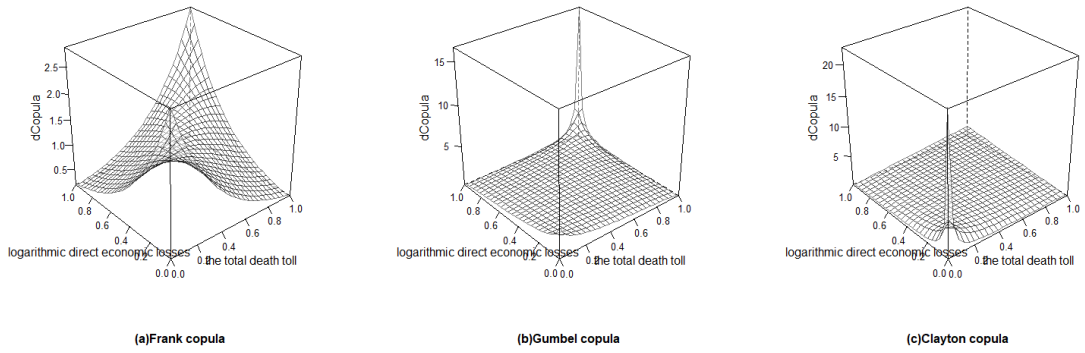


Figure 3: Plot of fitted probability density function (PDF)

According to the estimated values of the linear correlation parameters of the three copula functions and Figure 3, it can be found that they all have good linear correlation.

### 4.2.2. Selection of the optimal copula

① Test and Evaluation of Fitting Effect

In order to test the goodness of model fitting of the above three copula functions, this paper first introduces binary histograms of the data to initially estimate the model fitting effect, and then calculates the AIC and BIC values of the three copulas to accurately evaluate the model fitting

effect.

According to the comparison between Figure 4 and Figure 3(a), (b), (c), it can be seen that the bivariate histogram of Figure 4 has a thick tail at the upper tail, which is consistent with the characteristics of Gumble copula in Figure 3(b).

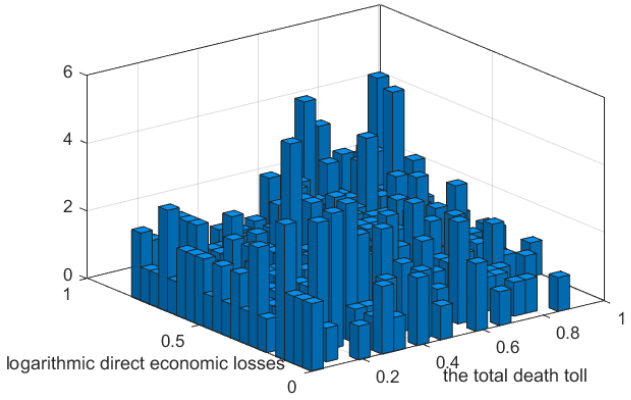


Figure 4: Binary Histogram of Data

The joint probability density plot of Gumble copula has obvious two-tailedness, and the graph shows a "J" shape. The upper tail is relatively high, and the lower tail is relatively low, indicating that the logarithmic direct economic losses and total death toll are more strongly correlated at the large and small extremes. The correlation is stronger at the extreme values. The preliminary indication is that the fitting effect of Gumble copula is the best among the three. In order to select the optimal copula function more accurately, the AIC and BIC values are further calculated to evaluate and compare the goodness of fit of different copula functions, and the results are shown in Table 8.

Table 8: Goodness of Fit of Copula Function

| | Frank copula | Gumbel copula | Clayton copula |
|---|---|---|---|
| AIC | -69.145921 | -79.310211 | -8.832996 |
| BIC | -65.319921 | -75.484211 | -5.006996 |

As can be seen from Table 8, the AIC and BIC of Gumbel copula are both minimized. Therefore, in this paper, Gumbel copula is used as the optimal copula to fit the data of torrential rain and flood disasters.The PDF fitted by Gumble copula is shown in Figure 3(b), in which there are two more obvious peaks at the upper and lower tails, respectively. The parameters are further estimated using the great likelihood method and obtained.

② Rank correlation coefficient test

Table 9: Results of rank correlation coefficient

| rank correlation coefficient | Gumbel copula |
|---|---|
| Spearman's $\rho$ | 0.4044855 |
| Kendall's $\tau$ | 0.2795389 |

From the results of the rank correlation coefficients in Table 9, the coefficients of the Spearman's rank correlation test and Kendall's rank correlation test for the Gumbel copula are 0.4044855 and 0.2795389, respectively, which are positive and indicate that there is a significant and positive correlation between logarithmic direct economic losses and total death toll.

## 5. Conclusion

In this paper, we use the statistics of major torrential rain and flood disasters in China from 2003 to 2019 as a research sample. First, we describe the risk of economic and fatal losses, which is due to major torrential rain and flood disasters. Then, we focus on the correlation between direct economic losses and death toll. Finally, we model the copula function.

(1) The distributions of logarithmic direct economic losses and total death toll both show obvious skewness and thick-tailed characteristics. We usually use the classical linear correlation method when the sample follows the normal distribution, while the total death toll does not follow the normal distribution. So the linear correlation method is no longer suitable for studying the relationship between the two. Hence the copula function can be used to draw more accurate conclusions.

(2) Based on the Gumbel copula function, we study the relationship between log direct economic losses and total death toll, and find that they are positively correlated, with a Spearman coefficient of 0.4044855 and a Kendall coefficient of 0.2795389. Therefore, we use the copula approach to more accurately characterize the correlation between direct economic losses and total death toll from a new perspective.

## References


*[1] Li Ying, Zhao Shanshan. Study on flood damage and disaster risk in China from 2001 to 2020[J]. Progress in Climate Change Research, 2022, 18(02): 154-165.*

*[2] Xia Jun, Wang Huijun, Gan Yaoyao et al. Research progress of heavy rainfall and flood forecasting methods in China [J]. Heavy Rainfall Disaster, 2019, 38(05): 416-421.*

*[3] China Meteorological Administration. China Meteorological Disasters Yearbook [M]. Beijing: Meteorological Press, 2004-2020.*

*[4] Zhao Shanshan, Gao Ge, Huang Dapeng et al. Characterization of meteorological disaster losses in China from 2004 to 2013 [J]. Journal of Meteorology and Environment, 2017, 33(01): 101-107.*

*[5] Liu Xinhong, Meng Shengwang, Li Zhengxiao. Copula mixture distribution model for earthquake loss risk and its application [J]. Systems Engineering Theory and Practice, 2019, 39(07): 1855-1866.*

*[6] Sklar, A. Fonctions de répartition àn dimensions et leurs marges [J]. Publications de l'Institut de Statistique de l' Université de Paris, 1959, 8: 229–231.*

*[7] Ye Yenting, Gong Junqiang, Zhang Haixia et al. Joint probability analysis of tropical cyclone wind and rain based on two-dimensional Archimedean copula function [J]. Journal of Tsinghua University (Natural Science Edition), 2023: 1-8.*

*[8] Joe, H. Multivariate models and dependence concepts [M]. Chapman& Hall: New York, 1997.*

*[9] He Shuhong, Huang Zhenxiong, Zheng Shangping. A study on the risk assessment of geological disasters in Yunnan Province based on copula model [J]. Journal of Yunnan University (Natural Science Edition), 2023, 45(02): 256-265.*

*[10] Hofert M, Kojadinovic I, Mächler M, et al. Elements of copula modeling with R [M]. Springer International Publishing, 2018.*