

Earth's Future

RESEARCH ARTICLE

10.1029/2024EF004775

Key Points:

- This study utilized the Sina Weibo platform to detect flood events in 370 cities in China from 2012 to 2023
- 52.43% of cities have experienced flood events, among which Xi'an, Chengdu, Tianjin, and Chongqing have had the most occurrences
- The detected and validated 674 flood events are mainly distributed in eastern and central China

Correspondence to:

D. Shen,
dingtaoshen@outlook.com

Citation:

Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., & Zhang, S. (2025). How many floods have occurred in China in the past decade? A perspective from social media. *Earth's Future*, 13, e2024EF004775. <https://doi.org/10.1029/2024EF004775>

Received 5 APR 2024
Accepted 24 MAR 2025

Author Contributions:

Conceptualization: D. Shen
Data curation: W. Chen, S. Zhang
Formal analysis: W. Chen, C. Zhang, S. Xiao
Funding acquisition: D. Shen
Investigation: W. Chen, C. Zhang, S. Zhang
Methodology: D. Shen, H. Gu, S. Xiao
Project administration: D. Shen
Resources: S. Xiao
Software: H. Gu, W. Chen, C. Zhang, S. Xiao, S. Zhang
Supervision: D. Shen
Validation: H. Gu, W. Chen, C. Zhang, S. Xiao, S. Zhang
Visualization: H. Gu, W. Chen, C. Zhang, S. Xiao
Writing – original draft: D. Shen, H. Gu
Writing – review & editing: D. Shen

© 2025. The Author(s).

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

How Many Floods Have Occurred in China in the Past Decade? A Perspective From Social Media

D. Shen^{1,2} , H. Gu^{1,2}, W. Chen³, C. Zhang⁴ , S. Xiao^{1,2}, and S. Zhang^{1,2}

¹Key Laboratory for Geographical Process Analysis & Simulation of Hubei Province, Central China Normal University, Wuhan, China, ²College of Urban and Environmental Sciences, Central China Normal University, Wuhan, China, ³Jiangsu Provincial Planning and Design Group, Nanjing, China, ⁴School of Information Engineering, China University of Geosciences in Beijing, Beijing, China

Abstract Global climate change has led to frequent and widespread flood disasters in China. Traditional flood disaster investigations mainly focus on major flood events, and small-scale flood events are often overlooked. This study utilized the Sina Weibo social media platform to detect flood events in 370 cities in China from 2012 to 2023. We downloaded 73.52 million Weibo posts and developed a two-step flood detection algorithm. In the first step, the algorithm initially identifies 956 flood events based on changes in posting frequency. In the second step, an LDA topic model is used to detect topics for these flood events and automatically filter out false events, resulting in 729 flood events. Verification of these events confirmed that 629 of the 729 were real flood events, achieving a detection accuracy of 86.28%. In the end, after excluding all false flood events and reinstating the mistakenly removed real ones, we obtained a total of 674 verified flood events. Among these 370 cities, 194 cities experienced flood disasters, accounting for 52.43% of the total. Additionally, we compared our findings with online news reports, as well as the flood data sets from the GDACS and EM-DAT. We found that our study had a high detection rate for urban waterlogging events. However, there were cases of missed detection for flash floods and small watershed flood disasters. Nevertheless, this study represents the most comprehensive publicly available detection of flood events in China to date, which is of great significance for the government's flood management and decision-making.

Plain Language Summary Climate change has led to a significant increase in extreme precipitation events in China, resulting in a rise in flood disasters across the country. However, our understanding of the frequency and distribution of flood events in China remains limited due to the lack of an effective monitoring and reporting mechanism. To address this gap, researchers from Central China Normal University analyzed data from the Sina Weibo to detect flood events in 370 cities in China between 2012 and 2023. They identified a total of 674 flood events from 73.52 million Weibo posts, indicating that over half of Chinese cities (52.43%) have experienced flood events in the past decade. The majority of these cities are located in the eastern and central regions of China, with 12 cities reporting 10 or more flood events. Specifically, Xi'an, Chengdu, Tianjin, and Chongqing have been the hardest hit, with the highest number of flood events. The northwest region of China, which has lower precipitation and population density, has seen fewer flood events. This study provides valuable insights into the current flood situation in Chinese cities, highlighting the need for effective monitoring and reporting mechanisms to mitigate the impact of climate change on flood disasters.

1. Introduction

Global climate change has led to an increasing frequency of extreme precipitation events at regional and local scales (Blöschl et al., 2019; Hirabayashi et al., 2013). Consequently, severe flood disasters have occurred worldwide, posing a common public safety challenge for many countries (Tellman et al., 2021). On the occasion of the 33rd International Day for Disaster Reduction on 13 October 2022, the Ministry of Emergency Management and the Ministry of Education (MEM & ME, 2022) of China, released the “2021 Global Natural Disaster Assessment Report.” The report indicates that in comparison to the 30-year average from 1991 to 2020, the total frequency of global natural disasters in 2021 exceeded by 13%, with floods being the most frequent, surpassing historical records by 48%.

The “Sendai Framework for Disaster Risk Reduction 2015–2030,” adopted by the United Nations in 2015, aims to significantly reduce disaster risks and losses and outlines seven global targets for reducing disaster losses (UNDRR, 2015). At the national level, the collection of flood disaster data and the statistics of disaster losses

become particularly crucial. Numerous natural disaster databases, including flood disasters, have been developed for specific countries or globally (Kellermann et al., 2020). Examples include the Global Disaster Alert and Coordination System (GDACS) by the United Nations and the European Union, the NatCatSERVICE Database by Munich Reinsurance Company in Germany, and the Emergency Events Database (EM-DAT) by the Center for Research on the Epidemiology of Disasters at KU Leuven in Belgium. China's Ministry of Emergency Management has also launched the Global Disaster Data Platform (GDDP).

These flood disaster databases have provided assistance and references for governments in conducting flood risk emergency management. However, flood disaster investigations mainly focus on major flood events, while small-scale flood events without casualties or significant property damage are often overlooked. These difficulties have led to deficiencies in the reliability and completeness of flood disaster information in current global disaster databases, particularly in terms of the inclusion of flood events in underdeveloped regions (Y. Wang et al., 2022). Most flood data sets have relatively few flood events and are typically at the provincial scale (EM-DAT, 2024). For example, we found that the Dartmouth Flood Observatory (DFO, 2024) data set includes only 45 flood events in China from 2012 to 2021, and the EM-DAT (2024) only records disaster events that exceed local capacity and require a request for external assistance at the national or international level. The scarcity of flood event data sets that are highly detailed at the city-scale, capturing more events and at higher frequencies, remains a significant challenge. Many current studies on flood impacts, including those on population exposure (Tellman et al., 2021) and economic inequality (Linderson et al., 2023), often rely on global data sets like DFO or EM-DAT. However, these data sets may lead to an underestimation of flood impacts due to their limited temporal and spatial resolution, which might not fully capture the frequency and localized effects of urban flood events.

In situ monitoring of rainfall, water level, and flow (A. Islam et al., 2022), as well as optical and microwave remote sensing observations (M. T. Islam & Meng, 2022; Zhang et al., 2021), can provide support for flood disaster investigations and provide detailed data on flood extent and water depth (Olthof & Svacina, 2020). However, in-situ monitoring is expensive and generally only sustained in large cities and important rivers, while remote sensing observations are limited by the long revisit cycle of satellites and are therefore unable to capture short-duration flood events lasting only a few days. For example, Tellman et al. (2021) used MODIS satellite data to map global flood events. However, this process still required support from the DFO data set. Due to limitations like cloud cover, only around 30% of flood events could have their inundation extents mapped effectively. This highlights the challenges in achieving comprehensive flood monitoring using satellite data alone, as certain environmental factors can significantly limit data coverage and accuracy.

The popularity of social media has provided a new perspective for the detection of flood events. When floods occur, people often post related information on social media platforms, discussing the social impacts of floods and sharing rescue information. Researchers have conducted numerous studies related to flood disasters using social media data, including flood mapping (Kankanamge et al., 2020), topic detection (Wu et al., 2020), community resilience (Kontokosta & Malik, 2018), and emergency response (Zhou et al., 2022). In particular, research on flood detection based on Twitter has attracted the attention of many researchers and has yielded fruitful results. Studies have been conducted in various regions, including the UK (Arthur et al., 2018), Japan (Shoyama et al., 2021), the Philippines and Pakistan (Jongman et al., 2015), Italy (Rossi et al., 2018), and a global flood event database has even been launched (de Bruijn et al., 2019).

To detect flood outbreaks (including location, start time, and end time), conducting topic classification on social media posts is a feasible approach. For instance, Resch et al. (2018) utilized LDA model to detect the fluctuation of flood disaster topics on Twitter, thereby capturing the footprint of flood disasters. In other studies, considering that people extensively discuss flood events on social media when disasters occur, many researchers detect disaster events by analyzing the temporal changes in social media data. These methods include Infinite State Automata (Kleinberg, 2003), comparison of the number of keywords in continuous time windows (Guzman & Poblete, 2013; X. Wang et al., 2013), Bayesian changepoint detection (Tartakovsky & Moustakides, 2010), and detection of temporal changes in text posting intervals within time windows (de Bruijn et al., 2019; Riley, 2008). Additionally, there are statistical methods specifically tailored for Twitter data such as E-divisive with Medians (EDM, James et al., 2016), which many scholars have applied for flood event detection (Shoyama et al., 2021; Theja Bhavaraju et al., 2019).

The above research on flood event detection is almost entirely based on Twitter. However, Twitter is difficult to cover the Chinese region comprehensively because most social media users in China do not use Twitter but have

their own Chinese community, which is Sina Weibo, the Chinese version of Twitter. Although not as globally popular as Twitter, Sina Weibo has a huge user base in China, with a monthly active user count of 598 million as of the end of 2023, which accounts for nearly half of China's population. Chinese researchers have conducted many studies on flood disasters using Sina Weibo. However, natural language processing for the Chinese context is not as mature as in English, and current research is mostly limited to topic classification and sentiment analysis for specific flood events. There remains a gap in research on flood event detection. Despite the growing perception among the public that flood disasters in China have become more severe in recent years, critical information such as the location, frequency, and impact of floods remains unclear. Furthermore, the flood disaster data released by the Ministry of Emergency Management of China through the GDDP is significantly incomplete. As a result, detailed, publicly accessible flood data sets for China are still largely absent.

This study aims to download Weibo posts related to heavy rainfall and flooding from 2012 to 2023 and develop a comprehensive flood event detection algorithm to identify historical flood events across 370 cities in China over the past 12 years. The key contributions of this research are as follows: (a) A city-scale flood event detection algorithm specifically designed for Sina Weibo, which can support the collection of current and future flood events in China; (b) A validated, ready-to-use flood event data set that will enrich global flood databases by adding crucial data on flood events in China; (c) The data set will also aid in research on the economic and social impacts of flood events in China, particularly contributing to the development of resilient cities and enhancing flood emergency management. The structure of this study is as follows:

The second section is Data Collection, which explains how to download and preprocess Weibo posts from the Sina Weibo platform. The third section is the Method, which focuses on disaster information extraction, flood event detection, and the filtering of false flood events. The fourth section is Results and Analysis, which presents the distribution of flood events and compares the results with other flood event data sets for validation, and analyzes the impact of different conditions on flood event detection performance. The fifth section is Discussion and Conclusion, which reflects on the limitations of this research and suggests potential directions for future improvements.

2. Data Collection From Sina Weibo

Using the keywords “heavy rain” and “flood” to download flood-related text and image data from social media platforms is a common practice (de Bruijn et al., 2019; Kharazi & Behzadan, 2021; Li et al., 2017; Tan & Schultz, 2021). However, compared to English, Chinese vocabulary is more diverse, and people often use idioms, proverbs, and short phrases to describe heavy rain and flood events. Therefore, using only the simple keywords “heavy rain” and “flood” may result in missing a significant amount of data. Bhabaraju et al. (2019) also recommend using keywords related to the consequences of natural disasters and emergency responses. The most appropriate method is to extract related and similar words for “heavy rain” and “flood” from social media sample data of a major flood event.

This study selected the extremely heavy rain and flood event that occurred in Henan Province, China on 20 July 2021, as a sample. Many scholars have conducted research on this flood event (He et al., 2023; Manandhar et al., 2023; Y. Yang et al., 2023). We used the API interface provided by Sina Weibo platform and a web crawler program to download 83,025 Weibo posts by setting the keyword as “Zhengzhou heavy rain” from Sina Weibo. Zhengzhou, the capital city of Henan Province, experienced the most severe flood damage compared to other cities during this flood event.

Mining similar and related words from text corpora has become quite mature, and this study utilized the Word2Vec model. Word2Vec is an efficient word embedding model proposed by Mikolov et al. (2013), widely applied in flood-related research (Lin & Wu, 2022; T. Yang et al., 2018). In this study, Word2Vec was used to learn from the downloaded 83,025 Weibo posts, extracting 24 similar words related to “heavy rain” and 20 related words to “flood,” as shown in Table 1. We placed these Chinese keywords in the brackets in the table, and due to the differences in expression between Chinese and English, there are cases where the English interpretations of multiple Chinese keywords are the same.

We used these keywords to download a total of 73.52 million Weibo posts from the Sina Weibo platform in China between 2012 and 2023. Figure 1 shows the monthly changes in the quantity of Weibo posts. It can be observed from the figure that the number of Weibo posts from 2012 to 2017 was relatively lower. With the gradual

Table 1

Heavy Rain and Flood Keywords Extraction

Heavy rain synonyms	Flood synonyms
rain (雨, 下雨), heavy rain (暴雨), heavy rainfall (大雨, 大暴雨, 强降雨, 大到暴雨), moderate rain (中雨), storm (暴风雨), gale (大风), Short-lived (短时), Overcast (阴天), Moderate to heavy rain (中到大雨), light rain (小雨), rain stop (雨停), lightning and thunder (电闪雷鸣), getting heavier (越下越大), dripping (淅淅沥沥), rainy (多雨), heavy precipitation (强降水), rainfall (降雨), rainfall amount (雨量, 降雨量), precipitation amount (降水量)	flood (洪灾, 洪水, 大水), torrential downpour leads to flooding (暴雨成灾), rain disaster (雨灾, 雨水), torrent (洪流), wash away (冲毁), water flow (水流), river water (河水), flash flood (山洪爆发), flood disaster (洪灾), standing water (积水), landslide (滑坡), collapse (塌陷, 坍塌, 塌方), water depth (水深), waist-high (齐腰), waterlogging (内涝)

popularization of smartphones, the quantity of downloaded Weibo posts significantly increased starting from 2018. From 2019 to 2023, more than 10 million Weibo posts were captured each year.

3. Method

3.1. Flood Damage Information Extraction From Weibo Posts

In the 73.52 million Weibo posts, the vast majority of the posts are actually not related to real flood events. Researchers use various tools to classify the downloaded social media posts, filtering out the posts unrelated to floods. Common text classification methods include direct keyword matching (Bhavaraju et al., 2019), machine learning methods such as Naive Bayes, SVM, Random Forest, and LightGBM (Tan & Schultz, 2021), as well as deep learning methods like BERT (de Bruijn et al., 2019). This study does not attempt to classify Weibo posts as either flood-related or not flood-related, but instead focuses on identifying whether the posts contain flood damage information. Flood events are detected based on the outbreak of such damage information, as this type of information is more prevalent during flood disasters, such as people being stranded, buildings and roads collapsing, and traffic congestion.

Classifying flood damage information from Weibo posts is much more complex than classifying flood information because flood damage includes many aspects such as harm to people, damage to buildings, damage to infrastructure, etc. As shown in Table 2, we constructed damage information classification indicators, which include eight primary categories and 31 secondary categories. We did not use machine learning or deep learning

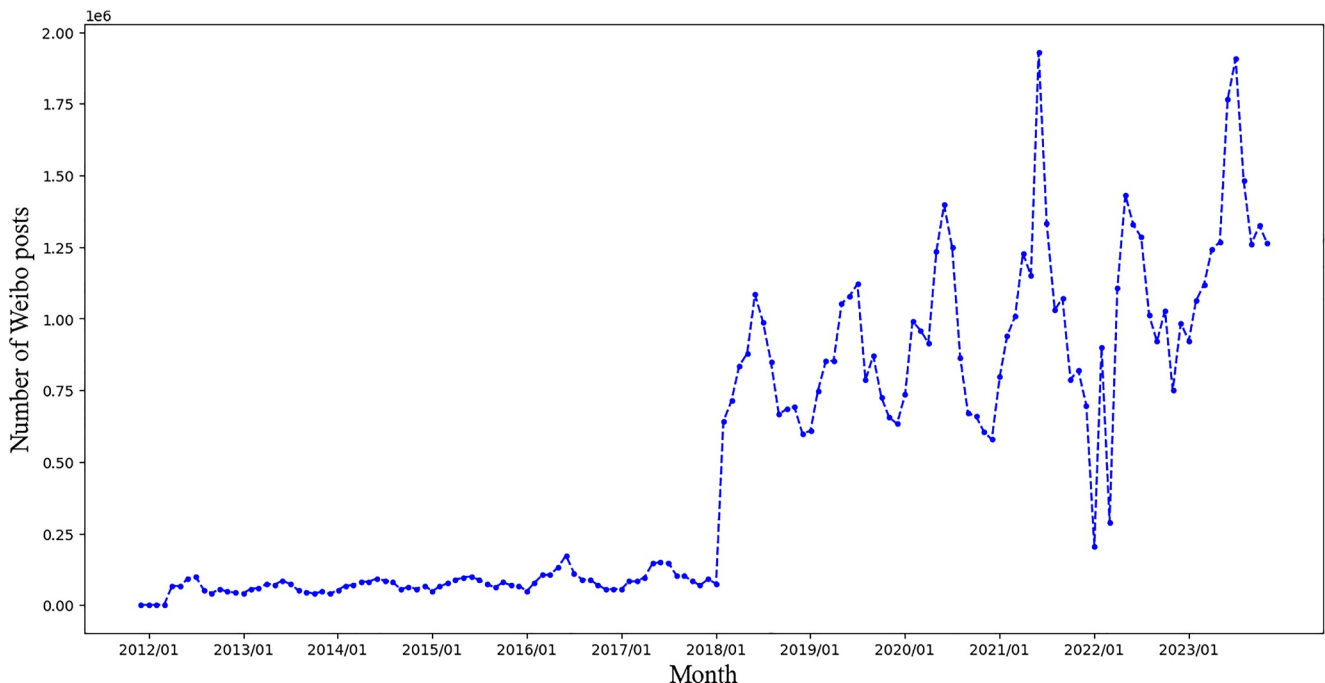


Figure 1. Monthly distribution of 73.52 million Weibo posts related to heavy rain and flooding between 2012 and 2023.

Table 2
Classification Indicators for Flood Damage Information

Level 1 indicators		Level 2 indicators	
Social impact (社会影响)	People (人)	Building (建筑物)	Basic necessities of life (衣食住行)
	Education, science, culture and health (教科文卫)	Overall impact (综合影响)	
Business impact (商业影响)	Shops (商铺)	Merchants (商户)	
Infrastructure impact (基础设施影响)	Bridges and tunnels (桥梁隧道)	Roads (道路)	
Communication impact (通信影响)	Communication carrier (通信运营商)	Communication base stations (通信基站)	Communication signals (通信信号)
	Communication terminals (通信终端)		
Electricity impact (电力影响)	Electrical facilities (供电设施)	Electrical equipment (供电设备)	Electricity supply (电力供应)
Water conservancy impact (水利影响)	Rivers, lakes and reservoirs (河流湖库)	Levees (堤防工程)	
Industry impact (工业影响)	Factories and workshops (工厂车间)	Industrial operations (企业运营)	
Traffic impact (交通影响)	Trees (树)	Drainage wells (排水井)	Roads (道路)
	Road traffic (道路交通)	Air transportation (航空运输)	Rail transportation (铁路运输)
	Subway station (地铁站)	Subway operation (地铁运营)	Automobile passenger transport (汽车客运)
	Vehicle (车辆)	Vehicle accessories (车辆配件)	

Table 3
Example of Damage Information Keywords Extraction

Level 1 indicators	Level 2 indicators	Word pairing	Keywords	Scenarios
Traffic impact (交通影响)	Road traffic (道路交通)	n. + v. or n. + adj	{Urban transportation (城市交通), Public transportation (公共交通, 公交), Urban roads (城市道路), Highway (公路), Road traffic (道路交通), Transportation system (交通系统), Road (道路), Order (秩序), Arterial roads (大动脉), Traffic lights (信号灯), Traffic order (交通秩序)} + {congested (拥堵), paralyzed (瘫痪), obstructed (受阻), blocked (阻断), halted (停摆), cascading (连锁反应), recover quickly (尽快恢复), alarmed (告警), stopped (停止)}	
	Rail transportation (铁路运输)	n. + v. or n. + adj	{train (火车), train station (火车站), high-speed train (动车), railway (铁路), high-speed railway (高铁), locomotive (列车)} + {delayed (晚点), start-stop (走走停停), turn back (折返), transfer (转车), speed limit (限速), block off (封锁), wait for clearance (待避), suspended (停运), trapped (被困)}	

Table 4
Accuracy Evaluation of Flood Damage Information Index Classification

Indicators	Precision	Recall	F1-score
Social impact	82.28%	93.92%	87.71%
Business impact	92.96%	95.20%	94.07%
Infrastructure impact	55.28%	97.51%	70.56%
Communication impact	84.93%	86.11%	85.51%
Electricity impact	90.51%	95.38%	92.88%
Water conservancy impact	75.00%	92.30%	82.75%
Industry impact	79.16%	73.07%	76.00%
Traffic impact	78.18%	89.06%	83.27%

methods for damage information classification because it involves multi-class classification problems. Firstly, it is limited by the number of samples, and secondly, a Weibo post may contain various types of flood damage scenarios, which can lead to confusion. Therefore, we adopted the keyword set matching method by constructing flood damage keywords to detect various damage information from Weibo posts. We placed the Chinese words of each flood damage information indicator in brackets in Table 2.

This study used the Word2Vec to construct synonyms related to secondary indicators from the Weibo post data set of heavy rain in Henan Province, and then explored matching words related to the damage of these secondary indicators, which are either verbs or adjectives. Table 3 shows examples of keywords for the secondary indicators “Road traffic” and “Rail transportation” in the category of transportation impact. From the keyword sets, it can be seen that the first group of keywords consists of synonyms related to

“Road traffic” and “Rail transportation,” while the second group of keywords are modifiers indicating the flood damage characteristics suffered by the first group of words. We use these {noun + verb} and {noun + adjective} word pairs to search the Weibo post in order to identify whether the post contains damage information and determine the type of damage.

To evaluate the classification accuracy of flood damage information, we used the Weibo data of heavy rain in Henan Province to create a sample data set with 5,000 posts, and manually labeled the data set to mark the damage information for each post. We categorize the Weibo posts into different level 1 indicator categories, as shown in Table 2. Then, we used the established keywords to perform machine classification on the sample data. Unlike manual classification, machine classification uses keywords to match the content of the post. Once a matching keyword pair appears in the post, the Weibo post is classified into the corresponding level 1 indicator category. The accuracy of the machine classification relative to the manual classification was evaluated using three indicators: Precision, Recall, and F1-score, as shown in formulas 1–3. Many scholars use these three indicators to conduct accuracy evaluations when using social media data to study flood disasters (Fu et al., 2022; Zhou et al., 2022). In the formulas, TP (true positive) represents Weibo posts that are labeled as damage information by both machine and manual classification, FP (false positive) represents Weibo posts that are labeled as damage information by the machine but not by manual classification, and FN (false negative) represents Weibo posts that are labeled as damage information manually but not detected by the machine. Table 4 shows the classification evaluation results of damage information for the sample data. The classification accuracy of different damage information is generally around 80%, which can meet the requirements for subsequent flood damage information extraction from 73.52 million Weibo posts.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Finally, this study used the constructed keyword set to classify flood damage information in 73.52 million Weibo posts, and found that 5.78 million Weibo posts contained flood damage information, accounting for 7.86% of the total number of Weibo posts. Figure 2 shows the monthly changes in the number of Weibo posts containing flood damage information. In July 2021, the most damage information was recorded, reaching 300,000 posts. That year, China experienced several major flood events, including the catastrophic rainstorm and flood disaster in Henan. The least damage information was recorded in 2012, with only 60,000 posts. The significant difference is partly due to the fact that Sina Weibo was launched in 2010 when there were fewer users and smartphones were not yet widely popular. Since 2018, the number of flood damage information has been consistently above 500,000 each year. The graph shows that flood damage information exhibits seasonal cyclic fluctuations, peaking during the flood season in July each year and reaching a trough in non-rainy December.

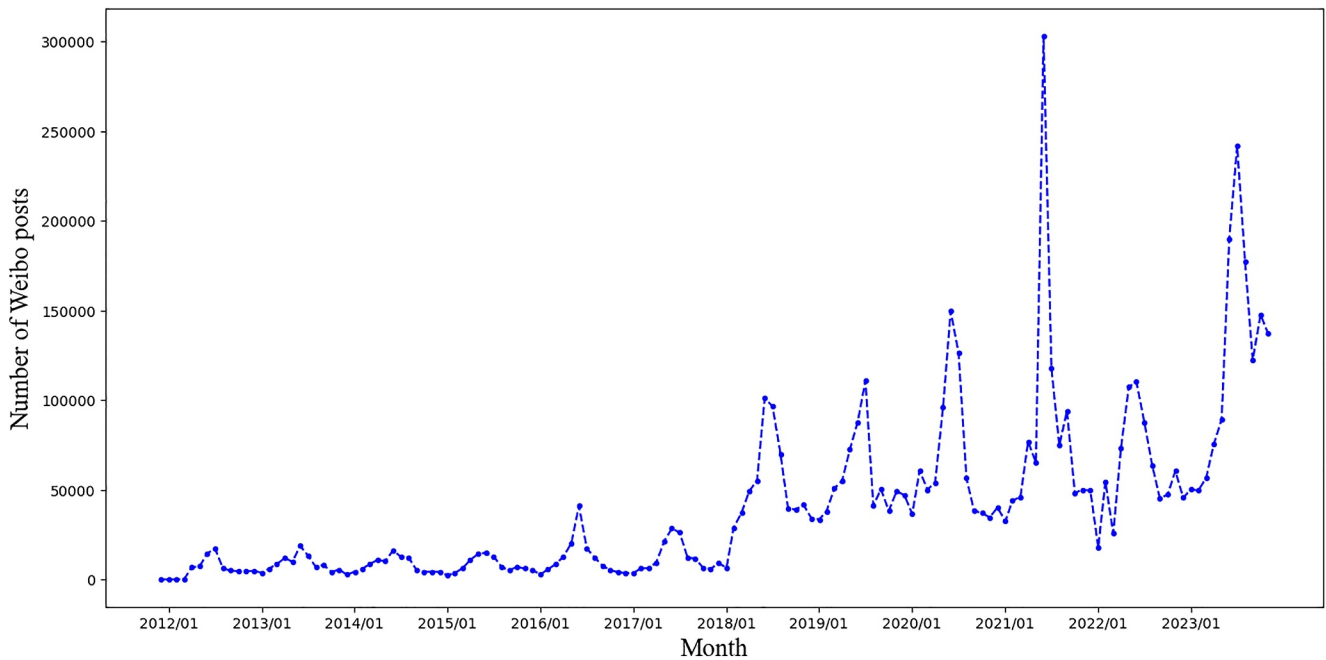


Figure 2. Monthly distribution of 5.78 million Weibo posts containing flood damage information between 2012 and 2023.

3.2. Categorize Flood Damage Information Into Different Cities

It is crucial to extract city information from the 5.78 million Weibo posts containing flood damage information. Without city information, a Weibo post containing flood damage information is useless. We used the “Chinese Province City Area mapper” (CPCA) library (DQinyuan, 2021), which is currently popular for resolving Chinese location addresses, to parse the city information in the Weibo posts. The CPCA library automatically extracts all city names from the Weibo posts and outputs the information in the format {province, city, county/district} as a list. In this study, from the 5.78 million Weibo posts containing flood damage information, we identified 3.15

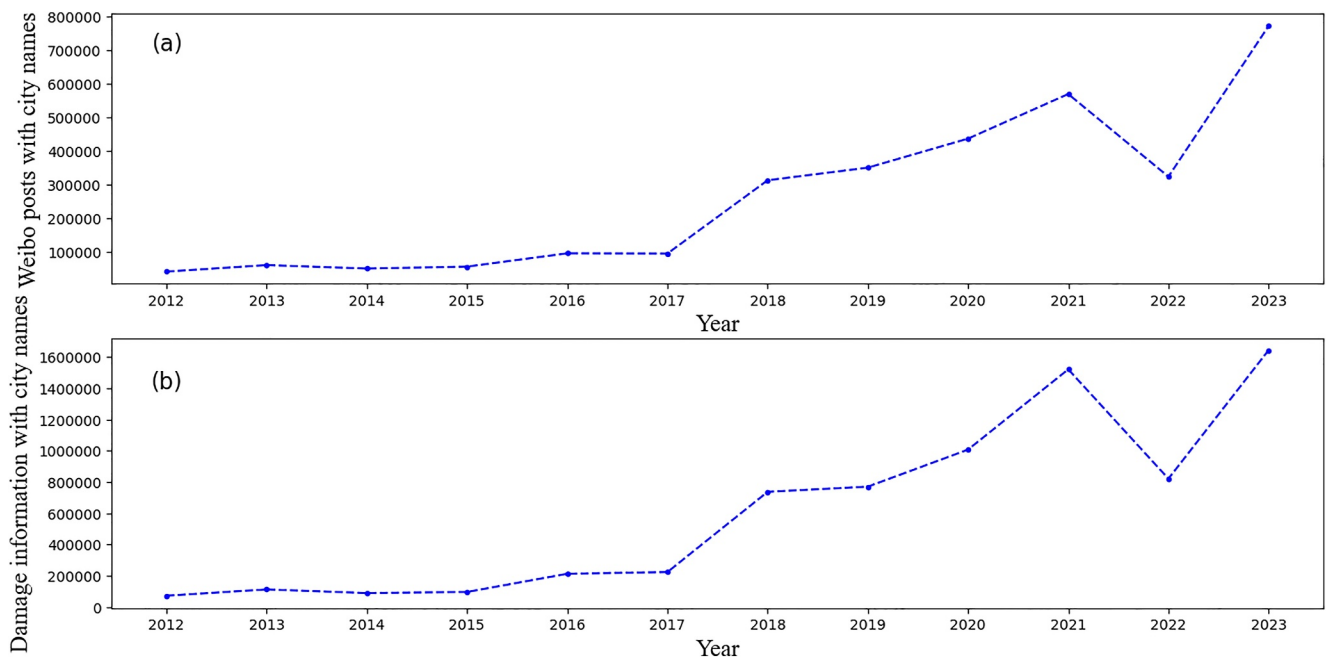


Figure 3. Annual changes in Weibo posts and damage information with city names: (a) Weibo posts; (b) Damage information.

million Weibo posts with city names. Figure 3a shows the change in the number of these 3.15 million Weibo posts from 2012 to 2023.

For a Weibo post containing flood damage information, it may contain multiple types of flood damage information (Level 1 indicators). For example, a Weibo post may not only describe the traffic congestion in a city but also mention the impact on infrastructure such as road collapses. Therefore, the number of flood damage information recorded in these 3.15 million Weibo posts will far exceed 3.15 million. In this study, 7.3 million instances of flood damage information were detected from these 3.15 million Weibo posts. Figure 3b illustrates the change in the number of these 7.3 million flood damage information instances from 2012 to 2023.

Divide these flood damage information according to the Weibo posting time into each city, so that each city stores a flood damage information list with a timestamp. The format of the list is {Weibo posting time, Weibo text, flood damage information type}. When a Weibo post contains multiple flood damage information, it will be converted into multiple records for the city, and the posting time in these records is the same. This can increase the frequency of Weibo postings, helping to detect the outbreak of flood events based on changes in Weibo posting frequency later on.

3.3. Correction of Flood Damage Information During Day/Night Intervals

The number of Weibo posts made by people at night is significantly smaller than during the day, resulting in longer intervals between Weibo posts at night compared to during the day. In order to eliminate the periodic fluctuations in the interval of flood damage information posts caused by day/night variations, it is necessary to correct the interval of flood damage information posts. This study refers to the tweet interval correction method proposed by de Bruijn et al. (2019) and corrects the interval between flood damage information posts. The specific method is as follows:

First, calculate the average number of flood damage information posts in the city for each hour. As shown in formula 4, sum the number of flood damage information posts for each hour, as well as the 2 hr before and after it, totaling 5 hr, then calculate the mean to obtain the average number of flood damage information posts for the h -th hour.

$$\bar{n}_h = \frac{n_{h-2} + n_{h-1} + n_h + n_{h+1} + n_{h+2}}{5} \quad (4)$$

Secondly, calculate the correction factor for each hour. As shown in formula 5, the correction factor c_h for the h -th hour is equal to the average number of flood damage information posts for that hour $\bar{n}_h$ multiplied by 24, then divided by the sum of the average flood damage information posts for 24 hr in that day.

$$c_h = \frac{\bar{n}_h}{\sum_{h=1}^{24} \bar{n}_h} * 24 \quad (5)$$

Finally, the corrected time interval Δt_c between two adjacent flood damage information posts can be calculated using formula 6. When the two records are in the same hour, c_h is the correction factor for the current time h ; when the two records do not belong to the same hour, the correction factor for the hour h of the subsequent record is used in the calculation.

$$\Delta t_c = \Delta t * c_h \quad (6)$$

The above correction factor c_h usually has a value greater than 1 during the day and less than 1 at night. Figure 4 shows the hourly variation of the correction factor c in Beijing on 10 July 2023. The correction factor c drops to below 0.25 between 1 and 4 a.m., starts to exceed 1 at 8 a.m., peaks at 10 a.m., and then falls below 1 starting at 9 p.m. This operation automatically compresses the time intervals of flood damage information during the night, while appropriately stretching the time intervals during the day.

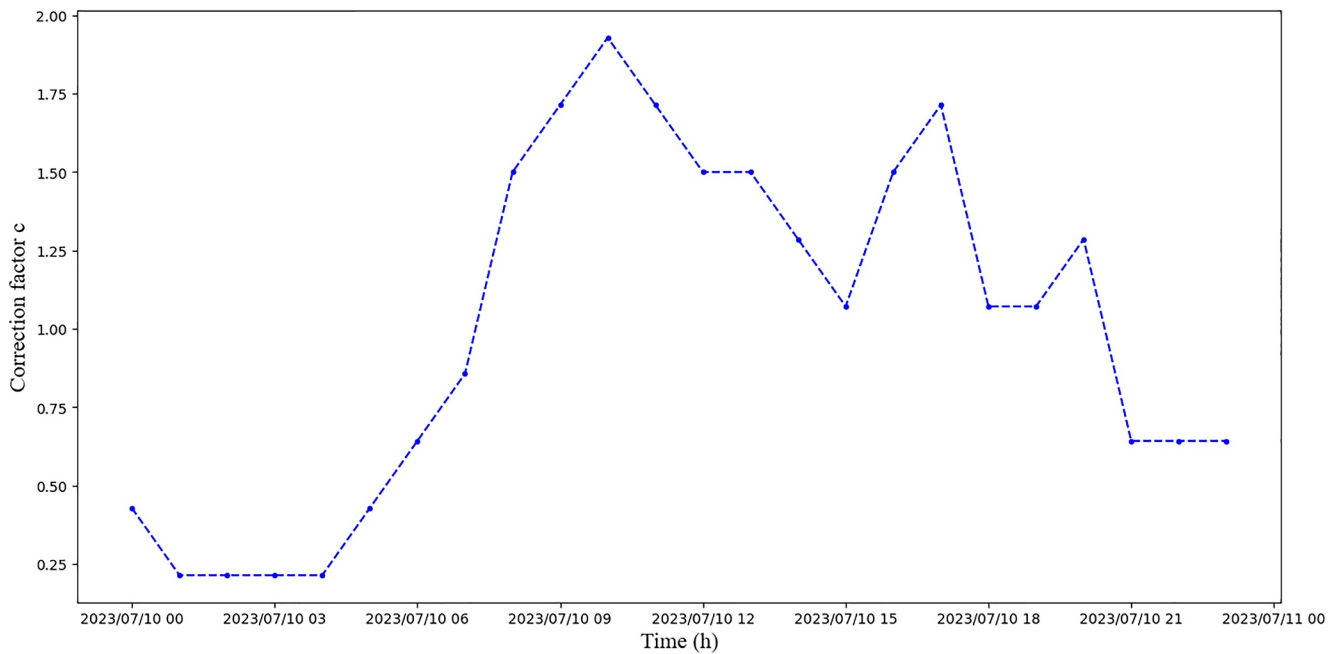


Figure 4. Hourly correction factor c in Beijing on 10 July 2023.

3.4. Detecting Flood Events by Monitoring Changes in Flood Damage Information Posting Frequency

Many scholars have attempted to use a fixed or variable time window moving along the timeline to real-time monitor the number of social media posts within the time window (Guzman & Poblete, 2013; X. Wang et al., 2013) or the changes in the time intervals between social media posts (de Bruijn et al., 2019; Riley, 2008) to detect the outbreak of flood events based on sudden changes in quantity or time. As shown in Figure 5, this study set a 24-hr time window to detect the outbreak of flood events by monitoring the changes in the time intervals of flood damage information posts within the time window. The reason for choosing a 24-hr time window is that regardless of the starting point in time, counting the flood damage information for the subsequent 24 hr covers exactly 12 hr of daytime and 12 hr of nighttime, minimizing the impact of fewer Weibo posts being published at night. The algorithm principle for detecting the outbreak of flood events is as follows:

First, to detect the sudden change in the time intervals of flood damage information posts, it is necessary to know the average interval time $\bar{t}$ of flood damage information posts without any sudden changes. Based on the hourly correction factor c , the corrected intervals between each pair of adjacent flood damage information posts are

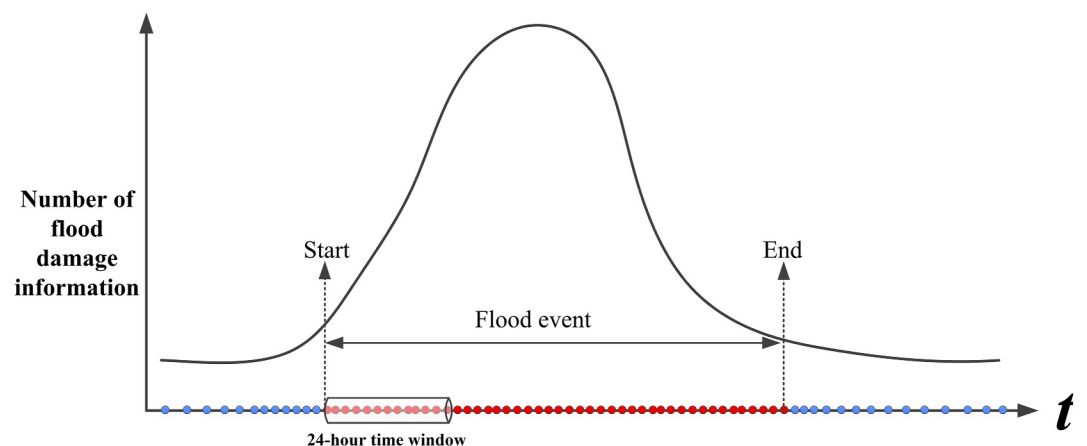


Figure 5. Schematic diagram of flood event detection based on changes in damage information release frequency.

Table 5

The Annual Start Threshold ($4.5\% \times \bar{t}_{\text{year}}$) and End Threshold ($30\% \times \bar{t}_{\text{year}}$) Settings for Flood Outbreak Detection in Beijing

Year	$\bar{t}_{\text{year}}$ (min.)	Start threshold (min., $4.5\% \times \bar{t}_{\text{year}}$)	End threshold (min., $30\% \times \bar{t}_{\text{year}}$)
2012	132.61	5.96	39.78
2013	146.16	6.57	43.84
2014	87.48	3.93	26.24
2015	129.87	5.84	38.96
2016	199.32	8.96	59.79
2017	147.09	6.61	44.12
2018	42.45	1.91	12.73
2019	54.42	2.44	16.32
2020	35.40	1.59	10.62
2021	25.59	1.15	7.67
2022	23.71	1.06	7.11
2023	14.55	0.65	4.36

calculated. Considering that the frequency of discussions on heavy rain and flood topics during the rainy season and non-rainy season is different, and the distribution of flood damage information varies, this study excludes data from the rainy season (May–September) when calculating the average interval time $\bar{t}$, and only uses data from January to April and October to December to calculate the average value. Additionally, as shown in Figure 3b, there are significant differences in flood damage information across different years, so the annual average interval time $\bar{t}_{\text{year}}$ is calculated separately for each year. The formula is as follows:

$$\bar{t}_{\text{year}} = \frac{\sum_{i=1}^{m-1} (|t_{i+1} - t_i| * c_{i+1}) + \sum_{j=1}^{n-1} (|t_{j+1} - t_j| * c_{j+1})}{m + n} \quad (7)$$

In the formula, m represents the number of flood damage information posts from January to April, n represents the number of flood damage information posts from October to December, c_{i+1} and c_{j+1} represent the correction factors at the hours of the $(i + 1)$ th and $(j + 1)$ th flood damage information posts, respectively.

Next, it is necessary to set the time interval threshold for the start and end of flood outbreaks. The start threshold and end threshold for flood outbreak detection have a significant impact on the number and accuracy of detected flood events. Therefore, it is necessary to set reasonable start and end thresholds. Instead of setting uniform start and end thresholds for all cities, we set these thresholds separately for each city in different years. This ensures that different cities use different start and end thresholds for detecting flood outbreaks in different years, thereby better ensuring detection accuracy. It can be inferred that over 12 years, the start and end thresholds for 370 cities amount to a total of $370 \times 12 \times 2 = 8,880$. Fortunately, because the $\bar{t}_{\text{year}}$ of each city in each year is different, we simplify the problem by setting the start and end thresholds as percentages of $\bar{t}_{\text{year}}$. To this end, we define the variable $\text{Percentage}_{\text{Start}}$ as the start threshold percentage and the variable $\text{Percentage}_{\text{End}}$ as the end threshold percentage. Ultimately, we obtain the relationship between the start and end thresholds and these two variables as shown in formula 8. As can be seen from formula 8, the problem of setting the start and end thresholds is transformed into how to set the start threshold percentage $\text{Percentage}_{\text{Start}}$ and the end threshold percentage $\text{Percentage}_{\text{End}}$.

$$\begin{cases} \text{Start Threshold}_{\text{year}} = \bar{t}_{\text{year}} \times \text{Percentage}_{\text{Start}} \\ \text{End Threshold}_{\text{year}} = \bar{t}_{\text{year}} \times \text{Percentage}_{\text{End}} \end{cases} \quad (8)$$

This study attempts to set a unified start threshold percentage and end threshold percentage for all cities. In Section 3.5, we will provide a detailed explanation of how to reasonably set the unified $\text{Percentage}_{\text{Start}}$ and $\text{Percentage}_{\text{End}}$ through sensitivity analysis and multi-objective optimization of flood detection in Beijing. In this study, the start threshold for flood outbreak detection is ultimately set to 4.5% of the annual average interval time

Table 6

List of 17 Real Flood Events in Beijing From 2012 to 2023 and Machine Detection Results (Percentage_{Start} = 4.5%, Percentage_{End} = 30%)

Index	Flood event	Date	Has the machine detected it?	Detected start time	Detected end time
1	Severe rainstorm in Beijing on July 21st	2012.07.21	Yes	2012.07.21	2012.08.07
2	Heavy rain in Beijing on July 16th	2014.07.16	Yes	2014.07.16	2014.07.17
3	Heavy rain in Beijing on July 17th	2015.07.17	Yes	2015.07.17	2015.07.18
4	Heavy rain in Beijing on July 20th	2016.07.20	Yes	2016.07.19	2016.07.26
5	Heavy rain in Beijing on July 20th	2017.07.20	Yes	2017.07.20	2017.07.21
6	Heavy rain in Beijing on August 2nd	2017.08.02	Yes	2017.08.02	2017.08.04
7	Heavy rain in Beijing on July 16th	2018.07.16	Yes	2018.07.15	2018.07.27
8	Heavy rain in Beijing on August 8th	2018.08.08	Yes	2018.08.10	2018.08.15
9	Heavy rain and hail in Beijing on May 17th	2019.05.17	Yes	2019.05.19	2019.05.21
10	Heavy rain in Beijing on August 12th	2020.08.12	Yes	2020.08.10	2020.08.14
11	July 2021 Beijing Heavy Rainfall	2021.07.11	Yes	2021.07.11	2021.07.31
12	Heavy rain in Beijing on August 16th	2021.08.16	Yes	2021.08.16	2021.08.20
13	Heavy rain in Beijing on July 31st	2023.07.31	Yes	2023.07.29	2023.08.08
14	Urban waterlogging in Beijing on 7 August 2015	2015.08.07	No		
15	Heavy rain in Beijing on June 22nd	2017.06.22	No		
16	Heavy rain in Beijing on July 6th	2017.07.06	No		
17	Heavy rain in Beijing on August 5th	2019.08.05	No		

$\bar{t}_{\text{year}}$ (Percentage_{Start} = 4.5%), and the end threshold is set to 30% of the annual average interval time $\bar{t}_{\text{year}}$ (Percentage_{End} = 30%). Taking Beijing as an example, Table 5 shows the start and end thresholds for flood outbreak detection in Beijing from 2012 to 2023, with each city independently storing such a table. From the table, it can be seen that as the frequency of users posting on Weibo increases, the annual average interval time $\bar{t}_{\text{year}}$ of flood damage information is also narrowing, indicating that the current data volume is much larger than in previous years.

We utilize a 24-hr time window to traverse the city's flood damage information on an hourly basis. Assuming we are currently traversing hour h_0 , the time window contains data for the 24 hr from h_0 to $h_0 + 23$. We calculate the average interval time $\bar{t}$ between flood damage information within the time window. If we find that the time $\bar{t}$ is less than 4.5% of the annual average interval time $\bar{t}_{\text{year}}$ and the current flood event is not yet marked, we label it as the start of a flood event. We continue to traverse the subsequent flood damage information on an hourly basis within the time window until the average interval time $\bar{t}$ exceeds 30% of the annual average interval time $\bar{t}_{\text{year}}$, at which point we label it as the end of the flood event. Thus, the detection of a flood event, including its start and end times, is completed. Unlike previous methods that detect the abrupt change point of the time window after it occurs, we have placed it beforehand. We mark the start and end of the flood using the leftmost point h_0 of the time window, rather than $h_0 + 23$ on the right. The advantage of this approach is that it allows for more sensitive detection of flood initiation and earlier termination of flood events. This is because based on our observations, we have noticed that even after a flood event has ended, there is still a significant amount of discussion about the flood for some time, and social media tends to prolong the duration of flood events.

Finally, we remove flood events with a duration of 1 hr and merge flood events with intervals of less than 24 hr, resulting in a data set of flood events for 370 cities in China.

3.5. Set the Start and End Thresholds for Flood Outbreak Detection

As Table 6 illustrates, we have collected a total of 17 flood events in Beijing from 2012 to 2023 through searching online news reports. The reason for selecting Beijing as the reference is because it is the capital of China, and flood events in Beijing are most likely to be reported by the news, making these reports the most credible. As shown in formula 8, we can change the start and end thresholds for flood detection in Beijing for different years by setting different Percentage_{Start} and Percentage_{End}. This allows us to analyze the sensitivity of flood event

Table 7

Accuracy Evaluation of Flood Event Detection in Beijing at Different End Threshold Percentages (With the $\text{Percentage}_{\text{Start}} = 4.5\%$)

$\text{Percentage}_{\text{End}}$	TP	FP	FN	Precision	Recall	F1-score
20.00%	13	3	4	0.8125	0.7647	0.7878
22.50%	13	2	4	0.8666	0.7647	0.8125
25.00%	13	1	4	0.9285	0.7647	0.8387
27.50%	13	1	4	0.9285	0.7647	0.8387
30.00%	13	0	4	1	0.7647	0.8666
32.50%	13	0	4	1	0.7647	0.8666
35.00%	13	0	4	1	0.7647	0.8666
37.50%	13	1	4	1	0.7647	0.8666
40.00%	13	1	4	1	0.7647	0.8666

formula 9, we hope to maximize the values of Precision, Recall, and F1-score, but it is difficult to achieve all of these at the same time. This multi-objective optimization problem actually involves two independent variables: the $\text{Percentage}_{\text{Start}}$ and the $\text{Percentage}_{\text{End}}$. After multiple rounds of trials, we preliminarily set the range of the $\text{Percentage}_{\text{Start}}$ between 3% and 7%, and the range of the $\text{Percentage}_{\text{End}}$ between 20% and 40%. Flood detection results outside these ranges are generally not satisfactory.

detection to different start and end thresholds. Finally, through multi-objective optimization, we evaluate the overall accuracy of Beijing's flood detection results relative to actual flood events and ultimately determine the appropriate $\text{Percentage}_{\text{Start}}$ and $\text{Percentage}_{\text{End}}$. These determined percentages will then be used for flood event detection in other cities to maintain optimal accuracy.

We can still use the three metrics, Precision, Recall, and F1-score from formulas 1 to 3, to evaluate the accuracy of flood outbreak detection. Similar to the accuracy evaluation of flood damage information classification, in these formulas, TP (true positive) represents the number of flood events that actually occurred and were also detected by the machine, FP (false positive) represents the number of flood events that did not actually occur but were detected by the machine, and FN (false negative) represents the number of flood events that actually occurred but were not detected by the machine.

Setting the $\text{Percentage}_{\text{Start}}$ and the $\text{Percentage}_{\text{End}}$ is a multi-objective optimization problem, which has three optimization objectives. As shown in

$$\begin{cases} \max F(1) = \text{Precision} \\ \max F(2) = \text{Recall} \\ \max F(3) = F1 - \text{score} \\ 3\% \leq \text{Percentage}_{\text{Start}} \leq 7\% \\ 20\% \leq \text{Percentage}_{\text{End}} \leq 40\% \end{cases} \quad (9)$$

By fixing the $\text{Percentage}_{\text{Start}}$ and the $\text{Percentage}_{\text{End}}$ separately, we tested the flood event detection results under different threshold percentages. As shown in Table 7, with the $\text{Percentage}_{\text{Start}}$ fixed at 4.5%, the $\text{Percentage}_{\text{End}}$ was gradually increased from 20% to 40%. It can be seen from the table that the optimal $\text{Percentage}_{\text{End}}$ was reached when it was gradually increased to 30%, correctly detecting 13 flood events, missing four flood events, and not detecting any false flood events.

Table 8

Accuracy Evaluation of Flood Outbreak Detection in Beijing With Different Start Threshold Percentages (With the $\text{Percentage}_{\text{End}} = 30\%$)

$\text{Percentage}_{\text{Start}}$	TP	FP	FN	Precision	Recall	F1-score
3.00%	10	0	7	1	0.5882	0.7407
3.50%	10	0	7	1	0.5882	0.7407
4.00%	11	0	6	1	0.6470	0.7857
4.50%	13	0	4	1	0.7647	0.8666
5.00%	14	1	3	0.9333	0.8235	0.875
5.50%	14	2	3	0.875	0.8235	0.8484
6.00%	14	3	3	0.8235	0.8235	0.8235
6.50%	14	4	3	0.7777	0.8235	0.8
7.00%	14	4	3	0.7777	0.8235	0.8

As shown in Table 8, we fixed the $\text{Percentage}_{\text{End}}$ to 30%, and gradually increased the $\text{Percentage}_{\text{Start}}$ from 3% to 7%. It was found that as the $\text{Percentage}_{\text{Start}}$ increased, 14 flood events could ultimately be correctly identified. However, too high a $\text{Percentage}_{\text{Start}}$ would lead to false detections, with the maximum false positives (FP) reaching four events. On the other hand, when the $\text{Percentage}_{\text{Start}}$ was too low, it was difficult to initiate flood outbreak detection, missing many flood events, with the maximum false negatives (FN) reaching seven events. This indicates that both too high and too low start thresholds for flood event detection can lead to significant errors, and it is essential to set the start threshold reasonably.

From Tables 7 and 8, it is evident that the setting of the $\text{Percentage}_{\text{End}}$ is relatively simpler because it primarily affects the duration of flood events (either extending or shortening the end time). This contrasts with the $\text{Percentage}_{\text{Start}}$, where variations in its setting can significantly alter the total number of flood events detected. Since flood event detection is not sensitive to the $\text{Percentage}_{\text{End}}$, this study has simplified the solution to the multi-objective problem by uniformly setting the $\text{Percentage}_{\text{End}}$ to 30%. When

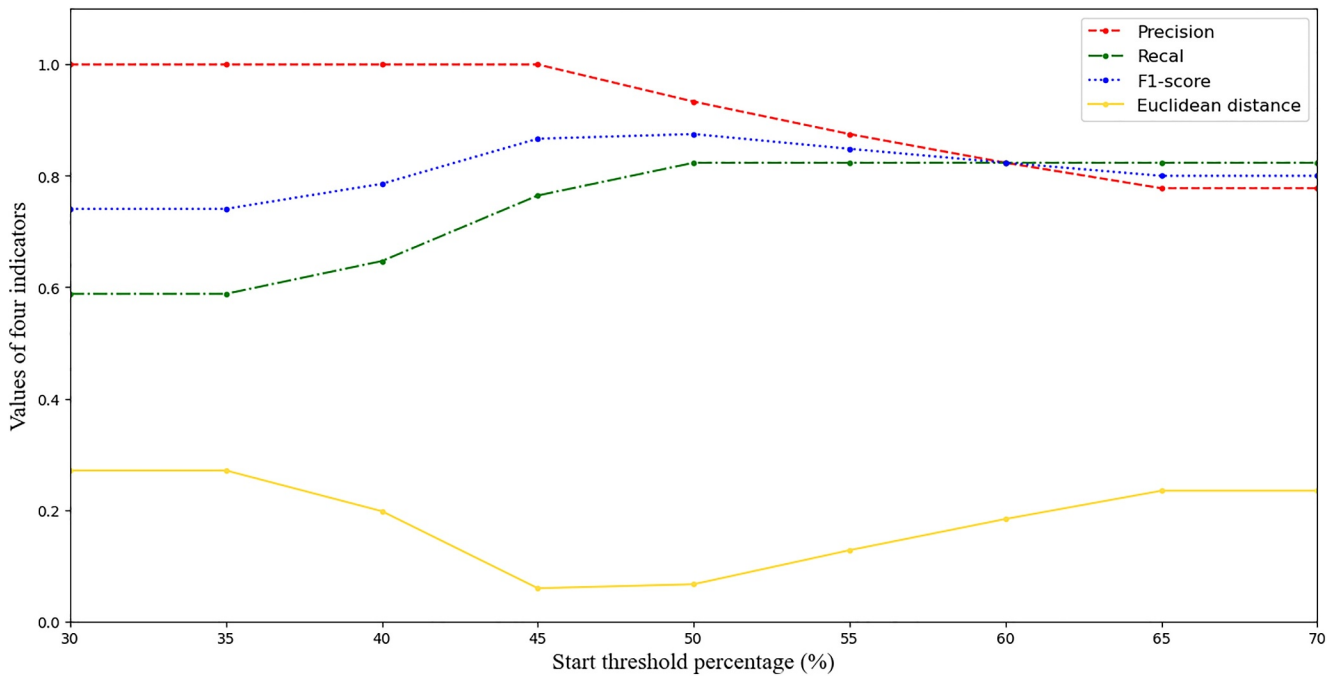


Figure 6. Precision, Recall, F1-score, and the Euclidean distance from general values to ideal values under different start threshold percentages.

comparing the two different start threshold percentages, 4.5% and 5%, in Table 8, it can be seen that the Precision is superior at a start threshold percentage of 4.5%, but the Recall and F1-score are better when the start threshold percentage is at 5%.

To obtain the optimal solution for the $\text{Percentage}_{\text{Start}}$, this study employs the ideal point method (Yuan et al., 2019) to solve the established multi-objective optimization model. For the objective functions $F(1)$, $F(2)$, $F(3)$, the ideal values $\{F_1^*, F_2^*, F_3^*\}$ are first calculated. From Table 6, it is known that the ideal values for Precision, Recall, and F1-score are $\{1, 0.8235, 0.875\}$. Using different start threshold percentages, the general values for these three indicators are calculated, and the Euclidean distance from the general values to the ideal values is calculated, as shown in Equation 10:

$$E = \sqrt{\sum_{r=1}^3 (F_r - F_r^*)^2} \quad (10)$$

The solution that minimizes the value of E is considered the optimal solution for the $\text{Percentage}_{\text{Start}}$. Figure 6 displays the Precision, Recall, F1-score, and the Euclidean distance from the general values to the ideal values under different start threshold percentages. When the start threshold percentage is 4.5%, the value of E reaches its minimum, at 0.0594. Therefore, setting the start threshold percentage to 4.5% yields the most balanced result. As shown in Table 6, when the start threshold percentage and the end threshold percentage are set to 4.5% and 30%, respectively, the machine detects a total of 13 flood events, and all 13 detected events are among the 17 actual flood events, resulting in a precision rate of 100%. Additionally, there are 4 flood events that were not detected, giving a flood event detection rate of 76.47%.

3.6. Automatic Filtering of False Flood Events With the LDA Topic Model

To download Weibo data using 24 rainfall-related and 20 flood-related keywords, non-flood-related events may sometimes be included. For instance, when the keyword “collapse” is used, Weibo posts related to a major bridge collapse incident in a city might also be downloaded, making it challenging for the flood damage classification process to filter out such unrelated information accurately. Additionally, because flood damage classification has

inherent uncertainties, some damage information that appears frequently within a short period may create a false peak, which the flood event detection algorithm might mistakenly identify as a flood event.

Since each flood event is tagged with start and end times, it's crucial to analyze the relevant Weibo posts during that timeframe to understand what topics are being discussed. If these posts are not discussing heavy rainfall or flooding, the detected flood event might be a false positive. To address this, we use the LDA topic model to identify topics within the Weibo posts corresponding to each flood event, effectively filtering out false flood events. LDA, an unsupervised machine learning algorithm, can reveal underlying themes within large document sets or corpora and has been widely applied in social media research for tasks like topic detection and classification (Du et al., 2023; Resch et al., 2018) and public opinion dissemination (W. Wang et al., 2024). The gensim library (<https://radimrehurek.com/gensim/>), which includes the LDA module, is used in our algorithm and can be directly invoked.

In this study, we configured the LDA model to analyze only one topic for each flood event's Weibo posts, extracting the top 20 most frequently appearing words associated with that topic. By examining whether these 20 words include terms related to heavy rain and flooding, our algorithm can determine if a flood event is false. We established a set of six Chinese detection keywords: ["rain (雨)," "flood (洪)," "waterlogging (涝)," "inundated (淹)," "flood stage (汛)," and "water pooling (积水)"]. After the algorithm performs topic detection for each flood event, it automatically compares the 20 detected words with these six keywords. If none of the six keywords are found in the 20 words, the detected flood event is flagged as false.

4. Result and Analysis

4.1. Flood Event Detection Results for 370 Cities

We conducted flood event detection in 370 cities based on the start threshold and end threshold of 4.5% and 30% of the annual average interval time $\bar{t}_{\text{year}}$, respectively, and initially detected a total of 956 flood events. Then, the LDA topic model was applied to automatically filter these 956 flood events, ultimately removing 227 false flood events and retaining 729 real flood events. However, we still cannot confirm that the remaining 729 flood events are entirely accurate and that the 227 false flood events are 100% incorrect. The process of detecting and filtering flood events, especially using automatic systems like LDA topic model, involves inherent uncertainty. While the model may significantly reduce the number of false positives, there is always a possibility that some genuine flood events might be incorrectly discarded, or some false positives may still remain.

Since LDA generates a topic consisting of 20 words for each flood event, we can easily identify what people are talking about from these 20 words, such as whether they are discussing floods, earthquakes, or something else. Most topics can be accurately identified, with only a few uncertain cases requiring verification of the corresponding Weibo posts for the event. This judgment is also very simple; it's not necessary to read all the Weibo posts corresponding to the event. Only a small number of posts are needed to make the determination. Ultimately, after validation, we found that among the 729 flood events retained, 100 were actually false positives, while 45 of the 227 discarded flood events were real. After applying the LDA topic model, the number of detected events dropped to 729, with 629 confirmed as genuine, resulting in a significantly higher detection accuracy of 86.28%. This demonstrates that the LDA model enhances the detection precision by effectively filtering out false flood events. In the end, after excluding all false flood events and reinstating the mistakenly removed real ones, we obtained a total of 674 verified flood events. The data set of validated 674 flood events can be downloaded from Shen et al. (2024f).

Among the 370 cities in China, 194 cities experienced flood disasters, accounting for 52.43%, indicating that nearly half of the cities in China are plagued by flood disasters. In terms of the number of flood events, 12 cities experienced more than 10 flood events, among which Xi'an and Chengdu suffered the most frequent flood disasters, with 24 and 15 flood events occurring over 12 years, respectively. Table 9 lists the top 20 cities in China with the most flood events, most of which are China's first-tier and second-tier cities, with 14 cities being provincial capitals or municipalities directly under the central government. These major cities have a higher susceptibility to floods, especially serious problems with urban waterlogging.

Figure 7 is a distribution map of flood events in various cities in China from 2012 to 2023. From the figure, it can be seen that flood disasters in China mainly occur in the central and southeastern parts. There are two main

Table 9
Top 20 Cities in China With the Highest Number of Flood Events

Ranking	Province	City	No. of flood events
1	Shannxi	Xi'an	24
2	Sichuan	Chengdu	15
3	Tianjin	Tianjin	14
4	Chongqing	Chongqing	14
5	Zhejiang	Ningbo	13
6	Beijing	Beijing	13
7	Hebei	Shijiazhuang	12
8	Henan	Zhengzhou	12
9	Guangdong	Shenzhen	12
10	Hainan	Haikou	12
11	Guangdong	Shanwei	11
12	Yunnan	Kunming	10
13	Guangdong	Zhuhai	9
14	Anhui	Hefei	8
15	Shandong	Qingdao	8
16	Guangdong	Guangzhou	8
17	Sichuan	Mianyang	8
18	Jiangsu	Nanjing	7
19	Fujian	Fuzhou	7
20	Fujian	Quanzhou	7

distribution intervals: one is the southeastern coastal cities, including many cities in Guangdong Province, Fujian Province, Jiangsu Province, and Zhejiang Province. Among the top 20 cities listed in Tables 9 and 8 cities are located in these four provinces. These cities are particularly susceptible to the impact of typhoons, for example, Shenzhen, Shanwei, and Quanzhou are almost annually affected by typhoons; the second is inland cities, mainly along Chengdu, Chongqing, Xi'an, and Zhengzhou. These cities are located at the dividing line between the north and south of China, where the climate changes drastically. Additionally, in China's arid and semi-arid regions such as Qinghai, Tibet, Xinjiang, Gansu, and Inner Mongolia, the number of flood events is significantly less, almost negligible, mainly characterized by sporadic flash floods.

As shown in Figure 8, this study statistically analyzed 674 flood events on a yearly basis. The number of flood events in 2020 and 2021 was higher than in other years, which were also the years with more extreme rainfall events in China. Due to the significantly lower volume of Weibo posts from 2012 to 2017 compared to 2018 to 2023, the detected flood events in these two periods also vary greatly. For example, the annual detected flood events from 2012 to 2015 were less than 10, while in 2023, the detected flood events reached 107, indicating that the scarce volume of Weibo posts in the early years had a certain impact on the total number of detected flood events. On the other hand, we also found that online news reports on early flood events were less frequent, with significantly fewer reports on flood events from 2012 to 2017 compared to 2018 to 2023 (Table 8). Nonetheless, our analysis of the detected early flood events revealed that many of these floods were historically significant major flood disasters. Taking 2012 as an example, a total of three flood events were detected in that year, including the severe rainstorm on 21 July 2012, in Beijing and Tianjin, where the rainfall in Beijing reached the

highest level since meteorological observations began in 1951. Another example is the severe rainstorm on July 9 in Shangqiu City, Henan Province, where the maximum local rainfall reached 540.8 mm, breaking the historical record since meteorological records began in 1953. This rainstorm affected over 400,000 people.

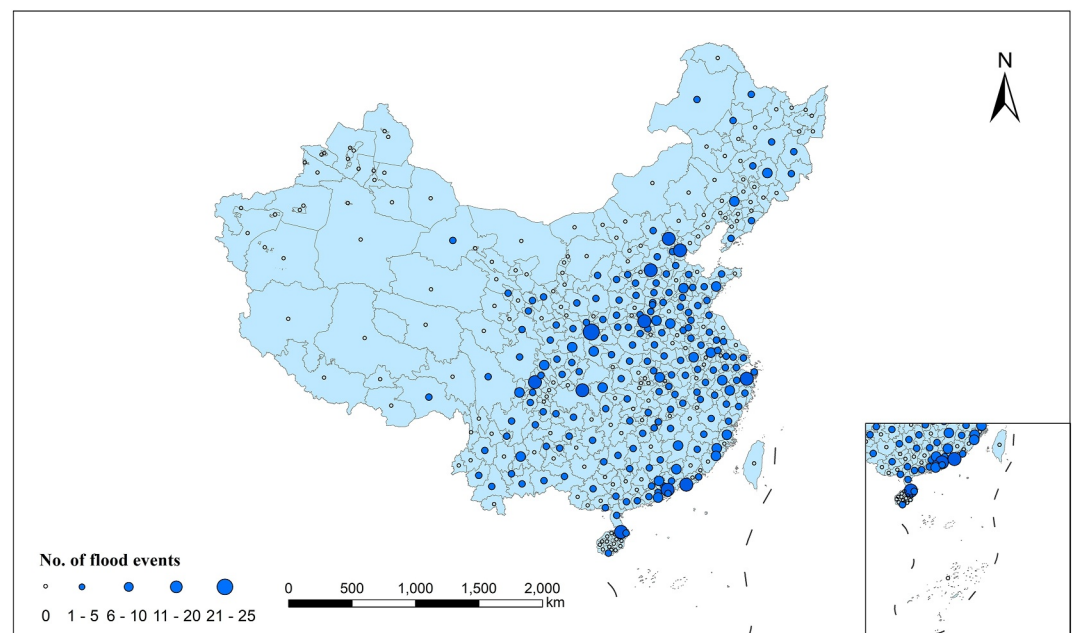


Figure 7. Distribution map of flood event occurrences in Chinese cities (2012–2023).

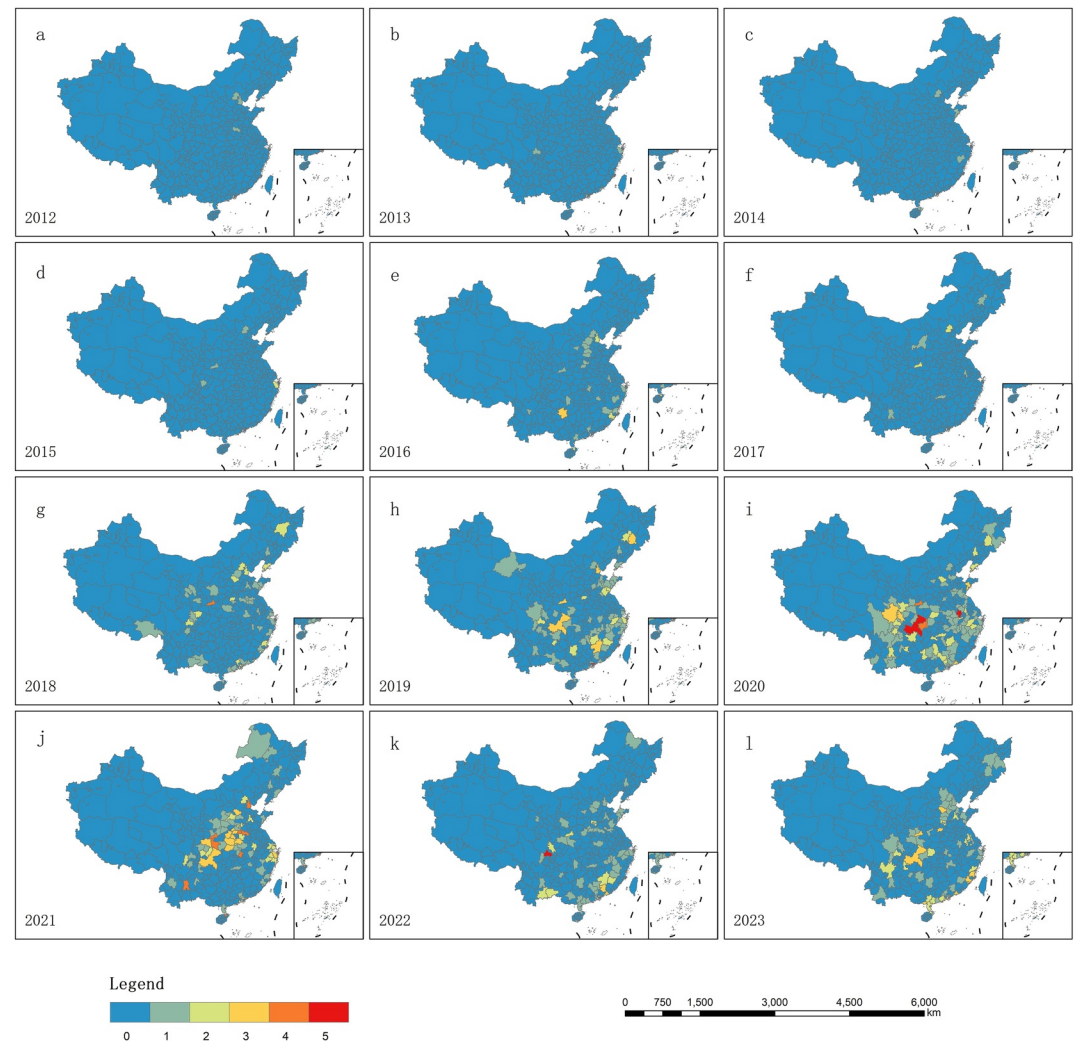


Figure 8. Distribution of annual flood event occurrences in Chinese cities from 2012 to 2023.



Figure 9. The disaster scene in Mentougou District, Beijing, after the heavy rain event on 31 July 2023 (Image source: https://www.thepaper.cn/newsDetail_forward_24054576).

Although the 674 flood events were only pinpointed at the city scale, a more detailed analysis of flood damage information for each event enables a finer-grained assessment of flood severity in different urban areas. We conducted a case study on an extreme rainstorm that occurred in Beijing on 31 July 2023 (Figure 9). From 8 p.m. on July 29 to 7 a.m. on August 2, the average rainfall in Beijing reached 331 mm, with 60% of the annual average rainfall occurring within 83 hr. The average rainfall in Mentougou District was 538.1 mm, and in Fangshan District, it was 598.7 mm. The maximum rainfall occurred in Wangjiayuan Reservoir in Changping District, reaching 744.8 mm, the highest recorded rainfall in Beijing area since instrumental measurements began 140 years ago. The maximum hourly precipitation intensity was 111.8 mm/hr at Qianling Mountain in Fengtai District (from 10 to 11 a.m. on July 31). On August 1, the rain intensity in Beijing significantly weakened, and the typhoon's impact was nearing its end. In the early morning of August 2, the thunderstorm clouds that affected Beijing weakened and moved out of the city.

As shown in Figure 10, we recorded flood damage information associated with “Beijing” and “Beijing districts” separately. Here, “Beijing districts”

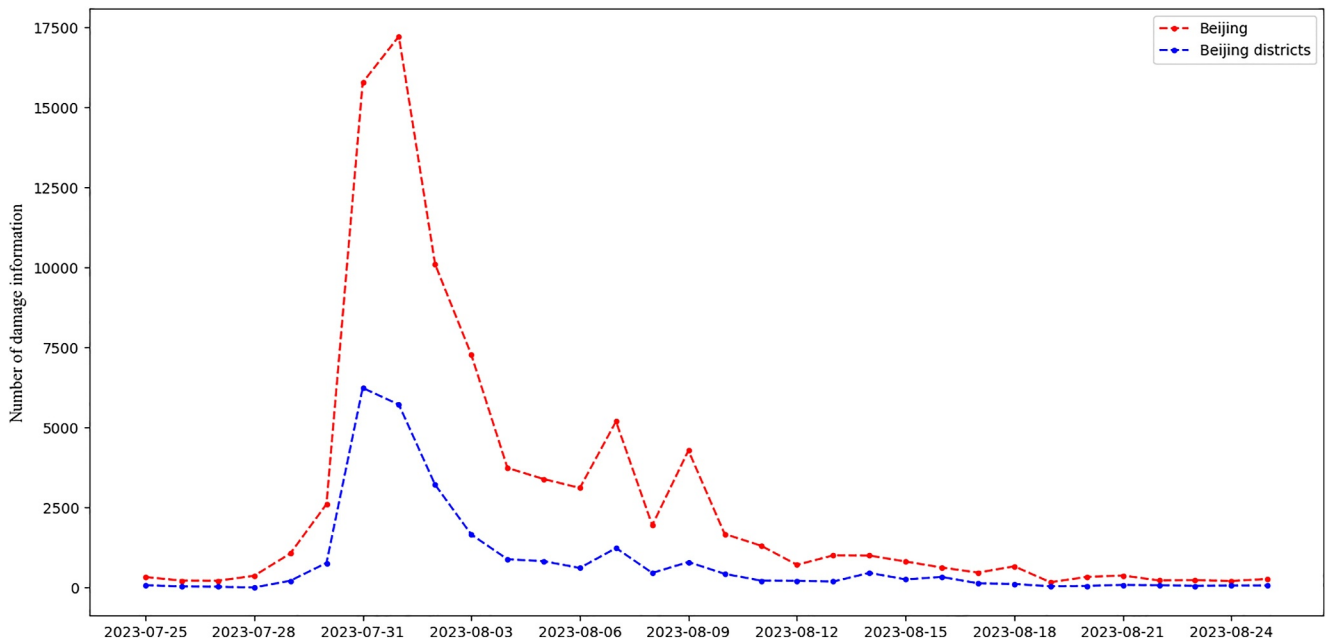


Figure 10. Daily variations in flood damage information in Beijing from 25 July 2023 to 25 August 2023.

refers to posts explicitly mentioning specific districts within Beijing, while “Beijing” refers to posts that could only be located to the city level without specific district details. Reports of flood damage in Beijing began to increase on July 29, surging dramatically with the heavy rainfall on July 31, resulting in 15,759 flood damage records for the city. This reached a peak of 17,201 records on August 1. By August 2 and 3, the flood damage

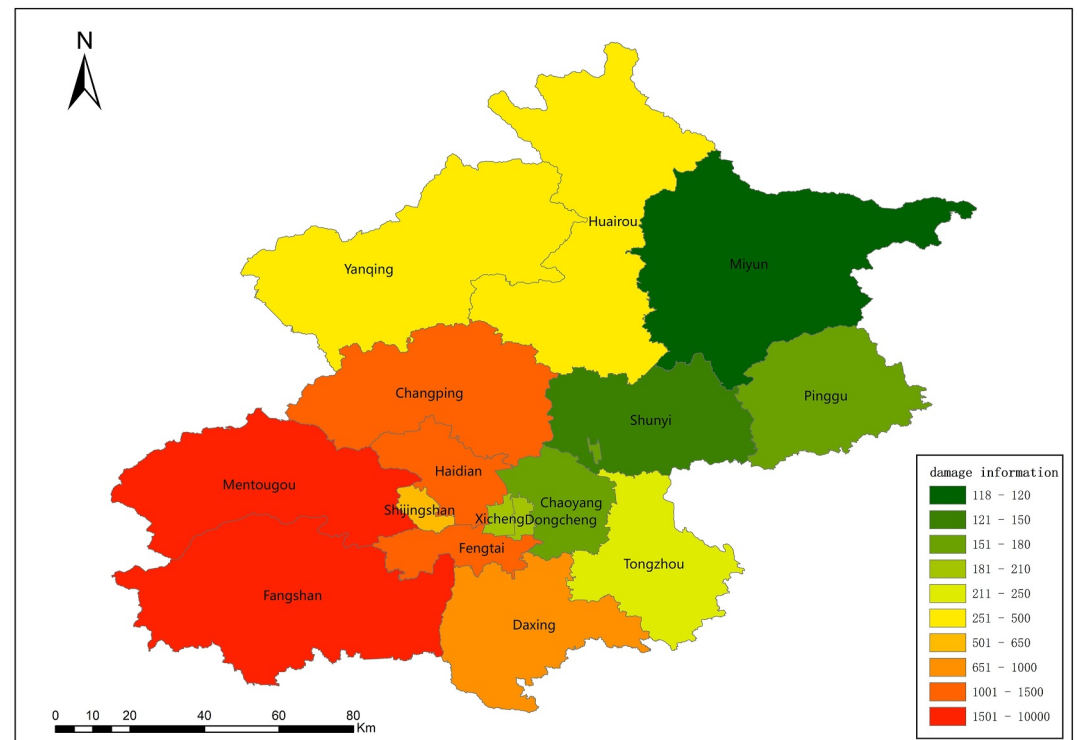


Figure 11. Statistics of flood damage information in various districts of Beijing during the heavy rainfall event on 31 July 2023.

Table 10
Detection Rate of Flood Events in 20 Key Cities

Index	City	No. of flood events (2012–2023)	No. of flood events detected by machine (2012–2023)	Detection rate (2012–2023)	No. of flood events (2018–2023)	No. of flood events detected by machine (2018–2023)	Detection rate (2018–2023)
1	Xi'an	25	20	80.00%	19	17	89.47%
2	Chengdu	14	13	92.85%	12	12	100%
3	Tianjin	11	10	90.9%	9	8	88.88%
4	Chongqing	28	12	42.85%	16	12	75%
5	Ningbo	13	11	84.61%	7	7	100%
6	Beijing	17	13	76.47%	8	7	87.5%
7	Shijiazhuang	14	11	78.57%	12	10	83.33%
8	Zhengzhou	8	8	100%	7	7	100%
9	Shenzhen	13	9	69.23%	9	9	100%
10	Haikou	15	10	66.66%	10	8	80%
11	Shanwei	22	11	50.00%	16	11	68.75%
12	Kunming	15	10	66.66%	12	9	75%
13	Zhuhai	6	3	50%	4	3	75%
14	Hefei	8	6	75%	6	6	100%
15	Qingdao	7	6	85.71%	6	5	83.33%
16	Guangzhou	14	8	57.14%	8	7	87.5%
17	Mianyang	12	8	66.66%	10	8	80%
18	Nanjing	10	6	60%	6	4	66.66%
19	Fuzhou	10	7	70%	7	6	85.71%
20	Quanzhou	7	5	71.42%	5	3	60%

records decreased to 10,091 and 7,270, respectively, and by August 18, the count dropped to 662, indicating that the flood event had completely subsided and public attention was fading. Flood damage records in the districts followed a similar trend, peaking on July 31 with 6,229 records and decreasing to 110 by August 18. Most remote sensing-based flood disaster monitoring methods require a time lag of 48–72 hr (Resch et al., 2018) to obtain flood disaster information. In contrast, social media platforms can almost monitor urban flood disasters in near real-time by frequently downloading flood-related textual data. This capability is of significant importance for disaster management.

Using the CPCA tool, flood damage information can be localized to various districts in Beijing, revealing the location of flood. As shown in Figure 11, Fangshan and Mentougou districts recorded 8,850 and 9,641 pieces of damage information, respectively, significantly higher than other districts, indicating that the majority of the heavy rainfall and flood damage occurred in these two areas. In contrast, the damage information from Miyun, Shunyi, Pinggu, Dongcheng, Xicheng, and Chaoyang districts was all below 200, suggesting these areas were almost unaffected, which aligns with reports from the media.

Furthermore, the flood event data set obtained in this study is expected to be refined to a street-level scale in future research. Many researchers have already started related work, such as using deep learning methods to detect all flood inundation points from each flood event (Liu et al., 2021), or extracting flood depth information from Weibo texts to create flood maps for urban waterlogging (Yan et al., 2024). This will provide a valuable supplement to remote sensing flood mapping research.

4.2. Comparative Validation With Online News Reports

To validate the accuracy of the flood events detected in this study, we selected 20 key cities from Table 9 and searched for news reports on flood disasters in these cities from 2012 to 2023 using the Baidu search engine. We search for flood events on Baidu using city names and different year/month keyword combinations, such as entering “Beijing Heavy Rain August 2023” or “Beijing Flood August 2023” to search for flood events in Beijing

Table 11*The 20 Flood Events That Occurred in Different Cities in China From 2019 to 2023 Recorded in the GDACS Database*

Index	Year	Location of the flood occurrence	Start time	End time	Has the machine detected it?
1	2023	Hong Kong, China	2023/9/07 0:00	2023/9/11 0:00	Yes
2	2023	Shenzhen, China	2023/9/07 0:00	2023/9/11 0:00	Yes
3	2023	Meizhou, China	2023/9/07 0:00	2023/9/11 0:00	No
4	2023	Xi'an, Shaanxi Province, China	2023/7/31 0:00	2023/8/15 0:00	Yes
5	2023	Hepu County, Beihai City, Guangxi Province, China	2023/6/01 0:00	2023/6/12 0:00	No
6	2023	Sanmenxia City, Henan Province, China	2023/1/23 0:00	2023/1/24 0:00	No
7	2022	Datong Hui and Tu Autonomous County, Longnan City, Gansu Province	2022/8/10 0:00	2022/8/19 0:00	No
8	2022	Xuwen County, Zhanjiang City, Guangdong Province	2022/8/9 0:00	2022/8/11 0:00	No
9	2022	Wudalianchi City, Daxing'anling Prefecture, Heilongjiang Province	2022/7/12 0:00	2022/7/13 0:00	No
10	2022	Gongshan Dulong and Nu Autonomous County, Nujiang Lisu Autonomous Prefecture, Yunnan Province	2022/4/2 0:00	2022/4/3 0:00	No
11	2022	Lianping County, Heyuan City, Guangdong Province	2022/5/26 0:00	2022/6/14 0:00	Yes
12	2022	Shaoguan City, Guangdong Province	2022/6/20 0:00	2022/6/26 0:00	Yes
13	2021	Qixian County, Jinzhong City, Shanxi Province	2021/9/24 0:00	2021/10/7 0:00	Yes
14	2021	Hanzhong City, Shaanxi Province	2021/9/24 0:00	2021/10/7 0:00	Yes
15	2021	Taiyuan City, Shanxi Province	2021/6/20 0:00	2021/6/30 0:00	No
16	2021	Yixian County, Baoding City, Hebei Province	2021/7/16 0:00	2021/7/19 0:00	No
17	2021	Wutaishan City, Shanxi Province	2021/7/2 0:00	2021/7/12 0:00	No
18	2020	Danba County, Garze Tibetan Autonomous Prefecture, Sichuan Province	2020/6/16 0:00	2020/6/17 0:00	Yes
19	2019	Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province	2019/7/31 0:00	2019/8/20 0:00	Yes
20	2019	Harbin City, Heilongjiang Province	2019/7/30 0:00	2019/7/30 0:00	No

in August 2023. Table 10 shows the detection rates of flood events for these 20 cities, where the detection rate refers to the proportion of flood events reported in online news that were also present in our flood data set. A higher detection rate indicates better performance of the algorithm in detecting flood events in that city. Considering the scarcity of Weibo posts from 2012 to 2017, this study adopted two statistical approaches: one calculates the detection rates for each city from 2012 to 2023, and the other calculates the detection rates from 2018 to 2023.

As shown in Table 10, the detection rate from 2018 to 2023 is significantly higher than that from 2012 to 2023. The median detection rate for the period 2012 to 2023 is 70%, while for 2018 to 2023, it is 83.33%. This indicates that, with the gradual increase in the quantity of Weibo posts in recent years, the flood event detection algorithm has become more efficient and accurate. Between 2012 and 2023, a total of 187 flood events occurred in these 20 cities, as identified through searching online news reports, with 159 of these events occurring between 2018 and 2023, accounting for 70.75%.

4.3. Comparative Validation With the Global Flood Disaster Database

Although there are many global natural disaster databases now, there is a serious problem of missing data in the records of flood events in China. A large number of actual flood events are not included in the database, and many flood events occur at the provincial level, making the location of the flood events relatively vague. In particular, urban waterlogging disasters have almost no records in multiple databases. The government's grasp of flood data is seriously inadequate, which hinders the management and decision-making of flood disasters.

1. GDACS Database. GDACS (2024) is a collaborative framework between the United Nations, the European Commission, and global disaster management agencies, aimed at improving early warning, information exchange, and coordination in the aftermath of major emergencies. Flood event alerts for different countries and regions can be directly accessed through the official GDACS website. The system has been recording flood disaster information from China since 2019. We have queried all flood records in the Chinese region from

Table 12

35 Flood Events Recorded in Different Cities in China From 2018 to 2023 in the EM-DAT Database

Index	Year	Location of the flood occurrence	Start time	End time	Has the machine detected it?
1	2023	Shenzhen City, Guangdong Province, China	2023/9/03 0:00	2023/9/08 0:00	Yes
2	2023	Xi'an City, Shaanxi Province, China	2023/8/11 0:00	2023/8/16 0:00	Yes
3	2023	Jilin City, Jilin Province, China	2023/8/01 0:00	2023/8/07 0:00	Yes
4	2023	Beijing City, China	2023/8/01 0:00	2023/8/09 0:00	Yes
5	2023	Leshan City, Sichuan Province, China	2023/6/01 0:00	2023/6/05 0:00	Yes
6	2023	Yiliang County, Zhaotong City, Yunnan Province, China	2023/6/01 0:00	2023/6/05 0:00	No
7	2022	Xining City, Qinghai Province, China	2022/8/17 0:00	2022/8/19 0:00	Yes
8	2022	Chengdu City, Sichuan Province, China	2022/8/17 0:00	2022/8/19 0:00	Yes
9	2022	Zhanjiang City, Guangdong Province, China	2022/8/10 0:00	2022/8/10 0:00	No
10	2021	Xiangyang City, Hubei Province, China	2021/8/12 0:00	2021/8/13 0:00	Yes
11	2021	Suizhou City, Hubei Province, China	2021/8/12 0:00	2021/8/13 0:00	Yes
12	2021	Xiaogan City, Hubei Province, China	2021/8/12 0:00	2021/8/13 0:00	Yes
13	2021	Zhengzhou City, Henan Province, China	2021/6/01 0:00	2021/8/30 0:00	Yes
14	2021	Hebi City, Henan Province, China	2021/6/01 0:00	2021/8/30 0:00	Yes
15	2021	Anyang City, Henan Province, China	2021/6/01 0:00	2021/8/30 0:00	Yes
16	2021	Xinxiang City, Henan Province, China	2021/6/01 0:00	2021/8/30 0:00	Yes
17	2020	Chongqing City, China	2020/5/21 0:00	2020/7/30 0:00	Yes
18	2020	Zunyi City, Guizhou Province, China	2020/6/22 0:00	2020/6/24 0:00	Yes
19	2020	Huishui County, Qiannan Buyei and Miao Autonomous Prefecture, Guizhou Province, China	2020/6/22 0:00	2020/6/24 0:00	No
20	2020	Tongren City, Guizhou Province, China	2020/6/22 0:00	2020/6/24 0:00	No
21	2020	Qiandongnan Miao and Dong Autonomous Prefecture, Guizhou Province, China	2020/6/22 0:00	2020/6/24 0:00	No
22	2020	Shijiao Town, Qingcheng District, Qingyuan City, Guangdong Province, China	2020/6/22 0:00	2020/6/24 0:00	No
23	2020	Chongqing City, China	2020/6/22 0:00	2020/6/24 0:00	Yes
24	2020	Mianning County, Liangshan Yi Autonomous Prefecture, Sichuan Province, China	2020/6/30 0:00	2020/7/05 0:00	Yes
25	2019	Enshi Tujia and Miao Autonomous Prefecture, Hubei Province, China	2019/8/04 0:00	2019/8/05 0:00	Yes
26	2019	Chongqing City, China	2019/6/13 0:00	2019/7/01 0:00	Yes
27	2019	Shenzhen City, Guangdong Province, China	2019/4/11 0:00	2019/4/12 0:00	Yes
28	2018	Shanghai City, China	2018/8/15 0:00	2018/8/17 0:00	No
29	2018	Tianshui City, Gansu Province, China	2018/7/10 0:00	2018/7/11 0:00	Yes
30	2018	Zhangye City, Gansu Province, China	2018/7/10 0:00	2018/7/11 0:00	No
31	2018	Pingliang City, Gansu Province, China	2018/7/10 0:00	2018/7/11 0:00	No
32	2018	Deyang City, Sichuan Province, China	2018/7/07 0:00	2018/7/07 0:00	Yes
33	2018	Mianyang City, Sichuan Province, China	2018/7/07 0:00	2018/7/07 0:00	Yes
34	2018	Guangyuan City, Sichuan Province, China	2018/7/07 0:00	2018/7/07 0:00	Yes
35	2018	Chongqing City, China	2018/5/05 0:00	2018/7/31 0:00	No

2019 to 2023, totaling 33 events. In these flood events, many of them are at the provincial level and do not provide specific city information. As shown in Table 11, we have organized the descriptive information of these flood events and identified 20 events that occurred in different cities. Among these 20 flood events, we have detected nine of them, accounting for 45%. The undetected flood events often belong to mountainous areas with debris flow disasters. For example, the 6th flood event was a debris flow disaster in Sanmenxia City, resulting in the death of two villagers. Overall, the GDACS database has serious gaps in recording flood events in China, especially those occurring in urban areas.

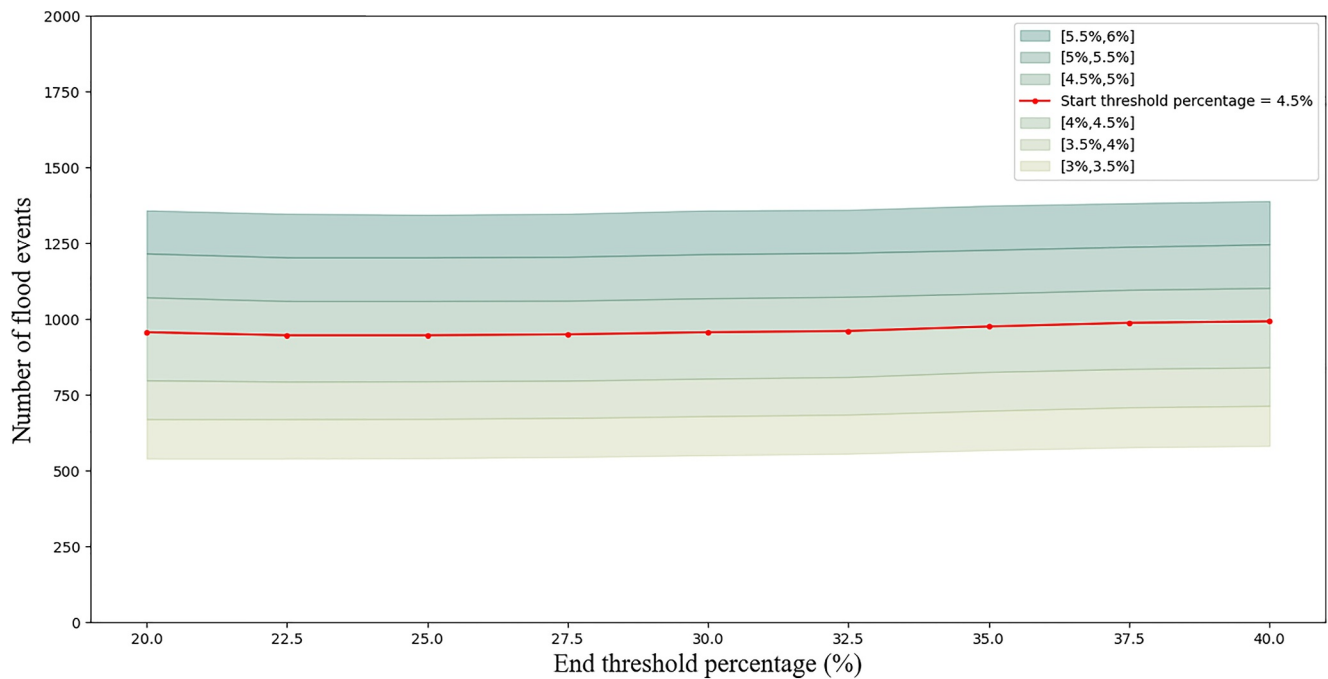


Figure 12. Number of flood events with different start and end threshold percentages.

2. EM-DAT Database. In 1988, the World Health Organization and the Center for Research on the Epidemiology of Disasters (CRED) jointly created the Emergency Events Database (EM-DAT, 2024), which is maintained by CRED. As a global disaster database, EM-DAT provides a wealth of data on natural and man-made disasters for international programs and scientific research, including data on major disaster events and their impacts worldwide since 1900. The database is updated daily, and new data is made publicly available within a month after verification. These data are compiled from various sources, including the United Nations, international organizations, governments, non-governmental organizations, insurance companies, research institutions, and the media.

We downloaded data on flood disaster events that occurred in China between 2018 and 2023 from the EM-DAT database. We excluded flood events that were only accurate to the provincial level and retained complete records of those events that clearly indicated the cities where the floods occurred. We considered flood events occurring within a 2-day interval as the same flood event. After data processing, we finally obtained information on 35 flood events as shown in Table 12. These flood events have clearly identified cities, start times, and end times. By comparing them with the flood event data we detected, we found that out of these 35 flood events, we detected 25 flood events, accounting for 71.42%.

By comparing Tables 11 and 12, it can be seen that both the GDACS database and the EM-DAT database recorded three flood events simultaneously. These events are the flood in Shenzhen in September 2023, the flood in Xi'an in August 2023, and the flood in Zhanjiang in August 2022. These global disaster databases generally operate on a larger scale and typically only record flood events at the provincial level. Records of flood events at the city level are severely incomplete. Additionally, different databases have different data standards, which leads to a lack of consistency between databases. This inconsistency makes comparison and use of the databases quite challenging.

4.4. Analysis of the Number of Flood Events With Different Threshold Settings

This study uses different start and end threshold percentages to conduct detection tests on flood events in China and to count the number of flood event occurrences. The start threshold percentages are set at seven different levels: [3%, 3.5%, 4%, 4.5%, 5%, 5.5%, 6%], and the end threshold percentages are set at nine different levels: [20%, 22.5%, 25%, 27.5%, 30%, 32.5%, 35%, 37.5%, 40%]. With the different combinations of start and end

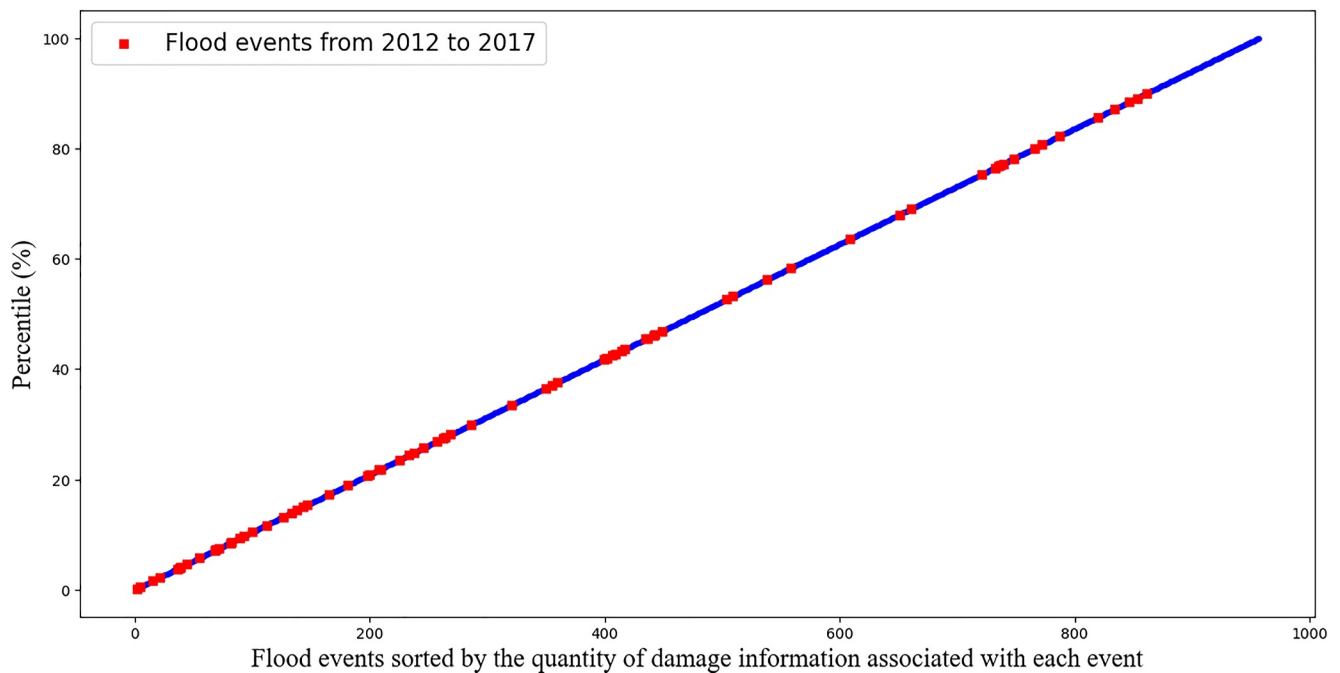


Figure 13. Percentile distribution of flood events from 2012 to 2017.

threshold percentages, a total of 63 sets are formed. The results of the flood event occurrences are shown in Figure 12.

In Figure 12, the red line represents the number of detected flood events with the start threshold percentage set at 4.5%, and the end threshold percentage gradually increasing from 20% to 40%. This number fluctuates between 946 and 992, indicating that the number of flood events is minimally affected by the end threshold. The lower edge of the polygonal frame in the figure represents the setting of the start threshold percentage at 3%, while the upper edge represents the setting of the start threshold percentage at 6%. When the start threshold percentage is set at 3%, the number of detected flood events ranges between 539 and 581, whereas when the start threshold percentage is set at 6%, the number of detected flood events ranges between 1,344 and 1,389. With a fixed end threshold percentage, different settings of the start threshold percentage lead to significant variations in the detected flood events. As the start threshold percentage increases from 3% to 6%, the number of flood events grows from over 500 to more than 1,300. A lower start threshold percentage results in a higher authenticity of detected flood events but may fail to detect many flood events. Conversely, a higher start threshold percentage, while detecting as many potential real flood events as possible, may also lead to a higher number of false flood events being detected. This becomes a burden for the LDA topic model, as the proportion of false flood events that need to be filtered increases.

It is evident that the start threshold percentage needs to be carefully set. This study has generalized the multi-objective optimization results of flood event detection in Beijing to 370 cities nationwide. A comparative analysis with flood events reported in online news indicates that the accuracy of flood event detection in this study is commendable. In subsequent work, this study will attempt to continue adding flood event information for multiple cities. By increasing the number of flood occurrences and refining the multi-objective optimization process, it aims to achieve more precise settings of the start threshold percentage and more accurately capture flood events.

As shown in Table 5, in earlier years, due to the limited popularity of Weibo, the volume of Weibo data was relatively small, and the intervals between Weibo posts were longer. With the widespread adoption of smartphones and the dramatic increase in the number of Weibo users, the volume of Weibo data has shown explosive growth in recent years, leading to higher accuracy in flood detection. Nevertheless, we analyzed the results of flood event detection from 2012 to 2023 and calculated the volume of damage information during each flood event. We found that the difference in the volume of damage information during the duration of flood events

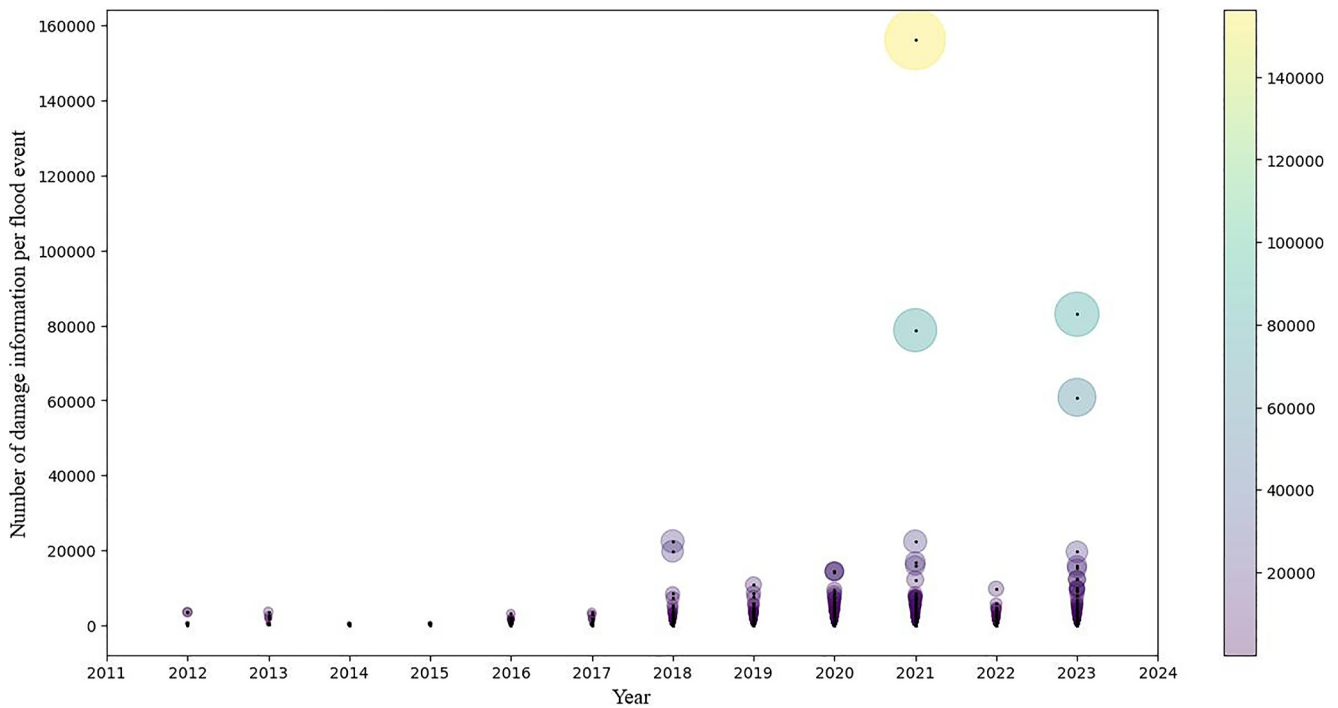


Figure 14. The amount of damage information for each flood event from 2012 to 2023.

between early years (2012–2017) and later years (2018–2023) was not significant. Since we have categorized the damage information by city and organized it according to the release time, we can calculate the total volume of damage information during a flood event in a particular city based on its start and end times. As illustrated in Figure 13, this study ranked flood events according to the quantity of damage information and calculated the percentile for each flood event. It was observed that the distribution of percentiles for flood events between 2012 and 2017 was relatively balanced, without a concentration at lower percentiles. Therefore, for early flood events, as long as they could be detected, the volume of damage information was also substantial. For those missed flood events, it is more likely that the event did not become a sufficiently popular topic on the Sina Weibo platform, rather than a lack of detection capability in the algorithm.

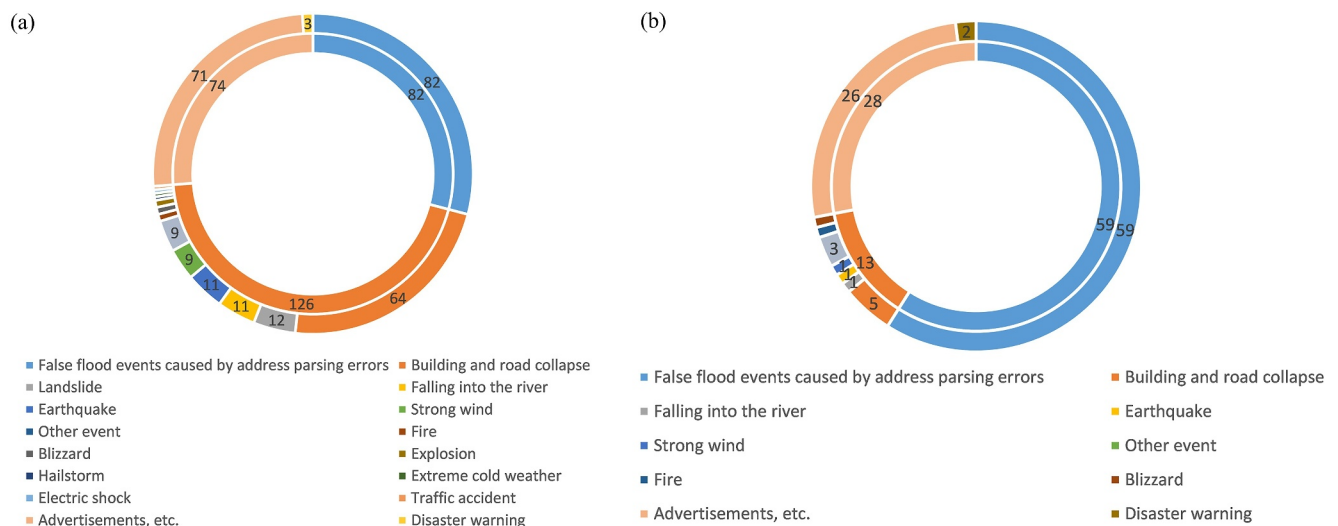


Figure 15. Changes in the number of different categories of false flood events before and after filtering with the LDA topic model: (a) 282 false flood events before filtering; (b) 100 false flood events after filtering.

Figure 14 shows the difference in the volume of damage information for each flood event from 2012 to 2023, with the size of the circles representing the quantity of damage information contained in each flood event. Except for a few extreme flood events in 2021 and 2023, the volume of damage information for most flood events is in the hundreds. However, for nationally known major flood events, the volume of damage information in the early years is far less than in recent years. For example, the rainstorm event in Beijing in July 2012 had a damage information volume of 3,577, while nearly 10 years later, in 2021, the major rainstorm event in Henan Province, with its provincial capital city of Zhengzhou, reached an astonishing volume of over 150,000 damage information entries (indicated by the yellow circle).

4.5. Analysis of LDA Topic Model's Effectiveness in Automatically Filtering False Flood Events

Among the 956 flood events initially detected by the algorithm, 674 are real flood events, while the remaining 282 are false. As shown in Figure 15a, we classified the 282 false flood events into three categories: the first category consists of flood events that were redundantly detected due to address parsing errors; the second category includes various other emergency events; and the third category consists of events that are neither flood events nor other emergencies and are unrelated to any incidents.

We found a total of 82 false flood events caused by city parsing errors. Due to the existence of some duplicate city and district names in China, the CPCA library (DQinyuan, 2021) struggles to distinguish these cases, which can result in Weibo posts being assigned to two different cities. A flood event detected in one city may create a “copy” in another city. For example, when the post contains the characters “河南平顶山,” the CPCA may parse out “平顶山” as a city, but it also concurrently identifies “南平” as another city, which can interfere with the flood detection algorithm. Additionally, we found that 126 false flood events were primarily other emergency incidents. The most common type of event included various collapses, such as building collapses, ground subsidence, and bridge failures, totaling 64 events. Other natural disasters, such as landslides, earthquakes, and strong winds, each accounted for about 10 events. The third category is completely unrelated to any incidents and primarily consists of false detections caused by frequent posts about various advertisements and celebrity concert announcements that contain keywords matching “damage information.” In this study, only three false flood events were triggered by flash flood and geological disaster warning messages, indicating that using damage information can effectively mitigate the interference from various disaster warning messages.

As shown in Figure 15b, after automatically filtering with the LDA topic model, the three categories of false flood events were reduced to varying extents. The filtering was most effective for “other emergency incidents,” reducing their count from 126 to 13, as these events were easier to exclude due to the absence of actual heavy rainfall on the day in question. Additionally, the model achieved a significant reduction in false alerts caused by non-event-related bursts, decreasing from 74 to 28. However, the hardest to eliminate were false flood events caused by address parsing errors, with a reduction from 82 to 59. Ultimately, after applying the LDA topic model, 59% of the 100 remaining false flood events were still due to address parsing errors.

5. Discussion and Conclusion

5.1. Limitations of the Study

Although the proposed algorithm demonstrates promising performance in detecting flood events using Sina Weibo data, several limitations must be acknowledged. Firstly, the algorithm heavily relies on data from Weibo, a social media platform predominantly used in China. This restricts its applicability to flood detection within China, as the platform's user base, geographical coverage, and linguistic context are inherently tied to the region. Consequently, the model's generalizability to other countries or regions is limited, as cultural, linguistic, and social media usage patterns may differ significantly. For example, the keyword-matching strategy employed in this study is tailored to Chinese language patterns, which may not be directly applicable to other languages such as English, where syntactic structures and vocabulary usage differ substantially. This makes it challenging to directly adapt the algorithm to platforms like Twitter or Facebook. Additionally, the algorithm's performance may be influenced by biases inherent in Weibo data, including uneven geographical representation and platform-specific user behavior. These factors could hinder the model's ability to accurately detect flood events in areas with limited Weibo activity or where alternative social media platforms dominate. Future work could explore adapting the algorithm to other platforms and languages, potentially leveraging multilingual natural language

processing techniques to enhance its global applicability. Addressing these limitations would further strengthen the model's robustness and broaden its potential impact.

5.2. Potential Improvements to the Methodology

The flood event detection method proposed in this study still has substantial room for improvement, particularly in reducing the detection of false flood events. Possible solutions include:

1. Use large language models for place names and address parsing. The CPCA lacks sufficient analysis and understanding of text context, which is a common limitation of all non-large language models. When the CPCA performs city name resolution, it functions more like keyword matching and does not effectively comprehend what the text is describing. Advanced large language models, such as ChatGPT and ChatGLM3 (2024), can enhance the accuracy of place name and address parsing by reading contextual information within the text. Utilizing large language models for place name and address parsing should effectively eliminate this type of false flood event.
2. Incorporate city weather information into the flood event detection algorithm for filtering false flood events. We observed that those non-flood-related emergency incidents, as well as various non-disaster-related noise detected in the data, are mostly unrelated to rainfall and can be effectively filtered using city weather information to assist the flood detection algorithm. For example, once an event is detected, we can immediately check whether there was any rainfall in the city during the week prior to the event date. If there was no rainfall, the flood event can be directly eliminated.

These enhancements could significantly improve the accuracy and reliability of the flood event detection method.

5.3. Future Research Directions

Due to the fact that social media data is mainly used to detect floods through changes in the number of Weibo posts, some errors are still difficult to avoid. In future work, it is necessary to combine other data sources to further improve the accuracy of flood event detection. For example, combining satellite precipitation data to correct flood detection results through rainstorm movement, or using SAR and other satellite remote sensing data for cross-validation, can help filter out some erroneous detection results. Additionally, with the rapid development of AI large models such as DeepSeek, ChatGPT, Janus, and other multimodal large models, there is significant potential to leverage these advanced technologies to enhance flood event detection capabilities. Specifically, future research could explore using AI large models to extract flood-related scenarios from social media texts and images, thereby improving the utilization of social media data and enhancing the perception of flood events. These models could provide deeper insights into flood dynamics by analyzing both textual and visual content, offering a more comprehensive understanding of flood impacts.

In summary, this research achievement allows us to examine flood disasters in China from another perspective and gain a deeper understanding of the problems faced by various cities in dealing with flood disasters. How to improve urban resilience and coexist with floods is a challenge that every city needs to face.

Data Availability Statement

The annual Sina Weibo data is publicly accessible. Weibo posts on heavy rain and floods from 2012 to 2023 can be found in Shen et al. (2024a). Additionally, the Weibo data for the Henan flood event can be accessed in Shen et al. (2024b). The flood damage information keywords can be accessed in Shen et al. (2024c). The flood event detection results at different thresholds can be found in Shen et al. (2024d). The flood event detection code is available in Shen et al. (2024e). The data set of detected and validated 674 flood events can be downloaded from Shen et al. (2024f). The GDACS flood event data are available in GDACS (2024). The EM-DAT flood event data are available in EM-DAT (2024). The CPCA library can be found in DQinyuan (2021).

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grants 42077438, 42371425) and the Fundamental Research Funds for the Central Universities (Grants CCNU24JC029, CCNU22QN019).

References

- Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLoS One*, 13(1), e0189327. <https://doi.org/10.1371/journal.pone.0189327>
- Bhabaraju, S. K. T., Beyney, C., & Nicholson, C. (2019). Quantitative analysis of social media sensitivity to natural disasters. *International Journal of Disaster Risk Reduction*, 39, 101251. <https://doi.org/10.1016/j.ijdrr.2019.101251>

- Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., et al. (2019). Changing climate both increases and decreases European river floods. *Nature*, 573(7772), 108–111. <https://doi.org/10.1038/s41586-019-1495-6>
- ChatGLM3. (2024). [Software]. Retrieved from <https://github.com/THUDM/ChatGLM3>
- de Bruijn, J. A., de Moel, H., Jongman, B., de Ruiter, M. C., Wagemaker, J., & Aerts, J. C. J. H. (2019). A global database of historic and real-time flood events based on social media. *Scientific Data*, 6(1), 311. <https://doi.org/10.1038/s41597-019-0326-9>
- DFO (Dartmouth Flood Observatory). (2024). [Dataset]. Retrieved from <https://floodobservatory.colorado.edu/>
- DQinyuan. (2021). chinese_province_city_area_mapper: Built to be recognize Chinese province, city and area in simplified Chinese string, it can automaticall map area to city and map city to province, 2021 [Software]. Retrieved from <https://pypi.org/project/cpcpa/>
- Du, W., Ge, C., Yao, S., Chen, N., & Xu, L. (2023). Applicability analysis and ensemble application of BERT with TF-IDF, TextRank, MMR, and LDA for topic classification based on flood-related VGI. *ISPRS International Journal of Geo-Information*, 12(6), 240. <https://doi.org/10.3390/ijgi12060240>
- EM-DAT (Emergency Events Database). (2024). [Dataset]. Retrieved from <https://www.emdat.be/>
- Fu, S., Lyu, H., Wang, Z., Hao, X., & Zhang, C. (2022). Extracting historical flood locations from news media data by the named entity recognition (NER) model to assess urban susceptibility. *Journal of Hydrology*, 612, 128312. <https://doi.org/10.1016/j.jhydrol.2022.128312>
- GDACS (Global Disaster Alert and Coordination System). (2024). [Dataset]. Retrieved from <https://www.gdacs.org/Alerts/default.aspx>
- Guzman, J., & Poblete, B. (2013). On-line relevant anomaly detection in the Twitter stream. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description ODD'* (Vol. 13, pp. 31–39). <https://doi.org/10.1145/2500853.2500860>
- He, J., Ma, M., Zhou, Y., & Wang, M. (2023). What we have learned about the characteristics and differences of disaster information behavior in social media—A case study of the 7.20 Henan heavy rain flood disaster. *Sustainability*, 15(6), 4726. <https://doi.org/10.3390/su15064726>
- Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Climate Change*, 3(9), 816–821. <https://doi.org/10.1038/nclimate1911>
- Islam, A., Ghosh, S., Barman, S. D., Nandy, S., & Sarkar, B. (2022). Role of in-situ and ex-situ livelihood strategies for flood risk reduction: Evidence from the Mayurakshi River Basin, India. *International Journal of Disaster Risk Reduction*, 70, 102775. <https://doi.org/10.1016/j.ijdrr.2021.102775>
- Islam, M. T., & Meng, Q. (2022). An exploratory study of Sentinel-1 SAR for rapid urban flood mapping on Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, 113, 103002.
- James, N. A., Kejariwal, A., & Matteson, D. S. (2016). Leveraging cloud data to mitigate user experience from “breaking bad”. In *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA (Vol. 2016, pp. 3499–3508). <https://doi.org/10.1109/bigdata.2016.7841013>
- Jongman, B., Wagemaker, J., Romero, B. R., & de Perez, E. C. (2015). Early flood detection for rapid humanitarian response: Harnessing near real-time satellite and twitter signals. *ISPRS International Journal of Geo-Information*, 4, 2246–2266. <https://doi.org/10.3390/ijgi4042246>
- Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., & Kamruzzaman, M. (2020). Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets. *International Journal of Disaster Risk Reduction*, 42, 101360. <https://doi.org/10.1016/j.ijdrr.2019.101360>
- Kellermann, P., Schröter, K., Thieken, A. H., Haubrock, S.-N., & Kreibich, H. (2020). The object-specific flood damage database HOWAS 21. *Natural Hazards and Earth System Sciences*, 20(9), 2503–2519. <https://doi.org/10.5194/nhess-20-2503-2020>
- Kharazi, B. A., & Behzadan, A. H. (2021). Flood depth mapping in street photos with image processing and deep neural networks. *Computers, Environment and Urban Systems*, 88, 101628. <https://doi.org/10.1016/j.compenvurbsys.2021.101628>
- Kleinberg, J. B. (2003). Hierarchical structure in streams. *Knowledge Discovery & Data Mining*, 7(4), 373–397.
- Kontokosta, C. E., & Malik, A. (2018). The Resilience to Emergencies and Disasters Index: Applying big data to benchmark and validate neighborhood resilience capacity. *Sustainable Cities and Society*, 36, 272–285. <https://doi.org/10.1016/j.scs.2017.10.025>
- Li, Z., Wang, C., Emrich, C. T., & Guo, D. (2017). A novel approach to leveraging social media for rapid flood mapping: A case study of the 2015 South Carolina floods. *Cartography and Geographic Information Science*, 45(2), 1–14. <https://doi.org/10.1080/15230406.2016.1271356>
- Lin, X., & Wu, S. (2022). Typhoon disaster network emotion analysis method based on semantic rules and word vector. *Journal of Geo-information Science*, 24(1), 114–126.
- Linderson, S., Raffetti, E., Rusca, M., Brandimarte, L., Mard, J., & Di Baldassarre, G. (2023). The wider the gap between rich and poor the higher the flood mortality. *Nature Sustainability*, 6(8), 995–1005. <https://doi.org/10.1038/s41893-023-01107-7>
- Liu, H., Hao, Y., Zhang, W., Gao, F., & Tong, J. (2021). Online urban-waterlogging monitoring based on a recurrent neural network for classification of microblogging text. *Natural Hazards and Earth System Sciences*, 21(4), 1179–1194. <https://doi.org/10.5194/nhess-21-1179-2021>
- Manandhar, B., Cui, S., Wang, L., & Shrestha, S. (2023). Post-flood resilience assessment of July 2021 flood in Western Germany and Henan, China. *Land*, 12(3), 625. <https://doi.org/10.3390/land12030625>
- MEM & ME (Ministry of Emergency Management & Ministry of Education). (2022). 2021 global natural disaster assessment report, 13 October, 2022, China. Retrieved from <https://www.gddat.cn/WorldInfoSystem/production/BNU/2021-EN.pdf>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Computer Science*, 1–12.
- Olthof, I., & Svčina, N. (2020). Testing urban flood mapping approaches from satellite and in-situ data collected during 2017 and 2019 events in eastern Canada. *Remote Sensing*, 12(19), 3141. <https://doi.org/10.3390/rs12193141>
- Resch, B., Uslander, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 45(4), 362–376. <https://doi.org/10.1080/15230406.2017.1356242>
- Riley, W. J. (2008). Algorithms for frequency jump detection. *Metrologia*, 45(6), S154–S161. <https://doi.org/10.1088/0026-1394/45/6/s21>
- Rossi, C., Acerbo, F. S., Ylinen, K., Juga, I., Nurmi, P., Bosca, A., et al. (2018). Early detection and information extraction for weather-induced floods using social media streams. *International Journal of Disaster Risk Reduction*, 30, 145–157. <https://doi.org/10.1016/j.ijdrr.2018.03.002>
- Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., & Zhang, S. (2024a). Weibo texts on heavy rain and floods from 2012 to 2023 [Dataset]. *figshare*. <https://doi.org/10.6084/m9.figshare.26117794.v1>
- Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., & Zhang, S. (2024b). Weibo text data for the Henan flood event [Dataset]. *figshare*. <https://doi.org/10.6084/m9.figshare.25366453.v1>
- Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., & Zhang, S. (2024c). Damage information keywords [Dataset]. *figshare*. <https://doi.org/10.6084/m9.figshare.26084737.v1>
- Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., & Zhang, S. (2024d). The flood event detection results at different thresholds [Dataset]. *figshare*. <https://doi.org/10.6084/m9.figshare.25366645.v1>
- Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., & Zhang, S. (2024e). The flood event detection code [Software]. *figshare*. <https://doi.org/10.6084/m9.figshare.28201499>

- Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., & Zhang, S. (2024f). The dataset of detected and validated 674 flood events [Dataset]. *figshare*. <https://doi.org/10.6084/m9.figshare.27718746>
- Shoyama, K., Cui, Q., Hanashima, M., Sano, H., & Usuda, Y. (2021). Emergency flood detection using multiple information sources: Integrated analysis of natural hazard monitoring and social media data. *Science of the Total Environment*, 767, 144371. <https://doi.org/10.1016/j.scitotenv.2020.144371>
- Tan, L., & Schultz, D. M. (2021). Damage classification and recovery analysis of the Chongqing, China, floods of August 2020 based on social-media data. *Journal of Cleaner Production*, 333, 127882. <https://doi.org/10.1016/j.jclepro.2021.127882>
- Tartakovsky, A. G., & Moustakides, G. V. (2010). State-of-the-art in Bayesian changepoint detection. *Sequential Analysis*, 29(2), 125–145. <https://doi.org/10.1080/07474941003740997>
- Tellman, B., Sullivan, J. A., Kuhn, C., Kettner, A. J., Doyle, C. S., Brakenridge, G. R., et al. (2021). Satellite imaging reveals increased proportion of population exposed floods. *Nature*, 596(7870), 80–86. <https://doi.org/10.1038/s41586-021-03695-w>
- Theja Bhavaraju, S. K., Beyney, C., & Nicholson, C. (2019). Quantitative analysis of social media sensitivity to natural disasters. *International Journal of Disaster Risk Reduction*, 39, 101251. <https://doi.org/10.1016/j.ijdr.2019.101251>
- United Nations Office for Disaster Risk Reduction (UNDRR). (2015). *Sendai framework for disaster risk reduction 2015–2030* (Vol. 6). Retrieved from https://www.unisdr.org/files/43291_chinesesendaiframeworkfordisasterri.pdf
- Wang, W., Zhu, X., Lu, P., Zhao, Y., Chen, Y., & Zhang, S. (2024). Spatio-temporal evolution of public opinion on urban flooding: Case study of the 7.20 Henan extreme flood event. *International Journal of Disaster Risk Reduction*, 100, 104175. <https://doi.org/10.1016/j.ijdr.2023.104175>
- Wang, X., Zhu, F., Jiang, J., & Li, S. (2013). Real time event detection in twitter. In *International Conference on Web-age Information Management* (pp. 502–513).
- Wang, Y., Yang, S., Zhang, L., Cao, Y., & Yin, Y. (2022). Comparative analysis and outlook of three global databases for meteorological disasters. *Climate Change Research*, 18(2), 253–260.
- Wu, W., Li, J., He, Z., Ye, X., Zhang, J., Gao, X., & Qu, H. (2020). Tracking spatio-temporal variation of geo-tagged topics with social media in China: A case study of 2016 Hefei rainstorm. *International Journal of Disaster Risk Reduction*, 50, 101737. <https://doi.org/10.1016/j.ijdr.2020.101737>
- Yan, Z., Guo, X., Zhao, Z., & Tang, L. (2024). Achieving fine-grained urban flood perception and spatio-temporal evolution analysis based on social media. *Sustainable Cities and Society*, 101, 105077. <https://doi.org/10.1016/j.scs.2023.105077>
- Yang, T., Xie, J., Li, Z., & Li, G. (2018). A method of typhoon disaster loss identification and classification using micro-blog information. *Journal of Geo-information Science*, 20(7), 906–917.
- Yang, Y., Yin, J., Wang, D., Liu, Y., Lu, Y., Zhang, W., & Xu, S. (2023). ABM-based emergency evacuation modelling during urban pluvial floods: A “7.20” pluvial flood event study in Zhengzhou, Henan Province. *Science China Earth Sciences*, 66(2), 282–291. <https://doi.org/10.1007/s11430-022-1015-6>
- Yuan, Y., Yang, X., Chen, L., Yuan, X., Dong, H., & Yu, Y. (2019). Optimization of the basin hydrologic network based on multi-objective criteria. *Journal of Hohai University (Natural Sciences)*, 47(2), 102–107.
- Zhang, H., Qi, Z., Li, X., Chen, Y., Wang, X., & He, Y. (2021). An urban flooding index for unsupervised inundated urban area detection using Sentinel-1 polarimetric SAR images. *Remote Sensing*, 13(22), 4511. <https://doi.org/10.3390/rs13224511>
- Zhou, B., Zou, A., Lin, B., Yang, M., Gharaibeh, N., Cai, H., et al. (2022). VictimFinder: Harvesting rescue requests in disaster response from social media with BERT. *Computers, Environment and Urban Systems*, 95, 101824. <https://doi.org/10.1016/j.compenvurbsys.2022.101824>