Dear Reviewer,

Thank you very much for your time involved in reviewing the manuscript and providing valuable feedback. Those comments are constructive for revising and improving our manuscript. We have taken the time to think through all of your comments and will carefully revise the manuscript to address each comment:

General Comment 1. Flood Query Keywords

The flood query was limited to "flood" and "flood disasters" (L142, L154), while many other terms could hint at flood events in news items, e.g., "typhoon," "cyclone," "mud," "heavy rainfall," "inundated areas,"… Query terms are an essential aspect of event detection and this could be seen as a restriction limiting the detection power of the proposed approach. It raises some questions: Should this be documented as a limitation? Is it a decision to limit the size of the corpus? Does the Q&A approach prevent that concern?

Thanks for bringing up this important point. However, the other keywords included may raise the dataset too large. For example, we tried using “heavy rainfall” as the query term and found that only around 10% news returned reported flood events. Most of these news texts are related to meteorological early warning information. Therefore, the current query was determined to limit the corpus to the most relevant content. Even if the Q&A approach can distinguish between relevant and irrelevant information, the benefits of large corpus are far less than the burden of running the model.

General Comment 2. Flood Types and Multi-Hazard Concerns

The paper focuses on urban floods, excluding other types of floods, yet flood types are interrelated and very often not mutually exclusive. Hence, referring, for instance, to the Hazard Information Profiles (HIPs, https://www.preventionweb.net/drr-glossary/hips), an urban flood could also be related to a flash flood (despite the exclusion of the query of "flash flood," L151), a riverine flood, a coastal flood, a groundwater flood. Floods are also secondary hazards associated with other hazards, such as a flood that could result from a Typhoon, heavy rainfall, a storm surge, an intense monsoon etc. Floods are also associated with geo-hazards such as landfall (See GLC studies). I found the Typhoon case study in the paper interesting. It also illustrates the multi-hazard nature of floods well. As in GLC studies, I would be interested in having the authors' view on multi, cascading, and co-occurring type issues, the possibilities of detecting multi-type floods, and the challenges, limitations, and perspectives concerning their proposed approach.

Though we agree with this perspective, this article mainly focuses on urban flooding, especially its temporal and spatial information. There are two considerations regarding the reviewer's comment.

First, in future studies, it can be continuously mined as new contents in our database about whether it is transformed by other flood types, and its complex causes. We could extract multi-hazard information to add a column in the dataset to show what weather event caused the floods and a column to show the floods resulting in what geo-hazards such as landfall. We believe that the Q&A method can effectively identify the causal relationships between floods and other hazards only if news data can include this kind of information. For example, we manually checked 100 samples describing 52 events and 6 events mentioned that this flood caused by a typhoon. Therefore, the feasibility of disaster causality analysis based on news data needs to be further studied and confirmed. Our future research will also add other data sources to increase the data potential.

Second, some recent studies have used news media reports to extract information on various meteorological and geological disasters. However, most of these studies just classify news by rules rather than analyze the causality between disasters and did not subdivide flood into different flood types. For instance, Yang et al. (2023) applied a rule-based approach to extract 15 types of disaster information from news texts. Specifically, the rule implies that if any of these disaster names appear in the text, the news is categorized accordingly, and then the prefecture-level administrative names are used to match the location information in the news. Another example (Liu et al., 2018) also utilized keyword positioning and rule-based named entity recognition methods to identify disaster types and locations in the news. Both of above studies used this rule: if one news report mentions multiple disaster types at the same time, it is determined that the news event is multi-disaster co-occurrence. This method will introduce biases when a news just mentions two hazards but in different events that have no direct relationship. In the future, we could employ the language model to test the efficiency of extracting multi-type floods and other related hazards form news-based data, but the performance should be examined.

For these two reasons, addressing single- or multi-hazard information from the dataset is challenging and would require considerable thought to overcome these limitations.

This is a valuable point raised by the reviewer. So, in the revised version, we will add further discussion on this issue.

Yang, Chenchen, Han Zhang, Xunhua Li, Zongyi He, and Junli Li. "Analysis of spatial and temporal characteristics of major natural disasters in China from 2008 to 2021 based on mining news database." Natural Hazards 118, no. 3 (2023): 1881-1916.

Liu, Xiao, Haixiang Guo, Yu-ru Lin, Yijing Li, and Jundong Hou. "Analyzing spatial-temporal distribution of natural hazards in China by mining news sources." Natural Hazards Review 19, no. 3 (2018): 04018006.

General Comment 3. A More Balanced Discussion: Trend Analyses vs. Gap Filling Potential

The manuscript extensively discusses spatiotemporal trend analysis, necessitating more caution and clarity on trends influencing factors. I understand the need to illustrate trends in the resulting dataset, but, in my opinion, this matter could be more efficiently summarized, and the paper could be more descriptive and less assertive in the interpretation. Some analyses are simplistic and do not go deep enough. Rather than make the paper even longer, I invite the authors to distinguish more between the essential and the accessory and, if anticipated, to cover in greater depth the spatiotemporal analysis of events and cross-referencing with third-party data in other papers (see GLC studies).

Some figures may be grouped, e.g., maps in different pannels of one figure, allowing not only to focus on the trends of the output data but also on how the output data compares to other datasets, which is currently limited to Figure 4, despite the numerous datasets being listed in the introduction. The reader has little clue as to what gap is being filled. In particular, the Chinese bulletin appears as a more exhaustive dataset (although coarser). This point may be worth further discussion.

Note regarding temporal trends:

Trends in hazard occurrences are complex, influenced by variations in hazard intensity and alteration of environmental susceptibility, as well as demographic shifts that alter exposure or vulnerability. Moreover, climatic cycles (e.g., ENSO or other climate indices) can distort linear trend estimations over brief periods due to their cyclical nature.

The complexity is further compounded when analyzing trends from news data. Changes in reporting capacity, especially in remote areas, along with new communication technologies like satellite and social media, may introduce significant biases. The proliferation of the internet during the 1990s and 2000s has notably impacted flood event reporting (Gall et al., 2009; Kron et al., 2012; Delforge et al., 2023). Kron et al., 2012 illustrate well the challenges in building a hazard database with flood examples. These works underscore the necessity for standardized flood event definitions to mitigate discrepancies in reporting scales. In the case of news scraping, the framing by journalists can significantly alter the perceived frequency, spatial representation, and the type of events.

In conclusion, the total number of flood events is a highly relative figure. It is essential to acknowledge that while flood hazards are natural phenomena, flood disasters and their reporting are social phenomena with potentially distinct and diverging trend patterns. Given these complexities, attributing trends depicted in the news (i.e., social variables, not physical ones) to climate change or land use changes requires careful consideration.

We greatly appreciate the reviewer's detailed and insightful feedback. Your comments are invaluable in refining our analysis and ensuring our conclusions are both accurate and impactful. In response to your comments mentioned above, we have taken the following considerations:

First, regarding distinguishing more between the essential and the accessory, we will focus on highlighting the characteristics of the spatial distribution while streamlining the discussion of temporal trends, particularly simplifying the analysis of the influence of natural factors. In addition, our study focuses on urban floods, and the fundamental data is derived from news reports, which have a strong social dimension. Therefore, it is necessary to analyze the flood trend in different population density and economically developed areas to provide conclusions from an urban and social perspective. We will include this information in the revised version, with a detailed explanation provided in the latter part of this response.

Second, cross-referencing with third-party data in other papers or comparing to other datasets is challenging because of the absence of proper data. Therefore, we can only find some relevant data for comparison in certain regions. We have created a line chart for reference (Figure 1 below), to analyze the correlation of the direct economic losses provided by the Guangxi Provincial Government website due to floods after 2016, and the scale of disaster represented by the number of news-extracted flood-affected counties. These two indicators exhibit relatively consistent trends, which can to some extent suggest that the coverage of news data in certain regions is fairly good. However, these two indicators do not represent the same physical quantity, we think this figure may not suit for inclusion in the main text but can be provided as supplementary material for reference.

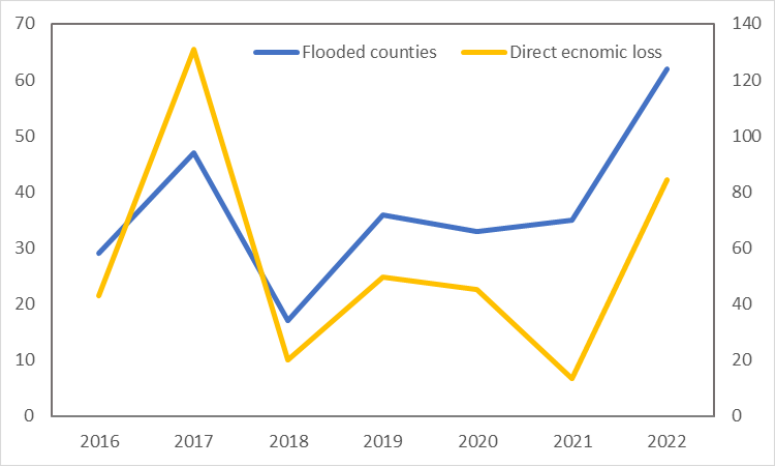


Figure 1. The time series of the number of news-extracted flooded counties and direct economic loss in Guangxi from 2016-2022.

Third, regrading what gap we have filled, it should be explained first that the *China Flood and Drought Bulletin* only provides the number of flooded cities in a general overview paragraph, without presenting their spatial distribution or specific inventory. The spatial distribution of flood loss information in the bulletin is limited to the province level, which encompasses multiple city-level areas. While our dataset is not comprehensive, it is the first county-level dataset on a national scale, and its time trends are largely consistent with authoritative data.

As for the temporal analysis, we agree that there are inherent limitations to using media data for temporal analysis. Overall, we will make the following adjustments in the revision:

In Section Temporal distribution of flood events and the relevant part of other sections, we will revise our statements on the temporal trends to reduce subjective interpretations and to clarify the limitations of news media data：

The temporal distribution of urban flood events in our dataset reveals an overall increasing trend over time. While this may reflect broader patterns of environmental change, such as the increase in extreme rainfall events driven by global warming and the effects of rapid urbanization, these trends should not be interpreted in isolation. Media data, which forms the basis of our dataset, is subject to various biases, as introduced by previous studies such as Gall et al. (2009) and Kron et al. (2012).

From the perspective of media communication studies, agenda-setting theory posits that by choosing which events to report on, the media effectively signals to the public which issues are important (Leidecker-Sandmann et al., 2023). Through the quantity and depth of coverage, the media can shape the level of public attention given to certain events. In the context of disaster reporting, the

government may influence the direction of media coverage to control public attention on specific disasters (Bai, 2022). For example, during the COVID-19 pandemic, research on government crisis communication showed that media agenda-setting was significantly influenced by government press conferences (Hayek, 2024). Crisis communication theory further explains the government can swiftly steer public opinion in the aftermath of a disaster, reducing the spread of negative emotions and maintaining social stability (Zhou et al., 2023). As a result, the variability in disaster reporting by the media may be influenced by multiple factors, including government policies, public interest, and the media's own resource allocation, leading to a situation where the volume of media reports is not necessarily consistent with the actual number of disaster events.

In Section Spatial distribution of flood events and relevant part in Section Discussion, we will streamline the results from different sub-regional analyses and group Figure 10, 11, and 12 into one figure as following picture (Figure 2) according to your suggestion:

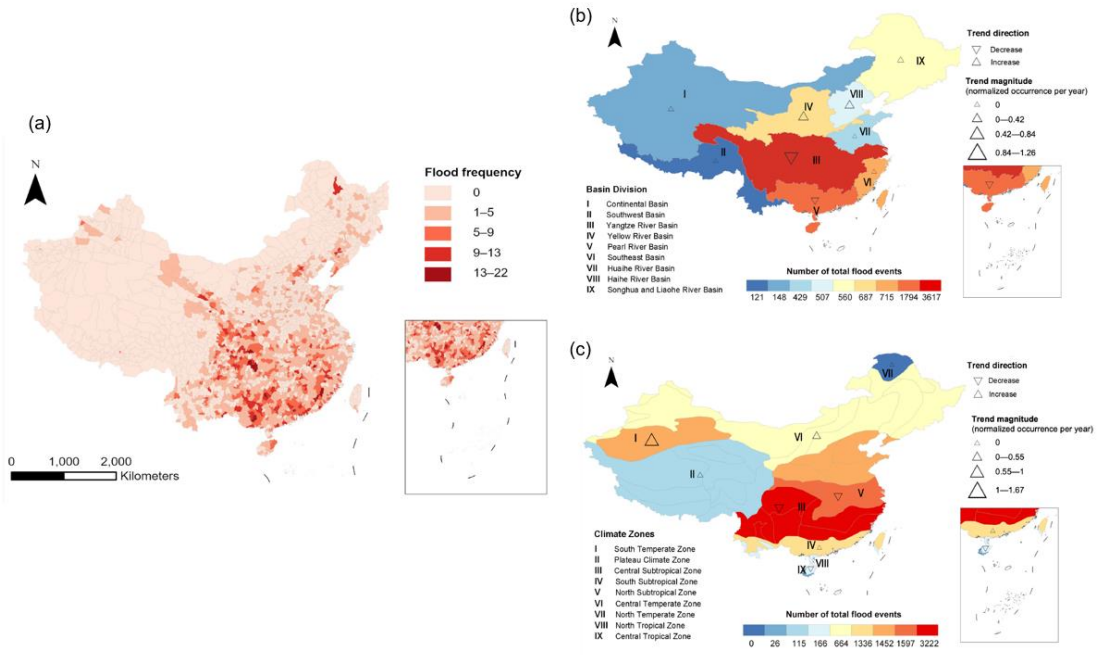


Figure 2. The spatial distribution of flood occurrence.

Furthermore, additional analysis on population density and Gross Regional Product (GRP) will be included as follows:

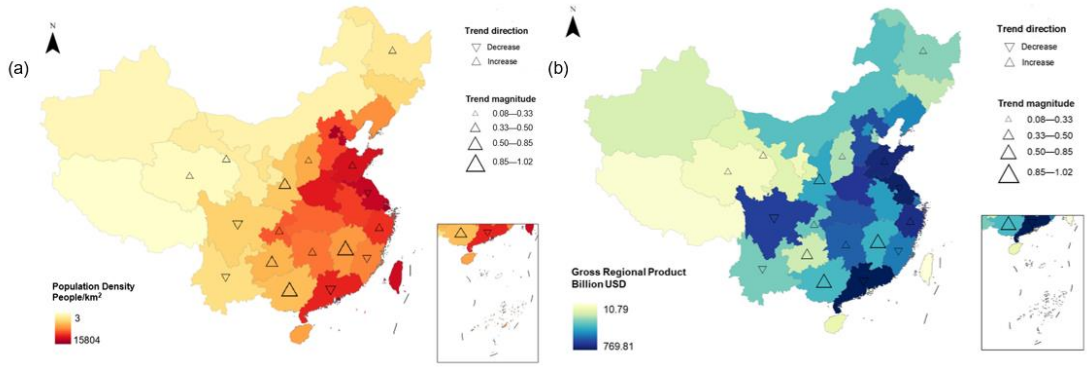


Figure 3. The analysis of flood event trends across Chinese provinces from 2000 to 2022, shown in relation to (a) population density and (b) Gross Regional Product (GRP).

The background maps display average annual Gross Regional Product (GRP) in billion USD and population density in people per square kilometer, respectively, with darker shades indicating higher values. Overlaid on these maps are Theil-Sen estimated trends for the number of flood events, where the direction of the triangle represents whether the trend is increasing or decreasing, and the size of the triangle corresponds to the magnitude of the trend. Provinces without a significant trend are not marked.

Overall, most provinces exhibit an increasing trend in flood events, particularly in the northern, and western regions of China. These areas, including provinces such as Heilongjiang, Shandong, and Chongqing, are characterized by varying levels population density, both higher and lower, according to Figure 3(a). The provinces that exhibit a decreasing trend in flood events are primarily located in the central and southeastern regions, particularly in provinces like Jiangsu, Fujian, and Guangdong, which are notable for their higher population densities. This suggests that the rising flood events are not strictly tied to population density.

As for the trends in relation to economic output in Figure 3(b), the provinces with increasing flood trends are mostly those with lower to moderate GRP, such as those in the northern and western parts of China, despite Shandong and Zhejiang. These regions may not have received the same level of economic investment in flood control infrastructure as the more developed eastern provinces, which might explain the rising trend in flood events. On the other hand, the central and eastern provinces showing a decreasing trend, such as Jiangsu, Guangdong, and Sichuan, are among the most economically developed in China. This suggests that the availability of economic resources has allowed for more comprehensive flood management strategies, reducing the frequency of flood events in these areas.

It is important to note that several provinces with high population densities and significant economic development, specifically Jiangsu and Guangdong, exhibit a decreasing trend in flood events. These regions have experienced a high number of flood events over these years, with a notable peak around 2010. The estimated decrease in flood trends may be related to this peak, where the number of flood events was significantly higher than in other years, possibly skewing the trend calculations downward. Additionally, as regions frequently affected by flooding and characterized by high economic output and population density, substantial investments in flood management infrastructure and policies may have been made, also contributing to the observed decline in flood events. Jia et al. (2022) have highlighted the significant investments in flood management infrastructure in China's economically developed regions. They compared the 1998 and 2020 floods in Yangtze River Basin regions, which are economically developed regions in China. Their analysis reveals that significant improvements in risk management, including engineering defenses, environmental recovery, forecasting and early warning, and emergency response have led to a substantial reduction in flood disaster losses in Yangtze River Basin regions.

Gall, M., Borden, K. A., and Cutter, S. L.: When Do Losses Count?: Six Fallacies of Natural Hazards Loss Data, Bulletin of the American Meteorological Society, 90, 799–810, https://doi.org/10.1175/2008BAMS2721.1, 2009.

Kron, W., Steuer, M., Löw, P., and Wirtz, A.: How to deal properly with a natural catastrophe database – analysis of flood losses, Natural Hazards and Earth System Sciences, 12, 535–550, https://doi.org/10.5194/nhess-12-535-2012, 2012.

Leidecker-Sandmann, Melanie & Koppers, Lars & Lehmkuhl, Markus. (2023). Correlations between the selection of topics by news media and scientific journals. PloS one. 18. e0280016. 10.1371/journal.pone.0280016.

Bai, Sheng. Mainstream Media Agenda Setting in Disaster Events. Journal of Emergency Management and Disaster Communications 3, no. 2 (2022): 83-98. https://doi.org/10.1142/S2689980922500038.

Hayek, Lore. (2024). Media Framing of Government Crisis Communication During Covid-19. Media and Communication. 12. 10.17645/mac.7774.

Zhou S, Yu W, Tang X, Li X. Government crisis communication innovation and its psychological intervention coupling: Based on an analysis of China's provincial COVID-19 outbreak updates. Front Psychol. 2023 Jan 26;13:1008948. doi: 10.3389/fpsyg.2022.1008948. PMID: 36778169; PMCID: PMC9909028.

Jia, Huicong & Chen, Fang & Pan, Donghua & Du, Enyu & Wang, Lei & Wang, Ning & Yang, Aqiang. (2021). Flood risk management in the Yangtze River basin

—Comparison of 1998 and 2020 events. International Journal of Disaster Risk Reduction. 68. 102724. 10.1016/j.ijdrr.2021.102724.

General Comment 4. Analyses of GDP

The manuscript highlights the GDP as the primary driver of media attention. However, the boxes in Figure 5 do not seem to show any significant difference between the occurrence of floods for different GDP groups. So, to highlight a possible effect of GDP on media attention, it is vital to use GDP per capita (see GLC studies).

The population is a critical factor in media attention and hazard exposure. More densely populated cities should receive more media attention in the event of a flood. It is likely the primary factor explaining the spatial patterns in the dataset. It is likely to be correlated with GDP, as well as other factors such as elevation, distance to river or coast, or climate (see G5). Therefore, controlling that factor when investigating some effects is essential.

We agree with your perspective. Our initial motivation for conducting the GDP clustering analysis was to explain how regional economic development might influence the biases in media data. However, after carefully considering the reviewers' comments and reviewing literature on media communication themes, we have decided to remove this section. Relying solely on economic development or population density to explain the biases in media data is not convincing enough. In the revised version, we will modify our explanation of the biases introduced by media data as mentioned in the response to G3.

Moreover, we will add the analysis of the flood trend in different population density and economically developed areas as mentioned in the response to G3.

General Comment 5. Analyses of Flood Susceptibility

Figure 7 and the underlying analysis of flood susceptibility present some issues and do not bring much to the paper. The proposed pattern is not very neat (the points also overlap with no transparency), likely because the chosen indicators are quite remote proxies of flood susceptibility and should not be presented as acknowledged indicators in hydrology (the supporting references are weak).

Average daily precipitation depicts a hydrological equilibrium rather than an extreme event. Naturally, arid regions are less susceptible (also less populated, hence, exposed). However, the indicator becomes less relevant to other hydrological systems with higher precipitation averages (a mixture of blue and red dots). Likewise, elevated areas are also likely to be less populated and then less exposed, and the elevation effect tends to disappear at a lower elevation.

Flow accumulation or topographical wetness indices could have been more reliable indicators of flood susceptibility.

I would recommend removing this analysis given its low informative value and also because these variables are related to climate variability, which is already pictured in Figure 12. See GLC studies for comparisons.

Thank you for pointing out the issue with the selection of flood susceptibility factors. We agree that the factors initially chosen were not appropriate. Average rainfall reflects the general characteristics of a region, but flood disasters are often associated with extreme rainfall. Additionally, discussing the impact of elevation alone is not convincing given the large extent of the study area. We will remove this section in the revision.

General Comment 6. Flood Events Dataset Resolution

While the final dataset is reported at the county-month level, the reader is left with little insight into the level of detail directly resulting from the information extraction process, which remains unclearly described. Based on Figures 4 and 6, it appears that information at the city-daily level was collected. It seems that a much more precise dataset could have been shared without much additional effort, raising questions about the motivation behind disaggregating the data to such a coarser level.

We are sorry that our description may confuse readers especially the term "county". First, we think the administrative level in China should be introduced:

The provincial level is the highest level of administrative division in China, and it consists of: Provinces, Autonomous Regions, Municipalities, Special administrative Regions (Hong Kong and Macau); The second level is prefectural level including: Prefecture-level Cities (just cities in the usual sense), Autonomous Prefectures, Leagues (found in Inner Mongolia); The third level is county level including: Counties, County-level Cities (smaller cities under the jurisdiction of a prefecture-level city), Districts, Banners (found in Inner Mongolia); The forth level is township level including: Towns, Townships (typically more rural areas), Subdistricts; And the last level is village level including: Villages, Communities.

Therefore, a county is a finer administrative division than a city, with one city typically comprising several county-level areas.

The locations extracted from news reports typically include only the county-level area name or the county name with the specific flooded street or building. Therefore, we standardized the spatial information by using county names.

Second, most of the data can be extracted to specific day information, but some can only be extracted to month, so at first, in order to unify the data set, we set the time resolution as month. In the revised version, we will change the events with day information to be accurate to day.

As for the figures you mentioned, Figure 6 is indeed the flood events with daily information within two typhoon event months. However, in Figure 4, we used a line plot just to show the temporal trends of news-reported flooded cities amount and those reported in bulletins. The data is aggregated annually rather than daily.

General Comment 7. Data Content, FAIR Principles, and Reusability

Also, given that a central outcome of the paper is a dataset, alignment with FAIR principles (https://www.go-fair.org/) should be particularly encouraged. Regarding the data shared, GitHub is not considered FAIR as it does not allow for persistent identifiers. Also, a few additional data could greatly increase the reusability of the dataset, e.g., precise column descriptions in the readme, the reference for the administrative unit shapefile to link the data with the post-code or administrative units as described in the paper (L275-278), using international time standards, and possibly translate region names to English to maximize reuse in the global context.

Regarding reproducibility, the data and code availability section could be improved. Input news data and their conditions of (re-)use are not described in this section. Tools and libraries being used to develop the approach are not referred to (except references to the Python "Re" module at L187). There is no comment about whether or not the developed models are accessible and under which conditions of use.

There are no links or references to the news articles that have been used to construct the dataset. Sharing the links could drastically increase the paper's outreach and support future research and NLP applications to extract additional information, such as flood impact variables or associated hazard types, without redeveloping an NLP flood event detection model. Annotated corpora are also valuable datasets in the context of NLP for future benchmarking. Consider commenting on that dataset as well.

Thanks for your helpful suggestions. We will change the dataset sharing website to Zenodo, which is an open-access repository that allows researchers to share and preserve their datasets. It is operated by CERN and OpenAIRE and provides features like DOIs for citations, which supports the FAIR principles. Furthermore, we will add a column describing post-code and precise column descriptions in the readme, and translate region names to English. As for the administrative unit shapefile to link the data with the post-code, we will add the reference in Section Data availability.

About the input news data, we will check the link of the news platform and share it in Section Data availability. Then, readers can retrieve the news using the query as we described in Section News data and download the data according to the data management rules of the WiseNews platform. We can share the annotated flood-related corpora in our open-access dataset.

We will refer to the tools and libraries being used to develop the approach, such as Tensorflow in the revised version. Besides, we could provide the developed model to the readers who contact to us.

Specific Comment 1. *L8: "similar" could be more nuanced.*

We will re-organize this sentence to describe the comparison between news-based floods and bulletin data:
Our analysis reveals that while there are notable differences in the magnitude reported events in peak years, the temporal trend of flooded cities in the news-based dataset broadly aligns with that in the *China Flood and Drought Bulletin*.

Specific Comment 2. *L9:10: "the connection between…": the connection does not support accuracy and the analysis is oversimplistic (See G5).*

We agree that the flood susceptibility indicators were insufficiently appropriate. As response to Comment G5, we will remove this analysis in the revision.

Specific Comment 3. *L43 (and after): "natural disaster" is a controversial terminology often avoided by Disaster Risk experts, acknowledging that a disaster is not natural (as opposed to natural hazards).*

We will correct it to 'natural hazard' in the revised version of the manuscript.

Specific Comment 4. *L43-L52: Table 2 could distinguish between catalogs from remote and social sensing, e.g., that DFO is based on remote sensing, EM-DAT on the collection of text documents and manual extraction of the information. Some missing recent initiatives could be worth mentioning, e.g., a global remote sensing catalog is the global flood database and a global catalog obtained from social media:*

- *Tellman, B., Sullivan, J.A., Kuhn, C. et al. Satellite imaging reveals increased proportion of population exposed to floods. Nature 596, 80–86 (2021). https://doi.org/10.1038/s41586-021-03695-w*
- *J.A. de Bruijn, H. de Moel, B. Jongman, M.C. de Ruiter, J. Wagemaker, J.C.J.H. Aerts. A global database of historic and real-time flood events based on social media. Scientific Data, 6 (1) (2019), p. 311, 10.1038/s41597-019-0326-9*

- *G.R. Brakenridge. Global Active Archive of Large Flood Events. Dartmouth Flood Observatory, University of Colorado, USA. http://floodobservatory.colorado.edu/ Archives/ (Accessed xxx)*
- *Delforge, D., Wathelet, V., Below, R., Lanfredi Sofia, C., Tonnelier, M., Loenhout, J. van, and Speybroeck, N.: EM-DAT: the Emergency Events Database, preprint, https://doi.org/10.21203/rs.3.rs-3807553/v1, 2023.*

We will add a column to distinguish between databases from remote and social sensing and the recent datasets in Table 2 as follows:

	Name	**Period**	**Flood Records**	**Update Frequency**	**Source**
Social sensing	The Emergency Events Database (EM-DAT)	1900--	Time, location and damage of global flood events that resulted in a certain number of deaths or economic losses	Continuously	Centre for Research on the Epidemiology of Disasters
	Natural Disaster Data Book	2002--	Statistical and analytical perspectives of flood events in Asia (data retrieved from EM-DAT)	Annual	Asian Disaster Reduction Center
	Global Flood Monitor	2014-2023	A real-time overview of ongoing flood events based on filtered Twitter data	Pause	IVM - VU University Amsterdam and FloodTags
	Floodlist	2016--	Dates, locations, magnitude and damages of each flood events based on news	Real-time	FloodList (funding from Copernicus)
Remote sensing	Dartmouth Flood Observatory (DFO)	1985--	Time, location and extent of global flood events using satellite observations	Continuously	University of Colorado Boulder
	Global Flood Awareness System	Real-time	Ongoing and upcoming flood events information from satellites to support flood forecasting at national, regional and global levels	Real-time	Copernicus Emergency Management Service (CEMS)

Global Flood Monitoring System (GFMS)	Real-time	Flood inundation extent and depth based on precipitation satellite data and flood model simulation	Every 3 hours	University of Maryland and NASA
The Global Flood Database	2000-2018	flood extent and population exposure for 913 large flood events	unknown	Floodbase

Specific Comment 5. *L65: Beyond cloud cover for optical imagery, mapping urban flood is challenging per se.*

It is true that mapping urban flooding is inherently challenging, and we just listed one source of uncertainty. In the revised version, we will modify this sentence to emphasize that mapping urban flooding is already a technical challenge per se.

Specific Comment 6. *L75: "Yang et al. (2023)" Such a paper of high relevance should be re-discussed later in the discussion section, among others, to identify (see Overview).*

Thank you for the helpful suggestion. As the response to G2, we will add the discussion on multiple hazards, particularly those related to flooding, and reviewing these highly relevant papers.

Specific Comment 7. *L77: The authors acknowledge the multi-hazard nature of floods here and after, but the issue is not discussed in light of their own work (see G2).*

We will add the discussion on the multi-hazard nature of floods in the revised version (see the response to G2)

Specific Comment 8. *L90: "Conditional Random Fields (CRF) layer" appears to be a central part of the methodology appearing multiple times in the paper; however, it lacks a clear explanation of what it is and why it is used.*

Sorry for this unclear statement. CRF model is a type of discriminative probabilistic model used to predict sequences of labels for sequences of input samples. It considers the context (i.e., neighboring labels) to make more accurate predictions. The CRF layer was part of the named entity recognition (NER) method in our approach.

We used BERT to extract initial answers including spatiotemporal information of floods and then, adopted an NER method called BiLSTM-CRF model to identify the location names in the answers. In the NER model, a BiLSTM layer is adopted to extract features from the input character vectors. And then, the CRF layer uses

the output from BiLSTM to compute the most likely sequence of labels considering the dependencies between labels.

Specific Comment 9. *L110:116: since the paper follows a conventional structure, it is unnecessary to detail it in the introduction.*

We will remove these explanations in revised version according to your suggestion.

Specific Comment 10. *Table 2: EM-DAT is continuously updated (see Delforge et al., 2023). I would also refer to the Global Flood Awareness System (https://global-flood.emergency.copernicus.eu/), the flood component of CEMS, instead of CEMS. See also S4.*

We will update Table 2 according to your comments (See response to S4).

Specific Comment 11. *L134: check url link (404 error).*

We will check this issue and re-share the data link (https://www.wisers.com/wisesearch)

Specific Comment 12. *Figure 1: I appreciate the availability of an example. However, consider selecting a more topic-appropriate example or asking for a where/when the question for more relevance.*

We will provide a more topic-appropriate example in the revised version, such as followings:

Context: Faced with a once-in-50-years "7·23" catastrophic flood, the city committee and government treated the disaster as a command and time as life, mobilizing the entire city, resettling the affected people, and actively conducting post-disaster epidemic prevention. In this event, many houses in Luzhou's Linjiang were submerged, and at least 100,000 to 200,000 people needed to be relocated, which is unprecedented in Luzhou's recent decades of history.

Question1: What disaster event occurred?

Answer1: "7·23" catastrophic flood.

Question2: When did it happen?

Answer2: "7·23".

Question3: Where did the disaster occur?

Answer3: Luzhou.

Specific Comment 13. *L142, L151, and L154: See G1.*

We could add sentences to explain our query term determination if needed:

Our study focuses on the mining of flood events, although other meteorology-related terms such as "typhoon," "cyclone," "heavy rainfall," may be related to flood events, but there are very few flood event news only mentioning flood-causing terms like typhoon. At the same time, we examined the results of a separate query of “heavy rain”, and only 10% were reported flood events, most of which were meteorological warnings. Therefore, in order to control the relevance of corpus and improve the efficiency of model, this study limited the current search terms.

Specific Comment 14. *L145-148: The description of the data and its processing, including test/train split, may be confusing. It may be more appropriate to move to the method section.*

We will move the description of the data processing to the method section.

Specific Comment 15. *L157: "Validation" unless China Flood and Drought Bulletin is considered a gold standard, I think referring to comparative data and cross-comparison instead of validation is more appropriate.*

We agree with you. We will replace “Validation” in revised version according to your suggestion.

Specific Comment 16. *L168-L174: oversimplistic view of hydrology and weak references. See G5.*

We will remove this part.

Specific Comment 17. *L190-199: This section could indicate the total/train/test sample sizes more clearly.*

Sorry for unclear explanation. The total of the CNKI news samples was 633, and these samples were divided into three parts: 402 samples for fine-tuning BERT (alongside with CMRC2018); 101 samples as validation set for adjusting hyperparameters; 130 samples for testing. We will re-organize the data-processing part in method section and remove current related descriptions in data section.

Specific Comment 18. *L235: words should be singular in "and does contain the words 'will'…". Also, I wonder if this approach successfully separated actual events from forecasts? Is there any language specificity in Chinese invoved here?*

We don’t think it is related to the language specificity in Chinese, the forecasts should include the words representing future state. Therefore, we can take two steps to distinguish actual events and forecasts. Firstly, the answer to Question 1 could contain the flood-related events. The answer is usually just one short sentence which defines the events described in the news. Secondly, we identify the words representing future state and remove the corresponding events.

Specific Comment 19. *Figure 3: Is [SEP] a requirement given the specificity of the Chinese language?*

No, [SEP] is a special token used for BERT model not just for Chinese language tasks. In a Question-Answering (Q&A) task using BERT, the [SEP] token is essential. It separates the question from the context or passage from which the answer needs to be extracted. The typical input format for BERT in a Q&A task is: [CLS] Question [SEP] Context [SEP]

This structure helps BERT understand the boundaries and relationships between the question and the context, facilitating accurate extraction of the answer.

Specific Comment 20. *L243: In the first sentence, correct "flood information extraction" into "(i) flood event detection and (ii) flood information extraction" for clarity.*

We will address this issue in the revised version.

Specific Comment 21. *L259: it is not clear to me how Exact Match behaves in case of multiple locations, zero if any error? What is it clearly meant by the location data? City? County? How is location handled before the flood location recognition is explained in section 3.2? Perhaps 3.2 should be explained before.*

Yes, if any one of multiple locations was not identified then the score for this sample is zero.

The location data specifically refers to county-level region name.

Sorry for unclear structure. We did not do any further processing of the answer information before flood location name recognition. We agree that Section 3.2 should be explained before and will address this issue in the revised version.

Specific Comment 22. *L276: consider adding the reference of the used administrative unit shapefile. See also G7.*

We will add this reference in the revised manuscript.

Specific Comment 23. *L285, section 4.1. The performance seems good in an absolute manner, but the reader has no clue how this performs in relation to the*

context of social sensing of flood or in the context of Chinese NLP. This is quite important to document.

Thanks for pointing that out. We will add the discussions on the performance of our NLP method.

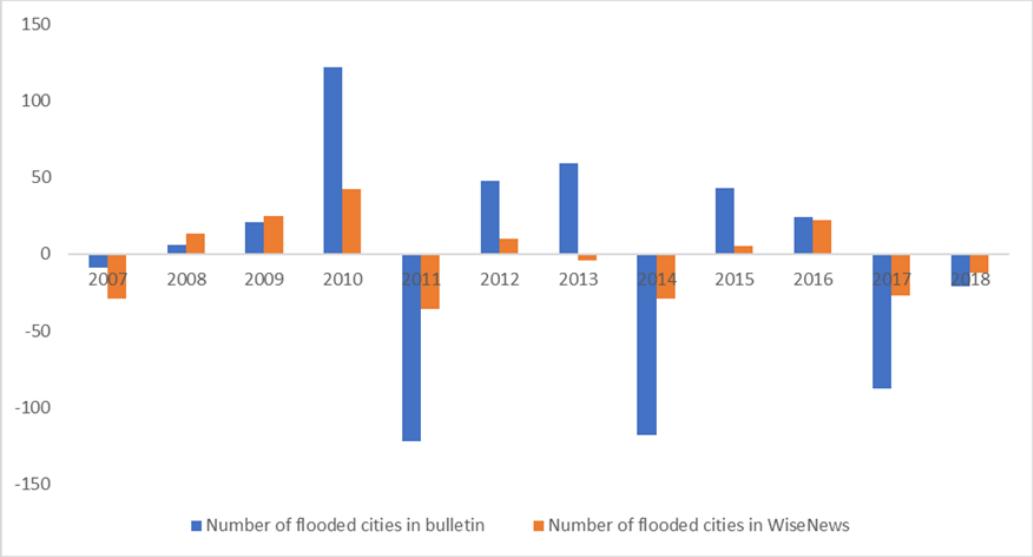
The performance of BERT model in this current study are competitive within the broader field of information extraction and Chinese NLP. For instance, Yang et al. (2022) adopted a BERT-based model for Chinese named entity recognition (NER) and achieved 94.78% and 62.06% F1 values on the MSRA (created by Microsoft Research Asia, is a well-structured and annotated collection of text for NER tasks) and Weibo (A Chinese social media platform) datasets, respectively. This significant disparity in performance highlights the challenges in semantic understanding in social media data compared to more structured datasets like MSRA. In addition, Kim et al. (2022) developed a question answering method for infrastructure damage information retrieval from textual data using BERT and achieved F1-scores of 90.5% and 83.6% for the hurricane and earthquake datasets, respectively.

Yang, Ruisen, Yong Gan, and Chenfang Zhang. 2022. "Chinese Named Entity Recognition Based on BERT and Lightweight Feature Extraction Model" Information 13, no. 11: 515. https://doi.org/10.3390/info13110515

Kim, Y., Bang, S., Sohn, J., & Kim, H. (2022). Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers. Automation in construction, 134, 104061.

Specific Comment 24. *Figure 4: Bulletin seems more exhaustive. This could be discussed more and the authors could highlight better complementarities between data collection approaches, e. g., how would the proposed approach improve Chinese bulletin?*

The spatial distribution of flood loss information provided in the bulletin is only at the provincial scale, and the number of flooded cities is mentioned in the paragraph describing the overall extent of the disaster. However, it does not provide a specific list of flooded cities or time information of each event. We have provided a more detailed list of affected counties. Additionally, we visualized the year-on-year differences in the data to offer a clearer view of interannual variations. As shown in the figure below, despite some degree of underestimation, the temporal trends in our data align closely with those reported in the Bulletin.



Specific Comment 25. *L298-L308: The analysis of media attention due to GDP biases is not significat and do not control for the population bias (see G4).*

We will remove the analysis of media attention due to GDP biases and detailed explanation is in response to G4.

Specific Comment 26. *L313-314: The two case studies were selected as the author assumed a good coverage because of their important hazard magnitude and impact. This is a known bias and an issue worth mentioning, as small-impact disasters tend to be less well-covered and documented. See Kron et al., 2012, Gall et al. 2009, and Delforge et al. 2023 and references therein for more insights about hazard catalog biases.*

We quite agree with you. We selected these two cases to show that our dataset can cover the more impactful events. However, it is true that small-impact events receive much less media attention, which is one of the limitations of our data set based on media data. We also appreciate the references you provided, and we will discuss the bias caused by media data as the response to G3.

Specific Comment 27. *L328-339 + Figure 7. These selected indicators are bad proxies of flood susceptibility, and I do not see how this analysis validates something about the spatial distribution of floods (see G5). Consider removing.*

We will remove this part.

Specific Comment 28. *L340: how the information was structured prior to harmonizing the data into the urban flood dataset is unclear. See also G6.*

The detailed description is in the response to G6. We will add the explanation in the revision.

Specific Comment 29. *Figures 8 and 9, it would be great to have an additional column or a time series on the Y axis with the annual total. This could help identify pluriannual cycles as a result of climate indices. Consider adding the total number of occurrences and items in the figure caption.*

According to your suggestions, we will modify these figures as followings:

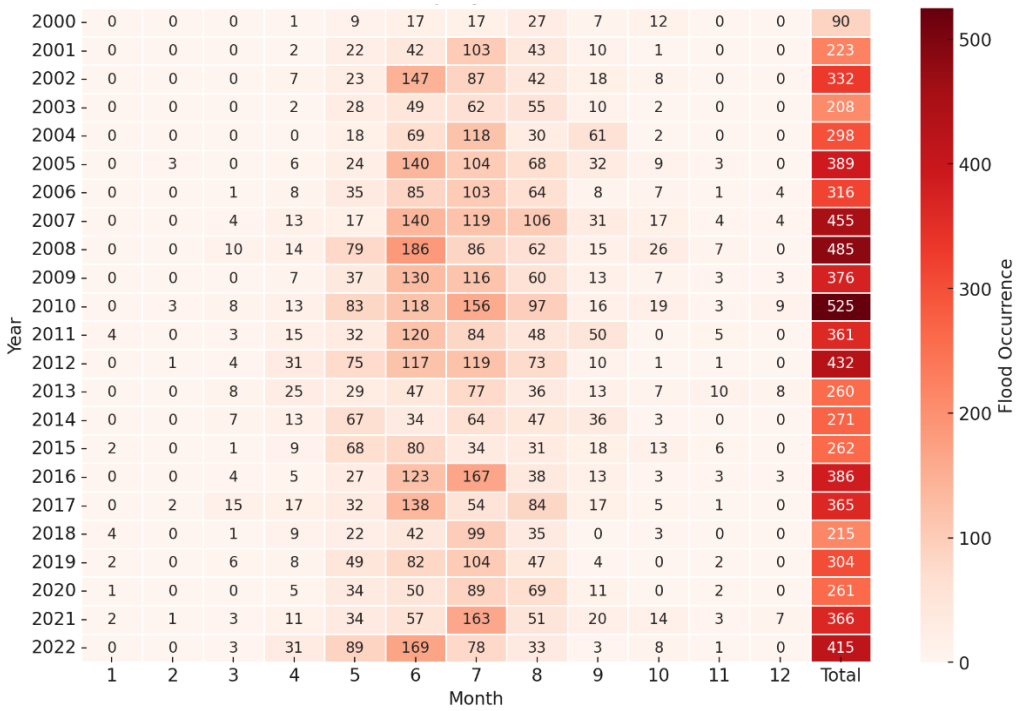


Figure 8. Flood occurrence heatmap by year and month.

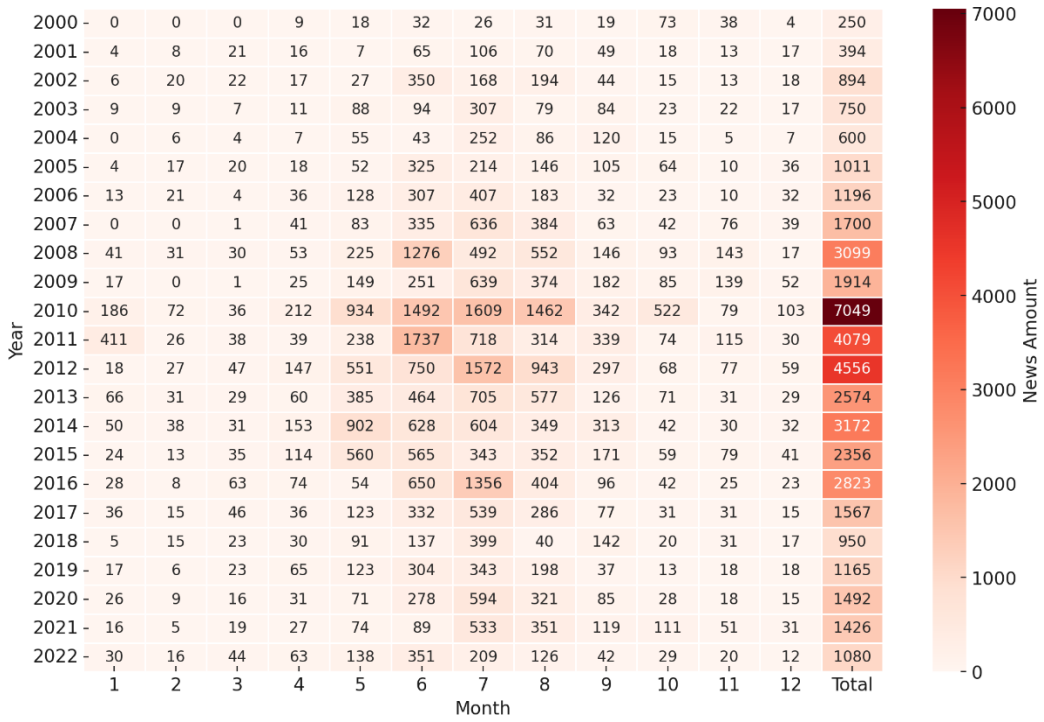


Figure 8. Flood-related news heatmap by year and month.

Specific Comment 30. *L354: "seasonality" instead of "climate's tendency" could be more appropriate.*

We will address this issue in the revised version.

Specific Comment 31. *L390: "exposure" or "susceptibility" (the environmental side of vulnerability) is maybe more appropriate than vulnerability because the latter also encompasses social vulnerability.*

We will replace vulnerability with susceptibility in the revised version.

Specific Comment 32. *Maps Figures 10, 11, and 12 could be grouped into a multipanel figure for conciseness. Consider adding population density as well since it drives hazard exposure. DEM and river networks may also be considered as information to include (parsimoniously).*

First, we will group these figures into a multi-panel figure as you suggested and include the population density and the Gross Regional Product (GRP) related analysis. The detailed description is in the response to G3.

Regarding the suggestion to include DEM and river networks, we appreciate the idea but believe these factors, while relevant, do not directly align with the primary focus of our analysis. Incorporating DEM and river networks would introduce additional complexity that may not substantially contribute to the core

findings or enhance the validation of our flood distribution data. We agree with your opinion in G5 that in areas with very high elevation, the low population exposure naturally leads to fewer reported flood events, so conducting spatial analysis with DEM as the sole base layer does not have significant meaning. Similarly, the independent analysis of river networks is not particularly meaningful.

Specific Comment 33. *L409: The comparison with other datasets is quite limited, and the Chinese bulletin seems more exhaustive if one can trace the original data. To what extent the proposed dataset fills gaps is thus not very well documented (see G1). Adding more than one catalog from Table 1 and 2 in Figure 4 for comparison can improve this discussion.*

There is no other Chinse national-level dataset describing the inventory of urban floods. The *Chinese Flood and Drought Bulletin* just shows the number of flooded cities for each year without specific flooded cities inventory and in recent years, even the numbers have not been published. Additionally, no other datasets from Table 1 and 2 could provide the number of flooded cities or counties across China so that we cannot add more than one catalog in Figure 4. The absence of such comparable data itself highlights that our dataset fills a gap in urban flood data on a national scale in China.

The spatial distribution of flood loss information in the bulletin is limited to the province level, which encompasses multiple city-level areas. Our dataset, despite its limitations, offers more granular information by identifying specific flooded areas at the county level, which is smaller than the city level. There may be biases inherent in the news data, but we believe that our dataset serves as a valuable reference in the absence of more detailed and comprehensive data sources.

Specific Comment 34. *L473: The data availability section does not include the input news data accessibility information. In line with HESS recommendations and FAIR standards, I also encourage the authors to share information about code and model availabilities.*

We will ensure the maximum possible sharing of data and code. The detailed explanation is in the response to G7.

Specific Comment 35. *L414-L416: this sentence (and the section in general) looks like the authors do their best to fit in the context of climate change and urbanization, even excluding some peak values to retrieve a positive trend. Trends, in particular for disaster news, are much more complex than trends observed on physical variables and include important social drivers and biases. The discussion is oversimplified, and the authors should take more distance and*

inquire about the biases arising from social sensing of hazards. See G3 and references.

We agree that trends for disaster news are much more complex than trends observed on physical variables and include important social drivers and biases. We will add the discussion on the biases arising from social sensing of flood hazard and the detailed description is mentioned in response to G3.

Specific Comment 36. *L445: Perspectives are neither exhaustive nor detailed. Consider adding more relevant perspectives, differentiating those related to the method (NLP-detection, extraction) and those related to the valorization of the resulting dataset.*

We will re-organize the discussions of limitations and future directions on the method and the resulting dataset according to your suggestion.

The current study employs a BERT model for question-answering tasks, which has proven efficient in information extraction. However, with the rapid advancement in large language models (LLMs), newer models such as ChatGPT offer significant improvements in various NLP tasks, including text classification, question-answering, and text generation, achieving state-of-the-art results. For instance, Colverd et al. (2023) have successfully used several LLMs, including GPT-3.5, GPT-4, and PaLM-Text-Biso, to generate flood disaster impact reports by extracting and curating information from the web. They found a notable correlation between the scores assigned by GPT-4 and human evaluators when comparing generated reports to human-authored ones. Furthermore, Hu et al. (2023) proposed a method fusing geospatial knowledge of locations with GPT models to extract location descriptions from disaster-related social media messages, demonstrating a 40% improvement over typically used Named Entity Recognition (NER) approaches. Given these advancements, our future research will explore the use of LLMs to extract nuanced information from flood-related text data, which includes distinguishing flood types, causes, and the specific losses associated with each flooding event.

The subsequent analysis of the resulting dataset (constructed from the extracted information) in this present study also has limitations, which fail to fully leverage the advantages of county-level data in revealing regional flood characteristics. Future research could involve attribution analysis of floods to explore the main contributing factors in different areas. Additionally, by analyzing changes in land use and urban planning in specific counties, a more comprehensive understanding of how various factors interact at the local level to cause flood events can be achieved. Moreover, leveraging advanced machine learning models, such as deep learning and ensemble methods, could enhance the predictive capabilities of flood risk evaluation. By addressing these aspects,

future studies can significantly improve the utility of county-level flood data, offering better-informed strategies for flood mitigation and resilience planning.

Colverd, Grace, Paul Darm, Leonard Silverberg, and Noah Kasmanoff. "Floodbrain: Flood disaster reporting by web-based retrieval augmented generation with an llm." arXiv preprint arXiv:2311.02597 (2023).

Hu, Yingjie, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou, and Kenneth Joseph. "Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages." International Journal of Geographical Information Science 37, no. 11 (2023): 2289-2318.

Specific Comment 37. *L473: data and code availabilities: see G7.*

The detailed explanation is in the response to G7.

Specific Comment 38. *Table A2: Same as Figure 4. It may be removed, in my opinion.*

We initially want to share the raw form data for Figure 4 incase that some readers are interested. However, the information indeed is duplicated. We will remove it in the revise version.