Comment 1: *From the abstract, but also the rest of the paper, the level of detail that this flood information dataset has is unclear. A spatial scale is mentioned as 'county-level', but that can vary quite a lot depending on where the reader is from. Connecting this to a typical length scale (1, 10, 100, … kilometres?) will make it clearer to a potential end-user whether this dataset is useful.*

Similarly: what kind of information is present about the flooding? Is it just spatial extent? Or also indications of amounts of water, timing or duration, damages done, etc etc. This should be immediately clear from the first reading, in both the abstract, as well as the results section.

Related to this, table 1 is an overview of current flood disaster reports, which also doesn't contain any information on the kind of data that's in there. Giving both your, and the existing datasets that level of detail can make it clear what the advantage of this new methodology is in comparison to the existing ones. Also, the validation data described in section 2.3 suffers from this lack of information.

First, we would like to explain the relationship between the different administrative levels in China:

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.

Regarding the specific spatial scope, the size of county-level administrative regions varies greatly across different parts of the country. The smallest county-level administrative region is Jing'an District in Shanghai, which covers only 8 square kilometers, while the largest is Ruoqiang County in the Xinjiang Uygur Autonomous Region, covering 206,903 square kilometers. The average area of a county-level administrative region nationwide is 3,459 square kilometers.

This is important for enhancing the readability of the article and dataset. We will include a brief explanation of the spatial scope in the revised version.

Second, our dataset provides information solely on the timing and names of the affected areas. Unfortunately, it does not include details such as the spatial extent, water volume, or damages caused by the flooding. This limitation arises from the nature of the data source—we relied on news reports rather than scientific papers, which typically do not provide the physical measurements or quantitative details often found in more specialized studies. In the future, we will introduce more data

source to update and correct the dataset and add these kinds of flood event details.

We will make this clearer in the abstract and results sections of the revised manuscript to ensure that readers understand the dataset.

Additionally, we will update the flood record information in Table 1 like followings:

Name	Period	Flood Records	Update Frequency	Source
Annual Report of Chinese Hydrology	2021--	Number of basin/river floods and flooded river list	Annual	Ministry of Water Resources of the People's Republic of China
China Flood and Drought Bulletin	2006--	The population, economic, and crop losses in each province	Annual	Ministry of Water Resources of the People's Republic of China
China Meteorological Disaster Yearbook	2004--	Time, flooded district, damage of major flood events, the record criteria as events causing over 50,000 hectares of agricultural damage, 10 deaths, or 14 million USD in direct economic losses	Annual	China Meteorological Administration
Reports on official website of China National Disaster Reduction Center	2011--	Records of the time, location and damage of flood events (Data prior to 2018 is not available)	Real-time	National Disaster Reduction Center of China

Comment 2: *The approach used seems quite specific for the Chinese language, using several specifically trained models and training input. It's worthy of discussion of your approach also works for a completely different language group to apply this methodology in other data-scarce regions (e.g. the Global South).*

Our approach was indeed fine-tuned using a Chinese corpus, which means that applying this methodology to a completely different language would require retraining the model with a suitable corpus in that language. This is because BERT

models are language-specific, and the fine-tuning process is critical to adapting the model to the nuances of the target language. Moreover, our method relies on the availability of news data, which may be less abundant or even scarce in certain regions. This potential lack of media data could limit the applicability of our approach in those areas.

This point is important, and we will include this discussion in the revision. While the model we trained cannot be directly transferred to regions with different languages, the technical approach we have developed can be applied in any region and serves as a reference.

Comment 3: *Also, regarding the approach: the media used are all newspaper databases, and only 2 different ones. Why is social media not included, or other sources of information? This seems to limit the potential of the method, since using one type of media source might be fairly uniform in its wording and phrasing, and perhaps not always covering all instances of floods. Furthermore, the restrictive choices on the keywords to select these articles might make the whole model biased: was there any form of testing with broader search terms, synonyms or other idioms for instance (like in L 152)? The model is strongly influenced by the choices for the training data obviously, but it seems to me like some additional testing of the influence of that training data is necessary.*

First, social media data is often non-public and may involve privacy issues, which can impose limitations on its use. On the other hand, due to concerns about data quality control, news articles were selected in the hope of obtaining more accurate results. Data with varying language styles might negatively impact the model's performance.

Regarding the problem of keywords, we tested other keywords as retrieval methods and found that 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. The idiom “Floods and beasts” was determined after analyzing CNKI news data, and other intrusive idioms are rarely seen.

We agree that the model is strongly influenced by the choices for the training data. In constructing the training set, we randomly selected samples rather than using data from consecutive years or a single newspaper source, aiming to help the model learn more diverse features.

Moreover, we realized this is an important suggestion, and we are currently experimenting with different random sample combinations for cross-validation. If possible, we will include the cross-validation results in the revised version.

Comment 4: *Reading through the methodology it seems like a lot of manual preprocessing is still required, including manually annotating news texts. Ow much of a bottleneck is that for operational purposes is that, if you really will have a constantly updating database? This requires some discussion since it directly impacts the applicability of this dataset.*

Thank you for raising this important point. The manual preprocessing, including the annotation of news texts, is primarily required during the initial fine-tuning of the model, as well as for adjusting hyperparameters. This step is essential for ensuring the accuracy and effectiveness of the model. However, once the model has been trained, it does not need to be retrained for future applications. Future data can be directly processed into the test set format and used as input for the model without additional manual preprocessing. Therefore, this process does not represent a significant problem for the operational application of a constantly updating database.

Comment 5: *L 235: the exclusion of any texts wit the word 'will' seems like it can introduce giant margins of error. I get the reasoning to exclude forecasts, but if 'will' is used in a different context in a text that is actually related to flooding (e.g. 'damaged roads will re-open in 4 days') is then the whole text still excluded?*

This screening measure is based on BERT's answer to what disaster event happened, which is a classification of the disaster events described in the news. For example, the statement you mentioned, 'damaged roads will re-open in 4 days,' is related to the impact of the event and would not typically appear in the answer content. The responses to the Question 1 are focused on the type of the event, and they are usually very brief and do not include the details of the event. Therefore, this exclusion will not introduce big biases.

Comment 6: *The choice of GDP as a clustering method is odd to me. Why not use population density instead? That does correlate somewhat with GDP (so you still get reports on economic losses) but the loss of human life also hugely matters in disaster reporting, I'd think.*

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 follows:

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.

Moreover, we will add the analysis of the flood trend in different population density and economically developed areas to provide insights from an urban and social perspective as following figure:

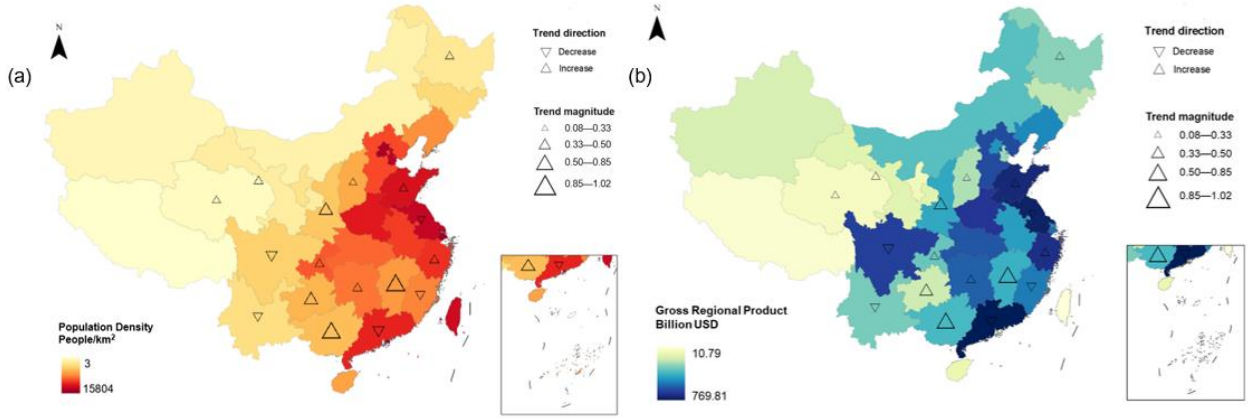

Figure x. The analysis of flood event trends across Chinese provinces from 2000 to 2022, shown in relation to (a) population density and (b) Gross Regional Product (GRP).

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 x(a). 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.

Comment 7: *Figure 6: This figure doesn't seem too relevant to the paper to warrant inclusion. A typhoon is certainly going to lead to flooding but the spatial scale is so wide that it's not a great verification in my opinion.*

We also agree that typhoons undoubtedly cause flooding, but the spatial scale is indeed quite large. Our intention in using Figure 6 was to demonstrate that our dataset successfully identifies disaster areas affected by typhoons, serving as evidence that our dataset can capture events with a broad impact. When considering the biases inherent in news data, a review of other literature revealed that 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. We speculate that this bias in media data is mainly due to the neglect of smaller-scale or less severe events. Therefore, we used these two significant cases to illustrate that larger-scale events are still likely to be reported. However, we understand that this figure may not adequately demonstrate the accuracy or completeness of our data. If it does not fit well with the structure of the paper, we are willing to move this analysis to the supplementary materials.

Comment 8: *Figures 8 and 9: occurrence is here shown without any distinction of severity of flooding, whereas the latter one might be more relevant for actual use of the dataset.*

Thank you for your suggestions regarding Figures 8 and 9. These figures represent the heatmaps of flood occurrences and the number of flood-related news reports by year and month. The primary purpose is to illustrate the temporal distribution of flood events across China. These visualizations help to highlight the years and seasons during which floods are most frequent, offering insights into the timing of flood events over the study period.

We acknowledge that distinguishing the severity of flooding could enhance the relevance of the dataset for certain applications. However, our current dataset does not provide detailed information on the severity of the flooding. Because this information is expressed with more variability and is more unstructured across different news sources, we will try to re-train the model or introduce new methods for extracting this type of information in future research.

Comment 9: *L230: I don't understand what the authors mean with '3 epochs' and*

a learning rate of 5 x 10^-5. Please elaborate.

The term "3 epochs" refers to the number of times the entire training dataset is passed through the model during the training process. In our case, we trained the model for 3 epochs, meaning the dataset was fed into the model three times, which is a standard approach to ensure that the model learns the patterns effectively without overfitting.

The learning rate of 5 x 10^-5 is a hyperparameter that controls the step size at each iteration while moving toward a minimum of the loss function. A smaller learning rate, such as the one we used, allows the model to converge more slowly and steadily, reducing the risk of overshooting the optimal parameters. This value was chosen based on preliminary experiments to balance the learning speed and model performance.

We will clarify these points in the revised manuscript to make them more understandable.

Comment 10: *L 275: 'verify and revise': is this part of the preprocessing? What does this mean, exactly?*

The process of "verifying and revising" is not part of the preprocessing but rather a post-processing step. The location names are generated by the BiLSTM-CRF model, and since the data spans from 2000 to 2022, it includes periods during which several regions in China underwent administrative adjustments or renaming. To ensure accuracy and relevance when associating these locations with the administrative division shapefile for spatial visualization in ArcGIS, I updated the names to reflect the most current administrative divisions. This step was crucial for maintaining consistency and ensuring that the visualizations accurately represent the latest geographical boundaries.

Comment 11: *Figure 4: Any idea what causes the large biases? This is hardly discussed.*

Identifying specific biases is challenging because the *China Flood and Drought Bulletin* provides only the total number of flooded cities, without listing specific locations. This limitation makes it impossible to pinpoint which specific events or regions are underreported in our dataset.

However, we can hypothesize that the biases stem from the intrinsic characteristics of news data. Some previous studies have also reflected the bias of constructing disaster catalogs with reports (Gall et al., 2009; 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.

Although our data consistently underestimates the number of flooded cities each year, likely due to the influence of the factors mentioned above, the trend in our data closely follows the overall temporal pattern observed in the *China Flood and Drought Bulletin*. This suggests that our dataset can still be effectively used to explore variations in flood events across regions. Then, it can be leveraged to analyze potential influencing factors of these changes, such as socioeconomic changes, climate change, alterations in land surface characteristics, and modifications in flood control measures, ultimately providing recommendations for flood management.

On the other hand, in future research, we plan to incorporate more available data sources to continuously update and validate this dataset. By expanding the dataset’s coverage and adding descriptions of damages, its comprehensiveness will be improved.

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.