Comment 1： *The authors provide a comprehensive introduction to existing natural disaster datasets that record flood events created by official sources, other governments, or organizations. However, the manuscript would benefit from a more detailed discussion on how this study specifically addresses the gaps in these existing datasets. It is essential to clearly state the novelty and significance of your work in the context of existing datasets. For instance, do the deficiencies in these existing datasets affect the analysis, modeling, and prediction of flood events to some extent? How does the new dataset you have developed alleviate these issues at both theoretical and practical application levels?*

We will add more statements of the novelty and significance of our dataset in the revision: We construct the first county-level urban flood inventory across China from 2000. The existing datasets cannot present the floods distribution at county-level or provide the information across China. The most comprehensive and authoritative data on flood disasters published annually in China is found in the *China Flood and Drought Bulletin*. However, this data primarily focuses on the economic losses, casualties, and agricultural damages at the provincial level. While the Bulletin provides an overall description that includes the number of affected cities, it does not present the specific inventory of these cities. Our dataset shows trends that are largely consistent with those reported in the *Bulletin*, indicating that it can reliably reflect changes in flood events across different regions. Our data is originally at the county level and it can also be resampled and aggregated to city or provincial levels for research on flood dynamics and their influencing factors across different spatial scales.

Comment 2： *Line 143 "After a manual review to remove duplicates and irrelevant entries, including those referring to flash floods which occur suddenly in mountainous areas and are not the focus of this study, the final dataset consisted of 253 relevant news articles". The data preparation section needs more details. Please explain the criteria used for manually reviewing and removing irrelevant news articles from the CNKI database. Additionally, discuss any potential biases or limitations introduced by this manual selection process.*

We will explain more in the data preparation section:

The lead author manually reviewed the news data from CNKI. During this process, duplicate reports of the same flood events from different regional newspapers were removed, as well as articles containing search keywords but actually reporting on flood prevention measures, flood season warnings, and other non-flood events. Additionally, since this paper focuses on urban flooding, events related to flash floods and landslides were also excluded. The researchers in our group has double-checked this reviewing results.

Although we have conducted a double-check, manual interpretation may still introduce biases due to subjective differences in cognition. In future research, we plan to incorporate large language models for correlation analysis or involve more domain experts to cross-validate the accuracy of our correlation assessments.

Comment 3： *Similarly, Line145 “These relevant news articles were then segmented into paragraphs and reorganized into 633 distinct samples. Among them, 503 samples were used to fine-tune the BERT model, alongside data from the CMRC2018 dataset, enhancing the model's stability to accurately extract flood disaster information. The remaining 130 samples served as a test set to evaluate the model’s performance.” Please clarify how the 503 samples were selected from the 633 distinct samples, and explain why the remaining 130 samples were used to evaluate the model’s performance. This selection process is currently unclear and confusing.*

Thank you for your insightful feedback. Regarding the selection of the 503 samples from the 633 distinct samples, we would like to clarify that this process was done through random sampling, without any subjective selection. The intent behind this random division was to ensure that the training and testing datasets were representative of the entire dataset, thereby minimizing any potential bias.

The remaining 130 samples were designated as the test set after the training samples were randomly selected, without any manual intervention or predefined criteria. This allows for an independent evaluation of the model's ability to generalize to new, unseen data.

Additionally, we are currently experimenting with randomly selecting different training samples for retraining. If possible, we will describe the results of cross-validation in the revised version.

Comment 4： *For the identification of flood locations, I have a general question. From my understanding, news media reports about flooding occurrences typically mention the affected city or, at most, the district. However, actual urban flooding can occur at the street level or even smaller scales. Could you please provide a detailed explanation of how the BiLSTM-CRF model was trained and applied to recognize flood locations?*

Thank you for raising this important point. It is indeed true that news reports often only mention the affected city or district, with only some reports specifying smaller-scale locations like streets or buildings. However, our goal is to obtain results at the county or district level rather than more granular details.

To clarify our approach, we used the BERT model to identify the flood-affected areas mentioned in each news report firstly. The answer sentences typically include cities, counties or districts, and in some cases, specific streets or buildings. After obtaining these mentions, the BiLSTM-CRF model was employed to extract the pure location names from BERT’s output.

The BiLSTM-CRF model was trained using the MSRA named entity recognition corpus, a widely used dataset developed by Microsoft Research Asia, which contains a large amount of annotated Chinese sentences with named entities such as location names, person names, and organization names. Then, we standardized the spatial information by using county/district names (in China, counties and districts are the same administrative level and both included in cities).

Comment 5： *Regarding the performance of the BERT model (Table 4), it appears that the authors have only examined results based on a binary classification (flood vs. non-flood). If this is the case, the task seems too simple and lacks sufficient novelty. Could the authors also provide an evaluation of the model’s performance in identifying the time and location of flood events?*

Table 4 does not only present the results of categorizing events as flood or non-flood. Only the first row is for classification. The second row shows the evaluation results of spatiotemporal information extraction. Two evaluation systems were used for the recognition of time and location information. This is explained in Section 3.1.3, Evaluation Metrics:

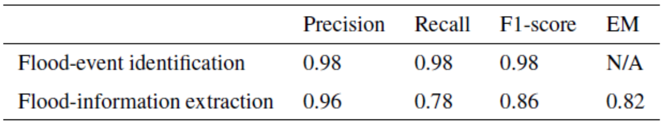
“The first index is called exact match, which measures the matching degree between the prediction and ground truths. The score is 1 for the EM of both the time and location information extracted. Otherwise, the score is 0. There is usually more than one disaster location in one flood event and maybe the model can output several but not completely accurate locations. Therefore, a fuzzy match was used to evaluate the location extraction using precision, recall, and F1 score. Unlike the classical formula, the precision and recall were calculated as:

$$Precision = \frac{P}{M}$$

$$Recall = \frac{P}{N}$$

Where P represents the number of accurately extracted flood locations, M is the total number of predicted flood locations and N is the total number of actual flood locations observed in the texts.”

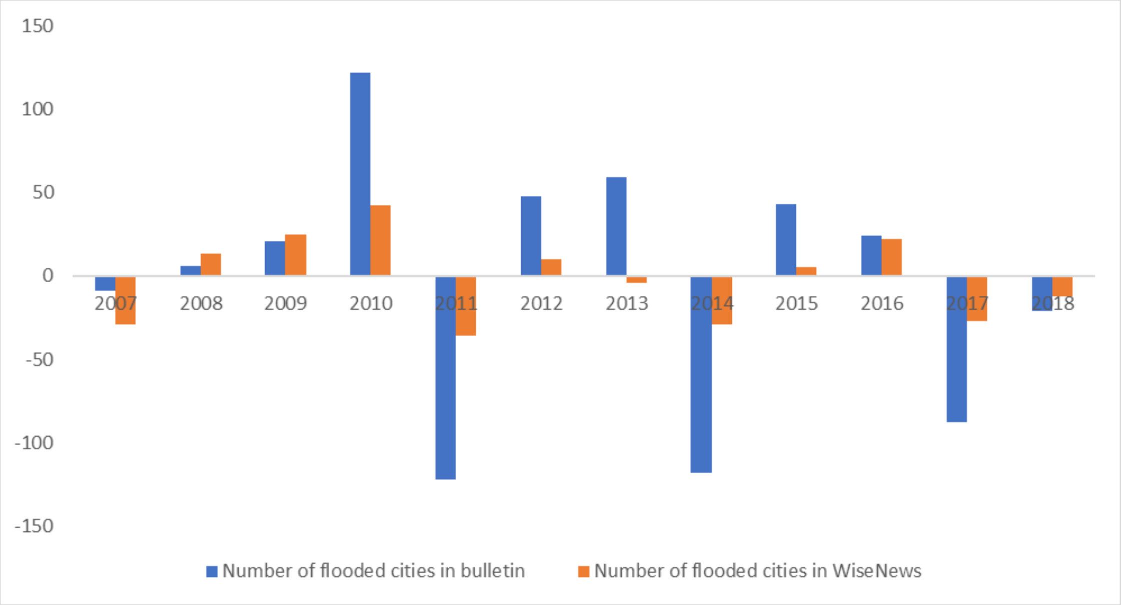
Table 4. The performance of the BERT model (EM index was not applied to evaluate event identification)



	Precision	Recall	F1-score	EM
Flood-event identification	0.98	0.98	0.98	N/A
Flood-information extraction	0.96	0.78	0.86	0.82

Comment 6： *It seems that the number of identified flooded cities is significantly underestimated by the news media compared to the China Flood and Drought Bulletin (Figure 4). The authors suggest this discrepancy is related to the low attention given to low GDP areas. However, this raises a significant concern about the reliability of the developed dataset. As mentioned in section 4.3, the dataset records urban flood events reported in news articles from 2000 to 2022. If the news media is so inaccurate that it fails to record a large number of flood events, how can the authors ensure the reliability of the data generated from these news sources?*

We visualized the year-on-year difference in the number of flooded cities between the two datasets to provide a more intuitive representation of interannual variations. As shown in the figure below, the trend in our data is largely consistent with that reported in the Bulletin. This suggests that our dataset can reliably reflect changes in flood events across different regions, though news media may underestimate the number of cities affected by floods.



On the other hand, while the *China Flood and Drought Bulletin* provides a summary of the number of flooded cities, it does not offer a detailed inventory of specific locations. The spatial distribution of flood disaster loss information in the bulletin is limited to the province level, which encompasses multiple city-level areas. There may be biases inherent in the news data, but we contend that our dataset serves as a valuable reference in the absence of more detailed and comprehensive data sources.

In the future, we will introduce more data source to improve the data coverage, such as social media data, available disaster reports in some cities or provinces and so on.

Comment 7：*Figure 6 is not directly related to your results, I think you can put it into supplementary materials.*

We will put this part into supplementary materials.