https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Open Access Earth System Science Data Discussions

CFDID: a China Flood Disaster Impact Database from 1994 to 2023

Shibo Cui[1,2], Ni Li[3,4], Shiruo Hu[1,2], Zhiqi Wan[1,2], Jiaru Wang[1,2], Jinchao Tang[1,2], Ruixi Zhang[1,2], Yingjia Li[1,2], Xinpeng Yu[1,2], Yiqian Tan[1,2], Wenbo Shan[1,2], Jianshi Zhao[1,2]*

[1]State Key Laboratory of Hydroscience and Engineering, Tsinghua University, Beijing, 100084, China
[2]Department of Hydraulic Engineering, Tsinghua University, Beijing, 100084, China
[3]Department of Water and Climate, Vrije Universiteit Brussel, Brussels, 1050, Belgium
[4]Department of Hydro Sciences, TUD Dresden University of Technology, Dresden, 01062, Germany

Correspondence to: Jianshi Zhao (zhaojianshi@tsinghua.edu.cn)

Abstract. China is one of the countries most severely affected by flood disasters worldwide. Long-term and systematic flood impact data are essential for flood risk assessment, flood impact model calibration, disaster prevention and reduction. However, an open, long-term, China-focused event-scale database with detailed flood disaster impact indicators remains lacking. Here we present the China Flood Disaster Impact Database (CFDID), which contains 1716 flood disaster event records from 1994 to 2023, including 1391 top-level (L1) event records. CFDID is compiled from 66 official documents and covers five flood-related disaster categories: rainstorm flood, typhoon, snowmelt flood, ice-jam flood, and dam-break flood. Each record includes date, location, disaster category, eight impact indicators, and data source. The database was constructed using a workflow that combines optical character recognition (OCR), large language model (LLM)-based information extraction, and manual verification, thereby improving the efficiency of converting unstructured text information into structured event records while ensuring data quality. Based on these L1 records, CFDID achieves total coverage ratios of 67.22% for deaths and missing persons, 72.25% for affected population, 60.70% for affected cropland area, 65.68% for collapsed dwellings, and 74.86% for direct economic losses relative to official annual flood disaster impact totals during 1994–2023. Compared with commonly used international flood impact databases, CFDID contains more flood disaster event records and has higher coverage for most impact indicators. CFDID also includes agricultural impact indicators, which are rarely available in international hazard impact databases. CFDID is available at https://doi.org/10.5281/zenodo.21699973 and will be continuously updated by year.

1 Introduction

Flood is one of the most widespread natural hazards globally (CRED and UNDRR, 2020). China is among the countries with the largest flood-affected population and direct economic losses worldwide (Dottori et al., 2018; Kang et al., 2025). With climate change and socioeconomic development, the spatiotemporal pattern of flood impacts in China may continue to evolve (Jiang et al., 2020; Winsemius et al., 2016; Zhang and Zhou, 2020). Scientific assessment of changes in flood impacts requires not only hydrometeorological, population, and asset data, but also long-term and systematic disaster impact records (Messori et al., 2026; Paprotny et al., 2025).

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Several international disaster databases record flood impacts across countries and provide important data for global disaster risk research. The Emergency Events Database (EM-DAT), one of the most widely used disaster databases, has recorded major disasters worldwide since 1900 and provides information on deaths, affected population, and economic losses (Delforge et al., 2025). However, small-scale and localized events are generally not included, which may reduce EM-DAT's coverage of flood impacts. The Wikimpacts database extracts global climate-related event impacts from Wikipedia using large language models (LLMs) and demonstrates the potential of LLMs for constructing disaster impact databases (Li et al., 2026). Nevertheless, the limited number of China-related flood entries in Wikipedia results in lower coverage for China in Wikimpacts than in EM-DAT. The Dartmouth Flood Observatory (DFO) database is mainly based on remote sensing and records the location and inundation extent of high-impact floods (DFO, 2026), but it is constrained by satellite observation capability and provides only limited socioeconomic impact indicators, mainly deaths collected from news. In addition, insurance and reinsurance companies maintain detailed natural catastrophe databases, such as the Sigma (Swiss Re, 2026) and the NatCat (Munich Re, 2026), but these datasets are not openly accessible. Overall, existing international databases remain limited for detailed flood impact studies of China, and their ability to represent flood impacts in China has not been systematically evaluated.

The Chinese government has long conducted official disaster statistics and annually publishes provincial flood impact totals, providing an annual-scale basis for flood loss assessment. Yet event-scale flood impact information is needed to characterize flood impacts more precisely and to support model calibration, validation, attribution, and prediction of flood impacts. Some recent studies have constructed flood event inventories for China using news or social media data (Fu et al., 2025; Shen et al., 2025), but these inventories generally lack detailed disaster impact information. Official annual documents from water resources, meteorological, civil affairs, and emergency management authorities contain authoritative and traceable flood impact information. However, these event-level data are scattered across multiple Chinese-language documents, limiting the public accessibility and applicability of existing flood disaster impact records in China.

To fill this gap, we developed the China Flood Disaster Impact Database (CFDID). The main contributions of this study are threefold. First, we integrate 66 official documents and compile 1716 event-scale flood impact records for China during 1994-2023. CFDID provides structured fields on date, location, disaster category, eight impact indicators, and traceable data sources, thereby substantially enriching the data basis for flood impact research in China. Second, we establish a database construction workflow that combines optical character recognition (OCR), LLM-based information extraction, and manual verification, which provides a methodological reference for constructing similar disaster impact databases. Third, using official annual flood impact totals as benchmarks, we assess the coverage differences between CFDID and several international databases, providing a quantitative basis for understanding the coverage and complementarity of different disaster impact databases.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

2 Methods

2.1 Definition and classification of flood disaster events

Flood generally refers to a hydrological phenomenon characterized by a rapid increase in water volume or water levels in rivers and lakes, or by excessive water accumulation on the land surface or in the soil. The flood disaster event records in CFDID are not equivalent to hydrological flood processes. Instead, CFDID adopts a source-based operational definition of flood disaster events. Because historical official documents do not consistently provide hydrological thresholds or fixed temporal windows for event delineation, event records in CFDID are delineated according to the event units used in the original official sources. Specifically, a flood-related disaster item was included when it was reported as a separate impact record in the source document and explicitly provided socioeconomic impacts, including casualties, affected population, affected cropland area, collapsed dwellings, direct economic losses, or other impact indicators. Records that describe only water level, streamflow, surface water accumulation, or soil moisture anomalies without socioeconomic impacts are not included.

To preserve the reporting structure of the original sources, CFDID uses a three-level event hierarchy. L1 records refer to parent-event records, representing independent flood disaster events or broad flood disaster events reported in the original sources. L2 records refer to sub-event records nested within L1 records, usually describing regional, local, or component impact records that belong to a broader L1 event. L3 records refer to more detailed nested records within L2 records. When the original sources reported both a broad disaster event and nested regional or local impact records, CFDID retained this hierarchy rather than redefining events using a uniform duration threshold. Therefore, L1, L2, and L3 should be interpreted as hierarchical record levels rather than three independent event types.

Referring to the definitions in the 2024 Natural Disaster Statistical Investigation System issued by the Ministry of Emergency Management of China, CFDID classifies flood-related disasters into five categories according to the primary triggering mechanism: rainstorm flood, typhoon, snowmelt flood, ice-jam flood, and dam-break flood. These categories represent flood disasters mainly caused by heavy rainfall, typhoons, snowmelt, ice jams, and dam or dike failures, respectively. Outburst floods caused by the failure of barrier lakes are also classified as dam-break floods in CFDID. Because historical Chinese flood disaster records generally do not provide sufficient information to separate impacts among fluvial floods, pluvial floods, flash floods, and coastal floods, CFDID does not further classify events according to these common flood subtypes.

Several classification rules require further clarification. First, CFDID assigns each event record to a single primary disaster category according to the classification used in the original official sources. When multiple hazards occurred in the same disaster process, the category follows the main cause reported by the source document. Therefore, CFDID does not define a separate category for compound flood disasters. Second, the typhoon category should be interpreted as the event category reported in the original official sources, rather than as flood-only losses. Typhoon events may cause damage through strong winds, heavy rainfall, flooding, storm surges, and related secondary hazards. However, the source documents usually report the impacts of a typhoon event in aggregate and do not consistently separate the contributions of wind, rainfall, and flooding.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Therefore, CFDID retains the reported aggregate impacts under the typhoon category. Third, CFDID does not systematically include all landslides, debris flows, or collapses. These geological hazards are included only when the original source explicitly states that they were triggered by rainstorms, floods, or typhoons and reports them as part of a flood- or typhoon-related disaster event. If the source does not provide a direct link to rainfall, flooding, or typhoon processes, these geological hazard impacts are not included in CFDID.

2.2 Data sources

CFDID (1994-2023) is mainly compiled from 66 Chinese official documents. The core sources include the China Water Resources Yearbook, the Yearbook of Meteorological Disasters in China, and the annual list of the "Top Ten Natural Disasters in China" (Table 1). Examples of the three data sources are shown in Appendix A. These sources record the dates, locations, and main impact indicators of flood disasters in the corresponding years. The China Water Resources Yearbook and the Yearbook of Meteorological Disasters in China are mainly paper publications and were obtained through library access or purchase. The "Top Ten Natural Disasters in China" records are available from government websites. The China Water Resources Yearbook recorded event-scale flood impacts from 1994 to 2018, but systematic reporting of flood disaster information ceased after 2018 due to changes in government responsibilities. The Yearbook of Meteorological Disasters in China has provided relevant records since its first publication in 2002, while the "Top Ten Natural Disasters in China" list has been compiled since 2005.

Table 1. Data sources of CFDID.

No.	Data source	Year	Number of sources	Responsible authority
1	China Water Resources Yearbook	1994-2018	25	Water resources authorities
2	Yearbook of Meteorological Disasters in China	2002-2023	22	Meteorological authorities
3	Top Ten Natural Disasters in China	2005-2023	19	Civil affairs authorities (2005-2017) / Emergency management authorities (2018-2023)

2.3 Database fields

CFDID contains 22 fields, which can be grouped into event hierarchy and identifiers, temporal information, spatial information, disaster category, typhoon information, impact indicators, and data sources. The event level field records the hierarchical position of each record as L1, L2, or L3, while the event ID field provides a unique identifier that preserves this hierarchy. Temporal fields record the beginning and ending dates of direct disaster impacts. Spatial fields record affected provincial-level administrative regions and, when available, more specific place names reported in the original sources.

Table 2 lists the field names, descriptions, data types, units, and requirement levels in the released database. The "Required" column indicates whether a field is mandatory, semi-required, or optional. A "Yes" field must be recorded for each event record. The "semi-required" typhoon field means that the typhoon ID or name should be recorded when the disaster category is typhoon. The "semi-required" impact fields do not mean that every event must contain all impact indicators; rather, a record

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

must contain at least one of the five core impact indicators: deaths and missing persons, affected population, affected cropland area, collapsed dwellings, or direct economic losses. Fields marked as "No" are optional and are recorded only when the original source provides the corresponding information. Missing values are recorded as null, indicating that the original source did not report the information; they should not be interpreted as zero.

125 **Table 2. All fields of CFDID.**

No.	Field Name	Description	Data type	Unit	Required
1	Event level	Hierarchical record level. L1 represents a parent-event record, L2 represents a sub-event record nested within an L1 record, and L3 represents a more detailed record nested within an L2 record.	Categorical	/	Yes
2	Event ID	Unique identifier of each event record. L1 records use "YYYY-NNN"; L2 records use "YYYY-NNN-SS"; and L3 records use "YYYY-NNN-SS-GG", where "NNN", "SS", and "GG" indicate the event, sub-event, and sub-sub-event sequence numbers, respectively.	String	/	Yes
3	Start year	Year when flood impacts began.	Integer	/	Yes
4	Start month	Month when flood impacts began.	Integer	/	Yes
5	Start day	Day when flood impacts began.	Integer	/	No
6	End year	Year when direct impacts ended.	Integer	/	No
7	End month	Month when direct impacts ended.	Integer	/	No
8	End day	Day when direct impacts ended.	Integer	/	No
9	Province	31 provinces in the Chinese mainland; Hong Kong, Macao, and Taiwan are not included.	List	/	Yes
10	Specific location	Original place names such as cities, counties, townships, villages, or river basins.	String	/	No
11	Disaster category	Rainstorm flood, typhoon, snowmelt flood, ice-jam flood, or dam-break flood.	Categorical	/	Yes
12	Typhoon information	Typhoon number or name.	String	/	Semi-required
13	Deaths and missing persons	Number of people killed or missing as a direct result of the flood disaster.	Non-negative integer	persons	Semi-required
14	Affected population	Number of people directly affected by the disaster.	Non-negative decimal	thousand persons	Semi-required
15	Affected cropland area	Sown area of crops with yield reduction of more than 10%.	Non-negative decimal	thousand ha	Semi-required
16	Damaged cropland area	Sown area of crops with yield reduction of more than 30%.	Non-negative decimal	thousand ha	No
17	Failed cropland area	Sown area of crops with yield reduction of more than 80%.	Non-negative decimal	thousand ha	No
18	Collapsed dwellings	Number of dwelling rooms that collapsed or were severely damaged and required reconstruction.	Non-negative decimal	thousand rooms	Semi-required
19	Damaged dwellings	Number of dwelling rooms structurally damaged but repairable.	Non-negative decimal	thousand rooms	No
20	Direct economic losses	Direct economic losses reported at current prices.	Non-negative decimal	million CNY	Semi-required
21	Direct economic losses (2023 constant prices)	Direct economic losses adjusted to 2023 prices using CPI.	Non-negative decimal	million CNY	Semi-required
22	Data source	Source document(s) from which the information was obtained.	List	/	Yes

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

CFDID retains direct economic losses at current prices as reported in the original sources, and provides losses adjusted to 2023 constant prices. The adjustment uses annual consumer price index (CPI), which is consistent with the treatment in EM-DAT and Wikimpacts (Delforge et al., 2025; Li et al., 2026). The CPI data used for this adjustment were obtained from the China Statistical Yearbooks for the corresponding years (NBS, 2025). For event i occurring in year t, the current-price loss is $Impact_{i,t}^{loss}$ and the 2023 constant-price loss is calculated as:

$$Impact_{i,2023}^{loss} = Impact_{i,t}^{loss} \times \frac{CPI_{2023}}{CPI_t} \quad (1)$$

where CPI_t and CPI_{2023} are the CPI indices for year t and 2023, respectively.

2.4 Database construction workflow

The construction workflow of CFDID consists of five steps: document preprocessing, rule-based segmentation, data extraction, data fusion, and data post-processing, as shown in Fig. 1.

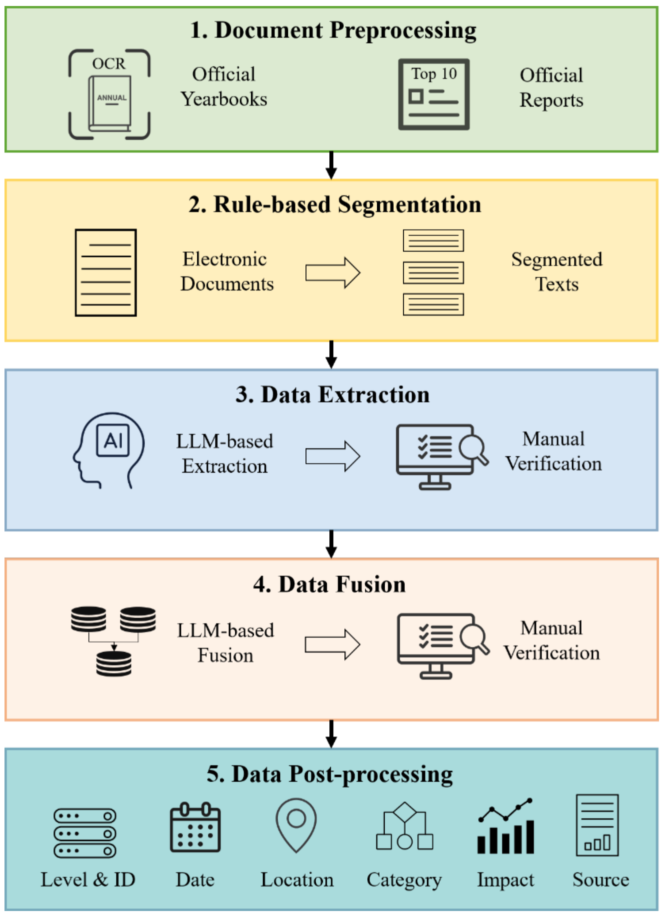


Figure 1. The workflow for constructing CFDID.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

2.4.1 Document preprocessing

The China Water Resources Yearbook and the Yearbook of Meteorological Disasters in China used in this study are paper publications. We first scanned the chapters that contain flood disaster event information, then used an OCR tool "MinerU" (https://mineru.net/) to convert the scanned pages into editable Word documents. For the annual "Top Ten Natural Disasters in China" reports, we directly collected and organized the textual information from government websites.

2.4.2 Rule-based segmentation

Because LLMs have limits on the content length that can be processed accurately in a single call, overly long texts may reduce information extraction accuracy. We therefore segmented each yearbook or report into shorter text blocks using rule-based procedures. Since the original official sources are relatively regular in structure, most texts can be segmented by paragraph breaks, numbered items, or other explicit markers. In total, 962 text blocks were generated, with the longest containing 4,585 Chinese characters and the shortest containing 31 Chinese characters.

It should be noted that the segmentation step only provides text blocks with appropriate length and sufficient context for the LLM, and does not directly define the final events. A single text block may still contain multiple events and impact records.

2.4.3 Data extraction

We called the API of the "deepseek-v4-pro" LLM (https://api-docs.deepseek.com/) to automatically extract temporal and spatial information, disaster category, typhoon information, and impact indicators from the 962 segmented texts. The prompt required the model to output JSON objects following a fixed field structure; the core prompt is provided in Appendix B. The LLM outputs were then manually checked and corrected by multiple team members. To ensure consistency in data verification, all checked records were finally reviewed and confirmed by the first author. Data source information was then assigned to each extracted record based on the corresponding source document.

2.4.4 Data fusion

Duplicate, overlapping, and nested records may exist both within the same source and across the China Water Resources Yearbook, the Yearbook of Meteorological Disasters in China, and the "Top Ten Natural Disasters in China". We therefore conducted data fusion in two stages. First, within-source consistency checks were performed to identify duplicate or nested records and to verify the hierarchical relationships. Second, cross-source fusion was conducted to integrate complementary records describing the same or nested flood disaster events across different source documents in the same year.

For both within-source checking and cross-source fusion, candidate duplicate or nested record pairs were first identified using event date and location. The screening criteria were: (1) the same year, (2) start dates differing by no more than 10 days, and (3) provinces that were identical, overlapping, or nested. The "deepseek-v4-pro" LLM then analyzed the original texts of

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

candidate pairs and classified their relationships as the same event, different events, nested events, or uncertain relationships. The LLM-assisted results were manually reviewed by the first author.

When two records were determined to describe the same event, complementary information was fused. When conflicting values were reported for the same event, we resolved the conflicts based on a source-priority hierarchy reflecting the authority and responsibilities of the reporting departments: "Top Ten Natural Disasters in China", Yearbook of Meteorological Disasters in China, and China Water Resources Yearbook. When records were identified as potentially nested or uncertain by the LLM, we manually reviewed the original texts to determine the hierarchical relationships. Confirmed nested records were then assigned hierarchical levels as L1, L2, or L3 according to the parent–sub-event structure.

2.4.5 Data post-processing

After data fusion, only event records containing at least one of the five core impact indicators were retained, including deaths and missing persons, affected population, affected cropland area, collapsed dwellings, and direct economic losses. Event IDs were then assigned according to the chronological order of occurrence and the hierarchical event level. Specifically, L1 records were numbered chronologically within each year, and L2 and L3 records were assigned suffixes based on their corresponding parent records.

2.5 Coverage assessment

The China Water Resources Yearbook (1994-2001) and the Yearbook of Meteorological Disasters in China (2002-2023) report annual national totals for five core flood impact indicators during 1994–2023. We used the annual coverage ratio (ACR) and total coverage ratio (TCR) to quantify how much of the official annual totals are covered by CFDID and by international disaster impact databases. Different databases use different event structures. For CFDID, only L1 records were used in the numerator, and nested L2 and L3 records were excluded from the coverage calculation to avoid double counting between parent events and their sub-events. For Wikimpacts, which also has a hierarchical structure, we used L2 national-level records for China. For EM-DAT and DFO, which do not have hierarchical event levels, the original event records were used directly. Coverage ratios were calculated only for indicators available in each database. For database *d*, indicator *k*, and year *t*, ACR is defined as:

$$ACR_t^{d,k} = \frac{\sum_{i=1}^{n_t} Impact_{i,t}^{d,k}}{Official_t^k} \times 100\% \tag{2}$$

TCR evaluates the overall coverage during 1994–2023 and is calculated as:

$$TCR^{d,k} = \frac{\sum_{t=1994}^{2023} \sum_{i=1}^{n_t} Impact_{i,t}^{d,k}}{\sum_{t=1994}^{2023} Official_t^k} \times 100\% \tag{3}$$

where $Impact_{i,t}^{d,k}$ denotes the reported value of indicator k for event i in year t in database *d*, and $Official_t^k$ denotes the corresponding official annual total of indicator k in year *t*.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

3 Results

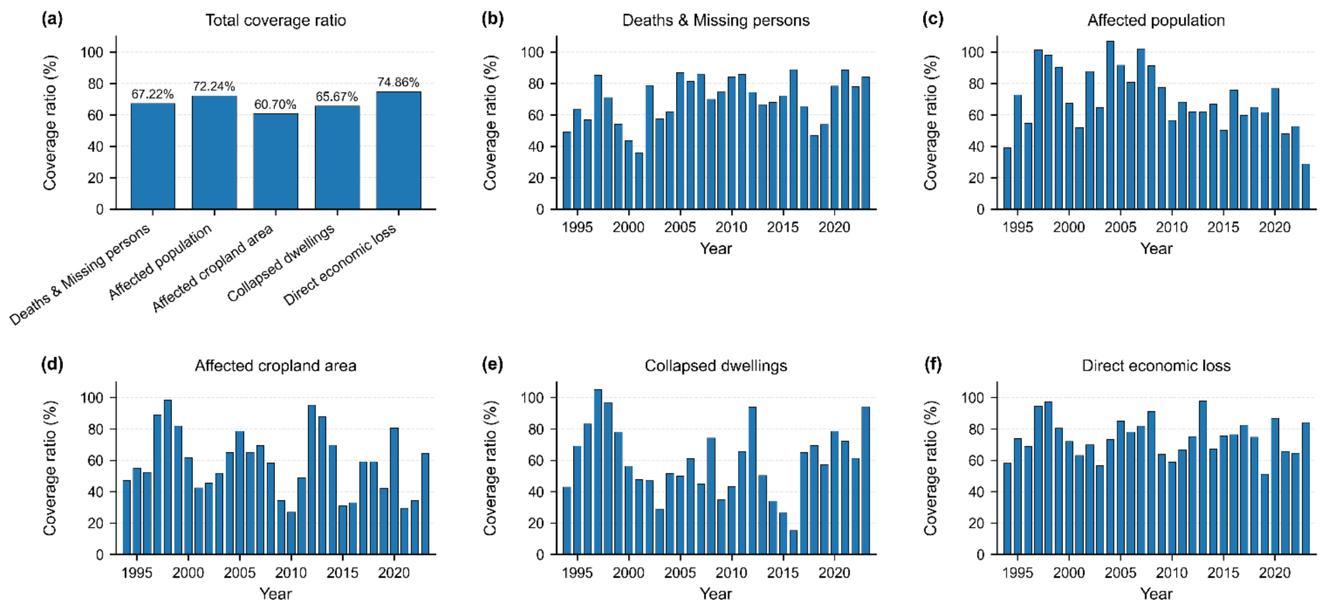
3.1 Overview of CFDID

From 1994 to 2023, CFDID contains 1716 flood disaster event records across the Chinese mainland, including 1391 L1 event records, 300 L2 event records, and 25 L3 event records. Each event record contains start-year, start-month, and province information. Among the 1716 records, 1205 records contain deaths and missing persons; 1419 records contain affected population; 1241 records contain affected cropland area; 87 records contain damaged cropland area; 397 records contain failed cropland area; 969 records contain collapsed dwellings; 397 records contain damaged dwellings; and 1443 records contain direct economic losses.

Among the 1391 L1 event records, 143 contain nested L2 event records, and 21 of these L2 event records further contain nested L3 event records. The relatively small number of L2 and L3 records reflects the fact that the source documents used in this study are national-level government publications. These documents mainly report flood impacts at the provincial level, while sub-provincial details are recorded less systematically. Therefore, L2 and L3 records are mainly retained as supplementary information for detailed analyses of specific local events, whereas the following analyses are based on L1 event records unless otherwise stated.

3.2 Coverage against official annual statistics

Compared with official annual national flood disaster totals, the L1 records in CFDID provide substantial coverage for the five core impact indicators. During 1994–2023, the total coverage ratios for deaths and missing persons, affected population, affected cropland area, collapsed dwellings, and direct economic losses are 67.22%, 72.25%, 60.70%, 65.68%, and 74.86%, respectively (Fig. 2a). Comparisons with international disaster impact databases are provided in Sect. 4.1.



https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Figure 2. Coverage ratios of CFDID against official annual national totals for five core flood impact indicators. Coverage ratios are calculated using L1 records only. (a) Total coverage ratios during 1994–2023. (b–f) Annual coverage ratios for deaths and missing persons, affected population, affected cropland area, collapsed dwellings, and direct economic losses.

The annual coverage ratios for the five core impact indicators vary substantially among indicators and years (Fig. 2b–f). The annual coverage ratio ranges from 35.70% to 88.80% for deaths and missing persons, from 28.83% to 107.00% for affected population, from 28.98% to 98.41% for affected cropland area, from 15.25% to 105.19% for collapsed dwellings, and from 51.22% to 97.70% for direct economic losses. Some annual coverage ratios exceed 100%, which may reflect differences in statistical approaches between event-scale records and official annual totals. Similar discrepancies are also observed in EM-DAT when event-level records are compared with official annual totals (Fig. 7).

3.3 Composition by disaster category

Across disaster categories, L1 event records in CFDID are dominated by rainstorm floods and typhoons (Fig. 3a). Among the 1391 L1 event records, 779 are rainstorm flood events, accounting for 56.00%, and 601 are typhoon events, accounting for 43.21%. Snowmelt floods, ice-jam floods, and dam-break floods are much less frequent, with 5, 2, and 4 events, respectively, together accounting for less than 1% of all L1 event records.

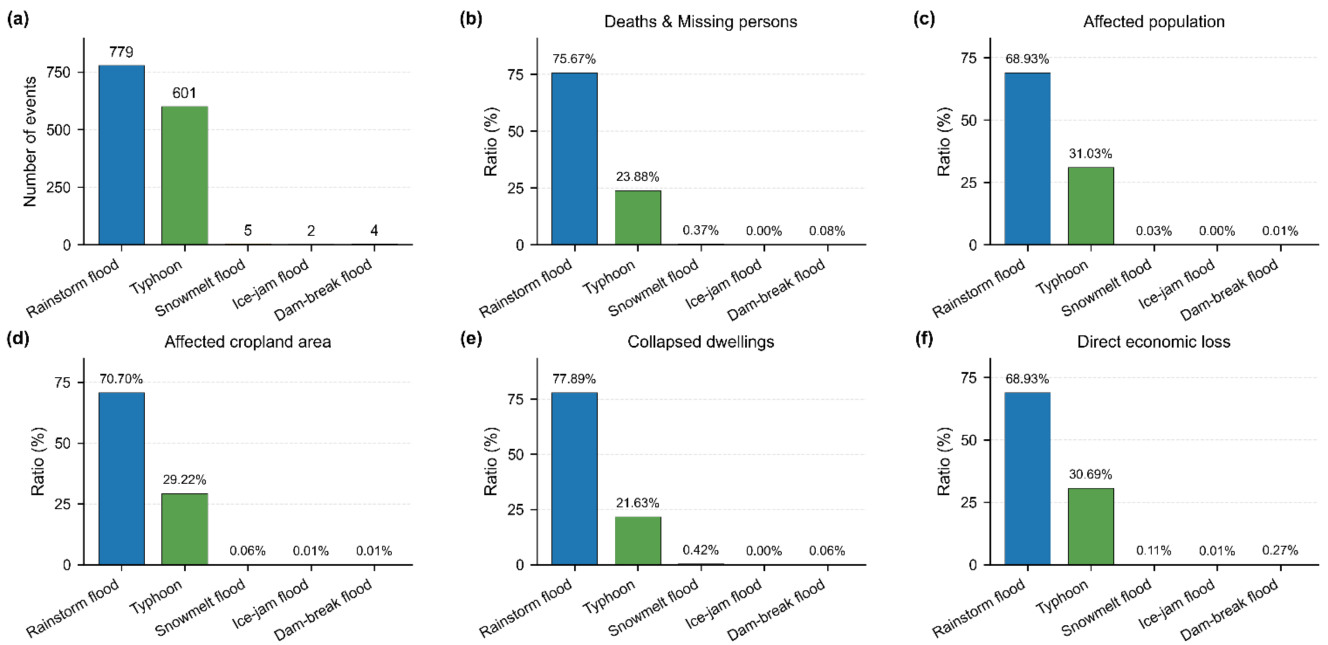


Figure 3. Number of L1 event records and shares of impact indicators by flood disaster category during 1994–2023. (a) Number of L1 event records. (b–f) Shares of deaths and missing persons, affected population, affected cropland area, collapsed dwellings, and direct economic losses by disaster category.

Different disaster categories also contribute differently to disaster impacts. Rainstorm floods account for the largest shares of all five core impact indicators: 75.67% of deaths and missing persons, 68.93% of affected population, 70.70% of affected cropland area, 77.89% of collapsed dwellings, and 68.93% of direct economic losses (Fig. 3b–f). Typhoon events rank second, contributing 23.88% of deaths and missing persons, 31.03% of affected population, 29.22% of affected cropland area, 21.63% of collapsed dwellings, and 30.69% of direct economic losses. Snowmelt floods, ice-jam floods, and dam-break floods contribute only small shares to the five core impact indicators.

3.4 Temporal distribution

The annual number of L1 flood disaster event records in CFDID shows substantial interannual variability (Fig. 4a). During 1994–2023, the largest number of records occurred in 2008, with 106 events, followed by 2009 and 2007, each with 102 events, and 2005, with 94 events. Overall, the number of recorded events was relatively high during the second decade, 2004–2013, with an average of 76.3 events per year. By contrast, the averages for the first decade, 1994–2003, and the third decade, 2014–2023, were lower, at 27.2 and 35.6 events per year, respectively. This pattern likely reflects differences in the level of detail across the source documents and should not be interpreted as direct evidence of long-term changes in hydrological flood frequency.

The annual distribution also differs among disaster categories (Fig. 4b–f). Rainstorm floods are the dominant category in most years, with the highest annual count of 80 events in 2009. Typhoon-related flood records vary considerably among years, with relatively high numbers in 2005, 2013, 2016, and 2018. Snowmelt floods, ice-jam floods, and dam-break floods are rare and occur intermittently.

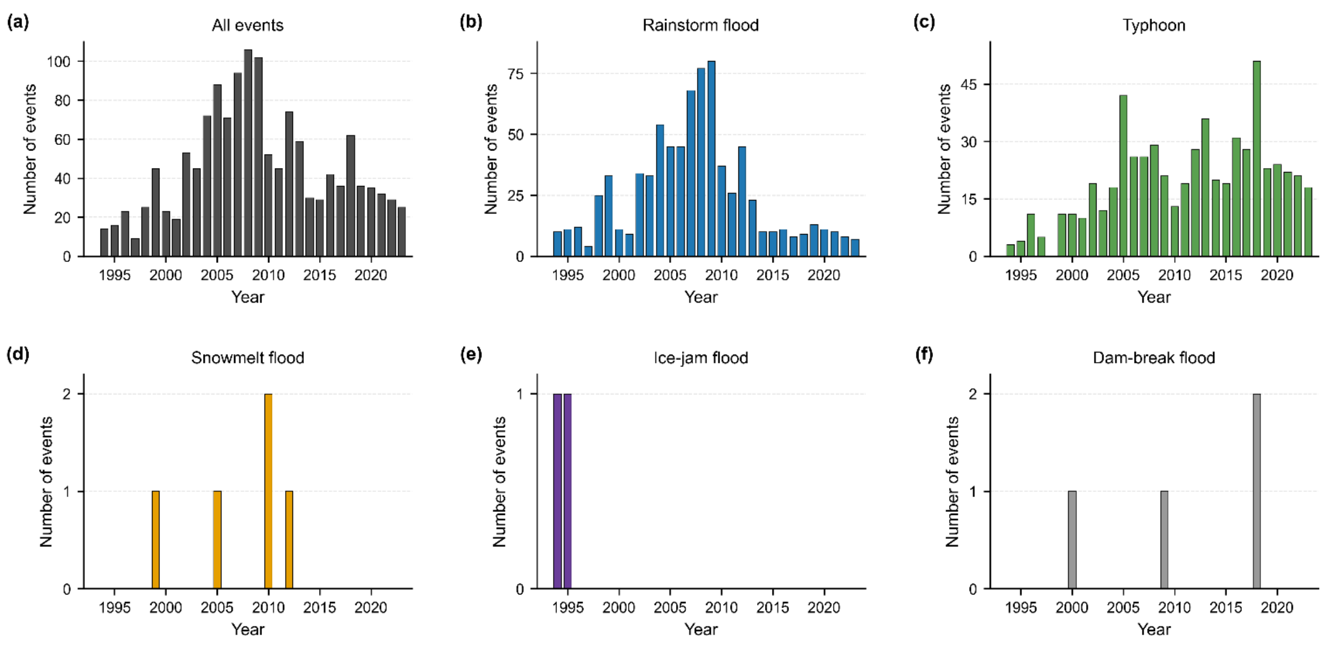


Figure 4. Annual numbers of L1 flood disaster event records during 1994–2023. (a) All flood-related disaster categories. (b–f) Five disaster categories: rainstorm flood, typhoon, snowmelt flood, ice-jam flood, and dam-break flood.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Seasonally, L1 flood disaster event records are highly concentrated from May to October (Fig. 5a). During 1994–2023, 1335 events occurred during this period, accounting for 95.97% of all L1 event records. July has the largest number of records, with 384 events, followed by August (381), June (222), September (181), May (94), and October (73). Seasonal patterns differ among disaster categories. Rainstorm floods mainly occur from May to August and peak in July, with 238 events. Typhoons mainly occur from June to October and peak in August, with 229 events. This seasonality is consistent with the concentrated summer rainfall in the East Asian monsoon region and the seasonal influence of western North Pacific typhoons on China (Shan et al., 2023).

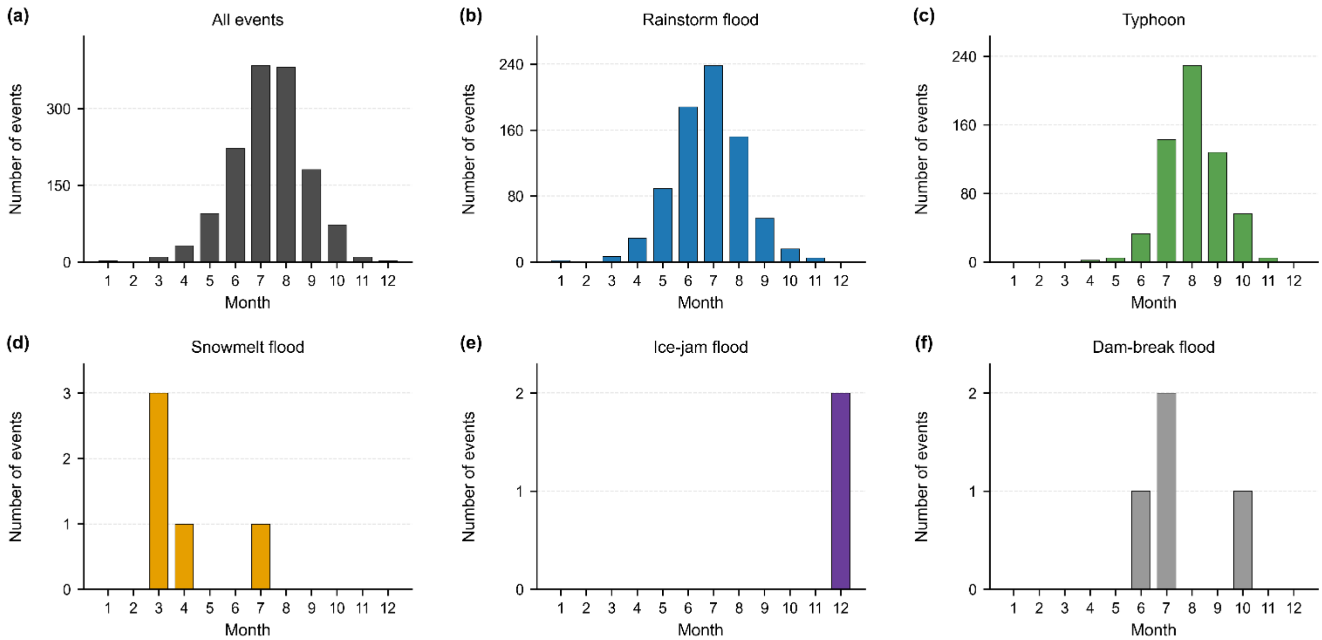


Figure 5. Monthly distribution of L1 flood disaster event records during 1994–2023. (a) All flood-related disaster categories. (b–f) Five disaster categories: rainstorm flood, typhoon, snowmelt flood, ice-jam flood, and dam-break flood.

3.5 Spatial distribution

Provincial statistics show clear spatial differences in L1 flood disaster event records (Fig. 6a). From 1994 to 2023, the 1391 L1 event records contain 1820 province-level records. Guangdong has the largest number of province-level records, with 155 records, followed by Guangxi (135), Sichuan (120), Fujian (115), Hunan (115), Jiangxi (113), Zhejiang (106), and Hubei (106). These high-count provinces are mainly located in the Pearl River Basin, the Yangtze River Basin, and the Southeast Rivers Basin, all of which are largely situated in southern China. The total number of province-level records is larger than the number of L1 event records because some L1 flood disaster events affected multiple provinces.

Different disaster categories show strong spatial heterogeneity. Rainstorm floods are mainly distributed in the Yangtze River Basin and the Southwest Rivers Basin. Sichuan has the highest number of rainstorm flood records, with 119 records, followed by Hubei (98), Hunan (95), and Guizhou (85) (Fig. 6b). Typhoon records are concentrated in the Pearl River Basin

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

and the Southeast Rivers Basin, with the largest number of records in Guangdong (100), followed by Fujian (71), Guangxi (70), Hainan (67), and Zhejiang (63) (Fig. 6c). Snowmelt flood records are found only in Xinjiang, with five records (Fig. 6d). Ice-jam flood records are mainly distributed along the Yellow River, including Shaanxi and Inner Mongolia (Fig. 6e). Dam-break flood records are mainly found in western China, with three records in Xizang and one record each in Xinjiang, Yunnan, and Qinghai (Fig. 6f).

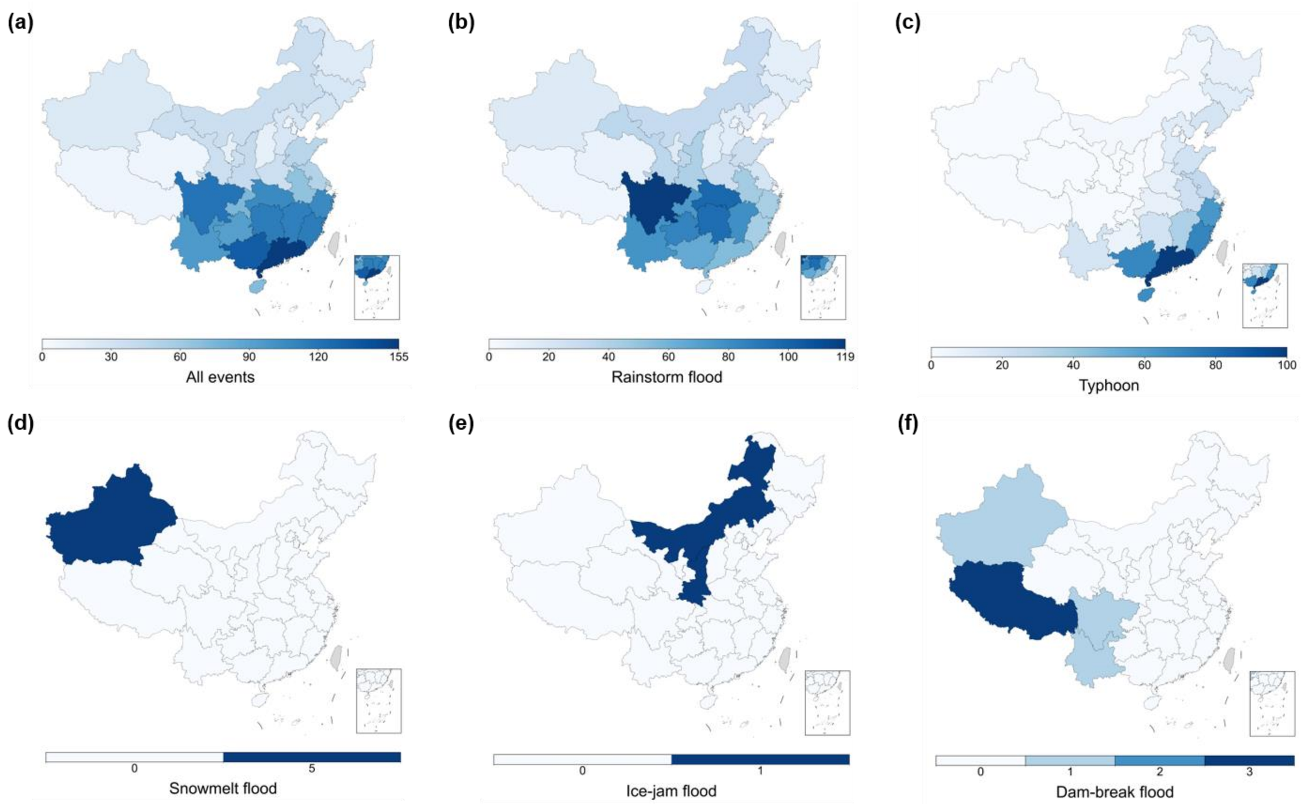


Figure 6. Provincial distribution of the number of L1 flood disaster event records during 1994–2023. (a) All flood-related disaster categories. (b–f) Rainstorm flood, typhoon, snowmelt flood, ice-jam flood, and dam-break flood.

4 Discussion

4.1 Comparison with international flood impact databases

Compared with major international flood impact databases, CFDID records substantially more flood disaster events in the Chinese mainland (Table 3). To ensure consistency with the event hierarchy used in Wikimpacts, the comparison uses L1 records for CFDID and L2 national-level China records for Wikimpacts. During 1994–2023, CFDID contains 1391 L1 event records, exceeding the number of flood records for the Chinese mainland in EM-DAT, DFO, and Wikimpacts. For the five core impact indicators, CFDID has higher coverage of affected population, collapsed dwellings, and direct economic losses than the international databases. In addition, CFDID provides agricultural impact indicators, including affected cropland area,

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

damaged cropland area, and failed cropland area, which are rarely available in global databases. These indicators are particularly useful for evaluating crop-specific flood impact models (Qiu et al., 2025; Zhang et al., 2026).

The international databases each have their own strengths and can complement CFDID. EM-DAT has slightly higher coverage of deaths and missing persons than CFDID. Table 3 also shows that EM-DAT is the most comprehensive among the international databases considered here, outperforming DFO and Wikimpacts in the number of records and in the coverage of deaths and missing persons, affected population, and direct economic losses. DFO provides spatial information derived from remote sensing and is valuable for locating high-impact flood events and characterizing affected or inundated areas. Wikimpacts provides a rich set of impact categories, although its China-related event counts and total impact coverage remain relatively low at present. This may partly reflect the limited representation of Chinese-language flood-related news and reports in Wikipedia-based sources (Li et al., 2026).

Table 3. Comparison of event records and total coverage ratios among CFDID and international flood impact databases during 1994–2023. CFDID is compared using L1 records, whereas Wikimpacts is compared using L2 national-level China records.

Indicator	CFDID (L1)	EM-DAT	DFO	Wikimpacts (L2)
Number of event records	1391	434	276	146
Deaths and missing persons	67.22%	70.82%	35.06%	16.49%
Affected population	72.25%	46.49%	/	38.02%
Affected cropland area	60.70%	/	/	/
Collapsed dwellings	65.68%	/	/	7.64%
Direct economic losses	74.86%	51.07%	/	38.84%

We further compare CFDID and EM-DAT with official annual totals for deaths and missing persons, affected population, and direct economic losses (Fig. 7). Overall, CFDID, EM-DAT, and official annual totals show broadly consistent interannual variability, indicating that both databases capture major fluctuations in flood impacts. CFDID and EM-DAT are relatively close for deaths and missing persons, whereas CFDID records higher affected population and direct economic losses than EM-DAT in most years.

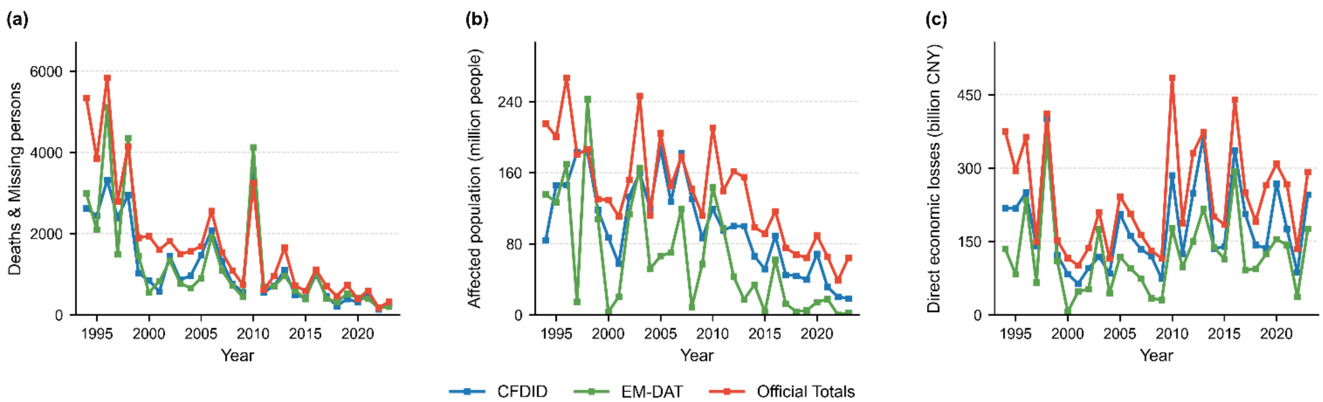


Figure 7. Annual comparison of selected flood impact indicators from CFDID, EM-DAT, and official annual national totals during 1994–2023. CFDID values are based on L1 records. (a) Deaths and missing persons. (b) Affected population. (c) Direct economic losses.

4.2 Potential applications

The main contribution of CFDID lies in transforming flood impact information scattered across official documents into an open, systematic, event-scale structured database, which has several potential applications.

First, CFDID can be used to assess changes in flood impacts in China, including trends in different impact indicators and the contributions of different drivers. Similar analyses have been conducted in Europe (Paprotny et al., 2018, 2025) and the United States (Davenport et al., 2021; Rashid et al., 2023). Second, CFDID can be used to validate flood impact assessment models. Although many studies have simulated or assessed flood impacts, their results often remain highly uncertain because of the lack of systematic validation data (Messori et al., 2026; Wing et al., 2020). Event-scale impact records can help improve model evaluation and support more reliable projections of future flood impacts in China. Third, CFDID can serve as an event inventory for investigating how extreme flood events lead to social and economic impacts, thereby supporting future flood prevention and risk reduction strategies (Garcia et al., 2025). CFDID can also be integrated with other global or regional datasets to provide a more reliable China component for future global studies of flood impacts.

Beyond scientific analysis, CFDID can also support practical risk management. For example, it can provide historical loss information for actuarial pricing and portfolio optimization by insurance and reinsurance companies, thereby contributing to the development of flood insurance and the improvement of insurance coverage in China (Cui et al., 2025).

4.3 Limitations and future improvements

Although CFDID includes more event records and provides higher coverage for most impact indicators than existing international databases, it still has several limitations.

First, CFDID currently relies mainly on central-government official records. Local government yearbooks, news reports, social media, academic papers, insurance company reports, international databases, and remote sensing products have not yet been systematically integrated. As a result, some small-scale flood disaster events may still be omitted, which is a common limitation of most natural disaster impact databases. Expanding data sources could further improve event counts and impact coverage (Fu et al., 2025; Qiu et al., 2025; Shen et al., 2025).

Second, some flood disaster records cover broad spatial extents across multiple administrative regions, and their impacts cannot yet be fully disaggregated to known subregions. Current records also lack precise coordinates, which limits their spatial resolution and spatial analytical potential. Future work may use additional data sources, remote sensing products, and model simulations to support spatial disaggregation and improve geographic attribution. In addition, known spatial information, such as affected provinces, prefectures, counties, and river basins, could be further converted into spatial layers (Rosvold and Buhaug, 2021; Teber et al., 2026), to facilitate spatial analysis and integration with hydrological, socioeconomic, and remote-sensing datasets.

Third, typhoon impacts in CFDID cannot be separated into wind and water components, and flood-related geological hazards cannot always be separated from flood impacts. In addition, CFDID classifies flood disasters according to triggering

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

causes rather than common flood subtypes such as fluvial floods, pluvial floods, flash floods, and coastal floods. This limits subtype-level comparisons with databases from the United States, Japan, Europe, and other regions.

Fourth, CFDID is currently provided in tabular form. Future work will develop a visualization platform based on the spatially enriched database to present geographic information more effectively, add narratives of major flood disasters, and gradually build a knowledge base of flood disaster events in China.

5 Data availability

CFDID is available from the public Zenodo dataset at https://doi.org/10.5281/zenodo.21699973 and will be continuously updated by year (Cui et al., 2026).

6 Code availability

The source code developed in this study is available from the same Zenodo dataset at https://doi.org/10.5281/zenodo.21699973. We also provide an example based on rainstorm disaster records from the 2022 Yearbook of Meteorological Disasters in China, which demonstrates the main processing workflow.

7 Conclusions

In this work, we developed CFDID, a China Flood Disaster Impact Database that systematically compiles event-scale flood impact records for 1994–2023. Based on 66 official documents, CFDID combines OCR, rule-based segmentation, LLM-based information extraction, cross-source data fusion, and manual verification to produce 1716 event records with eight disaster impact indicators.

CFDID shows that flood disaster event records in China are dominated by rainstorm floods and typhoons and exhibit strong seasonality and spatial heterogeneity. Events are concentrated mainly from May to October, and high event counts occur in Guangdong, Guangxi, Sichuan, Fujian, Hunan, Jiangxi, Zhejiang, and Hubei. Compared with major international flood impact databases, CFDID contains more event records and provides higher coverage for most commonly used impact indicators. It also provides agricultural impact indicators that are rarely available in global databases.

Nevertheless, CFDID still has several limitations. It currently relies mainly on national-level official documents, and some small-scale or local flood disaster events may remain omitted. Some records also cover broad spatial extents and cannot yet be fully disaggregated to finer administrative regions or precise geographic coordinates. In addition, typhoon-related records generally report aggregate disaster impacts, and the losses caused by wind, rainfall, flooding, storm surge, and related secondary hazards cannot yet be separated. This should be considered when using the typhoon category for flood-specific impact analysis.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Overall, CFDID provides an open and systematic data basis for flood risk research in China and demonstrates the feasibility of integrating OCR, LLMs, and manual quality control in the construction of disaster impact databases. With future expansion of data sources, improved spatial disaggregation, visualization platform development, and continuous updates, CFDID is expected to support flood risk assessment, model validation, disaster risk reduction decision-making, and flood insurance development.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Open Access Earth System Science Data Discussions

Appendix A: Examples of data sources

Figure A1. Examples of data sources. (a) China Water Resources Yearbook in 2000. (b) Yearbook of Meteorological Disasters in China in 2003. (c) "Top Ten Natural Disasters in China" in 2023.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Appendix B: Core prompt for LLM-based data extraction

You are an expert in extracting flood disaster impact records from Chinese official texts.

Your task is to identify all flood-related disaster event records in the input text and output them as a valid JSON array. Each event record should contain temporal information, spatial information, disaster category, typhoon information, and disaster impact indicators.

Do not infer missing values unless they are explicitly reported in the text. If a field is not reported or cannot be determined, use null. Do not output any text outside the JSON array.

Each event record must contain the following fields:
["Province", "Specific location", "Start year", "Start month", "Start day", "End year", "End month", "End day", "Disaster category", "Typhoon information", "Deaths and missing persons (persons)", "Affected population (thousand persons)", "Affected cropland area (thousand ha)", "Damaged cropland area (thousand ha)", "Failed cropland area (thousand ha)", "Collapsed dwellings (thousand rooms)", "Damaged dwellings (thousand rooms)", "Direct economic losses (million CNY)"]

Event segmentation rules:
1. If different regions were affected during the same period, extract them as separate event records when the text reports them separately.
2. If the same region was affected during different periods, extract them as separate event records.
3. If the text reports both an aggregate event and nested regional or local sub-events, extract both the aggregate event and the sub-events as independent records.

Field rules:
1. Province: extract affected provincial-level regions and translate them into English, such as "Anhui", "Guangdong", "Inner Mongolia", "Xizang", and "Xinjiang". Multiple provinces should be separated by English commas.
2. Specific location: extract and translate original place names such as cities, counties, townships, villages, river basins, and river reaches. For multiple provinces, retain province-level grouping when necessary.
3. Dates: extract start and end year, month, and day. If only year or month is reported, use null for missing date components. If only the start time is reported, set all end-time fields to null.
4. Disaster category: choose one of:
"rainstorm flood", "typhoon", "snowmelt flood", "ice-jam flood", "dam-break flood".

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Classify the event according to the primary triggering mechanism reported in the source text. Typhoon-related rainfall, flooding, storm surge, strong winds, and secondary hazards should be classified as "typhoon". Barrier-lake outburst floods should be classified as "dam-break flood". Landslides, debris flows, or collapses should be included only when explicitly linked to rainfall, flooding, typhoons, snowmelt, ice jams, or dam/dike failures.

5. Typhoon information: if the event is related to a typhoon, record the typhoon number and name as reported. Otherwise, use null.

6. Impact indicators:

- Deaths and missing persons (persons): sum deaths and missing persons if both are reported.
- Affected population (thousand persons): convert all population values to thousand persons.
- Affected cropland area (thousand ha): crops with yield reduction greater than 10%.
- Damaged cropland area (thousand ha): crops with yield reduction greater than 30%.
- Failed cropland area (thousand ha): crops with yield reduction greater than 80%.
- Collapsed dwellings (thousand rooms): collapsed or severely damaged dwelling rooms requiring reconstruction.
- Damaged dwellings (thousand rooms): structurally damaged but repairable dwelling rooms.
- Direct economic losses (million CNY): report current-price direct economic losses and convert all monetary values to million CNY.

Author contributions

SC, NL, and JZ designed the study. SC led the data collection, database construction, analysis, and manuscript preparation. NL and JZ supervised the study and contributed to manuscript revision. SH, ZW, JW, JT, RZ, YL, XY, YT and WS contributed to data collection, extraction, checking, and database revision. All authors discussed the results and contributed to the final manuscript.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Acknowledgements

We thank Hongrui Xiong, Hailong Liu, Weiyu Zhou, and Xiyu Chen of Tsinghua University for their assistance in checking the database records.

Financial support

This research has been supported by the National Key R&D Program of China [2023YFF0804901] and the National Natural Science Foundation of China [52539001].

References

CRED and UNDRR: Human cost of disasters: An overview of the last 20 years (2000-2019), 2020.

Cui, S., Wang, J., Hu, S., Wang, Y., Wang, X., and Zhao, J.: Reducing flood insurance costs by employing geographic risk complementarity, Water Resour. Res., 61, e2024WR038128, https://doi.org/10.1029/2024WR038128, 2025.

Cui, S., Li, N., Hu, S., Wan, Z., Wang, J., Tang, J., Zhang, R., Li, Y., Yu, X., Tan, Y., Shan, W., and Zhao, J.: CFDID: China Flood Disaster Impact Database (1994–2023) (Version 1.0), Zenodo [code and data set] https://doi.org/10.5281/zenodo.21699973, 2026.

Davenport, F. V., Burke, M., and Diffenbaugh, N. S.: Contribution of historical precipitation change to US flood damages, Proc. Natl. Acad. Sci. U.S.A., 118, e2017524118, https://doi.org/10.1073/pnas.2017524118, 2021.

Delforge, D., Wathelet, V., Below, R., Sofia, C. L., Tonnelier, M., Van Loenhout, J. A. F., and Speybroeck, N.: EM-DAT: the Emergency Events Database, Int. J. Disaster Risk Reduct., 124, 105509, https://doi.org/10.1016/j.ijdrr.2025.105509, 2025.

DFO: Global Flood Records (0.9.0), https://doi.org/10.5281/zenodo.19288171, 2026.

Dottori, F., Szewczyk, W., Ciscar, J.-C., Zhao, F., Alfieri, L., Hirabayashi, Y., Bianchi, A., Mongelli, I., Frieler, K., Betts, R. A., and Feyen, L.: Increased human and economic losses from river flooding with anthropogenic warming, Nat. Clim. Change, 8, 781–786, https://doi.org/10.1038/s41558-018-0257-z, 2018.

Fu, S., Schultz, D. M., Lyu, H., Zheng, Z., and Zhang, C.: Creating a national urban flood dataset for China from news texts (2000–2022) at the county level, Hydrol. Earth Syst. Sci., 29, 767–783, https://doi.org/10.5194/hess-29-767-2025, 2025.

Garcia, H. M., Sebastian, A., Fitzmaurice, K. P., Hino, M., Collins, E. L., and Characklis, G. W.: Reconstructing repetitive flood exposure across 78 events from 1996 to 2020 in North Carolina, USA, Earth's Future, 13, e2025EF006026, https://doi.org/10.1029/2025EF006026, 2025.

Jiang, T., Su, B., Huang, J., Zhai, J., Xia, J., Tao, H., Wang, Y., Sun, H., Luo, Y., Zhang, L., Wang, G., Zhan, C., Xiong, M., and Kundzewicz, Z. W.: Each 0.5°C of warming increases annual flood losses in China by more than US$60 billion, Bull. Amer. Meteor. Soc., 101, E1464–E1474, https://doi.org/10.1175/BAMS-D-19-0182.1, 2020.

Kang, S., Yin, J., Slater, L., Liu, P., Sun, F., Liu, D., and Xia, J.: Global flood projection and socioeconomic implications under a deep learning framework, Water Resour. Res., 61, e2024WR037139, https://doi.org/10.1029/2024WR037139, 2025.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Li, N., Thiery, W., Zahra, S., Madruga de Brito, M., Worou, K., Kurfalı, M., Lampe, S., Muñoz, P., Flynn, C., Trigoso, C., Nivre, J., Zscheischler, J., and Messori, G.: Wikimpacts 1.0: a new global climate impact database based on automated information extraction from Wikipedia, Nat. Hazards Earth Syst. Sci., 26, 2609–2636, https://doi.org/10.5194/nhess-26-2609-2026, 2026.

Messori, G., Boyd, E., Nivre, J., and Raffetti, E.: Challenges and opportunities for understanding societal impacts of climate extremes, Earth Syst. Dynam., 17, 199–208, https://doi.org/10.5194/esd-17-199-2026, 2026.

Munich Re: NatCatSERVICE: https://www.munichre.com/en/solutions/for-industry-clients/natcatservice.html, last access: 28 July 2026.

NBS (National Bureau of Statistics of China): China Statistical Yearbook, China Statistics Press, Beijing, 2025.

Paprotny, D., Sebastian, A., Morales-Nápoles, O., and Jonkman, S. N.: Trends in flood losses in Europe over the past 150 years, Nat. Commun., 9, 1985, https://doi.org/10.1038/s41467-018-04253-1, 2018.

Paprotny, D., Tilloy, A., Treu, S., Buch, A., Vousdoukas, M. I., Feyen, L., Kreibich, H., Merz, B., Frieler, K., and Mengel, M.: Attribution of flood impacts shows strong benefits of adaptation in Europe since 1950, Sci. Adv., 11, eadt7068, https://doi.org/10.1126/sciadv.adt7068, 2025.

Qiu, J., Yang, X., Zheng, Z., and Tarolli, P.: High-resolution mapping of China's flooded croplands, Science Bulletin, 70, 1165–1173, https://doi.org/10.1016/j.scib.2025.01.053, 2025.

Rashid, M. M., Wahl, T., Villarini, G., and Sharma, A.: Fluvial Flood Losses in the Contiguous United States Under Climate Change, Earth's Future, 11, e2022EF003328, https://doi.org/10.1029/2022EF003328, 2023.

Rosvold, E. L. and Buhaug, H.: GDIS, a global dataset of geocoded disaster locations, Sci. Data, 8, 61, https://doi.org/10.1038/s41597-021-00846-6, 2021.

Shan, K., Lin, Y., Chu, P.-S., Yu, X., and Song, F.: Seasonal advance of intense tropical cyclones in a warming climate, Nature, 623, 83–89, https://doi.org/10.1038/s41586-023-06544-0, 2023.

Shen, D., Gu, H., Chen, W., Zhang, C., Xiao, S., and Zhang, S.: 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, 2025.

Swiss Re: Sigma Explorer: https://sigma.swissre.com/home, last access: 28 July 2026.

Teber, K., Weynants, M., Gans, F., and Mahecha, M. D.: Geo-Disasters: Geocoding climate-related events in the international disaster database EM-DAT, Big Earth Data, 10, 303–318, https://doi.org/10.1080/20964471.2025.2576274, 2026.

Wing, O. E. J., Pinter, N., Bates, P. D., and Kousky, C.: New insights into US flood vulnerability revealed from flood insurance big data, Nat. Commun., 11, 1444, https://doi.org/10.1038/s41467-020-15264-2, 2020.

Winsemius, H. C., Aerts, J. C. J. H., van Beek, L. P. H., Bierkens, M. F. P., Bouwman, A., Jongman, B., Kwadijk, J. C. J., Ligtvoet, W., Lucas, P. L., van Vuuren, D. P., and Ward, P. J.: Global drivers of future river flood risk, Nat. Clim. Change, 6, 381–385, https://doi.org/10.1038/nclimate2893, 2016.

Zhang, S., Zhou, L., Liang, H., Obulkasim, O., and Dai, Y.: Underestimated agricultural losses due to flooding, Sci. Adv., 12, eaed2754, https://doi.org/10.1126/sciadv.aed2754, 2026.

https://doi.org/10.5194/essd-2026-607
Preprint. Discussion started: 12 August 2026

Open Access Earth System Science Data Discussions

Zhang, W. and Zhou, T.: Increasing impacts from extreme precipitation on population over China with global warming, Sci. Bull., 65, 243–252, https://doi.org/10.1016/j.scib.2019.12.002, 2020.