China Scientific Data
Vol.3, No.3, 2018

A dataset of scientific literature on floods, 1990 – 2017

Zhang Hongyue[1,2,3], **Li Guoqing**[4*], **Huang Mingrui**[2,3], **Qing Xiuling**[5], **Zhang Huarong**[6]

1. Minjiang University, Fuzhou 350108, P. R. China;
2. University of Chinese Academy of Sciences, Beijing 100049, P. R. China;
3. Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, P. R. China;
4. Key Laboratory of Earth Observation Hainan Province, Sanya 572029, P. R. China;
5. National Science Library, Chinese Academy of Sciences, Beijing 100090, P. R. China;
6. Shandong Water Conservancy Vocational College, Rizhao 276826, P. R. China

* Email: ligq@radi.ac.cn

ARTICLE DOI:
10.11922/csdata.2018.0020.en

DATA DOI:
10.11922/sciencedb.591

SUBJECT CATEGORY:
Earth sciences

RECEIVED:
May 29, 2018

RELEASED:
June 4, 2018

PUBLISHED:
September 6, 2018

Abstract: With an increasing number of scientific achievements published, it is particularly important to conduct literature-based knowledge discovery and data mining. Flood, as one of the most destructive natural disasters, has been the subject of numerous scientific publications. On January 1, 2018, we conducted literature data collection and processing on flood research and categorized the retrieved paper records into Whole SCI Dataset (WS) and High-Citation SCI Dataset (HCS). These data sets can serve as basic data for bibliometric analysis to identify the status of global flood research during 1990 – 2017. Our study shows that while the Chinese Academy of Sciences was the most productive institution during this period, the United States was the most productive country. Besides, our keyword analysis reveals the potential popular issues and future trends of flood research.

Keywords: literature data sets; flood; WS; HCS

Dataset Profile

Chinese title	1990–2017 年全球洪涝灾害研究 SCIE 文献数据
English title	A dataset of scientific literature on floods, 1990 – 2017
Data corresponding author	Li Guoqing (ligq@radi.ac.cn)
Data authors	Zhang Hongyue, Li Guoqing, Huang Mingrui, and Qing Xiuling
Time range	1990 – 2017
Data volume	6.46 MB (including 8115 records in Whole SCI Dataset and 150 records in High-Citation SCI Dataset)
Data format	.xls

Data service system	<http://www.sciencedb.cn/dataSet/handle/637>
Sources of funding	National Key Research and Development Program of China (2016YFE0122600)
Dataset composition	This dataset consists of two compressed ZIP files, namely, “WS.zip” and “HCS.zip”. The data are saved in XLS format. ● “WS.zip” refers to the Whole SCI Dataset that stores a full list of the papers collected. ● “HCS.zip” represents the High-Citation Dataset that stores a collection of papers, each with over 100 citations.

1. Introduction

In recent years, floods have been among the most frequent disasters worldwide. According to the German Reinsurance Company statistics, flood has been one of the most significant natural disasters in the world.[1] Due to its long time span and wide geographical scope, it is difficult to define a flood event with a specified time and location. In most cases, a series of flood events happened together, which requires scientific publications to be used as a group to illustrate empirical findings.

Literature-based knowledge discovery has been applied in several research domains, such as medical and biological research,[2,3] as well as information and science development studies.[4,5] Due to the increasing body of texts and the open-access policies of numerous journals, literature mining is becoming useful for both hypothesis generation and scientific discovery.[3] However, due to semantic heterogeneity, literature-based data acquisition is restricted. To the best of our knowledge, there are few studies collecting scientific publications on disaster events.

To unearth the hidden knowledge on flood research, our study adopts literature-based knowledge discovery to obtain Whole SCI Dataset and High-Citation Dataset by means of data retrieval and processing.

2. Data collection and processing

2.1 Overview

Among the most popular Web-based literature databases, such as Google Scholar, Web of Science (WoS), Scopus, and PubMed, WoS (Thomson Reuters) was one of the most frequently used by scientists from natural sciences in recent years.[6] The WoS Core Collection has several citation indexes, including Science Citation Index (SCI), Social Sciences Citation Index, and Arts and Humanities Citation Index.[7] Despite the continued emergence of bibliometric databases, SCI remains arguably the most reliable one for retrieving scientific output.[8]

In this study, we collected data from the SCI database at the Web of Science (WoS, formerly known as ISI Web of Knowledge) Core Collection. Although WoS guarantees a relatively stable search environment with clearly defined lists of indexed journals,[9] its searching conditions are restricted to metadata such as titles, keywords, and abstracts.[9] To obtain a full and accurate list of texts, we narrowed down our queries by setting four parameters.

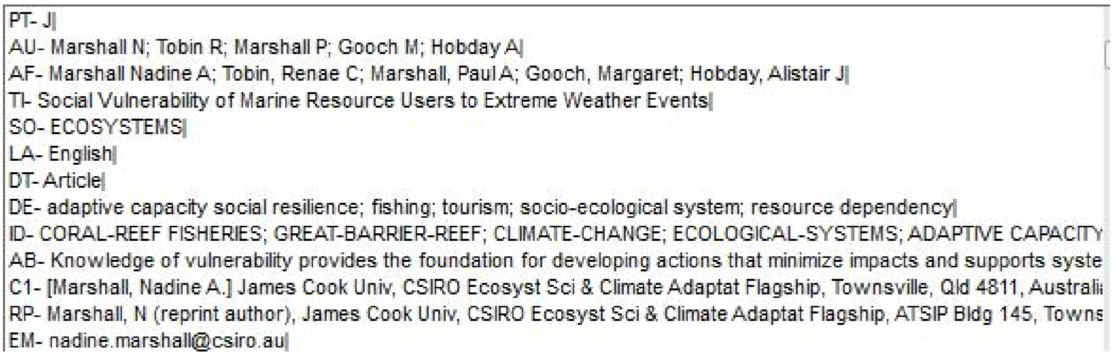
The first parameter is research topic. In advanced search, the "topic" field tag was set as "TS = (((flood near (event or hazard or disaster)) or (flood near/3 (inundation or damage or risk or zone))) not (volcano or basalt))," where "near/n" was used to find records containing all terms within a certain number of words (n) of each other (i.e., up to n words can be inserted between two terms). Considering that volcano and basalt papers are often published in the category of flood, "not (volcano or basalt)" was added to the searching formula to exclude irrelevant records.

The second parameter is time span. A time span was set to restrict the publication time from January 1990 to December 2017.

The third parameter is publication type. "Article" was selected as the only publication type of this search. The reason is that other publication types (e.g., discussion, biographical item, and editorial) do not provide sufficient attribute information and thus cannot meet our data requirements.

The last parameter is additional citation indexes. Science Citation Index Expanded was included in the scope of our search.

For a full description of the papers, full records of each paper, including cited references, were downloaded and stored in TXT format. Descriptive fields include abstract, authors, country, publication year, institution, research area, WoS subject category, journal, title, and source. An example of the description fields is presented in Figure 1.



```
PT- J|
AU- Marshall N; Tobin R; Marshall P; Gooch M; Hobday A|
AF- Marshall Nadine A; Tobin, Renae C; Marshall, Paul A; Gooch, Margaret; Hobday, Alistair J|
TI- Social Vulnerability of Marine Resource Users to Extreme Weather Events|
SO- ECOSYSTEMS|
LA- English|
DT- Article|
DE- adaptive capacity social resilience; fishing; tourism; socio-ecological system; resource dependency|
ID- CORAL-REEF FISHERIES; GREAT-BARRIER-REEF; CLIMATE-CHANGE; ECOLOGICAL-SYSTEMS; ADAPTIVE CAPACITY
AB- Knowledge of vulnerability provides the foundation for developing actions that minimize impacts and supports syste
C1- [Marshall, Nadine A.] James Cook Univ, CSIRO Ecosyst Sci & Climate Adaptat Flagship, Townsville, Qld 4811, Australi
RP- Marshall, N (reprint author), James Cook Univ, CSIRO Ecosyst Sci & Climate Adaptat Flagship, ATSIP Bldg 145, Towns
EM- nadine.marshall@csiro.au|
```

Figure 1 Descriptive fields for the articles collected

Notes: In the figure, the fields were named according to the criteria of Web of Science. PT denotes publication type (J = journal; B = book; S = series; P = patent); AU – authors; AF – authors' full name; TI – document title; SO – publication name; LA – language; DT – document type; DE – author keywords, which denotes keywords given by authors; ID – keywords plus, which denotes keywords generated by ISI; AB – abstract; C1 – authors' address; RP – reprint address; EM – e-mail address.

A total of 15,935 records were obtained on January 1, 2018.

2.2 Data processing

The processing tools adopted by this study included Thomson Data Analyzer (TDA)[10] and Microsoft Excel. TDA is a strong text mining software able to mine text information from multiple fields and offer visualized comprehensive analyses. Moreover, TDA enables a systematic organization of the literature information retrieved.[11] Microsoft Excel was used to rearrange the exported data.

Figure 2 presents the data processing flowchart used by this study, which includes two major steps: record removal and field cleaning. Natural language processing (NLP) was adopted to perform data cleaning. NLP mainly involved tokenization, stop word removal, stemming, lemmatization, and field merging.

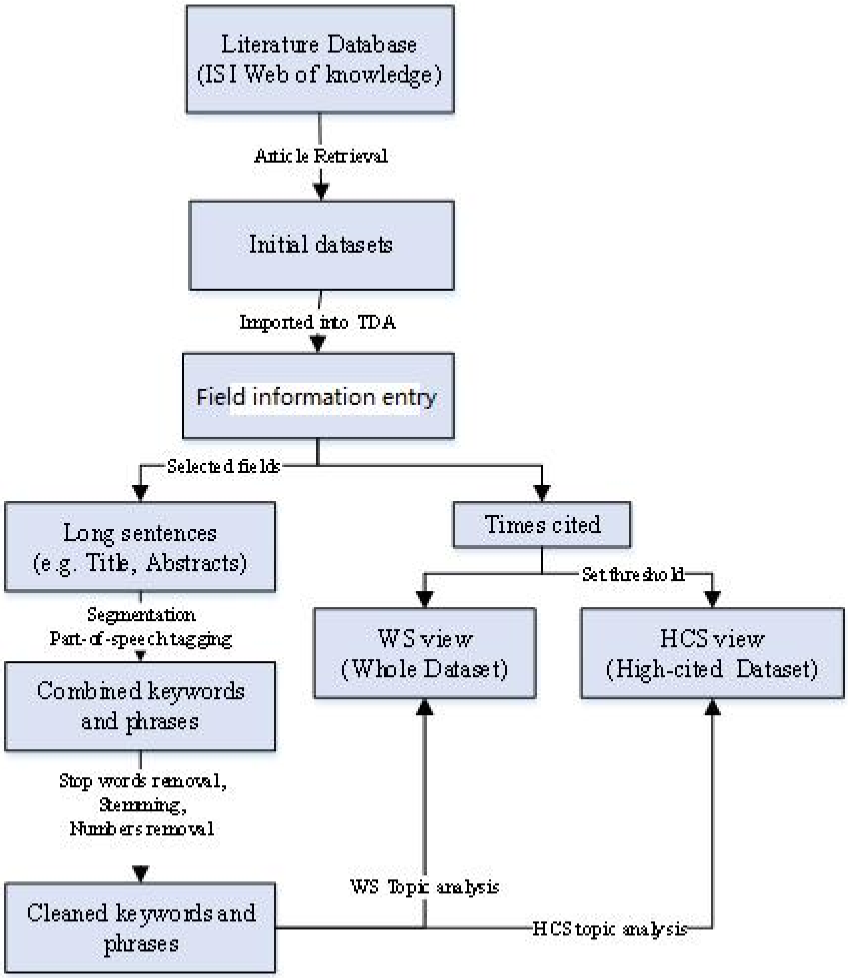


Figure 2 Processing flowchart of initial datasets (raw data)

Tokenization is the process of breaking up a given string into a series of subsequences, such as words, keywords, phrases.[12] Each of the subsequences is called a “token”. In this process, some special symbols, such as punctuation, etc., will be removed. In some cases, some common words are of little value when the document matches the user’s needs and thus need to be completely removed from the vocabulary. These words are called *stop words*.[12] In literature retrieval, stemming and lemmatization are different in meaning. *Stemming* usually refers to a very crude heuristic process that removes the affixes at both ends of the word. This process often involves the removal of derived affixes.[12] *Lemmatization* usually refers to the process of using the vocabulary and the morphological analysis to remove the inflection affixes,[12] thereby returning the original form of the words or the words in the dictionary, and the returned result is called a lemma. Raw data of this dataset were downloaded from the WoS database and stored in text format, which were then organized into several attribute fields, including title, authors, abstract, keywords, journal, publication year, and country. As each attribute field reflects distinct paper information, researchers can select specific fields to perform knowledge mining.

The raw records obtained were first imported into TDA by using an import filter, and the records were then segmented and stored in respective fields. A literature analysis was performed by using NLP and statistical methods. NLP modules in TDA were utilized to process the metadata fields of initial literature datasets, including title, abstract, authors, keywords, and keywords plus. The goal of NLP is to use rules to process text for specific purposes (such as translation, extraction of assertion, and summarization), where the rules may be predefined or learned through supervised or unsupervised methods.[13] First, long sentences, including the title and abstracts, were segmented, and part-of-speech tagging was conducted on the segmented words and phrases. The segmented words and phrases were then further processed. Regular expression is an efficient tool for extending retrieval. The Fuzzy Matching Editor allows the user to tailor TDA's cleanup algorithms to suit the requirements of data sources. These two modules were adopted to perform the processing task. Preprocessing includes de-duplication and empty record removal. De-duplication refers to the removal of duplicate records so that the abstract of each record could be used as a unique identifier. If two records shared exactly the same abstract, then one of them would be removed from the dataset. De-duplication ensures the uniqueness of each record. Empty record removal refers to the exclusion of paper records with null fields. If a record did not contain full information on title, abstract, or keywords, then the record would be removed.

A total of 15,935 papers were initially retrieved from the ISI WoS Core Collection, of which 7,820 were removed and the remaining 8,115 records were included. Key attribute fields include title, abstract, keywords, country, document type, journal, publication year, research area, and times cited.

However, the field data included inconsistencies ranging from spelling differences – whether intentional or accidental, to synonyms (e.g., "happy" and "glad"). As accurate analysis relies on minimizing these inconsistencies, the keywords were first preprocessed through data cleaning, by using such tools as number filter, punctuation eraser, stop word filter, English stemmer, and self-defined regex filter. Machine-assisted and rule-based recognitions were then adopted to merge synonyms and reduce the size of keyword list.

2.3 Whole SCI (WS) and High-Citation SCI (HCS)

To emphasize high-citation papers among all the articles retrieved, we grouped the papers into two datasets: WS and HCS (Figure 3). WS data refers to all the retrieved records after data processing, while HCS data contains selected papers, each with over 100 citations.

The user community of a research paper includes three stakeholders: authors who write the paper, editors who review and decide on the publication of the paper, as well as readers of the published article. A published paper reflects the authors' research interests and the editors' recommendations. As citation index reflects the impact factor of a published paper, high-citation papers are those with greatest popularity among readers. In this sense, WS represents authors' interests and editors' views, whereas HCS shows readers' preferences.

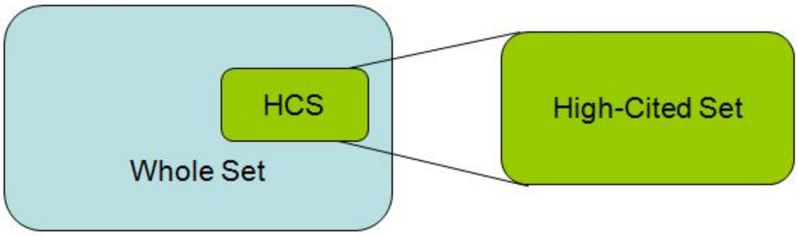


Figure 3 Relationship between HCS and WS

3. Data field analysis

The WS and HCS datasets were stored as WS.xls and HCS.xls, with 8115 and 150 records, respectively.

3.1 Statistics of attribute fields

Each record consists of 11 attribute fields, including article ID, title, abstract, publication country, times cited, keywords (authors), keywords plus, research area, document type, journal, and publication year. Table 1 shows the field statistics of WS.xls.

Table 1 Field statistics of WS paper records

Field	Number of items	Coverage (%)	Data type	Meta tags
ISI unique article identifier	8,115	100%	Number	Identity number
Title	8,114	100%		Paper title
Abstract	8,115	100%		

Authors 1st	6,085	100%		
Number of authors	29	100%	Number	
Countries	133	98%		Country
Times cited	188	100%	Number	
Keywords (authors)	16,692	84%		
Keywords plus	10,898	87%		
Keywords (authors and plus)	21,975	100%		
Research area	119	100%		
Document type	6	100%		Document type
Journal	1,280	100%		
Publication year	28	100%	Year	Date

3.2 Most productive journals and institutions

The flood papers of this database mainly came from the following 10 journals, each of which exceeded 100 publications: *Natural Hazards* (499 papers), *Journal of Hydrology* (385), *Natural Hazards and Earth System Sciences* (256), *Journal of Flood Risk Management* (212), *Hydrological Processes* (206), *Geomorphology* (154), *Hydrology and Earth System Sciences* (142), *Water Resources Research* (139), *Hydrological Sciences Journal* (114), and *Water* (109).

Each of the following four institutions published more than 100 papers: Chinese Academy of Sciences (163), Vrije Universiteit Amsterdam (131), University of Bristol (130), and Delft University of Technology (127).

3.3 Keyword analysis

“Author keywords” was used to denote keywords provided by authors in an article, while “keywords plus” referred to those generated by ISI on the basis of each article’s citations and references. 84% and 87% of the total 8,115 publication records retrieved contained “author keywords” and “keywords plus”, respectively. A total of 24,927 keywords were obtained after the two types of keywords were merged, which can be used to illustrate the trends of flood research.

21,975 keywords were obtained after data cleaning. Among the most frequently used keywords, “flood” and “climate change” were ranked first with 27 records for each, followed by “model” (18 records), “flood risk” (14 records), and “precipitation”, “rainfall”, “river” and “uncertainty” (13 records for each). Figure 4 shows a word cloud of the top 50 keywords generated based on their frequency.

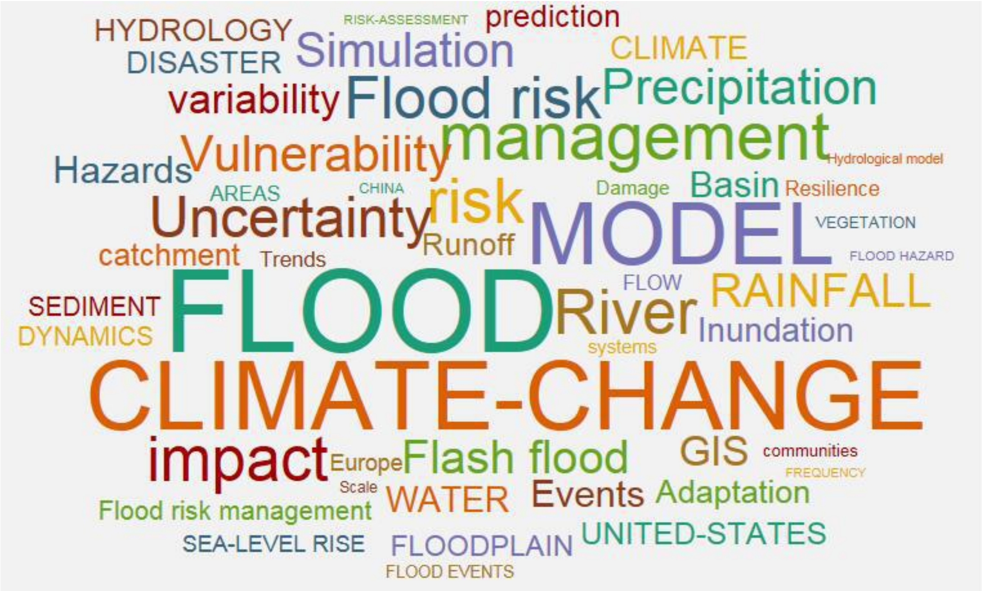


Figure 4 Word cloud of the top 50 keywords

4. Quality control and assessment

To guarantee the relevance of each record, we excluded those whose titles or keywords did not contain “flood.” In addition, duplicate records and records with empty fields were removed from the dataset. Stop words, punctuation, and number were deleted during processing. After data collection, we manually checked data validity and removed incomplete entries as well as entries irrelevant to flood disasters.

5. Value and significance

With an increasing number of publications added to the already large volume of scientific literature, it becomes significant more than ever to perform bibliometric analysis on specific research themes. In recent years, scientometric methods have been applied in global remote sensing,[12,13] night-time light remote sensing,[14] and the remote sensing of human health.[15] To the best of our knowledge, no literature-based datasets for flood research are available hitherto, and our dataset effectively fills this research gap. The literature-based knowledge mining model can be applied in disaster research, where the keywords can be used as core knowledge for topic analysis. The dataset presented here can be used to analyze major issues of flood research. A comparative analysis of WS and HCS datasets is helpful in illustrating the potential issues and future trends of flood research.

Acknowledgments

This work is supported by the Hainan Provincial Department of Science and Technology under Grant No.ZDKJ2016021. We thank Dr. Huang Mingrui from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences for her support on the collection of this dataset. Thank Qing Xiuling from the National Science Library, Chinese Academy of Sciences for her suggestion on data retrieval and processing.

References

1. Syvitski JPM, Overeem I, Brakenridge GR et al. Floods, floodplains, delta plains – A satellite imaging approach. *Sedimentary Geology* 267 – 268 (2012): 1 – 14.
2. Hristovski D, Peterlin B, Mitchell JA et al. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics* 74 (2005): 289 – 298.
3. Jensen LJ, Saric J & Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics* 7 (2006): 119 – 129.
4. He L & Li F. Topic discovery and trend analysis in scientific literature based on topic model. *Journal of Chinese Information Processing* 26 (2012): 109 – 115.
5. Zins C. Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology* 58 (2007): 479 – 493.
6. Vieira E & Gomes J. A comparison of Scopus and Web of Science for a typical university. *Scientometrics* 81 (2009): 587 – 600.
7. Bakkalbasi N, Bauer K, Glover J et al. Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries* 3 (2006): 7.
8. Kostoff R. The underpublishing of science and technology results. *Scientist* 14 (2000): 6.
9. Perianes-Rodriguez A, Waltman L & van Eck NJ. Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics* 10 (2016): 1178 – 1195, DOI: 10.1016/j.joi.2016.10.006
10. Feng H & Fang S. Research on the application of Thomson data analyzer to analyze the patent intelligence of scientific institutions. *Information Science* 26 (2008): 1833 – 1843.
11. Yang Y, Akers L, Klose T et al. Text mining and visualization tools – Impressions of emerging capabilities. *World Patent Information* 30 (2008): 280 – 293

12. Manning CD Raghavan P & Schütze H. *Introduction to Information Retrieval. Vol. 39*. Cambridge: Cambridge University Press, 2008.
13. Saffer JD & Burnett VL. Introduction to biomedical literature text mining: Context and objectives. *Biomedical Literature Mining*. NY: Humana Press, 2014: 1 – 7.
14. Zhang H, Huang M, Qing X et al. Bibliometric analysis of global remote sensing research during 2010 – 2015. *ISPRS International Journal of Geo-Information* 6 (2017): 332.
15. Zhuang Y, Liu X, Nguyen T et al. Global remote sensing research trends during 1991–2010: a bibliometric analysis. *Scientometrics* 96 (2013): 203 – 219.
16. Hu K, Qi K, Guan Q et al. A scientometric visualization analysis for night-time light remote sensing research from 1991 to 2016. *Remote Sensing* 9 (2017): 802 – 809.
17. Viana J, Santos JV, Neiva RM et al. Remote sensing in human health: A 10-year bibliometric analysis. *Remote Sensing* 9 (2017): 1225 – 1235.

Data citation

1. Zhang HY, Li GQ, Huang MR et al. A dataset of scientific literature on floods, 1990 – 2017. *Science Data Bank*. DOI: 10.11922/sciencedb.591

Authors and contributions

Zhang Hongyue, PhD; research area: natural language processing, disaster information mining. Contribution: literature data collection and processing, paper writing.

Li Guoqing, PhD, Professor; research area: geospatial data infrastructure, remote sensing, big data. Contribution: advice on dataset design and data check, paper writing.

Huang Mingrui, PhD Student; research area: literature retrieval, bibliometric analysis. Contribution: advice on data collection and processing, paper writing.

Qing Xiuling, PhD, Associate Professor; research area: literature retrieval, bibliometric analysis. Contribution: advice on literature retrieval and data processing, paper writing.

Zhang Huarong, MSc; research area: drone mapping, geographic information processing. Contribution: manuscript editing.

How to cite this article: Zhang HY, Li GQ, Huang MR et al. A dataset of scientific literature on floods, 1990 – 2017. *China Scientific Data* 3 (2018), DOI: 10.11922/csdata.2018.0020.en