Hydrol. Earth Syst. Sci., 26, 1165–1185, 2022
https://doi.org/10.5194/hess-26-1165-2022

Hydrology and Earth System Sciences
Open Access
EGU

Building a methodological framework and toolkit for news media dataset tracking of conflict and cooperation dynamics on transboundary rivers

Liying Guo, Jing Wei, Keer Zhang, Jiale Wang, and Fuqiang Tian

Department of Hydraulic Engineering, State Key Laboratory of Hydroscience and Engineering, Tsinghua University, Beijing 100084, China

Correspondence: Fuqiang Tian (tianfq@tsinghua.edu.cn)

Received: 19 July 2021 – Discussion started: 21 July 2021
Revised: 13 January 2022 – Accepted: 13 January 2022 – Published: 1 March 2022

Abstract. The management of transboundary rivers will be one of the great political and environmental challenges of the 21st century if knowledge of conflict and cooperation is not fully developed. Transboundary river conflict and cooperation are critical for the sustainable development of river basins, regional security, and stability and have significant scientific and practical implications. The construction of a dataset of transboundary water events – individual conflictive or cooperative interaction between riparian regions – provides an important data support and a factual basis for the study of transboundary rivers. However, the most representative research, the Transboundary Freshwater Dispute Database, is built by means of manual reading for information extraction, is, thus, difficult for fast updating, and also does not cover the global changes in the past decade. This research aims to build a methodological framework for news media datasets to track conflict and cooperation dynamics on transboundary rivers, provide a mass of relevant data for the research of transboundary rivers on the globe, prepare a potent research toolkit, lay a solid foundation for further data mining research, and better suit the big data age. In order to test the effectiveness of the methodological framework and toolkit for dataset construction, this research analyses the spatial coverage, both in terms of continental and national, temporal coverage from 1953 to 2019, and content coverage and conducts relevance screening of the articles in the four representative river basins in the datasets. The results show that the datasets built by this framework can capture comprehensive content of transboundary water conflict and cooperation in both spatial and temporal coverage with acceptable data quality.

1 Introduction

Globally, there are 310 transboundary river basins, covering 47.1 % of the land area, except Antarctica (McCracken and Wolf, 2019), and accounting for approximately 60 % of global freshwater discharge (Wolf et al., 1999). The population of the basins comprises 52 % of the world's total (McCracken and Wolf, 2019). Transboundary river basins not only support the lives of the people in the basins but also connect the various economic sectors and ecosystems in the basin into an organic whole; transboundary water management not only affects the development of riparian countries in all aspects but also intertwines social, economic, environmental, and political sectors of each riparian country and increases interdependence in between (United Nations and UNESCO, 2019). Riparian countries have divergent demands and priorities for transboundary water resources, different development agendas for water resources, and different water governance regimes and water resource cultures (Sadoff and Grey, 2005), which make the management of transboundary water resources more complex than that of domestic water resources. Transboundary river basins are, thus, prone to conflicts of various forms, forming a complex situation in which conflicts and cooperation develop intertwined. Therefore, research on water conflict and cooperation in transboundary rivers has important theoretical value and practical significance. The exploration of dynamics of conflict and cooperation as social sectors in a human–water coupled transboundary system is especially prominent.

Among the extant studies on transboundary rivers, transboundary water event datasets – individual conflictive or co-

operative interactions between riparian regions – provide factual data support for the formation of a global, generalized understanding, which is of great significance. The most representative research – the Transboundary Freshwater Dispute Database (TFDD) developed by Oregon State University (Wolf, 1999) – has compiled more than 6400 historical transboundary water events, both conflictive and cooperative (3813 left after we removed duplicated records from their original data), on the global scale from the year of 1948 to 2008 (Transboundary Freshwater Dispute Database, 2008). The data came from existing political science datasets and news media articles, which were manually screened, interpreted, and coded to extract the detailed information of the water event (Yoffe and Larson, 2001). Building upon these event data, the Basins at Risk (BAR) project (Yoffe and Larson, 2001) further classified water events by the level of intensity of conflict or cooperation, ranging from -7 to $+7$, to identify potential sociopolitical threats and provided a brief summary of the detailed information of the event. The results included very few examples of full cooperation and extreme conflicts but identified river basins that are at potential risk for further conflict. TFDD has built up foundation of this methodological framework for tracking transboundary river water events and allows for further identification of the conflict/cooperation dynamics and possible analysis of its complex driving mechanism.

Manual reading and coding processes were adopted in TFDD, which largely limit the implication of this method in the era of big data. The explosion of digital news data, whose discussion of transboundary water events has grown exponentially, made it more difficult to manually track all published water events and the dynamics of conflict and cooperation. While manual reading excels in extracting latent and detailed content, it is much more time and labor consuming. Therefore, it is necessary to revise the methodological framework to meet the current need for a more comprehensive and detailed dataset which can be updated in a more efficient manner. Meanwhile, it can also provide the basis for further analysis, i.e., to reflect the concerns of different stakeholders, and obtain a global law of transboundary water conflict and cooperation (Bernauer and Böhmelt, 2020).

This paper aims to provide such a revised methodological framework for news media tracking of conflict and cooperation dynamics on transboundary rivers and provide a toolkit when applying the framework in the corresponding research. The theory that inspired our framework is from Lasswell's model of communication (Lasswell, 1948), which focused on communication as a process to conduct a problem-oriented inquiry of the news report through content analysis with the seven fundamental elements of who, with what intentions, in what situations, with what assets, using what strategies, reaches what audiences, with what result? Our design of the search keyword generator follows to the line of theoretical principles by Lasswell closely and intends to track conflict and cooperation dynamics on transboundary rivers by answering Lasswell's questions involved with the seven elements. This study can help to reveal the evolutionary dynamics and patterns of transboundary water conflicts and cooperation on a global scale, by collecting news media datasets with an automated approach and minimizing the manual workload of screening, reading, and understanding the relevant news media articles, and provides researchers with powerful tools to retrieve useful information in related fields. It can serve as the foundation for further analysis via text mining and as a methodological foundation of quantifying the social dimension of transboundary river systems.

2 Data and method

This study attempts to build a revised methodological framework that reflects the dynamics of water conflicts and cooperation among all the transboundary rivers on the globe. Overall procedures in the revised framework are illustrated in Fig. 1. The method can be divided into the following three steps: step 1 – select database; step 2 – keyword determinants; step 3 – data cleaning and processing. More specifically, the method begins with selecting news database in step 1, and detailed criteria to select news databases are stated in Sect. 2.1.1. Search keywords are generated in step 2 with five blocks of keywords determinants. These five blocks are concerned with river basin characteristics and the research question and determine the validity and relevance of the data to be collected. Using generated keywords in step 2, an original dataset is downloaded for data cleaning and processing in step 3, which include rough manual reading and sorting to check the result relevance in order to feedback on further keyword modifications in step 2. Trial and error between steps 2 and 3 promise satisfactory keywords setting for the research. In Sect. 2.4, several potential areas for analysis in the future are introduced, which are extended applications for this methodological framework.

2.1 Step 1: select database

2.1.1 News media as data source

Choice of media sources should accord closely with the research goal. Our research goal is to track conflict and cooperation dynamics on transboundary rivers, which requires the data to cover water events and public opinion over a relatively long period of time. Also, newspapers (both print news and web news) published by professional journalists and editors are more suitable to use as data sources to reflect opinions of communities than social media (e.g., Twitter), which is better suited to reflections of individual opinions. News media reflect what is important for the individual country/sector that they are published within (Cooper, 2005); it has, thus, increasingly been studied by researchers to gain insight into transboundary water issues. The local news media is the first-hand material that reflects the attitude/perception

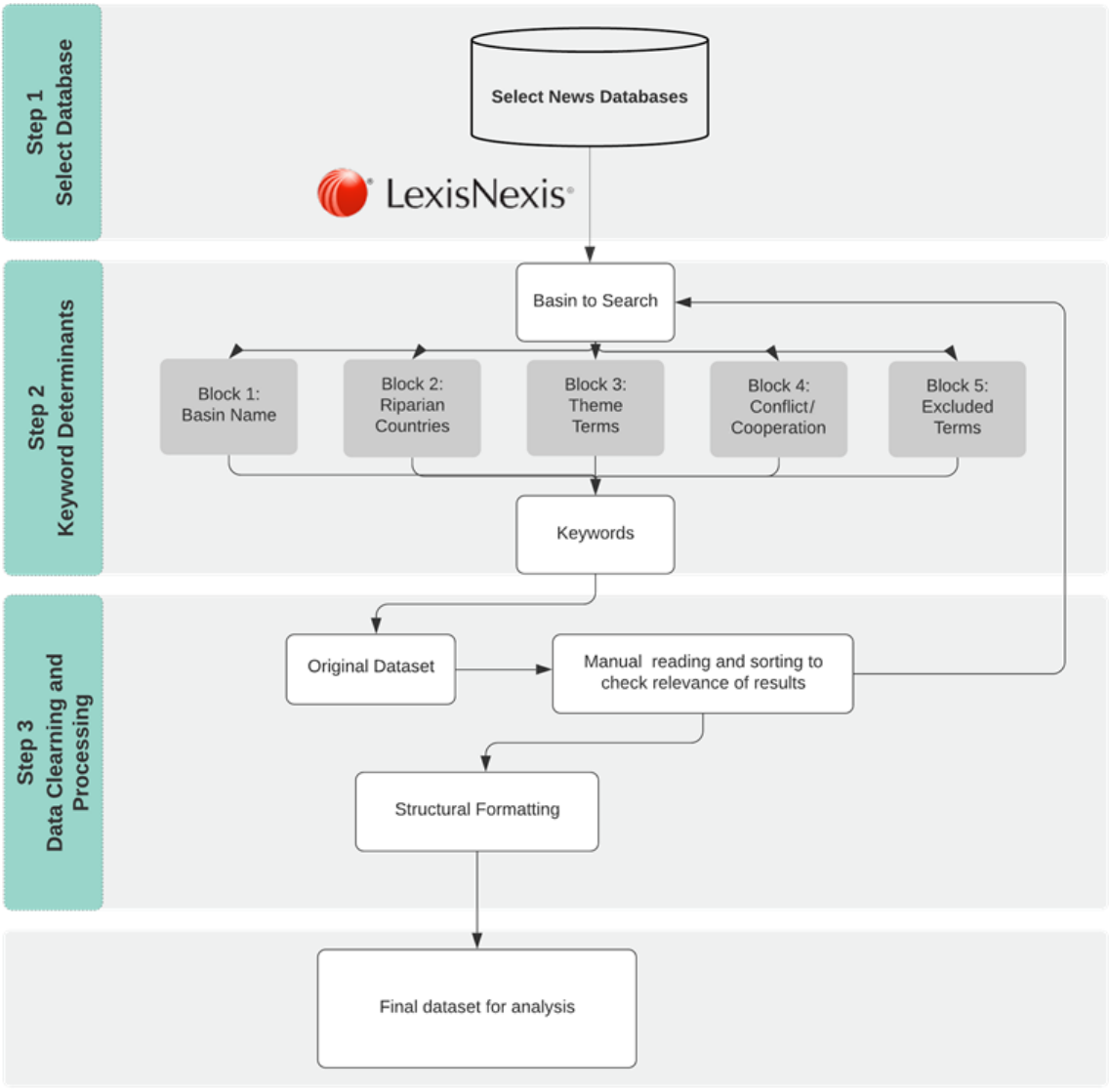

Figure 1. Method flow chart.

riparian countries held for their shared water and the involved stakeholders when discussing the water events in the transboundary river basin. In parallel, international news media serve as a good source of information to understand viewpoints from the international audience that are outside of the river basins. Together, text analysis of both regional and international news for water events in transboundary rivers can reveal the full picture of the ongoing dynamics in the river basin.

2.1.2 Select news database

The very first step of this method involves selecting a news database that covers comprehensive news sources spanning the globe. The selected media databases should include longitudinal coverage (i.e., can be traced back to decades) and updated in a timely manner, such as Lexis Advance (a product of the LexisNexis corporation), ProQuest LLC., Factiva, etc. Lexis Advance covers more than 6000 mainstream news media in most countries and regions around the world, with a long-term coverage, and is one of the most commonly used news sources in the field of social sciences (Weaver and Bimber, 2008; Racine et al., 2010). Therefore, Lexis Advance is taken as an example of a news media database to demonstrate the process of obtaining news media data of transboundary water conflicts and cooperation, and other suitable databases can, of course, be feasible options. Although the temporal coverage is affected by the level of media development in different regions, the covered time frame spans over 100 years to date, providing good data support on tracking media coverage of transboundary water conflict and cooperation research. The scope of the research is limited to En-

glish language newspapers only, due to our limitation of language processing, which is considered as sufficient enough to meet the requirements for extensive coverage of transboundary water conflicts and cooperative research.

2.2 Step 2: keyword determinants

2.2.1 Select rivers to search

The scope of rivers to search in this study are 286 transboundary rivers (Transboundary Waters Assessment Programme, 2016), as identified in Table 1. It is understood that the total number of transboundary rivers has recently been updated to 310 (McCracken and Wolf, 2019), which is due to the advancement of remote sensing technology. Remote sensing can examine the two fundamental characteristics of transboundary rivers (common terminus and perennial); thus, finer resolution of hydrologic data assists in discovering new transboundary rivers. In general, the majority of the 24 newly added basins are small in area (less than 10 000 km^2; McCracken and Wolf, 2019) and are considered as being inactive in conflicts and cooperation dynamics. Therefore, this study refers to 286 transboundary rivers in the procedure of selecting rivers to search, which can be extended to 310 in the future. In total, four river basins were taken as case studies, namely the Mekong, Nile, Columbia, and Ganges–Brahmaputra–Meghna (hereafter GBM) rivers, and used as the global hotspots of water events.

2.2.2 Search keyword generator

The search terms are one of the key determinants of the coverage and relevance of the data to be retrieved. This study develops a keyword generator that allows efficient generating of keywords terms which are applicable to all transboundary river basins (286 rivers basins) in the world. The keyword determinants are developed on the basis of TFDD (Yoffe and Larson, 2001) and further revised to include five blocks of terms (as shown in Fig. 2). These five blocks aim to include in which river basin (Block 1), who are involved (riparian countries – Block 2), regarding what issues (Block 3), and resulting in a conflict/cooperation status (Block 4). More specifically, Blocks 1 and 2 are basic information about the river basin, such as the name of the river basin, and various name formats of the riparian countries, and retrieved articles need to discuss the conflictive or cooperative aspects of the events involving at least one of riparian countries. Block 3 contains theme terms regarding various functions of the water body, topics discussing hydraulic infrastructure, water quality, agriculture/fishing, or any other specific topics with associated terms. Block 4 include keywords indicating conflict or cooperation, and Block 5 consists of keywords to be excluded as they bring in irrelevance. The above five blocks can narrow down the search to the desired scope, with the list of unwanted words to further screen out irrelevant topics, after which the search results can achieve a balance between coverage and relevance, that is, neither too much relevant information is missed nor too much irrelevant information is included.

(1) Block 1: basin name

This study customizes relatively general algorithms to generate search strings for river basins with different attributes and conducts special treatments for individual river basins so that each river basin is under the general search rules, resulting in a considerable number of search results with a balance of coverage and accuracy. The aim of Block 1 is to obtain a searchable list of the basin name, including various formats, and consider special treatments for specific categories of basin names. There are several categories identified for different variations of basin names (see below for specific information).

a. *The basin name is same as the name of a certain riparian country or state.* The search results are likely to contain many articles about the internal affairs and diplomacy of the country or state. The detailed list of this type of basin is shown in Table 1. When talking about transboundary water issues, people usually focus on interactions on the scale of local communities and riparian states rather than intercontinental issues and do not refer to the continent names. Therefore, raising the frequency of the continent name in the search keywords will only compress the data volume of the relevant articles significantly and not improve the data relevance pertaining to the research goal. However, river basins with the same names but located in different continents have different riparian countries. Adding the frequency setting of riparian countries will filter out articles about the river on the other continent effectively. For example, St. John rivers appear both in Africa (flowing through Côte d'Ivoire, Guinea, and Liberia) and North America (flowing through the United States and Canada). The rising frequency of riparian countries rather than continent names contributes more to the data relevance.

b. *The basin name contains commonly used words.* For example, Amazon not only refers to the Amazon River basin but also to an e-commerce company in the United States. More filters will be adopted in this case to ensure a good relevance rate. See Table 1 for a detailed list of this type of river basin.

c. *The basin name contains words such as "lake" or "sea".* The word frequency setting for "river" in the search string needs to be modified and either that for "lake" needs to be increased or "sea" needs to be removed from the list of noisy keyword. See the detailed list of this type of river basin in Table 1.

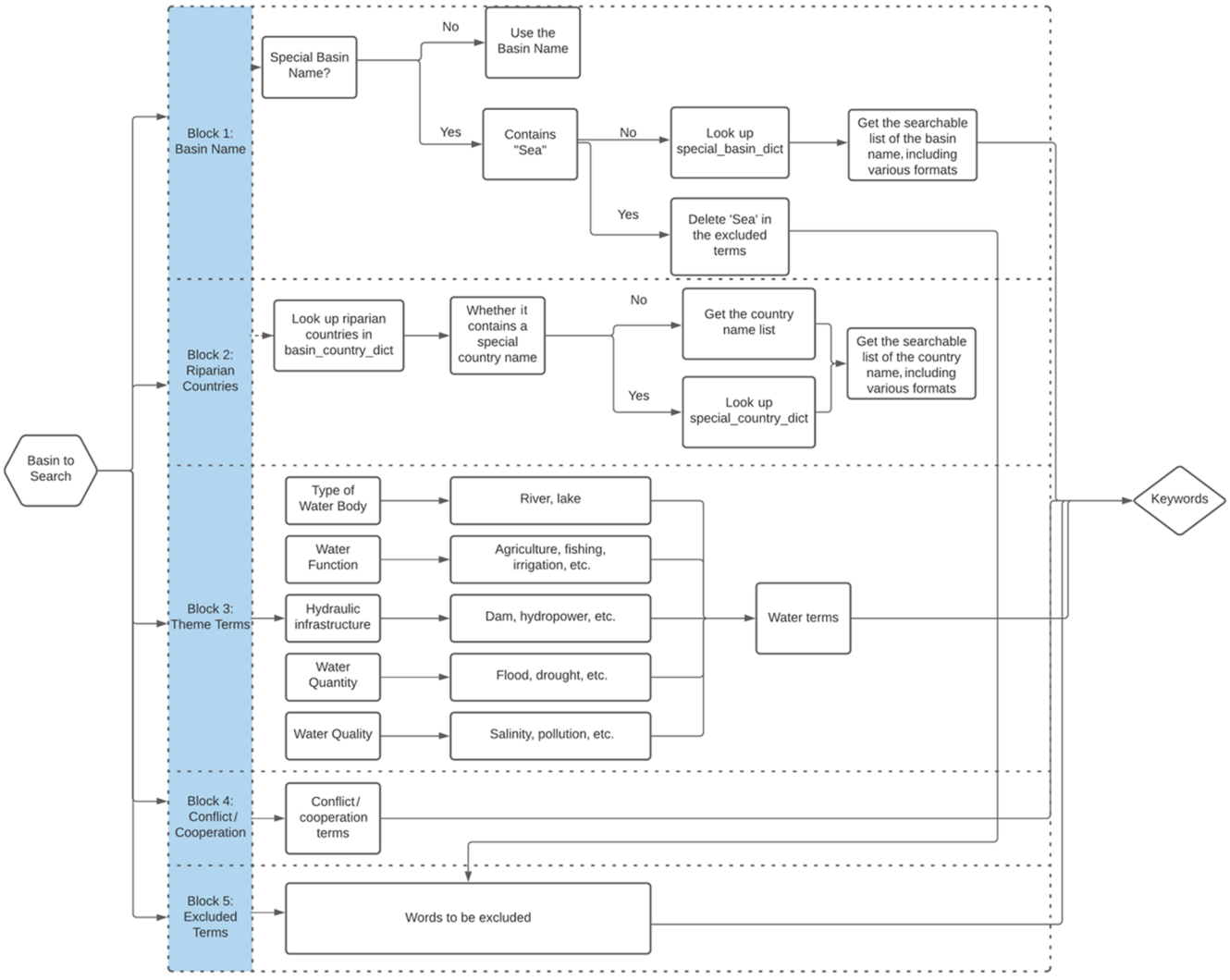

Figure 2. Search keyword generator flow chart.

d. *Other categories of basin names that require special treatment*. River basins have different names, e.g., upstream and downstream rivers are designated with different names, or the river basin contains multiple rivers. Sometimes rivers in the basin have different names, the basin name is composed of multiple words, or similar basin names exist on different continents. In other cases, the basin name contains "St.", but may be referred to with the word "saint" in media articles (see Table 1 for details).

The special_basin_dict in the toolkit in Block 1 is a Python dictionary uploaded on Zenodo, for which the keys are basin names with multiples words or with special characters (e.g., backslash, dash, or parenthesis), and the values are all searchable formats of the related basin names and river names. Given the original basin name to search, special_basin_dict can feedback its corresponding searchable keywords. If searching without special_basin_dict and using the original basin name to search, few or no results can be found. The coverage of the retrieved results is enhanced by the special_basin_dict. When using the dictionary, import it to your script first, and then call it up easily.

(2) Block 2

Block 2 is information concerned with the riparian countries within the transboundary river basin. The aim of Block 2 is to obtain the searchable list of the riparian country names, including various formats. To fulfill the task, two helpful dictionaries – basin_country_dict and basin_country_dict – are developed and provided in the toolkit of this study.

The basin_country_dict in the toolkit in Block 2 is a Python dictionary uploaded on Zenodo, for which the keys are basin names, and the values are all riparian countries located in the transboundary basin. Given the basin name to

Table 1. Categories of basins needing special treatment.

Categories of basins needing special treatment	Basin names	Treatment
Basin name includes a state or district name	Belize; Columbia; Congo/Zaire; Corredores/Colorado; Gambia; Jordan; La Plata; Mississippi; Nelson–Saskatchewan; Niger; Senegal; Tigris–Euphrates/Shatt al-Arab	Raise the frequency setting for "water", "river", etc. to filter out the geopolitical articles as there are many
Basin name includes a common word	Amazon; Baker; Columbia; Cross; Don; Fly; Han; Lagoon Mirim; Lotagipi Swamp; Massacre; Negro; Oral/Ural; Orange; Rhone; Red/Song Hong; San Martin; Seno Union/Serrano; Vanimo–Green; Whiting	Raise the frequency setting for "water", "river", etc. to filter out water-unrelated articles, as there are many, or delete a certain percentage of the articles from the end of the results list
Basin name includes "lake" or "sea"	Lake Chad; Lake Fagnano; Lake Natron; Lake Prespa; Lake Titicaca–Poopó System; Lake Turkana; Lake Ubsa-Nur; Aral Sea	The word frequency setting for "river" in the search string needs to be modified, that for "lake" needs to be increased, and/or "sea" needs to be removed from the list of noise keywords
Basin name includes multiple formats (might consist of multiple rivers)	Asi/Orontes; BahuKalat/Rudkhanehye; Bei Jiang/Hsi; Benito/Ntem; Ca/Song-Koi; Cancoso/Lauca; Carmen Silva/Chico; Coco/Segovia; Congo/Zaire; Corantijn/Courantyne; Corredores/Colorado; Cuvelai/Etosha; Douro/Duero; Gallegos/Chico; Ganges–Brahmaputra–Meghna; Hamun-i-Mashkel/Rakshan; Hari/Harirud; Ili/Kunes He; Jenisej/Yenisey; Juba–Shibeli; Kura–Araks; Lava/Pregel; Mana–Morro; Nelson–Saskatchewan; Oder/Odra; Oiapoque/Oyupock; Oral/Ural; Red/Song Hong; Seno Union/Serrano; Shu/Chu; Tagus/Tejo; Tigris–Euphrates/Shatt al-Arab; Tjeroaka–Wanggoe; Torne/Tornealven; Vanimo–Green; Vistula/Wista	Contain all formats of related basin/river names in the search keywords
Basin name consists of a river with multiple names	Muhuri (also known as Little Feni)	Contain all formats of related river names in the search keywords
Basin name includes multiple words	An Nahr Al Kabir; Astara Chay; Coatan Achute; El Naranjo; Great Scarcies; Har Us Nur; Kowl E Namaksar; La Plata; Lagoon Mirim; Lotagipi Swamp; Lough Melvin; Nahr el Kebir; Oued Bon Naima; Pu Lun T'o; Rio Grande (N. America); Rio Grande (S. America); San Martin; Song Vam Co Dong; St. Croix; St. John (Africa); St. John (N. America); St. Lawrence; St. Paul; Wadi Al Izziyah	Add quotation marks to the basin name in the search keywords to search it as a whole term and prevent the basin name from being tokenized
Same basin names exist in multiple continents	Great/Little Scarcies; Rio Grande (N. America/ S. America); St. John (Africa/N. America)	Usually, articles do not contain the continent name when talking about rivers. Therefore, adding continent names into search keywords compresses data volume significantly and does not help with relevance. Adding a frequency setting of riparian countries will filter out articles about the river on the other continent effectively.
Basin name includes "St." (saint)	St. Croix; St. John (Africa); St. John (N. America); St. Lawrence; St. Paul	Put "saint" and "St." into search keywords together

search, basin_country_dict can feedback the list of riparian countries. Another Python dictionary used in Block 2 is special_country_ dict, for which the keys are country names with various formats or with special characters (e.g., dot), and the values are all the searchable formats of the country name. Given the special country name to search, special_country_dict can feedback the list of all searchable formats of the country name.

When given a basin name to search, first looking up riparian countries in the basin_country_dict obtains the list of riparian countries. Then one can check whether there is a special country name in the list of riparian countries. If yes, through looking up special_country_dict, a searchable list of the country name, including various formats, is generated in Block 2.

(3) Block 3

Block 3 contains terms concerning various themes of transboundary water resources, as shown in Table 2. These include, for example, the type of water body, function of water body (agriculture, fishing, etc.), hydraulic infrastructure, water quantity, water quality, and other specific topics which arouse certain research interests.

(4) Block 4

Block 4 contains conflict-/cooperation-related keywords, adopted from the TFDD search keywords (Yoffe and Larson, 2001), as shown in Table 2. If you focus on a certain type of conflict/cooperation, the keywords in Block 4 can be modified accordingly. In addition, the UNBIS Thesaurus (UNBIS Thesaurus, 2021) provides lists of related keywords for conflict and cooperation which can be referred to.

(5) Block 5

Block 5 contains excluded terms, most of which are adopted from TFDD searching keywords (Yoffe and Larson, 2001) shown in Table 2. These terms, seemingly relevant to our topic, occur in media articles often and easily bring in lots of data noise. For example, the terms "sea" and "ocean" bring masses of irrelevant articles referring to marine rights and navigational utilization; "nuclear" refers to "nuclear power" and "nuclear threat", which is not the main concern of transboundary water conflict and cooperation; and as for "flood of refugees", despite containing the keyword "flood", it is regarded as irrelevant to our topic. These terms that are prone to bringing in noise should be excluded in the search results, and thus, a list of excluded terms is included in Block 5. If researchers employ our framework in their own study fields in the future, then the excluded terms to avoid noise in Block 5 should be modified accordingly to fit their own research field based on results of trial and error between steps 2 and 3 and combined with their experience and knowledge background. For example, when collecting data for the Aral Sea, the term "sea" should be deleted from the excluded terms in Block 5 to prevent a great loss of data coverage.

2.2.3 Term frequency setting of keywords

The setting of the term frequency of keywords comes from the recursive trial and error in the search process, which makes the search results for most transboundary river basins relatively satisfactory. For individual river basins, a universal setting of the rules of the term frequency will cause the search results to sharply drop to zero or return too many to cope with, and the accuracy of the search results cannot be guaranteed. For example, when collecting data on the Jordan River basin, given that Jordan is not only the name of the river basin but also the name of a riparian country in the basin, there are too many articles that meet all the search requirements but will purely be about regional politics. Therefore, the setting of term frequency for the keywords "water" and "river" needs to be increased to five times to highlight the theme of transboundary water resources and ensure that the search results have similar accuracy to other river basins.

Taking the Lancang–Mekong basin as an example, the search keywords used in this study are shown in Table 3. During the trial-and-error process, we found that the results' relevance rate is far below an acceptable level (less than 30 %); therefore, we revised the keyword terms to increase the frequency of certain terms until satisfactory results are produced. For example, the number of times the name of the basin appears in the article were increased to at least five times, and the name of any riparian country in the basin (either an official name or an abbreviation) should appear in the article at least two times. Water-related words are divided into three sub-blocks, namely the type of water body, function of water body, and infrastructures for water conservancy. Among them, the terms "water" and "river" appear at least three times, respectively, and the rest of the keywords of the water block appear at least once; words related to conflict or cooperation also appear at least once. Recordings of trial-and-error process for the Mekong, Nile, and Jordan river basins are provided in the Appendix to demonstrate the effects of various groups of frequency settings of keywords and to show how a balance between relevance and coverage is being approached. Although the term frequency settings of keywords and the justification of the balance between relevance and coverage in this study may not be optimal, with a certain degree of coexisting subjectivity and objectivity, they can also serve as a reference for other researchers.

2.3 Step 3: data cleaning and processing

Before finalizing the refined datasets for further analysis, data cleaning and processing is indispensable. The first stage in step 3 is rough manual reading and sorting to check result relevance, which aims to provide feedbacks on how to modify keywords in step 2. Rough manual reading can be done by

Table 2. Example of keywords in Blocks 1–5.

Block 1: basin name	Basin name (5)	
Block 2: riparian countries	Each riparian country (2)	
Block 3: theme terms of transboundary water resources	Type of water body	Water (3), river (3), lake, stream, tributary, etc.
	Function of water body	Irrigation, fish, fish rights, water rights, water diplomacy, water hegemony, etc.
	Hydraulic infrastructures	Dam, diversion, channel, canal, hydroelect*, hydropower, reservoir, etc.
	Water quantity	Flood, drought*, water allocation, water sharing, etc.
	Water quality	Salinity, pollution, etc.
Block 4: conflict/cooperation terms	Conflict	dispute*, conflict*, disagree*, war, troops, "letter of protest", hostility, "shots fired", boycott, protest*
	Cooperation	Treaty, agree*, convention, "framework directive", negotiat*, resolution, commission, secretariat, "joint management", "basin management", peace, "accord", "peace accord", settle*, cooperat*, collaborat*, bilateral, multilateral, sanction*
Block 5: excluded terms	Sea, ocean, navigat*, nuclear, water cannon, light water reactor, mineral water, hold water, cold water, hot water, water canister, water tight, water down*, flood of refugees, oil, drugs, a stream of, flood of	

Note: the asterisk (*) indicates the root of a word, and the numbers in parentheses (i.e., 5, 2, or 3) indicate at least how many times the keywords should appear in a search result.

Table 3. Search keywords in the study (using Lancang–Mekong as an example).

Keyword search	Lexis Advance database
Must include the basin name (at least five times)	Mekong (5)
Includes at least one of the following country names (at least twice)	Thai* (2), Cambodia* (2), China (2), Chinese (2), Laos (2), Myanmar (2), Burm* (2), Vietna* (2)
Includes at least one of the following words related to water	Same as Block 3 (see Table 2)
Includes at least one of the following words related to conflict/cooperation	Same as Block 4 (see Table 2)
Does not include any of the following noisy words	Same as Block 5 (see Table 2)

Note: the asterisk (*) the indicates root of a word, and the numbers in parentheses (5, 2, or 3) indicate at least how many times the keywords should appear in a search result.

random sampling or, more conveniently, from back to front. Since lists of news results by news media databases usually have options to sort by relevance, the frontlines displayed in the front of the list of search results are ranked as more relevant to search terms than those of the backlines of the list. (Sorting by relevance is one of the sorting functions provided by Lexis Advance, which also provides options to sort by date and by document title. Among the three options, sorting by relevance works best for us to read roughly to change the frequency setting of keywords by trial and error. Therefore, sorting by relevance was chosen before downloading the data from Lexis Advance. Usually, news databases have similar functions for readers to read roughly and conveniently.) A

proper percentage, like 80 % of the results which are relevant among them all, can be set to meet our expectation.

To better facilitate future analysis, all downloaded text data will go through a structure formatting process. A data structuring program is developed for Lexis Advance to download and organize the text data into a structured format. The relevant media articles are processed in order of relevance, and detailed information, such as the publication time of the articles, media source, author, article length, etc., are stored in a structured manner. An example of structured media data is shown in Table 4. As for data integration, any news data downloaded from suitable data sources (not only from Lexis Advance) can be arranged and structured in the format of Table 4 through a data cleaning and processing procedure. After data processing, the toolkit provided by this research can be applied to the integrated data, regardless of the original data sources.

2.4 Potential analysis

The news media dataset of water conflict and cooperation on transboundary rivers allows for a variety of analysis at a later stage. This study lists several examples of potential analysis, including event extraction, stakeholder analysis, sentiment analysis, and topic analysis.

Event extraction from news articles is a conventional application of the water conflict and cooperation dataset. Similar to what has been achieved by TFDD, water events that are both conflictive and cooperative were extracted from relevant political science datasets and news articles (Yoffe and Larson, 2001). Event extraction requires concise and accurate information recognition and extraction from latent content in text data. Since human coders perform better than machine programming (Howland et al., 2006), human coding event extraction is recommended.

Stakeholder analysis for transboundary rivers is a way to identify who has been involved in transboundary water issues and the roles they play in the game, i.e., understanding the demands and expectations of the major stakeholders inside and outside the basin, based on a typical definition of stakeholder analysis (Smith, 2000). News media represent or reflect the interests of the home country; thus, via analysis of news media sources in a transboundary basin, political positions and economic interrelationships between riparian countries and other extraterritorial countries lying outside the basin are uncovered. Longitudinal analysis has the capability to depict the trajectories of a stakeholder country's interests and reveal the evolution of stakeholder countries in transboundary water issues.

Sentiment analysis on the news media dataset of transboundary rivers can bring the implicit information to the surface (Jiang et al., 2016), since the willingness for cooperation and the hostility of conflicts often hide behind the news articles. Positive and negative sentiments are close to the dynamics of conflict and cooperation in transboundary water issues, which serve as precursors of significant situational changes. Sentiment lexicons (Khoo and Johnkhan, 2018) or machine learning (Neethu and Rajasree, 2013) are major methods for sentiment analysis in text mining.

Topic analysis tells the story about main the interests and concerns of the news media and even the stakeholders over time (Jacobi et al., 2016). Topics concerned along with society development and evolutionary trajectories of transboundary water issues can be uncovered through popular algorithms of topic modeling analysis, such as latent Dirichlet allocation (LDA; Alsumait et al., 2009).

3 Results

This section overviews the global datasets statistically, both in terms of spatial coverage and content coverage, which aim to show the datasets telling stories of conflict and cooperation on transboundary rivers from all aspects on a global scale. To demonstrate the effectiveness of the methodological framework and toolkit, manual reading to check the improvements of data relevance was conducted on four representative basins including the Nile, Mekong, GBM, and Columbia.

3.1 Overview of the global datasets

3.1.1 Spatial coverage

(1) Continental coverage

With the customized search strings for each transboundary river, and the data structured program developed for Lexis Advance to organize the data, as of 10 March 2019, the data volume results of 286 transboundary river basins around the world are shown Fig. 3. In Fig. 3, the base map of transboundary river basins around the world was downloaded from TFDD in the format of geographic information system (GIS) shapefiles (Transboundary Freshwater Dispute Database, 2008).

The data volume of news articles reflects the prominence of the conflict and cooperation events discussed in transboundary river basins. Enough data volume promises statistical significance. The mainstream application of this news media dataset is further text mining to track conflict and cooperation dynamics on transboundary rivers. For text mining purpose, this study assumes arbitrarily that 100 media articles are the minimum data volume to track the dynamics of transboundary rivers over time.

Overall, there are 60 river basins with more than 100 media articles, which are considered as the key research basins of transboundary water conflict and cooperation in our research. The number of news articles discussing these 60 key research basins reached more than 41 000. Among the 60 key research basins, 16 river basins have more than 850 data records, as shown in Table 5, which attract more attention and are considered as being heated basins. Note that the defi-

Table 4. Example of structured data.

Paper index	1
Title	The 1997 water rights settlement between the state of Montana and the Chippewa Cree tribe of the Rocky Boy's Reservation: the role of community and of the trustee.
Source	Art and Science Academic Publications (ASAP) II database
Date	22 December 1998
Pp; ISSN; vol.; no.	Pp. 255(1); ISSN 0733-401X; vol. 16; no. 2
Word count	18 256 words
Author	Barbara A. Cosens
Body	Introduction established on 7 September 1916 "for Rocky Boy's Band of Chippewas and … other homeless Indians". (1) The Rocky Boy's Reservation is home to over 3000 tribal members. The Reservation's annual population growth rate is in excess of 3% (the original data are too long for a demonstration; here is an excerpt).

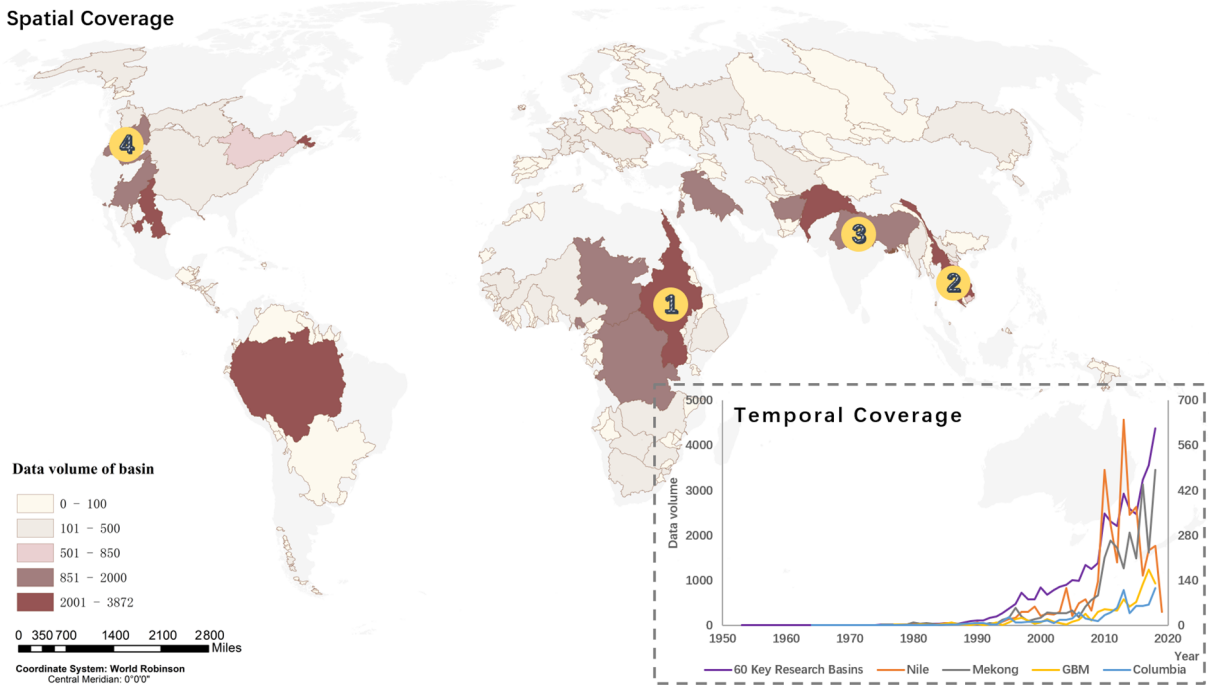

Figure 3. Spatial coverage and temporal coverage in the basin scale (1 – Nile; 2 – Mekong; 3 – GBM; 4 – Columbia).

nition criteria of key research basins (more than 100 articles) and heated basins (more than 850 articles) are flexible and adaptive according to specific research demands.

Most studies of conflict and cooperation on transboundary rivers focus on individual basins, which seek solutions to dealing with local challenges on transboundary water resources (Bernauer and Böhmelt, 2020). Therefore, the formation of a general understanding of conflict and cooperation on transboundary rivers needs global data support besides expert on-site experience from the research of individual basins. Many most-discussed transboundary river basins, such as the Nile, Mekong, Indus, GBM, and Tigris–Euphrates/Shatt al-Arab, are located in regions featured with frequent tensions and armed conflicts (Pohl et al., 2014) and are well-known by people. However, this study finds that there are also some river basins, from the authors' point of view, for which less

Table 5. 16 Most-discussed basins with more than 850 records.

Order	Basin name	Continent	Number of records	Countries
1	Nile	Africa	3872	Burundi, Central African Republic, Egypt, Hala'ib Triangle, Eritrea, Ethiopia, Kenya, Rwanda, Sudan, Abyei Area, South Sudan, United Republic of Tanzania, Uganda, Democratic Republic of the Congo
2	Mekong	Asia	3253	China, Cambodia, Lao People's Democratic Republic, Myanmar, Thailand, Vietnam
3	Rio Grande (N. America)	N. America	2718	Mexico, United States of America
4	Indus	Asia	2404	Afghanistan, China, India, Nepal, Pakistan
5	St. John (N. America)	N. America	2356	Canada, United States of America
6	Amazon	S. America	2078	Bolivia, Brazil, Colombia, Ecuador, French Guiana, Guyana, Peru, Suriname, Venezuela
7	Colorado	N. America	1975	Mexico, United States of America
8	Jordan	Asia	1816	Egypt, Israel, Jordan, Lebanon, West Bank, Syrian Arab Republic
9	Congo/Zaire	Africa	1391	Angola, Burundi, Central African Republic, Cameroon, Congo, Gabon, Malawi, Rwanda, Sudan, South Sudan, United Republic of Tanzania, Uganda, Democratic Republic of the Congo, Zambia
10	Lake Chad	Africa	1353	Central African Republic, Cameroon, Algeria, Libya, Niger, Nigeria, Sudan, Chad
11	Ganges– Brahmaputra–Meghna	Asia	1183	Bangladesh, Bhutan, China, India, Myanmar, Nepal
12	Helmand	Asia	1168	Afghanistan, Islamic Republic of Iran, Pakistan
13	Cross	Africa	1110	Cameroon, Nigeria
14	Tigris–Euphrates/Shatt al-Arab	Asia	939	Islamic Republic of Iran, Iraq, Jordan, Saudi Arabia, Syrian Arab Republic, Turkey
15	Columbia	N. America	859	Canada, United States of America
16	Tijuana	N. America	853	Mexico, United States of America

attention has been paid in the past in terms of transboundary water conflict and cooperation research, e.g., the St. John River (North America) and Tijuana River.

The data volume of transboundary water conflicts and cooperation news articles in the datasets of 60 key research basins on different continents is as follows: 14 454 for Asia, 11 306 for North America, 10 734 for Africa, 2674 for Europe, and 2498 for South America. It could possibly be attributed to the discrepant levels of economic development of major countries on each continent or varied attention being paid to the discussion of the management of transboundary rivers. The other important reason could be the linguistic variations. Since this paper chose English language newspapers as the search scope, the large amount of data in North America and the small amount of data in Europe could be due to system bias caused by language preferences.

There are notably large amounts of transboundary water conflicts and cooperation events reported in Asia and Africa, which indicates that transboundary water management is a major topic of peace and development in both Asia and Africa. Taking into consideration that most countries on these two continents do not speak English as their first language, the existence of a large number of news media articles on transboundary water conflicts and cooperation between Asia and Africa, on the one hand, reflects the fervent concerns about the transboundary water resources and

the desires for peace and development. On the other hand, it also reflects that people around the world are more involved in transboundary water issues in Asia and Africa and have invested heavily in the development and construction and pay close attention to these two rapidly developing and eye-catching continents.

(2) National coverage

News media data volumes in the datasets of 60 key research basins from different countries in the world are shown in Fig. 4. In Fig. 4, the base map of countries around the world was downloaded from ArcGIS Hub in the format of GIS shapefiles (Esri Data and Maps, 2021). It is seen that the United States of America contributes 11 515 news articles on transboundary water conflict and cooperation, ranking as number one, both as a riparian stakeholder in the transboundary water issues with Canada and Mexico and as an extraterrestrial international stakeholder involving in the transboundary water issues on continents other than the North America. Since a country's development and utilization of transboundary freshwater resources inevitably involves relations with other riparian countries, and transboundary water cooperation and conflicts often involve broader economic and social ties between riparian countries, transboundary freshwater management is an important component of the diplomacy of riparian countries. On the other hand, due to factors such as global hegemony, transnational investment, colonial history, and other factors, transboundary freshwater management often involves countries outside the region, thereby becoming a stage for great powers to play (Mirumachi, 2015).

3.1.2 Temporal coverage

The temporal coverage of the datasets of 60 key research basins (stated in Sect. 3.1.1) and four case study basins are shown in Fig. 3, which shows how many news media articles have been released over the years on transboundary water conflict and cooperation. Note that, due to differences in the order of magnitudes, the data series of 60 key research basins uses the major vertical axis, which ranges from 0 to 4500, and the four case study basins share the minor vertical axis, which ranges from 0 to 700. The datasets cover the years 1953 to 2019. A boom of news articles on transboundary water conflict and cooperation emerges from 1990s onwards and potentially continues in the future. This emphasizes the necessity to revise the methodological framework and toolkit for news media dataset tracking of conflict and cooperation dynamics on transboundary rivers to cope with the era of big data. For the four case study basins, the changing trends of data volume display strong variance, which may be affected by certain water events and geopolitical relations in the river basins at the moment.

The temporal coverage of four representative countries, which are the United States of America (USA; using the vertical minor axis on the left), China, Australia, and Egypt (using the vertical major axis on the right), is shown in Fig. 4. The USA contributes the largest volume of data among countries in the world. China has promoted transboundary cooperation in Mekong River basin actively in the recent years. Australia does not have a transboundary river with other neighboring countries but releases lots of news articles on transboundary water issues, and Egypt is one of the major countries in Nile River basin, which is representative in transboundary water conflict and cooperation. Similar to the temporal coverage of basin analysis, country datasets also cover the years 1953 to 2019. Data volume took off from 1990s and potentially continues in the future as well. For the four representative countries, the overall trends of data volume go up over time and are affected by contextual events in the country to show strong variance.

3.1.3 Content coverage

Word frequency analysis demonstrates that this study has generated good datasets tracking of conflict and cooperation dynamics on transboundary rivers. In the datasets, words concerned with water body function, hydraulic infrastructure construction, basin management, national power, civic rights, jointed research, and water conflict and cooperation appear in a high frequency, consistent with the related keywords in TFDD (Yoffe and Larson, 2001) and relevant words provided in UNBIS Thesaurus (UNBIS Thesaurus, 2021). This indicates that the datasets are closely corresponding to the research question, providing data as needed.

3.2 Relevance screening

The major advancement of this methodological framework is that it allows the efficient and effective tracking of transboundary rivers conflict and cooperation events. The keyword generator developed in this study could result in an acceptable level of relevance without too much manual coding intervention. To demonstrate the effectiveness of this methodological framework, four river basins, the Mekong, Columbia, Nile, and GBM, were taken as case studies to conduct a manual coding process. There were two manual coders, who were trained beforehand, involved to work independently for the four basins in the coding process. Each one undertook half of the total workload, and articles in the datasets were randomly divided into two groups. Before starting, an inter-coder reliability test was conducted. The test randomly selects 50 articles from the datasets for two coders to read; differences were then discussed, and definitions were given to reach a common understanding. Krippendorff's alpha reliability was calculated as 0.81, which is considered as valid and consistent (Krippendorff, 2004).

The total number of downloaded articles, after removing duplicates with the function of removing duplicates in the data panel of Microsoft Excel, and the remaining number of

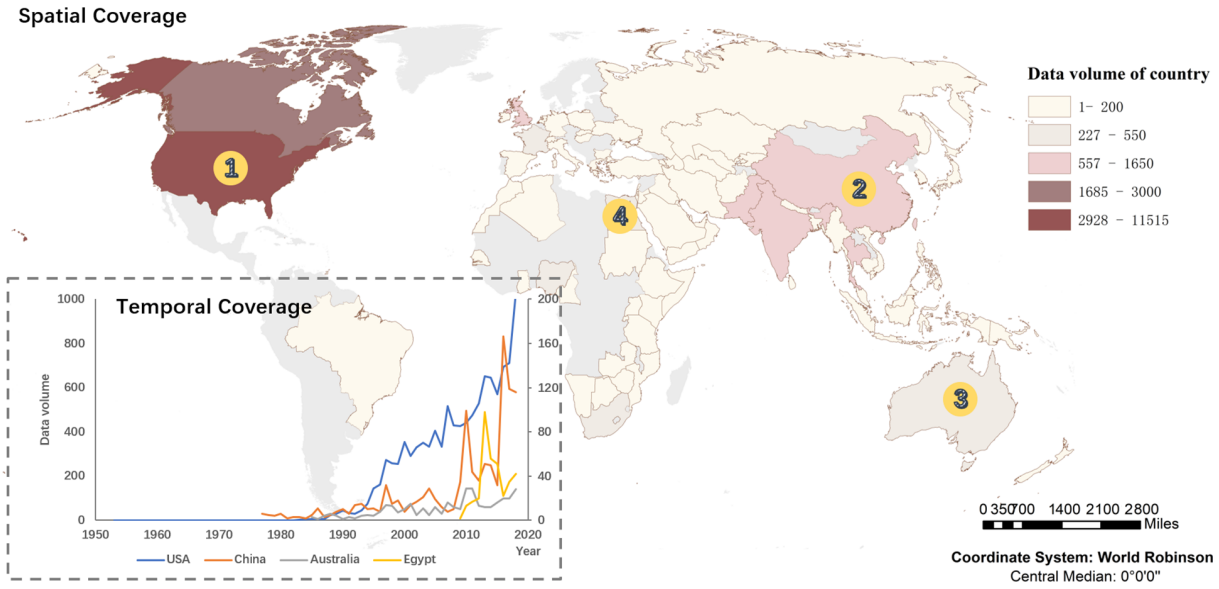

Figure 4. Spatial coverage and temporal coverage on a country scale (1 – USA; 2 – China; 3 – Australia; 4 – Egypt).

relevant articles with removal of the duplicates are shown in the Table 6. The calculation equation of the relevance percentage is shown as Eq. (1).

$$\text{Relevance percentage} = \frac{\text{Number after removing irrelevant}}{\text{Number after removing duplicates}} \times 100\%. \quad (1)$$

The last column of Table 6 shows the relevance percentage for the four river basins in a descending order. The relevance percentage of the Nile, Mekong, and GBM are at acceptable level and that of Columbia is less satisfying. This is due to Columbia belonging to special basin name category (details shown in Sect. 2.2.2 (1)), where the basin name is same as the name of a certain riparian country or state. To further investigate the relevance percentage of the four basins, the relevance percentage in 10 % stepwise increments is calculated for each basin, using Eq. (2). The relevance percentage in 10 % stepwise increments for the four basins is shown in Fig. 5. In Fig. 5, the horizontal axis is every 10 % stepwise segment of the news media articles data, and the vertical axis indicates the relevance percentage of that segment of data.

$$\text{Relevance percentage in 10 \% stepwise} = \text{Relevance percentage for every 10 \% of the total number after removing duplicates.} \quad (2)$$

The purpose of Fig. 5 is to demonstrate the necessity to apply special treatment for some river basins. Since the datasets retrieved from Lexis Advance are sorted by relevance, the frontlines are naturally more relevant than the backlines, and

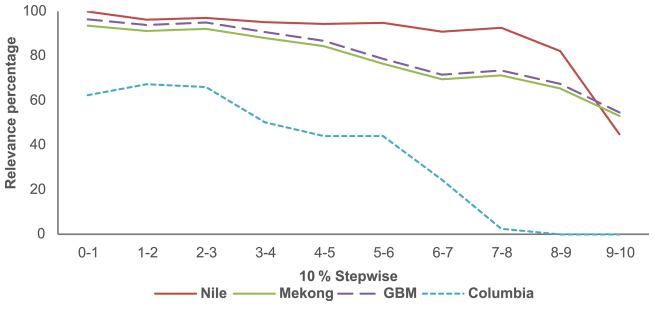

Figure 5. Relevance percentage in 10 % stepwise increments for the four basins.

the relevance percentage in 10 % stepwise increments displays descending trend lines. However, the slopes of the trend lines of the relevance percentage in 10 % stepwise increments between basins reflect heterogeneity of data quality. The relevance percentage for Columbia is unsatisfactory, even in the first 10 % of the article list, since Columbia is both a district's name and a commercial brand (see Table 1). It makes sense that the data quality of Columbia River basin is not as good as others. Special treatment for Columbia should be adopted here to improve the data quality, and special treatments are needed for certain categories of basins and their corresponding treatments, as mentioned in Table 1. To do so, usually the enforcement of the frequency constraints, shown in Sect. 2.2.3 (i.e., raise the frequency setting for "water" and "river" to filter out the geopolitical articles as many), or the removal of the most irrelevant articles in the end of the dataset work well. With an anticipation of relevance percentage in mind, random sampling or manual reading of the last percentage of articles was often undertaken to check the data

Table 6. Manual reading results of representative river basins. Note: GBM is Ganges–Brahmaputra–Meghna.

Basin name	No. of articles downloaded	No. of articles after removing duplicates	No. of articles after removing irrelevant ones	Relevance percentage (%)
Nile	3872	3563	3164	88.80
Mekong	3253	2917	2291	78.54
GBM	1183	1092	724	66.30
Columbia	859	817	295	36.11

quality. For example, given the relevance percentage in 10 % stepwise increments for Columbia, raising the frequency setting of "water" and "river" to five times or removing of the last 40 % of the data retrieved in the original dataset, due to its low relevance in general, are feasible solutions to improve data relevance. For other basins with satisfactory data relevance, no further operation is needed, and for the other basins, similar operations as for the Columbia River basin can be adopted before further potential analysis.

4 Summary

The management of transboundary rivers is challenging both in terms of the political and environmental context in the 21st century. Data support is crucial for research of conflict and cooperation on transboundary rivers. The conventional construction of datasets by manual reading and extraction cannot meet the requirement for fast updating in the big data era. This study brings up a revised methodological framework, based on the conventional, and toolkit for news media dataset tracking of conflict and cooperation dynamics on transboundary rivers. The design of the framework follows Lasswell's communication model (Lasswell, 1948) closely and is involved with seven elements (who, with what intentions, in what situations, with what assets, using what strategies, reaches what audiences, and with what result). Basic search keywords were adopted from TFDD and further revised to include five blocks of terms to make it extensible and adjustable according to a certain research topic. Through Blocks 1 and 2, with a corresponding toolkit (shown in Fig. 2), a dataset covering transboundary rivers on a global scale can be generated, which is an improvement of the results of TFDD. All of the special treatments for basin names (shown in Table 1), country names (shown in Block 2), and term frequency setting of keywords (stated in Sect. 2.2.3) are crucial measures to enhance data quality and save manual efforts, which are an improvement beyond the achievements of TFDD. Following the methodological framework, a dataset with good tradeoffs between data relevance and coverage is generated. This study demonstrates the effectiveness of the framework and the potency of our toolkit. This framework possesses extensibility and compatibility to other research topics besides transboundary water resources management, since the search terms are adaptive and the toolkit is transplantable for related future research. With this revised framework and toolkit, research using news media tracking of conflict and cooperation dynamics on transboundary rivers will be much easier and more practicable.

The implications of this research can be manifested through how we can use the news media dataset generated by the methodological framework and toolkit tracking of conflict and cooperation dynamics on transboundary rivers. The dataset can serve as the foundation for further analysis, e.g., to study the attitudes, the topics of concerns, and the relationship between the evolution of the water governance network and the level of integrated water management along the evolution of water conflict and cooperation in the transboundary river basins. Ultimately, it can contribute to an understanding of the driving mechanisms and transformation laws of water conflict and cooperation. By capturing the characteristics of the life cycles of water conflicts and cooperation, future researchers can explore the temporal evolution trend and spatial distribution law of global transboundary water conflicts and cooperation events, as well as the guiding significance of appropriate policy intervention, and improve the level of global water security.

Meanwhile, this research and the dataset can also serve as a methodological and statistical foundation for quantifying the social dimension in sociohydrological approaches of understanding transboundary river systems. A recent attempt has been made to use a sociohydrological approach to tackle the feedback mechanism of co-evolved sub-systems (Lu et al., 2021). While a sociohydrological model can contribute to an understanding the complexity of the intertwined nature of transboundary river systems, quantifying the social variable has been challenging in general. There has been increasing recognition that news media provide a valid proxy to reflect the changing values and interests of each riparian country (Wei et al., 2021). Conflict and cooperation sentiments that are reflected in news articles have been adopted in sociohydrological models as the willingness for cooperation to validate the social sector of the model (Lu et al., 2021). When expanded to other river basins, this study could provide a methodological support in measuring the social sector of transboundary river systems more effectively.

Still, this study has some limitations which could be overcome.

1. The absence of newly registered rivers – the list of transboundary rivers adopted in this study includes 286 rivers, which could be expanded to 310 rivers in the near future.

2. The language limitation – the scope of this study is limited to English language newspaper only due to our limitation of language processing, which could be expanded to include more main languages and local languages in transboundary river basins.

3. Absence of tributary information – in the keywords generator, the tributaries of transboundary rivers are not included, which may lose content coverage to some extent. Future work can add more details concerning tributaries of transboundary rivers.

Appendix A

Records of the trial-and-error process are provided, as follows, to demonstrate the effects of various groups of frequency settings of keywords and how balance between relevance and coverage is approaching. The following two justification indicators of data relevance are adopted:

1. *Indicator 1 – the number of articles relevant to our research topic within 20 articles at 60 % of total data volume.* For example, there are 10 000 articles retrieved by the certain frequency setting of search terms in Lexis Advance, and we locate the article at exactly 60 % of total data volume, which is the 6000th article, and read 20 articles from there. Therefore, Indictor 1 is how many articles are relevant among the 6001st to the 6020th articles.

2. *Indicator 2 – the number of articles relevant to our research topic within 20 articles at 80 % of total data volume.* Similar to the algorithm of Indicator 1, if the total data volume is 10 000, Indictor 2 is how many articles are relevant among the 8001st–8020th articles.

Table A1 presents the results of trial-and-error process of frequency settings of the keywords for the Mekong, Nile, and Jordan river basins and shows that strong frequency settings enhance data relevance prominently and, at the same time, reduce the data volume to a large extent. To promise a balance between data relevance and coverage, proper frequency settings of search keywords should be adopted. In this study, test 6 is adopted as the final setting. Notice that Nile and Jordan river basins have an overwhelmingly large volume of data if no additional constraints are exerted. Therefore the search terms of ("Nile River" or "Nile Basin" or "Nile Water") or ("Jordan River" or "Jordan basin" or "Jordan water") are added to the basic search terms to limit data volume to an acceptable extent. While conducting trial-and-error processes, topics of irrelevant articles are also recorded to show the potential causes of irrelevance, and these may provide us some hints to modify the search terms for a better performance (see Table A2).

Table A1. Trial-and-error process of the frequency settings of keywords for the Mekong, Nile, and Jordan rivers.

Test	Frequency settings				Mekong			Nile			Jordan		
index	Basin name	Riparian country	Water	River	Data volume	Indicator 1	Indicator 2	Data volume	Indicator 1	Indicator 2	Data volume	Indicator 1	Indicator 2
1	1	1	1	1	27979	5	2	16227	8	4	28604	3	0
2	3	1	1	1	7536	13	10	6707	17	15	14028	4	1
3	5	1	1	1	4036	16	15	4157	19	16	13284	7	3
4	5	2	1	1	3695	16	16	4124	20	16	13263	7	1
5	5	2	2	2	3316	18	17	4017	20	18	5267	5	3
6	5	2	3	3	3102	19	17	3912	20	18	3830	7	10

Table A2. Recording of potential irrelevant topics for the Mekong, Nile, and Jordan rivers.

Mekong test index	Data volume	Indicator 1	Indicator 2
1	27979	Paper index 16787–16806	Paper index 22383–22402
		16789 – Coastal monument; 16790 – Catfish; 16791 – Missing American servicemen; 16792 – Missing people; 16793 – Business plan; 16795 – Life-style; 16796 – America navy river corps; 16798 – Life-style; 16799 – Grade national; 16800 – Travel to Vietnam and Cambodia; 16801 – Travel; 16802 – Travel along the river; 16803 – Travel; 16804 – Riots in Thailand; 16805 – Riots in Thailand	22383 – Family; 22384 – Soldiers; 22385 – Vietnam annexed Cambodia through reconciliation; 22387 – Vietnam sentences followers after trial of Buddha; 22388 – Culture; 22389 – Judicial delays in four cases; 22390 – Ban on swill feed; 22391 – Combine harvester race; 22392 – Relocate 7 million people to relieve pressure on overcrowded areas; 22393 – The man who reformed the United States Navy; 22394 – Books about Vietnam; 22395 – China plans to establish a national park system; 22397 – Image consulting; 22398 – Songkran Festival of Thailand; 22399 – Southeast Asian refugees; 22400 – Bow movement; 22401 – dueling event in Denver; 22402 – Cambodian adoptees
2	7536	Paper index 4521–4540	Paper index 6029–6048
		4522 – Former Vietcong say fighting on the Mekong River; 4524 – Elephant; 4528 – Luang Prabang; 4533 – Laotian culture and beauty; 4534 – Mekong River Journey; 4537 – Hero for Children's Rights; 4540 – Travel to Cambodia	6029 – Crossing the Mekong; 6030–6031 – Vietnam's rice exports; 6032 – DNA catch; 6033 – Vietnam cruise; 6041 – Inland river cruise; 6043 – National Geographic researcher Reno; 6044 – Mekong River Tourism; 6046 – Escape the molecular; 6048 – Buddha
3	4036	Paper index 2422–2441	Paper index 3229–3248
		2428 – A cruise ship on the Mekong River; 2434 – Mekong River travel; 2440 – Illegal timber trade; 2441 – Mekong River travel	3231 – Mekong Animals; 3235 – A search for missing American soldiers in Laos; 3236 – First impressions of Cambodia and Vietnam; 3239 – Travel to Vietnam and Cambodia; 3240 – Illegal logs are cut down
4	3695	Paper index 2217–2236	Paper index 2956–2975
		2218 – Travel to Vietnam and Cambodia; 2220 – Travel to Laos; 2231 – Pacific Command disaster response exercise in Vietnam; 2236 – Vientiane, capital of Laos	2962 – Visit Southeast Asia; 2963 – Cambodia is trying to save a rare Mekong river dolphin; 2968 – A search for missing American soldiers in Laos; 2975 – Travel to the Mekong
5	3316	Paper index 1990–2009	Paper index 2653–2672
		1997 – Mekong Tourism; 2007 – Mekong Prize winner	2662 – Mekong River travels in Thailand, Cambodia, and Vietnam; 2663 – Travel to Thailand; 2671 – Mekong navigation
6	3102	Paper index 1860–1879	Paper index 2482–2503
		1861 – Mekong Adventure	2483 – Laos arrests American manhunt for missing man; 2486 – Roaming along the Mekong river; 2488 – Drifting
Nile test index	Data volume	Indicator 1	Indicator 2
1	16227	Paper index 9736–9755	Paper index 12982–13001
		9736–9738 – Rebels in South Sudan; 9739–9740 – Archaeologist; 9741 – Israel may withdraw from the West Bank; 9744 – Slavery in ancient Egypt; 9745 – Kurdish rebels in Turkey; 9748 – Travel to Egypt; 9750–9752 – Egypt's interior minister refused to allow "militias" to enter	12982 – Farmers and the Egyptian government fought for years in a legal battle; 12983 – Cursed Island; 12984 – Detectives use modern science to solve a 3300-year-old murder mystery; 12985 – Violence in Egypt; 12987 – Tagore festival in Egypt; 12988 – Bossi language learning courses; 12989–12992 – Russian plane crash; 12993 – Egypt's population grows; 12995 – Reviving Egypt's tourism industry; 12996–12997 – Anti-government demonstrations in Sudan; 12998 – Tourism landscape; 13001 – Egyptian court holds second mass trial

Table A2. Continued.

Nile test index	Data volume	Indicator 1	Indicator 2
2	6707	Paper index 4024–4043	Paper index 5366–5385
		4037 – Lake Victoria renamed; 4038 – Egyptian journalist Resigns; 4041 – Travel along the Nile	5371–5373 – Luxor Temple; 5378 – Egyptian history; 5380 – The uprising in Egypt is resurgent
3	4157	Paper index 2494–2513	Paper index 3326–3345
		2509 – Changes along the Nile	3326 – Conflict in South Sudan; 3328 – Travel along the Nile; 3333 – Sudanese refugees; 3343 – Sudan Peace Conference
4	4124	Paper index 2474–2493	Paper index 3299–3318
		No irrelevant articles	3301 – Egypt will withdraw 3.4 million from federal reserves to meet food needs; 3302 – Nile culture; 3305 – Sudan earthquake sequence 1990–1991 and the extent of the East African Rift Valley system; 3317 – Ethiopia – Lakeside cities overcome Africa's tourism crisis
5	4017	Paper index 2410–2429	Paper index 3214–3233
		No irrelevant articles	3216 – Displaced South Sudanese; 3232 – Sudan earthquake sequence 1990–1991 and the extent of the East African Rift Valley system
6	3912	Paper index 2347–2366	Paper index 3130–3149
		No irrelevant articles	3134 – An English man crosses the Nile on foot; 3136 – Displaced South Sudanese
Jordan test index	Data volume	Indicator 1	Indicator 2
1	28604	Paper index 17162–17181	Paper index 22883–22902
		17162 – Gaza's prison; 17164 – Israel security Separation Wall; 17165 – Jesus through Anne Rice's eyes – A book review; 17166 – The king of Morocco visited the United States; 17167 – Jewish terrorist group; 17168 – Palestinians demonstrated in Israeli-occupied territory; 17169 – Jordan's king urged the Palestine Liberation Organization to recognize Israel; 17170–17172 – United States: Israeli-Palestinian peace agreement; 17173 – Riding; 17174 – The PLO will not meet with American officials; 17175 – There has been violence in Jerusalem; 17176 – Israel's democracy and Arab population; 17177 – Women go to Palestine to resolve violence; 17178 – Music; 17181 – Hotel prices in Australia have fallen along with Asian growth	22883 – Joshua Myron, zionist who fought the Turks, died; 22884 – Did Netanyahu explain why the Palestinians did not reach a deal; 22885 – Ottawa ordered compensation for disabled first Nations children; 22886 – Sacramento State University student arrested in terrorist ring; 22887 – South Africa: Two white-owned farms to be confiscated in land reform; 22888 – State Department envoy meets with Palestinian Christians who oppose Israel; 22889 – Three-year-old girl shot dead in Gaza; 22890 – Mormon Temple renovation; 22891 – Yahsat hosted a forum on the humanitarian use of satellite broadband; 22892 – Humanitarian work; 22893 – Jenkins' death; 22894 – Interfaith activity; 22895 – The ultimate consultant; 22896 – Exhibition in the West Bank; 22897 – The Jewish population in the occupied West Bank is set to more than double this year; 22898 – Police have arrested a suspect in the Maverick shooting; 22899 – Disturbing violence; 22900 – The etymology of the names Israel and Jacob; 22901 – Terrorism; 22902 – The city of Midville's first citywide master plan for trails

Table A2. Continued.

Jordan test index	Data volume	Indicator 1	Indicator 2
2	14028	Paper index 8417–8436	Paper index 11222–11241
		8417 – Missionary trip to the West Bank; 8418 – Against Islamic State militants; 8419 – Israel imposed sanctions on Gaza; 8420 – Arafat; 8421 – Israeli withdrawal; 8422 – The Israeli military has questioned an Arab mayor in the West Bank; 8423 – It is widely believed in Israel that the current situation in Judea, Samaria, and Gaza cannot and should not continue; 8424 – Arafat rose up; 8425 – Palestinians say Washington accepts their approach to ending attacks on Israel; 8426 – The Likud trounced Sharon; 8429 – Pope endorses Palestinian Aspirations; 8430 – The Israeli authorities have jailed three senior Palestinian leaders without trial; 8433 – Defense of the Jewish State; 8434 – Australia's relationship with Israel; 8435 – Former commander of the Arab Legion Grubb Pasha has died; 8436 – A leadership void is holding Egypt back	11222 – Manitoba – The Assembly of First Nations supports the Declaration of the Manitoba Chiefs; 11223 – The Sheikh Hussein Bridge across the Jordan River was completed; 11224 – The Israelis shot at two Jordanians; 11225 – The sick Menachem Begin; 11226 – Top election official supports south Jordan petition; 11227 – How did a high school student in Nuremberg talk about crossing Israel; 11229–11230 – Protesters in Amman burn an Israeli flag after the judge's killing; 11231 – SCR 591 recognizes São Paulo's historic landmarks and museums; 11232 – Expand light rail public transport; 11233 – Update from AFPTV on Tuesday; 11234–11236 – Palestinians in the West Bank are under increasing economic pressure; 11237–11239 – Sharon army; 11240 – Silt diversion walls in East Jordan; 11241 – Traveler's cheque
3	13284	Paper index 7970–7989	Paper index 10627–10646
		7970 – Israel faces a demographic threat; 7971 – Israel and Palestine live side by side in peace; 7972 – Peace negotiation; 7973 – Palestinian Elections postponed; 7974 – The countryside camping; 7975 – A week of news; 7976 – The American president meets with Jordan's king; 7977 – Watch the Pope's Middle East pilgrimage online; 7979 – A secret meeting of Arab and Israeli writers; 7983 – Top story on Tuesday; 7984–7985 – Better support for Aboriginal children; 7986 – Israel Archives; 7963 – Catholics and Muslims seek dialogue	10627–10629 – Individual account; 10630 – Witness: The Jordanian defendant had ties to Osama bin Laden; 10631 – The wounded mayor vowed to continue the fight for Palestinian rights; 10632 – Former U.S. President: Middle Eastern leaders must tell their people that compromise is honorable; 10633 – Jordan baptism site sells bottled holy water tender; 10637 – A letter from Israel; 10638 – Public money spent on Park Avenue; 10639 – The Israeli prime minister has proposed the creation of an independent Palestinian state; 10640 – The European Commission has issued a final warning to the UK over repeated violations; 10641 – Vote split; 10642–10643 – Jordan's parliament failed to overthrow the government; 10644 – Jordan Valley Trail; 10645 – Possible hazards caused by pumping water near rivers; 10646 – Task biography
4	13263	Paper index 7958–7977	Paper index 10610–10629
		7958–7959 – Terrorism; 7960 – The new chief rabbi is a firebrand nationalist; 7962 – Silt diversion wall; 7963 – Catholics and Muslims seek dialogue; 7964 – The Palestinian government's plan; 7965 – Palestinian refugees; 7966 – Jordan bridge; 7968 – The United States is pushing for a Middle East peace plan; 7969 – The Arabs conspired to blockade Israel; 7970 – Jordan River Cycle Path; 7971 – Mr Netanyahu linked peace to Palestinian recognition of Israel as a Jewish state; 7976 – Jordan refugee woman craftsman	10610–10611 – Jordanian-British student volunteers seek positive change in Western and Arab societies; 10612 – Jerusalem Liberation Army; 10613–10614 – Palestinian textbooks versus Israeli textbooks; 10615 – Some rare right whales like winter in Maine; 10616 – The PLO is preparing to move to Gaza and Jericho; 10617–10618 – Arab Bank employees volunteer in Aguilón; 10619 – East Jordan Commissioner Thomas Breney asked the board to vote on hiring a full-time fire chief; 10621 – The FBI is demanding payment for the local youth's treatment; 10622 – The federal government is seeking to appeal the ruling on medical costs; 10623–10624 – Negotiations between Israel and Egypt; 10625 – The Conservative Party is appealing against the treatment ruling; 10626 – Fatah's al-Aqsa Brigades killed a woman in Nablus accused of collaborating; 10627 – Mr Hotovely bemoans Likud's "schizophrenia" over two countries; 10628 – Kiryat Arba population; 10629 – In the Likud debate, the two-state solution is schizophrenic

Table A2. Continued.

Jordan test index	Data volume	Indicator 1	Indicator 2
5	5267	Paper index 3160–3179	Paper index 4214–4233
		3160 – Mr Netanyahu does not fear being blamed if the London meeting is inconclusive; 3161 – Forty-eight hours in Amman; 3162 – Israel has begun clearing land mines at the site of Jesus' baptism in the West Bank; 3164 – Covenant of Israel; 3166 – Israel's efforts to win Over Christian tourists; 3168 – Christian sites are covered in land mines; 3170 – Community Notes: Volunteer; 3171 – Edit the letter in the pouch; 3173 – Travel; 3174 – Mormons and non-Mormons; 3175 – Immigrants find peace and opportunity in Corona; 3176–3177 – Hymns to Haaznu on a biblical urn; 3178 – A joyous gathering of prominent Israeli and PLO officials in the three years since the Signing of the Oslo Accords; 3179 – The new chief rabbi is a firebrand nationalist	4214 – Jordan sewage pump; 4215 – Across the border; 4216 – In memory of a distinguished journalist; 4217 – Travel manuscript; 4218 – Some news; 4219 – Joint Technology Center; 4220 – Leprosy; 4221 – Israeli soldiers pass through barbed wire in Wazzani; 4223 – Storm hits northern Michigan; 4224 – Jon and Martha Jensen of Petoskey gave birth to daughter Ruby Susan at Northern Michigan Hospital; 4225 – Peter and John stood before the Sanhedrin; 4228 – Obituary; 4229 – The UAE supports Jordan in implementing its development plans; 4230 – Russia and the United States are set to reach a new agreement by next summer on deep cuts in strategic offensive weapons; 4231–4232 – Hussein riots; 4233 – The East Jordan Chamber of Commerce distributes community awards
6	3830	Paper index 2298–2317	Paper index 3064–3083
		2298 – The establishment of a provisional Palestinian state; 2299 – Ways to promote Jordanian and British military cooperation; 2301 – Israeli Prime Minister Benjamin Netanyahu has said he will see the final results of a peace deal; 2303 – Epiphany; 2306 – Winners of the first Tourism Promotion Peace Prize have been announced; 2308–2309 – Better support for Aboriginal children; 2310 – Jordan has "ridiculed" Israeli ministers' efforts to block Palestinian statehood; 2311 – Pope Francis is making a visit to the Holy Land; 2314 – The Jordan River has long been a source of entertainment for Wasatch central city dwellers; 2315 – Community news; 2316 – South Jordan celebrates its 150th anniversary; 2317 – Jordan Trail	3064 – American troops were sent from California to Utah during the Civil War; 3065 – A Seattle moving company is offering spring deals; 3066–3067 – Seattle Moving Company; 3068 – Opponents of sewage treatment plants; 3069 – Soldiers swept away by the Jordan River; 3073 – Civil servants are considering a one-day strike on Monday; 3074 – Pilgrims mark baptism traditions in the Jordan River; 3075 – Word games; 3076 – People's Fund grant project

Code and data availability. The data and code used in this study are publicly available on Zenodo (including the basin–country dictionary, dictionary of country names with different formats or special country dictionary, dictionary of basin names with different formats, and the Python code of searching term generator; https://doi.org/10.5281/zenodo.5112624, Guo et al., 2021).

Author contributions. LG, JnW, and FT designed the research framework. LG collected the data and conducted the data analysis. LG, JnW, KZ, and JlW conducted manual reading for the case studies. LG, JnW, and FT composed the paper, with contributions from KZ and JlW.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Special issue statement. This article is part of the special issue "Socio-hydrology and transboundary rivers". It is not associated with a conference.

Financial support. This research has been supported by the NSFC (grant no. 51961125204) State Key Laboratory of Hydroscience and Engineering (grant no. 2022-KY-03).

Review statement. This paper was edited by Murugesu Sivapalan and reviewed by two anonymous referees.

References

Alsumait, L., Barbará, D., Gentle, J., and Domeniconi, C.: Topic Significance Ranking of LDA Generative Models, in: Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg, 6782, 2009.

Bernauer, T. and Böhmelt, T.: International conflict and cooperation over freshwater resources, Nature Sustainability, 3, 350–356, https://doi.org/10.1038/s41893-020-0479-8, 2020.

Cooper, S.: Bringing Some Clarity to the Media Bias Debate, Communications Faculty Research, https://mds.marshall.edu/communications_faculty/2 (last access: 7 May 2021), 2005.

Esri Data and Maps: World Countries (Generalized), ArcGIS Hub, https://hub.arcgis.com/datasets/2b93b06dc0dc4e809d3c8db5cb96ba69_0, last access: 14 April 2021.

Guo, L., Wei, J., Zhang, K., and Tian, F.: Toolkit for news media dataset tracking of conflict and cooperation dynamics on transboundary rivers, Zenodo [code], https://doi.org/10.5281/zenodo.5112624, 2021.

Howland, D., Becker, M. L., and Prelli, L. J.: Merging content analysis and the policy sciences: A system to discern policy-specific trends from news media reports, Policy Sci., 39, 205–231, https://doi.org/10.1007/s11077-006-9016-5, 2006.

Jacobi, C., van Atteveldt, W., and Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling, Digital Journalism, 4, 89–106, https://doi.org/10.1080/21670811.2015.1093271, 2016.

Jiang, H., Qiang, M., and Lin, P.: Assessment of online public opinions on large infrastructure projects: A case study of the Three Gorges Project in China, Environ. Impact Ass., 61, 38–51, https://doi.org/10.1016/j.eiar.2016.06.004, 2016.

Khoo, C. S. and Johnkhan, S. B.: Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons, J. Inf. Sci., 44, 491–511, https://doi.org/10.1177/0165551517703514, 2018.

Krippendorff, K.: Reliability in Content Analysis, Hum. Commun. Res., 30, 411–433, https://doi.org/10.1111/j.1468-2958.2004.tb00738.x, 2004.

Lasswell, H.: The Structure and Function of Communication in Society. The Communication of Ideas, edited by: Bryson, L., The Institute for Religious and Social Studies, New York, 14 pp., 1948.

Lu, Y., Tian, F., Guo, L., Borzì, I., Patil, R., Wei, J., Liu, D., Wei, Y., Yu, D. J., and Sivapalan, M.: Socio-hydrologic modeling of the dynamics of cooperation in the transboundary Lancang–Mekong River, Hydrol. Earth Syst. Sci., 25, 1883–1903, https://doi.org/10.5194/hess-25-1883-2021, 2021.

McCracken, M. and Wolf, A. T.: Updating the Register of International River Basins of the world, Int. J. Water Resour. D., 35, 732–782, https://doi.org/10.1080/07900627.2019.1572497 2019.

Mirumachi, N.: Transboundary Water Politics in the Developing World, Routledge, ISBN 9780415812962, 2015.

Neethu, M. S. and Rajasree, R.: Sentiment analysis in twitter using machine learning techniques, 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 1–5, https://doi.org/10.1109/ICCCNT.2013.6726818, 2013.

Pohl, B., Carius, A., Conca, K., Dabelko, G., Kramer, A., Michel, D., Schmeier, S., Swain, A., and Wolf, A.: The rise of hydrodiplomacy. Strengthening foreign policy for transboundary waters, adelphi, Berlin, https://doi.org/10.13140/2.1.4035.5848, 2014.

Racine, E., Waldman, S., Rosenberg, J., and Illes, J.: Contemporary neuroscience in the media, Soc. Sci. Med., 71, 725–733, https://doi.org/10.1016/j.socscimed.2010.05.017, 2010.

Sadoff, C. W. and Grey, D.: Cooperation on International Rivers: A Continuum for Securing and Sharing Benefits, Water Int., 30, 420–427, https://doi.org/10.1080/02508060508691886, 2005.

Smith, L. W.: Stakeholder analysis: A pivotal practice of successful projects, Project Management Institute Annual Seminars & Symposium, Houston, TX, Newtown Square, PA, https://www.pmi.org/learning/library/stakeholder-analysis-pivotal-practice-projects-8905 (last access: 22 April 2021), 2000.

Transboundary Freshwater Dispute Database: Program in Water Conflict Management and Transformation, Oregon State University, https://transboundarywaters.science.oregonstate.edu/content/transboundary-freshwater-dispute-database (last access: 4 February 2021), 2008.

Transboundary Waters Assessment Programme: Transboundary River Basins-Status and Trends, http://twap-rivers.org/assets/GEF_TWAPRB_FullTechnicalReport_compressed.pdf (last access: 30 November 2020), 2016.

UNBIS Thesaurus: UNBIS Thesaurus, http://metadata.un.org/thesaurus/?lang=en, 2021.

United Nations and UNESCO: Progress on Transboundary Water Cooperation 2018: Global Baseline for SDG 6 Indicator 6.5.2, Paris, France, https://doi.org/10.18356/f6afa45b-en, 2019.

Weaver, D. A. and Bimber, B.: Finding News Stories: A Comparison of Searches Using Lexisnexis and Google News, Journalism Mass Commun., 16, 515–530, 2008.

Wei, J., Wei, Y., Tian, F., Nott, N., de Wit, C., Guo, L., and Lu, Y.: News media coverage of conflict and cooperation dynamics of water events in the Lancang–Mekong River basin, Hydrol. Earth Syst. Sci., 25, 1603–1615, https://doi.org/10.5194/hess-25-1603-2021, 2021.

Wolf, A. T.: The Transboundary Freshwater Dispute Database Project, Water Int., 24, 160–163, https://doi.org/10.1080/02508069908692153, 1999.

Wolf, A. T., Natharius, J. A., Danielson, J. J., Ward, B. S., and Pender, J. K.: International River Basins of the World, Int. J. Water Resour. D., 15, 387–427, https://doi.org/10.1080/07900629948682, 1999.

Yoffe, S. and Larson, K.: BASINS AT RISK: WATER EVENT DATABASE METHODOLOGY, Oregon State University, chap. 2, 36, https://transboundarywaters.science.oregonstate.edu/sites/transboundarywaters.science.oregonstate.edu/files/

Database/Data/Events/Yoffe&Larson-EventCoding.pdf (last access: 20 June 2020), 2001.