Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

journal homepage: www.elsevier.com/locate/ijdrr

Towards improved knowledge about water-related extremes based on news media information captured using artificial intelligence

Joao Pita Costa [a], Luis Rei [b,e], Nejc Bezak [d,*], Matjaž Mikoš [d], M. Besher Massri [b], Inna Novalija [b], Gregor Leban [c]

[a] International Research Centre for Artificial Intelligence under the auspices of UNESCO (IRCAI), Ljubljana, Slovenia
[b] Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia
[c] Jožef Stefan International Postgraduate School, Jamova Cesta 39, 1000 Ljubljana, Slovenia
[d] University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia
[e] Jožef Stefan International Postgraduate School, Jamova Cesta 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Climate change resilience
Web-based observatory
Hydrometeorological extremes
Artificial intelligence
Machine learning
Multilingual news
Twitter (X) nowcasting
Text mining

ABSTRACT

Recent advances in machine learning have enabled near real-time retrieval of information from textual documents impacting a wide range of knowledge domains. This advantage makes of the insight extracted from text-based documents (e.g., reports and news) on water-related events an invaluable source of information that complements traditional approaches. By leveraging machine learning, we can not only characterize the event and determine its magnitude or phases of extreme weather event, but also identify its core elements. This is especially crucial in the current era of climate change, where extreme weather events such as floods or thunderstorms are becoming more frequent and unpredictable. By improving our ability to detect and analyse such events, we can enhance our alert systems and take more effective action to mitigate their impact. In this paper we discuss the role of worldwide media observation in extracting and estimating hydrological characteristics of floods, droughts, and heat waves, through the analysis of three case studies, complementing the information provided by traditional monitoring and measurement methods as an earlier but weaker signal. The results presented in this study indicate that the news media signal can be regarded as relatively good proxy of flood dynamics. It can capture the temporal dynamics of the event, and, in some cases, there could be a clear up to 1-day lag between the peak discharge values (i.e., the most extreme flood situation) and the peak in the number of published news. This lag can be attributed to the time needed by journalists to respond to the situation in publishing related news articles covering the event. Our result show that national and regional news can cover well local events. When compared to floods, drought conditions are less explicitly detected from the media. Our result show that European April 2022 drought did not produce much activity in the media while the combination of drought and extreme heat in July 2022 yielded a significant media coverage throughout the Europe. Hence, this can be attributed to the fact that hydrological drought such as low river flows do not attract much attention by the media unless there is a significant impact on the society. Therefore, media signal can be regarded as a relatively good proxy of the hydro-meteorological conditions in case there is a significant impact on the society such as extreme floods causing many casualties and large economic damage.

* Corresponding author.
E-mail addresses: luis.rei@ijs.si (L. Rei), nejc.bezak@fgg.uni-lj.si (N. Bezak).

https://doi.org/10.1016/j.ijdrr.2023.104172
Received 11 April 2023; Received in revised form 16 October 2023; Accepted 1 December 2023
Available online 7 December 2023

1. Introduction

In the current era of open data and shared research, we can derive valuable information to help the stakeholders in the water sector to complement their established approach by incorporating artificial intelligence (AI) capabilities to fit the specific challenges of hydrological activities. The exploration of knowledge from text-based data using machine learning has been recently successfully applied to other scientific disciplines, from the semantic enrichment analysis of environmental policies [1], to the recommender systems in healthcare [2] and the COVID observatory [3]. Particularly, the exploration of the potential and the limitations of social media to observe local and global feeds and generate alerts is explored in a diversity of case studies, many of which are showing the importance of the domain knowledge to fine tune the usefulness of this signal [4].

The investigations on extreme hydrological events using social media data have been focusing on natural disaster detection based on the analysis of the public input associated, e.g., with the frequency of keywords. Some recent studies have investigated this relationship, mostly from the scope of Twitter (X) data analysis. For example [5], showed that combined geo-social media (i.e., tweets) can be used for streamflow estimation in relation to flood monitoring. The preceding work in Ref. [6] goes towards the engagement of citizen science in the context of the usage of citizen reported images and computer vision techniques for flood early detection. That crowdsourcing approach using social media is later used in Ref. [7] to build near real-time flood maps using geo-statistical methods to overcome the inherently unreliability of tweets [8]. compared geotagged topics in social media with rainstorm event in China, while [6] focused on the usage of a deep learning framework to estimate flow-water levels from images shared in tweets. Similarly, also [9] analysed the natural hazard monitoring and social media during floods and [10] investigated the relevance participatory citizenship through social media when integrated to hydrological models. This knowledge is enhanced with the information provided by satellite imagery and in the context of geographical information systems (GIS) [11]. The potential use of social media for disaster risk reduction has been investigated also in other similar studies [12,13,14,15]. Social media was also used to assess road status during floods [16]. An interesting aspect of that work is that they point out that this information is often reported by local news media. This ties into [17]; which used social media to estimate the severity of the 2010–2011 South East Queensland Floods. The authors analysis shows that the most active social media accounts were news channels. Their work showed evidence that their media data could be used to identify fluctuations in the severity of the disasters and estimate the area impacted. In the context of news articles [18], used Named Entity Recognition (NER) to extract locations from news and assess urban flood susceptibility in Dalian, China. They reported no significant difference between using their methodology and the official planning report [19]. were able to track the long-term impact of droughts in Germany between 2000 and 2021 by analyzing news articles. They used supervised classification models to detect socioeconomic impacts and hierarchical clustering to detect outlier locations extracted using named entity recognition. They noted that their methodology requires a significantly lower workload than conventional impact assessment methods. Recognizing the importance of news media in the response to disasters, MediaEval 2019 Multimedia Satellite Task [20] had the stated purpose of improving the ability to estimate flood severity from news. The task involved analysing online news articles and their associated images with the goal to create a classifier that determines whether the image shows a person standing in water above the knee. Participants were encouraged to use both text and image analysis.

In this paper, we focus on the signal of worldwide news media and discuss how text mining algorithms can be made useful in exploring the impact of extreme weather events, and extracting from its signal a timeline of important event features that can serve as input for timely alerts. In contrast to previous work, we use the news articles published in the worldwide media aided by multilingual technologies [21] to understand through machine learning what news articles relate to the same event. Unlike previous work, which relies on monolingual named entity recognition, we rely on multilingual entity linking to directly provide disambiguated entities, namely locations [22]. Another methodological difference is the use of publication time and linked entities in clustering together with a multilingual text representation which allows us to both handle multiple languages and create clusters based on time, and to do so as articles are published, rather than rely on an archive. Like in Ref. [19]; this clustering also allows for the detection of outlier locations. We also explore multiple different extreme weather events: a local flood event with a short duration in Spain, a European-wide flood event, and a western European drought event. This provides a broader disaster type coverage than previous work under the umbrella of a single methodology. Also different from much of the most closely related work is that the primary ground truths for our methodology are discharge values, the Soil Moisture Anomaly Index and the Standardized Precipitation Index. In the context of prior research, we offer additional evidence of the significance of news analysis. We do this by employing a distinct methodology, concurrently studying multiple dissimilar disasters, and by comparing the results of media analysis with different hydrological measurements. From news articles we can: (i) extract information directly from the text, using a machine learning-based methodology that allows for name entity recognition and disambiguation in multiple languages; and (ii) estimate parameters (e.g., extension of the damage caused by an extreme event) having enough data in the collected articles about the same event to ensure a certain hypothesis.

The main contribution of this paper is the novel approach, methodology and technology for the usage of text-based data in the analysis of hydrological extreme events that are fed by earlier (but weaker) signals from news media, anticipating the classical hydrological methods are more expensive and time consuming. Thus, we show that this approach can help overcome the difficulties to assess the impact of the extreme weather events by using open data, following the recent research work [23].

In this paper we consider two types of events: (a) the news event, i.e., an occurrence or happening that is considered to be of significant interest or importance to the public, reported by multiple news media outlets; and (b) the classical hydrological events, defined as i) a flood, being an overflow of water that submerges land that is usually dry, covering many flood types such as river (fluvial), coastal floods, storm surges, inland flooding and flash floods among others [24,25], or defined as ii) a drought, being either meteorological, climatological, agricultural, hydrological, or socioeconomic drought. The first four drought types are defined by physical, hydrometeorological, or biological parameters, while the fifth centres on the impacts of drought on the society [26,27].

We will show how we can extract the various levels of severity based on the news and describe the characteristics of these separate phases in the identified events. Notice that not always the severity reported by the news or the change of phases in that severity corresponds directly to the magnitude of the hydrological event. This is since the severity in the event reported in the news articles is generated by the "journalistic happenings" related to the event that do not always correspond to hydrological phenomena. Thus, the signal from news is to be considered a weak signal, relative to the accurate and strong signal captured and/or computed by the classical hydrological observational methods and models.

To better understand the impact of the event and the succession of its phases, we need to explore the characteristics of the news that more accurately reflect characteristics of the event itself. In doing so, we must also consider the values (also called criteria or factors) [28], that influence the selection and presentation of events in the news and are assumed to be qualities of news media than inherent characteristics of an event itself [29]. [30] showed, in a large-scale manual analysis, that story prominence was influenced by elite characteristics, the value that states that events concerning global powers and famous people receive more coverage [31]. showed that frequency, the value that states that events that occur suddenly and fit well with the news organization's schedule are more likely reported than those that occur gradually, was discernible in Event Registry. Thus, we can expect that a flood in a large and rich country will get more coverage than a flood in a small and poor country, that a celebrity being caught in a flood will significantly change the coverage of an event, and that long term effects of an event might often be underreported.

For this study, we mostly focus on two types of events: floods and droughts (also in relation to heat waves), but with different temporal resolution (from an hourly temporal scale to daily or weekly time scales) and spatial scales (local, regional, national, and transnational levels). In that, we consider the definition of a flood from news and hydrology points of view: (i) localized flash flood; (ii) regional large-scale flood events occupying part of a country; (iii) multi-country event in, e.g., one month with infrequent coverage in news.

To complement this approach, we use a text mining classifier based on text similarity to offer another layer of extracted information, where now the source is the feed of published scientific articles from the Microsoft Academic Graph (MAG) dataset [32], representing a much improved level of information retrieved, in relation to the knowledge extracted from media articles that often depend on a journalistic perspective. The more than 120 million articles in MAG represent, in this study, a worldwide scientific perspective about a certain water-related topic, allowing to pursue further investigation over a certain event considering in the characteristics of such an event as identified by text mining methods in the analysis of news articles about it extracted as named entities. From this knowledge we derive best practices that can complement the strategies of hydrology practitioners and explore alternative approaches for specific problems consequent from these extreme water-related events. Furthermore, we reuse this intelligence advantage to refine the social media nowcasting analysis that serves to better understand how to improve the noise of this weak signal, learning from the acquired historical data on events that relate to those to be forecasted, based on identified characteristics.

Following a section on used data and methods and the used technologies (section 2), we present the main results together with their discussion, distinguished from content point of view, between the relations of worldwide and local news to hydrological events (section 3), the insights of the latter on the improvement of the social media signal (section 3), and the use of machine learning in high quality information extracted from published science to complement the statistical information of global and local indicators (section 3). The main objectives of this study are:

I. To capture new previously unidentified aspects of an extreme water-related event (i.e., information extracted from the news media) that can complement classical hydro-meteorological methods.
II. To measure aspects of an extreme weather event that are harder to measure through classical methods and learn from the past events from the perspective of disaster risk reduction.
III. To estimate the magnitude of an event (i.e., return period) from the news and provide information on the spatial extension of the event (i.e., extracted location and/or magnitude).

The proposed approach shows results across three case studies of different nature, corresponding to three well-known recent extreme events either based in a specific location or with multinational extent, relating to flood or drought:

1. Sequence of heavy precipitation and flash flooding of 12 and September 13, 2019 in Eastern Spain [33].
2. July 2021 floods in Europe, starting on July 13, 2021 [34,35].
3. Drought in May 2022 in France, Spain, Portugal that prolonged to June, July and extended to United Kingdom, Germany, Scandinavia, prolonged again into August [36,37].

We will use the above extreme events as running examples to achieve the above stated objectives of this study.

2. Data and methods

2.1. News articles data

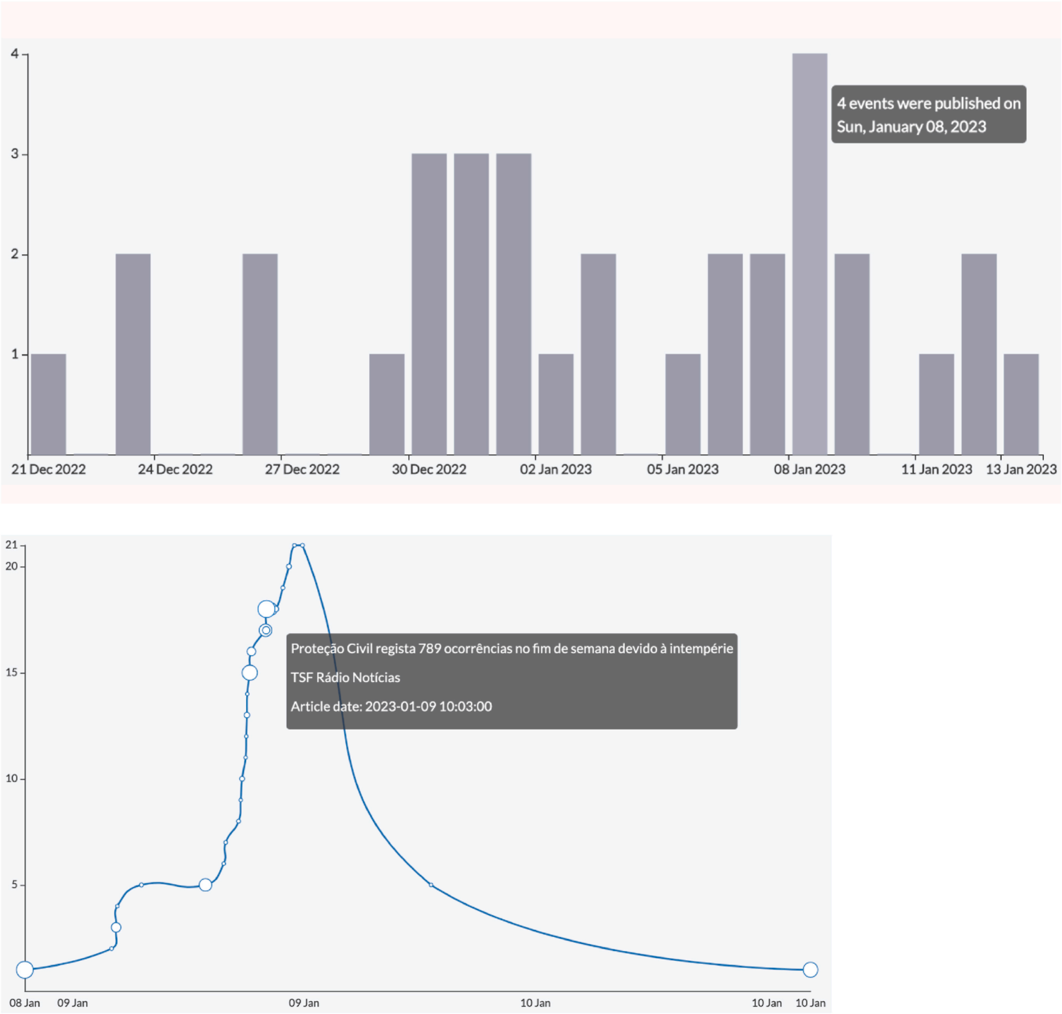
In this work we use online news articles. These articles are crawled using the news collection engine IJS Newsfeed (available at http://newsfeed.ijs.si/) [38]. The crawler periodically fetches updates from online news websites. The update time is automatically adjusted for each source to decrease the time between publication and fetching operation. When a new article is fetched, the crawler extracts the article text as well as useful metadata present in the HTML such as, e.g., canonical URL and time of publication. At the time of writing, over 75,000 sources are crawled, resulting in around 150,000 articles per day, distributed over 50+ languages. The core of this study is based on that collected news articles, allowing to generate visualization modules to analyse their insights (e.g., main concepts and relation between them, frequency, and subcategories of the search topic, etc). It allows the user to explore that data over a variety of data visualization modules that help identify concepts and entities common to a sample of news articles generated by a query on, e.g., “flood”. The example in Fig. 1 shows the historical data on floods in Spain for a time window between 2010 and 2020 identifying 482 news articles across 50+ news outlets. We can use this data to understand the common characteristics of such water-related events and extract concepts related to them that could help us identify news articles on similar events and partial solutions that could be planned and applied when such an event takes place. From this data we can extract directly some of the characteristics of the event including, e.g., the event type (matching a configurable predefined categorization including “flood”, “drought”, “heat wave” or “heavy rain”), and also date and location (extracted directly from the article or derived from the news venue when the latter is not available in the text [39]).

2.2. Scientific articles data

To explore the information provided by the dataset of scientific articles and accepted patents we ingested the Microsoft Academic Graph data with 126 million articles in a wide variety of topics, extended by the also ingested PubMed/MEDLINE dataset, covering over 28 million articles hand-annotated by health experts, and particularly relevant in the context of the automated classification of text for water contamination as long-term flood impacts. The latter is particularly relevant in the context of the analysis of the long- and medium-term impact of the event in public health.

2.3. Social media data

In this study we collected 7,407,856 tweets worldwide, through the scope of a Twitter (X) research account focusing on water event-related topics including “flood” (593,469), “storm” (952,547), “drought” (2,252,277) and “heat wave” (773,734), between 21 Nov 2019 and 9 Oct 2021. We identified 3,124 posts talking about flood events, geolocated in Germany throughout 2021 mostly in German and in English, with substantial discussions between 12 and 19 of July, reporting on the extreme event discussed in the case



4 events were published on Sun, January 08, 2023

Proteção Civil regista 789 ocorrências no fim de semana devido à intempérie

TSF Rádio Notícias

Article date: 2023-01-09 10:03:00

Fig. 1. Identifying the flood incidents of early 2023 in Portugal from historical news data allowing to retrieve specific news articles (above); and exploring the distinct phases of a related event from the published news about it (below).

study (2). Due to the limitations of the Twitter research account, we did not collect enough data to cover the cases studies (1) and (3) happening before and after the research contract established with Twitter, respectively.

2.4. Hydrological measurements data

For the comparison of the media signal with actual data we used the Soil Moisture Anomaly Index and the Standardized Precipitation Index (SPI) [40,41] as a proxy for the identification of drought throughout Europe. The soil moisture anomaly maps are prepared based on the 30-day moving window with a spatial resolution of 0.1° and updated every 10 days meaning that this index is available for every 10 days [41]. It is computed based on the hydrological model LISFLOOD root zone soil moisture, MODIS land surface temperature and ESA CCSI remote sensing skin soil moisture (Copernicus, 2022). A detailed description of this product is available in previous publications [40,41]. For this study, we checked minimum and mean values of this index over the selected country boundary. The lower the value of this index, the more extreme the drought conditions assumed are. Moreover, we also checked the SPI for different accumulation periods (1 and 3 months) where the SPI-1 (1 month accumulation) can be regarded as an indicator of the quick impact (Copernicus, 2022). The data resolution is one decimal degree and monthly precipitation data is used to calculate the SPI (Copernicus, 2022). Additionally, for the evaluation of the Water Observatory output in relation to floods we also used the river discharge data from the Global Flood Awareness System [42] where river discharge in the last 24 h is available. Hence, this is an average discharge value of the 24-h interval. Maximum and mean specific discharge values were computed per country boundary. Additionally, we also used hourly precipitation data from the ERA5-Land reanalysis product and the aggregated precipitation data from the same product [42] to compare the media signal to actual hydro-meteorological characteristics. For the comparison purposes the Pearson correlation coefficient was used, which measures linear dependence between variables [43]. For the visual representation we used line plots for the comparison of media and hydrological behaviour. The main assumption behind this analysis is that there will be a different relationship between media response and hydrological situation for the selected events (Table 1) since the hydrological characteristics of these events are different.

Table 1 describes the coverage of the media and social media captured for the purpose of this study across the three selected case studies. Regarding the direct comparison to flood and drought events in the hydrological perspective, it is hard to quantify and moreover to distinguish between one multinational event and several local events (e.g., the large European flood in 2021 was a unique flood event driven by spatially extensive extreme rainfall, but with several significant instances).

2.5. Wikification & categorization

The process of semantic annotation referred to as wikification, is utilizing Wikipedia as a source for potential semantic annotations. In this context, Wikipedia is viewed as a comprehensive and general ontology, with each page representing a particular concept. The connections between these concepts are represented by internal hyperlinks between different Wikipedia pages, as well as by Wikipedia's categorization and cross-language links. We use the tool Wikifier (https://wikifier.org/) [22] to annotate each article with relevant concepts. Wikifier matches the phrases found in the article to link text found in Wikipedia generating candidate annotations. Second, based on both the article and the Wikipedia data, it uses several algorithms and heuristic to disambiguate between multiple options, including the null option (no annotation). Wikifier supports over 130 languages, thus, an English language news article describing a flood, once wikified, will be annotated with the concept “Flood”, corresponding to the Wikipedia article https://en.wikipedia.org/wiki/Flood, even if it uses only the word “Inundation” and not “flood”. A Spanish news article will be annotated with the Wikipedia entry for “Inundación”, corresponding to the Wikipedia article https://es.wikipedia.org/wiki/Inundaci%C3%B3n. However, both those Wikipedia articles are linked together based on Wikipedia interwiki linking and thus the semantic annotation will be the same, i.e., “Flood”. Crucially, it is also important to know that newsworthy people, locations, and organizations, i.e., Named Entities, usually have Wikipedia pages in one or more languages. Thus, Wikification can be considered as performing Named Entity Recognition and Linking although restricted to entities present in Wikipedia.

Each article is automatically assigned a taxonomy category based on its content (utilizing the well-established DMOZ taxonomy, see dmoz-odp.org). We only consider the top 3 levels, which amounts to 5,000 categories. We use a fastText classifier [44] that has as input the wikifier semantic annotations of the article rather than its text. Since these semantic annotations all have an English surface form (“Flood”) even when annotating articles written in different languages, this classifier is effectively cross-lingual.

2.6. Event clusters & text similarity

Journalists tend to agree on which events are newsworthy [45] and thus any event reported by an outlet will also be reported by others. This is also true in the case of “exclusive” stories such as investigative journalism stories [46]. Nevertheless, a precise and widely agreed definition of “event” does not exist [47]. However, given that it is agreed that any event deemed newsworthy will be reported by multiple news outlets, we can define it as any significant happening in the world that was reported in at least a few arti-

Table 1
The number of news and tweets, compared to the number of geolocated events.

Case study	Number of News Events	Number of News Articles	Number of tweets
1 - Flash Flooding of 12 and 13 September, 2019 in Southeastern Spain	244	670	No Data
2 - International floods in Europe in July 2021	2,746	20,972	34,686
3 - International droughts in Europe in 2022	7,052	43,324	No Data

cles from various sources (3–6 articles depending on the language). We automatically group articles into multilingual events using the clustering method described in Ref. [48]; which makes use of semantic annotations provided by Wikifier.

Each Event is assigned a location based on the content of the articles that form it. While some news outlets use a dateline, a brief piece of text at the beginning of the news article that describes where and when the event happened, some do not, and some use the location where the story was written rather than where it happened. To assign a location more accurately to an event, we use a support vector machine (SVM) [49] classifier where each mentioned city is considered a candidate for the event location. For each candidate, both the number times it is mentioned in the article body and the number of times it is mentioned in article datelines are used as features. Results in Ref. [39] showed this classifier to have an accuracy of 98 %.

All concepts (Wikifier semantic annotations) present in any of the articles that form the event will also be assigned to the event. Thus, an event mostly about heavy rains can be brought up when searching using the concept “Flood” if it was present in a single article that forms the event cluster, which can be the last article to be published in time and still assigned to that event cluster. Similarly, the categories of an event are determined by the categories assigned to each article in an event.

In this work when we use text similarity explicitly or implicitly as part of text clustering or search. Note that text search is different from concept-based search. These texts can be, for example, news articles, queries and news articles, or scientific papers and news articles. Unless otherwise specified, text similarity is calculated using the “standard” or “textbook” approach within the field of information retrieval [50]. Namely, we calculate the similarity between texts using the cosine distance between TF-IDF vectors [51,52].

Searchpoint (http://searchpoint.ijs.si) [53] is a search engine that uses BM25 [54,55] for ranking documents according to a query. The documents are then automatically sorted into clusters created using K-Means++ [56,57] and the centroid vectors are used to extract the most representative keywords for each cluster in a 2-dimensional visualization of the search space. The search user can then manually shift the search results towards a particular set of coordinates in this 2D space. This is accomplished by adding the keywords in those coordinates to the original user query and adjusting the BM25 wt accordingly. We use this tool to refine literature review on news articles, tweets, scientific articles, and accepted patents.

2.7. User interface and interactive data visualization using Event Registry

Given a water-related event such as, e.g., a flood of large dimensions, one can distinguish its progress of intensity across time over the news that are published about that event. First the warnings from the national institutions, then the rise of the waters, the cuts in electricity and immediate response, the missions to save the most vulnerable, and (in large events) potential health implications. Surely, we can go deeper in detail about the description of these topics and find common insightful information that could help us in managing such an event.

To search for an event based on news we can use keywords and key phrases, applied in several languages to ensure specific language coverage, though one of the key advantages is to search using concepts, that being based on the Wikipedia terms, already cover the number of languages where the concept is present, recurring to the text classification algorithms that allow to retrieve the news based on text similarity [48]. We can thus identify the main topics in flood events around the world in 2022, and the size of their importance by the number of news articles associated with them: “rain” (34,978), “river” (18,273), “monsoon” (14,045), “climate change” (12,389), “agriculture” (11,164), “precipitation” (9,869), “drought” (9,485), “cyclone” (9,047) and “landslide” (8,874).

From the historical data collected over more than ten years, it is possible to derive an insightful interactive data visualization module. An example of this is the Article Categories, allowing the user to distinguish the weight of other topics in the sample generated by the initial query. For instance, in eleven years (2010–2020) of news about floods in Spain, 7.88 % are related to Climate Change (see Fig. 2). The automated classification of news underlying this capability is based on the implementation of a proprietary algorithm learning from historical news data and classifying it.

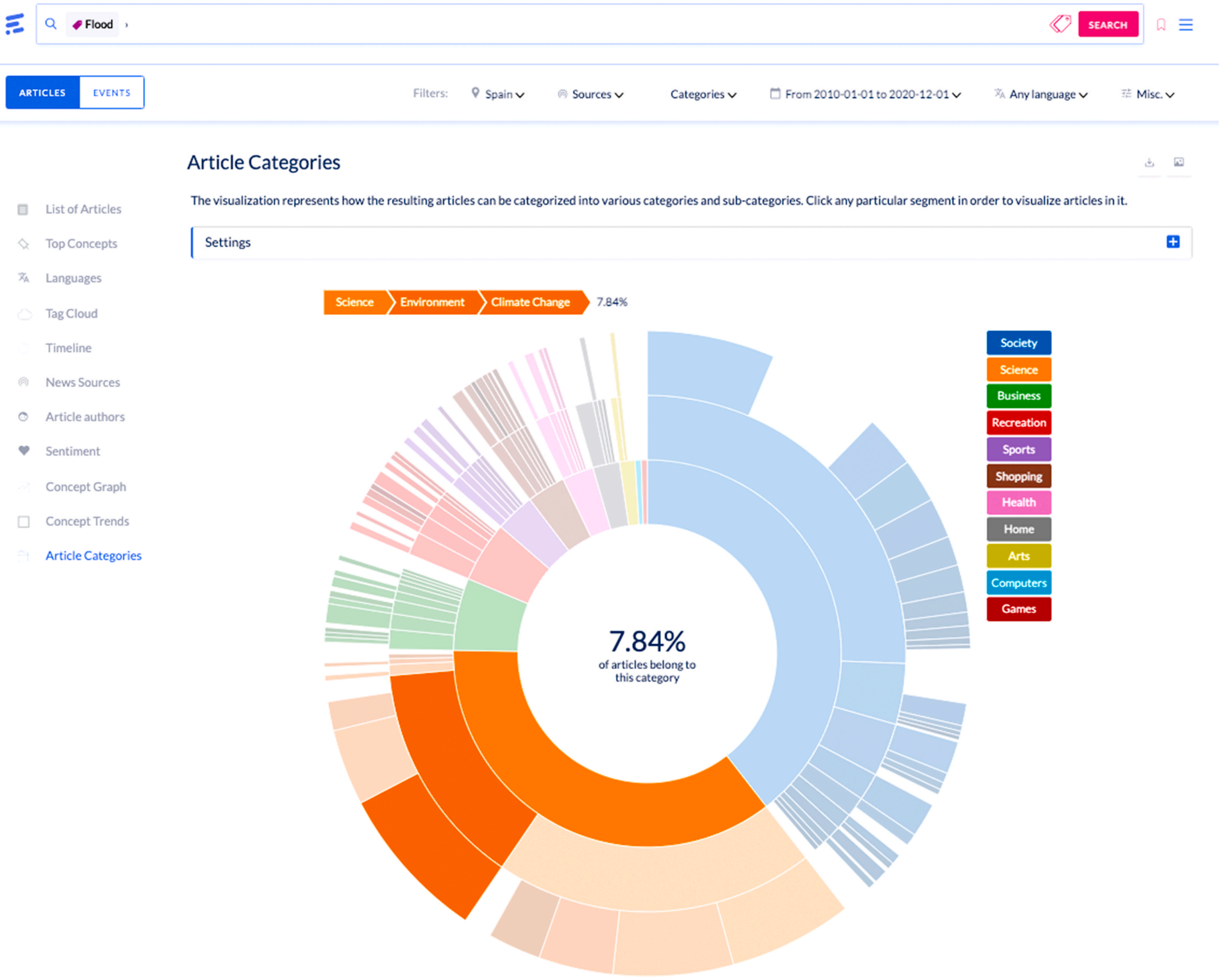
2.8. Water Observatory

To allow domain experts in the water sector to explore the complementarity of the perspectives provided by (i) the worldwide media, in relation to (ii) the published scientific articles, and (iii) the Twitter (X) feed, we built a publicly available web-based system, where some of the interactive data visualization modules can be accessed (available at naiades.ijs.si). There is novelty in this web-based observatory per se, being fed on data of heterogeneous nature, advancing the state of the art in relation to existing systems usually based on local parameters over single data sources [58], GIS technology [59] or yearly reported data and indicators [60].

This knowledge platform, named NAIADES Water Observatory, targeted domain experts, researchers, and water education institutions, offering machine learning technology that allows for text-based data exploration on a wide range of water-related topics (from extreme events like floods and droughts, to their potential consequences as water contamination and water stress level, with a climate change preparedness angle [61]). It was built in the context of the NAIADES project funded by the European Commission, as part of the holistic water ecosystem aiming to improve the digitisation of urban water sector, including with two water management utilities and one municipality feeding the research and development with the needs of the sector and domain knowledge.

To investigate the potential of the combined information extracted from the news and published science, we make available an interactive exploration tool including official indicators obtained from local and national open data portals. An example of this is exposed by the comparison in time of the water usage efficiency in Spain and its exploitation index and stress level (see Fig. 3) to infer a potential reasoning and causality in climate change-related variables, and plan strategies that can help avoid the cases of extreme weather disasters.

The user can also identify using those chosen parameters other countries that follow similar trends and learn from their responses recurring to text similarity applied to the news and scientific articles to investigate their best practices and success stories. The system can be queried on topics in research over a certain time window, and technologies used in the context of a water-related topic or to



Filters: Spain · Sources · Categories · From 2010-01-01 to 2020-12-01 · Any language · Misc.

Article Categories

The visualization represents how the resulting articles can be categorized into various categories and sub-categories. Click any particular segment in order to visualize articles in it.

Science › Environment › Climate Change 7.84%

7.84% of articles belong to this category

Fig. 2. News categorization for floods in Spain (2010–2020) pointing out that in over ten years of collected news data, 8 % of the news articles on floods are associated to climate change.

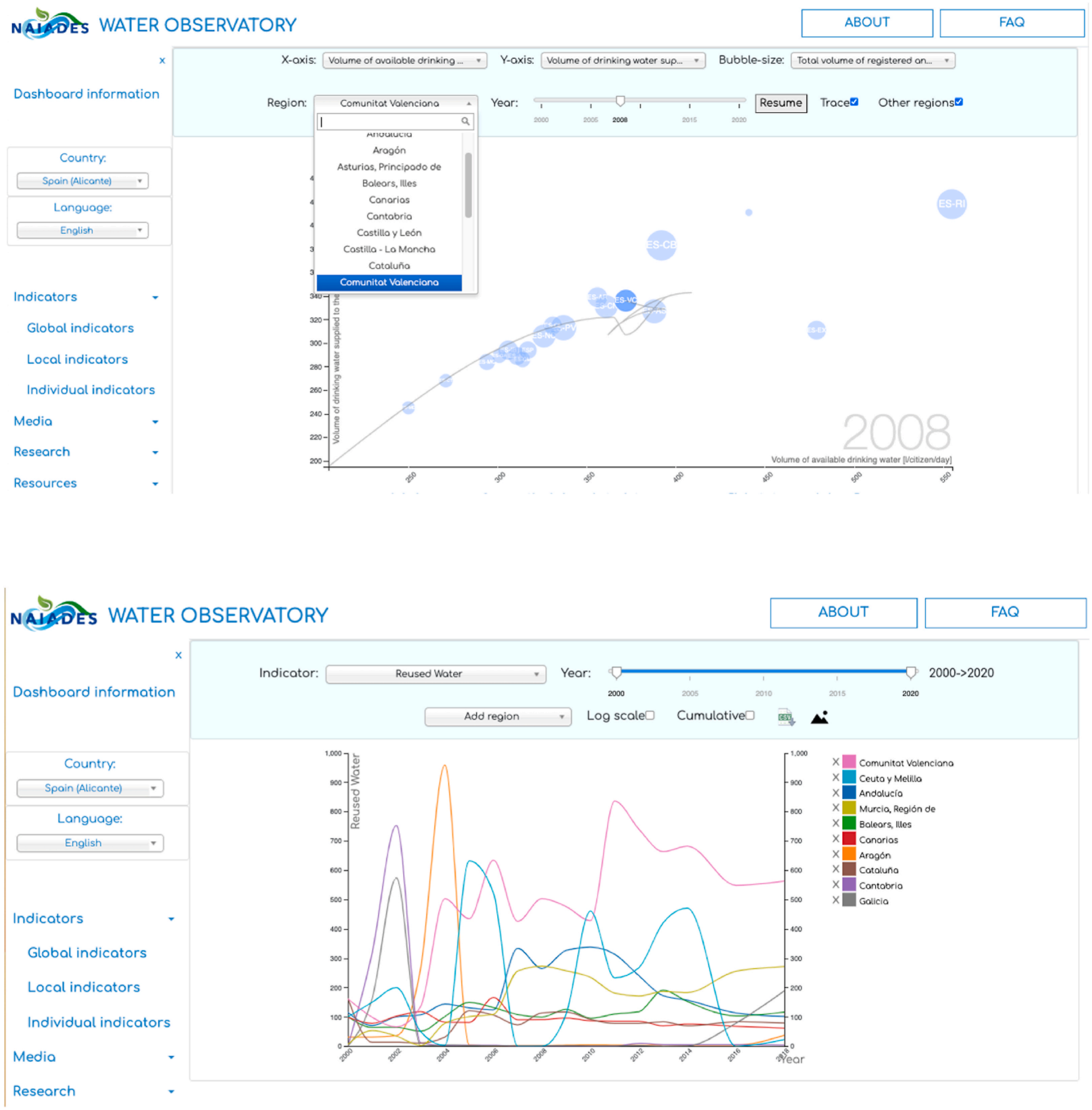
the solution of a specific problem. In Fig. 4 we show such a query aiming to identify flood disaster information published articles retrieved from MAG and, with that, refining a literature survey on floods by dragging the mouse over the coloured clusters and reordering the search results. The first three results were occupying the positions 80, 57 and 114, respectively, but seem to be more relevant in the context of flood disaster information.

3. Results and discussion

In the following section, we will be discussing the achievements of the proposed approach across three case studies established in the introduction section corresponding to specific hydrological events. This analysis was done engaging state-of-the-art machine learning and text mining taking into consideration the specific challenges of hydrological data, and realistic problems in extreme weather event detection and exploration.

3.1. First case study: 2019 flood in the southeast Spain

For the first case study, we analyse the flash flood happening between 12 and 14 September 2019, in the southeast regions of Spain described in Ref. [33]. We use simulated discharge data from the Global Flood Awareness System in the geographical area (i.e., Valencia and Murcia) that was most severely hit by the extreme precipitation event that generated floods, to see the graphical correspondence to the behaviour of the curve representing the journalistic events over time (Fig. 5). Taking into consideration the characteristics of the hydrological event that can be defined as a flash flood event [33], the reporting news (in Spanish language) can follow the dynamic of the event very well (Fig. 5). More specifically, the Pearson correlation coefficient between both datasets (i.e., simulated discharge data and reported news articles) equals to 0.96 (very strong correlation according to Ref. [43] and it is statistically significant with the significance level of 0.05. It should be noted that the news data was aggregated to daily time step since the 24h simulated discharge was used. For the selected event, the news articles can be regarded as a good proxy of the hydrological situation in the investigated area, meaning that published news articles have similar temporal dynamics than discharge data. Additionally, relating to



X-axis: Volume of available drinking ...
Y-axis: Volume of drinking water sup...
Bubble-size: Total volume of registered an...
Region: Comunitat Valenciana
Year: 2008
Trace ☑
Other regions ☑
Country: Spain (Alicante)
Language: English

Indicator: Reused Water
Year: 2000->2020
Log scale □
Cumulative □
Country: Spain (Alicante)
Language: English

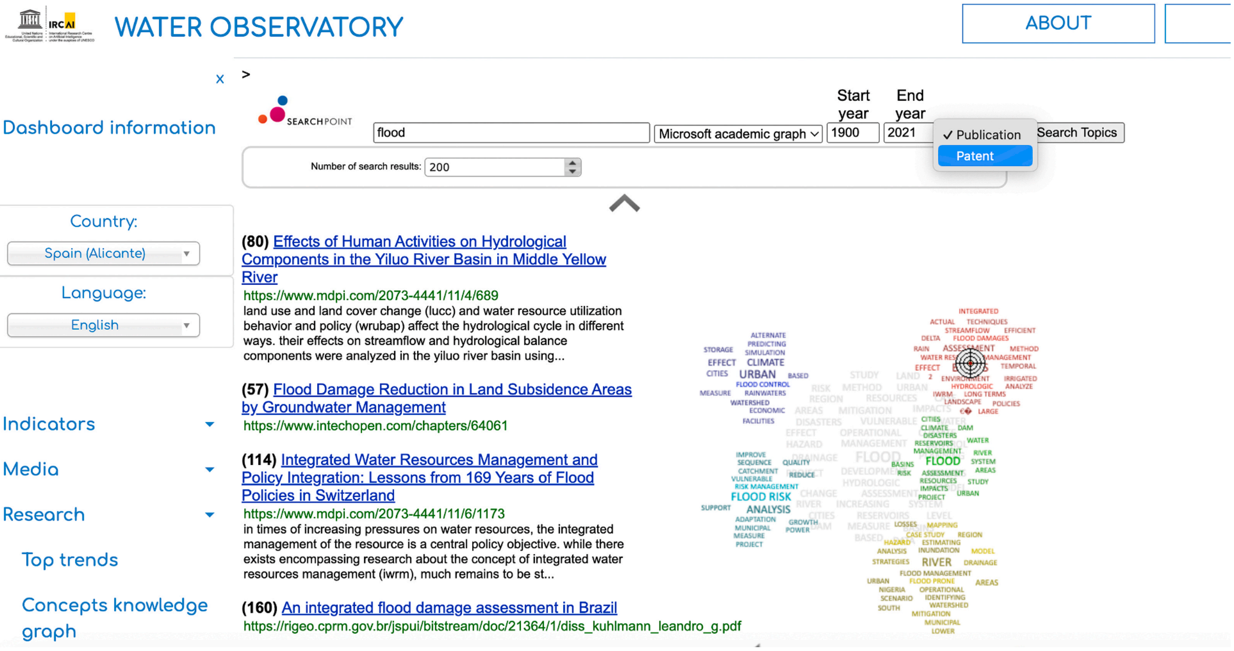
Fig. 3. Tracking the progress of Spain when analysing the efficiency of water usage against its exploitation index and stress level in three dimensions across time (above) and comparing over a specific parameter at local and global levels (below).

the damage caused by the event depicted in Fig. 5 is identified in three main peaks: (i) the first one happening on the 12 of September at 17:55 deriving from initial impact flood event causing two casualties and much (undefined) destruction; followed by (ii) a jump deriving from 300 homeless reported in the same day at 23:43; and finally (iii) the reporting of the event in wider news venues on the 13 of September at 9:20 with an overview of the damaged caused and a final count of three casualties.

Moreover, in the articles composing this event, mostly in Spanish language, the main concepts captured in the reported news are “rain” (100 %), “wind” (95 %), “cold drop” (66 %) later identified as related to the event, and “army” (32 %) called to help the population. This event is specific to this area and other events generated from the news refer to the reporting of damage in other locations of the region, or later generated by articles summarising the flood phenomena during those days.

3.2. Second case study: 2021 European floods with focus on Germany

The second selected case study was an extreme weather event whereby a storm caused heavy rainfall in multiple neighbouring countries for a prolonged period in July 2021 with severe floods across several European countries [34,35]. Compared to the previous



flood | Microsoft academic graph | 1900 | 2021

Number of search results: 200

Country: Spain (Alicante)

Language: English

(80) Effects of Human Activities on Hydrological Components in the Yiluo River Basin in Middle Yellow River
https://www.mdpi.com/2073-4441/11/4/689
land use and land cover change (lucc) and water resource utilization behavior and policy (wrubap) affect the hydrological cycle in different ways. their effects on streamflow and hydrological balance components were analyzed in the yiluo river basin using...

(57) Flood Damage Reduction in Land Subsidence Areas by Groundwater Management
https://www.intechopen.com/chapters/64061

(114) Integrated Water Resources Management and Policy Integration: Lessons from 169 Years of Flood Policies in Switzerland
https://www.mdpi.com/2073-4441/11/6/1173
in times of increasing pressures on water resources, the integrated management of the resource is a central policy objective. while there exists encompassing research about the concept of integrated water resources management (iwrm), much remains to be st...

(160) An integrated flood damage assessment in Brazil
https://rigeo.cprm.gov.br/jspui/bitstream/doc/21364/1/diss_kuhlmann_leandro_g.pdf

Fig. 4. Flood disaster information published articles retrieved from MAG and refining a literature survey on floods.

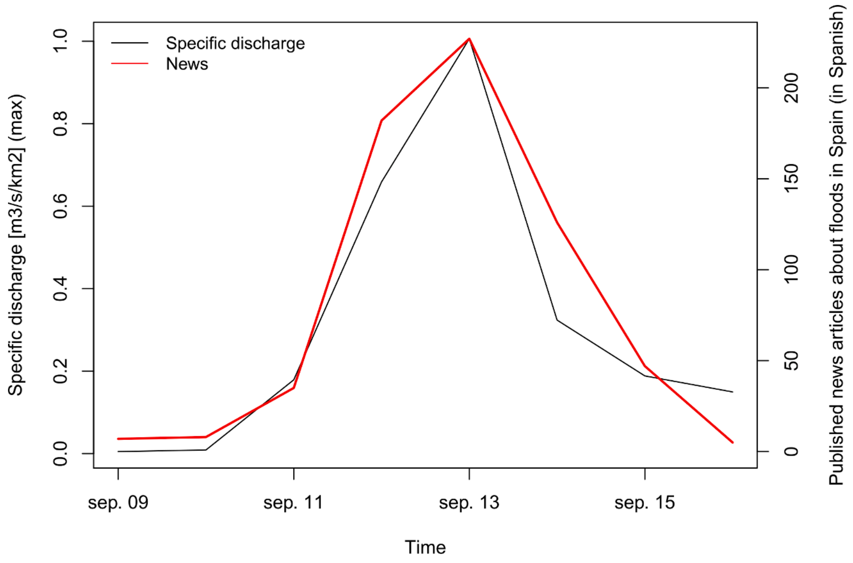


Fig. 5. Maximum specific discharge values for Spain (regions of Valencia and Murcia) during the month of September 2019 in comparison to the published news articles about floods and heavy rain in Spain during the same period while only the news articles published in Spanish language were taken into consideration.

case study (Section 3.1), this one had a much larger geographical spread and human consequences (including 243 dead and more than $11B in estimated damages [62]). If we look at Germany, one of the most affected countries, during July 2021 [34,35], we can notice that the extreme event with the peak discharge values on July 14 is captured in that day with 463 news articles followed by 1,242 news articles in the following day (July 15), until it reaches the mainstream news on July 16 rising to 1,324 articles and decreasing from then on (Fig. 6). In this case study there is lag of 1 day between maximum specific discharge values for Germany and peak in the number of published news (Fig. 6). Although, despite this lag, the Pearson correlation coefficient between the two-time series (i.e., discharge values and published news articles) equals to 0.44 (i.e., moderate correlation according to Ref. [43] and it is statistically significant with the selected significance level of 0.05. Moreover, considering the 1-day time lag, the Pearson correlation coefficient becomes 0.67 (i.e., moderate correlation according to Ref. [43] and is again statistically significant with the significance level of 0.05. In comparison to the first case study (i.e., local flood in Spain) the agreement between temporal dynamics of published news and discharge values is worse and there is a clear 1-day lag between the peak discharge and peak in published news. Hence, even in local events (i.e., case study 1, flood in Spain) local news can capture the dynamics of the flood event, even better than for large-scale

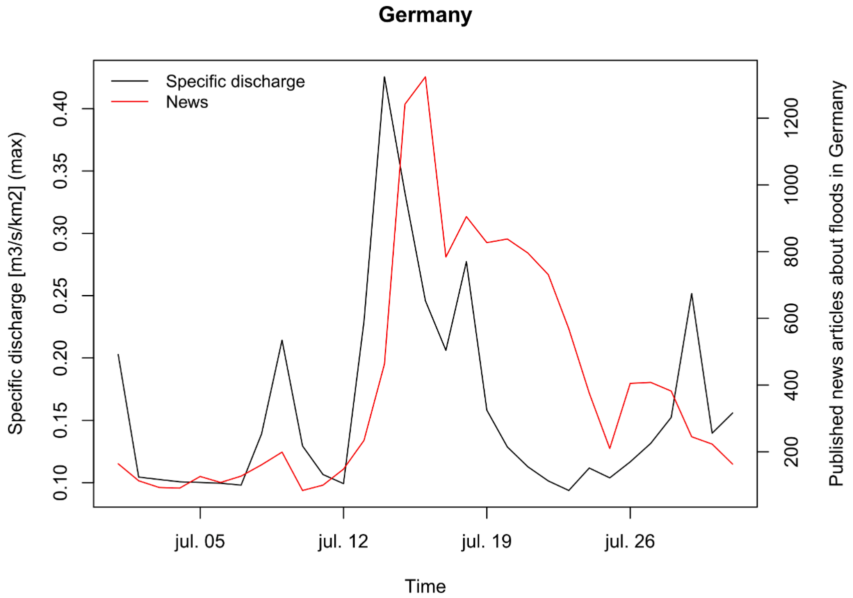


Fig. 6. Maximum specific discharge for Germany during the month of July 2021 in comparison to the published news articles about floods and heavy rain in Germany during the same period.

events. Additionally, in this case study, a specific journalistic event reporting flood in Bavaria initiating at 14:30 of July 13 [63] rises intensity in that day until a maximum peak at 16:57 in that day. Another example of a journalistic event retracting the damage caused by the flood phenomenon in Hagen, Germany, is initially reported at 10AM in July 14 [64] reaching a maximum peak at 21:33 that day, where we can quickly access some of the impact from the related concepts extracted from the text including “retirement home”, “childcare”, “road transport” or “electric current”, and relate these concepts with locations extracted from the text of the collected news articles on this event (Fig. 7).

This geolocation of news is extracted as a location mentioned in the text (it can be multiple when the news article describes an event extending geographically), and if not mentioned, it is assumed to be the location of the news venue. This allows us to relate the estimated impact in relation to the extracted locations (with city and sometimes neighbourhood granularity) as discussed above. Given the extended multinational geographical localization of the event across time, it can be difficult to capture the data to produce a complete analysis, and that might be another opportunity to leverage the signal from the news media. Fig. 8 shows the monthly precipitation (July 2021) sum for central Europe according to the ERA5-Land reanalysis product and the geolocation extracted from the news providing the multinational extension of the event throughout central Europe with intensity levels expressed by the areas that are mentioned more in relation to the tracked journalistic event. The maximum number of news were related to the location close to the border between Germany, Luxemburg, and Belgium, where there was also a high accumulated precipitation amount in July 2021 (Fig. 8). Specifically, more than 200 mm of rainfall fall in this region in this period, which is around 1/6 of the total average annual precipitation in this area. Additionally, some news can also be related to the Alpine area (e.g., Switzerland) that also received quite significant rainfall amounts in July 2021 (Fig. 8). Furthermore, it can also be seen that for areas where there were no flooding issues during that time (e.g., western part of France) there is almost no news reported in this area and the total precipitation in this month was lower compared to other areas (Fig. 8).

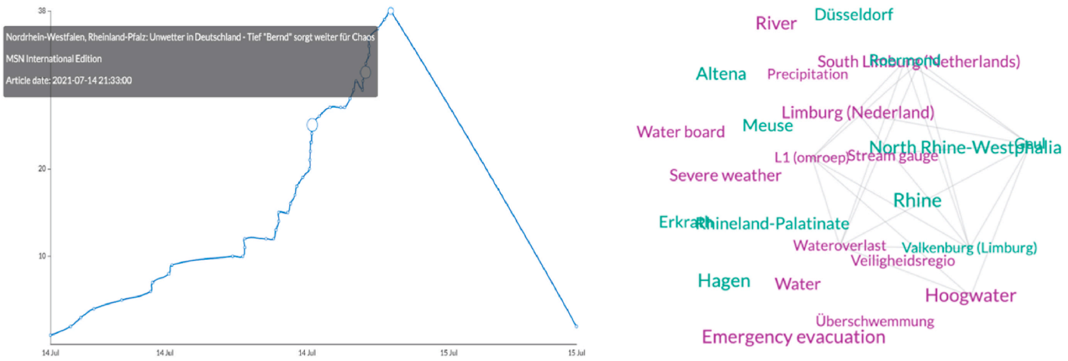

Nordrhein-Westfalen, Rheinland-Pfalz: Unwetter in Deutschland - Tief "Bernd" sorgt weiter für Chaos
MSN International Edition
Article date: 2021-07-14 21:33:00

Fig. 7. The evolution of events of the flood in Hagen, Germany (on the left) and the mentioned name entities extracted from the news on the impact of the event (on the right).

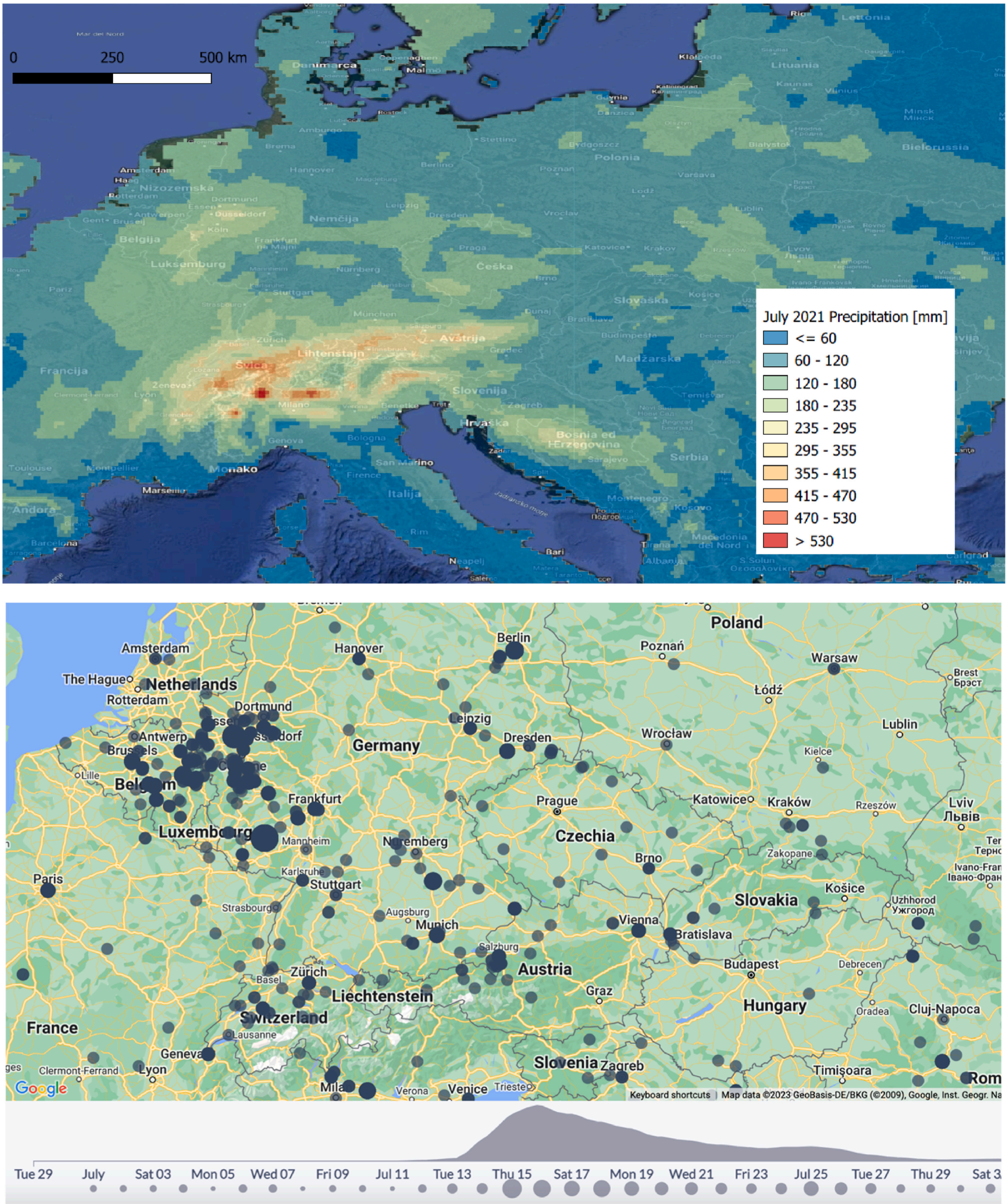


Fig. 8. The monthly precipitation sums for July 2021 for central Europe according to the ERA5-Land reanalysis product (above) and the geolocation from the news providing the multinational extension of the event throughout central Europe (below).

3.3. *Third case study: 2022 European drought*

In the third and final case study, we focus on a multinational event with similar range, long duration, and high return period, but this time over a drought event. We have investigated the drought in Europe during the year of 2022, starting in April 2022 in several European countries such as France, Spain, and Portugal, that prolonged to June, July and extended to the United Kingdom, Germany, Scandinavia, and again into August [65]. During the summer of 2022, drought in combination with heat waves' alerts were issued across the continental Europe, the British Isles, north Africa, and Turkey [65]. The intensity of the drought led to water restrictions in several European countries such as France, Germany, Italy, Romania, Spain, Portugal, and UK. It also affected hydropower generation, nuclear power plant cooling, agriculture, and river shipping across the continent [65]. It should be noted significant precipitation deficit was already observed in the first part of the year 2022 and that severe drought conditions could also be detected in April 2022 [65]. These conditions can be well observed from the Soil Moisture Anomaly Index. Furthermore, the conditions got worse during May and June with some early heat waves [65] and were the most extreme in July 2022 when the combination of drought and heat wave impacted several water-related sectors. The event can be measured on one side from the measured soil moisture anomaly or SPI indexes and on the other from the perspective of news (Fig. 9). For France there is a good agreement between the lowest SPI-1 value and the peak in the number of published news in July. A similar behaviour can be seen for Slovenia (Fig. 9). In both cases (i.e., Slovenia and France) the Pearson correlation coefficient between two datasets is around −0.28 and not statistically significant with the selected significance level of 0.05. On the other hand, the situation in Portugal was obviously more complex and there is no clear temporal agreement between the peak in the published news and minimum value of the SPI-1 (Fig. 9). Hence, it should be noted that SPI-1 can be regarded as an indicator for immediate impact such as reduced snowpack, soil moisture or flow [41]. In the case of Portugal, the Pearson correlation coefficient is −0.06 and not statistically significant with the selected significance level of 0.05. Therefore, drought conditions are harder to detect from the news compared to the case of floods, earlier discussed. Moreover, one could argue that what is reported in the news as drought it is in most cases a combination of drought and heat wave while in this section we compared news signals with SPI-1 and SMAI that are both not taking into consideration heat wave explicitly. Therefore, we argue that better agreement between published news and the actual situation could be obtained considering also some heat wave or extreme heat index [66]. This can be nicely seen from Fig. 10 that shows the situation in Europe where there was a clear peak in the number of events in July 2022 while, as discussed above, hydrological drought was already quite significant in April and May 2022 [65], and this cannot be detected from published news and events.

3.4. Opportunities and limitations in the usage of the signal from media

3.4.1. Main strengths and limitations of the study

The novelty of the research presented in this paper unveils the potential role of news media in the early assessment to the impact of extreme events such as floods and droughts. Considering the related work discussed in the Introduction section of this paper, there seems to be a knowledge gap in regarding he usage of news media, although social media being well exploited. The main results in

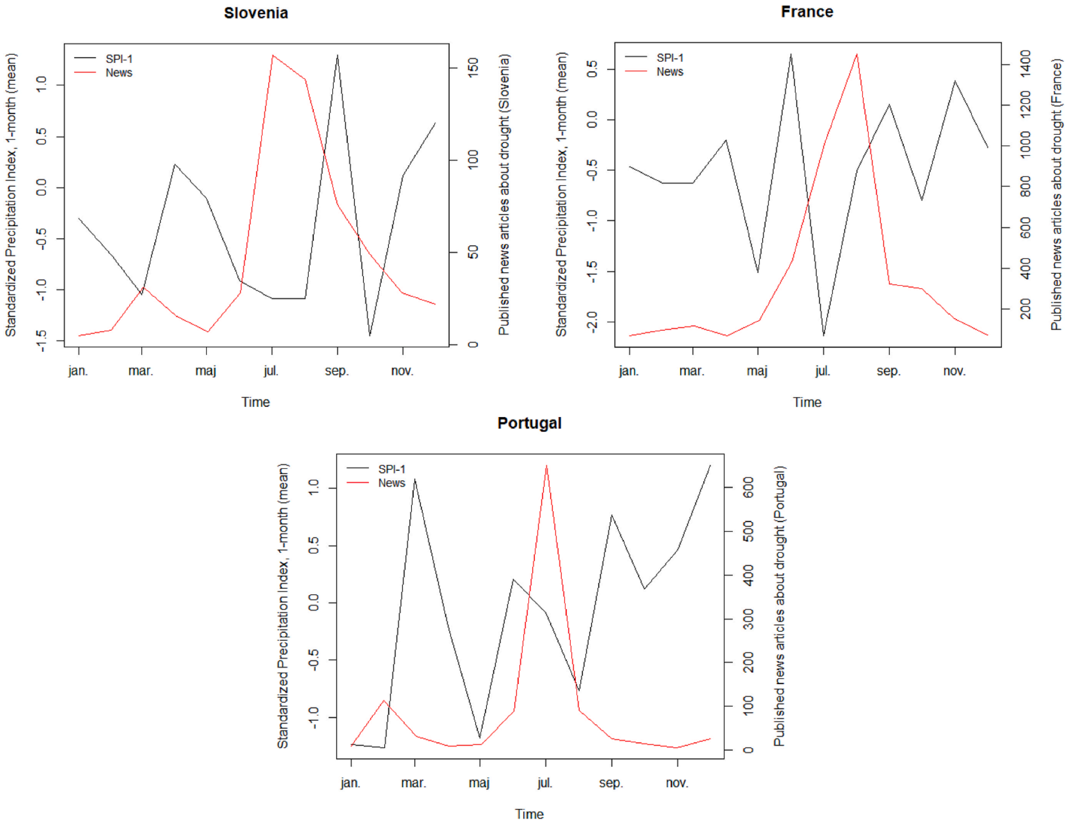


Fig. 9. Comparison between the hydrological signal based on the standardized precipitation index (SPI-1) and the media signal for drought in Slovenia (above), France (in the middle) and Portugal (below) across 2022. It should be noted that news was aggregated to monthly time step.

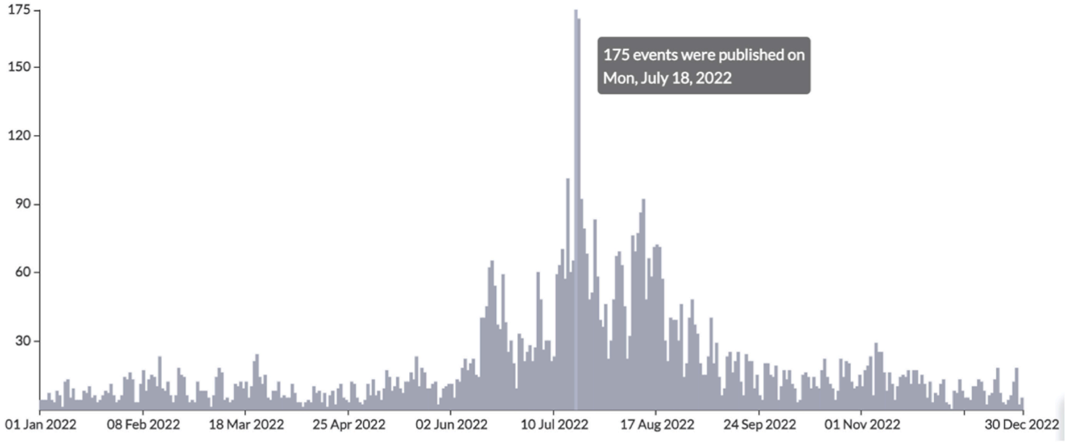


Fig. 10. The European heat wave reported by the drought news events during 2022 captured through the amount of detected news articles.

this study show the potential of the used approach in offering an informed overview of an extreme weather event that can follow its occurrence in an almost real time, given that the time of publication of such impactful events tends to be short and their coverage frequent. The main strengths of this study are thus: (i) the facilitated identification of extreme events in almost real-time news data (that can be automatized) and historical data based on text similarity approaches; (ii) the fast access to the details of the extreme weather event, the engaged entities and the scope of the impact in the population from what is reported in the news; and (iii) the transborder scope of the assessment, which is often more limited to a national scale considering the access to data.

On the other hand the main limitations of this study are: (i) tied to the accuracy of what is understood and reported by the journalists and the data they can access to build the news articles; (ii) the granularity of the location of the event that is also derived from the text reported in the news or the location from where it is being reported; and (iii) the difficulty of establishing a straightforward relation between the information retrieved from the news and typically used water-related extreme event parameters, e.g., the standardized precipitation index due to the different nature of both measures.

In the following subsection, we discussed the opportunities deriving from the results in this study with respect to: (i) the improved social media signal from the information retrieved with the news media signal, potentially automating the identification of specific characteristics of the event and entities involved in it; and (ii) the exploration of worldwide best practices in, e.g., the management of similar events identified thorough text similarity approaches across more than 60 languages.

3.4.2. Improving the social media signal on hydrological events

The main benefit of the usage of social media data is the real-time polling on weather-related events just before they happen, a procedure known as nowcasting [67], allowing alerts to be triggered by the real-time ingestion of Twitter data often labelled with geolocation. Although there is a lack of Twitter data regarding urban water events (e.g., water loss and leakage), a substantial amount of Twitter data can be collected for hydrological events of bigger dimension (e.g., a large flood consequence of heavy rain). With a good enough Twitter signal on extreme weather events (i.e., with low noise capturing mostly posts directly related to the event), we can further analyse their impact in the context of causality [68] combining the information obtained from: (i) the monitoring and exploration of news articles and social media feeds; (ii) the analysis of a combinations of indicators through time and what stories can they tell.

We used concepts characterising the specific event detected from the information in the news to extract the related tweets based on timestamp and filtered search by hashtags and keywords, improving the often-noisy signal of social media allowing for more efficient nowcasting, and cleared causality relations. The news media and social media capture the same trends within the location and time specifications of the case study 2, as expressed in Fig. 11, presenting the distribution of news vs tweets with similar curve behaviour and peaks.

3.4.3. Deriving best practices for climate change preparedness from the advanced queries based on text similarity

To leverage the capabilities of text mining technologies and methodologies in information extraction from text, we add to the layer of extracted information from news and social media (providing a journalistic or public perception perspective on events and their outcomes) a trustworthy scientific perspective provided by the content of published science ingested in the system and sourced at the Microsoft Academic Graph (MAG). From this combined knowledge, we can elaborate on the characterisation of events and their sequence of happenings, using text similarity to identify best practices from similar events. It is particularly important to automate this information retrieval when, in a climate change context, we can be observing an increasing number of similar events across the globe, leveraging from the multilingual capabilities of the news engine and technology to overcome the language and national level limitations often present in traditional methods.

This capability is enhanced by the topics and relations identified in the scientific research, through text mining algorithms enabling technological capabilities in the identification of major research concepts in the textual documents, highlighting new concepts that were not yet present, and complementing the perspective extracted from historical news data. In Fig. 12 we can see how the data

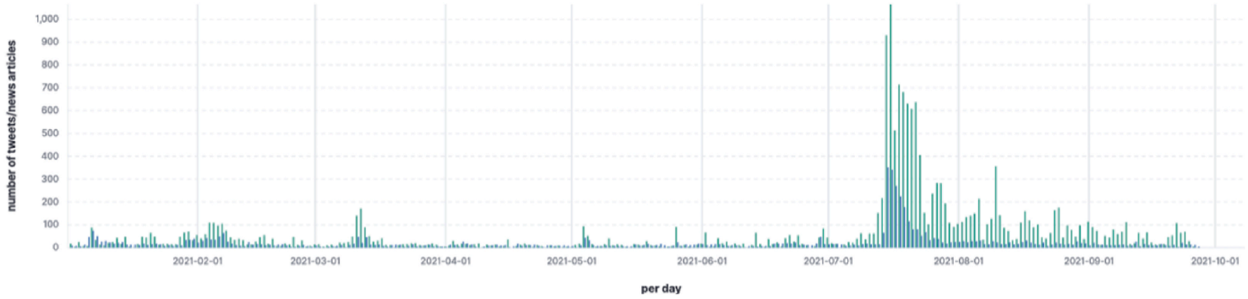


Fig. 11. The signal of news (in green) and tweets (in blue) aligned over time, showing the correspondence of coverage under the example disaster type: "flood" and "storm" in Germany throughout 2021, under the specifications of the case study 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

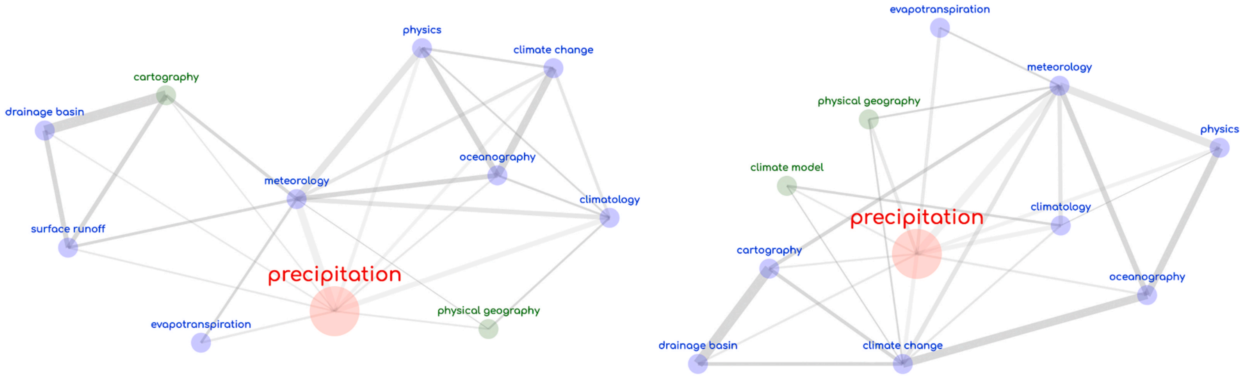


Fig. 12. The water-related research topics in relation to the research on precipitation in 2009 (on the left) and 2019 (on the right). The items marked in green is identified as new research trends. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

visualization generated from the research topics identified (present as MAG categories) and their relation in time (only the main relations are represented) can expose reasoning in a research perspective on climate change in a certain time frame, and can propose trends that must be further analysed and compared to understand the dynamics of climate change research and of its related water topics and enrich the query to news with new research trends (e.g., in the domain of "precipitation", directly related to floods, "climate change" has been a significant topic prior to 2000 while "cartography" is identified as a relevant related topic in 2009 being still within the main related research trends today). This summarized information can hint to topics that need to be added updating the filtering generating the data collection on news articles and tweets to engage a climate change perspective on the ongoing observation of water topics and extreme events. Examples of this are the climate change-related consequences of floods and droughts worldwide, both in their increasing frequency and in the impacts of their return (considering that climate change is bringing new phenomena and behaviours that differ to what traditional approaches can expect).

4. Conclusions and further steps

The research work reported in this paper explores the potential and limitations of the application of machine learning on text-based documents sourced in media and social media as a signal for extreme hydrological events such as floods and droughts, complementary to the traditional hydrologic approaches. The exploration of historical data with significant coverage in time and demographics allows for insights on the magnitude of the events, but also feeds new concepts automatically identified from the text that can be valuable to update search criteria in the context of climate change where recurring new patterns and behaviours might affect the efficiency of traditional hydrologic methods. Based on the results presented in this paper it can be concluded that news can capture the dynamics of the extreme hazards such as floods well, even in the case of an event with smaller spatial extent. Additionally, in some cases a lag between the actual situation and number of published news could be observed, which can be attributed to the fact that journalists need some time, e.g., to travel to the location of the event and prepare a news article. However, we argue that these kinds of time lags mostly do not exceed 1-day. On the other hand, the situation with capturing the drought dynamics is more complex and news articles are not able to capture the drought dynamics. Moreover, we argue that the combination of drought and extreme heat is what can be better captured in news since the extreme heat has similarly as flood have a more direct impact on the society and can be more easily detected by journalists. While on the other hand, droughts (e.g., hydrological, or even agricultural) are visible in the society. For example, in some European countries there were extreme drought conditions (e.g., low river flows) already in April 2022 but this phenomenon could not be detected in published news. As indicated, in the information extracted from the published news, we were able to detect combined drought and heat wave extremes occurring in July 2022 across Europe.

To proceed with the presented research in this paper, the authors envision a few areas of improvements. Firstly, improving the media and social media mining methods that could be more appropriate for other extreme weather events not covered by the scenarios in this paper. Secondly, testing of this text mining approach for various categories of extreme weather events, considering complexity and different temporal and spatial resolution of sub-categories of floods (i.e., debris floods, flash floods, pluvial floods, and coastal floods) or also to cover rainfall-induced mass movement events (e.g., debris flows, shallow landslides, and deep-seated landslides). Thirdly, improving complementarity in the triangle news-media-science towards hydrological science. The procedure presented in this paper is aligned with the Data Information Knowledge Wisdom Structure with a decision-making process at the top of this pyramid. It is essential for hydrology experts to extract best-knowledge based on all available data and information held in them from a variety of sources, to support decision makers in their responsible governance of weather or water-related extreme events. Knowledge development by including information conveyed by news is an added value in this regard.

Funding

J.P.C., B.M. and I.N. thank the support of the European Commission on the H2020 NAIADES project (GA nr. 820985). B.N. and M.M. would like to acknowledge the financial support of Slovenian Research and Innovation Agency by grants P2-0180 and J6-4628, and of the University of Ljubljana by Development Fund for UNESCO Chair on Water-related Disaster Risk Reduction. European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 820985).

Ethics statements

Not applicable.

Credit author statement

J.P.C.: Conceptualization, Methodology, Data Curation, Writing - Original Draft, Project administration, Funding acquisition. L.R.: Writing - review & editing, Data Curation. I.N. and B.M.: Writing - review & editing, Data Curation. N.B., M.M.: Conceptualization, Methodology, Data Curation, Writing - review & editing. M.B.M.: Writing - review & editing. I.N.: Writing - review & editing. G.L.: Data Curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

J.P.C., B.M. and I.N. thank the Twitter research contract that allowed the data collection of tweets on water events-related topics, from November 2019 to October 2021. Critical and useful comments from two anonymous reviewers are acknowledged by the authors.

References

[1] M.B. Massri, S. Brezec, E. Novak, K. Kenda, Semantic enrichment and analysis of legal domain documents, Ljubljana, in: IKDD 2019: Slovenian KDD Conference on Data Mining and Data Warehouses, 2019, pp. 1–4.

[2] J. Pita Costa, L. Rei, L. Stopar, F. Fuart, M. Grobelnik, D. Mladenić, I. Novalija, A. Staines, J. Pääkkönen, J. Konttila, J. Bidaurrazaga, O. Belar, C. Henderson, G. Epelde, M.A. Gabaráin, P. Carlin, J. Wallace, NewsMeSH: a new classifier designed to annotate health news with MeSH headings, Artif. Intell. Med. 114 (2021) 102053, https://doi.org/10.1016/j.artmed.2021.102053.

[3] J. Pita Costa, M. Grobelnik, F. Fuart, L. Stopar, G. Epelde, S. Fischaber, P. Poliwoda, D. Rankin, J. Wallace, M. Black, R. Connolly, P. Davis, Meaningful Big data integration for a global COVID-19 strategy, IEEE Comput. Intell. Mag. 15 (2020) 51–61, https://doi.org/10.1109/MCI.2020.3019898.

[4] N. Alnajran, K. Crockett, D. McLean, A. Latham, Cluster analysis of twitter data: a review of algorithms, in: ICAART 2017 - Proceedings of the 9th International Conference on Agents and Artificial Intelligence, 2017, pp. 239–249, https://doi.org/10.5220/0006202802390249.

[5] C. Restrepo-Estrada, S.C. de Andrade, N. Abe, M.C. Fava, E.M. Mendiondo, J.P. de Albuquerque, Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring, Comput. Geosci. 111 (2018) 148–158, https://doi.org/10.1016/j.cageo.2017.10.010.

[6] P. Chaudhary, S. D'Aronco, J.P. Leitão, K. Schindler, J.D. Wegner, Water level prediction from social media images with a multi-task ranking approach, ISPRS J. Photogrammetry Remote Sens. 167 (2020) 252–262, https://doi.org/10.1016/j.isprsjprs.2020.07.003.

[7] D. Eilander, P. Trambauer, J. Wagemaker, A. van Loenen, Harvesting social media for generation of near real-time flood maps, Procedia Eng. 154 (2016) 176–183, https://doi.org/10.1016/j.proeng.2016.07.441.

[8] W. Wu, J. Li, Z. He, X. Ye, J. Zhang, X. Cao, H. Qu, Tracking spatio-temporal variation of geo-tagged topics with social media in China: a case study of 2016 hefei rainstorm, Int. J. Disaster Risk Reduc. 50 (2020) 101737, https://doi.org/10.1016/j.ijdrr.2020.101737.

[9] K. Shoyama, Q. Cui, M. Hanashima, H. Sano, Y. Usuda, Emergency flood detection using multiple information sources: integrated analysis of natural hazard monitoring and social media data, Sci. Total Environ. 767 (2021) 144371, https://doi.org/10.1016/j.scitotenv.2020.144371.

[10] M. Mazzoleni, M. Verlaan, L. Alfonso, M. Monego, D. Norbiato, M. Ferri, D.P. Solomatine, Can assimilation of crowdsourced data in hydrological modelling improve flood prediction? Hydrol. Earth Syst. Sci. 21 (2017) 839–861, https://doi.org/10.5194/hess-21-839-2017.

[11] N. Said, K. Ahmad, M. Riegler, K. Pogorelov, L. Hassan, N. Ahmad, N. Conci, Natural disasters detection in social media and satellite imagery: a survey, Multimed. Tool. Appl. 78 (2019) 31267–31302, https://doi.org/10.1007/s11042-019-07942-1.

[12] E. Daniel, L. Adriana, T.A.M. Tavares, Using social media for economic disaster evaluation: a systematic literature review and real case application, Nat. Hazards

Rev. 23 (2022) 5021020, https://doi.org/10.1061/(ASCE)NH.1527-6996.0000539.

[13] Q. Khan, E. Kalbus, N. Zaki, M.M. Mohamed, Utilization of social media in floods assessment using data mining techniques, PLoS One 17 (2022) 1–18, https://doi.org/10.1371/journal.pone.0267079.

[14] Re Mariano, Leandro D. Kazimierski, Pablo E. Garcia, Nicolás E. Ortiz & Marina Lagos, Assessment of crowdsourced social media data and numerical modelling as complementary tools for urban flood mitigation, Hydrological Sciences Journal 67 (9) (2022) 1295–1308, https://doi.org/10.1080/02626667.2022.2075266.

[15] C. Zhang, C. Fan, W. Yao, X. Hu, A. Mostafavi, Social media for intelligent public information and warning in disasters: an interdisciplinary review, Int. J. Inf. Manag. 49 (2019) 190–207, https://doi.org/10.1016/j.ijinfomgt.2019.04.004.

[16] L. Lopez-Fuentes, A. Farasin, M. Zaffaroni, H. Skinnemoen, P. Garza, Deep learning models for road passability detection during flood events using social media data, Appl. Sci. 10 (2020), https://doi.org/10.3390/app10248783.

[17] N. Kankanamge, T. Yigitcanlar, A. Goonetilleke, M. Kamruzzaman, Determining disaster severity through social media analysis: testing the methodology with South East Queensland Flood tweets, Int. J. Disaster Risk Reduc. 42 (2020) 101360, https://doi.org/10.1016/j.ijdrr.2019.101360.

[18] S. Fu, H. Lyu, Z. Wang, X. Hao, C. Zhang, Extracting historical flood locations from news media data by the named entity recognition (NER) model to assess urban flood susceptibility, J. Hydrol. 612 (2022) 128312, https://doi.org/10.1016/j.jhydrol.2022.128312.

[19] J. Sodoge, C. Kuhlicke, M.M. de Brito, Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning, Weather Clim. Extrem. 41 (2023) 100574, https://doi.org/10.1016/j.wace.2023.100574.

[20] B. Bischke, P. Helber, S. Brugman, E. Basar, Z. Zhao, M.A. Larson, K. Pogorelov, The Multimedia satellite task at MediaEval 2019, in: M.A. Larson, S.A. Hicks, M.G. Constantin, B. Bischke, A. Porter, P. Zhao, M. Lux, L.C. Quiros, J. Calandre, G. Jones (Eds.), Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019, {CEUR} Workshop Proceedings, CEUR-WS.org, 2019.

[21] J. Rupnik, A. Muhič, P. Skraba, Multilingual document retrieval through hub languages, in: SiKDD, vol. 2012, 2012, pp. 1–4 Ljubljana.

[22] J. Brank, G. Leban, M. Grobelnik, Annotating documents with relevant Wikipedia concepts, in: Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017, 2017, pp. 1–4.

[23] M. Mikoš, N. Bezak, J.P. Costa, M.B. Massri, I. Novalija, M. Jermol, M. Grobelnik, Natural-hazard-related web observatory as a sustainable development tool, Issue 1, 2022, in: K. Sassa, K. Konagai, B. Tiwari, Ž. Arbanas, S. Sassa (Eds.), Progress in Landslide Research and Technology, vol. 1, Springer International Publishing, Cham, 2023, pp. 83–97, https://doi.org/10.1007/978-3-031-16898-7_5.

[24] T. Hartmann, L. Slavíková, S. McCarthy, Nature-Based Flood Risk Management on Private Land: Disciplinary Perspectives on a Multidisciplinary Challenge, Springer, 2019, https://doi.org/10.1007/978-3-030-23842-1.

[25] D.R. Maidment, Handbook of Hydrology, McGraw-Hill, New York etc, 1993.

[26] A.K. Mishra, V.P. Singh, A review of drought concepts, J. Hydrol. 391 (2010) 202–216, https://doi.org/10.1016/j.jhydrol.2010.07.012.

[27] A. Zargar, R. Sadiq, B. Naser, F.I. Khan, A review of drought indices, Environ. Rev. 19 (2011) 333–349, https://doi.org/10.1139/a11-013.

[28] J. Galtung, M.H. Ruge, The structure of foreign news: the presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers, J. Peace Res. 2 (1965) 64–90, https://doi.org/10.1177/002234336500200104.

[29] M. Bednarek, H. Caple, Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond, Discourse Soc. 25 (2014) 135–158, https://doi.org/10.1177/0957926513516041.

[30] M. Boukes, N.P. Jones, R. Vliegenthart, Newsworthiness and story prominence: how the presence of news factors relates to upfront position and length of news stories, Journalism 23 (2022) 98–116, https://doi.org/10.1177/1464884919899313.

[31] E. Belyaeva, A. Košmerlj, D. Mladenić, G. Leban, Automatic estimation of news values reflecting importance and closeness of news events, Inform 42 (2018) 527–533, https://doi.org/10.31449/inf.v42i4.1132.

[32] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft academic graph: when experts are not enough, Quant. Sci. Stud. 1 (2020) 396–413, https://doi.org/10.1162/qss_a_00021.

[33] F.J. Tapiador, C. Marcos, J.M. Sancho, C. Santos, J.Á. Núñez, A. Navarro, C. Kummerow, R.F. Adler, The September 2019 floods in Spain: an example of the utility of satellite data for the analysis of extreme hydrometeorological events, Atmos. Res. 257 (2021), https://doi.org/10.1016/j.atmosres.2021.105588.

[34] M.I. Brunner, Floods and droughts: a multivariate perspective on hazard estimation, Hydrol. Earth Syst. Sci. Discuss. 2023 (2023) 1–26, https://doi.org/10.5194/hess-2023-20.

[35] C.C. Ibebuchi, Patterns of atmospheric circulation in Western Europe linked to heavy rainfall in Germany: preliminary analysis into the 2021 heavy rainfall episode, Theor. Appl. Climatol. 148 (2022) 269–283, https://doi.org/10.1007/s00704-022-03945-5.

[36] D. Bonaldo, D. Bellafiore, C. Ferrarin, R. Ferretti, A. Ricchi, L. Sangelantoni, M.L. Vitelletti, The summer 2022 drought: a taste of future climate for the Po valley (Italy)? Reg. Environ. Change 23 (2023), https://doi.org/10.1007/s10113-022-02004-z.

[37] K.R. van Daalen, M. Romanello, J. Rocklöv, J.C. Semenza, C. Tonne, A. Markandya, N. Dasandi, S. Jankin, H. Achebak, J. Ballester, M. Nilsson, R. Lowe, The 2022 Europe report of the Lancet Countdown on health and climate change: towards a climate resilient future, Lancet Public Health 7 (2022) e942–e965, https://doi.org/10.1016/S2468-2667(22)00197-9.

[38] M. Trampuš, B. Novak, The internals of an aggregated web news feed, in: Proceedings of 15th Multiconference on Information Society, 2012, pp. 1–4.

[39] G. Leban, B. Fortuna, J. Brank, M. Grobelnik, Event registry - learning about world events from news, in: WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web, 2014, pp. 107–110, https://doi.org/10.1145/2567948.2577024.

[40] C. Cammalleri, J.V. Vogt, B. Bisselink, A. de Roo, Comparing soil moisture anomalies from multiple independent sources over different regions across the globe, Hydrol. Earth Syst. Sci. 21 (2017) 6329–6343, https://doi.org/10.5194/hess-21-6329-2017.

[41] Copernicus, Global drought observatory [WWW Document]. URL. https://edo.jrc.ec.europa.eu/gdo/php/index.php?id = 2112, 2023.

[42] CDS Climate, River discharge and related historical data from the global flood awareness system [WWW Document]. URL. https://cds.climate.copernicus.eu/cdsapp#!/dataset/cems-glofas-historical?tab = form, 2023.

[43] P. Schober, L.A. Schwarte, Correlation coefficients: appropriate use and interpretation, Anesth. Analg. 126 (2018) 1763–1768, https://doi.org/10.1213/ANE.0000000000002864.

[44] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, 2017, pp. 427–431, https://doi.org/10.18653/v1/e17-2068.

[45] T. Harcup, D. O'Neill, What is News?: news values revisited (again), Journal. Stud. 18 (2017) 1470–1488, https://doi.org/10.1080/1461670X.2016.1150193.

[46] H. de Burgh, P. Bradshaw, M. Bromley, M.D. Arcy, I. Gaber, R. Greenslade, M. Hanna, C. Horrie, P. Lashmar, G. Macfadyen, INVESTIGATIVE JOURNALISM, Investigative Journalism, second ed., 2008, https://doi.org/10.4324/9780203895672.

[47] D. Trilling, M. van Hoof, Between article and topic: news events as level of analysis and their computational identification, Digit. Journal 8 (2020) 1317–1337, https://doi.org/10.1080/21670811.2020.1839352.

[48] G. Leban, B. Fortuna, M. Grobelnik, Using news articles for real-time cross-lingual event detection and filtering, in: CEUR Workshop Proceedings, 2016, pp. 33–38.

[49] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1023/A:1022627411411.

[50] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008, https://doi.org/10.1017/CBO9780511809071.

[51] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21, https://doi.org/10.1108/eb026526.

[52] G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (1975) 613–620, https://doi.org/10.1145/361219.361220.

[53] L. Stopar, B. Fortuna, M. Grobelnik, Newssearch: search and dynamic re-ranking over news corpora, in: Conf. Proceedings of SiKDD, vol. 2012, 2012, pp. 1–4.

[54] S.E. Robertson, K.S. Jones, Relevance weighting of search terms, J. Am. Soc. Inf. Sci. 27 (1976) 129–146, https://doi.org/10.1002/asi.4630270302.

[55] K. Sparck Jones, S. Walker, S.E. Robertson, Probabilistic model of information retrieval: development and comparative experiments. Part 1, Inf. Process. Manag. 36 (2000) 779–808, https://doi.org/10.1016/S0306-4573(00)00015-7.

[56] D. Arthur, S. Vassilvitskii, K-Means++: the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, Society for Industrial and Applied Mathematics, USA, 2007, pp. 1027–1035.

[57] S.P. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theor. 28 (1982) 129–137, https://doi.org/10.1109/TIT.1982.1056489.

[58] GOWP, Global observatory for water and peace [WWW Document]. URL. https://www.genevawaterhub.org/platforms/global-observatory-water-and-peace, 2023.

[59] GSWE, Global surface water [WWW Document]. URL. https://global-surface-water.appspot.com/, 2021.

[60] UN Water, SDG 6 data portal [WWW Document]. URL. https://www.sdg6data.org/index.php/en, 2023.

[61] J. Pita Costa, Business intelligence, built from open data, Waterworld Mag 38 (3) (2022).

[62] S. Mohr, U. Ehret, M. Kunz, P. Ludwig, A. Caldas-Alvarez, J.E. Daniell, F. Ehmele, H. Feldmann, M.J. Franca, C. Gattke, M. Hundhausen, P. Knippertz, K. Küpfer, B. Mühr, J.G. Pinto, J. Quinting, A.M. Schäfer, M. Scheibel, F. Seidel, C. Wisotzky, A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe -- Part 1: event description and analysis, Nat. Hazards Earth Syst. Sci. 23 (2023) 525–551, https://doi.org/10.5194/nhess-23-525-2023.

[63] Welt, Deutscher Wetterdienst warnt vor Starkregen, 2021.

[64] Wetter, Feuerwehr Befreit Autofahrer in Überschwemmter Unterführung, 2021.

[65] Global Drought Observatory, GDO copernicus [WWW Document]. Reports Sev. Drought Events. URL. https://edo.jrc.ec.europa.eu/gdo/php/index.php?id=2050, 2023.

[66] N. Bezak, M. Mikoš, Changes in the compound drought and extreme heat occurrence in the 1961–2018 period at the european scale, Water 12 (2020) https://doi.org/10.3390/w12123543 (Switzerland).

[67] I. Novalija, M. Papler, D. Mladenić, Towards social media mining: twitter observatory, in: SIKDD, vol. 2014, 2014, pp. 1–4.

[68] M. Bunge, Causality and Modern Science, Causality and Modern Science, 2017, https://doi.org/10.4324/9781315081656.