Environmental Research Communications

LETTER

# Classifying the 2021 'Ahrtal' flood event using hermeneutic interpretation, natural language processing, and instrumental data analyses

**Michael Kahle**[1,*], **Michael Kempf**[1,2,3], **Brice Martin**[4] and **Rüdiger Glaser**[1]

[1] Department of Physical Geography, Institute of Environmental Social Science and Geography, University of Freiburg, Schreiberstr. 20, 79085 Freiburg, Germany
[2] Department of Archaeology and Museology, Faculty of Arts, Masaryk University, Arne Nováka 1, 60200 Brno, Czech Republic
[3] McDonald Institute for Archaeological Research, University of Cambridge, Downing Street, Cambridge CB2 3ER, United Kingdom
[4] UHA, FSESJ, Campus Fonderie, 16 rue de la Fonderie, 68093 Mulhouse Cedex, France
[*] Author to whom any correspondence should be addressed.

**E-mail:** michael.kahle@geographie.uni-freiburg.de

**Keywords:** natural language processing, ahrtal flooding, machine learning, documentary sources, risk management

OPEN ACCESS

RECEIVED
25 November 2021

REVISED
5 April 2022

ACCEPTED FOR PUBLICATION
11 April 2022

PUBLISHED
17 May 2022



## Abstract

Extreme weather events and severe flash floods during July 2021 caused numerous deaths and massive ecological disasters across Europe. The regionally overstrained environmental and socio-cultural resilience triggered an intensive discussion about cause and effect, responsibilities and public denouncement, and the financial consequences of climate-induced extreme events. In this article we analyze the flood event by four methodological approaches: (1) hermeneutics, with an analog interpretation of printed newspapers and sources; (2) text mining and natural language processing of digital newspaper articles available online; (3) precipitation and discharge models based on instrumental data; and (4) how the findings can be linked to the historical extreme floods of 1804 and 1910, based on documentary source analysis. These four approaches are used to compare and evaluate their consistency by tracking the course, consequences, and aftermaths of the flood disaster. The study shows a high consistency between the analog, digital, and instrumental data analysis. A combination of multidisciplinary methods and their application to historical events enables the evaluation of modern events. It enables to answer the question of return periods and intensities, which are indispensable for today's risk assessments and their social contextualization, a desideratum in historical and modern climatology.

## 1. Introduction

Extreme floods are among the most severe natural disasters. The event during the 14/15th of July 2021 with more than 200 victims, is one of the most hazardous in Europe since 1950. Rapid flash floods and extensive flooding events have dramatically increased in frequency and intensity over the past decades (Blöschl *et al* 2020, Ionita and Nagavciuc 2021, Moraru *et al* 2021). A central question is to which extent these events are amplified by climate change and how they can be evaluated in the context of risk management (Merz *et al* 2011). Here, the availability of comprehensive and long-term data series is considered a major recent achievement (Börngen and Tetzlaff 2000, Glaser and Stangl 2004, Herget and Meurs 2010, Blöschl *et al* 2020). Increasingly, digital portals have been created to provide comprehensive high-resolution flood information, such as ORRION (Himmelsbach *et al* 2015, Giacona *et al* 2019), the virtual research environment tambora.org (Riemann *et al* 2016), and UNDINE of the Federal Office of Hydrology and the Federal Ministry for the Environment, Germany. Furthermore, the European Flood Awareness Systems (EFAS) and the Global Flood Awareness Systems (GloFAS) provide forecasts and temporal analyses. In addition, social media rises in importance (Casagrande *et al* 2015, Jüpner 2016, Douvinet *et al* 2017, Jüpner *et al* 2018, de Bruijn *et al* 2019). The European Drought Impact Report Inventory (EDII) and the European Drought Reference (EDR) database (Stagge *et al*

2013, Stahl *et al* 2016) are used to comprehensively analyze textual material about the impact structures of droughts via preconceived categories.

In this paper, we investigate, how text harvesting of online media and natural language processing (NLP) can be used to explore and evaluate extreme events such as the catastrophic flood of July 2021 in the Ahrtal Region (Germany). In general, the analysis of current weather-, and climate-related information from digital media and formats has increased significantly in recent years (Niforatos *et al* 2017, Lin *et al* 2016, Zisgen *et al* 2014). Qualitative and quantitative evaluations using various media formats, including newspapers, have become scientifically established (Kirilenko and Stepchenkova 2012, Schmidt *et al* 2013, Boussalis and Coan 2016, Comby and Le Lay (2019)). A generally strengthened refinement and an increase in complexity can be noticed. Common methods from the field of machine learning include classification, thematic clustering, Natural Language Processing (NLP), vectorization and neural networks or even reinforcement learning (LeCuan *et al* 1998, Kohonen 2001, Mahdisoltani *et al* 2013, Schmidhuber 2015, Mikolov *et al* 2013). Their application to newspaper archives and information in the context of climatic issues has for example been demonstrated by Yzaguirre *et al* (2016) with their impressive analysis of over 2 million articles, and Kang and Park (2018) using Korean newspapers. Kim and Kang (2018) exemplified keyword generation via word vectorization. Topic extraction and different methods of classification had also been used by Saura *et al* (2021a, 2021b) to analyze security issues and privacy concerns when using smart living devices (Internet of Things IoT). They proved to be successful on such different media as short messages on the social network twitter and scientific articles covering these topics.

The approach presented in this article aims at comparing and evaluating the 2021 flood event using spatial analysis and NLP. Long term records play a major role in today's assessment of return periods and intensities. In order to evaluate this, the insights gained from the analysis are applied to two historical extremes of 1910 and 1804. We then analyzed to what extent the application of NLP to digitally available newspaper articles and to historical sources represents an additional value for risk assessment. The findings of the NLP are compared to an analog interpretation of printed articles during and after the flood event. Finally, the event was modelled using instrumental precipitation and runoff data. This opens-up further scientific and numerically reproducible levels of comparison. This approach is originally multidisciplinary in the way that it relates hermeneutic, analog processes of text interpretation to the digital approaches of NLP and instrumental analysis and modelling. This enables a number of scientific questions and hypotheses:

(i) To what extent is a content-related, spatial, and temporal comparison of the different methodological approaches feasible? The focus is initially put on the mutual evaluation of the individual methodological strands.

(ii) How do these methodological approaches complement each other, and can they be linked?

(iii) Can the manual interpretation of big data be adequately replaced by digital methods?

## 2. Data and methods

Comprehensive data wrangling of the harvested and data-mined articles in open access media was carried out. These were compared to manually collected newspaper articles from a quality daily newspaper in to identify main topics and keywords in their temporal course, to derive pathways, and for cross-validation purposes. Simultaneously, the identified topics were used for the systematic search in the news portal 'newsapi.org' (https://newsapi.org/, last accessed, 25th of March 2022) and for the initial extraction of the text corpus. Furthermore, historical records of flood events along the river Ahr and model simulations of the 1910 and 1804 events were integrated. The historical text corpus was analyzed manually as well as through NLP to derive topics and keywords, the temporal development of the events, and their pathways. In parallel, official numerical climate and gauge data from the German Weather Service (DWD) and the Federal Agency for Water Research (BfG) were used to carry out terrain and precipitation-based run-off simulations. The temporal and spatial outline were visualized in a GIS (Geographic Information System). Finally, the results were used for further cross-validation (figure 1).

### 2.1. Online news: harvesting, data mining, and NLP

The focus of the analysis was put on publicly available German online articles collected using the news search engine „newsapi.org'. The whole processing was done in Python V3.8.5 and used the modules pandas, nltk (Bird *et al* 2009), spacy, scikit-learn (Pedregosa *et al* 2011) and HanoverTagger (Wartena 2019) for data wrangling. The visualization was done using matplotlib (Hunter 2007), seaborn, networkx, and cartopy (Met Office 2015). The basis for the harvesting were defined keywords, which led to a total of more than 5,000 datasets containing the

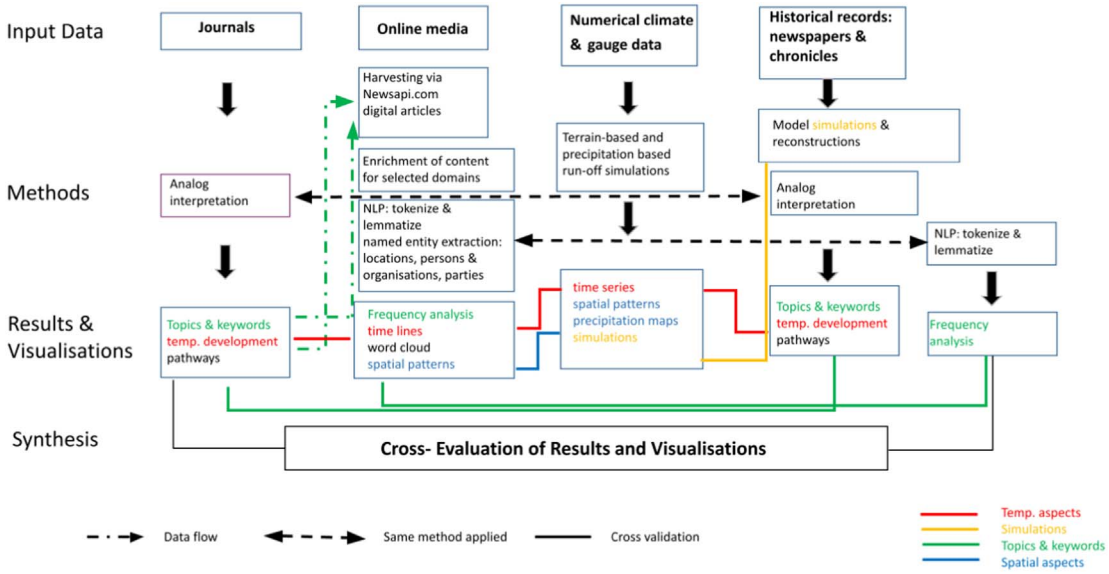



**Figure 1.** Conceptual outline and design of the workflow.

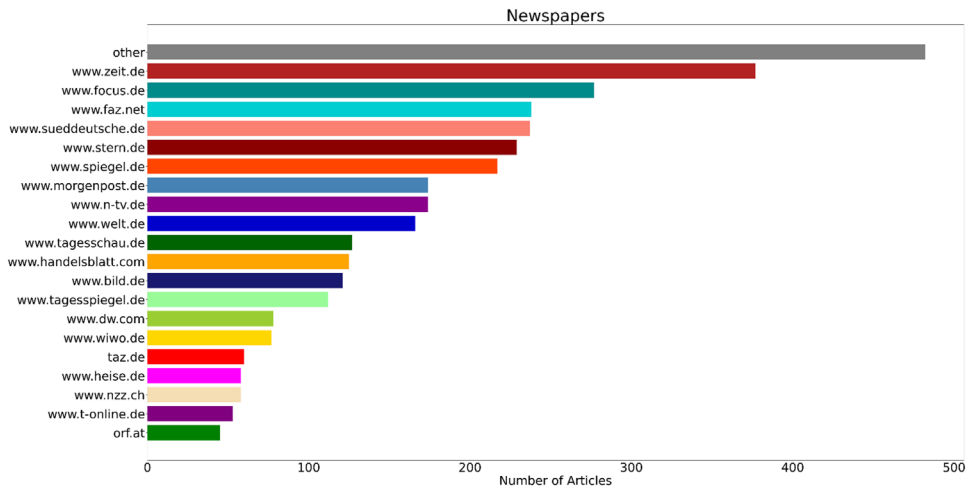



**Figure 2.** Number of articles in the 20 strongest media, with 40 or more contributions.

url, the title and a short description of the articles (Kahle 2021b ). These were sifted regarding their plausibility and duplicates were automatically removed. Thus, about 1500 of the articles were excluded from further processing and analysis. For the most 20 strongest domains mining routines for the html structure were applied using Beautiful Soup (Richardson 2020) to gain a more detailed extraction of the content. The remaining 3474 articles finally sum up to 115748 sentences and 2157390 words (Kahle 2022).

Significant shares are distributed among 20 different newspapers (figure 2), representing a broad political and content spectrum. The articles also include German-language titles from Switzerland and Austria. The shares remain fairly constant over time.

In a first step, the appearance of the keywords in the various newspapers was quantified. These counted appearances can be grouped according to further criteria, e.g. the date of publication or the media internet domain. For further refinement, the text corpus of each article was broken down to sentences, tokenized, and lemmatized on word level. This step also allowed distinguishing the type of word as nouns, adjectives, or verbs. Named entities were identified for organizations, persons, and locations. Extracted location names were cross validated using geonames.org (last accessed 04th of April 2022) and enhanced by their latitudes and longitudes and their corresponding countries. In a next step, the corpus was linked to the keywords using a supervised Bayesian classification (Zhang 2004). For each word, the probability of belonging to one of the classes was determined. This was done reciprocally to the distance of the words to the keywords. Finally, the temporal course of the topics can be mapped on different scales using the derived probabilities on word level. The topic

classification based on the multiplied probabilities can be done temporally via annual, monthly, weekly, or daily resolution, or textually via media domains, complete articles, individual sections, and sentences down to individual words. Optionally, the determined topics can be counted up to a different level. For example, it is possible to count the sentences dominated by a topic for each day. Since a subset of words were location names and are geo-referenced, their associated probabilities can further be spatially interpolated and mapped. Additionally, an unsupervised topic extraction was applied using non-negative matrix factorization NMF (Hoyer 2004) and Latent Dirichl *et al* location LDA (Blei *et al* (2003)) on extracted tf-idf features (Bun and Ishizuka 2001). The relation between the so extracted topics and their corresponding keywords to the manually given topics was determined by using the Bayesian classification from above and helps to get a theme for the untitled topics (Saura *et al* 2021a).

### 2.2. Modern newspaper articles and their analog interpretation

Furthermore, newspaper articles were analyzed analogously, building on articles from the Süddeutsche Zeitung (SZ), supplemented by other newspapers and magazines, public television and broadcasting news and reports. The SZ is considered a widespread quality daily newspaper with a subscription of more than 430,000 copies (1.41 million readers in 2011). It is most popular among rather highly educated and higher income groups, and middle-aged people and executives. For this study, a total of 73 SZ articles were evaluated in detail, written by 54 authors and collectives from different resorts. The articles appeared almost daily. In addition, a few articles were taken directly from the press agencies dpa and reuters. In some cases, several articles were published per issue and placed in various sections, including lead stories, economics, politics. Most of the articles were illustrated. From the articles, the main topics and their chronological progression were analyzed. The analog analysis was also used to identify the major topics for the evaluation and clustering and to extract initial keywords for the search and further evaluation of the digital media.

### 2.3. Numerical weather and gauging data

In addition to the text processing numerical data was analyzed. Daily and hourly time series were acquired from the national weather and federal hydrological services to evaluate the instrumental run-off, water level, and precipitation development during July 2021. GIS-based maps and time series analysis illustrate the spatio-temporal dimension and the correlation between the extreme precipitation and the flooding event from the 14th to the 15th of July. Furthermore, the data served as a cross-evaluation parameter. Precipitation values for July were acquired from the DWD open data portal for the reference period 1900–2021. The arithmetic mean value (m) and the standard deviation (SD) were calculated from all values. SD was subtracted from and added to the July monthly mean value (m-SD; m+SD) to create the range of the standard deviation for the reference period. Both raster sets were merged to create the precipitation anomaly range for July 2021. The rivers Ahr, Rhine, Rur, Inde, Erft and the river Ahr's tributaries were extracted from the geoportal of the BfG (WasserBLIcK 2021). Using the ASTER Global Digital Elevation Map V3 (DEM) from the NASA (NASA 2019, 2021) earthdata open science portal and SAGA GIS (2021) the drainage basin was simulated based on the topographic features of the DEM. In a next step, the DEM was clipped to the extent of the drainage basin and reclassified using QGIS and a classification into 100 m ranges. From the DWD RADOLAN data portal, hourly precipitation values were acquired and accumulated for the period between the 12th to the 15th of July 2021. The July precipitation values within the drainage basin were accumulated for each topographical range and a time-series was plotted, which represents topography-dependent precipitation variability within the extent of the drainage basin of the river Ahr over the long-term observation period 1900–2021. A comparison between the long-term precipitation records and the short-term totals for the period 12th of July—17th of July has been carried out based on hourly gridded precipitation data from the DWD RADOLAN dataset. The reprojected data has been clipped to the extent of the drainage basin and the summed-up total values for each hour were calculated.

### 2.4. Historical sources and records of flood events along the river Ahr

Historical data were used to enable long-term comparison. In the 19th century, a total of 17 severely damaging flood events were recorded in the river Ahr valley. Only four were related to snowmelt and ice drift in winter, most of them were triggered by heavy and continuous rainfall during summer (Frick 1933 1955, Seel 1983, Börngen and Tetzlaff 2000, Kohl 2007). For the 20th century, Seel (1983) listed 21 events, seven of which were summer floods. During the 18th century there were 15 reports and 10 for the 17th century. The decreasing number of reports reflects the declining density of records. Several approaches to quantify historical flood events have been worked out by Bürger *et al* (2006 2007), Glaser *et al* (2006), Sudhaus *et al* (2008a, 2008b), Herget and Meurs (2010), emphasizing the high suitability of documentary resource analysis for trend detection.

The flood disasters of 1804 and 1910 show remarkable parallels to 2021 (Börngen and Tetzlaff 2000, Kohl 2007). This applies to the weather conditions, the extensive damage, the high number of victims, and partly

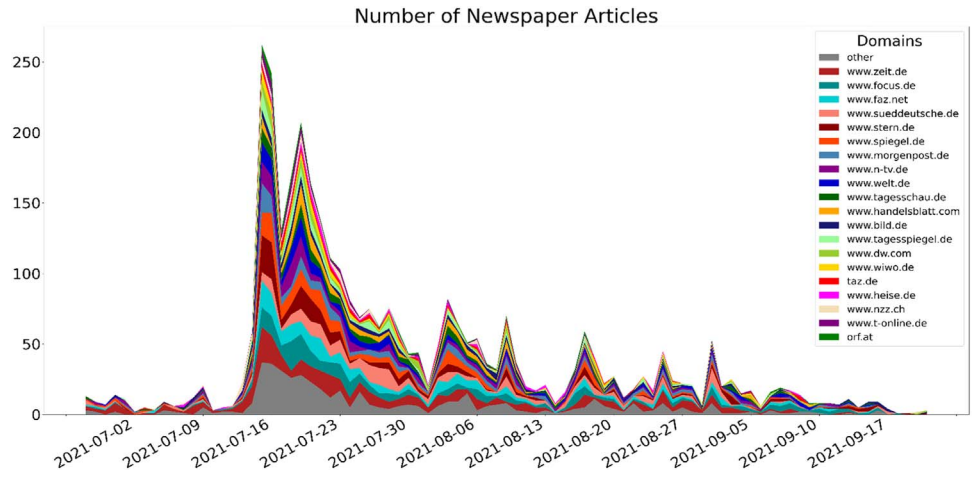



**Figure 3.** Publication date of the articles.

also to the societal, economic, and political reactions and debate (Frick 1933 1955). A further scientific evaluation of the two flood disasters was carried out by Grommes (1930), who determined the peaks of the 1804 and the 1910 floods based on flood marks. Roggenkamp and Herget (2014) reconstructed the peak discharges of the 1804, 1888, 1910, 1918, and 1920 events. The value for 1804 with 1200 $m^3 s^{-1}$ at Dernau is of particular interest, because it is the highest quantifiable value since 1800. The peak discharge in 2021 was estimated to about 1000 $m^3 s^{-1}$ (+/−200 $m^3 s^{-1}$) (Kreienkamp *et al* 2021), which highlights the likelihood of recurrence of exceptional flood events.

In addition to the documentary sources, we analysed historical newspaper from 1910 of the 'Bonner Zeitung' and 'Freiburger Zeitung' and from 1804 of the 'Allgemeine Intelligenz- und Wochenblatt für das Land Breisgau' (Kahle 2021a). We categorized and reclassified their information content to match the modern article structure of the 2021 event. This is based on the key topics obtained from recent articles about the 2021 flood event and the corresponding keywords, which were adapted to the historical terminology.

## 3. Results

In the following, the results of the methodological approaches are presented and discussed. A particular focus is put on the number and temporal appearance, the main topics including political parties, the spatial dimension, and the comparison of the 2021 flood to the 1910 and 1804 events.

### 3.1. Online media analysis

The harvested articles show significant temporal structures. Most articles appeared shortly after the event on the 15th of July (figure 3). A few secondary peaks can be identified. The general decline continues until mid-September.

In a further step, the articles were processed word by word, their frequency was counted, and consecutive nouns were connected. In the following figure 4, the main words are presented as a word cloud according to their frequency and scaled in size and connected as lines. Frequently connected words are shown in short distances, less frequent connections are shown further apart. The words with the most frequent connections center in the middle, rare connections appear towards the edge. The most frequent words are 'flood', 'catastrophe' and 'deluge'. The classification of location, person and organization obtained by named entity classification was colored.

The extraction of the main topics and keywords from the analog analysis and the corpus obtained by text mining and processed by NLP resulted in a broad spectrum of topics. A total of 17 main topics were manually defined. The corresponding keywords were determined using an NLP-supported frequency analysis of the entire text corpus. The frequencies and connections are shown in figure 5.

The unsupervised topic extraction using tf-idf features and NMF decomposition results in similar patterns. The weather conditions and their linkage to climate change, the damages and the coverage through insurance companies as well as the amount of victims are prominently featured. The involvement of politicians, rescue organizations, voluntary helpers, and charity events is visible, and their corresponding keywords are homogeneously related to a specific bayes class. The question of adequate warnings and responsibilities appears as separate subjects. Other topics, like health issues in conjunction with the corona pandemic or the typical local features of wine growing get only visible by exploring several dozens of topics. Some few topics likewise the pollution of drinking water or the occurrence of extremists, looters, and cheaters stays invisible - probably due to

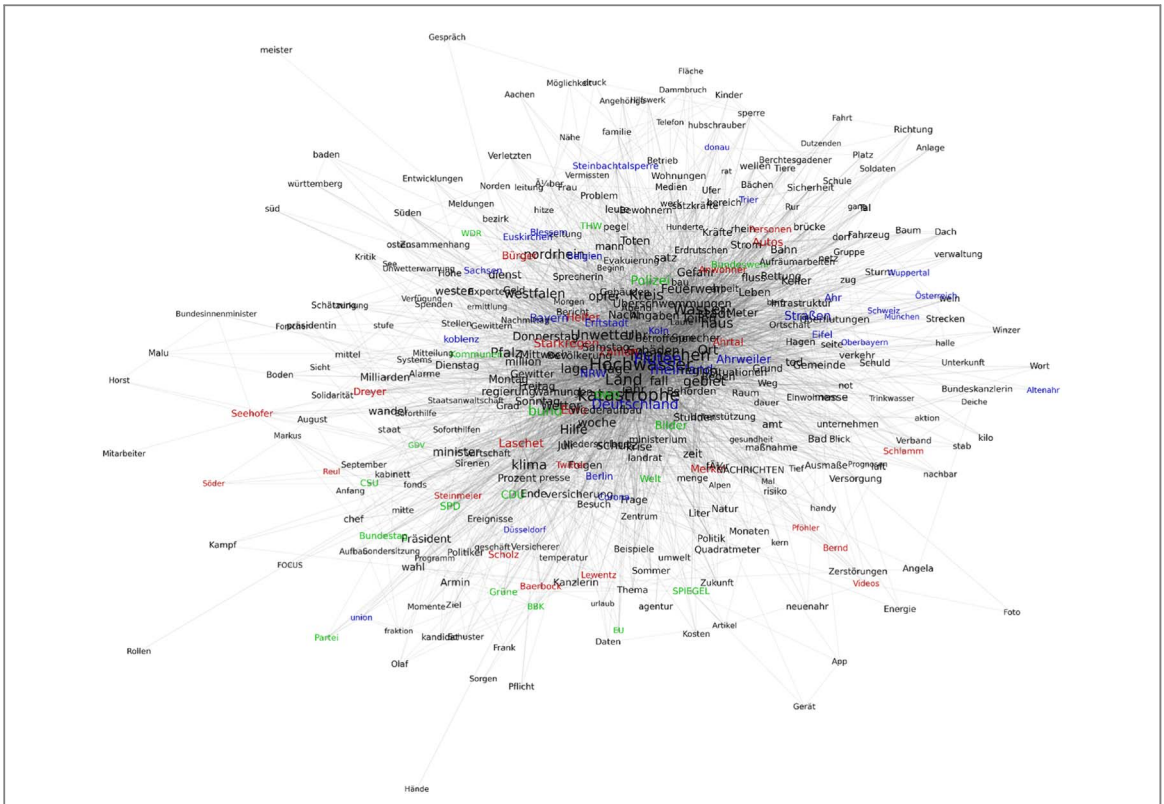



**Figure 4.** Word Cloud of the collected corpus. Detected names are given in blue, people in red, and organizations in green.

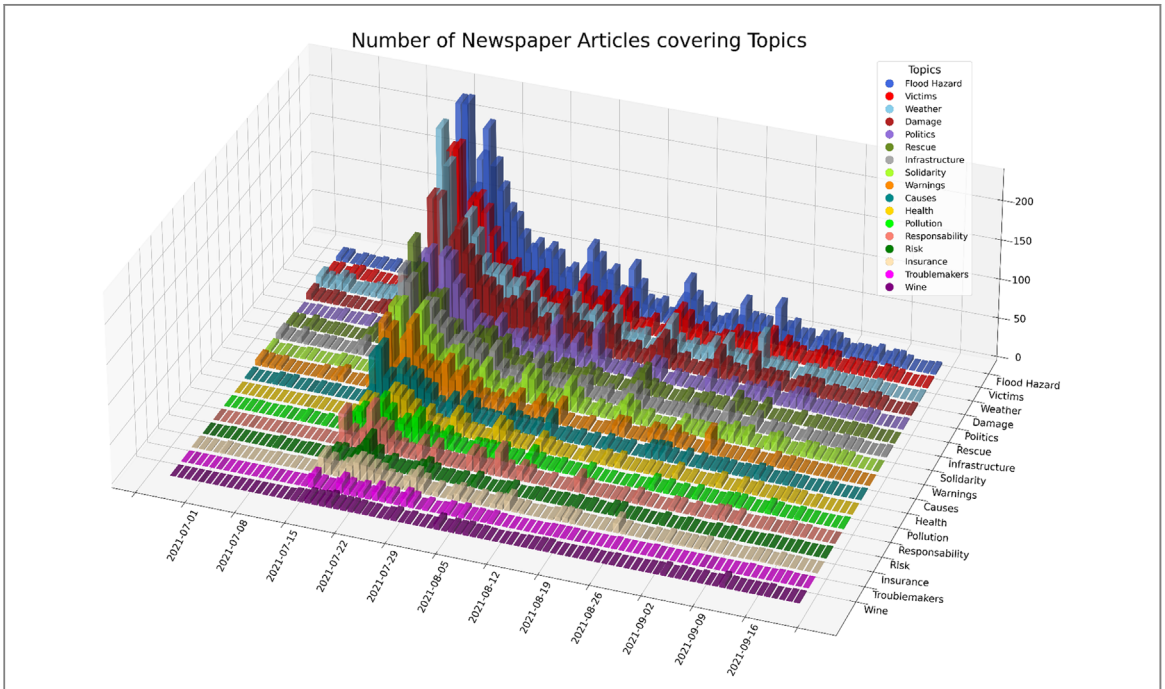



**Figure 5.** Number of articles versus identified keywords and publication date.

their strong interconnection to more prominent topics. Some groups also appear to be an inhomogeneous mixture covering several bayes topics. Passages of text dedicated to advertising and access restrictions can be identified and then ignored. In contrast, the use of LDA results in extracted topics that are mostly inhomogeneous regarding the expected themes and therefore less useful than the topics extracted by NMF (see figure 6).

As expected, the number of articles increased rapidly with the onset of the flood and reached a maximum during the first week. Afterwards the number of articles decreased continuously. The focal points remained in

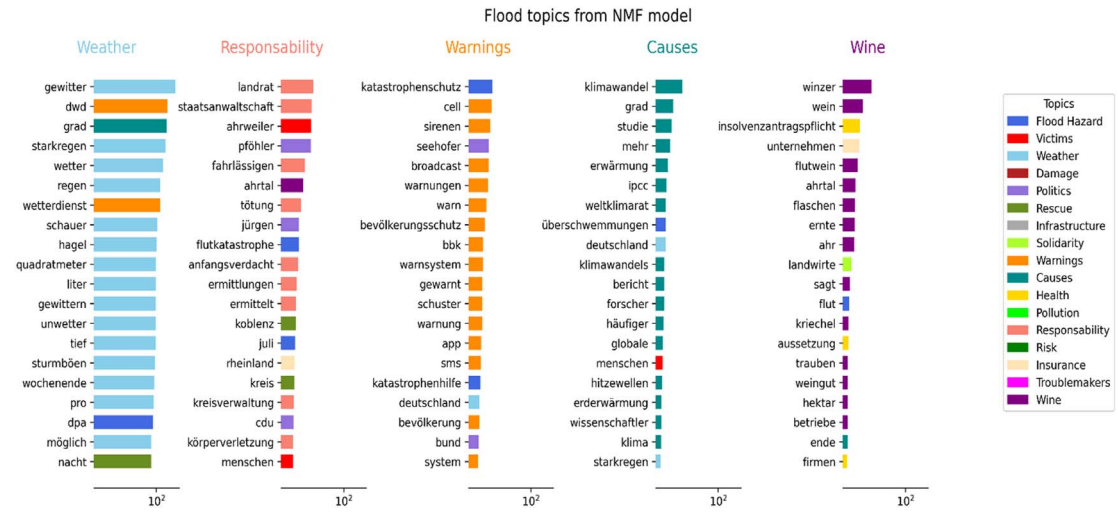



**Figure 6.** Selected flood topics from none negative matrix factorisation (NMF) model.

almost equal proportion to each other. Reports occurred almost daily until mid-September 2021. Most of the references referred to the flood disaster and the weather situation. They account for over 50% of the information. References to victims and damages account also for high proportions. The references to politicians, political parties, and the state and federal governments, here-in summarized as 'politics' further account for a large share. In this context, the federal election on 26th of September 2021 played a reinforcing role in the reporting frequency. All newspaper articles report in a broad spectrum. The focus sometimes varies, often depending on the target groups. For example, the business-oriented newspapers such as the 'Handelsblatt', the 'Wirtschaftswoche' and the 'NZZ' report more frequently on insurance issues. The distribution of place names in the relevant articles of the named entity analysis is shown in figure 7.

The map shows the spatial patterns of coverage of the flood event in the Ahrtal Region, Erftal, and parts of the Ruhr Area around Dortmund. It highlights the focal points of the flood event and coincides very well with the precipitation distribution (see figure 8). Since the flood disaster took place during the election campaign for the Bundestag on the 26th of September 2021, many politicians showed up on the spot, which is sometimes referred to as 'boot politics'. Regardless, this is good practice during severe disasters to get a picture of the damage, coordinate government emergency aid and launch additional relief funds, and express solidarity with those affected. If we analyze the topics on a weekly level as relative shares, further structures become apparent. Towards the end of the reporting period, for example, topics on 'wine' dominate, which can be explained by the upcoming grape harvest.

### 3.2. Analog interpretation of modern newspapers

The analog evaluation of the modern newspapers revealed a broad spectrum of topics. At first, the focus was put on weather conditions and the flood disaster itself, usually accompanied by reports of the severity of the damages and the proposed number of victims, including missing people (Hagen *et al* 2021, Müller-Arnod and Stegemann 2021, Scheuer 2021). Other articles highlighted topics such as oil contamination, failure of sewage treatment plants, and lack of clean water (Hermann 2021) as well as peculiarities caused by Covid-19. The long-term consequences were already starting to be evaluated, particularly due to discussions about destroyed schools and hospitals (KPF (2021)). The flooding restricted access to pharmacies and medical care, which was distributed among the victims (Knuth and Parnack 2021, Niewel 2021, Schwarz 2021). The local particularities of the wine-growing region and future economic alternatives in tourism were also included in the articles (Britzelmeier 2021).

The strong link to climate changes was emphasized, partly triggered by the publication of the latest IPCC report (IPCC 2021). Questions arose about the deficits of the early warning system and future responsibilities (Schneider 2021). The British hydrologist, Hannah Cloke, even spoke of a tremendous system failure. The federal disaster warning system and the lack of uniform nationwide management were repeatedly criticized. In contrast, individual handling at the administrative district and city levels was considered an advantage (Fahrenholz 2021, Hummel 2021, Reuters 2021, Schneider 2021). The role of insurance for handling the damages popped up and the introduction of an obligatory natural hazard insurance was discussed, as were relief funds for the victims (Bellmann 2021, Krieger 2021, REUTERS 2021).

The focused reports about the frequent appearances of politicians on the site were amplified by the simultaneous federal election campaign. Massive media presence with benefit events and appeals for donations

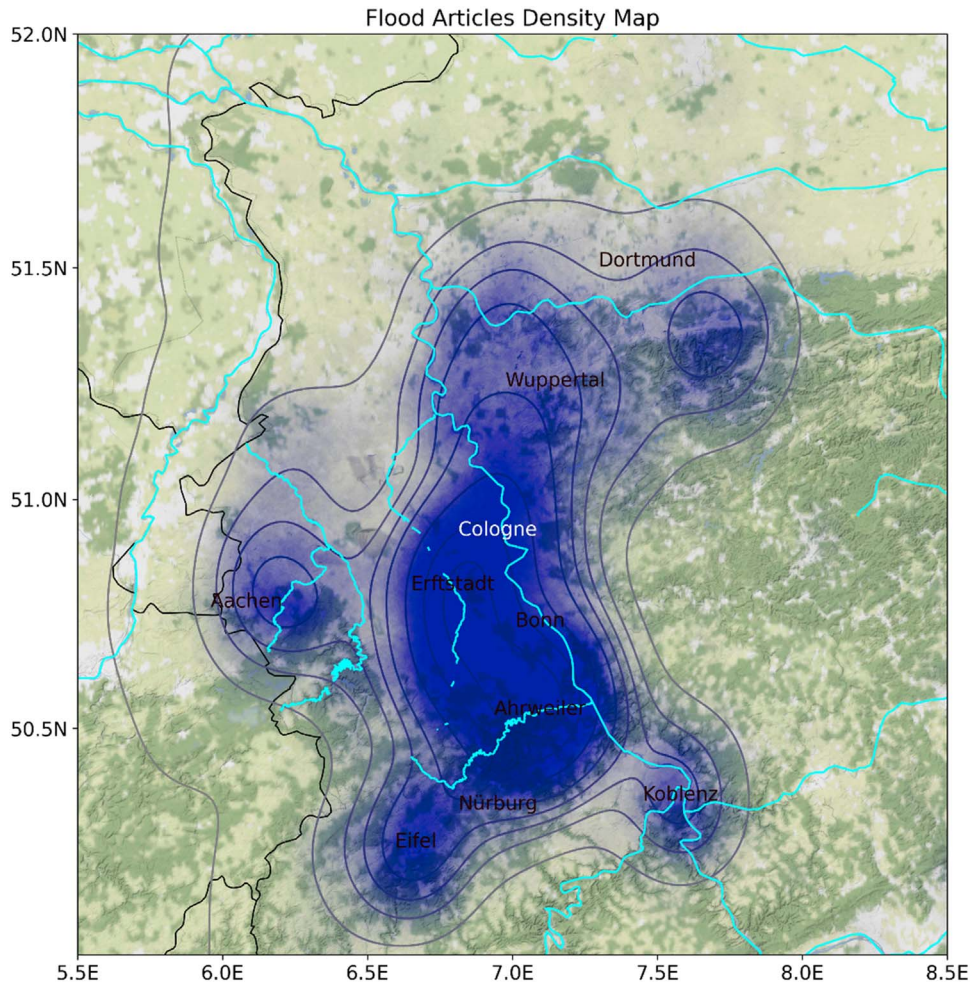



**Figure 7.** Spatial pattern of the place names in the relevant articles of the named entity analysis. Place names in black and white are only for orientation.

underlined the willingness to support (Bauchmüller 2021, Deininger 2021, Gammelin 2021). The wide-ranging factual information was constantly enriched with very emotional images and reports. The contra-productive and politically self-serving activities of so-called 'Reichsbürger' and so-called 'Querdenker', politically motivated by right-wing views as well as looting, were further covered by the press (Grossmann 2021, Steinke 2021).

Finally, issues of administrative consequences, planning directives, and infrastructural reconstruction were addressed. Improvement of technical flood protection, incorporating innovative concepts of the sponge city and blue-green infrastructure with infiltration swales, more greenery and unsealing have further been reported (Rühle 2021, Weiss 2021).

### 3.3. Instrumental data findings

Heavy rain events are considered the main driving factor behind run-off maxima (e.g., Bürger *et al* 2006, Bronstert *et al* 2018). Compared to the long-term period of 1900–2021, July 2021 total precipitation reached a maximum of the instrumental record (figure 8). The calculated precipitation anomaly highlights the regional concentration of rainfall during the month in western Germany with a secondary regional peak in southern and eastern Germany. The magnitude of the event, however, is masked by the monthly total precipitation values and needs to be visualized using high-resolution temporal precipitation values within the drainage basin. Figure 9 highlights the strenght of the precipitation anomaly during the 14th of July 2021, compared to the previous three days. Between approximately 9.00 a.m. and 6.00 p.m., the precipitation remained at very high totals, which triggered an immediately rising water level in the river Ahr valley. With a response time of approximately 6 h, the accumulated rainfall from the tributaries and the headwater rivulets entered the main river discharge volume and led to a dramatic rise in total run-off in the valley. At Bad Bodendorf, which is located at the lower parts of the valley and in topographically rather gentle terrain, the discharge volume reached a maximum of over 400 $m^3 s^{-1}$, compared to the long-term average at the station of 6.24 $m^3 s^{-1}$. Considering the terrain roughness and the run-off characteristics at Bad Bodendorf, the upstream discharge volume and velocity would have been greatly intensified.

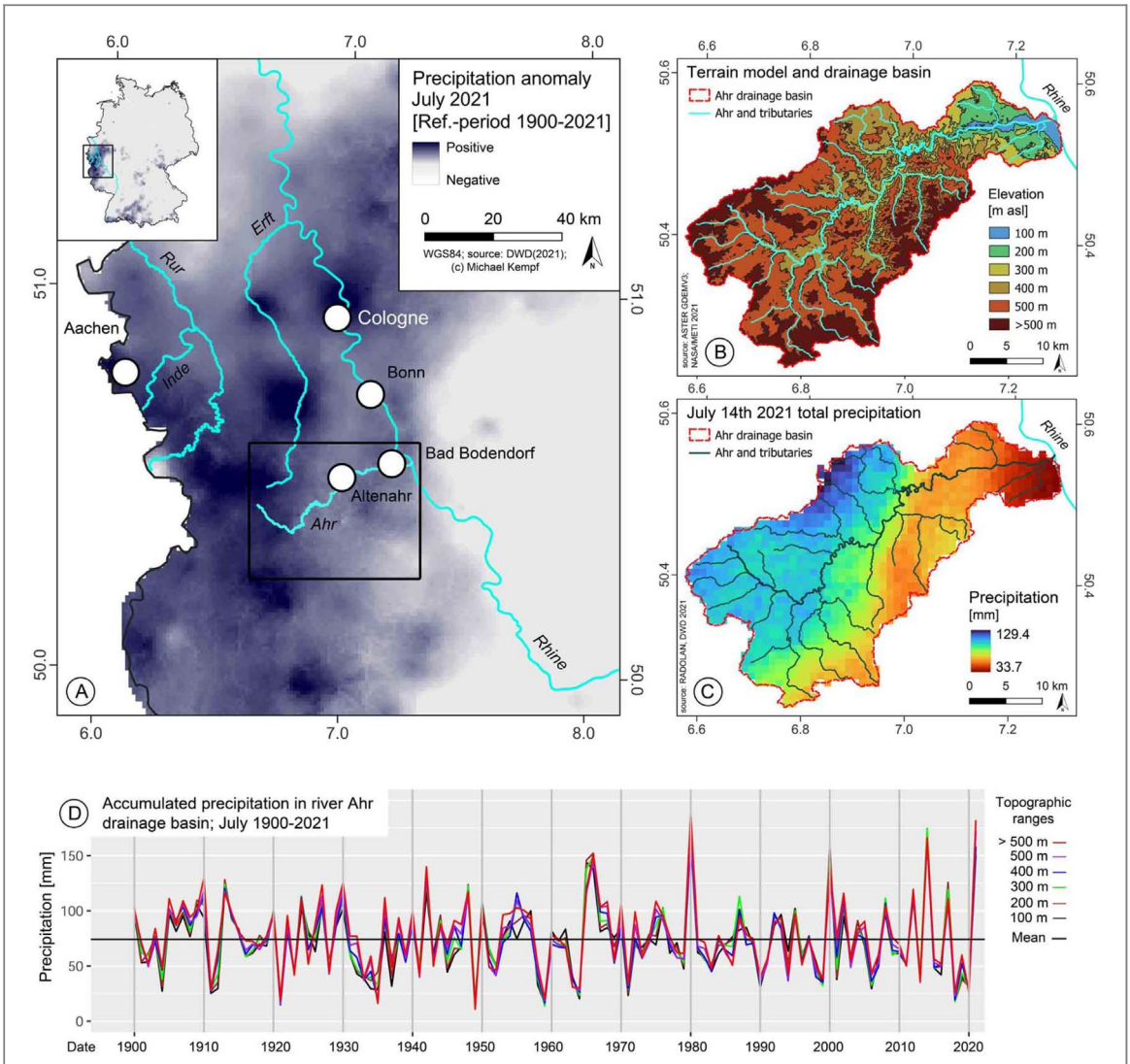



**Figure 8.** July 2021 precipitation anomaly in the study area compared to the reference period 1900–2021 (A). Reclassified digital terrain model and calculated drainage basin of the river Ahr and its tributaries (B). Accumulated total precipitation within the river Ahr drainage basin during the 14th of July 2021 (C). Long-term July precipitation variability in the river Ahr drainage Basin using topographical ranges (D).

Given the extent of the drainage basin of the river Ahr and its topographic heterogeneity, maximum rainfall before the flooding event can be localized in the northern part of the headwaters of the river Ahr. The rather remote location would most likely have contributed to the higher response time of the flooding event during the 15th of July. Although forest-covered, the uppermost soil layers were saturated due to pre-event precipitation rates during the 13th of July 2021 and would thus have favoured increased surface runoff, particularly amplified by channeling effects from the tributaries (e.g., Dreisbach and Armuthsbach, north of Schuld). The clayey soil composition could have further contributed to the superficial run-off, amplifying the discharge volume and counteracting the retention of precipitation in vegetation and forested areas in the upstream and headwater area. In addition, the simulated run-off, based on topography, shows hot-spots of flow accumulation in the western part of the drainage basin-most likely linked to the junction of the southern drainage area and the north-western part (figure 9(a)). The precipitation-based run-off model by Irving and colleagues and Domisch and colleagues (2018) further highlights the amplification of the flooding event at and after the confluence, which acted like an additional channeling effect (figure 9(b)).

### 3.4. Comparison of the 1804, 1910, and 2021 events

To evaluate the change of topics over time, the three catastrophic flood events of 1804, 1910, and 2021 were compared. The documentary sources about the historical events of 1804 and 1910 were classified according to the modern articles covering the 2021 event (figure 10). This is based on the key topics obtained from the recent articles and the corresponding keywords, which were adapted to the historical terminology.

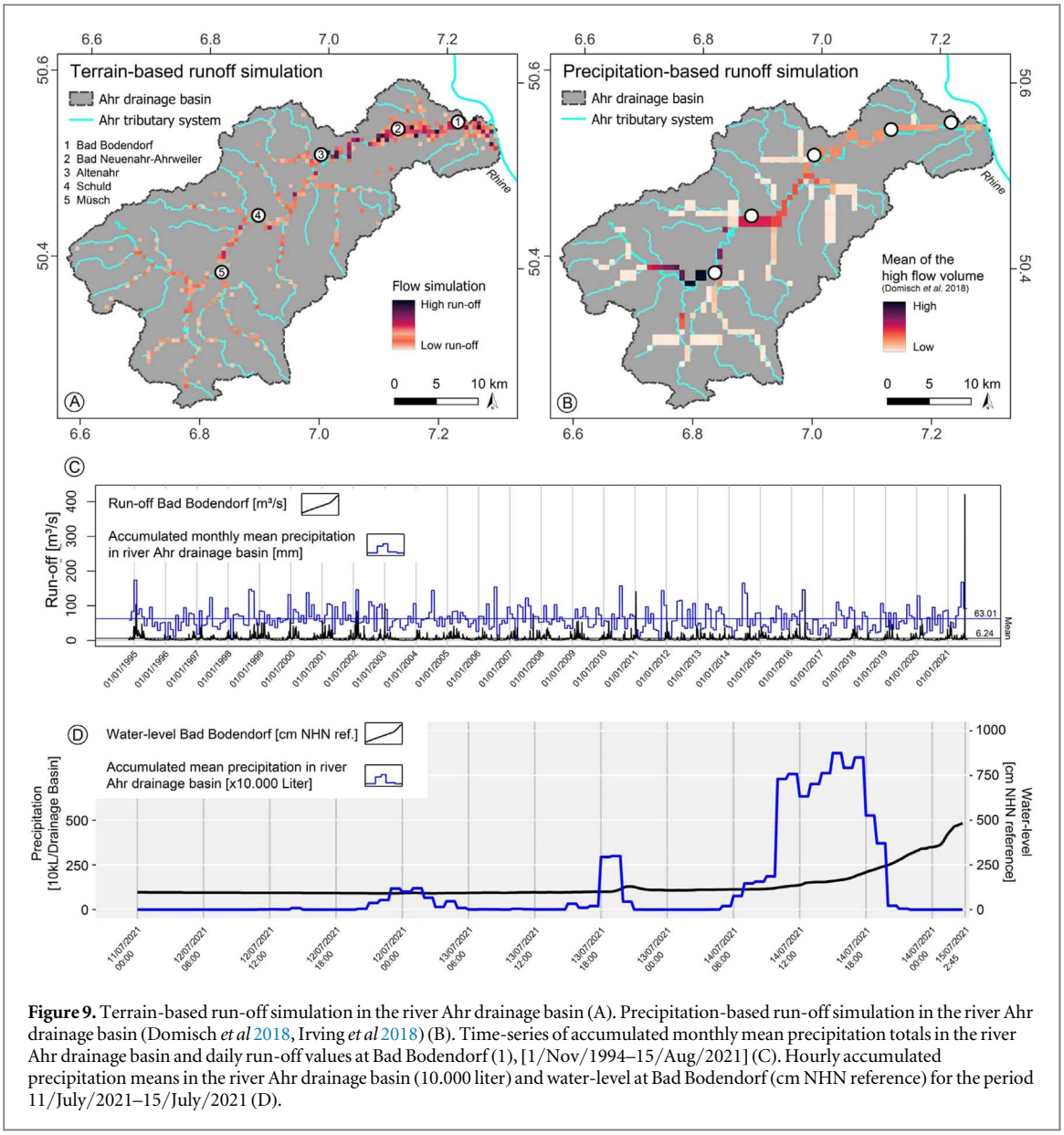
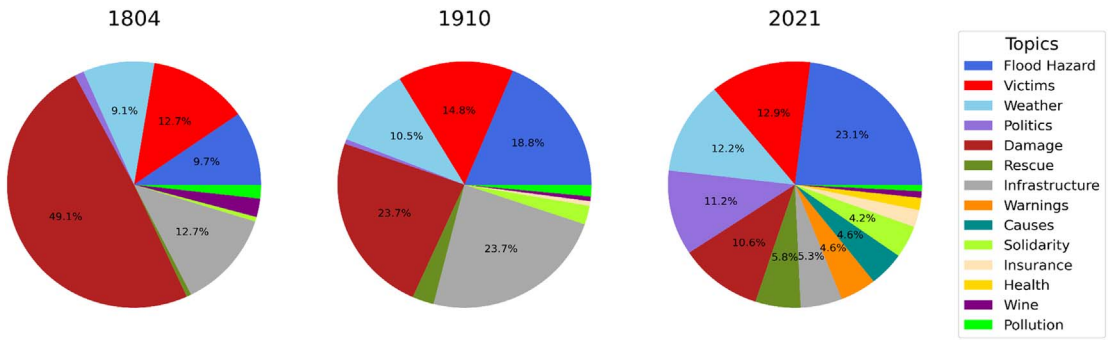



**Figure 9.** Terrain-based run-off simulation in the river Ahr drainage basin (A). Precipitation-based run-off simulation in the river Ahr drainage basin (Domisch *et al* 2018, Irving *et al* 2018) (B). Time-series of accumulated monthly mean precipitation totals in the river Ahr drainage basin and daily run-off values at Bad Bodendorf (1), [1/Nov/1994–15/Aug/2021] (C). Hourly accumulated precipitation means in the river Ahr drainage basin (10.000 liter) and water-level at Bad Bodendorf (cm NHN reference) for the period 11/July/2021–15/July/2021 (D).

The results show a shift in topics: 1804 was dominated by general descriptions of the damages (49%), followed by those relating to infrastructure (13%). They account for over half of the information, because mostly buildings were affected during the event. The numbers highlight the enhanced flooding vulnerability of the Ahrtal in historical perspectives. Subsequently, weather and flooding descriptions account for another 19%. During the 1910 flood disaster, the emphasis differs from that in 1804 by about 29%. In contrast to 1804, information about the damages accounts for about only 23%. Infrastructural losses were strongly emphasized in 1910 due to the development of the railway and the casualties among the constructors. In 2021, information about the flood event and the weather conditions accounted for about 35% whereas infrastructural damages and casualties were covered by about 29%. The most evident differences are now in topics of vulnerability, spatio-temporal scale, complexity of the disaster and shifts in perception of the catastrophe as a global event. However, a temporal bias is inherent in the data due to technical standards of communication, infrastructure, and economic development in 2021, compared to 1804 or 1910.

## 4. Discussion and conclusion

Multidisciplinary approaches, including digital methods of NLP, analog text interpretation, and instrumental data modelling have proven to be useful tools to analyze extreme events. Cross evaluation highlights the potential and strength of each individual methodical approach and the synergic surplus. This opens-up possibilities to evaluate communication of disasters and particularly the changing focus on topics over time. Building on this, the integration of historical events enables to link long-term analysis to modern flood risk management.




**Figuree 10.** Main topics in 1804, 1910, and 2021.

### 4.1. Potential of NLP to evaluate climatic extremes

The method of harvesting digital articles from online portals via text mining procedures is suitable to capture extreme flood disasters. It confirms the findings obtained by Brito *et al* (2020) using text mining procedures to analyze the 2018 and 2019 droughts. The procedure quickly yields large amounts of data that can be analyzed statistically and further exploited via NLP. This represents an immense advantage over the analog analyses, which are evidently slower and rather subjective in nature. This applies to the recently available information in digital newspapers as well as to the increasing number of archives containing digitized historical sources.

The digitally extracted texts reflect both the temporal development and the spatial coverage of the event. The relevant topics are equally represented. This can be concluded in particular from the joint analyses of the digitally harvested records and the analog interpretation of modern newspapers as well as from cross-validation using numerical data. Essential event inherent features, such as the rapid increase of the streamflow run-off characteristics and the subsequent exceeding of the maximum measurable gauge level are clearly visible in the time series data. Together with the numerous victims and the massive material damages, it emphasizes the frequently used designation of an 'event of historical magnitude'. Remarkable is the broad spectrum of topics, which is taken up almost simultaneously. Obviously, they include climatic events in combination with the actual flooding, issues of solidarity, rescue organizations, political and administrative failures and the follow-up including responsibilities and insurance capacities. In addition, the framing of the event likewise the political election in Germany or the worldwide corona pandemic gets visible. As expected, the number of contributions increases strongly with the event, however, it thins out during the following two months. The focal topics and the relationships among them remain the same. Unlike the numerical data, the text analysis highlights the emotional impact and conveys true-to-life images of the disaster, making the event vivid and tangible. This allows unique insights into the social, mental, and economic dimensions that are important for coping with future events. It further highlights issues of accountability and shortcomings in the reporting structure. These aspects are essential elements to improve prevention, which has become increasingly important in the aftermath of the event.

### 4.2. Analog text interpretation

The analysis of modern and historical text evaluation enables to estimate the impact paths and damage patterns. With these, both the aspects of inundation areas and intensities as well as return periods can be interrelated. It combines the numerical intensity analysis, which is based on modern run-off data and precipitation values, and the results of the text interpretation. Such damage patterns and impact paths cannot be mapped using numerical data alone. This applies to infrastructural damage, such as railways and bridges, which were particularly affected, but also to the number of victims. However, these are related to modern population density, land- use strategies, and other socio-cultural variables that change over time. This represents a certain bias of the data. In general, the multi-layered social, mental, and economic effects can be traced through the evaluation of documentary data and provide a useful information source for future disaster management.

### 4.3. Numerical data analysis

Instrumental data analyses provide an objective background for large-scale environmental models. However, spot-light data records are prone to technical failure during extreme weather events, which emphasizes the importance of cross-validation with hermeneutic and documentary sources to generate long-term risk assessment tools. The data underlying the approach presented here clearly shows the potential to transfer *in situ* records to other regions, independent of linguistic and socio-cultural contexts and developments. However, historical land-use and landcover change altered the landscape composition and the alluvial floodplain

characteristics through massive channeling and drainage activities during the 19th and 20th century. These anthropogenic overprints limit the comparability of historic extreme events and current climate-related environmental disasters.

### 4.4. Transferability to historical extremes

We can link the current developments to the comparative cases of the 1804 and 1910 events, revealing the differences in the focal themes that resulted from contemporary social-political and socio-economic conditions. The infrastructural features that did not exist in 1804, such as railroads, electricity, and telecommunication, were of course missing in the articles. However other important features were particularly vulnerable, such as water mills, which are no longer present today. By means of a correspondingly coarser classification, they could be integrated, for example, into the topic 'infrastructure'. A large part of the facilities such as bridges, roads, residential and farm buildings, schools, and administration facilities remain the same. Thus, the presented method is also suitable for retrospective analyses and corresponding comparisons.

### 4.5. Risk assessment and management

Recent results have shown that the designation of floodplains according to HQextrem is underestimated in modern flood risk maps (Thieken *et al* 2021). Intensities and return periods are incorrectly estimated because historical events are not adequately included, and the modern time series analysis do not reach back far enough (Roggenkamp and Herget 2014). In other Central European catchments, such as the river Neckar in southern Germany, which has a similar dimension compared to the Ahr catchment, it was demonstrated that the integration of historical events led to a reassessment of current HQextrem and HQ100 discharges and intensities and thus improved flood risk management (Bürger *et al* 2006, Sudhaus *et al* 2008a 2008b). This method offers important perspectives to prevent future disasters and further highlights aspects such as administrative and cultural boundaries.

In general, the integration of historical sources can help to improve modern flood risk management. Further insights into the social, mental, economic, and ecological implications can be gained through text analysis. The quantitative methods of NLP enable it to rapidly open-up large data volumes and to represent the diversity of aspects more objectively than using manual methods (Assael *et al* 2022). Since large data collections covering hydrological extremes are available for Central Europe (e.g., tambora.org, ORRION, or the Weikinn Collection, or digitized historical newspaper archives), there are multiple opportunities to explore them for an improved flood assessment.

## Acknowledgments

We are very grateful to three anonymous reviewers for their detailed and constructive suggestions and critics that helped to reshape the approach behind the article. We would further like to thank Yvonne Henrichs and Michael Göller, Hydrologischer Dienst der oberirdischen Gewässer, Landesamt für Umwelt, Rheinland-Pfalz who provided daily and hourly gauge data. M Ke received funding from the Masaryk University (grant number CZ.02.2.69/0.0/0./18_053/0016952; Postdoc2MUNIm order number 21 0053).

This publication is a contribution to the Past Global Changes (PAGES) Floods Working Group. In memory to its co-founder Bruno Wilhelm.

## Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.6094/UNIFR/222040, https://doi.org/10.6094/UNIFR/223318, https://doi.org/10.5281/zenodo.6406244.

## Author contributions

M K and R G. designed the study. M K, R G and M Ke wrote the initial version of the paper. M K provided harvesting and analysis of the online news. R G and B M provided the analog newspaper collection and interpretation. M Ke provided instrumental climate and gauge data analysis and the simulations. R G and M K collected and analyzed the historical data. M K, M Ke, and R G edited and wrote the final version of the manuscript.

## ORCID iDs

Michael Kahle https://orcid.org/0000-0001-8571-2821
Michael Kempf https://orcid.org/0000-0002-9474-4670
Brice Martin https://orcid.org/0000-0003-1515-2712
Rüdiger Glaser https://orcid.org/0000-0001-6819-2764

## References

Assael Y *et al* 2022 Restoring and attributing ancient texts using deep neural networks *Nature* **603** 280–3
Bauchmüller M 2021 Grüne verlangen mehr Katastrophenvorsorge *SZ* **172** S.5
Bellmann C 2021 Nach der Flut könnten Policen teurer werden *SZ* **189** S.19. 18.8.21
Bird S, Edward L and Klein E 2009 *Natural Language Processing with Python* (Sebastopol: O'Reilly Media Inc) 978-0-596-51649-9
Blei D M, Ng A Y and Jordan M I 2003 Latent dirichl*et al*location *Journal of Machine Learning Research* **3** 993–1022
Blöschl G *et al* 2020 Current European flood-rich period exceptional compared with past 500 years *Nature* **583** 560–6
Börngen M and Tetzlaff G (H) 2000 *Quellentexte zur Witterungsgeschichte Europas von der Zeitwende bis zum Jahr 1850* (Berlin: Stuttgart, Borntraeger)
Boussalis C and Coan T G 2016 Text-mining the signals of climate change doubt *Global Environ. Change* **36** 89–100
Brito M M, de Kuhlicke C and Marx A 2020 Near-real-time drought impact assessment: a text mining approach on the 2018/19 drought in Germany *Environ. Res. Lett.* **15** 1040a9
Britzelmeier E 2021 Das Winzerdorf, das sich selbst hilft *SZ* **172** S.8
Bronstert A *et al* 2018 Forensic hydro-meteorological analysis of an extreme flash flood: the 2016-05-29 event in Braunsbach *SW Germany The Science of The Total Environment* **630** 977–91
Bun K K and Ishizuka M 2001 Emerging Topic tracking system *Proc. Third Int. Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems. WECWIS* pp 2–11
Bürger K, Dostal P, Seidel J, Imbery F, Barriendos M, Mayer H and Glaser R 2006 Hydrometeorological reconstruction of the 1824 flood event in the Neckar River basin (southwest Germany) *Hydrol. Sci. J.* **51** 864–77
Bürger K, Seidel J, Dostal P, Glaser R, Sudhaus D and Mayer H 2007 Extreme floods on the 19th century in southwest Germany *Houille Blanche* **1** 67–73
Casagrande D G, McIlvaine-Newsad H and Jones E C 2015 Social networks of Help-seeking in different types of disaster responses to the 2008 mississippi river floods *Human Organization* **74** 351–61
Comby E and Le Lay Y 2019 Les trajectoires discursives et politiques des inondations du fleuve Sacramento : entre risque et catastrophe, entre ici et ailleurs *Annales de géographie* **726** 31–57
de Bruijn Jens A., de Moel Hans, Jongman Brenden, de Ruiter Marleen C., Wagemaker Jurjen and Aerts Jeroen C. J. H. 2019 A global database of historic and real-time flood events based on social media *Scientific Data* **6** 311
Deininger R 2021 Mit Merkels Hilfe *SZ* **165** S.2
Domisch S, Kuemmerlen M, Irving K, Kiesel J, Kakouie K and Jähnig S C 2018 MH21 - High Flow Volume, dataset *Scientific data* **5** 180224
Douvinet J, Gisclard B, Sekedoua Kouadio J, Saint-Martin C and Martin G 2017 Une place pour les technologies smartphones et les Réseaux Sociaux Numériques (RSN) dans les dispositifs institutionnels de l'alerte aux inondations en France? *Cybergeo : European Journal of Geography* **801** 1 (http://journals.openedition.org/cybergeo/27875)
Fahrenholz P 2021 Die flut und die folgen *SZ* **168** S.2
Frick H 1933 *Quellen zur Geschichte von Bad Neuenahr (Wadenheim/Beul/Hemmessen), der Grafschaft Neuenahr und der Geschlechter Ahr, Nauenahr und Saffenberg. Als Festschrift zum 75jährigen Jubiläum des Bades Neuenahr.* (Bad Neuenahr: Selbstverlag der Gemeinde Bad Neuenahr) pp 1–693
Frick H 1955 *Das Hochwasser von 1804.- Heimatjahrbuch des Kreises Ahrweiler* **12** 43–51
Gammelin C 2021 Jetzt muss geholfen werden *SZ* S.1
Giacona F, Martin B, Furst B, Glaser R, Eckert N, Himmelsbach I and Edelblutte C 2019 Improving the understanding of flood risk in the Alsatian region by knowledge capitalization: the ORRION participative observatory *Natural Hazards and Earth System Sciences* **19** 1653–83
Glaser R, Bürger K, Sudhaus D, Dostal P, Mayer H, Imbery F and Seidel J 2006 The historical flood events of 1824, 1845 and 1882 in Germany - their integration in an actual flood risk management by means of the extreme flood in 1824 *Publications S.H.F* 113–20
Glaser R and Stangl H 2004 Climate and floods in Central Europe since AD 1000: data, methods, results and consequences *Surv. Geophys.* **25** 485–510
Grommes G 1930 *Das Ahrtal-Eine anthropogeographische Studie* (Osnabrück: Buchdr. M. Steinbacher) 17ff
Grossmann V 2021 Querdenker im Flutgebiet *SZ* **166** S.10
Hagen P, Krieger F and Müller-Arnold B 2021 Hilflos zwischen Trümern *SZ* **163** S.2
Herget J and Meurs H 2010 Reconstructing peak discharges for historic flood levels in the city of cologne, Germany *Global Planet. Change* **70** 108–16
Hermann B 2021 Bombenfund und bestialischer Gestank.- Main Post **176** S. 9
Himmelsbach I, Glaser R, Schönbein J, Riemann and Martin B 2015 Reconstruction of flood events based on documentary data and transnational flood risk analysis of the upper Rhine and its French and German tributaries since AD 1480 *Hydrol. Earth Syst. Sci.* **19** 4149–64
Hoyer P O 2004 Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research* **5** 1457–1469
Hunter John D. 2007 Matplotlib: A 2D Graphics Environment *Computing in Science & Engineering* **9** 90–95
Hummel T 2021 Die Stunde der Landräte *SZ* **164** S.2
Ionita M and Nagavciuc V 2021 Extreme floods in the eastern part of europe: large-scale drivers and associated impacts *Water* **13** 1122
IPCC 2021 *AR6. Summary for Policymakers.-Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge: Cambridge University Press)
Irving K, Kuemmerlen M, Kiesel J, Kakouei K, Domisch S and Jähnig SC 2018 A high-resolution streamflow and hydrological metrics dataset for ecological modeling using a regression model *Scientific data* **5** 180224

Jüpner R 2016 *Hochwasserrisikomanagement 2030—ein Ausblick. - Jüpner R, Müller U (Hrsg) Tagungsband zur 8. Veranstaltung des Forums zur EG-HWRM-RL 23.06.2016 in Mainz. Ber. Forums z. Europ. Hochwasserrisikomanagement-Richtlinie* (Aachen: Shaker Verlag) 113–21
Jüpner R *et al* 2018 Resilienz im Hochwasserrisikomanagement.- KorrespondenzWasserwirtschaft *Heft* **11** 656–63
Kahle M 2021a Newspaper Articles covering Floods in the Ahr Valley (Germany) in 1804 and 1910
Kahle M 2021b Newspaper Articles covering Floods in the Ahr Valley (Germany) in 2021
Kahle M 2022 News whisperer/floods2021: News-Harvester and Classifier (v1.0.0) (https://doi.org/10.5281/zenodo.6406244)
Kang Y and Park C-S 2018 A multi-risk approach to climate change adaptation, based on an analysis of south korean newspaper articles *Sustainability* **10** 5
Kim D-Y and Kang S-W 2018 Analysis of Recognition of ClimateChanges using Word2Vec *Int. J. of Pure and Applied Mathematics* **120** 5793–807
Kirilenko Andrei P. and Stepchenkova Svetlana O. 2012 Climate change discourse in mass media: application of computer-assisted content analysis *Journal of Environmental Studies and Sciences* **2** 178-191
Knuth H and Parnack C 2021 Von unbezahlbarem Wert *Die Zeit* **31** S. 9
Kohl M 2007 *Ahrhochwasser 1804.- Heimatjahrbuch des Kreises Ahrweiler* **64** 161–4
Kohonen T 2001 Self-Organizing Maps *Springer Berlin S* (*SSINF* 30) (Berlin: Springer) 1-501978-3-642-56927-2
KPF F F U and HAK 2021 Wenn der Hang ins Rutschen gerät *SZ* **163** S.2
Kreienkamp F *et al* 2021 Rapid attribution of heavy rainfall events leading to the severe flooding in Western Europe during July 2021.- World Weather Attribution *Scientific Report* **53** 1-51 (https://worldweatherattribution.org/wp-content/uploads/Scientific-report-Western-Europe-floods-2021-attribution.pdf)
Krieger F 2021 Welche Versicherung zahlt? *SZ* **161** S.17
LeCuan Y, Leon B, Bengio Y and Haffner P 1998 Gradient based learning applied to document recognition *Proc. IEEE* **86** 2278–324
Lin Y, Bates J and Goodale P 2016 Co-observing the weather, co-predicting the climate: Human factors in building infrastructures for crowdsourced data *Science and Technology Studies* **29** 10–27 (https://sciencetechnologystudies.journal.fi/article/view/59199)
Mahdisoltani F, Biega J and Suchanek F M 2013 $YAGO_3$: a knowledge base from multilingual wikipedias *CIDR, Jan 2013, Asilomar, United States* (https://hal-imt.archives-ouvertes.fr/hal-01699874)
Merz B *et al* (ed) 2011 Management von Hochwasserrisiken *Stuttgart* (Stuttgart: Schweizerbart) 978-3-510-65268-6
Met Office 2015 Cartopy: a cartographic python library with a Matplotlib interface (https://scitools.org.uk/cartopy)
Mikolov Tomas, Chen Kai, Corrado Greg and Dean Jeffrey 2013 Efficient Estimation of Word Representations in Vector Space (https://arxiv.org/abs/1301.3781) 1301.3781
Moraru A, Pavlíček M, Bruland O and Rüther N 2021 The Story of a Steep River: Causes and Effects of the Flash Flood on 24 July 2017 in Western Norway *Water* **13** 1688
Müller-Arnod B and Stegemann J 2021 Die Hochwasserkatastrophe. Hilflos zwischen Trümmern *SZ* **162** S.2
NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team 2021 (https://search.earthdata.nasa.gov/downloads/9635679842?p=C1711961296-LPCLOUD!C1711961296-LPCLOUD&pg[1][v]=t&pg[1][gsk]=-tart_date&pg[1][m]=download&q=aster&sb[0]=4.76367%2C49.02887%2C8.05078%2C51.84298&tl=1629358537!3!!)
NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team 2019 *ASTER Global Digital Elevation Model V003* distributed by NASA EOSDIS Land Processes DAAC
Niewel G 2021 Gute Nachrichten *SZ* **222** S. 40
Niforatos E, Vourvopoulos A and Langheinrich M 2017 Understanding the potential of human machine crowdsourcing for weather data *International Journal of Human Computer Studies* **102** 54-68
Open Data 2021 DWD open data portal (https://opendata.dwd.de/, last accessed 19th of August 2021)
Pedregosa *et al* 2011 Scikit-learn: machine learning in python *Journal of Machine Learning Research* **12** 2825–30
REUTERS 2021 Zahlungen nach Hochwasser *SZ* **223** S.16
Reuters L 2021 Für eine wichtige Durchsage *SZ* **195** S.27
Richardson L 2020 Beautiful Soup 4.9.3 (https://crummy.com/software/BeautifulSoup/)
Riemann D, Glaser R, Kahle M and Vogt S 2016 The CRE tambora.org—new data and tools for collaborative research in climate and environmental history *Geoscience Data Journal* **2** 63-77
Roggenkamp T and Herget J 2014 *Reconstructing Peak Discharges of Historic Floods of the River Ahr* (Germany: Erdkunde) 68, 49–59
Rühle A 2021 Klimawandel, das war immer woanders *SZ* **165** S. 11
SAGA GIS 2021 (System for Automated Geoscientific Analyses version 7.8.2 (http://saga-gis.org last accessed 19th of August)
Saura J R, Ribeiro-Soriano D and Palacios-Marqués D 2021a Using data mining techniques to explore security issues in smart living environments in Twitter *Computer Communication* **179** 285–95
Saura J R, Ribeiro-Soriano D and Palacios-Marqués D 2021b Setting privacy 'by default' in social IoT: Theorizing the challenges and directions *Big Data Research, Big Data Research* **25** 100245
Scheuer N 2021 In den Trümmern von Kall *SZ* **164** S.13
Schmidhuber J 2015 Deep learning in neural networks: an overview *Neural Networks* **61** 85–117
Schmidt A, Ivanova A and Schaefer M S 2013 Media attention for climate change around the world: a comparative analysis of newspaper coverage in 27 countries *Global Environ. Change* **23** 1233–48
Schneider J 2021 Profil. Armin Schuster. Bedrängter Chef des Amtes für Bevölkerungsschutz *SZ* **165** S.4
Schwarz E 2021 Wir leben *SZ* **175** S.3
Seel K A 1983 *Die Ahr und ihre Hochwässer in alten Quellen.- Heimatjahrbuch des Kreises Ahrweiler* **40** 91–102
Stagge J H, Tallaksen L M, Kohn I, Stahl K and van Loon A F 2013 *A European Drought Reference Database: Design and Online Implementation* Universitetet i Oslo, Norway (UiO); Albert-Ludwigs-Universität Freiburg, Germany (ALU-FR); Wageningen Universiteit (WU) (https://edepot.wur.nl/291635)
Stahl K *et al* 2016 Impacts of European drought events. Insights from an international database of text-based reports *Nat Hazards Earth Syst Sci Discuss* **3** 801–19
Steinke R 2021 Lügen aus dem Lautsprecher *SZ* **173** S.6
Sudhaus D, Seidel J, Bürger K, Dostal P, Imbery F, Mayer H, Glaser R and Konold W 2008a Discharges of past flood events based on historical river profiles hydrol *Earth Syst. Sci. Discuss.* **5** 323–44
Sudhaus D, Seidel J, Bürger K, Dostal P, Imbery F, Mayer H, Glaser R and Konold W 2008b *Determing Discharges of Past Flood Events Using Historical River Profiles Hydrol Earth Syst. Sc* **12** 1201–9

Thieken A, Kemter M, Vorogushyn S, Berghäuser L, Sieg T, Natho S, Mohor G S, Petrow T, Merz B and Bronstert A 2021 *Extreme Hochwasser bleiben trotz integriertem Risikomanagement eine Herausforderung* GFZ Potsdam, PIK Potsdam NatRiskChange (https://uni-potsdam.de/fileadmin/projects/natriskchange/Taskforces/Flut2021_StatementThiekenEtAl.pdf)

Wartena C 2019 A probabilistic morphology model for German lemmatization *Proc. of the 15th Conf. on Natural Language Processing (KONVENS 2019) Long Papers Pp 40–49 Erlangen* (https://doi.org/10.25968/opus-1527)

WasserBLIcK/BfG & Zuständige Behörden der Länder 2021 datasource der Oberflächengewässer (https://geoportal.bafg.de/inspire/download/HY/waterbody/datasetfeed.xml last access: 18 August 2021)

Weiss M 2021 Bis die Bilder verblassen *SZ* **168** S.1

Yzaguirre A, Smit M and Warren R 2016 Newspaper archives + text mining = rich sources of historical geo-spatial data *IOP Conf. Ser.: Earth Environ. Sci.* **34** 1–8

Zhang H 2004 The optimality of naive bayes *FLAIRS2004 conference*

Zisgen Julia, Kern Julia, Thom Dennis and Ertl Thomas 2014 #Hochwasser – Visuelle Analyse von Social Media im Bevölkerungsschutz / #Hochwasser – Using Visual Analytics of social media in civil protection *i-com* **13** 37–44