

Response to Referee#2

Manuscript title: **From text to geoinformation – A modular approach for extraction of disaster information from web text data**

RC2: '[Comment on egusphere-2026-2361](#)', Anonymous Referee #2, 16 Aug 2026

Referee comments are written in black italics; author's responses are written in blue and normal font; and proposed changes are highlighted in red italics, Times New Roman.

We would like to sincerely thank the referee for their time and valuable feedback and careful reading of the manuscript. The comments provided have been instrumental in improving the clarity and rigor of the manuscript. We will revise the manuscript accordingly and address all your comments below. Please find attached our detailed responses and proposed changes.

If you have any further questions, we would be happy to discuss them.

Kind regards,

Vanessa Rittlinger on behalf of all Co-authors

This manuscript presents a modular workflow for extracting and geolocating disaster-relevant information from web text data, using the 2021 flood event in western Germany as a case study. The topic is relevant to disaster information extraction and situational awareness, and the integration of text mining with geospatial analysis is potentially useful. However, the methodological contribution and validation of the proposed workflow need to be strengthened before the manuscript can be considered for publication. The following issues should be addressed.

RC2 Comment 1. *The workflow mainly combines existing components, including dictionary-based extraction, Flair NER, gazetteer/Nominatim geocoding, and rule-based entity linking. What is the main methodological novelty of this study compared with existing disaster-related geoparsing and information extraction approaches?*

RC2 Answer Comment 1:

We thank the referee for prompting us to articulate the methodological contribution more precisely. We agree that the individual components of our pipeline, dictionary-based extraction, Flair NER, gazetteer and Nominatim geocoding, and rule-based entity linking, are established methods. Our novelty does not lie in any single algorithmic advance but in two interrelated aspects that, to the best of our knowledge, have not yet been combined in the disaster-information-extraction literature. First, we shift the unit of analysis from the document or topic level, which dominates existing GDELT-based and crisis-NLP studies, to the entity level: each extracted record links a specific toponym to a specific disaster-relevant keyword and its disaster-cycle class within a single sentence, yielding

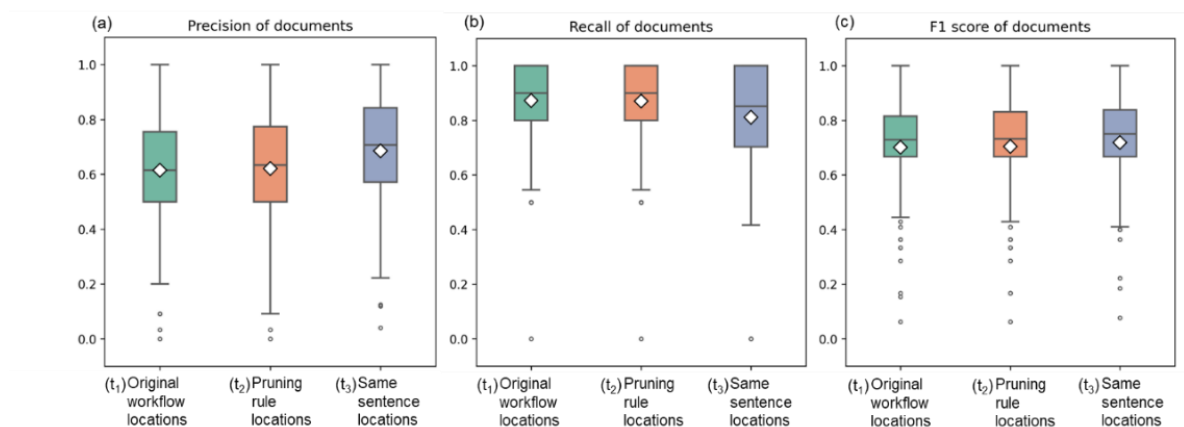
coordinates rather than article-level labels. Second, we target full-length, formally structured web news articles rather than short-form social media posts, which requires handling longer contexts, higher lexical variation, and multi-level administrative references that short-text pipelines do not encounter. The architectural integration of these steps into a single, configurable, end-to-end workflow is the focus. In the revised manuscript, this positioning will be stated explicitly in the final paragraph of the Introduction and revisited in greater detail in the discussion.

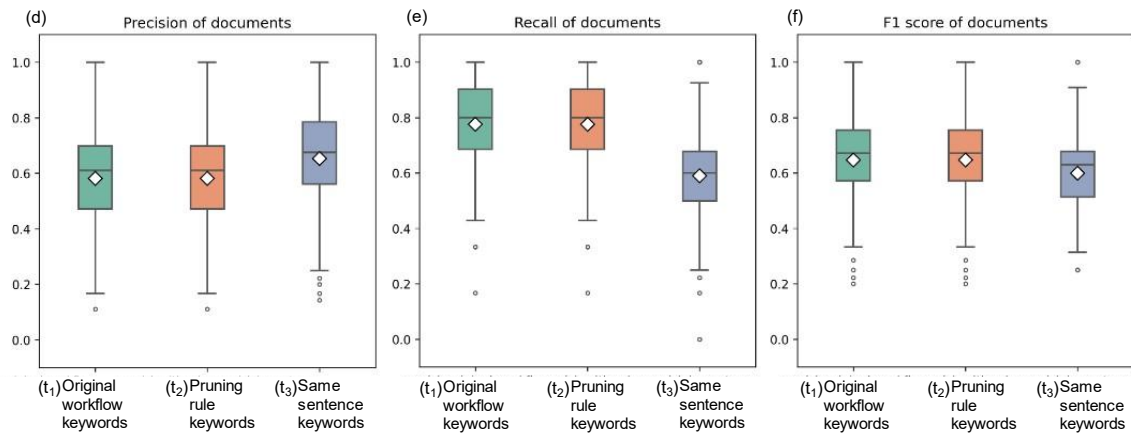
RC2 Comment 2. *The evaluation focuses mainly on the final toponym–DRI relation pairs, while errors from DRI extraction, toponym recognition, and geocoding are not evaluated separately. It would be useful to report the performance of the major modules individually.*

RC2 Answer Comment 2:

We agree with the referee’s recommendation to add the evaluation metrics of the toponym and DRI identification. Individual module performance metrics for DRI extraction and toponym recognition will be calculated and added to the text, as well as the referring plots added to the appendix. These results are expected to demonstrate that foundational components (dictionary-based keyword identification and Flair NER) achieve substantially higher F1-scores than the final relational linking stage. The integrated evaluation remains central to assessing end-to-end utility, but the disaggregated metrics provide clearer insight into error propagation and optimization priorities. *We propose to add this text to the section 4.1: “Additionally, we conducted a separate evaluation of the individual entity extraction components, toponym and DRI identification, on the same reference dataset, assessing their precision, recall, and F1 scores independently. Showing overall high results from the Toponym identification model FlairNER and from the dictionary based DRI information identification. These results are presented in the appendix to provide a comprehensive view of the system’s performance at the entity level.”*

And the following plots to the appendix:





Appendix A1: Evaluation metrics separately for toponym (top) and flood disaster keyword dictionary (bottom)

RC2 Comment 3. *The F1 scores of the three linking strategies are relatively low, ranging from 33.4% to 42.6%. How does this level of accuracy affect the reliability of the subsequent spatial analysis and its potential use for disaster management?*

RC2 Answer Comment 3:

We thank the referee for this critical assessment of the evaluation. We are aware that this F1 score indicates a moderate level of performance and would be considered low in many other domains (Detroja et al., 2023; Nasar et al., 2021). Our approach, however, addresses the novel challenge of simultaneously performing domain-specific entity recognition, context-based disambiguation, spatial inferences and precise coordinate alignment, which, to the best of our knowledge, has not yet been achieved to a comparable extent in the literature. We have therefore established a baseline, and regard this paper as an introduction to a method that provides actionable insights for disaster managers and lays the groundwork for future developments. That being said, we are aware that, given these F1 score values, this data should not be taken as a standalone data source for making statements in a disaster context. Nor should it be used for detailed, well-founded geospatial analyses based solely on this data set. For applications in disaster management, the workflow is positioned more as a supplementary screening tool than as a definitive verification system. Results with moderate accuracy are valuable for flagging potential clusters of incidents, guiding targeted resource allocation, or identifying areas requiring verification via satellite or field sensors. We will highlight this added value and our intentions more clearly in the new publication, examine the results more critically and further substantiate them through scientific references.

RC2 Comment 4. *The same-sentence rule (t_3) achieves the highest F1 score, but the authors finally use the pruning rule (t_2). Please provide a clearer quantitative justification for this choice.*

RC2 Answer Comment 4:

We thank the referee for this observation. We will update our manuscript with an explicit justification for adopting the t_2 pruning strategy. Our selection criterion prioritized balanced performance across precision, recall, and F1-measure rather than maximizing a single metric. Rule t_2 yielded moderate, stable values across all three indicators, minimizing false positives while maintaining adequate recall for downstream spatial aggregation. Rule t_3 , despite a higher peak F1, exhibited greater variance and inflated precision driven by restricted syntactic proximity, which artificially limited valid relation discovery in longer descriptive passages.

In section 4.1 we propose to add: “Given t_2 ’s moderate performance across all three metrics and its recall, precision and F1 score values, which all fall within the top three ranges, we opted for t_2 rather than t_3 , which shows significantly lower recall values.”

RC2 Comment 5. *The evaluation of geocoding mainly reports whether a toponym can be matched to a location. However, successful matching does not necessarily mean correct geocoding, as illustrated by the erroneous allocation of “Qatar” to an embassy in Bonn. Geocoding accuracy should be evaluated against manually verified locations.*

RC2 Answer Comment 5:

We acknowledge this important limitation raised by the referee and implement an additional accuracy evaluation of the geocoding. We manually cross-verified the random subset of 10% toponyms against authoritative administrative gazetteers and OpenStreetMap, checked in the documents for the true location mentioned and measured the haversine distance and accuracy@161km and accuracy@15km distance from the original place (Gritta et al., 2020). We propose to include this geocoding accuracy evaluation in the paragraph addressing geocoding in section 5.1.

“Our geoparsing returns location information detailed up to city districts and specific localities (like Nürburgring). Nevertheless, it has uncertainties as toponyms can be missed and words can be incorrectly recognized as toponyms (Gritta et al., 2018b; Hu et al., 2024). Overall, 1 356 unique toponyms were not recognised by the NER model (e.g., “Großraum Ahrweiler”, the suffix-type “Eifeldorfs”, or fine-grained localities such as “Felberntauerstraße”). Consequently, these entities could not be included in the final DRI dataset. Conversely, some non-toponyms were falsely identified and passed to the geocoder, producing missing or incorrect coordinates and incorrect ambiguity resolution (Gritta et al., 2018a). Our approach first matches toponyms with an official geographic database and then passes unmatched toponyms to the Nominatim geocoding service, using high-quality local public reference information for places that can be matched to this database (Acheson et al., 2017). Testing each

single database match results, we extracted a random sample of 230 toponyms (10% of the unique toponyms) different numbers per database are matched. Within the national local database alone, 35 places are not found, with the global open-source datasets "populated places" and GADM, 208 and 88 respectively are not found. Using solely Nominatim, with a parameter set to the study area, 8 toponyms cannot be found. However, it commits some notably erroneous allocations, like Qatar as the embassy of Qatar in Bonn. By merging all the databases and carrying out successive comparisons, we were able to assign all 230 place names to the sample. Furthermore, we verified the accuracy of the geocoding by manually cross-referencing 10 per cent of the geocoded place names with their original longitude and latitude coordinates as intended in the text (Gritta et al., 2020). The accuracy @161km (Gritta et al., 2020) measurements is 85 per cent. When comparing locations within a 15km distance, we achieved an accuracy of 81 per cent. This is due, amongst other things, to incorrect disambiguation of places with the same name, the assignment of places to the same administrative unit (such as a town or district), or places being assigned to units of the same name, such as bodies of water. We would also like to draw attention to the resulting distortion in the assignment of administrative units, which stems from incorrectly geocoded toponyms."

RC2 Comment 6. *The temporal analysis uses the publication or registration date of web documents rather than the actual time of the reported disaster information. Therefore, Figure 7 may represent media attention dynamics rather than the temporal evolution of the flood event. This distinction should be clarified.*

RC2 Answer Comment 6:

We appreciate the referee's suggestion to improve the timestamp clarification. We agree that the timestamp attribute reflects media publication timelines rather than hydrological event chronology. In the revised manuscript, Figure 7 and associated analyses are explicitly labelled as representing "reported media activity timelines," and the methodology section defines the time attribute and state the information given more clearly.

Our proposed modified text in section 4.2: "Figure 7 shows the temporal trends for all thematic categories, as mentioned in news articles published on the web. The figures are shown in absolute terms throughout the disaster phase (dark grey), defined here as the period during which the rainfall and flooding occurred, and the subsequent recovery phase, during which intensive emergency and damage-mitigation measures were taken (light grey)."

Figure 1: Media attention dynamics of DRI information from the text at the time the text information was registered in the GDELT database, with the 9 disaster-cycle classes aggregated to five classes over time

RC2 Comment 7. *The comparison with CEMS building-damage data in Section 4.4 is mainly visual. A quantitative spatial comparison, such as correlation, hotspot agreement, or affected-area classification, would provide stronger evidence for the usefulness of the extracted information.*

RC2 Answer Comment 7:

We agree with the referee that quantitative validation forms the basis for comparative conclusions. However, we see a limitation in quantitative spatial comparisons (such as a correlation) of different absolute values from the two data products. Comparing the absolute values of damage data for houses and the extent of this damage in the flood-affected areas recorded by CEMS with the absolute values of damage data at local authority level, which do not fall within a comparable range or distribution of values, which makes direct comparisons without transformations extremely difficult and potentially misleading. Therefore, in the revised manuscript, we will describe the hotspots mentioned previously in Section 4.4 in more detail with regard to the hotspot-matching analysis, and calculate the proportion of overlap between matching locations and the differences (total absolute number of locations mentioned in the study area versus overlaps between the two datasets). To enable readers to make a visual comparison, in the revised manuscript we will present a bivariate spatial overlay analysis between the extracted DRI clusters and the damage polygons from the Copernicus Emergency Management Service (CEMS) (formerly in the Appendix). To this end, we will compare a categorisation of the respective value ranges, grouped by quantiles. Such a visual analysis compels the viewer to take the geographical context and patterns into account, which is why we argue that it is preferable as the primary tool for presenting our results.

We propose to change this paragraph in section 4.4: “A bivariate representation of damage information at local authority level – normalized per 1,000 inhabitants and categorized by quantiles – demonstrates how DRI can complement geospatial data on an event. We have chosen this approach because we must bear in mind that, with the CEMS damage data, we prioritize data at the smallest spatial level within a defined area, which we then compare with mentions of damage information at the municipal level. However, this enables us to provide information on the frequency of incidents for municipalities across both data sources, such as in the Ahr Valley and Erfstedt. With the exception of text-based data, most mentions occur within the districts that declared a state of emergency, i.e. within our study area. In both cases, the results of the text data analysis, however, encompass not only damage information relating to various building types according to our classification, but also to bridges and other damaged infrastructure components. The absolute and spatial distribution of the two different datasets is shown in Appendix3.”

RC2 Comment 8. *Toponyms referring to municipalities, counties, states, rivers, and specific facilities are all represented as points. Using administrative centroids may introduce artificial spatial precision and could influence the geographic-centre analysis in Figure 9. This uncertainty should be considered.*

RC2 Answer Comment 8:

We are pleased that the referee highlighted this aspect, which was a point of substantial discussion also within our team. We recognize that point-based representation of a variation administrative features introduces positional approximations. Despite their imperfect precision, centroids were utilized as the default geocoding target due to standardized availability in all of our integrated gazetteers. The resulting centroid-approximation uncertainty will explicitly be mentioned in the in the relevant section 4.3 Spatial dissemination of information. *"This spatial shift is subject to uncertainties, as our data assigns administrative units to the coordinates of their geographical centre of gravity; this is a necessary abstraction that does not correspond to reality and may not be suitable for every application, leading to an approximation of the change in the geographical centre of gravity."*

RC2 Comment 9. *The manually annotated reference dataset contains 122 documents and 5,087 toponym–DRI pairs. Please provide more information about the annotation procedure, number of annotators, annotation guidelines, and inter-annotator agreement.*

RC2 Answer Comment 9:

We thank the referee for encouraging us to add these methodological details. As requested, we will provide further information on the annotation process and amend the manuscript accordingly in Section 3.3. Due to the subject-specific requirements and the limited pool of qualified annotators, the annotation was carried out by a single trained specialist. As only one annotator was involved, it was not possible to calculate inter-annotator agreement. We acknowledge this as a limitation and recommend a second independent annotation for validation in future. The annotation process was carried out using the Doccano annotation tool (doccano, 2026) for both entity tagging and entity linking.

We propose to change the last paragraph in section 3.3 to the following: "To assess the validity of our workflow, we created a manually annotated reference dataset. We use a random sampling method to ensure that all classes are covered in the reference corpus (Thompson, 2012). In total, 122 documents were randomly selected in this way and subsequently annotated manually, resulting in 5,087 toponym-DRI pairs. The toponym and DRI entities are annotated individually, and a relational link is then established between the toponym and the associated DRI information. The annotation process is carried out using the Doccano annotation tool (doccano, 2026). The resulting reference corpus comprises relationship pairs with a correct identification of the toponym, the DRI and the associated disaster cycle class, as well as a correct link between the toponym and the DRI. This evaluation, which focuses on the entity

pairs, is similar to the approach presented in Schiersch et al. (2020). We use the statistical metrics of precision, recall and F1-score (Warto et al., 2024). In the case of recall, false-negative results are defined as missing, labelled toponyms and their missing links to DRIs. The evaluation focuses on the correct identification of toponyms and the associated DRIs, which are also classified into the nine predefined disaster cycle classes. Errors or biases arising from the earlier steps (toponym identification, geocoding of toponyms and DRI identification) are addressed in the discussion. “

RC2 Comment 10. *The disaster-relevant dictionary is a core component of the method, but its construction is not sufficiently reproducible. Please clarify the TF-IDF selection procedure, manual keyword addition, suffix expansion, and preferably provide the complete dictionary as supplementary material.*

RC2 Answer Comment 10:

We appreciate the referee’s emphasis on reproducibility and clarification on the disaster-relevant dictionary, which undeniably is a crucial input to our approach. We intend to amend the text to facilitate a better understanding of the dictionary, specifically our application of the TF-IDF method to pre-selected German reports and documents on flooding and heavy rainfall from the official websites of the BKG or regional authorities. From the keywords extracted from the documents, the experts then selected those relevant to the pipeline for inclusion in the dictionary, taking into account suffix variants of the words. If required, we can make the full disaster-related dictionary available in the appendix.

We propose to adjust the first paragraph of section 3.1 to the following: “*Our aim is to identify flood-related keywords and categorise them according to the disaster cycle and the progression of an event (initial phase, response phase, mitigation and adaptation measures) (Bosher et al., 2021) (Table 1). A key challenge in this process is the lack of an existing, subject-specific and language-optimised dictionary for German terms relating to the disaster cycle and progression. As the identification of relevant key terms is standard practice in the context of crisis management, we have adopted a pragmatic approach to create a robust and practical keyword dictionary. As part of our dictionary compilation process, we utilise a corpus compiled by our subject matter experts, consisting of German reports and documents on flooding, flood events and heavy rainfall sourced from the official websites of the BKG or regional authorities. To identify relevant terms and keywords from these documents, as well as from additional expert knowledge, we apply a TF-IDF (Term Frequency–Inverse Document Frequency) approach (Senaratne et al., 2023), which filters out relevant words and keywords. As our database consists of documents in German, we must take into account variations in word suffixes depending on gender, case and singular or plural form (Kharis et al., 2021). We have opted for an approach in which we include variations of the keywords in the dictionary based on their lemma, rather than lemmatising all words and the entire text (Table 1). “*

RC2 Comment 11. *The method is evaluated only for one flood event and German-language web texts. The claims regarding scalability and transferability should therefore be moderated, or an additional event or study area should be used for independent validation.*

RC2 Answer Comment 11:

We agree with the referee that, with regard to transferability and scalability, an additional event or study area is required. In response, we will moderate our claims but also include a second validation case study: the flood event of May 2024 in Bavaria (southern Germany). Applying the (identical) pipeline to this temporally and spatially independent case yields consistent extraction patterns and confirms cross-regional generalisability. We will incorporate the results of the additional case study into the revised manuscript.

We propose to add a section 3.4 *Transferability study*

To assess the transferability of our pipeline, we have also applied it to another flood event: The 2024 floods in southern Germany (Bayerisches Landesamt für Umwelt (LfU), 2025). To this end, we have likewise extracted web text data from a collection of GDELT news data covering the period from 29th May 2024 to 13th June 2024 of the event and presented the results. We therefore analyse 1125 documents for this period and focus area.

We propose to add a section 4.5 *Transferability assessment*

Our method is applied to another flood event in Germany. To this end, we extracted data from the GDELT database, processed it using the same procedure, and fed it into the pipeline. For this event we can extract in total 41590 DRI points. As can be seen from Figure 12, information on damage reports and the disaster itself can also be extracted from the texts relating to this event, providing details of the damage and the course of events in the affected area as well as during the flood period. These are particularly high in areas with large inundated areas, but also show correlations in areas with smaller inundated areas. However, not all areas (grey/reddish) are mentioned in our text data in connection with the flood.

Aggregation of the absolute number of media mentions of infrastructure damage per municipality between 29 May 2024 and 13 June 2024

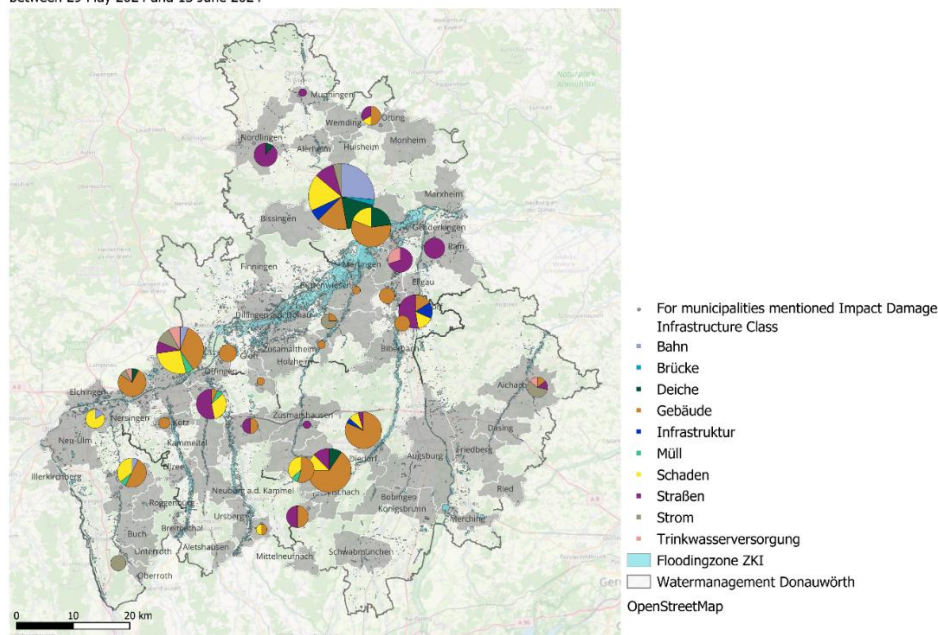


Figure 12: Map extract showing Donauwörth during the flood event in June 2024, illustrating the results of the workflow in a further region. At local authority level, the aggregated mentions of damage to infrastructure in web news reports for the event period are presented; these are broken down by the DRI damage data and the number of mentions in the pie chart, and visualised according to scale.

RC2 Comment 12. *GDELT-based news data may contain duplicate reports and uneven media coverage. Were duplicated or syndicated articles removed? The authors should also discuss whether the spatial distribution of DRI reflects actual disaster impacts or, partly, differences in media attention.*

RC2 Answer Comment 12:

Whilst our deduplication process removes exact duplicates (identical URLs and dates), we have retained articles with unique URLs, date and text, even if their content may overlap, as this provides a closer approximation of the actual volume of media coverage and the breadth of reporting on the event. We acknowledge that this means the DRI values sometimes reflect media attention rather than the number of individual events or locations affected. We will fundamentally revise the manuscript to address this aspect, and this limitation is discussed in more detail in Section 5.2. We would like to emphasise that we have not edited the text data from the documents; that is to say, no text has been altered. Textual similarities and their context, including with regard to media coverage, differences or the significance of certain semantic clusters, would certainly be extremely interesting to discuss; however, the many potential dimensions of texts that would need to be taken into account would go beyond the scope of this analysis. In Section 2.2, ‘Web Text Dataset’, we describe the data filtering process in greater detail, and the following filtering information is added: *‘In this dataset, duplicates are filtered out based on their text content, website URLs and dates.’* In Section 5.2, ‘Quality of Data Sources, Biases and

Misinformation’, we shall revisit the limitations associated with using information from web texts. Furthermore, we view the comparison of our data analysis with other geospatial data relating to the event as an opportunity to demonstrate the potential of this type of data. In the revised version of the manuscript, we therefore emphasise that we do not regard this dataset as a standalone data source capable of capturing the entire event, but merely as an additional data source that can provide supplementary information to the existing remote sensing data.

RC2 Comment 13. *The manuscript repeatedly emphasizes potential applications in situational awareness and disaster management. However, the case study is mainly a retrospective analysis. The authors should clarify whether the proposed workflow is intended for real-time operational use or primarily for post-event analysis.*

RC2 Answer Comment 13:

We would like to thank the referee for drawing our attention to this lack of distinction. In the revised manuscript, we will emphasise that this study employs a retrospective analytical approach to assess the feasibility of the workflow and the quality of data extraction, rather than testing operational performance in real time. Nevertheless, we will emphasize a potential operational use of the data after further research. Indeed, the modular architecture and automated processing steps inherently support near-real-time deployment, as confirmed by GDELT’s 15-minute update cycle. The revised text in section 5.3 will explicitly distinguish retrospective validation from prospective operational integration and outline the necessary infrastructural adjustments for live monitoring environments.

We therefore propose to rewrite the last part in the section 5.3: *“We carried out a retrospective analysis to test the process and methodology using an event that had already occurred. The approach provides information at specific spatial levels and enables the mapping of additional attributes alongside satellite-based data. However, we also recognise that the methodology holds potential for near-real-time application in disaster situations and for use as a supplementary data source to integrate news streams emerging in parallel with the course of the event into the analysis.”*

References

- Acheson, E., De Sabbata, S., and Purves, R. S.: A quantitative analysis of global gazetteers: Patterns of coverage for common feature types, *Comput. Environ. Urban Syst.*, 64, 309–320, <https://doi.org/10.1016/j.compenvurbsys.2017.03.007>, 2017.
- Akbik, A., Blythe, D., and Vollgraf, R.: Contextual String Embeddings for Sequence Labeling, in: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, 1638–1649*, 2018.
- Bayerisches Landesamt für Umwelt (LfU): Hochwasser Mai/Juni 2024 - Wasserwirtschaftlicher Bericht, 2025.

- Bosher, L., Chmutina, K., and Niekerk, D.: Stop going around in circles: towards a reconceptualisation of disaster risk management phases, *Disaster Prev. Manag. Int. J.*, ahead-of-print, <https://doi.org/10.1108/DPM-03-2021-0071>, 2021.
- doccano: GitHub - doccano/doccano: Open source annotation tool for machine learning practitioners., <https://github.com/doccano/doccano>, last access: 24 August 2026., 2026.
- Gritta, M., Pilehvar, M. T., and Collier, N.: A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics, *Lang. Resour. Eval.*, 54, 683–712, <https://doi.org/10.1007/s10579-019-09475-3>, 2020.
- Hu, X., Zhou, Z., Li, H., Hu, Y., Gu, F., Kersten, J., Fan, H., and Klan, F.: Location Reference Recognition from Texts: A Survey and Comparison, *ACM Comput. Surv.*, 56, 1–37, <https://doi.org/10.1145/3625819>, 2024.
- Kharis, M., Kisyani, Suhartono, Pairin, U., and Darni: How to Lemmatize German Words with NLP-Spacy Lemmatizer?., *International Seminar on Language, Education, and Culture (ISoLEC 2021)*, Malang, Indonesia, <https://doi.org/10.2991/assehr.k.211212.036>, 2021.
- Schiersch, M., Mironova, V., Schmitt, M., Thomas, P., Gabryszak, A., and Hennig, L.: A German Corpus for Fine-Grained Named Entity Recognition and Relation Extraction of Traffic and Industry Events, <https://doi.org/10.48550/arXiv.2004.03283>, 7 April 2020., 2020.
- Senaratne, H., Mühlbauer, M., Götzer, S., Riedlinger, T., and Taubenböck, H.: Detecting crisis events from unstructured text data using signal words as crisis determinants, *Int. J. Digit. Earth*, 16, 4601–4620, <https://doi.org/10.1080/17538947.2023.2278714>, 2023.
- Thompson, S. K.: *Sampling*, John Wiley & Sons, Incorporated, Newark, UNITED STATES, 2012.
- Warto, Rustad, S., Shidik, G. F., Noersasongko, E., Purwanto, Muljono, and Setiadi, D. R. I. M.: Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application, *Stat. Optim. Inf. Comput.*, 12, 907–942, <https://doi.org/10.19139/soic-2310-5070-1631>, 2024.