Wikimpacts 1.0: A new global climate impact database based on automated information extraction from Wikipedia

Response to reviewers

February 24, 2026

Contents

1 Christos Giannaros (NHESS Editor)

We would like to thank Dr.Christos Giannaros for his dedicated time in reviewing the manuscript and response letters. According to the following comments and suggestions by the editor, we propose the following changes in the database and manuscript (Notably, all line numbers refer to the revised manuscript):

1. Major revisions following the comments and suggestions by the editor and two reviewers.
2. Provide an explanation for longevity and homogeneity of the dataset.
3. Address the issues of footnotes, typos, citations, and cross references.

Additionally, we are building a website to make the database easier to visualize and access. This tool will be available soon.

Comment 1

Your manuscript "Wikimpacts 1.0: A new global climate impact database based on automated information extraction from Wikipedia" has now been evaluated by two reviewers.
Their detailed feedback underscores the necessity for major revisions before the manuscript can be reconsidered for publication in NHESS.

Response

The manuscript and database have been revised in accordance with the editor's and two reviewers' comments and suggestions, as reflected in the revised manuscript and the responses below.

Comment 2

In addition to the reviewers' comments, I would also like to point out the following concern: The Wikimpacts 1.0 extraction pipeline is sensitive to the "prompt version" used. Is it also sensitive to the specific LLM employed? Specifically, could different versions or models yield varying results for the exact same Wikipedia text input and prompt? Addressing this is crucial for the longevity and homogeneity of the dataset, especially as proprietary models like those in GPT family are subject to very frequent updates or even discontinuation.

Response

We agree with the editor's concern regarding the longevity and homogeneity of the dataset, given that our production pipeline relies on a specific version of a GPT model. The extracted information is sensitive to the *prompt version* used. In SI Section 4, we report evaluation results for different prompt versions; for the production run, we selected the prompt version with the best evaluation performance. The extracted information is also sensitive to the particular LLM employed. Based on our prior experiments with open-source LLMs, their performance is generally worse than GPT models (Li et al., 2024). We do not compare perfor-

mance across different GPT versions in this work. The production run of the Wikimpacts 1.0 database was generated using the state-of-the-art GPT-4o model available in 2024.

For future updates, we aim to fine-tune open-source LLMs rather than relying on commercial models for live updates. We also plan to compare extractions across multiple models and only incorporate information into the database when multiple models produce consistent results, thereby improving the longevity and homogeneity of the dataset.

Here, we add the following text in Section 6 (L669-672): Moreover, the extracted information is sensitive to the specific LLM employed. The full Wikimpacts 1.0 database was generated using the state-of-the-art GPT-4o model available in 2024. The GPT family has continued to evolve since then; however, newer GPT versions are not evaluated in this work because we aim to fine-tune the open source LLMs for the future live updates.

Comment 3

Also, some minor issues:

1. Footnotes: According to NHESS guidelines, footnotes should generally be avoided in the text. Please review whether they can be removed or at least reduced.
2. Typos and text issues: Fix the broken text "Appendix List ??" appearing at Line 444, and ensure that any other potential typos are corrected, including in the SI (e.g., "Conatins" in Table S3).
3. Citations: Verify the formatting for the citation "(UNISDR, n.d.)" at Line 44 and "(Antweiler)" at Line 270, as both lack a year. Check also the reference list for duplicate entries (e.g. Jones et al. (2022a) and (2022b)). Finally, as pointed out by the review file validation, the reference list includes a work âunder reviewâ. Such works can be cited upon submission if available to the reviewers.
4. SI cross-references: SI Section 6 refers to EM-DAT mapping, whereas information about the 2021 European Floods appears to be in Section 1 (Lines 87-88). Also, prompt design (Lines 203-204) is SI Section 4. Overall, a reference to specific SI content (e.g. Table S4 for prompt evaluation) would enhance readability.

Response We address these minor issues one by one as follows:

1. Footnotes: We convert all footnotes citing Wikipedia articles into in-text citations. We also reduce other footnotes wherever the information could be moved into citations or incorporated into the main text. Overall, we reduce the number of footnotes from 28 to 2.

2. Typos: on L508, we rephrase the cross reference to see SI Section 9. Beside that, we correct the typo "Conatins" to Contains in SI Table S4.

3. Citations: We have corrected the citations (UNISDR, n.d.) into (UNISDR, 2024), removed the duplicate entries (Jones et al. (2022a) and (2022b)), and updated the "under review" citation (d'Errico et al., 2020)

4. SI cross-references: we correct and enhance readability of the SI cross-references as follows:

Lines	Corrected SI reference
L110 / L147 / L580	SI Section 1
L172	SI Section 2
L181	SI Section 3
L247 / Table 5 caption	SI Section 4
L333	SI Section 5, Table S6
L348	SI Section 5, Table S7
L350	SI Section 5, Tables S8 and S9
L400	SI Section 5, Table S10
L412	SI Section 5, Table S11
L565	SI Sections 4 and 5
L531	SI Section 6
L141 / L252 / L300 / L339	SI Section 7
L313/ L323 / L618	SI Section 8
L508	SI Section 9

2 Reviewer 1

We would like to thank Reviewer #1 for their dedicated time in reviewing the manuscript and for their useful and constructive suggestions. We implemented the following changes in the database and manuscript:

1. Filter the multi-event article entries in Wikimpacts 1.0.
2. Update the statistics and figures along with the updated database.
3. Add explanation for the location definitions across different levels and interpretation of error scores.
4. Improve the Discussion section.

Many suggestions point to potential additional research, which incited us to rework the manuscript, and which indicate that our study could ignite and inspire new research.

Below, we would like to clarify our changes regarding all comments, which are repeated in grey boxes. Moreover, the following convention is applied to denote modifications that weimplemented in the original manuscript (Notably, all line numbers refer to the revised manuscript): new text.

Comment 1

General assessment
The topic is timely and relevant. A global, open, LLM-based impact database is of clear interest to NHESS readers. The manuscript reads as a data-and-methods paper describing a new dataset, its structure, extraction pipeline, evaluation, and example applications. This fits well within the journal's scope for data-oriented contributions. I recommend major revisions to improve clarity, strengthen the discussion of limitations, and help users understand how to interpret and apply the dataset.

Response

We thank Reviewer #1 for the positive assessment of the study topic and for the thoughtful and constructive review. Below, we carefully address each comment (with some comments addressed jointly) and describe the corresponding changes made to the manuscript.

Comment 2

Regarding event article mapping and potential misclassification.
While inspecting the public database, I noticed several cases where impacts appear to be drawn from broad multi-event Wikipedia pages (e.g. "Tropical cyclones in 2017") even when a dedicated single-event article exists (e.g. "Cyclone Numa"). This can lead (it does actually) to incorrect country lists, the inclusion of non-impacted areas, or duplicate entries for the same event. The current manuscript does not fully explain how cross-references within multi-event articles are handled (for example, when one system contributes to another) nor how potential double-counting or mis-allocation of impacts is prevented. I recommend adding a subsection clarifying the filtering logic, giving examples of typical failure modes, and explaining whether any automatic or rule-based deduplication is applied during consolidation.
Specific comments on the article
Regarding the example of the 2011 European floods, the manuscript notes that the main event was categorised as an extratropical cyclone, although a flood category may be more appropriate. This is a useful illustration of hazard-type ambiguity. It would help if the authors commented briefly on how common such cases are and whether simple rule-based corrections might reduce them.

Response

We agree with the reviewer that there is potential misclassification of event article mapping and event types in the database. During the document selection process, we first distinguished between single-event and multi-event articles. For each event, we then selected the best available article among those classified as describing that event. However, as the reviewer correctly illustrated, some articles still cover multiple events, such as "Tropical cyclones in 2017". In the updated version of the database, we therefore filtered out 195 multi-event article entries on tropical storms/cyclones from Wikimpacts 1.0. For other major event types, such as floods and wildfires, we did not remove articles with plural titles (e.g., 2021 European floods", 2015 Russian wildfires"), because these articles are not cross-linked to separate entries for individual flood or wildfire events. Notably, in Wikimpacts 1.0, we do not process tables or list items within articles; consequently, for articles such as "Tulsa tornadoes of 2017", information contained solely in tables is currently not included in our database. This procedure ensures that the Wikimpacts 1.0 database contains only single-event articles and avoids duplicated information arising from multi-event articles. Although we consider the likelihood of duplicated information within single-event articles to be relatively low, we will nevertheless address and resolve these potential issues in the next version of the database. We address this update in Section 3.1, L197-213: Moreover, within the 3,368 event articles, we identify 195 multi-event article entries on tropical storms/cyclones. To detect these misclassified multi-event articles, we perform a keyword search using ["list", "season", "cyclones", "hurricanes", "typhoons", "tornadoes"]. This procedure yields 191 entries con-

taining the keyword "season" (e.g., 1939 Pacific hurricane season and 2020-21 Australian bushfire season) and 4 entries containing the keyword "cyclones" (e.g., Tropical cyclones in 2013). For other keywords, such as "tornadoes", we find articles like "July 2009 Mid-Atlantic tornadoes" that describe multiple tornadoes but for which no separate Wikipedia articles exist for the individual events. For the remaining keywords, we do not identify any corresponding articles in the Wikimpacts 1.0 database. For other major event types, such as floods and wildfires, we do not remove articles with plural titles (e.g., 2021 European floods, 2015 Russian wildfires), because these articles do not provide information for multiple individual flood or wildfire events. It is important to note that Wikimpacts 1.0 does not process tables or list items within articles. Consequently, for articles such as "Tulsa tornadoes of 2017", information that appears only in tables is currently excluded from the database. This filtering procedure ensures that the Wikimpacts 1.0 database is restricted to single-event articles and mitigates duplication arising from multi-event entries. The risk of duplicated information within single-event articles is expected to be relatively low, and we plan to verify this more systematically in future updates of the database. In the remainder of the paper, we focus exclusively on those 3,173 events mapped to a single-event article to construct the Wikimpacts 1.0 database.

Along with the updated database, after removing the multi-event articles, we update all database statistics and the figures in the paper. We present the key changes below. For detailed statistical revisions, we kindly refer the reviewer to the revised version of our manuscript.

1. L6-7: Impact data is stored at the event, national, and sub-national levels, covering 2,733 events from 1034 to 2024, with 17,958 national and 32,567 sub-national entries.

2. L10-11: In comparing impact values, 32 out of 181 matched events have identical data for deaths, and 7 of 77 for injuries.

3. L430-431: The Wikimpacts database, version 1.0, encompasses a total of 2,733 events. These correspond to the subset of the 3,173 events, each mapped to a single-event article for which all mandatory fields were completed.

4. L499-503: In total, we have 17,958 such entries, with the US having the highest number (4,049 entries in total). They are followed by the Philippines with 1,176 data entries, Mexico with 1,170 , Japan with 1,006, China with 883, Australia with 511, Canada with 487, and India with 422. Most African countries only procured a limited number of impact data entries, with some exceptions such as Madagascar (208 entries) and Mozambique (112 entries).

5. L508-510: The US leads with 10,894 sub-national level entries, followed by Mexico, Japan, the Philippines, China, Australia, India, and Canada, with 2,384, 2,350, 1,848, 1,305, 1,010, 938, and 701 entries, respectively. Vietnam and Cuba have fewer entries, with 509 and 355 entries, respectively.

6. L516-519: In some countries, such as China, impacts predominantly occur in coastal regions, with Guangdong province having the highest number of data entries at 182, followed by Fujian province with 141 entries. In contrast, in the US, impacts are distributed across the entire country, with Georgia and Delaware having the highest number of entries at 99. In Mexico, a few states have a large number of data entries, like

Acapulco (116 entries).

7. L521-523: Most entries are of events that occurred in the US, with Outer Banks having the highest number of entries (34), followed by East Texas with 17 entries. Notably, Cabo San Lucas in Mexico also has a large number of data entries, with 26 entries.

8. L541-542:In total, the EM-DAT database contains 35,502 impact data entries, whereas the Wikimpacts 1.0 database comprises 31,608 data entries for the same period.

9. L556-560: Of the remaining events, 38 show higher values by 50% or more, and 48 show lower values by 50% or more compares to EM-DAT. In the injury category, 7 out of 77 events perfectly align with EM-DAT values; over one-third of the events exhibit values at least 50% higher than EM-DAT. For the homeless and damage categories, no events display the same impact values. More events in the homeless category have lower values than EM-DAT, while nearly 94% in the damage category show higher values than EM-DAT.

In response to the comment regarding the misclassification of the main event type, we found that 27 articles with titles containing the word "flood", such as "2011 European floods", were classified as extratropical storm/cyclone events. For example, the article "2011 European floods" states, "The 2011 floods in Europe were caused by a series of storms in the fall, including Cyclone Meeno and Tropical Storm Rolf." Consequently, our model considers this event as an extratropical storm/cyclone. In our database, the hazard column lists "flood" for all of these 27 events, except for the article "1999 Blayais Nuclear Power Plant flood". We acknowledge that this ambiguous classification is not an error or a limitation of our model; even human experts find it difficult to classify such events decisively. Therefore, we decided not to apply a keyword- or rule-based approach to modify the database for these events, and we have retained the original classification. Instead, we have revised the text in Section 5.1, lines 479-485:

However, a few entries are present for Africa. Further investigation reveals that the latter may be an ambiguous classification between the main event categories of flood and extratropical cyclone. For example, the event "2011 European floods" (Wikipedia contributors, 2025b) - despite the article's title - also impacted North Africa. The floods were caused by a series of storms, and in our database this main event is categorised as extratropical cyclone. We identify 27 such cases in our database for which our the hazard column lists "flood", except for the article "1999 Blayais Nuclear Power Plant flood" (Wikipedia contributors, 2025a) for which the hazard is "NULL". We argue that this is an example of an unavoidable classification ambiguity and is not an error or a limitation of our extraction pipeline.

Comment 3

Specific comments on the article
Regarding the role of Wikimpacts 1.0 relative to existing datasets, it would help to clarify early in the Introduction whether the authors position this work as complementary to curated impact databases such as EM-DAT and DesInventar, or as an alternative source. It would also be useful to specify for which types of analyses the dataset is most appropriate (for example, global multi-hazard comparisons or exploratory sub-national studies) and which applications require caution (such as completeness-sensitive national loss or other variable accounting).
Coverage and Wikipedia reliance: the requirement that an English Wikipedia article must exist implies notability and language biases. Small-scale or local events, or those in regions with limited Wikipedia activity, are surely under-represented. I suggest adding a short paragraph addressing this and explaining how users should interpret the absence of events. This also helps contextualise the patterns shown in Fig. 4.

Response

We position Wikimpacts as a complementary impact database to EM-DAT and DesInventar, with the multi-level georeferencing being a clear distinctive factor. However, the inclusion criteria and event definitions differ across these databases. As a result, not all events can be directly matched between Wikimpacts and EM-DAT/DesInventar, although a subset of events can be aligned and compared, as illustrated in Fig.10 of the preprint. Wikimpacts is therefore particularly useful as a complementary source of fine-scale impact information, for cross-database impact comparisons, and for exploring sub-national impact and damage functions. For comprehensive national loss estimates, however, the impact information extracted from Wikimpacts should be validated against nationally reported data or other established impact databases.

Similarly, to address the bias associated with English-language Wikipedia, we have added the following to the Introduction from L71:

The Wikimpacts database is designed to complement existing global disaster impact databases such as EM-DAT and DesInventar, notably by providing georeferenced information at sub-national level. Although the inclusion criteria and event definitions differ across databases, matching events by type, date, and location allows us to identify a set of shared events that can be jointly analysed. These overlapping records can support global multi-hazard comparisons and sub-national studies, particularly when combined with climate and other geospatial data. We find relatively few climate event articles in other languages that are not reported in English. Although we rely solely on English Wikipedia articles in the Wikimpacts 1.0 database, the English bias may exist, but it is not the main issue. We acknowledge that regions with limited Wikipedia activity and small or highly localized events are likely to be under-reported. Consequently, the Wikimpacts 1.0 database should not be used in isolation for complete or highly sensitive national loss assessments. For such applications, we recommend benchmarking Wikimpacts against other databases such as EM-DAT and officially reported government statistics.

Comment 4

Specific comments on the article
L1-L3 definitions (Sect. 2 and Sect. 3.3.3): the manuscript would benefit from a clearer and more consistent description of what "location" means at each level. At L80, levels are defined as event (L1), country (L2), and sub-national (L3). Later wording in Sect. 3.3.3 is less precise. Restating the definitions once, with consistent terminology, will help users interpret the later evaluation of location accuracy.

Response

The "location" definition is different across 3 levels; therefore, we would prefer to keep the technical definition table in the main text without moving it into the Supplementary Information as Reviewer #1 suggested in comment 14. Table 1 in preprint illustrates the definitions of each location-related items in the database and corresponding levels. To make the user better interpret the definition of location-related information, we add more explanation in Section 2, without additional information in Section 3.3.3.

From L101 in Section 2 (Database Structure), we add The location-related information is summarized in Table 1. For L1 and L2, the locations are specified in Administrative_Areas_Norm, which contains a list of affected countries. For L3, the location information comprises both Administrative_Areas_Norm and Locations_Norm, indicating one affected country and a list of affected sub-locations such as cities. In the remainder of the text, we use the general term "location" or "locations" to refer to both these fields.

Comment 5

Specific comments on the article
Hazard types (Sect. 3.4): the exclusion of events that cannot be mapped to the seven main hazard categories, such as landslides, should be made explicit as a limitation. It would help readers to understand whether landslide impacts are entirely lost or whether some are absorbed under parent storm or flood events.

Response

To address this concern, we have added the following sentence at L303 in Section 3.4: Nonetheless, these events may be implicitly accounted for in the impact data, for example if the reported impacts for a flood include the impacts of a landslide triggered by the flood.

Comment 6

Specific comments on the article
Phrase "extensive spatio-temporal coverage" (L58): this wording may overstate completeness given Wikipedia's known biases. Moderating this statement or directing the reader to the evaluation and limitations sections would avoid possible misinterpretation.

Response

We partially agree with the reviewer and recognize the limitation for extending the Wikimpacts database coverage within Wikipedia limits. However, we do not claim that extensive coverage will resolve the bias issue in the data, We have therefore revised the sentence in L64 to: extensive spatio-temporal coverage (within Wikipedia limits)

Comment 7

Specific comments on the article
Temporal coverage: the database extends from 1034 to 2024. The sharp rise in event counts after the 19th century makes it clear that older entries are sparse. Adding one sentence advising users how to interpret pre-1900 records (highly incomplete, not suitable for quantitative trend analysis) would improve clarity.

Response We thank the reviewer for this helpful suggestion and have added the following sentence in L446 in Section 5.1.1: Nonetheless, as records prior to 1900 are sparse due to limited reporting, we view them as unsuitable for quantitative trend analyses.

Further, we also acknowledge the value of including more pre-1900 records, and have added the following sentence to Section 6.3 (Limitations), L661: Moreover, the pre-1900 records in Wikimpacts 1.0 database are limited, and future versions of Wikimpacts database will aim to include available historical events prior to 1900.

Comment 8

Specific comments on the article
Evaluation (Sect. 4): Its structure and intent are valuable. It would help the reader if Sect. 3 indicated that the extraction quality is assessed in Sect. 4. Furthermore, Sect. 4 would benefit from a short explanation of how the 70 + 156 gold-standard events were sampled (randomly or stratified). This may affect the generalisability of the error rates (?).

Response

To address the first part of this comment, we added a brief paragraph at the end of Section 1 (Introduction) that outlines the structure of the paper as follows:

The remainder of this paper is organized as follows. Section 2, *Database Structure*, presents an overview of the database design and the technical definitions of all fields. Section 3, *Wikimpacts Processing Pipeline*, describes in detail the methodology used to construct the database, while Section 4, *Evaluation of the Pipeline*, reports the evaluation methods and results. Section 5, *Wikimpacts 1.0: Content of the Database*, presents the spatial and temporal distribution of the database content and compares it with EM-DAT. Section 6, *Discussion*, provides a detailed assessment of the pipeline and database, the comparison with EM-DAT, and the limitations of the database. Finally, Section 7, *Conclusion*, summarizes the main contributions of this work.

We address the second part of the comment jointly with the related comments from another reviewer concerning the annotation process, as follows in L326-330:

The gold standard events are randomly sampled from the original set of 5,046 classified ar-

ticles, and only the subset of single-event articles are annotated for Wikimpacts 1.0 database development, with the constraint that their event-type distribution is representative of that of the entire database. The annotation was conducted over the course of one year by two postdoctoral researchers in climate science and two researchers with a master's degree in water engineering.

Comment 9

Specific comments on the article
Regarding the interpretation of field-specific error rates, Table 6 reveals that location has a much higher error rate than other fields. The manuscript would be strengthened by explaining what types of errors dominate (for example, administrative-level mismatches, NULL penalties, or coordinate issues) and by offering guidance on how users should interpret L2 and L3 location fields. A short paragraph identifying which extracted fields are robust for most applications and which require caution would be particularly helpful.
The sentence "Within L1, event and timing data are highly accurate, while location data is less robust," is not intuitive because L1 represents the aggregated event level. It would be useful to clarify what "location" means at L1 and why its accuracy is lower.

Response

We address this point jointly with the corresponding comments provided by other reviewers on this section. We have updated the interpretation of the error scores across different fields in the discussion section, where we reflect on the evaluation and analysis of the results and provide targeted recommendations for potential application use cases. We prefer not to add further information to Section 4.3 (Evaluation Results), to make the quantitative evaluation easy to follow. We instead direct readers to Section 6.1 (Database Quality Assessment), where these aspects are discussed in detail. We also refer the reviewer to our reply to Comment 4, where we clarify the meaning of locations at each level.

We propose the following edits: L424-425 in Section 4.3: For an overview of the error score analysis, we refer the reader to Section 6.1. In Section 6.1, we implement the changes as follows: In L567, we add Overall, for L1, the error rate for location information is high, with a score of 0.48 in Administrative_Areas_Norm across 156 events. We find that 35 NULL penalties, each scoring 1 for the corresponding attribute, account for nearly 46% of the total error score. Similarly, NULL penalties in the L2 and L3 location information also lead to high error scores in these location-related fields. During the consolidation process, we filter out these entries. Moreover, we do not compare the results obtained after consolidation with the gold standard, because, as described in Section 3.4, the data are processed from L3 to L2 to L1, and the resulting processed data no longer match the originally annotated data.

Comment 10

Specific comments on the article
Comparison with EM-DAT: the manuscript would be improved by a more cautious framing of discrepancies. Differences may arise not only from extraction errors but also from different event definitions, thresholds, and loss components. I recommend explicitly positioning Wikimpacts as complementary to curated sources and offering guidance on how users might use them together.

Response We agree that differences in inclusion criteria, as well as in event and impact definitions, can contribute to discrepancies in the comparison. In line with the comments concerning the EM-DAT comparison raised by another reviewer, we have added the following text in L636 Section 6.2 to address this point:

Notably, the substantial differences between Wikimpacts and EM-DAT arise not only from extraction errors in our pipeline, but also from divergent event definitions, inclusion thresholds and criteria for impact data, differences in impact sources, and differences in the definitions and components of impact categories. Moreover, in this study we did not conduct a systematic comparison with other existing databases, such as DesInventar, but previous work shows large discrepancies between DesInventar and EM-DAT (Worou and Messori, 2025; Panwar and Sen, 2019). We therefore recommend using our database as a complementary resource to existing databases. Furthermore, when matching events across different databases, we suggest testing different matching algorithms, for example based on event names and by allowing temporal buffers in event dates, in order to more reliably identify corresponding events across databases.

Comment 11

Specific comments on the article
Fig. 4: The distribution of events confirms that Wikimpacts 1.0 reflects high-impact, media-visible disasters rather than a full record of climate extremes. Making this explicit in the Discussion would prevent users from interpreting the dataset as complete.

Response We partly agree with this statement, although we note that the diffuse contributor base to Wikipedia means that also events of local relevance that may not make it to the international news outlets may be reported. It is nonetheless true that Wikimpacts does not provide a complete record of climate extremes. Accordingly, we have added the following sentence to Section 5 L435 to clarify this point:

Notably, the Wikimpacts 1.0 database is constructed from events recorded in the English-language Wikipedia, and does not constitute an exhaustive record of all climate extremes.

Comment 12

Specific comments on the article
Discussion: The existing section is strong, but could more directly address certain limitations:

1. Notability and language biases linked to Wikipedia;
2. The deliberate exclusion of particular hazards (for example, landslides) and implications for multi-hazard studies;
3. Systematic weaknesses in LLM extraction for multi-country or compound hazards beyond aggregated error rates.

A short concluding paragraph offering explicit user guidance, indicating suitable and unsuitable use cases, would be valuable. For example, the database appears well-suited for global comparative studies, exploratory sub-national analyses where data exist, or cross-hazard synthesis, but less suitable for completeness-sensitive applications or studies focused on small, local events.

Response

Together with the comments from the other reviewers on this section, we have added the following text to Section 6.3.

1. In L655, we add We however find that most of the reported events in other language editions of Wikipedia are also reported in English. While English-language bias exists in our database, we thus do not view it as the primary source of overall bias. We also note that including articles on climate events from other-language Wikipedias, where there is no English-language reporting, would not fully address the location bias issue.

2. In L665, we add Moreover, certain hazards, such as landslides, are not included in our database.

3. In L667, we add Systematic errors introduced by the LLM-based extraction, such as misclassification of L2 and L3 information, may lead to duplicate entries in the database.

4. In L675, we add Nevertheless, the Wikimpacts 1.0 database is suitable for global impact-database benchmarking, exploratory sub-national impact assessment, and risk modeling in data-rich contexts (e.g., tropical storm events in the United States). However, it should be used with caution in applications that are highly sensitive to completeness or that focus on local or small-scale events.

Comment 13

Minor comments
Regarding language and clarity, a few specific examples illustrate areas where editing would improve readability:

1. The sentence at L32 about geolocating EM-DAT events is grammatically unclear and should be revised.
2. The phrase "administrative units at the same level can also be highly variable" (L35) needs clarification as to what variability matters for impact analysis.
3. The sentence beginning "Due to the categorization based on single hazards" (L40) would benefit from rephrasing for grammar and conceptual clarity.
4. The reference to DesInventar at L42 would be clearer if the specific shared limitations were briefly stated.
5. The last part of the Introduction (L55-70) blends methodological details that belong in Methods or Data Availability. Ending the Introduction with a clearer statement of objectives and contributions would strengthen the structure.

Response We adapt the text as follows based on the comments and suggestions above:

1. In L34, we rephrase the sentence to Researchers have attempted to geolocate disaster events from EM-DAT, yet the resulting data comes with limited temporal coverage (events from 1960 to 2018) and include some geoparsing errors (Worou and Messori, 2025; Lindersson and Messori, 2025; Teber et al., 2025). Although the geoparsing errors for EM-DAT events from 1990-2023 are handled in the new geocoding database, GeoDisasters (Teber et al., 2025), mapping the aggregated impact from national level to subnational scales remains challenging (Rosvold and Buhaug, 2021; Delforge et al., 2025).

2. In L38, we rephrase the sentence to Moreover, the level of administrative divisions used in the latter databases varies between countries, and administrative units at the same level can also be highly variable due to the differing resolutions, which complicates implementation of damage functions in impact assessment studies (Eberenz et al., 2021; Lüthi et al., 2021).

3. In L45, we rephrase the sentence to Because impact entries are categorized based on single hazards, thus, impacts from co-occurring such as droughts and heatwaves (Zscheischler and Seneviratne, 2017) or multi-hazard events may not be captured appropriately (Lee et al., 2024; Mithal et al., 2024).

4. In L46, we rephrase the sentence to Similar limitations, such as the reporting of impacts as single-hazard categories, also affect other global multi-hazard impact databases, such as DesInventar (UNISDR, 2024).

5. For L55-L70, we prefer to retain the current text, which provides a brief introduction to the method and explains how it addresses the database limitations discussed earlier in this section. With respect to the explicit statement of objectives and contributions, we have added a dedicated paragraph in the Introduction section, as indicated in Comment 3.

Comment 14

Minor comments

1. In Section 2, the opening paragraph focuses on repository and accessibility information rather than internal structure. Moving that material to a Data Availability section would allow Section 2 to begin more directly with the conceptual design. Similarly, some technical field-definition details could move to Supplementary Information.
2. Regarding abbreviations, SI at L88 should be defined on first use.
3. Regarding the example referring to 2025 Wikipedia data (L89), the manuscript could clarify how information "as of 2025" was obtained when the mining cutoff appears to be 2024. A brief explanation would prevent confusion. Consider a relevant note in Figure 1.

Response

We have addressed these comments as follows:

1. We move the first paragraph in Section 2 to the Data Availability section. Besides, the technical definitions are maintained in this section as reasons indicated in Comment 4.
2. We define Supplementary Material (SI) on L109.
3. In L110, we rephrase the sentence to In relation to the 2021 European flood event, the information on fatalities recorded in Wikipedia has been updated as of 2024 since it was accessed to construct the database and indicates 196 deaths in Germany, 39 in Belgium, 2 in Romania, and 1 each in Italy and Austria as of 2025 (information manually extracted on 15 January 2026).

3 Reviewer 2

We would like to thank Reviewer #2 for their dedicated time reviewing the manuscript and for their useful and constructive suggestions. We carefully addressed all comments by the Reviewer and the manuscript has strongly benefited from the proposed changes. In short, we implemented the following changes in the database and manuscript:

1. Add contribution of Wikimpacts database for climate impact analysis.
2. Add detailed description of annotation process.
3. Add more information for the comparison with EM-DAT.
4. Update the Figure 2, 7 and 8.
5. Improve the Discussion section.

Many suggestions point to potential additional research, which incited us to rework the manuscript, and which indicate that our study could ignite and inspire new research.

Below, we would like to clarify our changes regarding all comments, which are repeated in grey boxes. Moreover, the following convention is applied to denote modifications that we implemented in the original manuscript (Notably, all line numbers refer to the revised manuscript): new text.

Comment 1

General assessment
The work is of high scientific interest, given the need for transparent, harmonized and scalable datasets for climate risk assessment and loss and damage research. Despite its obvious potential, the manuscript requires substantial revisions to improve the clarity, structure, and articulation of its scientific contribution. Several aspects of the article would benefit from clearer explanation, greater integration between methods and results, better organization of figures and tables, and a more explicit discussion of both scientific limitations and advances.
If the authors adequately address the points presented below, the manuscript may be reconsidered for publication.

Response

We thank the Reviewer for appreciation of the study topic and for the thoughtful review. Below, we address every comment carefully (with some comments addressed jointly) and explain the corresponding changes in the manuscript.

Comment 2

Major comments
The abstract is clear and well structured, effectively presenting the objective, methodology, and main results. However, it does not fully convey the scientific relevance and broader implications of the work. The authors could slightly expand the final sentence to better highlight the contribution of Wikimpacts 1.0 to climate impact research and data integration.

Response

To address this, we rephrase the sentence on L12 and L13 as follows: Our public database serves as a complementary resource to existing impact databases, facilitates subnational climate impact assessments, and highlights the potential of natural language processing to enhance existing impact datasets and provide robust information on climate impacts.

Comment 3

Major comments
The introduction is clear, comprehensive, and well supported by recent literature. It effectively presents the relevance of climate impact data, the limitations of existing databases, and the general methodological approach. However, the section could place greater emphasis on the scientific innovations of the work - for example, by briefly explaining how Wikimpacts 1.0 improves on existing approaches or supports climate impact analysis. A brief description of the paper's structure at the end of the section would also enhance readability. Finally, the reference "Li et al., 2025a" appears to correspond to the DOI of the Wikimpacts 1.0 dataset. While it is appropriate to cite the dataset associated with the paper, the current phrasing could confuse readers into thinking it refers to a separate publication. The authors could clarify by writing "the accompanying Wikimpacts 1.0 dataset (Li et al., 2025a)" or similar.

Response

To address the request to clarify how *Wikimpacts 1.0* improves upon existing approaches and supports climate impact analysis, we add the following sentence on L65:

Our framework implements an automated, end-to-end pipeline for extracting and processing impact information from the sub-national up to the event level, thereby enabling multi-scale climate impact assessment studies.

As well as on L71: The Wikimpacts database is designed to complement existing global disaster impact databases such as EM-DAT and DesInventar, notably by providing georeferenced information at subnational level. Although the inclusion criteria and event definitions differ across databases, matching events by type, date, and location allows us to identify a set of shared events that can be jointly analysed. These overlapping records can support global multi-hazard comparisons and sub-national studies, particularly when combined with climate and other geospatial data. Although we rely solely on English Wikipedia articles in the Wikimpacts 1.0 database, we find relatively few climate event articles in other languages that are not also reported in English. Some English bias may exist, but we thus do not view it

as a main issue. We however acknowledge that regions with limited Wikipedia activity and small or highly localized events are likely to be under-reported. Consequently, the Wikimpacts 1.0 database should not be used in isolation for complete or highly sensitive national loss assessments. For such applications, we recommend benchmarking Wikimpacts against other databases such as EM-DAT and officially reported government statistics.

Besides, in response to the suggestion that "a brief description of the paper's structure at the end of the section would also enhance readability", we have added a short paragraph at the end of the Introduction section.

The remainder of this paper is organized as follows. Section 2, *Database Structure*, presents an overview of the database design and the technical definitions of all fields. Section 3, *Wikimpacts Processing Pipeline*, describes in detail the methodology used to construct the database, while Section 4, *Evaluation of the Pipeline*, reports the evaluation methods and results. Section 5, *Wikimpacts 1.0: Content of the Database*, presents the spatial and temporal distribution of the database content and compares it with EM-DAT. Section 6, *Discussion*, provides a detailed assessment of the pipeline and database, the comparison with EM-DAT, and the limitations of the database. Finally, Section 7, *Conclusion*, summarizes the main contributions of this work.

Lastly, we have incorporated the phrase the accompanying Wikimpacts 1.0 dataset (Li et al., 2025) into line 62, as suggested in the comment.

Comment 4

Major comments
In section 2.1, the normalization example ("more than 200 deaths...") introduces methodological details that belong to Section 4 and distracts from the structural description of the database.
Similarly, in Section 2.2, the forward reference to the case study later discussed in Section 6 is premature and breaks the logical flow; it could be rephrased more generally (e.g., "an example is provided later in the paper").

Response

We choose to preserve the original text as each of the examples serves a specific purpose, as detailed below:

In Section 2.1, the normalization example is specifically used to explain that, in L141, the "Approx" field is set to "True" under this case.

In Section 2.2, the case study of deaths in Germany during the 2021 European floods is employed to illustrate the database schema for L2 (national-level information), in conjunction with Fig. 1.

Comment 5

Major comments
In section 3.1, it should be specified that the text classifier is the fine-tuned English BERT model itself, not a separate classifier. The numerical results (30085 - 4900 - 5046 articles) are intermediate outcomes and would be better presented in the results section; this part should focus on the procedure.
Section 3.2 clearly describes the GPT-4o extraction process, though it is somewhat too detailed. Some prompt information could be moved to the Supplementary Material, and a brief justification of the model choice would improve clarity.
In section 3.3.3, the term "version 3.1/3.2" appears to refer to different configurations of the extraction pipeline or prompt templates rather than model versions. This could be stated explicitly to avoid confusion.

Response
Regarding the first suggestion, we rephrase the sentences in L176-179 to: we apply as text classifier a pre-trained English distilled Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019; Sanh et al., 2019), to filter non-climate-related articles. To this end, the BERT model is fine-tuned on a set of 300 Wikipedia articles, containing 248 relevant and 52 irrelevant articles.

Regarding the suggestion of moving the numerical results (30,085-4,900-5,046 articles) of the document selection process to Section 5 (Wikimpacts 1.0 Content of the Database), we prefer to retain this text in its original location. Relocating it may lead to confusion for readers in Section 5, where we focus exclusively on the single-event articles in Wikimpacts 1.0. The current section is intended to present the intermediate numerical results obtained during the selection process, which culminate in the final number of processed articles included in Wikimpacts 1.0.

Concerning the second suggestion, we have clarified the our model choice on Lines 216-218 as follows: we employ the GPT4o model, released in 2024, rather than the GPT4 model because its longer context window (Hurst et al., 2024) allows it to process a substantially larger number of input tokens, which is particularly beneficial for handling long Wikipedia articles.
Besides, in response to the reviewer's suggestion to move part of the prompt information to the SI (Supplementary Material), we would retain the original text in the main manuscript to guide readers through the design and implementation of the prompts, while the SI provides the full, detailed prompt specifications.

Lastly, for the third suggestion, we added "prompt" before v3.2 and v3.1 in Section 3.3.3.

Comment 6

Major comments
Section 4 presents a solid and well-structured description of the validation framework used to assess the Wikimpacts 1.0 pipeline. The creation of a manually annotated gold standard and the definition of customized normalized error metrics are strong methodological points that demonstrate transparency and reproducibility. However, the section mixes methodological description and quantitative outcomes. The numerical results (e.g., error scores by level and field) would be more appropriately presented in the results section, while this part should focus on explaining the validation design. Separating the methods from the results would improve clarity and logical flow. The rationale for selecting 70 events for the development set and 156 for the test set should be briefly explained (e.g., availability, annotation workload, or statistical considerations). The description of the annotation process could also specify the number and background of annotators and how disagreements were resolved.
The statement regarding the evaluated database fields (those without asterisks in Table 3) could be clarified: only fields directly extracted by GPT-4o are assessed, while post-processed or derived fields are excluded. This distinction is important and should be explicitly stated.

Response

First, we offer our response regarding the first comment stating that the section mixes methodological description and quantitative outcomes. In this particular case, we prefer to retain the evaluation results within the current section rather than moving them to a separate results section. Since this work presents a database, the manuscript does not follow the conventional structure with a distinct "Results" section. Instead, Section 5 describes the content of the database. To avoid potential confusion between the characteristics of the database itself and the results of the pipeline evaluation, we consider the current organization of the paper to be clearer.

Secondly, regarding the comments about the annotation process, we address them by adding the following explanation on L326: The gold standard events are randomly sampled from the original set of 5,046 classified articles, and only the subset of single-event articles are annotated for Wikimpacts 1.0 database development, with the constraint that their event-type distribution is representative of that of the entire database. The annotation was conducted over the course of one year by two postdoctoral researchers in climate science and two researchers with a masterâs degree in water engineering. As well as L355 :These variations in interpretation contribute to inter-annotator agreement errors, while the remaining discrepancies in the L2 and L3 annotations will be addressed in the forthcoming version of the database.

Lastly, we address the comment about the evaluated database fields by rephrasing the text in L382-384: In this paper, we evaluate only those fields that are directly extracted by GPT4o, excluding any post-processed or otherwise derived fields. Specifically, our assessment is limited to the fields in Table 3 that appear without an asterisk, as the asterisked fields are obtained through post-processing rather than representing the raw output of the LLM.

Comment 7

Major comments
The comparison with EM-DAT, reported in Section 5.4, is extremely useful but would benefit from a clearer explanation. The text should specify that the event-by-event comparison involves the main quantitative impact fields (deaths, injuries, damage, and homeless when available), and describe how events were matched between the two databases and how the percentage-difference classes (± 10 %, ± 10 %, ± 10 %) were defined. Figure 10 (a-d) effectively summarizes these differences, but the caption and text should explicitly explain the calculation, the meaning of the color scheme (blue = Wikimpacts < EM-DAT; red = Wikimpacts > EM-DAT), and the interpretation of the observed discrepancies-particularly the large deviations for damage data.

The qualitative comparison with EM-DAT presented in section 6.2 is relevant and adds useful context regarding differences in structure, granularity and traceability. However, Section 6.2 does not introduce enough additional material to justify a standalone subsection titled "Comparison with existing impact databases", especially since EM-DAT is the only database discussed. To improve coherence and avoid fragmentation, the authors might integrating this discussion directly into Section 5.4, making the EM-DAT comparison a single unified block or expanding the subsection to briefly include other relevant databases (e.g., DesInventar,), thereby matching the broader scope suggested by the title.

Response In conjunction with the related comments from another reviewer, namely:

"Comparison with EM-DAT: the manuscript would be improved by a more cautious framing of discrepancies. Differences may arise not only from extraction errors but also from different event definitions, thresholds, and loss components. I recommend explicitly positioning Wikimpacts as complementary to curated sources and offering guidance on how users might use them together."

we address these points as follows:

1. In L535-539, we rephrase the original sentences to Moreover, consistent with the characteristics of the EM-DAT database, the four impacts (deaths, injuries, homeless, and total damage) in L2 data are used in our database for the event-by-event impact value comparison. For each event and each impact variable, we compute the relative difference as ((Wikimpacts impact value - EM-DAT impact value) / EM-DAT impact value $\times$100%), which quantifies how much the Wikimpacts value differs from the corresponding EM-DAT value Besides, the caption of Fig 10, is rephrased to Impact value comparison between EM-DAT and Wikimpacts 1.0 over 01/01/1900 - 29/02/2024. Blue indicates cases where the Wikimpacts impact values are lower than the EM-DAT values, whereas red indicates the opposite. (a) The percentage difference between Wikimpacts 1.0 and EM-DAT in the death category, (b) injury category, (c) homeless category, and (d) damage category. This directly addresses the comment that the text should specify that the event-by-event comparison involves the main quantitative impact fields, clarifying how the percentage-difference values are defined, and the meaning of the color scheme in Fig 10.

2. In response to the reviewers' comments on: 1) the interpretation of the observed dis-

crepancies, especially the large deviations in the damage data, 2) the recommendation to move the comparison discussion from Section 6.2 to Section 5.4, and 3) the request to expand a subsection to briefly cover other databases (e.g., DesInventar) to match the title of Section 6.2, as well as the suggestions provided by the other reviewer mentioned above. We have decided to keep the EM-DAT comparison discussion in Section 6.2 that provides a deeper insight of the comparison discussion, revise the corresponding paragraph in that section, and add an explanation at L636 to reflect the broader scope of Section 6.2, as follows: Notably, the substantial differences between Wikimpacts and EM-DAT arise not only from extraction errors in our pipeline, but also from divergent event definitions, inclusion thresholds and criteria for impact data, and differences in the definitions and components of impact categories. Moreover, in this study we did not conduct a systematic comparison with other existing databases, such as DesInventar, because DesInventar predominantly covers events in the Global South, as illustrated in Worou and Messori (2025), whereas our database primarily documents events in the Global North. We therefore recommend using our database as a complementary resource to existing databases such as EM-DAT and DesInventar. Furthermore, when matching events across different databases, we suggest testing different matching algorithms, for example based on event names and by allowing temporal buffers in event dates, in order to more reliably identify corresponding events across databases.

Comment 8

Major comments
The subsection 6.1.1 highlights the main sources of error at L1. To improve readability, the authors could:

1. more clearly separate systematic error types from specific illustrative examples;
2. explicitly link these error types to the corresponding error magnitudes in Table 6.

The subsection 6.1.2 provides a clear explanation for why error rates increase at L2 and L3.
Table 9 is central to this discussion, but its current placement after section 6.1.1 disrupts the logical flow. It should be moved to follow section 6.1.2. The distinction between "LLM" (model-generated output) and "Gold" (manually annotated reference) is introduced earlier in the methodology, but briefly restating it here would improve clarity. A brief clarification of the NULL penalty mechanism would also enhance comprehension.

Response Together with the comments from the other reviewers on this section, these have led us to add the following text at L567. Overall, for L1, the error rate for location information is high, with a score of 0.48 in Administrative_Areas_Norm across 156 events. We find that 35 NULL penalties, each scoring 1 for the corresponding attribute, account for nearly 46% of the total error score. Similarly, NULL penalties in the L2 and L3 location information also lead to high error scores in these location-related fields. During the consolidation process, we filter out these entries. Moreover, we do not compare the results obtained after consolidation with the gold standard, because, as described in Section 3.4, the data are processed from L3 to L2 to L1, and the resulting processed data no longer match the originally annotated data. Beside that, we move the Table 9 after Section 6.1.2 as suggested.

Comment 9

Major comments
The consolidation step in the subsection 6.1.3, is presented as resolving many of the issues identified in 6.1.1 and 6.1.2, but the explanation is very brief. Since consolidation plays an important role in improving the final database, it would be useful to include at least one concrete example showing how an error is corrected during this process.

Response To address these points, we add the following at L618:

Here, we use the Typhoon Kate (1970) event in the evaluation-score computation to illustrate how the consolidation process can improve database quality. For this event, there are 631 recorded fatalities in total; this is the only information available in the annotated data. However, the LLM also extracts 631 deaths in the Philippines at L2 and 631 deaths in southern Mindanao, Philippines, at L3. Under our evaluation protocol, these additional details are assigned NULL penalties, although correct, because they were missed by the annotators and are not present in the ground-truth annotations. In the consolidation process, the L2 and L3 information is recognized as consistent with L1, so these more detailed records are preserved in the database.

Comment 10

Major comments
The limitations, described in the section 6.3, are clearly described, particularly those related to the reliance on English-language Wikipedia and the uneven availability of impact information across regions and hazard types. However, the section could be further strengthened by briefly explaining the practical implications of these limitations for potential uses of the database (e.g., risk modelling, regional comparisons, or loss-and-damage assessments). Even a short statement would help readers understand how these constraints may affect downstream analyses.

Response

Together with the comments from the other reviewers on this section, we have added the following text to Section 6.3. In L675, we add Nevertheless, the Wikimpacts 1.0 database is suitable for global impact-database benchmarking, exploratory sub-national impact assessment, and risk modeling in data-rich contexts (e.g., tropical storm events in the United States). However, it should be used with caution in applications that are highly sensitive to completeness or that focus on local or small-scale events.

Comment 11

Major comments
The conclusions effectively summarizes the structure and overall objectives of the Wikimpacts 1.0 database, but it does not explicitly acknowledge the main limitations discussed in section 6.3, nor does it highlight potential future applications of the database (e.g., climate risk modeling, vulnerability assessments, benchmarking of other impact datasets).
Furthermore, although the conclusion states that Wikimpacts 1.0 presents several innovations compared to the state of the art, these innovations are not explicitly listed.

Response

Based on the comment above, we update the conclusion section as follows:

1. In L683, we rephrase the sentence to There is, however, a clear bias towards events in the Global North, specifically tropical storm events, and those occurring from the 1950s onwards.

2. In L686, we update the sentence to The principal innovation is that, for each extreme event, the database provides hierarchical information on the impacts, enabling multi-scale impact analyses, climate risk and vulnerability assessments, as well as benchmarking against other impact databases. In L691, we rephrase the sentence to The innovative automated pipeline ensures that the database is readily updateable and expandable with the inclusion of additional textual sources.

Comment 12

Comments to figures and tables
Several figures and tables could be improved.
Figures

1. Figure 2 illustrates the three database levels (L1, L2, L3), but its current form is overly schematic and provides limited insight into the actual relational structure of the database. The figure does not clarify the type or direction of the relationships between levels (e.g., 1:N, N:N), nor does it explain the meaning of the arrows or how information is propagated or linked across levels. Improving the figure (e.g., clearer arrows, explicit relationship types, distinct fields at each level) and adding a more detailed explanation in the text would significantly help the reader understand the architecture of the Wikimpacts 1.0.
2. In Figures 7a, 7c, 7f and 8a-c, some color ranges are visually indistinguishable (e.g., 80-160 vs 160-350; 20-40 vs 40-75; 5-10 vs 10-15; 50-100 vs 100-200) and should be adjusted for better contrast.
3. In Figure 9a, the legend appears to refer only to storm events, although the text suggests that both tropical and extratropical categories are included; the legend and caption should be revised accordingly.

Response

Below, we address the Reviewer's points one by one.

1. For Fig. 2, we have revised the figure to clarify that, in the Wikimpacts 1.0 database, only a single L1 entry is permitted, which may be associated with zero or multiple L2 entries. which may be associated with zero or multiple L2 entries. For each L3 entry, the impact within the same country is aggregated into a single L2 entry. Therefore, each L2 entry can be linked to zero or multiple L3 entries by comparing the "Administrative_Area_GID" from the L3 entry with the "Administrative_Areas_GID" in the L2 entry. In L115 of Section 2 (Database Structure), we add As shown in Figure 2, the database permits only a single entry at the L1 level. Each L1 entry is associated with zero or more L2 and L3 entries. Each L2 entry can in turn be linked to zero or multiple L3 entries. If there are L3 entries, all L3 impacts within the same country are aggregated into a single L2 entry. There is no direct linkage between the L2 and L3 entries in the database; however, comparing the "Administrative_Area_GID" from the L3 entry with the "Administrative_Areas_GID" in the L2 entry enables identifying the corresponding entries. The updated figure is shown below:

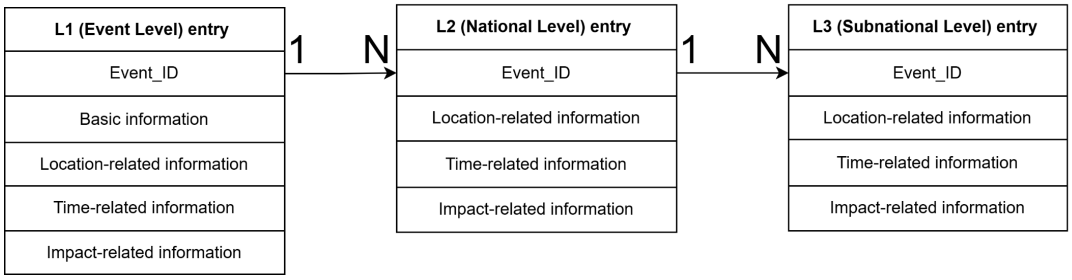

Figure 1: Updated Fig. 2 caption: Wikimpacts 1.0 database structure. L1 records the complete event metadata, including the basic information. L2 and L3 store only location-, time-, and impact-related fields. Entries across L1-L3 are linked by a shared Event_ID, with one L1 event potentially linked to multiple L2 entries, and one L2 entry potentially linked to multiple L3 entries (1:N). For further details on the information fields in the figure see Table 1.

2. For Figs. 7 and 8, we have revised the color scheme to enhance visual distinguishability, as shown below.

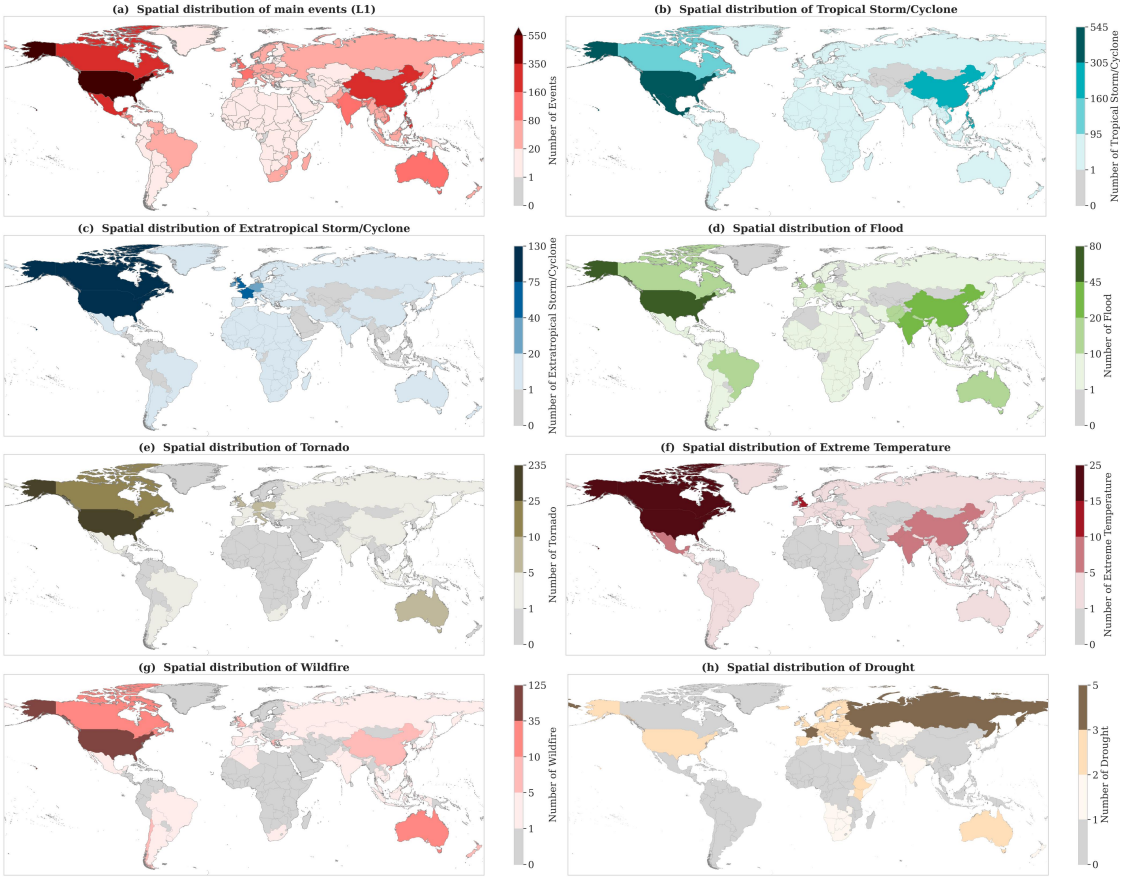

Figure 2: Original Fig. 7 caption. Spatial distribution of main events in the Wikimpacts 1.0 database, based on L1 entries: (a) overall spatial distribution of all main events, (b) spatial distribution of Tropical Storm/Cyclone events, (c) spatial distribution of Extratropical Storm/Cyclone events, (d) spatial distribution of Flood events, (e) spatial distribution of Tornado events, (f) spatial distribution of Extreme Temperature events, (g) spatial distribution of Wildfire events, and (h) spatial distribution of Drought events. Note the non-linear colour scale.

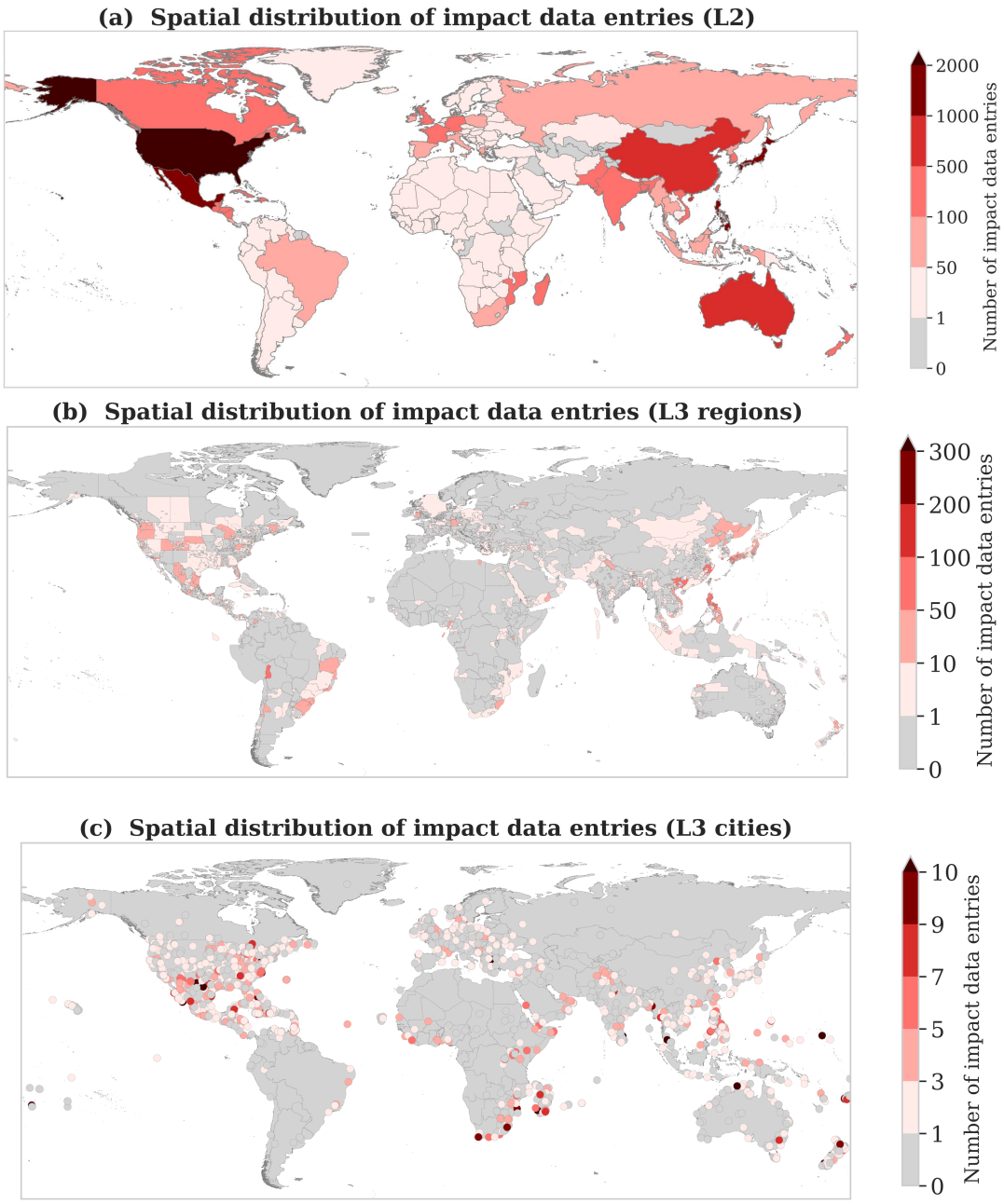

Figure 3: Original Fig. 8: Spatial distribution of L2 and L3 impact data entries in Wikimpacts 1.0. (a) Spatial distribution of impact data entries at national level(L2), (b) Spatial distribution of impact data entries at regional level (L3 polygons, see text), (c) Spatial distribution of impact data entries at city level ((L3 points, see text).

3. In Fig. 9a, we use all storm-related entries from EM-DAT for the comparison; therefore, we retain the original legend. In the caption, we add note that the EM-DAT Storm

category includes all storm-related entries, whereas for Wikimpacts the Trop. Cycl and Extrat. Cycl categories are included.

Comment 13

Tables

1. Table 1 presents the main information categories (base, temporal, spatial, impact, etc.) and their fields, but it is lengthy and not clearly referenced in the text. It should be explicitly cited and briefly discussed when introduced.
2. Table 2 is informative but visually dense. Reformatting (e.g., grouping hazards by type or using multi-column or bullet layouts) would enhance clarity.
3. Table 3 documents each database field, specifying data type, value format, mandatory status, and applicable level (L1-L3). While technically useful, its role relative to Table 1 is not explained. The authors should clarify that Table 1 summarizes data categories, whereas Table 3 provides the technical specifications.

Response Below, we address these points one by one.

1. For Table 1, together with comments from other reviewers, from L101 in Section 2 (Database Structure), we add The location-related information is summarized in Table 1. For L1 and L2, the locations are specified in Administrative_Areas_Norm, which contains a list of affected countries. For L3, the location information comprises both Administrative_Areas_Norm and Locations_Norm, indicating one affected country and a list of affected sub-locations such as cities. In the remainder of the text, we use the general term "location" or "locations" to refer to both these fields.
2. Regarding Table 2, we will retain its current format, as it presents the relationship between the main events and the associated hazards recorded in the database.
3. For the relationship of Table 1 and 3, we add a sentence in L114 Table 1 provides the characteristics of the fields stored in the Wikimpacts 1.0 database for L1, L2, and L3 and Table 3 details the technical specifications for such fields.

References

Delforge, D., Wathelet, V., Below, R., Sofia, C. L., Tonnelier, M., van Loenhout, J. A., and Speybroeck, N.: EM-DAT: the Emergency Events Database, International Journal of Disaster Risk Reduction, 124, 105 509, ISSN 2212-4209, https://doi.org/https://doi.org/10.1016/j.ijdrr.2025.105509, https://www.sciencedirect.com/science/article/pii/S2212420925003334, 2025.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, pp. 4171–4186, 2019.

Eberenz, S., Lüthi, S., and Bresch, D. N.: Regional tropical cyclone impact functions for globally consistent risk assessments, Natural Hazards and Earth System Sciences, 21, 393–415, https://doi.org/10.5194/nhess-21-393-2021, https://nhess.copernicus.org/articles/21/393/2021/, 2021.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card, arXiv preprint arXiv:2410.21276, 2024.

Lee, R., White, C. J., Adnan, M. S. G., Douglas, J., Mahecha, M. D., O'Loughlin, F. E., Patelli, E., Ramos, A. M., Roberts, M. J., Martius, O., Tubaldi, E., van den Hurk, B., Ward, P. J., and Zscheischler, J.: Reclassifying historical disasters: From single to multi-hazards, Science of the Total Environment, 912, 169 120, 2024.

Li, N., Zahra, S., de Brito, M. M., Flynn, C. M., Görnerup, O., Koffi, W., Kurfali, M., Meng, C., Thiery, W., Zscheischler, J., Messori, G., and Nivre, J.: Using LLMs to Build a Database of Climate Extreme Impacts, in: Natural Language Processing meets Climate Change @ ACL 2024, https://openreview.net/forum?id=h7o0qyQ0rt, 2024.

Li, N., Thiery, W., Zahra, S., de Brito, M. M., Worou, K., Kurfali, M., Lampe, S., Munoz, P., Flynn, C., Trigoso, C., Nivre, J., Zscheischler, J., and Messori, G.: Wikimpacts â A global climate impact database based on automated information extraction from Wikipedia Dataset version 1.0., https://doi.org/10.17043/li-2025-wikimpacts-1.0, 2025.

Lindersson, S. and Messori, G.: SHEDIS-Temperature: linking temperature-related disaster impacts to subnational data on meteorology and human exposure, Earth System Science Data, 17, 6379–6403, 2025.

Lüthi, S., Aznar-Siguan, G., Fairless, C., and Bresch, D. N.: Globally consistent assessment of economic impacts of wildfires in CLIMADA v2.2, Geoscientific Model Development, 14, 7175–7187, https://doi.org/10.5194/gmd-14-7175-2021, https://gmd.copernicus.org/articles/14/7175/2021/, 2021.

Mithal, V., Sillmann, J., and Zscheischler, J.: Linking regional economic impacts of temperature-related disasters to underlying climatic hazards, Environmental Research Letters, 19, 124 010, 2024.

Panwar, V. and Sen, S.: Economic impact of natural disasters: An empirical re-examination, Margin: The Journal of Applied Economic Research, 13, 109–139, 2019.

Rosvold, E. L. and Buhaug, H.: GDIS, a global dataset of geocoded disaster locations, Scientific data, 8, 61, 2021.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR, abs/1910.01108, http://arxiv.org/abs/1910.01108, 2019.

Teber, K., Weynants, M., Gans, F., and Mahecha, M. D.: Geo-Disasters: Geocoding climate-related events in the international disaster database EM-DAT, Big Earth Data, 0, 1–16, https://doi.org/10.1080/20964471.2025.2576274, https://doi.org/10.1080/20964471.2025.2576274, 2025.

UNISDR: DesInventar: United Nations Office for Disaster Risk Reduction, Retrieved in May 2024 from https://www.desinventar.net, 2024.

Wikipedia contributors: 1999 Blayais Nuclear Power Plant flood — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=1999_Blayais_Nuclear_Power_Plant_flood&oldid=1309899681, [Online; accessed 26-January-2026], 2025a.

Wikipedia contributors: 2011 European floods — Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=2011_European_floods&oldid=1310685507, [Online; accessed 26-January-2026], 2025b.

Worou, K. and Messori, G.: Compounding droughts and floods amplify socio-economic impacts, Environmental Research Letters, 20, 104 024, 2025.

Zscheischler, J. and Seneviratne, S. I.: Dependence of drivers affects risks associated with compound events, Science advances, 3, e1700 263, 2017.