https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

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

Ni Li[1,2], Wim Thiery[1], Shorouq Zahra[3,6], Mariana Madruga de Brito[4], Koffi Worou[5,6], Murathan Kurfalı[3,6], Seppe Lampe[1], Paul Muñoz[1], Clare Flynn[5,6], Camila Trigoso[1], Joakim Nivre[3,6,7], Jakob Zscheischler[2,8,9], and Gabriele Messori[5,6,10]

[1]Department of Water and Climate, Vrije Universiteit Brussel, Brussels, Belgium
[2]Department of Hydro Sciences, TUD Dresden University of Technology, Dresden, Germany
[3]RISE Research Institutes of Sweden, Sweden
[4]Department of Urban and Environmental Sociology, Helmholtz Centre for Environmental Research — UFZ, Leipzig, Germany
[5]Department of Earth Sciences, Uppsala University, Uppsala, Sweden
[6]Swedish centre for impacts of climate extremes (climes), Uppsala University, Uppsala, Sweden
[7]Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden
[8]Department of Compound Environmental Risks, Helmholtz Centre for Environmental Research — UFZ, Leipzig, Germany
[9]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany
[10]Department of Meteorology and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden

Correspondence: Ni Li (ni.li@vub.be)

Abstract.

Climate extremes like storms, heatwaves, wildfires, droughts and floods significantly threaten society and ecosystems. However, comprehensive data on the socio-economic impacts of climate extremes remains limited. Here we present Wikimpacts 1.0, a global climate impact database built by extracting information from Wikipedia using natural language processing. Our method identifies relevant articles, extracts the information using GPT4o, post-processes the information and consolidates the database. Impact data is stored at the event, national, and sub-national levels, covering 2,928 events from 1034 to 2024, with 20,186 national and 36,394 sub-national entries. The database shows low error scores (range from 0 to 1) for event-level information like timing (0.05), deaths (0.03), and economic damage (0.12), and slightly higher error scores for injuries (0.21), homelessness (0.25), displacement (0.29), and damaged buildings (0.28) compared to manually annotated data from 156 events. Wikimpacts 1.0 provides broader impact coverage on storms than EM-DAT at the sub-national level. In comparing impact values, 38 out of 234 matched events have identical data for deaths, and 7 of 94 for injuries. However, there are notable discrepancies in information on homelessness and damage. Our public database highlights the potential of natural language processing to complement existing impact datasets and to provide robust information on climate impacts.

1 Introduction

Climate extremes – such as storms, heatwaves, wildfires, floods, and droughts – cause substantial impacts to society, often leading to large losses of life and property. These consequences are expected to exacerbate in the future due to the ongoing

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

climate and land-use changes (Seneviratne et al., 2021; Ara Begum and Wester, 2022). Climate change has already led to an increase in the frequency, intensity, duration and geographical extent of many climate-related extreme events, a trend projected to continue in the coming decades (Seneviratne et al., 2021; Lange et al., 2020; Thiery et al., 2021; Muheki et al., 2024). A comprehensive understanding of the impacts of extreme climate events is crucial for improving impact forecasting, impact projections, early warning systems, and managing disaster risks (Thiery et al., 2017; de Brito et al., 2024; Hurlbert et al., 2019; Zommers et al., 2020). For example, accurate and geographically-resolved impact data is essential for pinpointing areas that are disproportionately affected by climate extremes (Hammond et al., 2015), allowing for targeted allocation of climate adaptation efforts. Climate impact data can also be used to assess the effectiveness of adaptation measures in reducing loss and damage from climate extremes (Kreibich et al., 2023).

However, currently available climate impact data suffer from a number of limitations. Many global impact databases are proprietary, and not openly available for researchers. Examples include NatCatSERVICE[1], Sigma[2], and PERILS[3], originating from the insurance sector (Jones et al., 2022a; Ahmadi Mazhin et al., 2022). The data in existing open-access global climate impact databases suffer from incompleteness, inconsistencies, and/or biases (Harrington and Otto, 2020; Tschumi and Zscheischler, 2020; Panwar and Sen, 2020; Mithal et al., 2024). One of the most widely used open databases for climate extreme event impact studies is EM-DAT[4] (Delforge et al., 2023, 2025). Although EM-DAT is a valuable database, its use for systematic climate impact studies presents several challenges. Researchers have attempted to geolocate disaster events from EM-DAT, this comes with limitations in temporal coverage, and mapping the impact to a subnational scale remains challenging (Rosvold and Buhaug, 2021; Delforge et al., 2025). Moreover, the level of administrative divisions used in the latter database varies between countries, and administrative units at the same level can also be highly variable. Similarly, temporal information in EM-DAT can be inconsistently documented as a range of days, months, or a single year. Furthermore, when a single physical event has a wide-ranging influence, it may be documented under multiple entries (Faiella et al., 2020). In addition, the number of events in both developed and underdeveloped countries is likely under-reported (Harrington and Otto, 2020). For those events that are reported, there is a substantial number of missing entries in the predefined impact categories, especially those pertaining to economic losses (Jones et al., 2022b). Due to the categorization based single hazards, the impacts from frequently co-occurring hazards such as droughts and heatwaves (Zscheischler and Seneviratne, 2017) but also other multi-hazard events are often not captured appropriately (Lee et al., 2024; Mithal et al., 2024). Similar limitations also affect other global multi-hazard impact databases, such as DesInventar (UNISDR, n.d.). While single-hazards databases (e.g. Papagiannaki et al. (2022); Paprotny et al. (2023), IFNet[5], Dartmouth[6], WISC[7]), and databases focusing on national spatial scales (e.g., Sodoge et al., 2023) have better coverage and completeness, they are generally difficult to expand to multiple hazards or other regions. Furthermore, they all use different impact categories and event definitions, hindering cross-database multi-hazard impact analyses. Lastly,

[1]https://www.munichre.com/en/solutions/for-industry-clients/natcatservice.html
[2]https://www.sigma-explorer.com/
[3]https://www.perils.org/products/industry-exposure-and-loss-database
[4]https://www.emdat.be/
[5]http://www.internationalfloodnetwork.org/index.html
[6]https://floodobservatory.colorado.edu/Archives/index.html
[7]https://climate.copernicus.eu/windstorm-information-service

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

many multi-hazard global databases and single-hazard or national databases are developed as a manual effort by small teams of researchers, which makes timely updates difficult. The manual process also limits traceability of the information, making it challenging to connect a given entry to a specific data source.

An alternative source of information on impacts from climate and weather extremes comes from digitalised textual records such as newspaper archives (de Brito et al., 2020; d'Errico et al., 2020; Stahl et al., 2016; Alencar et al., 2024) and Twitter (de Bruijn et al., 2019). This data can overcome some of the shortcomings of existing impact databases. They provide detailed impact records, which are typically associated with specific dates and locations. Despite the widespread digitalisation of text and the wealth of quantitative information available on impactful climate events, there is currently no global, multi-hazard, open and traceable climate impact database leveraging freely available online textual sources. Here, we present one such database: Wikimpacts 1.0 (Li et al., 2025a).

Wikimpacts 1.0 addresses some of the aforementioned database limitations, by providing extensive spatio-temporal coverage, standardized temporal, spatial, and impact information, and ease of updating for new events. Our automated multi-step pipeline extracts semi-structured data from English Wikipedia articles by utilizing GPT4o, a pre-trained Large Language Model (LLM). The data then undergoes a post-processing step in which different data points are refined, normalized, and stored in a relational database. As such, this database aims to reflect the information available in Wikipedia as accurately as possible, without evaluating the reliability of the underlying information in the article. Geo-parsing is a crucial step in our post-processing to connect place names to geographical entities and boost the database's usability for research.

This database has been developed in compliance with Article 3 of Directive (EU) 2019/790 regarding copyright and related rights within the Digital Single Market, utilizing lawful text and text mining techniques. The data encompassed within this database is derived from automated extraction and synthesis of information from legally accessible sources, including publicly available Wikipedia articles. The dataset exclusively comprises factual information (e.g., temporal data, geographic locations, event types, reported impacts) and does not replicate any protected expressions or copyrighted material from the original sources.

2 Database Structure

The Wikimpacts 1.0 dataset comprises approximately 1.5 GB of data in SQLite database format and is publicly accessible via an open-access database server under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Interested users can access the entire dataset at https://bolin.su.se/data/li-2025-wikimpacts-1.0 (Li et al., 2025a). Furthermore, we direct readers to explore the various database releases at https://doi.org/10.5281/zenodo.14730195, which include the raw outputs from the LLMs, as well as subsequent processing steps related to currency conversion and inflation adjustment.

To ensure ease of use, Wikimpacts 1.0 adopts impact categories similar to existing disaster databases, such as EM-DAT. Table 1 provides a detailed description of the information recorded in our database. We classify all events recorded in the database into 7 categories, hereafter referred to as Main Events, and subsequently assign one or more hazards to each main event category

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

(see Table 2). Our database provides impact information at three levels: event level (L1), national level (L2), and sub-national level (L3). L1 provides the total impacts associated with a given main event across all affected countries; L2 provides the national-level impacts, and is the same as L1 if the event affected a single country; and L3 includes impacts at the smallest available sub-national locations within each affected country. This structure is exemplified in Figure 1 for deaths caused by the severe flooding episode that affected parts of Western Europe in 2021. The L1 information is the total number of deaths across all countries affected by the event. L2 provides a breakdown of the number of deaths by country e.g., 196 deaths in Germany. L3 further details the number of deaths at a sub-national level, specifying either point locations (e.g., cities) or polygons (e.g., provinces). This is shown as an example for the city of Pepinster (point) and for the German state of Bavaria (polygon); the full deaths information for the 2021 European Floods can be found in the SI Section 6. It is important to note that the Wikimpacts 1.0 database, which utilizes article mining beginning in 2024, currently does not reflect updates to article information. Future developments will enable the database to undergo near-realtime updates. In relation to the European flood event, the information concerning fatalities is recorded in Wikipedia as of 2025, indicating 196 deaths in Germany, 39 in Belgium, 2 in Romania, and 1 each in Italy and Austria.

The Wikimpacts 1.0 database is stored in a relational format (Figure 2) and Table 3 details the characteristics of the fields stored in the Wikimpacts 1.0 database for L1, L2, and L3. For all three levels, we provide a detailed breakdown of the schema, structure, and permitted values as follows: (i) Field: the specific names assigned to each piece of information within the schema; (ii) Data Type: the data format for each field (integer, string, list, boolean); (iii) Permitted Values: the range or set of allowed values for each field (e.g., specific categories, numeric ranges); (iv) Mandatory: an indication of whether the field is required (e.g.Yes/No). All main events must include L1 information, while L2 and L3 are optional and included only when relevant information is available in the corresponding Wikipedia article.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

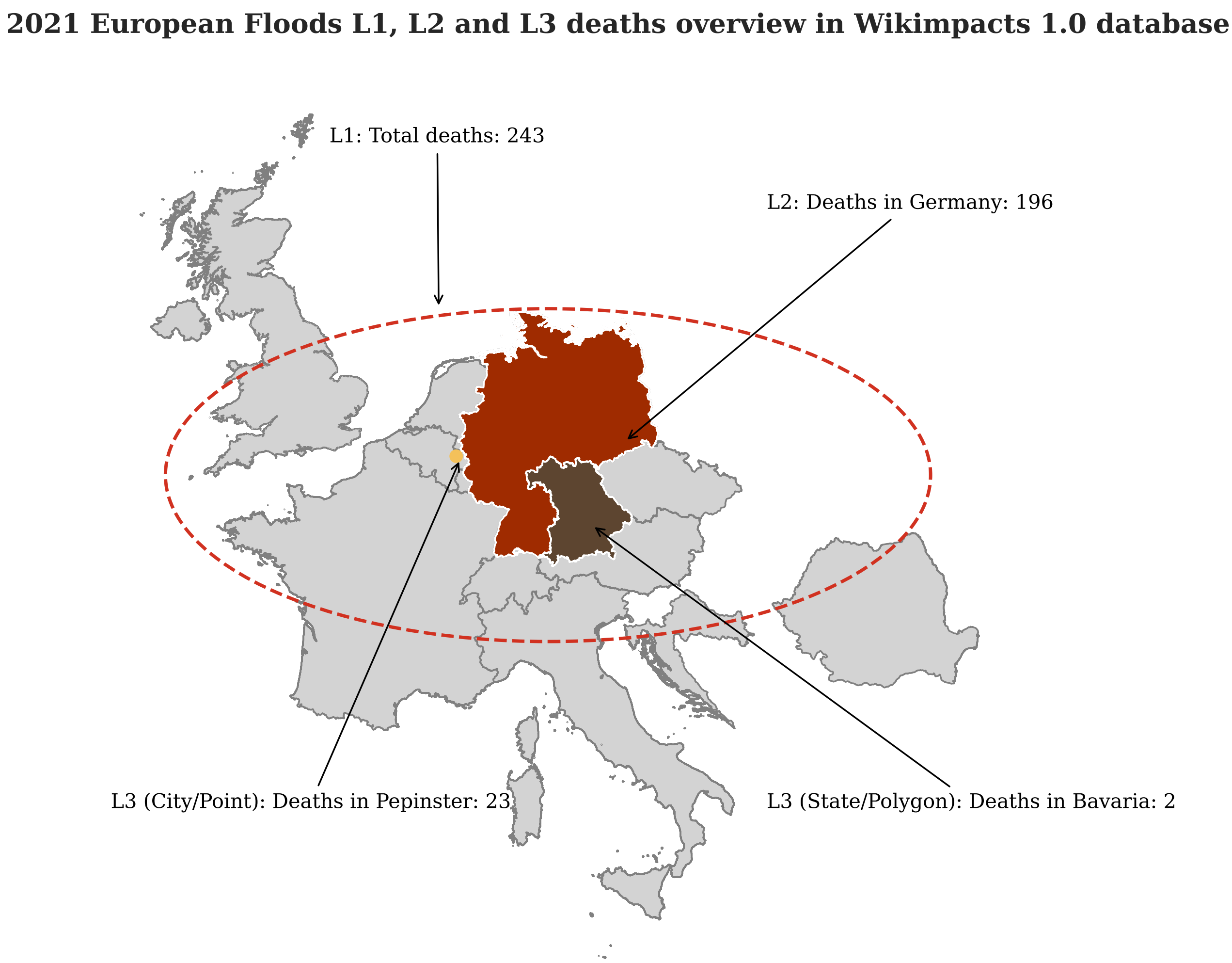

Figure 1. A simplified representation of deaths caused by the 2021 European Floods as reported in the Wikimpacts 1.0 database. For this flood event, the database includes information at L1 (event level): Total deaths in the 2021 European Floods, L2 (national level): 196 deaths in Germany, L3 (sub-national level, polygon): 2 deaths in the state of Bavaria, and L3 (sub-national level, point): 23 deaths in the city of Pepinster.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Table 1. List of the information included in the Wikimpacts 1.0 database and their definitions.

Information Type	Field	Definition
Basic	Event_ID	Unique event identifier, consistent across levels L1-L3.
Basic	Sources	Original Wikipedia link(s) of the event.
Basic	Event_Names	Name(s) of the event.
Basic	Main_Event	Unique categorisation of the event at L1 (see Table 2).
Basic	Hazards	Hazards associated with the Main_Event at L1, which refer to the potential occurrence of physical phenomena that may cause impacts (see Table 2).
Time-related	Start_Date_Day	Start day of the event at L1, or start day of the impact recorded at L2/L3.
Time-related	Start_Date_Month	Start month of the event at L1, or start month of the impacts recorded at L2/L3.
Time-related	Start_Date_Year	Start year of the event at L1, or start year of the impacts recorded at L2/L3.
Time-related	End_Date_Day	End day of the event at L1, or end day of the impacts recorded at L2/L3.
Time-related	End_Date_Month	End month of the event at L1, or end month of the impacts recorded at L2/L3.
Time-related	End_Date_Year	End year of the event at L1, or end year of the impacts recorded at L2/L3.
Location-related	Administrative_Areas_Norm	Affected countries from GADM at L1/L2.
Location-related	Administrative_Areas_Type	Administrative types of affected countries from OpenStreetMap (OSM), United Nations Statistics Division (UNSD), or Global Administrative Unit Layers (GAUL 2015) at L1/L2.
Location-related	Administrative_Areas_GeoJSON	GeoJSON format of the affected countries at L1/L2.
Location-related	Administrative_Areas_GID	GADM Global Administrative Areas IDs of the affected countries at L1/L2.
Location-related	Administrative_Area_Norm	Affected country at L3.
Location-related	Administrative_Area_Type	Administrative type of affected country from OSM , UNSD, or GAUL 2015 at L3.
Location-related	Administrative_Area_GeoJSON	GeoJSON format of the affected country at L3.
Location-related	Administrative_Area_GID	GADM Global Administrative Areas ID of the affected country at L3.
Location-related	Locations_Norm	Affected sub-national area names at L3.
Location-related	Locations_Type	Affected sub-national types from OSM at L3.
Location-related	Locations_GeoJson	GeoJSON format of the affected sub-national areas at L3.
Location-related	Locations_GID	GADM Global Administrative Areas IDs of the affected sub-national areas at L3.
Impact-related	Deaths	The number of deaths in the event. Missing people are not included as deaths.
Impact-related	Injuries	The number of non-fatal injuries in the event.
Impact-related	Homeless	The number of people made homeless by the event.
Impact-related	Displaced	The number of people displaced by the event.
Impact-related	Affected	The number of people affected by the event.
Impact-related	Buildings_Damaged	The number of buildings damaged by the event.
Impact-related	Insured_Damage	Damage from physical harm or loss to property, assets, or individuals covered under an insurance policy in the event.
Impact-related	Damage	The economic damage caused by the event.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Table 2. L1 Main event categories and associated hazards.

Main Event	**Associated Hazard(s)**
Flood	Flood
Extratropical Storm/Cyclone	Wind, Flood, Blizzard, Hail
Tropical Storm/Cyclone	Wind, Flood, Lightning
Extreme Temperature	Heatwave, Cold Spell
Drought	Drought
Wildfire	Wildfire
Tornado	Wind

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

2.1 Event Level (L1)

L1 includes both direct impacts (Deaths, Injuries, Homeless, Displaced, Affected, and Buildings Damaged) and monetary impacts (Damage and Insured Damage). All impact fields are nullable, but at least one impact must be present to report the event in our database. For the basic information about the event, fields such as Event_ID, Event_Names, Sources, Start_Date_Year, and Administrative_Areas_Norm are mandatory at this level. Here we use the Database of Global Administrative Areas (GADM) level 0 administrative area to denote the Administrative_Areas_Norm field in our database (Global Administrative Areas, 2012). This area representation may contain countries or other geographic entities. For simplicity, we hereafter refer to such a representation as either a country or as a national-level location, yet we remain neutral regarding to jurisdictional claims made in any material presented in this paper and the associated database. The impact information refers to the event's overall impact (e.g. 243 deaths in Figure 1). Whenever possible, impact information from Wikipedia articles is sourced from parts of the text which explicitly state the total impact. If aggregated impacts for the main event are not explicitly stated in the article, we aggregate data from L2 to present the total impact in L1. Fields with names ending in "Approx" indicate whether the information extracted from the Wikipedia article is precise (Table 3). Returning to our example of the 2021 European floods, the article specifies "243 deaths", and the "Approx" field this number in the database is thus marked as "False". Conversely, if the article had stated "more than 200 deaths", then the data would have been normalized to [201, 301] in the database using predefined normalization rules (see SI Section 4), and the related "Approx" field would have been marked as "True". Similarly, if the L1 impact information is inferred from L2, the related "Approx" field is also marked as "True".

2.2 National Level (L2)

L2 breaks down the impact information at national level. The Administrative_Areas_Norm field is a list, typically containing one country (20,041 entries in the database) where the total impact in that country occurred or, in rare cases, a list of countries (45 entries in the database) if the impact could not be dissociated between a subset of countries. Figure 1 provides as example the total number of deaths in Germany during the 2021 European Floods, and SI Section 6 shows that L2 contains corresponding information for other countries affected by these floods. Spatial information on the impacts is mandatory in order for them to be included in the database, while temporal information is not mandatory as the impact will likely fall within the time span specified in L1 (Table 3). National-level impact information is not always available. In some cases, the overall impact at this level is unknown, but impact information for specific locations within a country is provided. In these cases, we aggregate the information from L3 to present it in L2.

2.3 Sub-national Level (L3)

L3 provides a detailed breakdown of impact information at a sub-national level (e.g., federal state, municipality) (Global Administrative Areas, 2012; OpenStreetMap contributors, 2017a). Same as L1, we use the GADM level 0 administrative area to denote the Administrative_Area_Norm fields in this level. Figure 1 provides as example the information on number of deaths in L3 for the city of Pepinster (shown as a point) in Belgium and the state of Bavaria (shown as a polygon) in Germany. In

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

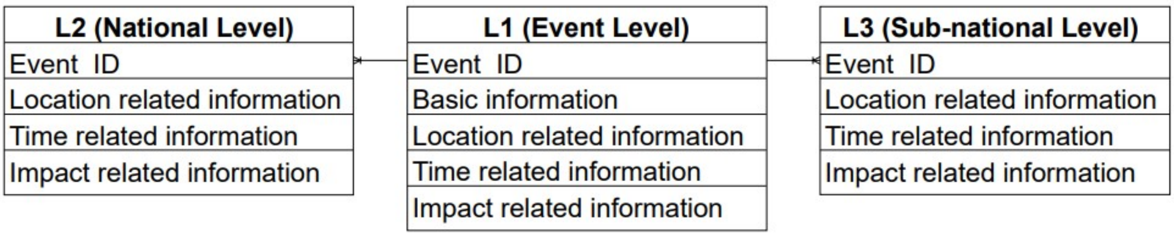

Figure 2. Wikimpacts 1.0 database structure. L1, L2 and L3 share consistent Event_ID entries. The basic information for L2 and L3 are identical to that of L1; therefore, they are only recorded in L1. For further details on the information fields in the figure see Table 1.

most cases, the impact is confined to a single location, yet there are instances where the impact spans multiple places within the country. In that case, the field Administrative_Area_Norm contains the country, and the field Locations_Norm contains a list of specific locations within that country where the impact occurred. Like for L2, spatial information is mandatory at this level, while temporal information remains optional (Table 3).

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Table 3. List of information fields in the Wikimpacts 1.0 database, their properties and the relevant database levels. DIRECT IMPACT includes deaths, injuries, displaced, homeless, affected and buildings damaged; MONETARY IMPACT includes damage and insured damage. Asterisks indicate fields that are not included in the database evaluation process.

Field	**Data Type**	**Permitted values**	**Mandatory**	**Applied Level(s)**
Event_ID	UUID	Short uuid, 7 characters	Yes	L1, L2, L3
Hazards	List[String]	String(s) from Table 2	Yes	L1
Main_Event	String	String from Table 2	Yes	L1
Event_Names	List[String]	String(s)	Yes	L1
Sources	List[String]	Valid URL(s)	Yes	L1
Administrative_Areas_Norm	List[String]	National-level administrative area names from OSM or UNSD	Yes	L1, L2
* Administrative_Areas_Type	List[String]	National-level administrative area types from OSM, UNSD, or GAUL 2015	Yes	L1, L2
* Administrative_Areas_GID	List[String]	National-level administrative area GIDs from GADM	Yes	L1, L2
* Administrative_Areas_GeoJson	List[JSON]	National-level administrative area GeoJSON objects from OSM	Yes	L1, L2
Administrative_Area_Norm	String	The national-level administrative area name from OSM or UNSD	Yes	L3
* Administrative_Area_Type	String	The national-level administrative area type from OSM, UNSD or GAUL 2015	Yes	L3
* Administrative_Area_GID	String	The national-level administrative area GID from GADM	Yes	L3
* Administrative_Area_GeoJson	JSON	The national-level administrative area/division **GeoJSON** objects from OSM	Yes	
Locations_Norm	List[String]	Area names within the specified national-level administrative area	Yes	L3
* Locations_Type	List[String]	Area types (from OSM) within the specified national-level administrative area	Yes	L3
* Locations_GID	List[String]	Area GADM GIDs within the specified national-level administrative area	Yes	L3
* Locations_GeoJson	List[JSON]	GeoJSON objects within the specified national-level administrative area	Yes	L3

continued on next page

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

continued from previous page

Field	**Data Type**	**Permitted values**	**Mandatory**	**Applied Level(s)**
Start_Date_Day	Non-negative integer	1-31	No	L1, L2, L3
Start_Date_Month	Non-negative integer	1-12	No	L1, L2, L3
Start_Date_Year	Non-negative integer	1034-2024	Yes	L1, L2, L3
End_Date_Day	Non-negative integer	1-31	No	L1, L2, L3
End_Date_Month	Non-negative integer	1-12	No	L1, L2, L3
End_Date_Year	Non-negative integer	1034-2024	No	L1, L2, L3
Total_DIRECT_IMPACT_Min	Non-negative integer	0-inf	No	L1
Total_DIRECT_IMPACT_Max	Non-negative integer	0-inf	No	L1
*Total_DIRECT_IMPACT_Approx	Boolean	True, False	No	L1
Total_MONETARY_IMPACT_Min	Non-negative integer	0-inf	No	L1
Total_MONETARY_IMPACT_Max	Non-negative integer	0-inf	No	L1
*Total_MONETARY_IMPACT Approx	Boolean	True, False	No	L1
Total_MONETARY_IMPACT_Unit	String	ISO 4217 currency	No	L1
Total_MONETARY_IMPACT Inflation_Adjusted	Boolean	True, False	No	L1
Total_MONETARY_IMPACT Inflation_Adjusted_Year	Non-negative integer	1034-2024	No	L1
Num_Min	Non-negative integer	0-inf	Yes	L2, L3
Num_Max	Non-negative integer	0-inf	Yes	L2, L3
* Num_Approx	Boolean	True, False	Yes	L2, L3
Num_Unit	String	ISO 4217 currency code	Yes	L2, L3
Num_Inflation_Adjusted	Boolean	True, False	No	L2, L3
Num_Inflation_Adjusted_Year	Non-negative integer	1034-2024	No	L2, L3

3 Wikimpacts Processing Pipeline

The Wikimpacts processing pipeline comprises four modules (Figure 3). First, we select relevant articles for processing. Next, we apply a list of prompts to extract the necessary information. After extraction, we post-process the raw output to represent the data in standardized formats. Finally, we check for data consistency across the three levels, address missing information, convert currencies, adjust inflation, and format the database for ease of use. Each module is described in detail below.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

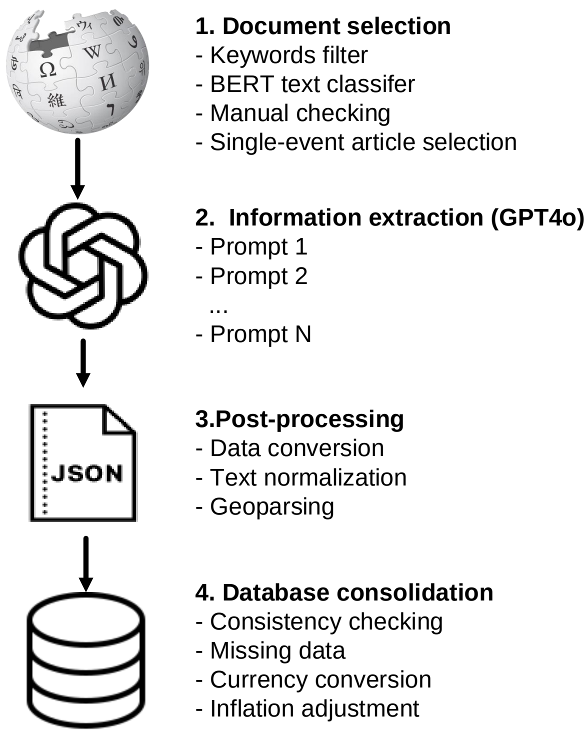

Figure 3. Full pipeline of Wikimpacts 1.0 database construction.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

3.1 Document Selection

A three-step approach is used to select relevant English articles. First, we craft a keyword list covering all major event categories in the database, which we then use to extract relevant English Wikipedia articles (see Supplementary Information (SI) Section 1 for keywords). Querying Wikipedia using these English keywords (with a cut-off date at 29/02/2024), results in 30,085 articles. However, not all articles retrieved through this keyword extraction process are related to climate events. For instance, some refer to topics like "Miami Hurricanes football"[8]. Therefore, in a second step, to quickly obtain the related articles automatically, we apply a text classifier[9] (Devlin et al., 2019; Sanh et al., 2019) to filter non-climate-related articles. To this end, the pre-trained English distilled Bidirectional Encoder Representations from Transformers (BERT) model is fine-tuned on a set of 300 Wikipedia articles, containing 248 relevant and 52 irrelevant articles.[10] Using 150 articles for training, 100 articles for validation, and 50 articles for testing (with the relevance distribution shown in Table 4), we obtain an F1-score of 98.8 on the test set (with a precision score of 97.7 and a perfect recall score of 100.0; see SI Section 2 for definitions of F1-score, precision, and recall). From the original 30,085 articles, we classify 4,900 as relevant in this second step. Thirdly, we manually check all these classified articles to confirm their relevance. We identify 184 false positives in the set of 4,900 articles and another 330 false negatives in the remaining 25,185 articles. In the end, we identify 5,046 English Wikipedia articles as relevant for further processing.

Table 4. Article relevance distribution for the 300 English Wikipedia articles used to fine-tune the BERT model for text classification.

Data Set	Relevant	Irrelevant
Training Data	128	22
Validation Data	78	22
Test Data	42	8

It should be highlighted that some Wikipedia articles describe only one event, e.g., Hurricane Ida[11], while other articles cover a series of events, such as the 2021 Atlantic hurricane season[12]. We refer to the former as the "single-event articles" and to the latter as "multi-event articles". Multi-event articles present specific challenges. For some events, they serve as the sole source of information on Wikipedia, while for others, there exist dedicated "single-event articles" that provide more detailed information and should be used as the basis for those events. Moreover, the structure of multi-event articles differs significantly from that of single-event articles due to the number of events they cover, requiring a further processing step. To address this, we post-process the 5,046 articles to identify single- and multi-event articles. Using the GPT4o Mini model[13], we extract relevant climatic events from the full set of 5,046 Wikipedia articles. This yields 6,625 events. We then conduct the reverse process, and search Wikipedia to locate the relevant articles for those events. In total, we identify 3,368 events mapped to a unique

[8] https://en.wikipedia.org/wiki/Miami_Hurricanes_football
[9] DistilBert for sequence classification, accessed via the HuggingFace platform at https://huggingface.co/docs/transformers/en/model_doc/distilbert
[10] For articles longer than 512 tokens, only the first 512 tokens are used.
[11] https://en.wikipedia.org/wiki/Hurricane Ida
[12] https://en.wikipedia.org/wiki/2021 Atlantic hurricane season
[13] gpt-4o-mini-2024-07-18

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Wikipedia article. The remaining 3,257 events are linked to Wikipedia articles already identified as sources for at least one other climate event, suggesting that those articles are multi-event articles. In the remainder of the paper, we focus exclusively on those 3,368 events mapped to a single-event article to construct the Wikimpacts 1.0 database.

3.2 Information Extraction Using GPT4o Model

A core component of our database construction pipeline is the application of the GPT4o model[14]. Different from initial trials (Li et al., 2024), we use the GPT4o model instead of the GPT4 model due to the longer context window (Hurst et al., 2024), which enables it to take in a larger number of input tokens to process. To extract information from Wikipedia articles, we feed the full text and the information box (if it exists) to the GPT4o model together with a set of prompts corresponding to the different fields of our database. To facilitate post-processing, we instruct the GPT4o model to provide output in JSON. However, one issue we faced was that the model sometimes cannot produce valid JSON objects – this can occur due to the longer output length that exceed the total number of permitted output tokens, as found by Li et al. (2024). This issue commonly occurs when we instruct the model to return the text segment where it identifies the relevant information as a traceable source in the output but the text segment, as taken verbatim from Wikipedia, is too long for the model to give a complete JSON output. To mitigate this issue, each Wikipedia article is presented as a JSON file where each key is the header title of the section in the Wikipedia article and each value is the verbatim text as it appears in Wikipedia. Compared to other text sources, Wikipedia articles are well-structured, enabling the extraction of the full text in the header-content pair format described earlier. In this setup, and for all fields, we instruct the model to provide a source section that includes the original section headers from the Wikipedia article where the information is found. This strategy also helps to prevent model hallucinations (Tonmoy et al., 2024) since the model does not need to return large blocks of source text when extracting information on extreme climate impacts.

To obtain location and time information about the event in L1, we build four prompts: two for the relevant information, and two for the source section of this information. To obtain the location information, we ask the model to capture all locations affected by the event and retrieve the affected countries during post-processing. To obtain the main event category and the reported hazards, we provide the model with the list of main event categories and the hazards associated with each category (Table 2). Following this, we pose four prompts: the first for the event category, the second for the source section of the event category, the third for the associated hazards based on the result of the event category, and the last for the source section of the hazards. We then use more complex prompts to extract information on different impacts. These prompts partly rely on keywords used for categorising the impacts. These same keywords are also used in the annotation process (Sect. 4.1). The L1 information represents the total impact of the event, for which we ask two questions: one regarding the total impact and another pertaining to its source section. This is followed by information extraction at L2 and L3, as well as specifics on times and locations, if such details are available. Additionally, we prompt the model to capture L1 impact information only when explicitly stated in the text, such as in the 2021 European Floods example: “At least 243 people died in the floods”. If this information is not provided explicitly, the model is tasked to return “NULL” rather than summing individual data entries to produce a total for L1. Similarly, for L2, the model is tasked to return the total impact for specific countries only when explicitly

[14]gpt-4o-2024-05-13

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

available. If unavailable, it should not aggregate data from various locations within a country but instead return "NULL". We also evaluate the performance of different prompt settings for information extraction. Our prompt design, evaluation and the full text of the prompts used for full run production can be found in SI Section 3. We use what is termed "prompt v3.1" in the SI for all information categories, except the L1 location, where we use "prompt v3.2".

3.3 Post-Processing

The outputs produced by the GPT4o model contain "raw" data (e.g. dates represented in a variety of formats or ambiguous location names) that often requires normalization i.e., conversion to a standardized format. We apply a set of normalization rules (see SI Section 4) to the raw data, ensuring that the post-processed model output can be evaluated and stored in a standardized format in the database. Overall, we normalise the main event, hazards, time, location, and numerical data prior to the evaluation. The detailed post-processing steps are listed below.

3.3.1 Main Event and Hazards

The model is prompted to identify the unique Main Event of each article, and the goal of the normalization is to validate that the model extracts a single Main Event belonging to one of the categorical variables, and that all extracted hazards are associated with that particular Main Event category (see in Table 2). In case of multiple relevant Hazards, we prompt the model to split with "|", and in the normalization process, we convert it to a list.

3.3.2 Time

Dates extracted by the LLM appear in various locales or formats with some components missing (for example, the month and year may be known but not the day). Dates in their different locales, whether partial or complete, are standardized using `dateparser` (DateParser contributors, 2024) in Python.

3.3.3 Location

The model is prompted to identify locations at different administrative levels (such as countries, cities, or regions) for the three database levels (L1, L2, and L3). In v3.2, the model is prompted to identify a list of countries affected by the main event in question, thus providing L1 information. In v3.1, For L2, the model is prompted to produce a list of national-level locations where an impact could be quantified. For L3, the model is prompted to identify a single administrative area at the national level and to capture smaller administrative areas associated with the impact within that country. The normalization pipeline tries to disambiguate these locations using the Nominatim API[15] to search for locations on OpenStreetMap (OpenStreetMap contributors, 2017a), an open geographical database.

Location names are extracted verbatim from the article by the LLM, making the output prone to contain locations expressed in various spelling conventions or colloquial names. Often, the retrieved text may refer to locations that are under dispute or not

[15] https://nominatim.org/release-docs/develop/api/Overview/

recognized internationally. To mitigate this, we normalize all locations so that they fit within a single standard representation. The standard format for a normalized location is split over 4 fields representing:

1. the location's official English name (`_Norm`) whenever available;
2. the administrative address level or type (`_Type`) as defined by OpenStreetMap and returned in the Nominatim API raw output as defined by OpenStreetMap contributors (2017b), or GAUL 2015 [16];
3. the GADM GID (`_GID`)(Global Administrative Areas, 2012)[17]; and
4. a GeoJson object (`_GeoJson`) to visually represent the area on a map with a valid GeoJson type (such as `Polygon`). In Wikimpacts 1.0 database, GADM 4.1 version is used.

Location and region names are disambiguated using OpenStreetMap (OpenStreetMap contributors, 2017a) and the UNSD M49 dataset[18]. When extracting GeoJson objects from OpenStreetMap, GeoJson shapes other than `Point` are preferred whenever available, but the pipeline falls back to `Point` if nothing better is available.

3.3.4 Numerical Data

Often, LLM output extracts numerical information with phrasing that renders it open to interpretation (e.g., "No less than 12 people were injured" or "Billions of dollars were paid in damages"). It may also extract single numbers expressed in different locales dictating whether decimals use periods or commas. The normalization process aims to transform such expressions into quantifiable and standardized formats. Normalized numbers are represented in a range spread over three columns: $\langle min, max, \texttt{approximation}\rangle$ where "approximation" is a boolean representing whether the information is an exact number or an approximation of the exact number.

In short, the normalization process for numbers automatically checks if an immediate conversion of the expression to a number or range is possible (e.g., "1,421" or "20-30" can quickly be converted to $\langle 1421, 1421, \texttt{False}\rangle$ or $\langle 20, 30, \texttt{True}\rangle$). If not, the normalization script checks for any quantifiers such as "tens of thousands of homes were destroyed" or "No less than 20 deaths" and converts them into a range (the two previously mentioned examples would be normalized to $\langle 20000, 90000, \texttt{True}\rangle$ and $\langle 20, 30, \texttt{True}\rangle$. A list of synonyms is used to determine whether or not a number is an approximation. This rule-based approach also employs part-of-speech tags and entities identified by SpaCy's [19] English transformer pipeline model to extract min and max values for more complicated expressions. The rules we applied in this step for normalizing different expressions into ranges are described in detail in SI Section 4.

[16] Verison of GAUL 2015 : https://data.apps.fao.org/catalog/dataset/gaul-code-list-global-admin-1
[17] Not all locations can be normalized to a GID given the limited level depth of GADM which may not represent small towns or may not group larger unofficial or disputed regions
[18] https://unstats.un.org/unsd/methodology/m49/
[19] https://spacy.io/

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

3.4 Database Consolidation

During the consolidation process, we filter out events that do not fall within our predefined main event or hazard categories, such as “geomagnetic storm”, or “landslide”. Upon constructing the initial database, we identify missing information at various levels. For instance, the attribute Total_Deaths for a particular event might be recorded as NULL at L1, while at L2, there could be an entry indicating 20 fatalities in Germany. Furthermore, the cumulative impact at L3 within a single country might exceed the documented impact at L2 for the same country, and similarly, aggregated L2 data might provide larger values than the information available at L1.

To address these discrepancies and ensure data consistency across the database, we adopt a bottom-up approach beginning with L3. We sum L3 impact values for a given country and compare these to L2. If L2 provides a range, we adjust the minimum and maximum of the range if necessary. If L2 provides a single number, we transform this into a range if it is lower than the aggregated information from L3. We repeat the same procedure for L1, by aggregating impact values from L2. Detailed rules and procedures are provided in SI Section 5.

In addition to addressing these data inconsistencies, we standardize currencies and adjust for inflation throughout the database, choosing USD as the base currency. Using the currency statistics from the Wikimpacts 1.0 database, we obtain conversion rates for most non-USD currencies from a publicly available resource (Antweiler). For periods before a currency’s available data, a constant currency conversion rate is applied as the earliest available year. Additionally, we include EUR as a secondary standardized currency, with USD-2024-inflation-adjusted values converted to EUR using the 2024 average conversion rate.

Our approach to inflation adjustments follows the same rules as those documented by EM-DAT.[20],[21] In Wikimpacts 1.0, all monetary values are adjusted to reflect 2024’s inflation rates, except for events occurring in 2024, which are left unadjusted for inflation. Detailed information on these adjustments is provided in SI Section 5.

4 Evaluation of the Pipeline

4.1 Data Annotation

As part of Wikimpacts 1.0, we develop a gold standard dataset by manually annotating Wikipedia articles. This gold standard dataset includes a development set containing 70 main events (used to develop the information extraction pipeline) and a test set containing 156 main events (used exclusively for evaluation of the GPT4o model output). The disaster type distribution of these two sets is shown in SI Section 8. Compared to preliminary results from Li et al. (2024), part of these annotated data now include L2 and L3 information. In the development set, we have 55 events with L2 and L3 information annotated, while in the test set, there are 97 events annotated with this additional level of impact information. An error rate for each field provided by the GPT4o model is calculated by comparing the LLM output with the gold standard.

[20] https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1800-
[21] https://doc.emdat.be/docs/protocols/economic-adjustment/

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

We recognise that manual annotation is not error-proof. To ensure consistent and robust annotation, we provided the annotators with a list of keywords for detecting impacts (see Table 5). We further established comprehensive logical normalization rules for the annotators to follow. These correspond to the post-processing rules for the Wikimpacts database (see SI Section 4). To verify consistency between different annotators, two annotators blindly double-annotate 10 articles. The internal annotator agreement scores are discussed in Sect. 4.1.1.

4.1.1 Internal Annotation Agreement Evaluation

The quality of our gold data is assessed using 10 articles annotated independently by two different annotators. One annotator provides annotations without classifying L2 and L3 information, from which we infer L2 and L3 levels based on location information annotated at either the national or sub-national level. The second annotator annotates these same articles, explicitly defining L2 and L3 information.

For L1, both annotators extract identical information, resulting in error rates of "0" across all fields, as shown in SI Section 8. However, in the L2 and L3 evaluations, some discrepancy between the two annotators is apparent. These discrepancies depend on differences in the number of annotated entries, resulting from the two annotators interpreting the text differently (see SI Section 8). For L2 annotations, some cases involve one annotator transferring information from L3 into L2. In L3 annotations, for instance, in the event "Cyclone Vayu"[22], one annotator creates an L3 entry with "Locations" as "Ullal&India|Gujarat&India" and "Buildings Damaged" as "15", while the other annotator records the same event with "Locations" as "Ullal&India". Upon examining the original text, we find that both interpretations are logical: one annotator retained "Ullal", a location mentioned in the impact sentence, while the other included "Gujarat", mentioned in the leading sentence of the related impact in the paragraph. These variations in interpretation contribute to inter-annotator agreement errors.

4.2 Evaluation Methods

We evaluate our database using the above-described test set from the gold standard data. The evaluation involves all three levels of information. The information extracted for each main event is complex since all three levels contain many fields, making evaluation challenging. To obtain an overall aggregated score for each event, as well as scores for specific fields, we define a difference error metric for each field, ranging from 0 to 1 (where lower values indicate better performance). We then calculate an aggregated score as a weighted sum of these field-specific scores:

$$D(a,r) := \frac{1}{n}\sum_{i} w_i d_i(a_i, r_i) \qquad (1)$$

$D(a,r)$ is the difference between a gold entry a and an LLM output entry r, with weights w_i and difference metrics d_i of fields i, where n is the number of fields. This approach allows to adjust the relative influence of each field by modifying its weight. Since the importance of the different fields is user-dependent, in this paper we present evaluation results with an equal weighting of all fields.

[22] https://en.wikipedia.org/wiki?curid=61000334

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Table 5. Keywords used for identifying impact fields in the annotation process of Wikimpacts 1.0 database and in some of the prompts (See SI Section 3).

Variable	**Keywords**
Deaths	die, dead, killed, fatality, lost lives, perished, passed away
Injuries	injured, hurt, wound, hospitalized
Homeless	lost home, homeless, household damage, household destroy, house damage, home destroy, unhoused, without shelter, houseless, shelterless
Displaced	evacuated, displace, transfer/move to shelter, relocated, flee
Affected	affect, impact, influence
Buildings_Damaged	home, house, household, building, apartment, apartment block, school, church, office buildings, retail stores, hotels, hospitals, dwellings, structures
Insured_Damage	insurance, insured
Damage	damage, economic, economy

The difference metrics for specific fields are defined based on metrics for the following basic types: numbers, strings, booleans, and lists.

– For (non-negative) numbers:

$$d_n(a,r) := \begin{cases} 0, & \text{if } a = r \\ \frac{|a-r|}{a+r}, & \text{otherwise} \end{cases} \tag{2}$$

– For strings and booleans:

$$d_{t,b}(a,r) := \begin{cases} 0, & \text{if } a = r \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

– For lists:

$$d_s(a,r) := 1 - \frac{|a \cap r|}{|a \cup r|} \tag{4}$$

Rather than using more conventional evaluation metrics (such as accuracy, recall, or precision), we opt to use metrics tailored to the database's specific application: representing climate extremes and their impacts. For instance, if the correct number of deaths is 10, a prediction of 11 would result in a minor error, whereas a prediction of 100 would be a substantial error. Under the current metric, these predictions receive a normalized error rates of 0.048 and 0.818, respectively.

In this paper, we evaluate the fields from GPT4o model output without an asterisk (the fields are derived through post-processing rather than representing the raw output from the LLM) in Table 3. For the evaluation of L1, the LLM output is

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

automatically matched with the gold standard using the Event_ID since only a single entry is retrieved per article in L1. However, for L2 and L3, the number of entries extracted by the LLM may vary, compared to the gold standard. For instance, for the same event there may be 5 entries in L2 from the gold standard, but 10 entries in the LLM output. To address this, we implement a matching algorithm that identifies the most similar entries to the LLM output and the gold standard during the L2/L3 evaluation process. The matching algorithm uses the same evaluation metrics as above to determine the overall similarity between all the entries from the LLM output and the gold standard. The best matching pairs for each entry in the LLM output and the gold standard is then selected. For non-matching entries (either in the gold data or the LLM output), we construct an empty entry padded with "NULL" values. This results in two lists of entries of equal length, which is the desired format for evaluation. The similarity between two entries is defined as $1 - d(a,r)$. In the matching algorithm, it is possible to select different values for some important parameters:

- the similarity threshold under which entries are not considered to be a good match
- the weights for different fields used when matching
- the null penalty, which is the error value assigned in the difference metrics when one of the two entries contains a "NULL" value

We use 55 events from the gold standard development set to select the algorithm setting. For the rest of the evaluation, we use the parameter set termed "Setting 2" in SI Section 8).

4.3 Evaluation Results

Table 6 shows that the model performs consistently well across all L1 fields. For basic information, the error rates for Main Event and Hazards are 0.0256 and 0.2004, respectively, indicating that the model effectively captures robust information for the Main Event and comparatively reliable information for the associated hazards. Regarding time-related information, the model achieves near-perfect performance, with error rates ranging from 0.0003 to 0.0463. This indicates that time information in our database is a highly robust representation of the information contained in Wikipedia. However, the location information exhibits a higher error rate of 0.4843. For the impact categories, Total Deaths has the lowest error rates, ranging from 0.0236 to 0.0374; Total Damage and Total Insured Damage also show low error rates of approximately 0.07 and 0.012, respectively. For other impact categories, error rates range from 0.2118 to 0.311, indicating that the LLM encounters difficulties in capturing this information from Wikipedia articles.

Tables 7 and 8 present the evaluation results for L2 and L3 on the test set. For these two levels, the model's performance on the test set is comparable to its performance on the development set (see SI Section 3 and 8). Next to the Weighted_Score in each impact category, the error rates for individual fields within the impact categories are presented in these tables. The location field Administrative_Areas_Norm exhibits the highest error rate across all impact categories in L2. Similarly, in L3, both Administrative_Area_Norm and Locations_Norm display higher error rates compared to other fields. Notably, time-related information in both L2 and L3 has relatively low error rates, which are generally lower than the Weighted_Score.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

For impact information fields such as Num_Min and Num_Max, L2 generally achieves lower error rates compared to L3. Referring to the Weighted_Score across all impact categories, the information in L2 is more robust than that in L3 within our database. Furthermore, across impact categories, Injuries, Homeless, and Displaced exhibit relatively lower error rates than other categories.

Overall, in our database, L1 information thus provides the most reliable representation of the underlying Wikipedia article, followed by L2 and L3. Within L1, event and timing data are highly accurate, while location data is less robust. The Deaths category in L1 has the lowest error, followed by Damage and Insured Damage, with other categories showing higher errors. In L2 and L3, the injuries, homeless, and displaced categories are more reliable.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Table 6. The L1 evaluation results on the gold standard test set. The Weighted Score represents the average across all fields, given an equal weighting. For instance, the score of 0.0256 in the "Main_Event" field corresponds to the average score for all 156 test set events within this category. Numbers closer to 0 indicate a close match between two entries, while numbers closer to 1 indicate a poorer match.

Field	**Score**
Weighted_Score	0.1431
Main_Event	0.0256
Hazards	0.2004
Start_Date_Day	0.0299
Start_Date_Month	0.0115
Start_Date_Year	0.0003
End_Date_Day	0.0463
End_Date_Month	0.0194
End_Date_Year	0.0066
Administrative_Areas_Norm	0.4843
Total_Deaths_Min	0.0374
Total_Deaths_Max	0.0236
Total_Injuries_Max	0.2118
Total_Injuries_Min	0.2115
Total_Homeless_Min	0.2559
Total_Homeless_Max	0.2528
Total_Displaced_Min	0.2950
Total_Displaced_Max	0.2963
Total_Affected_Min	0.3110
Total_Affected_Max	0.2993
Total_Buildings_Damaged_Min	0.2827
Total_Buildings_Damaged_Max	0.2797
Total_Insured_Damage_Min	0.1218
Total_Insured_Damage_Max	0.1218
Total_Insured_Damage_Unit	0.1474
Total_Insured_Damage_Inflation_Adjusted	0.1667
Total_Insured_Damage_Inflation_Adjusted_Year	0.0064
Total_Damage_Min	0.0706
Total_Damage_Max	0.0729
Total_Damage_Unit	0.0321
Total_Damage_Inflation_Adjusted	0.1026
Total_Damage_Inflation_Adjusted_Year	0.0128

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Table 7. Results of the L2 evaluation on the test set, for each field within the impact categories. The Weighted Score represents the average across all fields in a given impact category, which are given an equal weighting. For example, the Weighted Score "0.4221" for "Deaths" is the mean error of all the fields in this category.

	Deaths	Injuries	Homeless	Displaced	Affected	Buildings_Damaged	Insured_Damage	Damage
Weighted_Score	0.4221	0.3171	0.3928	0.3709	0.4695	0.5125	0.5308	0.4772
Start_Date_Day	0.2310	0.2479	0.3051	0.3321	0.4247	0.4860	0.8210	0.5542
Start_Date_Month	0.2310	0.2479	0.3051	0.3321	0.4219	0.4832	0.8198	0.5550
Start_Date_Year	0.5446	0.3277	0.3517	0.3931	0.4665	0.5902	0.8198	0.6906
End_Date_Day	0.2244	0.1849	0.2288	0.1870	0.2987	0.4039	0.7477	0.4830
End_Date_Month	0.2244	0.1849	0.2288	0.1870	0.2946	0.4012	0.7477	0.4868
End_Date_Year	0.2508	0.2101	0.2331	0.1985	0.3163	0.4037	0.7523	0.4981
Administrative_Areas_Norm	0.7665	0.9076	0.9576	0.9351	0.9617	0.9358	0.9775	0.7887
Num_Min	0.6618	0.2724	0.4619	0.3864	0.5208	0.4543	0.1570	0.4030
Num_Max	0.6643	0.2702	0.4633	0.3865	0.5208	0.4541	0.1573	0.4023
Num_Unit	NA	NA	NA	NA	NA	NA	0.1757	0.4151
Num_Inflation_Adjusted	NA	NA	NA	NA	NA	NA	0.1892	0.4415
Num_Inflation_Adjusted_Year	NA	NA	NA	NA	NA	NA	0.0045	0.0075

Table 8. The L3 evaluation results of the test set, presented following the same format as in Table 7.

	Deaths	Injuries	Homeless	Displaced	Affected	Buildings_Damaged	Insured_Damage	Damage
Weighted_Score	0.4896	0.4228	0.4582	0.4684	0.5022	0.6031	0.5788	0.5371
Start_Date_Day	0.3222	0.2486	0.2699	0.3190	0.3081	0.5314	0.8110	0.6202
Start_Date_Month	0.3194	0.2486	0.2752	0.3178	0.3100	0.5326	0.8110	0.6260
Start_Date_Year	0.5916	0.3728	0.2888	0.4289	0.3161	0.6591	0.8171	0.6279
End_Date_Day	0.2728	0.1909	0.1962	0.1907	0.2207	0.4558	0.7530	0.5465
End_Date_Month	0.2743	0.1908	0.1962	0.1929	0.2204	0.4552	0.7530	0.5543
End_Date_Year	0.2915	0.1965	0.1962	0.2031	0.2249	0.4692	0.7561	0.5562
Administrative_Area_Norm	0.7225	0.9162	0.9646	0.8858	0.9757	0.8629	0.9939	0.9806
Locations_Norm	0.8528	0.9552	0.9714	0.9387	0.9886	0.9194	0.9939	1.0000
Num_Min	0.6245	0.4541	0.6117	0.6056	0.7288	0.5734	0.1925	0.3450
Num_Max	0.6247	0.4546	0.6115	0.6016	0.7289	0.5721	0.1921	0.3450
Num_Unit	NA	NA	NA	NA	NA	NA	0.2134	0.3837
Num_Inflation_Adjusted	NA	NA	NA	NA	NA	NA	0.2317	0.3876
Num_Inflation_Adjusted_Year	NA	NA	NA	NA	NA	NA	0.0061	0.0097

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

5 Wikimpacts 1.0 Content of the Database

The Wikimpacts database, version 1.0, encompasses a total of 2,928 events. These correspond to the subset of the 3,368 events, each mapped to a single-event article (Sect. 3.1) for which all mandatory fields were completed. At the event level (L1), tropical cyclones are the dominant event type, constituting 59.39% of events in the dataset (Figure 4a). They are followed by floods (12.23%); tornadoes, wildfires, and extratropical storms also collectively account for a substantial portion of events. Droughts and extreme temperatures are less frequently recorded. The national level (L2) contains a total of 18,233 data entries and exhibits a similar distribution as L1 across event categories, although the share accounted for by tropical cyclones is reduced (Figure 4b). At the sub-national level (L3), there are 36,394 data entries, with tropical cyclones again comprising the largest share of recorded entries at 67.55%, followed by floods at 9.92% and tornadoes at 9.24% (Figure 4.c).

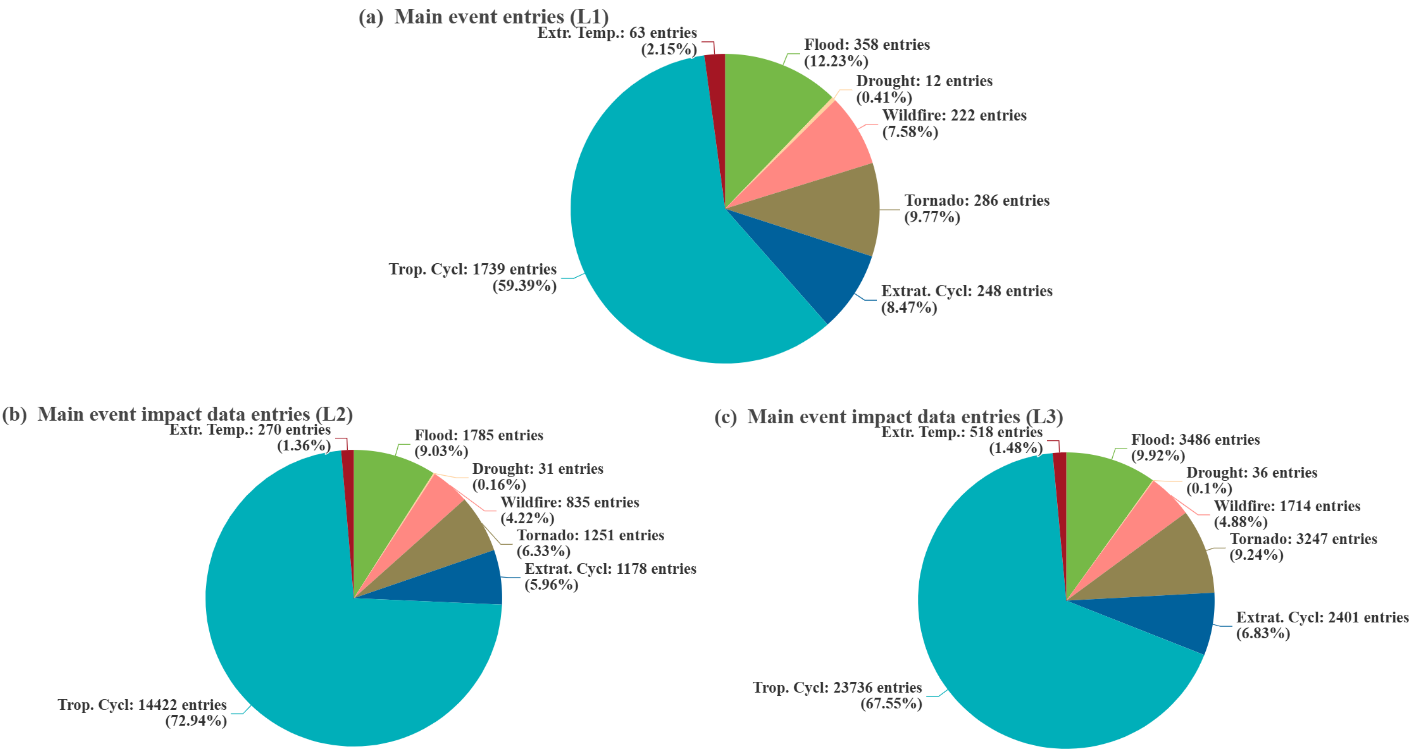

Figure 4. Statistics overview of Wikimpacts 1.0. (a) number of main events in L1, (b) number of impact data entries in L2 (national level), (c) number of impact data entries in L3 (sub-national level). Abbreviations are as follows: Extratropical Storm/Cyclone (Extrat. Cycl), Tropical Storm/Cyclone (Trop. Cycl), and Extreme Temperature (Extr. Temp). These abbreviations are also used in subsequent figures 5 and 6.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

5.1 L1 (Event Level)

5.1.1 Temporal Distribution

Figure 5a presents the decadal trends of the number of events in Wikimpacts 1.0, with our database encompassing data from the years 1034 through 2024. Although entries from the early period, spanning the 1030s to the 1890s (Figure 5b), are limited, a discernible upward trend emerges in the 1850s. The number of recorded events continues to increase steadily until the 2010s. However, due to the 2020s data only covering January 2020 to February 2024, the number of events for the current decade is lower (Figure 5a). This upward trend is evident across all main event categories. Further research is needed to disentangle the potential causes for this increase (e.g., improved reporting, rising number of events, increased exposure).

5.1.2 Impact Distribution

Tropical storms are the most frequently recorded events, and they indeed dominate the aggregated impacts for all impact categories except for number of injuries. (Figure 6). However, there are several cases in which much less frequent main event categories display comparable aggregated impacts. For instance, droughts contribute to almost as many deaths as tropical cyclones, despite being over 140 times less frequent in our database (Figure 6a). The single most severe drought event reported is the "1983–1985 famine in Ethiopia"[23], which led to approximately 1.2 million deaths. Floods and tropical storms also result in substantial numbers of deaths, and they are also among the most frequently recorded events in the database. Floods are also notable for causing the highest number of injuries (Figure 6b), followed by tropical storms and wildfires. There are no injury entries for droughts in our database. Floods rank second for the displaced and homeless impact categories, with extreme temperatures and tropical storms also playing a significant role (Figure 6c-d). Extreme temperatures rank second for total number of affected people and total damage (Figure 6e, h), while extratropical cyclones rank second for buildings damaged and insured damage (Figure 6f, g). Overall, tropical storms thus dominate the impacts recorded in our database, followed by floods. Nonetheless, all of the other main event categories also display substantial impacts in specific impact categories.

5.1.3 Spatial Distribution

The database encompasses events globally, with the US (1,245 events) exhibiting the highest number of occurrences (Figure 7a). Mexico, Canada, the Philippines, China, and Japan follow with 404, 337, 325, 300, and 279 events, respectively. Cuba has 182 recorded events, and Australia has 167 events, followed by Vietnam with 155, India with 150, and the United Kingdom with 142. There are comparatively fewer entries from the Global South [24], particularly in Africa and South America. While event entries remain limited also in Europe and Southeast Asia, they are more numerous than those for African countries.

The spatial distribution of main events for individual event categories (Figure 7b-h) mirrors the overall spatial pattern (Figure 7a). Tropical storms are documented in most countries, with exceptions including Argentina, Inner Mongolia, and several

[23] https://en.wikipedia.org/wiki/1983-1985_famine_in_Ethiopia

[24] Comprising Africa, Latin America and the Caribbean, Asia excluding Israel, Japan, and South Korea, and Oceania excluding Australia and New Zealand, according to UN Trade and Development

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

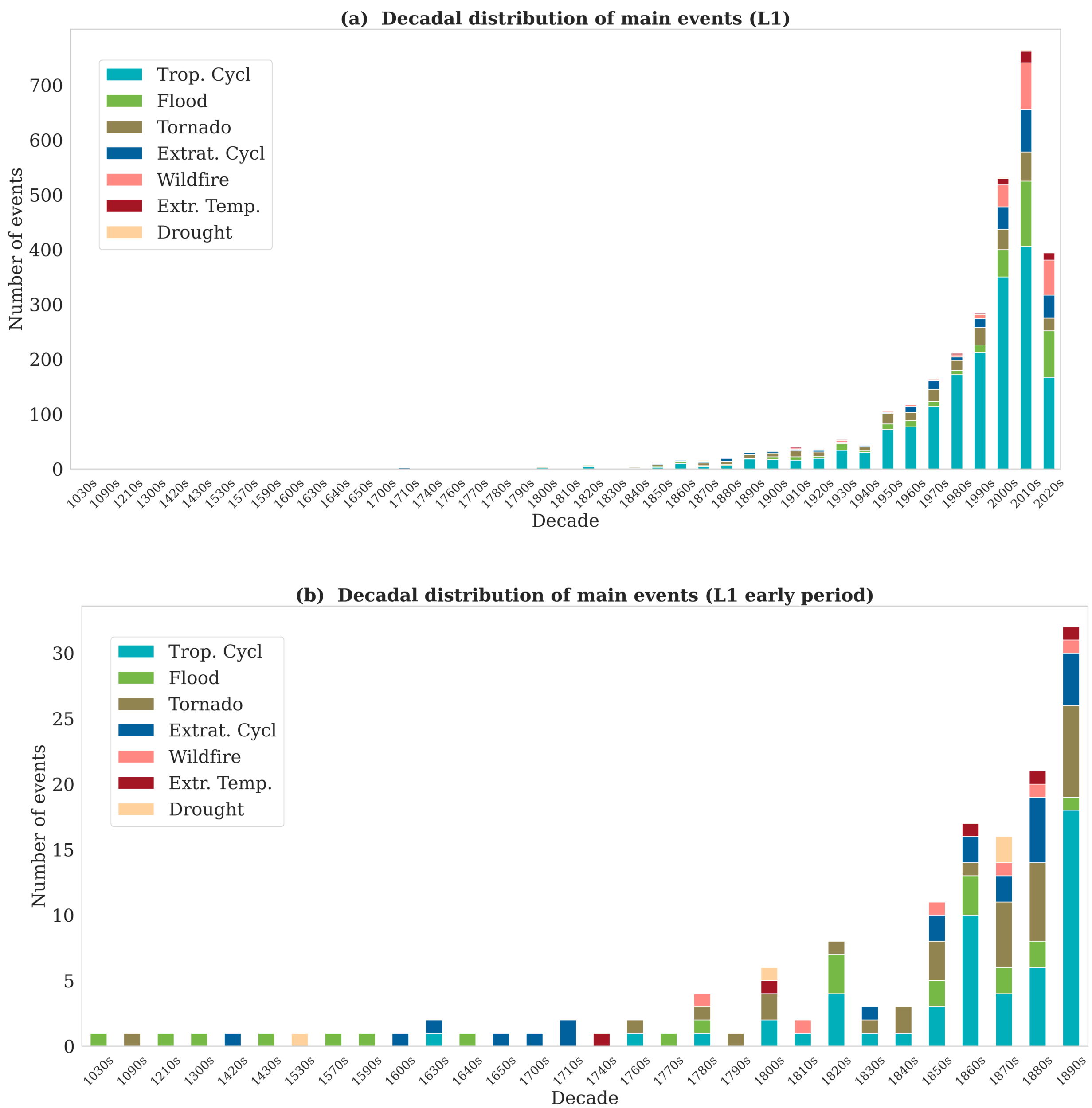

Figure 5. The temporal distribution of main events (L1) in Wikimpacts 1.0. (a) The decadal distribution of main events in the Wikimpacts 1.0 spanning all decades, from the 1030s to the 2020s, (b) The decadal distribution of main events in Wikimpacts 1.0 during the early period, covering the 1030s to the 1890s. Note the discontinuous x-axis scale for the 1030s–1760s in both panels and the fact that the 2020s only include data up to February 2024.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

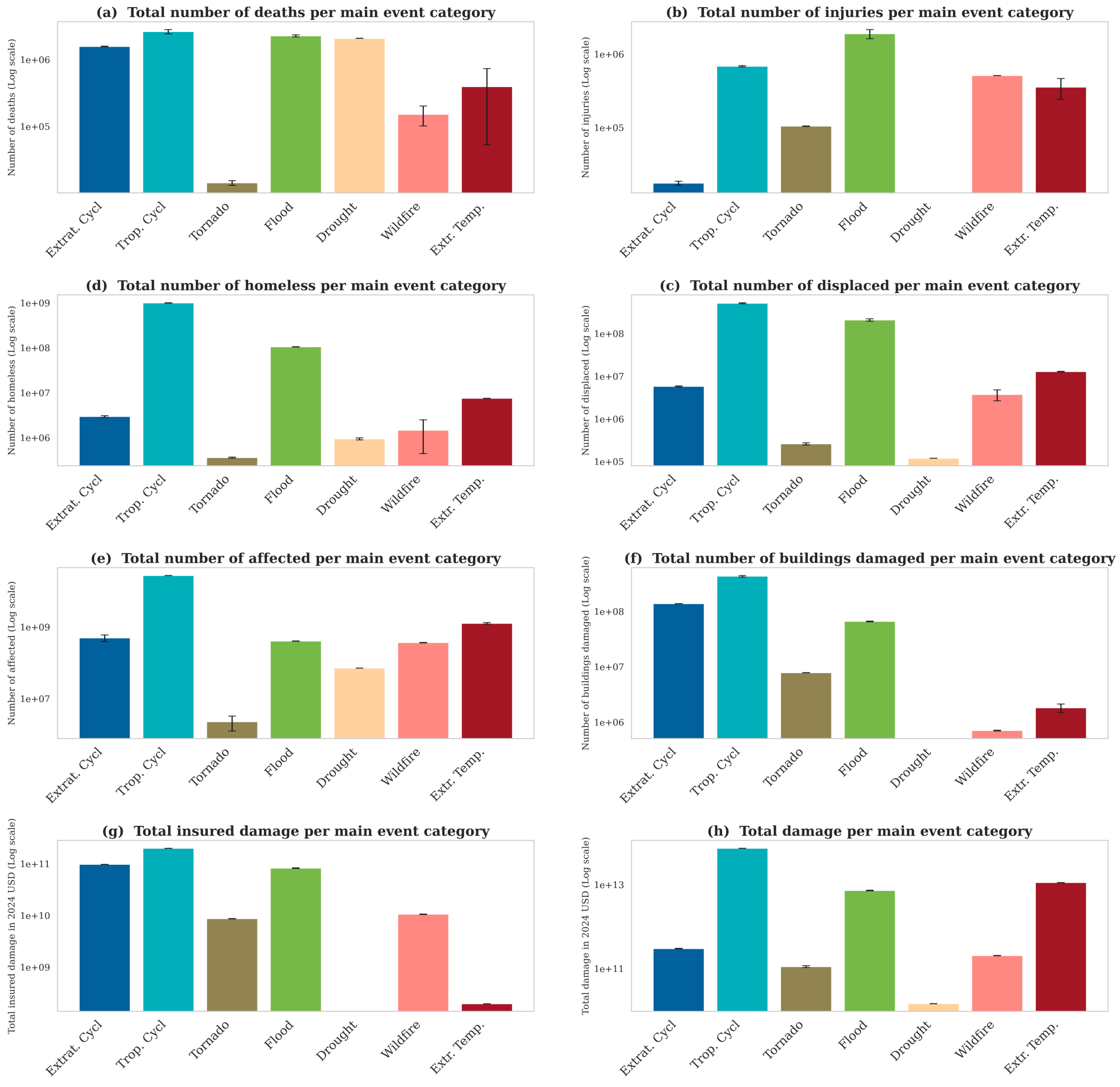

Figure 6. Total impact of each main event category in the Wikimpacts 1.0 database. Error bars represent the Max-Min range from L1 and the bar heights reflect the middle of the intervals from L1. (a) Deaths, (b) Injuries, (c) Homeless, (d) Displaced, (e) Affected, (f) Buildings Damaged, (g) Insured Damage, and (h) Total Damage. Note the logarithmic y-axis scale.

Central Asian nations. The US leads in tropical storm entries with 647 events, followed by Mexico (382 events), the Philippines (316 events), China (265 events) and Japan (264 events) (Figure 7b). While we report impacts from tropical storms in several mid-latitude countries, the number of such events is low. In most cases, they refer to tropical cyclones that underwent an extratropical transition. If the impacts of the cyclone following the extratropical transition are recorded in the Wikipedia article, we ascribe those impacts to the "Tropical Cyclone" main event category. Examples include “Hurricane Nate (2005)”[25] and “Hurricane Larry”[26] has even impacted Greenland. For extratropical storms (Figure 7c), the majority of entries are in the US (130 events) and Canada (77 events), followed by the UK (56 events), France (42 events), Germany (37 events), Ireland (32 events), the Netherlands (29 events), Belgium (21 events), Spain (19 events), Italy (17 events), and Switzerland (17 events). The database contains no entries for extratropical storms in Southeast Asia and some Central Asian countries. However, a few entries are present for tropical Africa. Further investigation reveals that the latter may be a misclassification by the model between the main event categories of flood and extratropical cyclone. For example, the event “2011 European floods”[27] – 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, yet classifying it as a flood would have been more appropriate. In terms of floods (Figure 7d), the US (83 events) and India (43 events) have the most entries, followed by China (20 events), Canada (18 events), the UK (15 events), Pakistan (15 events), Australia (13 events), and Germany and Afghanistan (11 events each). Many African and South American countries have only a single flood entry. Regarding tornadoes (Figure 7e), the US (237 events) has the highest number of entries, followed by Canada (24 events) and the UK (9 events). Most African countries lack recorded events for tornadoes. Extreme temperature events are primarily recorded in the US (26 events), Canada (17 events), and the UK (10 events). There are limited entries for extreme temperatures in African countries (Figure 7f), despite this continent being known for extreme heat episodes (Mora et al., 2017; Harrington and Otto, 2020; Thiery et al., 2021). For wildfires (Figure 7g), the US (123 events) and Australia (42 events) have the highest number of entries. Wildfire reports are sparse in African countries, Central Asia, Northern Europe, and South America, despite several of these regions being fire-prone (Burton et al., 2024). Lastly, droughts (Figure 7h), which have the fewest entries in the database (12 events) but often span several countries, are most frequently recorded in Russia and France (3 events each), followed by the US, Australia and some European countries like Italy and Luxembourg (2 events each).

5.2 L2 Impact Data (National Level)

We next investigate the spatial distribution of national-level (L2) impact data entries (Figure 8a). In total, we have 20,186 such entries, with the US having the highest number (4,475 entries in total). They are followed by Mexico with 1,377 data entries, the Philippines with 1,260, Japan with 1,058, China with 993, Australia with 592, Canada with 526, and India with 560. Most African countries only procured a limited number of impact data entries, with some exceptions such as Madagascar

[25]https://en.wikipedia.org/wiki/Hurricane_Nate_(2005)
[26]https://en.wikipedia.org/wiki/Hurricane_Larry
[27]https://en.wikipedia.org/wiki/2011_European_floods

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

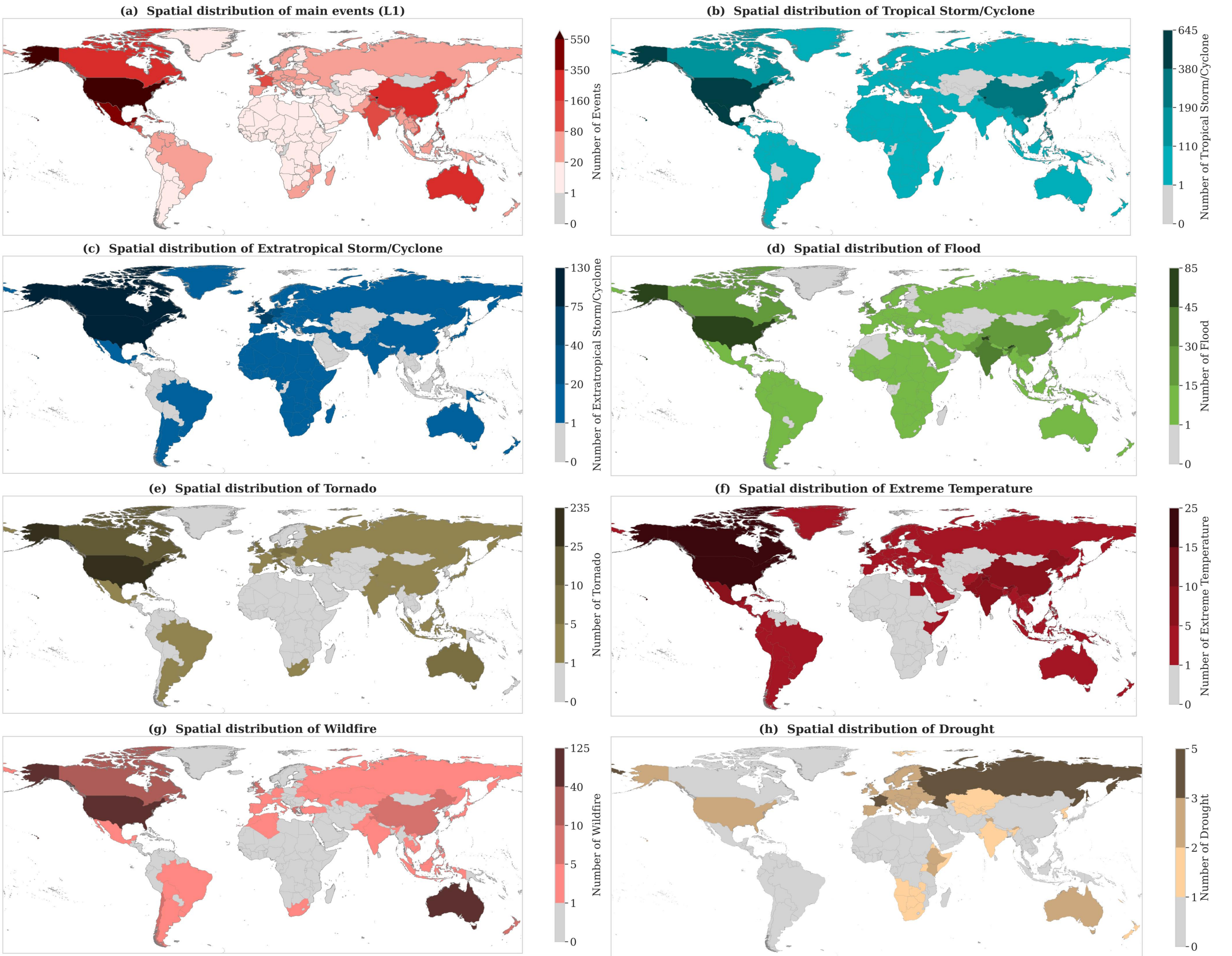

Figure 7. Spatial distribution of main events in the Wikimpacts 1.0 database, based on L1 entries: (a) overall spatial distribution of all main events, (b) spatial distribution of Tropical Storm/Cyclone events, (c) spatial distribution of Extratropical Storm/Cyclone events, (d) spatial distribution of 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.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

(259 entries) and Mozambique (130 entries). This largely reflects the spatial distribution of the main events (Section 5.1.3). Furthermore, a limited number of data entries is observed in parts of Western Latin America, Eastern Europe, and Central Asia.

5.3 L3 Impact Data (Sub-national Level)

The L3 information in our database reflects impact data reported at sub-national level, which we visualise in Figure 8b-c. In total, there are 36,394 entries in L3, and 9 entries contain unexpected GeoJSON shapes, such as ocean shapefiles of the Arabian Sea, which were subsequently removed for the visualization (see Appendix List **??**). The US leads with 11,894 sub-national level entries, followed by Mexico, Japan, the Philippines, China, Australia, India, and Canada, with 2,654, 2,425, 1,991, 1,511, 1,373, 1,165, and 810 entries, respectively. Vietnam and Cuba have fewer entries, with 581 and 415 entries, respectively. The GeoJSON files include both polygons and points, where polygons often represent larger administrative areas such as states or provinces (here referred to as “Regions"). Points typically represent cities, towns, or villages (here referred to as “Cities") for which OpenStreetMap could not find a GeoJSON object of a non-Point shape (such as Polygon or MultiPolygon). In some cases, when a region cannot be represented by a polygon or multi-polygon shape, it is recorded as a point location in our database. The “Regions” map (Figure 8b) indicates that not all states or provinces within a country have recorded impacts (Figure 8b). In some countries, such as China, impacts predominantly occur in coastal regions, with Guangdong province having the highest number of data entries at 191, followed by Fujian province with 154 entries. In contrast, in the US, impacts are distributed across the entire country, with Georgia having the highest number of entries at 118, followed by Delaware with 114 entries. In Mexico, a few states have a large number of data entries, like Acapulco (129 entries). In contrast, sub-national impact entries for African countries are limited, particularly in Central and Northern Africa. We observe a similar pattern in Western Latin America. In the “Cities” map, the distribution of impact data entries is more concentrated (Figure 8c). Most entries are of events that occurred in the US, with Outer Banks having the highest number of entries (63), followed by Grand Canyon with 21 entries and Cape Hatteras with 12 entries. Notably, Cabo San Lucas and San Jose del Cabo in Mexico also have a large number of data entries, with 46 and 21 entries, respectively. In Africa, the sub-national impact data entries are concentrated in coastal regions of South, East, and West Africa, whereas in South America, city-level impact data entries appear to be limited.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

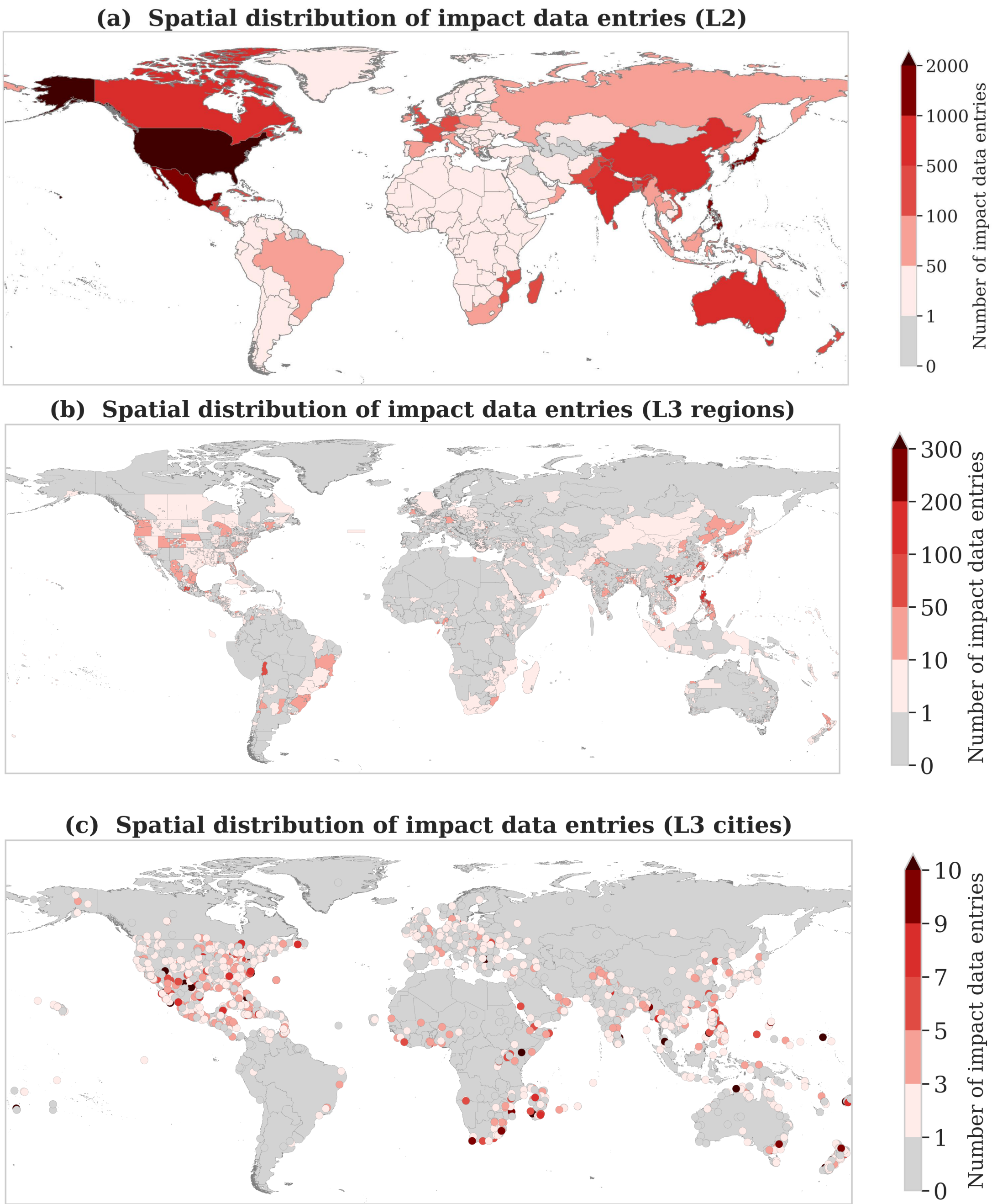

Figure 8. Spatial distribution of L2 and L3 impact data entries in Wikimpacts 1.0. (a) Spatial distribution of impact data entries at national level(L2), (b) Spatial distribution of impact data entries at regional level (L3 polygons, see text), (c) Spatial distribution of impact data entries at city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

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

5.4 Comparison With EM-DAT

We compare the coverage of Wikimpacts 1.0 to the widely-used EM-DAT impact database. As a first step, we align our database's time span with that of EM-DAT (01/01/1900 - 29/02/2024). To compare the most detailed available level in each dataset, We benchmark the number of impact data entries at L3 between our database and EM-DAT. To ensure a fair comparison, we assign one entry per available impact field in EM-DAT, and EM-DAT disaster subtypes are mapped to our main event categories (see SI Section 9). We specifically map the Tropical Cyclone category in EM-DAT to our Tropical Storm/Cyclone category, and the Extra-Tropical Storm category in EM-DAT to our Extratropical Storm/Cyclone category. Additionally, we aggregate all storm-related entries in EM-DAT and compare this with the combined total of Tropical Storm/Cyclone and Extratropical Storm/Clone categories in our database. For the spatial comparison, we utilize ISO codes from EM-DAT and the Administrative_Areas_GID identifiers from our database. Moreover, according to 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 impact value comparison. Events are precisely matched based on ISO code (Administrative_Areas_GIDs), main event type, and exact start/end year and month.

In total, the EM-DAT database contains 35,502 impact data entries, whereas the Wikimpacts 1.0 database comprises 33,904 data entries for the same period. Wikimpacts 1.0 database includes a greater number of data entries for main event types, such as tropical storms (15,002 more entries), extratropical storms (1,903 more), tornadoes (1,935 more), and wildfires (470 more) (Figure 9a). Notably, our database has 12,537 more entries for storms overall even when considering all storm-related entries in EM-DAT. However, our database contains substantially fewer data entries for floods (fewer by 14,581), as well as fewer data entries for droughts (fewer by 1,370) and extreme temperature events (fewer by 589). From a spatial distribution perspective (see Figure 9b), our database contains more impact data entries in the US (7,873 more entries), Mexico (1,988 more entries), Japan (1,545 more entries), Australia (518 more entries), and Canada (406 more entries), as well as in a few countries in Northern Europe and Africa. In contrast, Wikimpacts 1.0 contains fewer impact data entries in most countries in Africa, South America, and Asia. For example, there are fewer data entries in China (1,075 fewer), Indonesia (775 fewer), Bangladesh (600 fewer), Brazil (566 fewer), India (510 fewer), Vietnam (402 fewer), Thailand (363 fewer), South Africa (248 fewer), Kenya (200 fewer), Nigeria (178 fewer) and Tanzania (170 fewer). Despite these regional differences in coverage, both datasets overall suffer from a spatial reporting biased towards the Global North.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

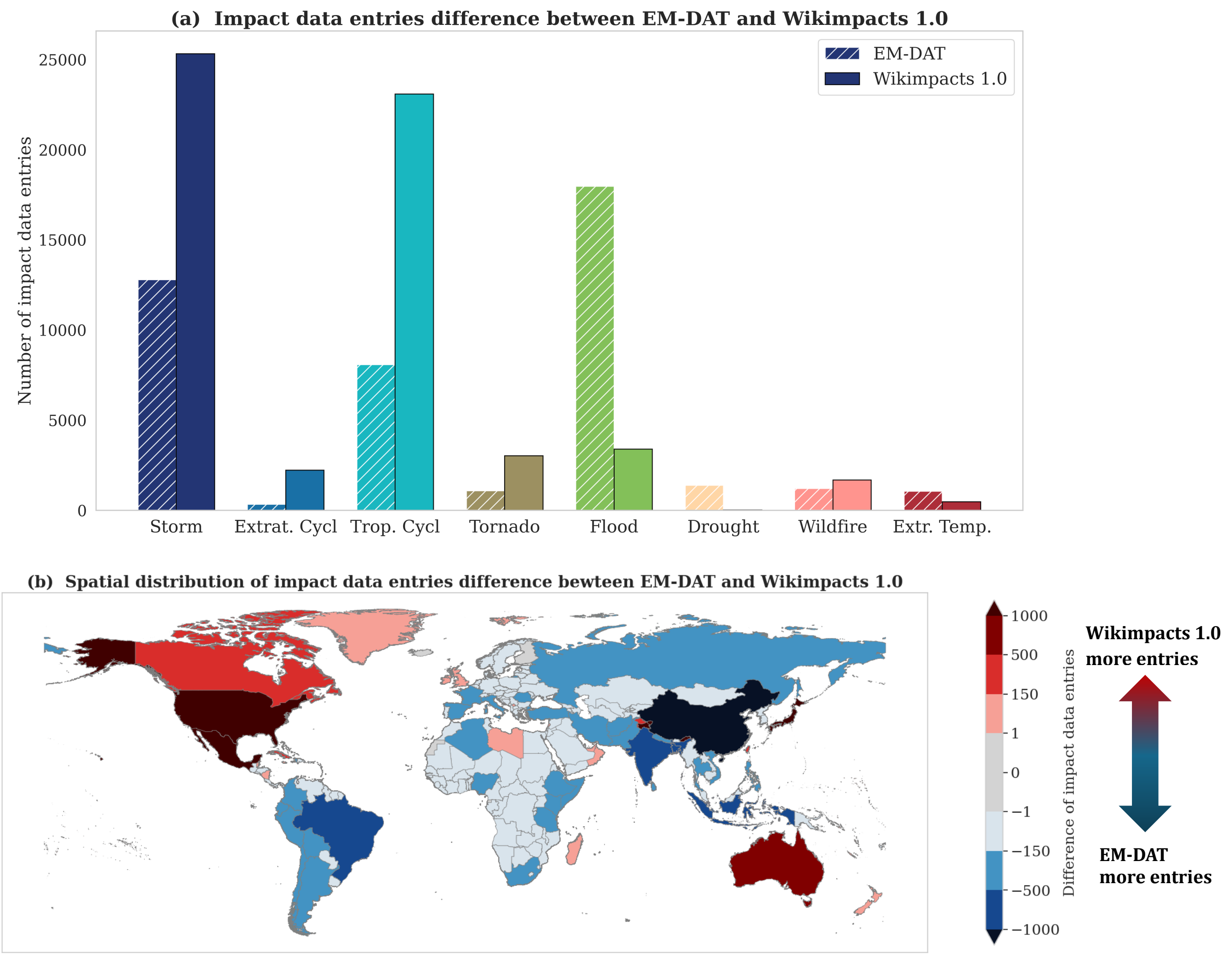

Figure 9. Impact data entry comparison between EM-DAT and Wikimpacts 1.0 from 01/01/1900 - 29/02/2024. (a) number of impact data entries in EM-DAT (in hatched colors) and Wikimpacts 1.0 (in full colors) for each main event category in Wikimpacts 1.0, (b) spatial distribution of the difference (Wikimpacts 1.0 minus EM-DAT) in number of impact data entries after aggregation across main event categories. Note the non-linear colour scale.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

We also perform an event-by-event matching between EM-DAT and Wikimpacts, and classify the events depending on whether the impact entries match, or by how much they differ. The comparison is illustrated in Figure 10. In the deaths category, 38 out of 234 matched events exhibit identical values with EM-DAT. However, 50 events show 50% higher values, with 47 events having values 50% lower than EM-DAT. In the injury category, 7 out of 94 events perfectly align with EM-DAT values; over one-third of the events exhibit values 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 75% in the damage category show higher values than EM-DAT.

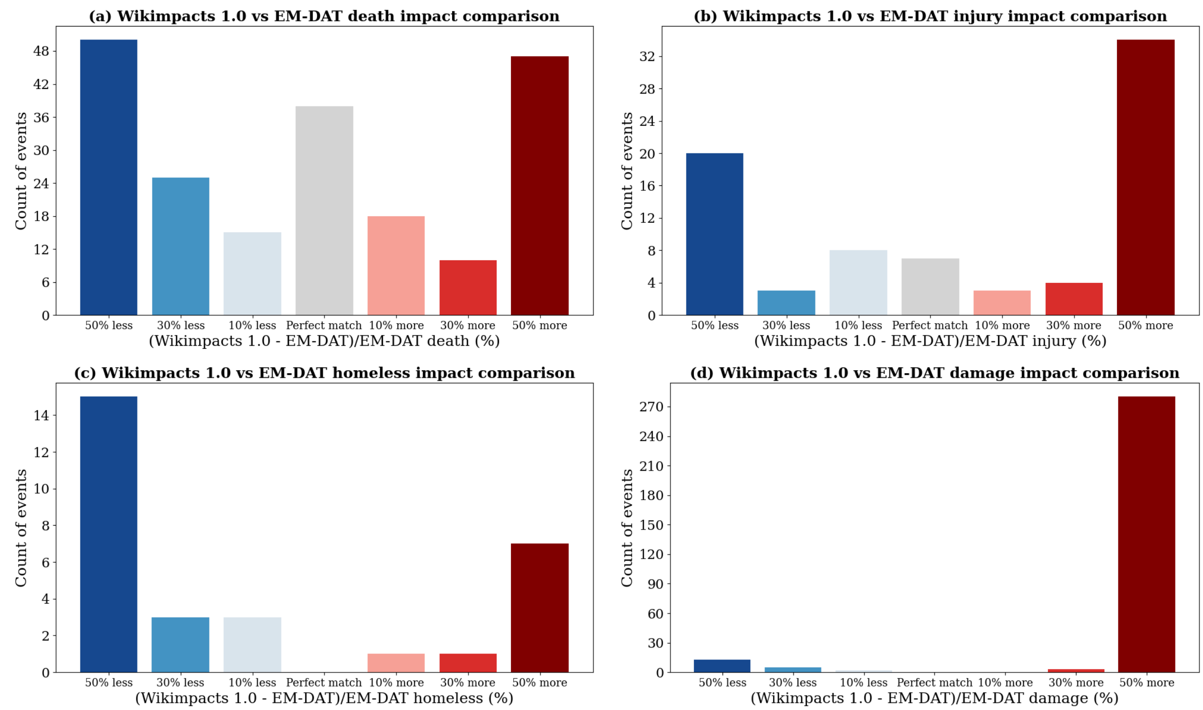

Figure 10. Impact value comparison between EM-DAT and Wikimpacts 1.0 from 01/01/1900 - 29/02/2024. (a) the percentage of difference between Wikimpacts 1.0 and EM-DAT in the death category, (b) injury category, (c) homeless category, (d) damage category.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

6 Discussion

6.1 Database Quality Assessment

The pipeline of the Wikimpacts 1.0 database is designed to capture information contained in Wikipedia articles as accurately as possible. In this respect, the performance of the GPT4o model is crucial for determining the database's quality and robustness. The GPT4o model exhibits strong performance across all three levels of information (see Section 4.3 and SI Section 8). L1 data is the most robust and reliable, displaying the lowest error rate among the three levels, with the errors increasing as the spatial scale of the reported impacts decreases. L3 data entries display the highest errors. The reasons for these differences are explored in detail in the following error analysis.

6.1.1 L1 (Event Level)

Fields like event timing, main event category, total deaths, and total damage exhibit very low error rates, closely aligning with the gold standard database (Table 6). Analysing the articles used for model input, we find that most articles in both the development and test sets contain an "Info_Box" that may often include information on the start and end date, the total number of deaths or injuries, or information on the total damage or insured damage. Specifically, 69 out of 70 articles in the development set and 151 out of 156 articles in the test set contain an Info_Box. An example of the Info_Box for the 2021 European Floods can be found in SI Section 6. In addition, the "Event_Name" is directly extracted from the article title and fed to the model. Consequently, for the aforementioned four categories, the model is able to extract information with ease.

For the Hazards field, the model occasionally captures undefined hazards, such as "Landslide", or mixes hazards from one Main Event type with another. For example, it captures "Flood|Lightning" in a flood event where only "Flood" is defined as the hazard. In the evaluation process, this is given an error rate of 0.5. Consequently, the hazard field exhibits a higher error rate than the Main Event field. For the Administrative_Areas_Norm field, the model performs worse in the test set than in the development set, with the error rate approximately doubling. Notably, in the test set, 35 instances of the "Administrative_Areas" output exhibit an invalid JSON structure that deviates from the prompt-designed output structure used in the development set. Consequently, these outputs cannot be normalized during the evaluation process and are assigned a NULL penalty score of 1. This accounts for the increased error rate for this field in the test set compared to the development set.

For the impact categories, a major source of error arises from the model incorrectly capturing information from other levels. For instance, in the event "Cyclone Vayu" [28], the article states that "Approximately 300,000 residents of coastal Gujarat were evacuated on 12 June in preparation for the system's arrival". The model captures "Approximately 300,000" as the L1 total displaced information, which according to our definition of levels represents sub-national impact data and should therefore be recorded in L3. For similar reasons, in the test set, the model incurs NULL penalty scores for 46 entries in the "displaced field", 33 entries in injuries, 39 in homeless, 47 in affected, 43 in buildings damaged, and 19 in total insured damage. This highlights

[28] https://en.wikipedia.org/wiki?curid=61000334

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Table 9. Comparison of the average number of entries at L2 and L3 levels for the Large Language Model (LLM) output and the gold standard data (Gold).

	L2		L3	
	LLM	Gold	LLM	Gold
Deaths	1.60	0.89	2.88	2.16
Injuries	1.47	0.22	1.96	0.48
Homeless	1.50	0.08	2.35	0.09
Displaced	1.62	0.19	2.28	0.58
Affected	1.96	0.14	4.21	0.12
Buildings Damaged	1.89	0.38	3.12	1.22
Insured Damage	1.42	0.03	2.09	0.03
Damage	1.56	0.51	2.85	0.78

that a large part of the L1 error rate for impact fields arises from impact information being correctly captured but assigned to the wrong spatial level.

6.1.2 L2 (National Level) and L3 (Sub-national Level) Information Extraction

The model performance for L2 and L3 is generally worse than those for L1 (Section 4.3). During the matching process, an empty entry is added to the LLM output if the model fails to capture information recorded in the gold standard data. Similarly, if the model captures information absent in the gold standard data, an empty entry is created in the gold standard data. Both cases result in the highest possible error rate – a NULL penalty score of "1".

To investigate this further, we analyse the number of entries in L2 and L3 for both the gold standard data and the LLM output (Table 9). We find that the LLM output contains more information on average than the gold standard data, which can be attributed to a variety of reasons. First, we observe a similar error type to that seen in L1 information extraction: the model sometimes assigns sub-national information to L2 instead of L3. A similar issue occurs in L3, where the model occasionally assigns country names to sub-national locations. Second, the model does not always adhere to the defined impact categories. For example, it may capture "about 600 houses without electricity" in the "Affected" field, which does not align with our definition that requires explicit mentions of keywords (see Table 5). We recognise that, in some cases, this may highlight shortcomings of our keyword list rather than erroneous information extraction by the model.

Finally, we examine the location-related fields, which have the highest error rates among all fields in each impact category. The model often outputs "NULL" in the "Num" impact field when location data is present, yet impact data is absent. This results in 2,188 "NULL" values present among 5,413 L2 and L3 output entries. For these entries, the model incurs a NULL penalty score for the location-related field, leading to a higher error rate.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

6.1.3 Quality Improvement with Consolidation

Overall, the model's capability to capture and classify impact information varies across levels and entry categories. We address this challenge by automatically filtering incorrect main event types and hazards and by completing missing information through aggregation from L3 to L2 and from L2 to L1 (Sect. 3.4 and SI Section 5). The consolidation does not address issues such as misclassification of impact categories. Nonetheless, thanks to the consolidation steps many of the issues detailed in Sections 6.1.1 and 6.1.2 are resolved in the final version of the database.

6.2 Comparison with Existing Impact Databases

We endeavoured to conduct a fair comparison between EM-DAT and Wikimpacts, even though the databases differ in structure and in the categorisation of main events. EM-DAT has some level of standardization, using ISO / UN regions, and refers to GAUL administrative units. The weakness of EM-DAT is that the impact is not disaggregated between identified sub-national units. Our database generally contains more detailed information than EM-DAT, such as standardised impact data at a sub-national level, despite having fewer impact data entries than EM-DAT in many countries. The two databases further display very different distributions of the impact information across the different event categories. The total number of impact data entries between the two databases is nonetheless comparable, as are their biases in geographical coverage. When assessing the impact values, matching records can be noted for death and injury information, although significant discrepancies are observed in the data for homelessness and damage.

We next return to the broader challenges that we outlined in the introduction related to the currently available impact data for climate-related hazards. We argue that Wikimpacts 1.0 presents clear advances in several of those respects. First, it addresses the issue of non-standardised geographical information by including event-level and national-level impact data, and standardised information for sub-national impact data. This also prevents the issue of the same large-scale main event, e.g. a heatwave, being included in the database as several distinct events if it affected different countries. Second, Wikimpacts 1.0 is readily expandable thanks to its highly automated pipeline. Third, data in Wikimpacts 1.0 is traceable, as our database includes a Sources field for each impact entry. We also assign a range to our quantitative data when exact information is not provided in the underlying data source, enabling uncertainty-aware impact analyses. Finally, the database is fully reproducible, since we openly share both the database itself and the source code of our processing pipeline.

6.3 Limitations

While Wikimpacts 1.0 innovates over existing databases in many aspects, it nonetheless comes with a number of caveats. First, the geographical coverage of the impact data remains uneven, likely at least in part due to the exclusive use of English-language Wikipedia articles. This limitation is particularly pronounced in the Global South, where there is generally less reporting of extreme events through Wikipedia and where English is not always widely used. Related to this, for the events that are reported, there are many missing data for the different impact categories. Second, the coverage across main event categories is uneven, with comparatively few data entries for some categories like extreme temperatures and droughts. This may be partly due to the

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

difficulty of assigning quantitative impacts to these events as these detrimental effects often occur on relatively long timescales and are often indirect. In addition, our pipeline, while scoring highly on the evaluation metrics, nonetheless introduces some errors relative to the original information provided by the Wikipedia articles we use. Furthermore, the database focuses on direct impacts, and overlooks indirect or cascading impacts unless these fit into one of the predefined impact categories. Finally, the database's reliability is inherently tied to the quality of Wikipedia articles, as we perform no additional verification of the sources beyond what Wikipedia does. Lastly, we only provide information on the hazard causing the impacts at L1 level, while the database lacks L2 and L3 hazard information. Further research could aim at addressing some of these limitations, for instance by expanding the database to multi-event Wikipedia articles, other Wikipedia languages or online textual sources beyond Wikipedia.

7 Conclusion

The resulting open access Wikimpacts 1.0 database encompasses 2,928 climate events spanning from the period 1034 to 2024, with global coverage. There is, however, a clear bias towards events in the Global North and occurring from the 1950s onwards. Wikimpacts 1.0 presents several innovations over the state of the art in multi-hazard global climate impact databases. For each extreme event, the database provides hierarchical information on the impacts, enabling multi-scale analyses. The data is provided at three different spatial levels: aggregated over the whole event (L1; 2,928 data entries), aggregated per affected country (L2; 20,186 data entries), and at the most highly spatially resolved information provided by the textual sources (L3; 36,394 data entries). The automated pipeline ensures that the database is readily updatable and expandable with the inclusion of additional textual sources. Finally, each impact information is linked to the original source, ensuring verifiability of the information provided.

8 Code Availability

Code is available at https://github.com/VUB-HYDR/Wikimpacts/tree/main DOI: 10.5281/zenodo.14726407 (Li et al., 2025b)

9 Supplementary Information

Please refer to the Supplementary Information file.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Author contributions. NL, WT, SZ, MMdB, SL, CF, JN and GM designed the analysis. NL conducted the LLM experiments, designed and plotted the figures and wrote the first draft of the manuscript under the supervision of WT. SZ wrote the database software pipeline with the support of NL and MK. KW and GM established the normalization rules for numeric information. KW supported part of their implementation in the software. JN and GM supervised and helped implementing the software and database structure. KW, CF, and CT coordinated and implemented the manual data annotation. PM designed the website with the support of WT and NL. JZ provided guidance and contributed to discussions. All authors provided guidance on the analysis and contributed to writing the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. The author group remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper.

Acknowledgements. We thank Davide Faranda, Aglaé Jezequel and Olof Görnerup for the valuable discussions about this project. We thank EM-DAT project team Niko Speybroeck, Valentin Wathelet and Regina Below for guidance on our database consolidation process. We thank Johan Tengholm and Jacob Nelsone for their work on annotating textual data. Compute and storage resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government, and by Research Institutes of Sweden AB. Ni Li is supported by the VUB Research Council in the framework of a EUTOPIA inter-university co-tutelle PhD between the Vrije Universiteit Brussel, Belgium, and TU Dresden, Germany. The EUTOPIA alliance is part of the European Universities Initiatives co-funded by the European Union. WT acknowledges funding from the European Research Council (ERC) under the European Union's Horizon Framework research and innovation programme (grant agreement No 101124572; ERC Consolidator Grant 'LACRIMA'). GM, JN and MK acknowledge funding from the Swedish Research Council Vetenskapsrådet (grant no. 2022-06599, climes). GM, JN and SZ acknowledge funding from the Swedish Research Council Vetenskapsrådet (grant no. 2022-03448). GM, CF and KW acknowledge funding from the European Research Council (ERC) under the European Union's Horizon Framework research and innovation programme (grant agreement no. 101112727). Lastly, we thank OpenAI Researcher Access Program for allocating 5000$ API credits for the GPT application and text rephrasing support.

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

References

Ahmadi Mazhin, S., Farrokhi, M., Noroozi, M., Roudini, J., Hosseini, S., Motlagh, M., Kolivand, P., and Khankeh, H. R.: Worldwide disaster loss and damage databases: A systematic review, Journal of Education and Health Promotion, 10, https://doi.org/10.4103/jehp.jehp_1525_20, 2022.

Alencar, P. H., Sodoge, J., Paton, E., and Madruga de Brito, M.: Flash droughts and their impacts–using newspaper articles to assess the perceived consequences of rapidly emerging droughts, Environmental Research Letters, 2024.

Antweiler, W.: PACIFIC Exchange Rate Service. University of British Columbia. Sauder School of Business, https://fx.sauder.ubc.ca.

Ara Begum, R., R. L. E. A. T. B. T. B. W. C. X. C. K. M. G. N. N. S. R. S. and Wester, P.: Point of Departure and Key Concepts, in: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by H.-O. Pörtner, D.C. Roberts, M. T. E. P. K. M. A. A. M. C. S. L. S. L. V. M. A. O. B. R., pp. 121–196, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009325844.003, 2022.

Burton, C., Lampe, S., Kelley, D. I., Thiery, W., Hantson, S., Christidis, N., Gudmundsson, L., Forrest, M., Burke, E., Chang, J., et al.: Global burned area increasingly explained by climate change, Nature Climate Change, pp. 1–7, 2024.

DateParser contributors: Dateparser – python parser for human readable dates, https://github.com/scrapinghub/dateparser/tree/master, 2024.

de Brito, M. M., Kuhlicke, C., and Marx, A.: Near-real-time drought impact assessment: a text mining approach on the 2018/19 drought in Germany, Environmental Research Letters, 15, 1040a9, 2020.

de Brito, M. M., Sodoge, J., Fekete, A., Hagenlocher, M., Koks, E., Kuhlicke, C., Messori, G., de Ruiter, M., Schweizer, P.-J., and Ward, P. J.: Uncovering the Dynamics of Multi-Sector Impacts of Hydrological Extremes: A Methods Overview, Earth's Future, 12, e2023EF003 906, 2024.

de Bruijn, J. A., de Moel, H., Jongman, B., de Ruiter, M. C., Wagemaker, J., and Aerts, J. C.: A global database of historic and real-time flood events based on social media, Scientific data, 6, 311, 2019.

Delforge, D., Wathelet, V., Below, R., Sofial, C. L., Tonneliere, M., van Loenhout, J., and Speybroeck, N.: EM-DAT: The Emergency Events Database, 10.21203/rs.3.rs-3807553/v1, 2023.

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.

d'Errico, M., Yiou, P., Nardini, C., Lunkeit, F., and Faranda, D.: A dynamical and thermodynamic mechanism to explain heavy snowfalls in current and future climate over Italy during cold spells, Sci. Rep., under review, hal-02367559, 2020.

Faiella, A., Tiberiu-Eugen, A., Luoni, S., Francisco, R. D., Ferrer, M. M., et al.: The risk data hub loss datasets-The risk data hub historical event catalogue, 2020.

Global Administrative Areas: GADM database of Global Administrative Areas, version 2.0. [online, URL: www.gadm.org, 2012.

Hammond, M. J., Chen, A. S., Djordjević, S., Butler, D., and Mark, O.: Urban flood impact assessment: A state-of-the-art review, Urban Water Journal, 12, 14–29, 2015.

Harrington, L. J. and Otto, F. E.: Reconciling theory with the reality of African heatwaves, https://doi.org/10.1038/s41558-020-0851-8, 2020.

Hurlbert, M., Krishnaswamy, J., Johnson, F. X., Rodríguez-Morales, J. E., and Zommers, Z.: Risk management and decision making in relation to sustainable development, 2019.

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.

Jones, R. L., Guha-Sapir, D., and Tubeuf, S.: Human and economic impacts of natural disasters: can we trust the global data?, Scientific Data, 9, ISSN 20524463, https://doi.org/10.1038/s41597-022-01667-x, 2022a.

Jones, R. L., Guha-Sapir, D., and Tubeuf, S.: Human and economic impacts of natural disasters: Can we trust the global data?, Scientific data, 9, 572, 2022b.

Kreibich, H., Schröter, K., Di Baldassarre, G., Van Loon, A. F., Mazzoleni, M., Abeshu, G. W., Agafonova, S., AghaKouchak, A., Aksoy, H., Alvarez-Garreton, C., Aznar, B., Balkhi, L., Barendrecht, M. H., Biancamaria, S., Bos-Burgering, L., Bradley, C., Budiyono, Y., Buytaert, W., Capewell, L., Carlson, H., Cavus, Y., Couasnon, A., Coxon, G., Daliakopoulos, I., de Ruiter, M. C., Delus, C., Erfurt, M., Esposito, G., François, D., Frappart, F., Freer, J., Frolova, N., Gain, A. K., Grillakis, M., Grima, J. O., Guzmán, D. A., Huning, L. S., Ionita, M., Kharlamov, M., Khoi, D. N., Kieboom, N., Kireeva, M., Koutroulis, A., Lavado-Casimiro, W., Li, H.-Y., LLasat, M. C., Macdonald, D., Mård, J., Mathew-Richards, H., McKenzie, A., Mejia, A., Mendiondo, E. M., Mens, M., Mobini, S., Mohor, G. S., Nagavciuc, V., Ngo-Duc, T., Nguyen, H. T. T., Nhi, P. T. T., Petrucci, O., Quan, N. H., Quintana-Seguí, P., Razavi, S., Ridolfi, E., Riegel, J., Sadik, M. S., Sairam, N., Savelli, E., Sazonov, A., Sharma, S., Sörensen, J., Souza, F. A. A., Stahl, K., Steinhausen, M., Stoelzle, M., Szalińska, W., Tang, Q., Tian, F., Tokarczyk, T., Tovar, C., Tran, T. V. T., van Huijgevoort, M. H. J., van Vliet, M. T. H., Vorogushyn, S., Wagener, T., Wang, Y., Wendt, D. E., Wickham, E., Yang, L., Zambrano-Bigiarini, M., and Ward, P. J.: Panta Rhei benchmark dataset: socio-hydrological data of paired events of floods and droughts, Earth System Science Data, 15, 2009–2023, https://doi.org/10.5194/essd-15-2009-2023, https://essd.copernicus.org/articles/15/2009/2023/, 2023.

Lange, S., Volkholz, J., Geiger, T., Zhao, F., Vega, I., Veldkamp, T., Reyer, C., Warszawski, L., Huber, V., Jägermeyr, J., Schewe, J., Bresch, D., Büchner, M., Chang, J., Ciais, P., Dury, M., Emanuel, K., Folberth, C., Gerten, D., and Frieler, K.: Projecting Exposure to Extreme Climate Impact Events Across Six Event Categories and Three Spatial Scales, Earth's Future, 11, e2020EF001 616, https://doi.org/10.1029/2020EF001616, 2020.

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 — 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, 2025a.

Li, N., Zahra, S., Worou, K., Kurfali, M., Görnerup, O., Nivre, J., and Messori, G.: Wikimpacts, GitHub repository, https://doi.org/10.5281/zenodo.14730195, 2025b.

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.

Mora, C., Dousset, B., Caldwell, I. R., Powell, F. E., Geronimo, R. C., Bielecki, C. R., Counsell, C. W., Dietrich, B. S., Johnston, E. T., Louis, L. V., et al.: Global risk of deadly heat, Nature climate change, 7, 501–506, 2017.

Muheki, D., Deijns, A. A., Bevacqua, E., Messori, G., Zscheischler, J., and Thiery, W.: The perfect storm? Co-occurring climate extremes in East Africa, Earth System Dynamics, 15, 429–466, 2024.

OpenStreetMap contributors: Planet dump retrieved from https://planet.osm.org , https://www.openstreetmap.org, 2017a.

OpenStreetMap contributors: TMC/Location Code List/Location Types, https://wiki.openstreetmap.org/wiki/TMC/Location_Code_List/Location_Types, 2017b.

Panwar, V. and Sen, S.: Disaster damage records of EM-DAT and DesInventar: a systematic comparison, Economics of disasters and climate change, 4, 295–317, 2020.

Papagiannaki, K., Petrucci, O., Diakakis, M., Kotroni, V., Aceto, L., Bianchi, C., Brázdil, R., Gelabert, M. G., Inbar, M., Kahraman, A., et al.: Developing a large-scale dataset of flood fatalities for territories in the Euro-Mediterranean region, FFEM-DB, Scientific data, 9, 166, 2022.

Paprotny, D., Terefenko, P., and Śledziowski, J.: An improved database of flood impacts in Europe, 1870–2020: HANZE v2.1, Earth System Science Data Discussions, 2023, 1–37, https://doi.org/10.5194/essd-2023-321, https://essd.copernicus.org/preprints/essd-2023-321/, 2023.

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.

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M., and Zhou, B.: Weather and Climate Extreme Events in a Changing Climate, p. 1513–1766, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.013, 2021.

Sodoge, J., Kuhlicke, C., and de Brito, M. M.: Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning, Weather and Climate Extremes, 41, 100 574, ISSN 2212-0947, https://doi.org/https://doi.org/10.1016/j.wace.2023.100574, https://www.sciencedirect.com/science/article/pii/S2212094723000270, 2023.

Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stagge, J. H., Tallaksen, L. M., Kampragou, E., et al.: Impacts of European drought events: insights from an international database of text-based reports, Natural Hazards and Earth System Sciences, 16, 801–819, 2016.

Thiery, W., Gudmundsson, L., Bedka, K., Semazzi, F. H., Lhermitte, S., Willems, P., van Lipzig, N. P., and Seneviratne, S. I.: Early warnings of hazardous thunderstorms over Lake Victoria, Environmental Research Letters, 12, 074 012, 2017.

Thiery, W., Lange, S., Rogelj, J., Schleussner, C.-F., Gudmundsson, L., Seneviratne, S. I., Andrijevic, M., Frieler, K., Emanuel, K., Geiger, T., et al.: Intergenerational inequities in exposure to climate extremes, Science, 374, 158–160, 2021.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., and Das, A.: A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, https://arxiv.org/abs/2401.01313, 2024.

Tschumi, E. and Zscheischler, J.: Countrywide climate features during recorded climate-related disasters, Climatic change, 158, 593–609, 2020.

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

https://doi.org/10.5194/egusphere-2025-4891
Preprint. Discussion started: 29 October 2025

Zommers, Z., Marbaix, P., Fischlin, A., Ibrahim, Z. Z., Grant, S., Magnan, A. K., Pörtner, H.-O., Howden, M., Calvin, K., Warner, K., et al.: Burning embers: towards more transparent and robust climate-change risk assessments, Nature Reviews Earth & Environment, 1, 516–529, 2020.

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