**Response letter: referee comment 1**

**Below, we address each comment in detail. Our answers to comments in bold.**
Comments by the Anonymous Referee are in normal style.

**We would like to thank the reviewer for the time and effort invested in evaluating the revised manuscript. We greatly appreciate the positive assessment of our responses and the recognition that most of the previous comments have been successfully addressed. The remaining suggestions and clarifications have been highly valuable and have helped us further improve the quality, transparency, and usability of both the HazMiner methodology and dataset. Below, we provide a detailed point-by-point response and describe the revisions implemented in the manuscript.**

# General Comments

I thank the authors for carefully replying to and successfully addressing most of the comments. After reading the rebuttal and the new version of the manuscript, several points are clearer. Although a few imprecisions and, in some cases, less than optimal performance in subtasks of the pipeline remain, they are also hard to prevent completely for large-scale data with current computational tools. The pipeline and resulting dataset ought to be a valuable contribution to the field as long as the readers and users are well aware of its limitations and do not treat it as a definite ground truth of the underlying hazard manifestations.

Below are a few adjustments that I still consider necessary for publication. Then, I list various minor details throughout the text that would profit from clarification or rephrasing at several points.

# Recommended Adjustments

- The introduction should finish with a prominent sentence pointing the reader specifically to the limitations section and recommending that users become familiar with potential influence on their results before using the dataset.

**We want to thank the reviewer for their suggestion. We have included it in L110-L114 as followed: “Before using the dataset, readers are strongly encouraged to consult the**

**evaluation and limitations of HazMiner, as its automated approach may result in false-positive events and other biases. Taking these limitations into account, we finally discuss the added value of the HazMiner dataset providing new insights in the geo-hydrological hazard distribution globally, and suggest avenues for future research, discussing the versatility of HazMiner and its integration with other hazard detection methods (Sect. 4).”**

- The authors successfully resolved the ambiguous use of the term LLM, but there is still a pending issue with NLP terminology. "Language models" is mainly used to refer to models that specifically *model language*; technically, models that take a context of words/tokens as input and output a probability distribution over *words/tokens* (either the word that comes next or a missing word in the middle). As far as I understand, none of the models in this paper are used for that task. The listed models do rely on text encoders but their output layer is performing classification, token labelling, translation etc. More suitable terms to refer to them, when applicable, are transformer-based models, text encoders, transformer-based classifiers, translation models or, more generally, just NLP models. Some (but not all) of such occurrences are pointed in the comments below.

**As suggested by the reviewer, we changed the terminology into NLP models.**

- In the spatial and temporal clustering (Section 2.4), paragraphs are ignored based on arbitrary thresholds. On the one hand, it helps remove noise, but on the other hand it also potentially loses information on events that got less media attention, exactly the ones that may not be in existing inventories.

**We thank the reviewer for this insightful comment. We agree that there is a delicate balance between including too much noise in the clustering algorithm and loosing events due to strict filtering. Nevertheless, we believe that improving recall of the filtering NLP models would result in detecting those lost events, rather than changing the clustering parameters. Furthermore, these clustering parameters set the definition of an event, i.e., restructuring clusters based on current paragraphs, and therefore their size and duration, rather than detecting more events.**

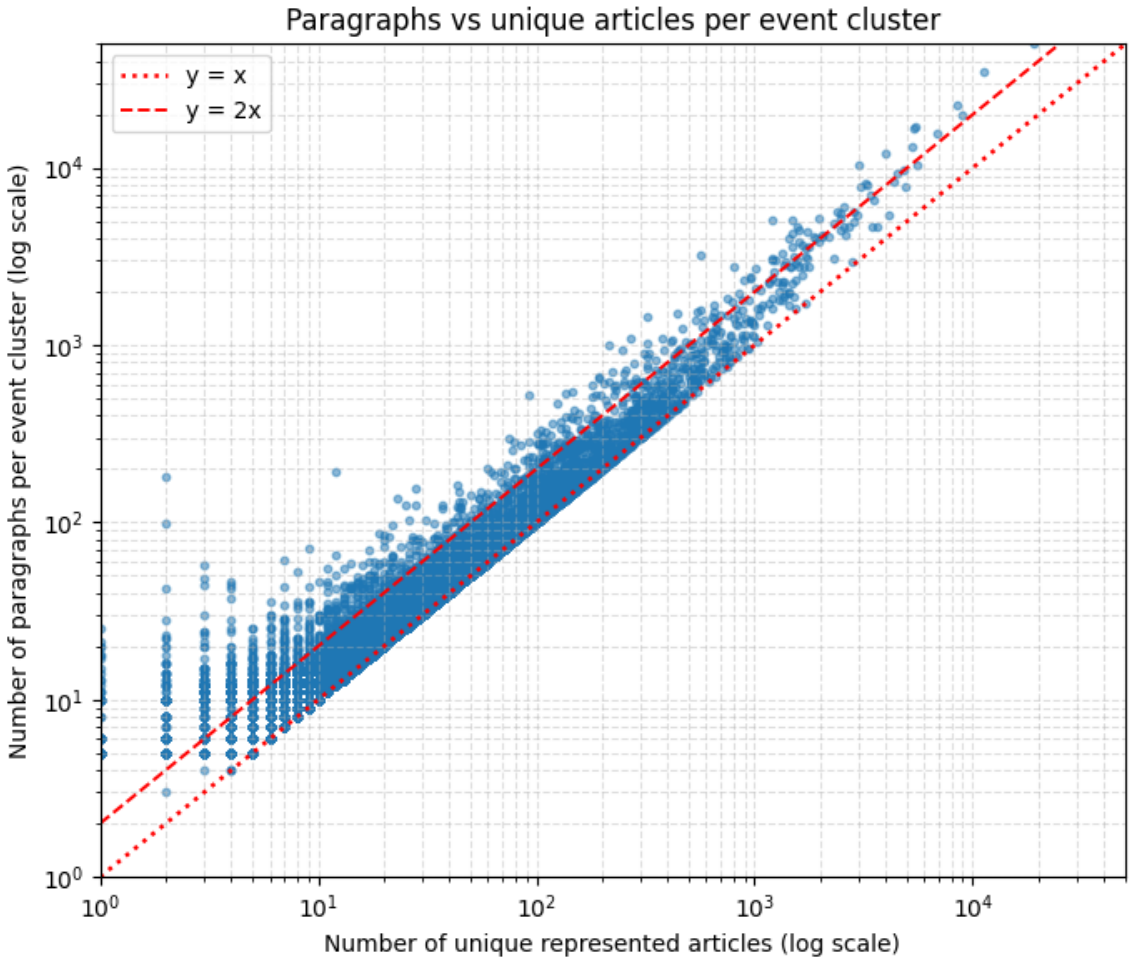
- Evaluation: the distribution of the labels in the annotated data samples used for the evaluation should be reported. Otherwise, it is not possible to correctly interpret the metrics' values. This could be part of Table B3, which should be part of the main text. The new classifier of "non-events" is an essential addition to the pipeline. However, it has not been evaluated in Section 2.5. Only an F1-score is reported without further information in Section 2.2. Please include its evaluation in Section 2.5.

**We agree with the reviewer that the evaluation was scattered over the manuscript. We moved Table B3 to the main text and updated it with the number of positive and negative annotated data used for the evaluation of the different tasks.**

**We furthermore moved the evaluation of the non-event filter from Section 2.2 to Section 3.2, centralizing all evaluation metrics together (L407-L409): "Non-events are filtered out by the "ValkB/all-MiniLM-L12-v2_HazMiner_nonevents" NLP model (Valkenborg et al., 2026) with a F1-score of 0.74, a recall of 0.78 and precision of 0.71. This step was evaluated against 89 positive and 79 negative annotated paragraphs (Table 3). The confusion matrix can be consulted in the appendix (Fig. A1)."**

- The unit of "media information" is still the article, even though the pipeline relies on paragraphs. A long article may have multiple paragraphs about the same event. Using the number of paragraphs per event cluster in the analysis can distort results in case many paragraphs are all coming from the same text. I suggest showing a scatter plot with the number of represented unique articles vs. the number of paragraphs per event cluster. If it not too uncommon for many paragraphs to belong to same article in a given event cluster, the number of unique articles should be used instead for the analyses and argumentations.

**As suggested by the reviewer, we plotted the number of articles per event over the number of paragraphs per events.**




**The plot suggests that most of the events contain either one paragraph or 2 paragraphs per articles. Furthermore, we calculated per event the ratio of the number of paragraphs over the number of events. These resulted in the following distribution:**

| Min | 5% quantile | 25% quantile | Median | 75% quantile | 90% quantile | Max | Mean |
|---|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 1.0 | 1.0 | 1.3 | 2.2 | 90.5 | 1.3 |

**These results suggest that only a small number of events contain multiple paragraphs from the same article, these cases are not representative for the overall distribution. Therefore, "media information" is unlikely to depend on whether paragraph counts or unique article counts are used.**

- Since each of the tasks in the pipeline are prone to errors, a final validation with an error analysis of the dataset would be very valuable. For instance, a sample of 100 events could be manually checked for their relevance (is it really an ongoing event of the hazard

category?), coherence (are all the paragraphs indeed talking about the same event), geolocation (does the location boundaries align with the paragraphs?), impact (does the final value really represent the numbers reported in the paragraphs?) and time (during the event really happen at the estimated period?).

**We want to thank the reviewer for this suggestion. We have evaluated 75 events as described in L307-L310: "Finally, we manually reviewed the paragraphs associated with 75 events to assess their relevance, coherence, and consistency with the documented event location, timing and impact. The reviewed events included, according to the HazMiner classification, 24 floods, 24 landslides and 27 flash floods. Their spatial distribution across continents per hazard type can be found in Table A5."**

**This resulted in the following results (L476-L484): "Of the 75 evaluated events, 70 (93%) corresponded to an actual geo-hydrological hazard event. The extracted hazard type was correctly identified in the paragraphs for 65 events (87%), while the paragraphs consistently referred to the same event for 59 events (79%). In terms of information extraction, the event location was correctly documented for 62 events (83%). Temporal information was present in at least one paragraph for 52 events (69%). Among these, the reported timing fell within the documented event duration (i.e., between the recorded start and end dates) for 37 events (71%). Moreover, for 28 of the 52 events (54%), at least one paragraph reported a hazard date that exactly matched the extracted event date. HazMiner events appear to under-detect impact information in the text. Although 42 events contained paragraphs describing at least one impact category documented by HazMiner, only 11 events had a reported impact. Detailed results for each impact category are provided in Table A7."**

# Minor Suggestions and Pending Clarifications

## General

- As far as I understand now after the rebuttal, only paragraphs that mention both a valid hazard AND a location can end up in the final sample. The text, e.g. at the beginning of the methodology at L118, could make it explicit. It would have avoided some confusion in my first reading.

**We further specified this by adapting L121-L122 into: "In this configuration, the output is a dataset on floods, landslides, and flash floods events, where each event is identified through a cluster of geocoded paragraphs.". Furthermore, we adapted Figure 1 so that it clearly shows that only paragraphs with a location are used in the event clustering algorithm.**

- Due to the selected keywords, the flood sample overlaps considerable with the flash flood sample, because the term "flood" also matches the same occurrences of the terms flash flood and sudden flood. Is that intended?

**From this comment we understand that the reviewer thinks that each hazard got his own dataset of URLs. Nevertheless, the request to the GDELT API was done through one structured query generating one dataset for all three hazards.**

**To clarify this in the text we removed table 1, which illustrated the query using three rows, generating the confusion to the reader.**

**Instead, we presented the query as a formula (L319-L324): "The keywords used to generate the HazMiner dataset are geo-hydrological terms related to floods, landslides, and flash floods, often referenced by English newspapers, they are presented to the API as one structured query (Eq. 1), creating one dataset for all three hazard types:**

**(flood OR inundation OR inundated OR "flash flood" OR deluge OR "sudden flood" OR landslide OR "debris flow" OR rockslide OR mudslide OR landslip OR "ground collapse" OR "soil sinking" OR subsidence) (1)**

- Table 2: The total counts should be reported in the same table to enable interpretation. It is also unclear what exactly the last 3 rows represent. For instance, Table B2 shows 185 HI floods in EM-DAT, whereas HazMiner has 8,298. How do these numbers relate to the reported 45? The header says "relative variation", but the computation is not clear.

**We added the total number of events to this table and removed table B2 from the appendix. In addition, we mentioned the formula for the relative variation in the methodology (L331-L335): "HazMiner is compared to the reference dataset in total number of events and in relative variation calculated as:**

$$\frac{n_{HazMiner}}{n_{ref}} \quad (2)$$

**where $n_{HazMiner}$ is the total number of HazMiner events and $n_{ref}$ is the total number of events in the reference dataset."**

## By Section

### Introduction

- L72-74: This sentence is not well-formed. The "while" should introduce a contrast; and either the "a" is spurious or a noun is missing after "reliable".

**We thank the reviewer for noticing this grammar mistake. We removed “while” and “a”**

- L79: A text corpora is already composed of its texts in some format. The "extraction" (news crawling) step is needed here because this work starts from a list of URLs instead, which is not always the case in text mining.

**We removed “and extracting its text”**

- L82: The task is typically called "Question Answering", not "Question and Answer".

**We adapted and changed “Question and Answer” into “Question Answering” everywhere**

- 88: Here, the term NLP seems more suitable than computational linguistics.

**We changed it to NLP**

- L88-90: The claim about the underuse of encoder-based models would require a citation or some evidence.

**We adapted our phrasing of the statement, focusing less on the underuse of encoder-based models. Instead, we focused on the advantages of using smaller NLP models for certain tasks (L89-L92): “However, these non-text generating NLP tasks, do not necessarily require text generating LLMs and can also be performed more efficiently by much smaller fine-tuned encoder models, such as BERT (Rogers et al., 2020). BERT, compared to LLMs, can run locally on commonly available computers without sharing copyrighted text data through Application Programming Interfaces (APIs).”**

### Methodology

- L119: The rebuttal explains why this step is named "text extraction". Still, it causes confusion due to the common usage of "information extraction". Besides, step 1 also includes splitting texts into paragraphs and translating them, which are not "extraction". Step 1 could thus be more generally called "text preparation" or "preprocessing".

**We thank the reviewer for further clarifying their previous comment, we understand it better now. We therefore adapted “text extraction” to “text preprocessing” everywhere in the document.**

- L120: just models, not language models (see above)

**As proposed in the first general comment, we changed the term “language models” to “NLP models”.**

- L122: Suggestion for clarity: "is designed to potentially document any type of natural hazards from text sources"

**We adapted the text based on the suggestion of the reviewer**

- Figure 1: in the output of step 3 and input of step 4, paragraphs are not only geocoded but also annotated with impact and time? Maybe use "annotated" instead, otherwise it seems that they are *only* geocoded at this point.

**We adapted the figure to the suggestion of the reviewer**

- L134: Please say explicitly where this collection of URLs is coming from. It is not clear it comes from GDELT. In other news databases, there might not even be URLs to crawl.

**We clarified in the text that users of HazMiner introduce their own URL’s and that we specifically extracted the URLs from GDELT, by adding L132-L134: “The text of the articles is downloaded from a user-introduced collection of URLs, using the newspaper3K package (Ou-Yang, 2014), specifically designed for web scraping and parsing news articles. Within our application of HazMiner we specifically used GDELT as a source for URLs (Sect. 2.6).”**

- L139: translation models instead of language models

**We adapted the text to the reviewer’s suggestion**

- L141: NLP models instead of encoder language models

**As proposed in the first general comment, we changed the term “language models” to “NLP models”.**

- L152: Is this performed with a multi-label classifier? Please specify. The use of the scores do not sound like it was a multiclass model. If it is indeed multi-label, what happens when a paragraph talks about both a flood and a landslide and both are above the thresholds: is it assigned only to the one with the highest score? Why not to both categories?

**We indeed used a multi-label classifier, we clarified this in the text by: “We apply a fine-tuned multi-label classification DeBERTa model (Laurer et al., 2024) (Table A1), which generates a score for each predefined category”.**

**HazMiner assigns one hazard per paragraph, i.e., the one with the highest score, exceeding the empirical thresholds, as was already stated in L156.**

**HazMiner currently supports only single-hazard types. Both categories could be included, creating a new hazard type namely a multi-hazard. Nevertheless, this would require large scale adaptations to the method.**

- L175: The previous paragraphs about the other two classifiers did not report any evaluation metrics. It is better to either report on all of them or move this information to the evaluation section.

**We wanted to make a distinction between self-tuned models and models selected from Hugging Face. We understand that this might have caused some confusion to the readers. Therefore, as suggested by the reviewer, we centralized all evaluation metrics of self-tuned and Hugging Face models to Section "3.2 Evaluation of HazMiner".**

- L175: Please be specific on which version of the F1 score was used. In this case, the binary F1 score version is more suitable. Is this the binary F1-score (i.e. considering relevant paragraphs as the positive class) or an average of both classes? Reporting precision and recall of the positive class would be even more informative.

**This sentence was moved to Section 3.2 (L406-409), where we reported the evaluation of all models, and added recall and precision.**

**As suggested by the reviewer we adapted the calculation of the F1-score, recall and precision for the non-event filter to that of the positive class instead of the weighted average. The new metrics are reported in Table 3 of the manuscript. We specified this in L274-L276: "The NLP models used in the paragraph filtering step were evaluated based on recall, precision, and F1-score of the positive classes".**

- L182: "for these articles" -> how are they identified among all instances?

**We rewrote the paragraph as follows (L179-L182):**

**"The dates mentioned in the paragraphs are extracted using the dateparser Python package (Shabir et al., 2025). These include actual dates, indicating a day, month and year, or date references, such as, 'yesterday' or 'last Wednesday'. Afterwards, these date references are used to calculate the actual date by correcting the article's publication date. For example, a reference date of 'yesterday' will result in a date, one day before the publication date."**

**By rewriting the paragraph, we clarified that paragraphs are only corrected for which the dateparser package extracted a date.**

- L201: What happens to paragraphs that do not mention any location? Are they excluded or information from other paragraphs are assigned to it? This information is relevant here but comes only much later on.

**We further clarified that only paragraphs containing at least one place name are geocoded (L200-L201): "Eventually, each paragraph, for which at least one place name is extracted, is assigned one location, which is defined by a latitude, longitude, and bounding box (WGS84)".**

**We agree that it should be mentioned in the methods to make the reader aware of this limitation. We therefore added this information to "Section 2.4 Event creation: clustering paragraphs into events", as this is the step where these paragraphs are excluded (L221-L222): "Only paragraphs for which a location was extracted (Sect. 2.3.2) are considered in the spatially based event clustering algorithm."**

- L236: Media attention should rather be counted in the number of *articles*. Here, it may be the case that many paragraphs about the same event come from a single long article (see comment above). If that is the case, this should rather be computed using the number of unique article identifiers of a cluster of paragraphs.

**Our answer to general comment 5 indicates that these findings are unlikely to depend on whether paragraph counts or unique article counts are used.**

- L237: Eisensee and Stromberg (2007) are cited as a source of such claim but they do not seem to separate floods from flash floods in their analysis.

**We agree with the reviewer. We replaced the citation with another one that studied flood and flash flood events in Malawi through news articles, documenting significantly more news articles on riverine floods. We further changed the sentence into (L235-L327): "Because floods and landslides seem to receive more media attention compared to flash floods (Bailon et al., 2025), we set a higher threshold for event creation, corresponding to at least 10 paragraphs per event for floods and landslides, and five paragraphs for flash flood events."**

- L239: "two separate events": the concept of event here depends on the temporal segmentation, which at this point has not been performed yet. Do you mean two separate regions?

**We agree with the reviewer that events depend on temporal segmentation. Therefore, we adapted our sentence to (L237-L239): "For all three extracted hazard types, the maximum epsilon equals 0.5°, as we perceive paragraphs further away than 0.5° to spatially belong to a different high-hazard activity zone (Gill and Malamud, 2014)."**

- L256: Why was a clustering algorithm applied for the temporal dimension? If periods are set with epsilon=2, simple segmentation would be enough and simpler.

**We want to thank the reviewer on pointing out this methodological issue. After reviewing the DBSCAN method we agree that our statement was incorrect. DBSCAN is a density-based approach, identifying temporally dense zones around a core-point characterised by the "minimum number of samples" variable within a certain distance. For example, the following distribution:**

| Time (days) | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Number of paragraphs | 1 | 1 | 2 | 1 | 0 | 1 | 1 |
| Number of paragraphs within 2 days | 4 | 5 | 5 | 5 | No data | 2 | 1 |
| Core point? | No | Yes | Yes | Yes | No data | No | No |
| Neighbour point? | Yes | No | No | No | No data | Yes | No |
| DBSCAN Cluster | Yes | Yes | Yes | Yes | Yes | Yes | No |
| 2-day gap segmentation cluster | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

**The above example the paragraph at day 7 is not included in the DBSCAN clustering as is not a core-point and neither a neighbour point (non-core point within 2 days from a core-point), therefore it will be excluded. While when applying a 2-day gap segmentation rule it would be considered as part of the cluster. Nevertheless, despite what was stated in our text, this event did not end with 2 consecutive days without paragraphs.**

**We therefore adapted our text as follows (L252-L256): "In this context, epsilon is set to two days, defining the temporal neighbourhood within which paragraphs can be considered part of the same event. However, event boundaries in DBSCAN are determined not only by this temporal threshold but also by the density of paragraphs within the resulting temporal region. Consequently, a later paragraph may not be assigned to the same event if it is insufficiently connected to a sufficiently dense temporal cluster even if the data gap is smaller than epsilon."**

- L276: What about the new "non-event" classifier?

**We added the non-event filter to the list.**

- L278: NLP models instead of language models.

**As proposed in the first general comment, we changed the term “language models” to “NLP models”.**

- L318: The names of the two translation models were mentioned in Section 2. They could be here instead, as they are specific to this assessment and not to the pipeline.

**We understand that including the model names here might be easier for the reader. Nevertheless, we did not do this change as the translation models are part of the HazMiner algorithm and not an implication of the application of GDELT to HazMiner. We did add a reference to the table in the appendix listing the different NLP models.**

### Results

- L370: Section 3.2 should preferably come before the actual resulting dataset starts to be discussed. First the pipeline is evaluated, then its outputs (events) are described and discussed.

**We want to thank the reviewer for the suggestion. The first section of the results illustrates that HazMiner reduces the data gaps on the documentation of geo-hydrological events, especially in low- and lower-middle-income countries. We wanted to start our results with addressing our main message, followed by a critical assessment of the underlying method, i.e., the evaluation of HazMiner.**

- L374: What was the distribution of the positive and negative label among these 750 instances? It is important to know how imbalanced the problem is when interpreting the metrics. One can infer it later from Figure 3, but an explicit reporting would be much appreciated.

**We adapted the sentence as follows (L387-L388): “As mentioned in Sect. 2.5, the classification of hazard-related text was evaluated against 750 manually classified paragraphs, of which 123 were positively annotated and 627 negatively (Table 3).”**

- L377: What was the distribution of the three labels among these 2,000 instances? One can infer it later from Figure 3, but an explicit reporting would be much appreciated.

**We adapted the sentence as follows (L392-L393): “The zero-shot hazard classification was evaluated against 2000 manually classified paragraphs, of which 477 were annotated as floods, 69 as landslides, 60 as flash floods and 1394 as “other hazard” (Table 3).”**

- L384: "correctly detects" -> not always, as it does not achieve recall = 1.

**We agree with the reviewer and adapted our sentence to (L399-L400): "This indicates that the model correctly identifies a large proportion of the positively annotated flood- or landslide-related paragraphs."**

- L386: What exactly was the precision? Table B6 belongs to the main text.

**We added the precisions in the text, resulting into (L400-L401): "But its low precision results in a considerable proportion of paragraphs predicted as floods (precision of 0.65) or landslides (precision of 0.64), that are in fact not related to these hazards (Table 3)." Furthermore, we moved Table B6 to the main text, which became Table 3.**

- Table 4: Why are the min and max so extreme? If an EM-DAT event and a news event are months apart, they probably do not refer to the same fact. The alignment procedure should handle that.

**The extreme minimum or maximum can be explained by**

1. **Events that are reported in HazMiner but not in either EM-DAT or GFM, or**
2. **A time extraction error in HazMiner. As HazMiner relies on the publication date when no time reference is mentioned in the text**

**We did not expect a perfect alignment between HazMiner and the evaluation datasets. Nevertheless, the 25%- and 75%-quantile illustrate that a large fraction of the events and paragraphs are within an acceptable time error.**

- L443: Maybe the underestimations are due to using the median? Maybe a larger quantile or the mode would work better?

**We thank the reviewer for this suggestion. We agree with the reviewer that using a larger quantile might reduce the difference in reported numbers between EM-DAT and HazMiner.**

**Nevertheless, we would like to point out the large variation within the reported HazMiner and EM-DAT numbers for affected, compared to fatalities, which shows smaller variations centralized around the 1:1 line. This difference is probably due to the variation in the definition of affected people, which is not standardised, especially in media sources. We therefore argue that changing the aggregation strategy will barely have an effect, when in fact the underlying data might not be reliable to extract the number of affected people.**

- L474: Here is another case where reporting the number of unique articles within the cluster of paragraphs is more reliable than the number of paragraphs, as more than one paragraph might have been extracted from the same text.

**Our answer to general comment 5 indicates that these findings are unlikely to depend on whether paragraph counts or unique article counts are used.**

### Discussion

- L598: "its multilingual approach significantly improves..." The paper has not provided evidence to generalise this, it only showed a few specific examples.

**We adapted this into (L636-L637): "First, our results suggest that a multilingual approach can enhance geo-hydrological hazard documentation and help identify hazard information that may be underrepresented in global datasets."**

- L663: "languages lack equivalents" -> it is still often possible to express the same concepts, despite the lack of a single term.

**We adapted this into (L701-L702): "Translation additionally complicates terminology, as certain language-specific hazard terminology has no direct translation into English."**

- L705: NLP models

**As proposed in the first general comment, we changed the term "language models" to "NLP models".**

### Appendix

- L805: Please also report the distribution of articles by language in the initial sample and of paragraphs by language in the final dataset.

**To the appendix we added a new table containing a list of the languages together with the total number of articles extracted through GDELT, and of these the number of articles and paragraphs that are associated with an event.**

- Table B1: Please add hyperlinks to the models' pages.

**We thank the reviewer for their suggestion. We have included the hyperlinks of the model.**

- Table B5: A paragraph that does not mention e.g. any injured people is different from a paragraph that states clearly that no one was injured. Were these cases treated separately in the classification and in the subsequent impact estimations?

**Yes, paragraphs that did not return any answer were recorded as “NaN”, while paragraphs indicating that no one was injured were corrected to zero. The same was done for the other impact categories.**

**Response letter: referee comment 2**

**Below, we address each comment in detail. Our answers to comments in bold.**
Comments by the Anonymous Referee are in normal style

I only have a minor comment concerning the source of the news data. The manuscript states that the global dataset was generated using URLs obtained from the GDELT Global Knowledge Graph through the GDELT API. It would be useful to provide a few additional details on the construction of this initial news corpus, for example the specific GDELT/API query and selection criteria, the version or retrieval date of the GDELT dataset, and how issues such as duplicate articles or inaccessible URLs were handled. Providing some information on the characteristics and coverage of the underlying news sources would also improve the reproducibility of the study and help readers better assess the transferability of the approach to other applications.

**We want to thank the reviewer for reading our manuscript and giving suggestions to further improve the quality of our research. We carefully assessed their suggestions.**

**As requested by the reviewer, we have now included the query used to consult the GDELT API as Eq. 1 of the manuscript in L322-L323. To avoid repeating information, we removed Table 1, which contained the keywords used for the extraction.**

**We furthermore added the retrieval date to the text (L324-L325): "The data was retrieved from the API on 1 July 2024 for the articles published between 1 January 2017 and 31 December 2023, while those published between 01 January 2023 and 01 January 2025 were retrieved on 04 April 2025."**

**Some URLs were indeed inaccessible. During the development of HazMiner, we did a small study on these URLs. These articles were inaccessible due to paywalls, obsolete URLs, web scraping limitations or GDPR restrictions. We tried to retrieve the text of news articles using proxies. We attempted to re-download 10,000 failed downloads using proxies, achieving a success rate of 7%. However, this marginal increase in successfully retrieved articles would come at a significant cost in proxy usage.**

**To provide more information on the sources, we have included Table A6 in the appendix showing for each language the number of articles retrieved from GDELT, and of those the number of articles and paragraphs associated with an event. We additionally added a map (Figure A3) on the number of articles published from a certain country for all articles retrieved from GDELT and those associated with an event. These should inform the reader on the spatial distribution and bias of the extracted sources.**