Thank you very much for your valuable suggestions. We have answered your questions one by one in our PDF attachment. However, the current situation of the journal does not allow us to modify the manuscript during the discussion or submit it as supplementary material. Therefore, we will directly mention or show our changes, such as modified pictures, text to be added, comparisons with the old text, etc.. For details, please refer to our PDF attachment.

Your question is marked in **purple color.**

Question 1

The manuscript contains several inconsistencies. The first concerns the number of publications. The authors state that they retrieved 150,000 publications from WoS and subsequently used an LLM to select 57,474 articles closely related to flood research. However, the wording in Section 2.2 suggests that the 57,474 records were obtained "after deduplication." This would imply that no records were excluded by the LLM-based screening process. This discrepancy requires clarification.

And, the study resembles an LLM-assisted bibliometric analysis more than a conventional systematic literature review. Several elements typically expected in a systematic review are missing, including a detailed description of the publication screening and selection process, exclusion criteria applied at successive stages, and an assessment of the quality of the included studies.

Thank you for your suggestion. Our manuscript lacks details of the literature screening process and inclusion procedures. Another expert in Referee #1 also raised a similar issue. In short, this process is a multi-layered collaborative strategy consisting of blacklists and whitelists, and includes expert review mechanism.

We will modify the text around paragraph **135 on page 6** as follows to fix the ambiguity of the text.

Old sentences (green color):

2.2 Metadata collection

The metadata collection was completed on 14 May 2025 using the Web of Science database, and the publication time span of the flood-related literature is from January 1985 to December 2024, totaling 40 years. To comprehensively cover the diverse topics and disciplinary perspectives of flood research, 29 search terms (e.g. "Flood" and "Flooding", details in Appendix A2) were defined and queried independently. To ensure that the included publications possess high academic originality and research value, the document type was restricted to "Article," thereby systematically excluding review articles, conference papers, book reviews, and other literature with limited originality. After deduplication, 57,474 article records were obtained, and core metadata including titles, abstracts, keywords, journal names, publication years, and citation counts were extracted. The complete text of the "Author Information" section in each publication was parsed using LLM to directly extract the first corresponding institutions, all author affiliations, and their associated countries.

2.3 In-depth information extraction

The in-depth information extracted by the LLM primarily includes the study areas, flood types, research topics, and research methods. After extracting the names of study areas using the LLM, they were converted into geographic coordinates (latitude and longitude) via the geocoding API of Google Maps. Flood types were categorized into urban flood, river flood, mountain flood, coastal storm surge, ice flood, dam break flood, snowmelt flood, tsunami, and

groundwater flood. This was a multi-label classification task, with any category not included in the predefined list classified as “other”. The same rule applied to research topics, which include management, simulation, monitoring, and risk assessment.

Latest sentence (mainly change showed by **blue color**):

2.2 Metadata collection

The metadata collection was completed on 14 May 2025 using the Web of Science database, and the publication time span of the flood-related literature is from January 1985 to December 2024, totaling 40 years. To comprehensively cover the diverse topics and disciplinary perspectives of flood research, 29 search terms (e.g. “Flood” and “Flooding”, details in Appendix A2) were defined and queried independently. To ensure that the included publications possess high academic originality and research value, the document type was restricted to “Article,” thereby systematically excluding review articles, conference papers, book reviews, and other literature with limited originality. After indexing, we initially obtained about 150,000 publications. By using prompts used in this study are provided in Appendix A1, core metadata including titles, abstracts, keywords, journal names, publication years, and citation counts were extracted. The complete text of the “Author Information” section in each publication was parsed using LLM to directly extract the first corresponding institutions, all author affiliations, and their associated countries.

Next, we built a screening pipeline based on LLM, which includes a multi-level collaborative pipeline consisting of the blacklists, whitelists, and expert review mechanism (details in **Fig. P1**):

In Step 1, we retrieved publication data from the Web of Science using flood-related keywords (detailed in Appendix A2), restricted the document type to Article in Web of Science retrieval settings, and finally acquired more than 150,000 publications as the initial whitelists.

In Step 2, we adopted Prompt 1 (detailed in Appendix A1) to guide the large language model (LLM) for primary screening. The LLM judged each publication one by one: authentic research articles closely associated with flood research were classified into the whitelist, while irrelevant literature was incorporated into the blacklist.

In Step 3, carried out iterative alternating filtering between the whitelist and blacklist to continuously optimize classification accuracy, including two sub-steps:

Sub-step 1: Re-screen the existing whitelist with LLM driven by Prompt 1. Misclassified non-flood articles found within the whitelist would be moved to the temporary blacklist, which would then be merged into the global blacklist to update and expand the blacklist library; correctly retained flood-related articles stayed in the whitelist.

Sub-step 2: Deploy the same Prompt 1 to recheck publications inside the blacklist. If flood-related valid publications were mistakenly putted in the blacklist, such publications would be extracted to whitelist and flood-irrelevant publications remained in the blacklist.

We iterated Sub-step 1 and Sub-step 2 alternately for 10 full cycles starting from Sub-step 1. After cyclic screening, only publications appearing in the whitelist over 50% of the iterations were reserved as candidate valid flood-related

articles, greatly lowering the probability of false inclusion and false exclusion caused by single LLM judgment. It is worth stating that each LLM discrimination is a new restart and does not fix the random seed and its controllable parameters.

In Step 4, we introduced manual expert verification as the final reliability guarantee. We randomly sampled 4,000 papers mixed from the whitelist and blacklist, and invited four doctors specialized in hydrology and flood disaster research to conduct manual review. If manual inspection discovered misclassified documents, we would go back to Step 3 to adjust the LLM prompt rules and repeat the alternating filtering iteration again. The cycle would not terminate until no obvious misclassification existed in the test set, and the final stable whitelist was confirmed as the reliable literature library for subsequent information extraction.

Based on the two-layer verification system of "large model cycle screening + expert manual calibration", we finally obtained a whitelist which contains 57,474 publications.

2.3 In-depth information extraction

The in-depth information extracted by the LLM primarily includes the study areas, flood types, research topics, and research methods. After extracting the names of study areas using the LLM, they were converted into geographic coordinates (latitude and longitude) via the geocoding API of Google Maps. Flood types were categorized into urban flood, river flood, mountain flood, coastal storm surge, ice flood, dam break flood, snowmelt flood, tsunami, and groundwater flood. This was a multi-label classification task, with any category not included in the predefined list classified as "other". The same rule applied to research topics, which include management, simulation, monitoring, and risk assessment. The screening of this process also went through a similar pipeline (details in Appendix J1).

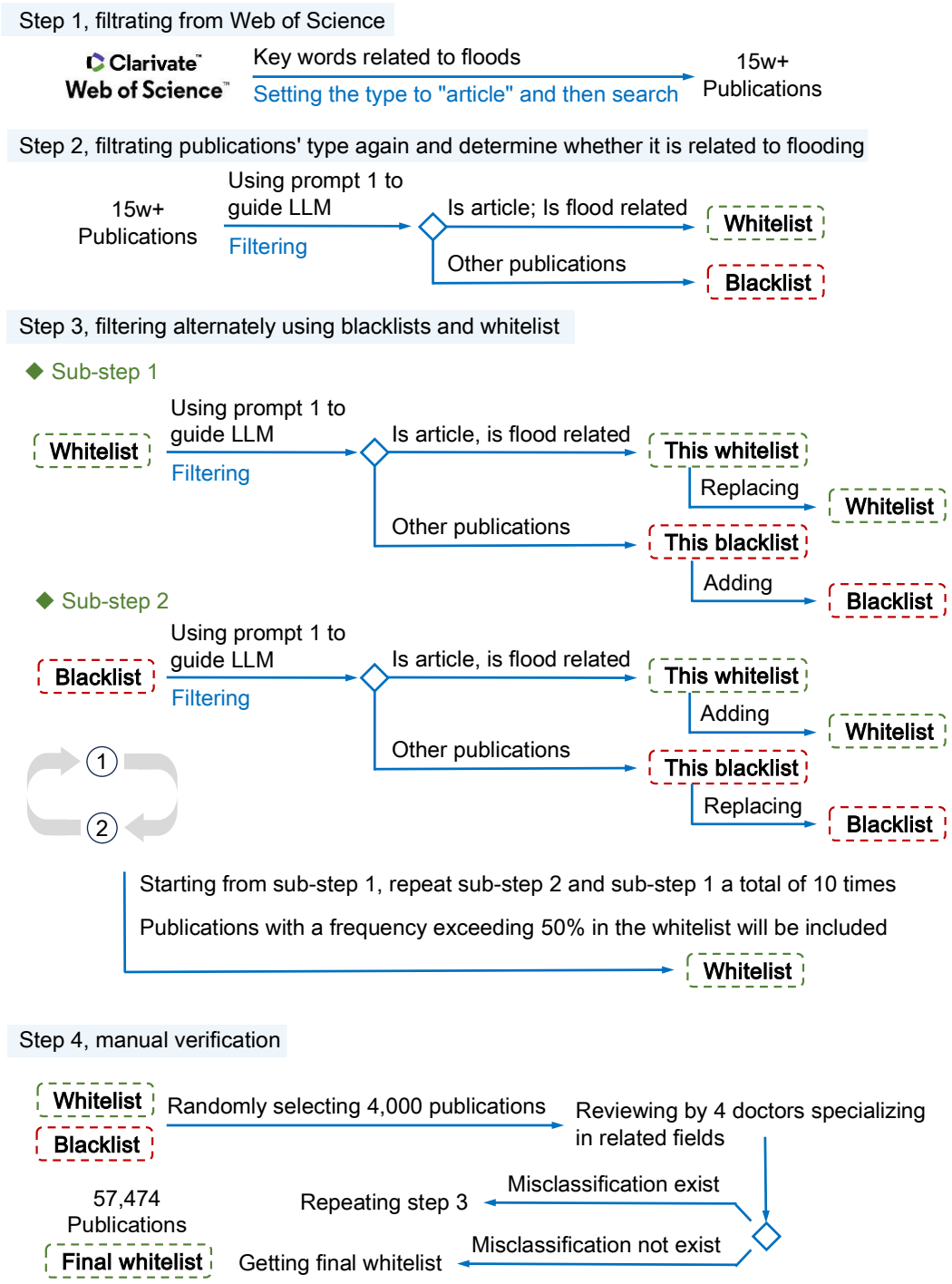

Fig. P1 Pipeline for screening research articles related to floods.

Appendix J1

It is similar to **Fig. P1**, this pipeline consisting of iterative black and white list screening and manual verification was also migrated and applied to the subsequent structured information extraction of fields such as research location, flood type, research topic, research method, affiliation, and country. Specifically, as shown in **Fig. P2**, The initial whitelist consisted of these 57,474 publications. The prompt words are replaced from Prompt 1 to Prompt 2−7 in Appendix A1, and the black and white lists are redefined based on whether valid target information is successfully extracted. Only after the cycle iteration process ends, the publication can successfully obtain feature information more

than 50% of the time, and the obtained information remains completely consistent in at least 5 rounds of extraction results, will it be determined as the valid feature information corresponding to the publication.

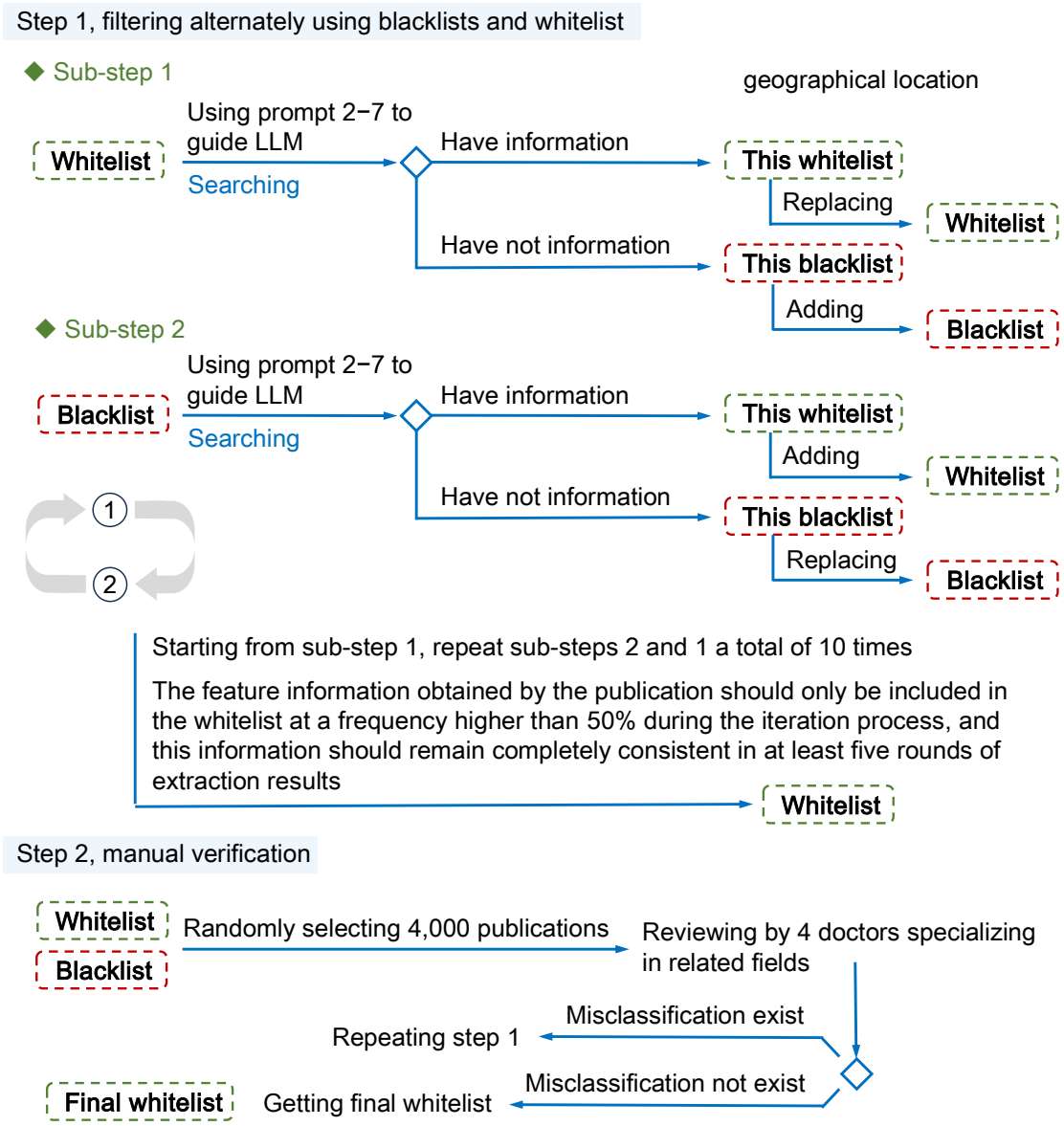

Fig. P2 Pipeline for extracting multi-dimensional characteristic information from publications, with all target features listed in Appendix A1. Prompts 2–7 in Appendix A1 are applied to extract structured information including study locations, flood types, research topics, methodologies, affiliated institutions and corresponding countries. Only features appearing in over 50% of iterations and holding identical outputs for at least five extraction rounds are retained as valid sample characteristics.

Question 2

Another concern relates to the concept of an "objective knowledge structure." The adopted structure is, at least to some extent, imposed by the authors. For example, the LLM is required to assign publications to predefined categories: management, simulation, monitoring, and risk assessment. Similarly, nine flood types are defined in advance. Therefore, the resulting knowledge structure cannot be regarded as entirely objective, as it is partly determined by the classification framework established by the authors.

Thank you for your great insights. We agree that a completely objective classification framework is difficult to achieve. The subject and flood type classification system of this study were indeed predefined by the research team. This classification system comprehensively refers to the classification standards provided by other researchers (Tarasova et al., 2019; Sikorska et al., 2015; Field, 2012), covering the major and frequent typical scenarios of global flood disaster research.

Tarasova, L., Merz, R., Kiss, A., et al, 2019. Causative classification of river flood events. WIREs Water, 6(4), e1353. https://doi.org/10.1002/wat2.1353

Sikorska, A. E., Viviroli, D., and Seibert, J. et al., 2015. Flood-type classification in mountainous catchments using decision trees. Water Resources Research, 51(10), 7959−7976. https://doi.org/10.1002/2015WR017326Digital Object Identifier

Field, C.B., Barros, V., Stocker, T.F., Dahe, Q., 2012. Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the intergovernmental panel on climate change, 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9781139177245

It is undeniable that any classification inevitably involves some degree of subjective judgment. To reduce the constraints caused by such subjectivity, we emphasize in this article that the task should be set as "multi-label classification". Additionally, a "other" category is designated specifically to accommodate those flood types and research topics that cannot be classified under the predefined categories.

We will explain our method, purpose, and limitations around paragraph 135 on page 6.

Old sentences (green color):

2.3 In-depth information extraction

The in-depth information extracted by the LLM primarily includes the study areas, flood types, research topics, and research methods. After extracting the names of study areas using the LLM, they were converted into geographic coordinates (latitude and longitude) via the geocoding API of Google Maps. Flood types were categorized into [urban flood, river flood, mountain flood, coastal storm surge, ice flood, dam break flood, snowmelt flood, tsunami, and

groundwater flood]. This was a multi-label classification task, with any category not included in the predefined list classified as “other”. The same rule applied to research topics, which include [management, simulation, monitoring, and risk assessment].

Latest sentence (mainly change showed by **blue color**):

The in-depth information extracted by the LLM primarily includes the study areas, flood types, research topics, and research methods. After extracting the names of study areas using the LLM, they were converted into geographic coordinates (latitude and longitude) via the geocoding API of Google Maps. Flood types were categorized into [urban flood, river flood, mountain flood, coastal storm surge, ice flood, dam break flood, snowmelt flood, tsunami, and groundwater flood]. The same rule applied to research topics, which include [management, simulation, monitoring, and risk assessment].

These tasks adopt a multi-label classification setting, allowing a single document to be simultaneously classified into multiple coexisting flood types or research topics, thereby reflecting the common reality that a single study can simultaneously involve multiple interrelated flood topics. We note that this taxonomy is predefined by the authors based on existing flood-related literature and cannot achieve absolute objectivity, the resulting knowledge structure is inherently constrained by our preset classification framework.

Question 3

The authors specify the Gemma 3-27B model and provide the prompts used; however, several details necessary for full reproducibility are missing. For example, the inference parameters of the model are not reported. Such information is important when using LLMs because, even with identical prompts, model outputs may not be completely deterministic.

Thank you for your valuable comment. We agree that inference parameters are critical for experimental reproducibility. But even with identical prompts and inference parameters, LLM outputs can still exhibit stochastic fluctuations. Therefore, instead of pursuing deterministic single-run outputs via a fixed random seed, we adopted a multi-round iterative black-and-white-list framework (see our previous response in Question 1). Random biases from individual inferences are mitigated by voting and aggregation over multiple independent model judgements.

In addition, we will provide the configured "options" arguments for "ollama.chat" when invoking Gemma 3-27B to facilitate reproducibility in Appendix A1. The seed for each inference run is a 31-bit random integer generated by Python's [secrets] module to ensure statistical independence across runs. This module produces cryptographically-secure random numbers based on hardware-level noise collected by the operating system (hardware interrupts, network timing, CPU clock jitter, etc.), with no user-configurable initial seed.

Specifically:

```
import secrets
seed = secrets.randbits(32) # Set it to 0-2147483647
options = {"temperature": 0.7, "top_p": 0.9, "top_k": 40, "seed": seed, "num_predict": 3072, "num_ctx": 32768, "repeat_penalty": 1.1, "stop": ["\n\n"], "tfs_z": 1.0, "min_p": 0.0}
response = ollama.chat(....., options=options)
```

Question 4

There is also an issue with the mutual exclusivity and conceptual consistency of the flood-type definitions. For example, a tsunami should not be treated as a cause of a storm surge. Moreover, tsunami is simultaneously defined as a separate category, creating a logical classification conflict. Similarly, mountain flood is defined using mudslide as an example. A mudflow or landslide is not simply a type of mountain flood.

The definition of urban flood refers to flooding caused by insufficient drainage capacity during rainfall in urban areas. However, the authors subsequently analyse "urban flood research" in which 71% of the cases are classified as river floods, 18% as mountain floods, and 10% as storm surges. This suggests that the authors may be conflating two different variables: the geographical setting of the study and the flood mechanism/type. These concepts should be clearly distinguished.

Thanks for your suggestion. We believe that the original description of storm surges and flash floods has a probability of causing the LLM model to fall into a classification trap, which means only the occurrence of these phenomena will be included in this classification, thus falling into the decision-making trap of crisp tree (Sauquet and Catalogne, 2011). However, we found that the data provided by LLM after modifying the prompt words was almost the same as when starting the analysis, which shows that our definition based on the fuzzy decision-making method of Sikorska et al. (2015) still gave LLM sufficient thinking space and made reasonable judgments under our multiple-choice conditions. This is the same for urban floods. Because of our vague concept of drainage (without emphasis on pipes) and the existence of the keyword "urban rainfall" at the beginning, LLM did not only identify urban floods as caused by drainage pipes. However, we realize that these prompt words do have definitional flaws or are too specific. In order to prevent readers from making mistakes by reusing keywords, we have modified some of the definitions in Appendix A1.

Sauquet, E., and C. Catalogne., 2011, Comparison of catchment grouping methods for flow duration curve estimation at ungauged sites in France, Hydrol. Earth Syst. Sci., 15, 2421–2435, doi:10.5194/hess-15-2421-2011.

Sikorska, A. E., Viviroli, D., and Seibert, J. et al., 2015. Flood-type classification in mountainous catchments using decision trees. Water Resources Research, 51(10), 7959−7976. https://doi.org/10.1002/2015WR017326

Prompt 4 of the new Appendix A1 is as follows (**blue color**):

Prompt 4: Identify the flood type(s) of the studies, including [urban flood, river flood, mountain flood, coastal storm surge, ice flood, dam break flood, snowmelt flood, tsunami, groundwater flood], multiple choices, or choose [other]:

You are now a professional scholar in the field of hydrology, with a focus on flood-related research and an in-depth understanding of flood and inundation literature. I will provide you with information from flood-related papers,

including the title, abstract, and keywords. Your task is to determine the main type(s) of flood studied in the article. Please choose the most appropriate one or more from the following categories: [urban flood, river flood, mountain flood, coastal storm surge, ice flood, dam break flood, snowmelt flood, tsunami, groundwater flood]. Herein, **urban flood, flood-related inundation events taking place within built-up urban areas; river flood, river water level exceeding the warning line and overflowing; mountain flood, sudden flood caused by intense rainfall in mountainous areas; coastal storm surge, abnormally high sea levels inundating coastal areas due to typhoons or strong storms;** ice flood, floods caused by ice breaking or melting in rivers; dam break flood, sudden flooding due to dam or levee failure; snowmelt flood, floods caused by rising river levels due to melting snow; tsunami, large sea waves caused by underwater earthquakes or volcanic eruptions resulting in coastal flooding; groundwater flood, surface flooding caused by rising groundwater levels; simulations, observations, or predictions related to floods. Only return the category name(s). Avoid selecting multiple options if possible. Choose only from the listed categories. Do not include any extra explanation. If you cannot determine the category, return “other” and do not select any of the other categories. Do not add any extra text, only return the category name(s) in English, separated by (;). Do not include any extra explanation.

Question 5

The proposed influence index, defined as the mean number of citations per article, does not appear to be a sufficiently robust measure. One major problem is that older publications have had considerably more time to accumulate citations. Moreover, comparisons among countries, journals, or institutions without normalization for publication year, research field, and publication age may lead to misleading conclusions. It should also be emphasized that citation impact is not equivalent to scientific quality. An article may be highly cited because it is fundamental, controversial, methodological, widely used, or even because it contains an important error. Citation counts measure bibliometric impact rather than scientific quality directly.

Thank you for your very pertinent suggestions. In this study, these indicators, such as the number of citations of a single publication in a journal, are mainly used for relative ranking and horizontal comparison, and we do not interpret them as an absolute evaluation of scientific quality. In fact, we have considered using the most recent year of publications as a consideration, but data on the daily or annual citations of each publication are not widely publicly available in various journals, especially for publications before IF became a mainstream definition (about 1990s). This makes some comprehensive indices (such as the average citations in the 5/10-year time window, Nature Index and Altmetric) are difficult to calculate. Therefore, in a similar study, Miao et al. (2024) also directly used the current total citations of publications as the basis for comparison.

Miao, C., Hu, J., Moradkhani, H., and Destouni, G., et al., 2024. Hydrological research evolution: A large language model-based analysis of 310 000 studies published globally between 1980 and 2023, Water Resour. Res., 60, e2024WR038077, https://doi.org/10.1029/2024WR038077

Additionally, we will clarify the above limitations of this metric in chapter [2.4 Data-driven interpretation] (**blue color**):

It should be noted that the influence index based on average citations is a bibliometric indicator affected by the year of publication and discipline. It is only used for relative comparison within flood research and is not used for absolute evaluation across disciplines.

It is worth adding that in our recent unpublished research, we obtained the monthly citation count of each publication in the past 10 years. This is the time range in which we can obtain relatively complete data for each journal, and we will try to introduce measurement methods such as fixed time window correction and field normalization. Thank you for your valuable advice.

Question 6

The study relies exclusively on Web of Science. Consequently, the results primarily describe global flood research as represented in Web of Science, rather than the entirety of global flood research. This limitation should be explicitly acknowledged when interpreting and generalizing the findings.

Thank you for your important reminder. We will add a clear statement of limitations in the conclusion section of future revised drafts to remind readers to pay attention to this constraint when interpreting the extrapolation of the results of this study (**blue color**):

It should be noted that all publications analysed in this work were retrieved solely from the Web of Science database. Our findings therefore represent only flood-research outputs indexed by Web of Science, and may not fully capture global flood-related studies. Some regional studies, non-English publications and local-journal articles could be under-represented. This database-specific bias should be kept in mind when interpreting and generalising the presented results.

Question 7

There are also errors in the prompts. For example, the location prompt contains the example "North America, USA, London," which is geographically incorrect if it refers to London in the United Kingdom. Furthermore, in the prompt concerning countries, the authors provide the examples "China, American, England, Japan." Here, "American" is not a country name, while "England" is not equivalent to the sovereign state of the United Kingdom. Such inconsistencies may affect the standardization and reliability of the extracted data and should be corrected.

Thank you for your suggestion. Our manuscript was compiled through the collaboration of multiple authors, which led to errors in classifying national sovereignty in the text. For example, we labeled it USA in the illustrations but might refer to it as America in the main text. Therefore, we made the following corrections:

(1) In the main text, wherever "America" actually referred to "United States of America (USA)" was corrected to USA, and similarly, irregular geographical names referring to "United Kingdom (UK)" were also corrected.

(2) Prompt 2 and Prompt 7 in Appendix A1 were updated, with changes mainly indicated **by blue colors.**

Prompt 2: Get the geographical location:

You will now receive information from a research paper, including its title, abstract, and keywords. Your task is to identify the geographical region studied in the article. Acceptable answers include rivers, lakes, oceans, cities, mountains, and other locations. The only requirement is to extract the most specific location possible. In your response, separate each hierarchical region with a comma, listing from the largest to the smallest. For example: Asia, China, Yellow River; Europe, France, Seine River; South America, Amazon; Africa, Niger River; Africa, Zambia, Lake Kariba; Asia, China, Beijing; Asia, China, Guangxi, Nanning; North America, USA, **New York;** Asia, Japan, Tokyo; Asia, Japan. Important rules: If no specific geographical location can be determined, return None. If the article only refers to a global scale, also return None. If multiple locations are mentioned, only return one. Only provide geographic names; do not include any extra text, quotation marks, spaces, or line breaks. Carefully assess the content before providing the location. If a level of location cannot be determined, skip it; if it can be, include it. Strictly follow these rules and do not mess things up.

Prompt 7: Obtain all agency addresses and their countries:

Now, I will give you a piece of text. Your task is to extract all institution names mentioned in it and remove any irrelevant non-primary unit information such as department names, faculty names, street addresses, or postal codes—these are not needed. Only retain the core names of universities, national research centers, scientific institutions, or companies. Then, convert each extracted name into its full official English name. For example, from "Free Univ Amsterdam, Fac Earth & Life Sci, Boelelaan 1085-1087, NL-1081 HV Amsterdam, Netherlands", extract "Free Univ

Amsterdam" and convert it to "Vrije University Amsterdam"; from "UNIV BOLOGNA, INST HYDRAUL CONSTRUCT, VIALE RISORGIMENTO 2, I-40136 BOLOGNA, ITALY", extract "UNIV BOLOGNA" and convert it to "University of Bologna"; from "Research Center for Eco-Environmental Sciences, CAS", extract "CAS" and convert it to "Chinese Academy of Sciences". Finally, return all identified institutions in the text along with their corresponding countries. Use country names in forms **like China, USA, UK, Japan,** etc.. Institution names must be written in full official form, without using any abbreviations, for example, do not use "UNIV" but write "University" instead. If no recognizable institution can be extracted, return None. Do not include any explanatory dialogue. Note, the correct return format is: xxx Institution (Country); xxx xxx Institution (Country); xxx Institution (Country).