

Thank you very much for your valuable suggestions. We have answered your questions (blue text) one by one in the reply below (black color). However, the current status of the journal does not allow us to modify the manuscript or submit the manuscript as supplementary material in the discussion. Therefore, we will directly mention or show our changes, such as modified pictures, text to be added and comparison with the old text, etc. For more details, please refer to our the PDF attachment.

Question 1

The study relies on Gemma 3-27B to identify relevant articles and extract study locations, flood types, topics, methods, institutions, and countries. Because these outputs form the basis of the results, the authors should explain how their reliability was verified.

Answer to question 1

Thanks for your advice. As we wrote in the Materials and Methods and Appendix, the article database was originally obtained from the Web of Science and further advanced the data based on the large language model, but the details of processing flow are not mentioned, so we will add our data processing details which include our methods and pipelines to ensure data reliability to our manuscript in the future.

In short, this is a multi-layered collaborative strategy consisting of blacklists and whitelists, and includes expert review mechanisms, and here I will provide instructions based on the pipeline content that shown 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.

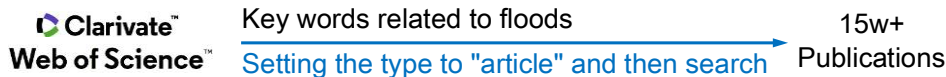
Sub-step 2: Deploy the same Prompt 1 to recheck publications inside the blacklist. If flood-related valid publications were mistakenly putted in the blacklist, such publications would be extracted to whitelist and flood-irrelevant publications remained in the blacklist.

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.

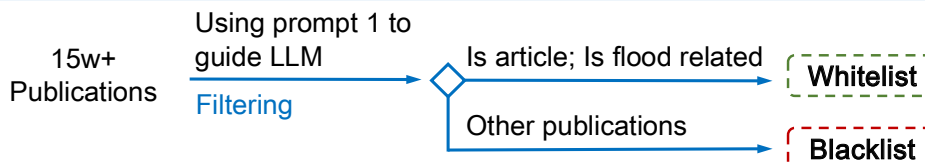
In Step 4, we introduced manual expert verification as the final reliability guarantee. We randomly sampled 4,000 papers mixed from the whitelist and blacklist, and invited four doctors specialized in hydrology and flood disaster

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.

Step 1, filtrating from Web of Science

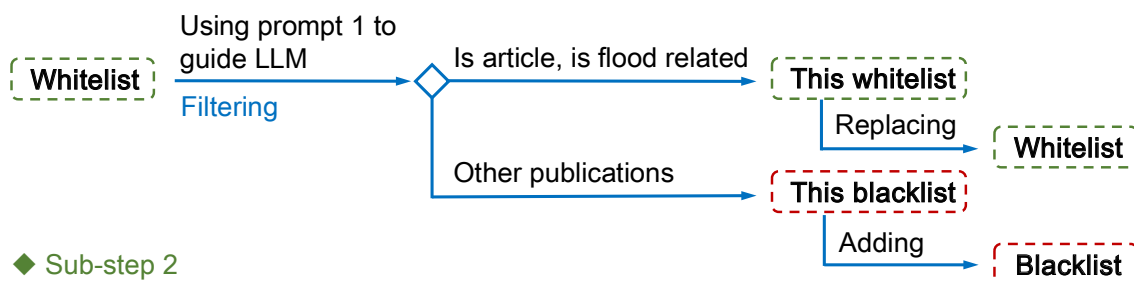


Step 2, filtrating publications' type again and determine whether it is related to flooding

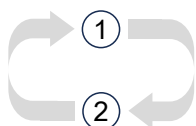
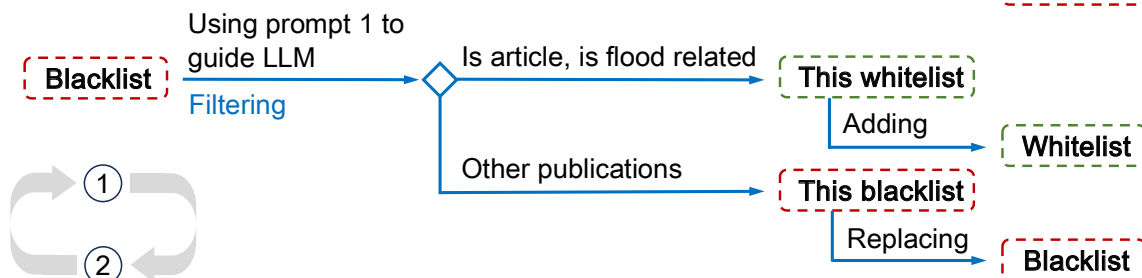


Step 3, filtering alternately using blacklists and whitelist

◆ Sub-step 1



◆ Sub-step 2



Starting from sub-step 1, repeat sub-steps 2 and 1 a total of 10 times

Publications with a frequency exceeding 50% in the whitelist will be included



Step 4, manual verification



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

Based on the two-layer verification system of "large model cycle screening + expert manual calibration", we finally

obtained the official publication whitelist. This complete process 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. P1, 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.

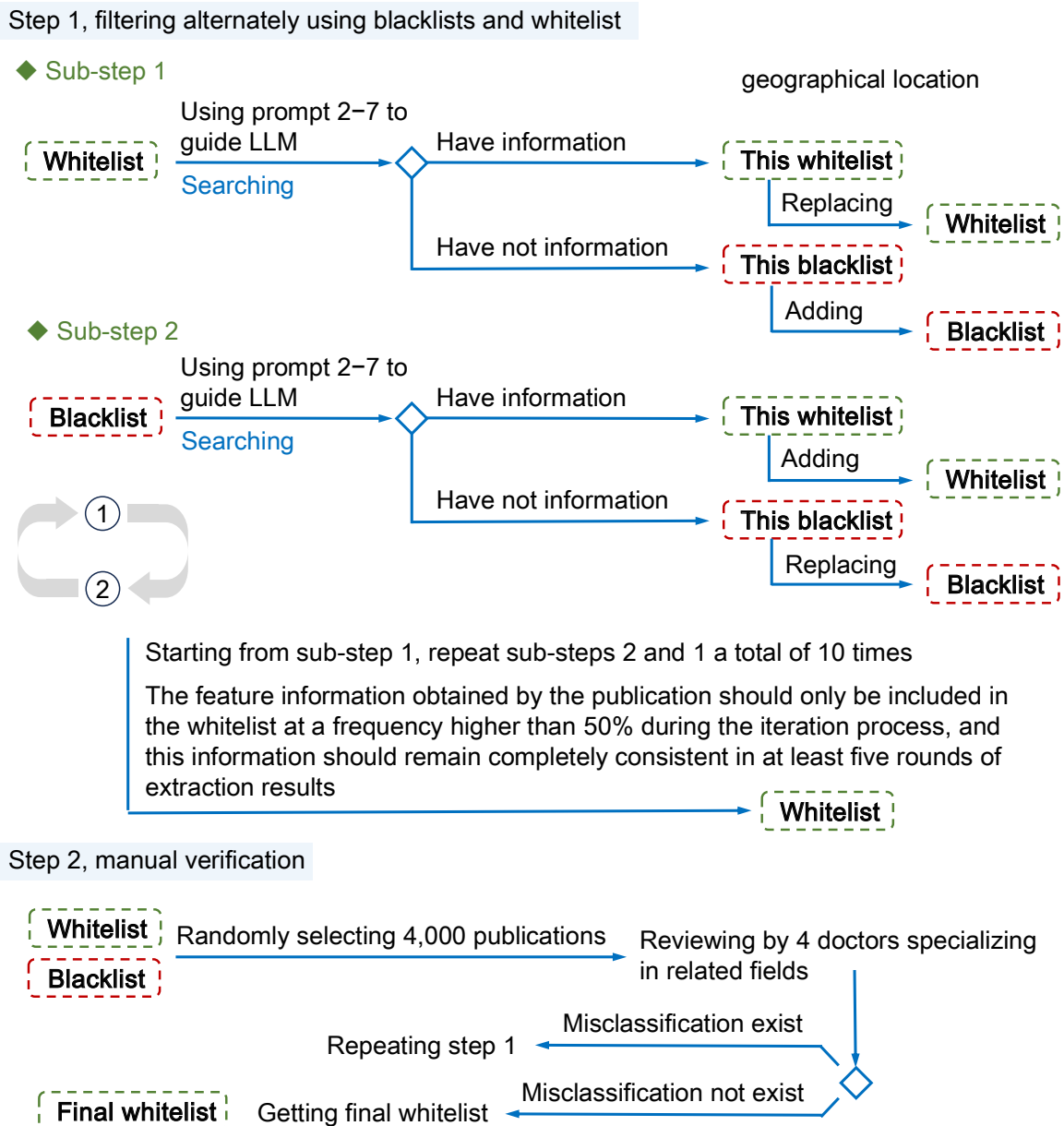


Fig. P2 Pipeline for extracting multi-dimensional characteristic information from publications, with all target features listed in Appendix A1. Prompts 1–7 in Appendix A1 are applied to extract structured information including study locations, flood types, research topics, methodologies, affiliated institutions 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

The Introduction discusses earlier LLM-assisted reviews, including a study analysing 310,000 hydrology publications. The authors should explain more clearly how the present manuscript differs from previous large-scale hydrological literature reviews.

Answer to question 2

Thank you for your valuable suggestions. The reason why we present these articles in the introduction is to demonstrate that we have analyzed the current research status and are fully aware that in the field of hydrology, we are not the first to use LLM-assisted literature reviews. The introduction here is as follows (green color):

Especially in the field of water resources, Miao et al. (2024) integrated an LLM with geographic coding to analyze the research characteristics of 310,000 publications related to hydrology from 1980 to 2023 across spatiotemporal scale, examining aspects like publications, keywords, institutions, countries, and basin attention. He et al. (2025) combined bibliometric analysis with an LLM to comprehensively review 3,106 literature from 1980 to 2025 on remote sensing applications in river obstruction research, analyzing features including publications, involved disciplines, keywords, countries, institutions, and basin attention. Jiang et al. (2022) proposed a natural language understanding-based deep clustering approach to conduct a meta-analysis of thematic evolution patterns in 130,000 lake-related research literature from 1900 to 2020.

Miao et al. <https://doi.org/10.1029/2024WR038077>

He et al. <https://doi.org/10.1016/j.ejrh.2025.102850>

Jiang, et al. <https://doi.org/10.1016/j.jhydrol.2022.128597>

However, this manuscript demonstrates significant differentiation and innovation in some aspects: the spatial analysis framework of the research scope, the depth of information extraction, and the model screening process. It clearly distinguishes from previous large-scale hydrological reviews that relied solely on LLM-assisted literature research to extract basic metadata such as the number of published articles and keywords. Further, Miao et al. (2024) was oriented towards broad hydrology and did not conduct discussions on the classification of any specific sub-directions. He et al. (2025) was only limited to the narrow subfield of river remote sensing, and Jiang et al. (2022) only analyzed lake literature. None of them could fully cover the research context of multiple types of global floods

This article is the first multi-dimensional panoramic quantitative review specifically focusing on forty years of global flood research. We take flood disasters as the main research object, and further extract information to divide the literature into nine types of floods including urban floods, river floods, flash floods, etc., and four research topics including flood management, simulation, disaster monitoring, and risk assessment. It can quantitatively analyze the changes in the proportion of various types of research over the past forty years, making up for the shortcomings of

existing research.

We will add the above explanations in future manuscripts when we enter the revision period.

Question 3

The definitions of several flood categories require clarification. For example, coastal storm surge, tsunami, ice flood, snowmelt flood, and mountain flood may overlap. The prompt also describes the task as multi-label classification but instructs the model to avoid multiple categories whenever possible.

Answer to question 3

We appreciate the professional comments raised by the reviewer regarding the definition of flood categories and the wording of extraction prompts. This study establishes a nine-category flood classification system, including urban flood, river flood, mountain flood, coastal storm surge, ice flood, dam break flood, snowmelt flood, tsunami, and groundwater flood. The boundary of each category is defined primarily by dominant disaster-inducing mechanisms. Although mountain flood, snowmelt flood, ice flood, coastal storm surge and tsunami may overlap in practical research scenarios, their core driving factors are fundamentally distinct. For example, flash floods are triggered by the rapid accumulation and runoff of water resulting from heavy rainfall in mountainous areas; snowmelt floods have seasonal snowpack and melting glacier water as their main sources of runoff; ice floods occur when ice blocks in rivers lead to backups and overflows; coastal storm surges are caused by cyclones and temperate storms, which result in abnormal rises in sea levels; tsunamis, on the other hand, are generated by geological activities such as underwater earthquakes and landslides.

Regarding the “multi-label classification” and “avoiding multiple categories as much as possible” in the prompt words, the two are logically consistent. In this set of prompt words, LLM prioritizes the core flood research objects of the literature, and only performs multi-label output when multiple flood types are the subject of equal weight analysis in the article. This strategy can filter out secondary disaster words that are only incidentally mentioned, effectively reducing redundant annotations in the data set.

In addition, all large models generate classification labels using the stratified sampling manual calibration process used in answer to question 1 to complete multiple rounds of cross-checking, and the classification of the data set is stable and reliable.

In the subsequent section 2.3 of the main text, the definitions of various types of floods based on their disaster-causing mechanisms will be supplemented. At the same time, the logical explanation of the hierarchical extraction of prompt words in Appendix A1 will be improved, and the priority rules of the main label and the conditions for enabling multiple labels will be clarified to reduce misunderstandings.

Question 4

Some interpretations connect publication patterns directly to policies, disasters, economic development, and major infrastructure projects. For example, the manuscript links research changes to the Rio Earth Summit, the EU Water Framework Directive, the Three Gorges Project, Hurricane Katrina, and the Sponge City initiative. These explanations may be reasonable, but the analysis mainly demonstrates temporal associations. The wording should therefore be more cautious. Expressions such as “may have contributed,” “coincided with,” or “could partly explain” would be more appropriate than direct causal statements. The terms “economically driven basins” and “religiously driven basins” should also be reconsidered. The description of the Indus and Ganges as “religiously driven” is broad and may oversimplify complex hydrological, political, economic, and cultural conditions. More neutral and evidence-based terminology is recommended.

Answer to question 4

Thank you for this advice. We carefully re-examined the relevant descriptions throughout the manuscript and revised the words to avoid overstating causal interpretations. In particular, statements linking changes in research patterns to specific policies, disasters, economic development, or infrastructure projects have been changed from direct causal expressions to more cautious formulations. We have also reconsidered the terminology used to characterize different river basins and replaced potentially oversimplified labels with more neutral descriptions based on the characteristics identified in our analysis. Some revisions are summarized as examples below.

(1) Revision of statements concerning the Rio Earth Summit

Before revision, the manuscript stated (green color):

“From 1990 to 1994, the convening of the Rio Earth Summit propelled ‘global change’ research (including climate change and land use change) to the forefront, driving a diffusion of flood research activity to surrounding basins.”

This wording implied that the Rio Earth Summit directly caused the subsequent geographical diffusion of flood research. We have therefore revised it to (orange color):

“From 1990 to 1994, the convening of the Rio Earth Summit coincided with the increasing prominence of ‘global change’ research, including climate change and land use change. This broader research context may have contributed to the subsequent diffusion of flood research activity to surrounding basins.”

This revision makes clear that the observed relationship is a temporal association rather than a demonstrated causal effect. The original analysis indeed identifies the increasing research attention to several river basins during 1990-1994 but does not quantitatively establish that the Rio Earth Summit caused this change.

(2) Revision of statements concerning the EU Water Framework Directive

Before revision, the manuscript stated:

“During the period of 1995–1999, against the backdrop of the development of the EU Water Framework Directive,

transboundary European river basins like the Danube and Rhine gained significantly elevated research status...”

and subsequently:

“Between 2000 and 2004, the EU Water Framework Directive officially entered into force, while the Three Gorges Project ... commenced impoundment in China. The pattern of secondary research areas expanded...”

We have revised these descriptions to emphasize temporal coincidence rather than causality. The revised text is:

“During the period of 1995–1999, the development of the EU Water Framework Directive coincided with increased research attention to transboundary European river basins such as the Danube and Rhine.”

and subsequently:

“Between 2000 and 2004, the EU Water Framework Directive officially entered into force, coinciding with continued growth in research attention to several European basins. During the same period, the Three Gorges Project commenced impoundment in China, while the Yangtze River Basin also emerged as an increasingly important research area. These concurrent developments may partly explain the observed expansion of secondary research areas.”

In this way, we retain the important historical context but avoid implying that the policy or infrastructure project was the sole or direct cause of the observed publication patterns. The original results show the temporal correspondence between these events and the changing research geography.

(3) Revision of statements concerning economic development

The manuscript previously used economic development as a relatively direct explanation for changes in research intensity. For example, the original text stated:

“During 2010–2014, sustained economic growth in countries like Brazil and India led to increased investment in scientific research. Their research outputs began gaining international recognition, resulting in the Paraná River Basin ... and the Ganges River Basin ... being elevated to secondary research areas.”

We have revised this statement to:

“During 2010–2014, sustained economic growth in countries such as Brazil and India coincided with increased research activity and growing international visibility of their research outputs. These developments contributed significantly to the increasing research attention to the Paraná and Ganges river basins.”

This modification is intended to avoid treating economic growth as a demonstrated causal mechanism, while preserving the observed temporal pattern.

Overall, these revisions did not alter the quantitative results or the observed spatial and temporal patterns, and improved the interpretation of these changes by using more neutral, evidence-based basin characteristic terms.

Question 5

The conclusion mainly repeats the descriptive findings. It should more clearly explain how the results can guide future flood research.

Answer to question 5

Thank you for your suggestions. We have revised the conclusion, interspersed it with an overview of the core results, in-depth discussions on disciplinary gaps, and targeted suggestions for future research. The revised conclusion is as follows:

Global flood research faces a serious problem of knowledge fragmentation due to the explosive growth of academic literature. Existing reviews are either oriented to the broad hydrological field or limited to a single flood subdivision, lacking large-scale review covering long time scales. The black-and-white list screening process provides a set of reusable large language model document processing processes for subsequent review development. By screening out 57,474 valid flood research papers based on Gemma 3-27B, we systematically quantified publication timing, institutional country output influence, keyword evolution, proportions of various types of floods, and dual spatial differentiation characteristics of river basins and cities, multi-layer structured information such as disaster types and research methods and built a complete spatiotemporal knowledge map of global flood research. The main conclusions of the review are as follows:

(1) The number of articles, authors and institutions related to flood research has significantly increased.. The annual number of publications has increased from roughly 100 four decades ago to more than 5,500 annually , with the average number of authors and institutions per publication increasing from 2.1 to 4.8 and from 1.4 to 2.6, respectively. *Journal of Hydrology* and *Water Resources Research* lead in terms of influence and productivity. Notably, journals outside the “water resources” category, such as *Climatic Change* and *Remote Sensing of Environment*, have gradually gained significant influence in the flood research field. This reflects a clear trend in the interdisciplinary research model and involves collaboration among multiple authors and different institutions.

(2) The top 15 countries in terms of publication output account for 70% of the global annual total. Chinese and American institutions lead in publication quantity, while European and American institutions dominate in research influence. Among them, the Chinese Academy of Sciences has become the most productive institution globally, while the University of Arizona ranks first nationally in terms of publication influence. The collaborative network for global flood research presents a pattern driven by leading international institutional clusters and interconnected regional institutional groups. However, the clustered cooperation structure creates obvious barriers to cross-continental academic exchange, which is a prominent bottleneck restricting balanced global flood progress. The unbalanced matching between publication quantity (China & US) and research influence (Europe) reveals that massive research outputs from East Asia lack high-impact theoretical innovations. Regional cluster segmentation separates African, South Asian high-risk

regions from core research networks, leading to insufficient local flood monitoring and modeling capacity in disaster-prone developing zones. To address these two prominent defects, future global hydrological communities shall launch cross-continental joint research programs. European high-impact institutions can share mature risk assessment methodologies, while Chinese and American teams transfer urban flood control technologies to understudied tropical and monsoon regions, jointly eliminating the gap between research capacity and actual flood risk across continents.

The changes in high-frequency keywords in different cycles constitute a complete research evolution sequence, which gradually evolves from eco-hydrological system interaction, hydrological and hydrodynamic process analysis, and comprehensive floodplain management to climate-oriented flood prediction and risk assessment. Among them, keywords such as climate change and smart technology have shown a clear upward trend in the past 10 years, while ecohydrology, hydrodynamic simulation and climate risk assessment are still popular research directions. This shift reflects the shift in global flood governance needs from exploring basic hydrodynamic mechanisms to pursuing climate adaptation and risk mitigation. Specifically, by deepening the coupling of artificial intelligence algorithms and traditional hydrophysical models, building a multi-scenario risk prediction framework that simultaneously incorporates long-term climate trends and rapid urbanization changes, and gradually extending mature intelligent monitoring and simulation tools to actual flood research scenarios.

(4) The quantity of urban flood research outputs and their proportion within relevant research topics show a sustained upward trend.. River floods remain a consistent focus (averaging 55% of research), while urban floods have seen a notable rise in attention (increasing from 8% to 15% over the past decade). Mountain floods and coastal storm surges constitute a secondary yet important share of the total research volume (with an average proportion of 8% and 7%, respectively). The primary themes of flood research focus on management, simulation, monitoring, and risk assessment, with relatively balanced proportions. Commonly used research methods are broadly categorized into three types: data collection, data analysis, and hydrological modeling and simulation techniques. Most of the existing simulation and monitoring technologies are designed for a single type of flood, and there is a lack of unified and integrated modeling tools suitable for mixed flood disaster chains To fill this systematic research gap, future studies should conduct targeted field investigations and mechanism experiments on neglected flood types, and construct a unified multi-type of flood coupling simulation system covering river, mountain, coastal and urban waterlogging scenarios.

(5) River basin flood research has shifted from being led by the Mississippi to a bipolar dominance shared with the Yangtze River, complemented by a multi-polar network of other active basins like the Danube, Indus and Saint Lawrence. Similarly, urban flood research has transitioned from leadership by Texas in the United States to a dual-core structure with Guangdong Province in China, complemented by multiple active urban areas such as Greater London in the UK, São Paulo in Brazil. Research themes in urban floods are relatively concentrated, primarily focusing on Sponge City infrastructure development and urban resilience assessment in the context of rapid urbanization, especially in

economically developed river basins and coastal provinces, while high-flood-risk basins (Nile, Ganges-Indus) and inland developing cities receive far less research investment. Cross-border transboundary river basins lack unified coordinated research frameworks, and inland regions cannot replicate mature urban flood control experience from coastal core provinces directly due to unique topographic and climatic conditions. In response to this severe spatial imbalance, future research shall allocate more field observation and numerical simulation resources to underrepresented high-risk zones. Transboundary joint monitoring platforms shall be built for international river basins, and localized low-cost flood mitigation technologies shall be developed for inland mountainous provinces and tropical developing cities.

Although the large model-assisted literature screening and extraction framework established in this paper has significantly improved the efficiency of quantitative analysis of large-scale flood literature, LLMS still have limitations in capturing the subtle semantic differences in domain professional descriptions. It is necessary to rely on multiple rounds of model iterative determination combined with expert manual calibration to ensure data reliability. Even so, the structured dataset produced by this study holds extremely high value for flood research and governance. It can quickly identify research gaps based on the global flood knowledge map of the past four decades and avoid low-level repetitive research. In addition, policymakers can adjust the allocation of regional scientific research resources according to the spatial imbalance pattern of research hotspots, and flood management departments in various regions can refer to river basins. The evolution law of urban differentiated themes formulates targeted flood control plans. The large model processing pipeline constructed in this paper can also be transferred to other water disaster literature review work, accelerating the systematic knowledge sorting of the entire hydrological field and providing theoretical support for global comprehensive flood risk management adapted to climate change.

Question 6

The manuscript contains valuable maps, networks, word clouds, and multi-panel figures, but several labels are small and difficult to read. The geographical figures in particular contain many panels, abbreviated locations, small legends, and crowded labels.

Answer to question 6

Thank you for your suggestions. To address this issue, we first tried to ensure that the color blocks and text in each image are presented in vector format. Additionally, we reduced the clutter caused by labels and adjusted those sections where the font size was too small. You can see these changes in some images below. We will make further modifications once it is possible to upload our updated versions of the manuscript to the system.

For example, we changed Fig. 1 to **Fig. P3** (vector format). Readers can zoom in to see each word in the word cloud diagram.

On the other hand, we change Fig. 9 to **Fig. P4**. We have increased the clarity of the image and adjusted the position of the labels (replacing the abbreviations of geography with their full names) and histograms on the original map to better present the content.

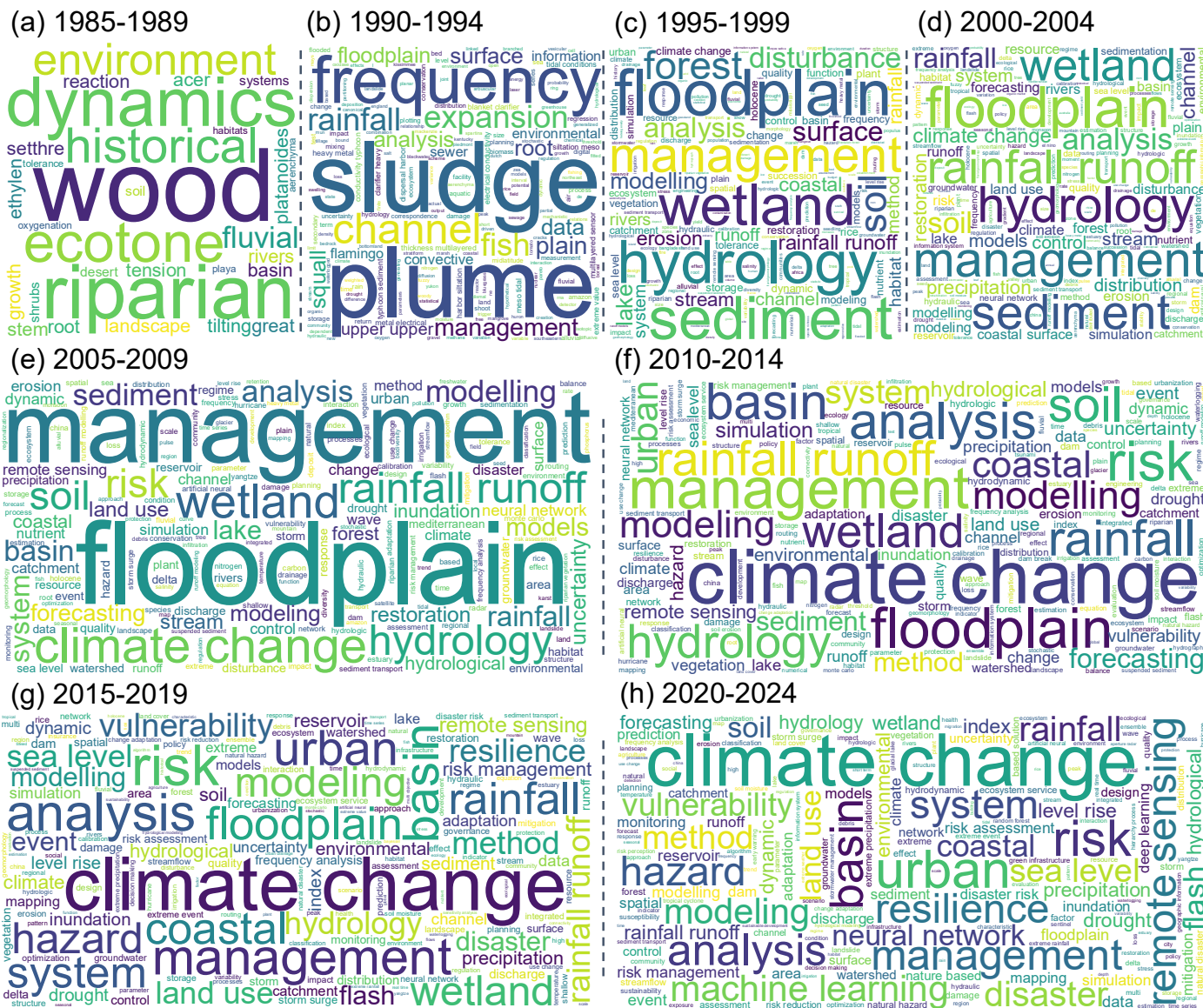


Fig. P3 Word clouds of keywords across different periods. The words in the figure represent the top 200 keywords with the highest frequencies of occurrence in the publications.

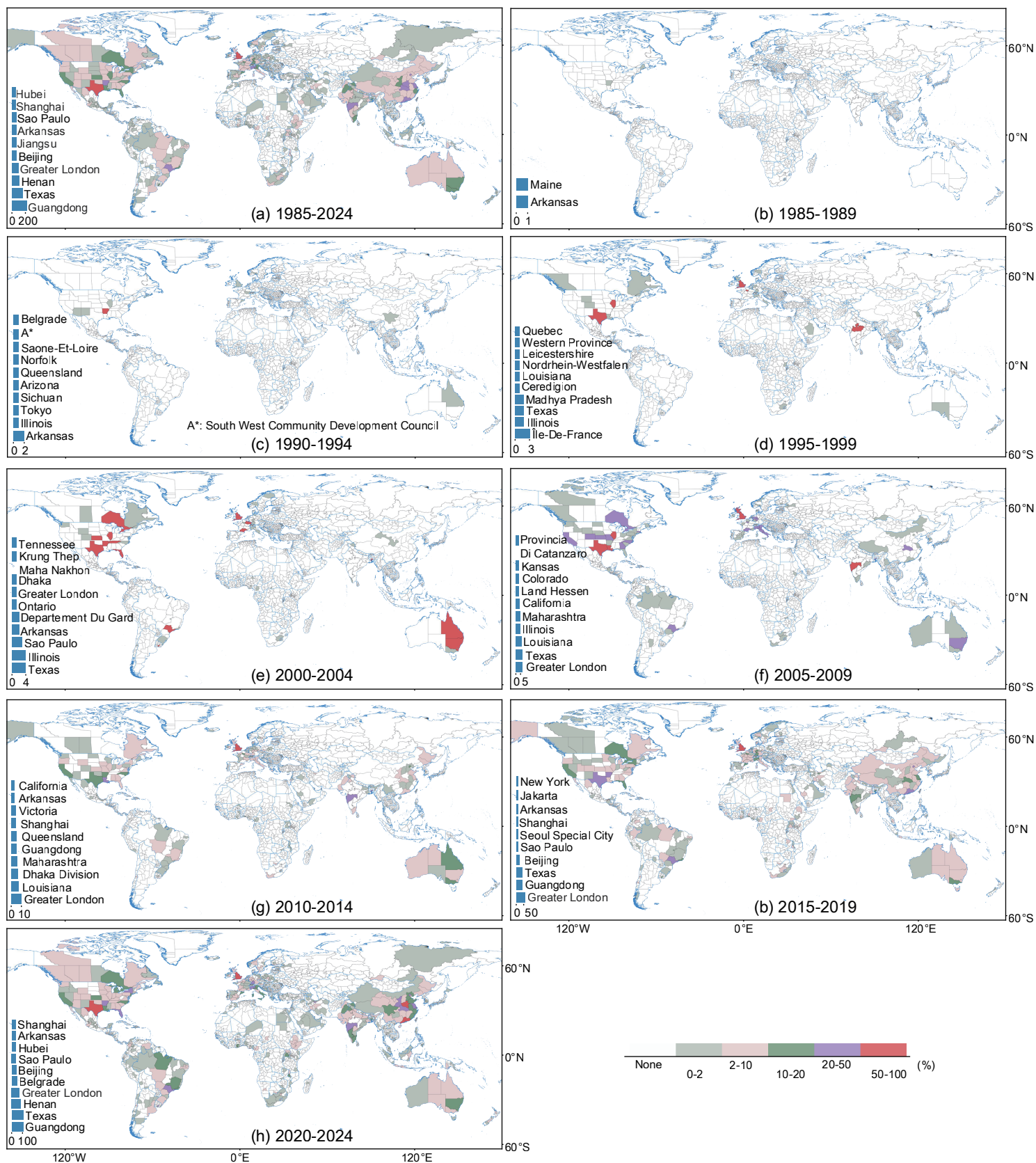


Fig. P4 The percentage of urban where the study area of urban floods located in different periods. The bar chart in the bottom left corner of each subgraph displays the top 10 urban, with long names replaced with letters.