Environmental Data Science (2026), 5: e10, 1–12
doi:10.1017/eds.2026.10037

CAMBRIDGE UNIVERSITY PRESS

POSITION PAPER

How reliable are retrieval-augmented and standard ChatGPT models to support flood susceptibility mapping?

Ali Pourzangbar and Mário J. Franca

Institute for Water and Environment (IWU), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
Corresponding author: Ali Pourzangbar; Email: ali.pourzangbar@kit.edu

Received: 24 September 2025; **Revised:** 05 March 2026; **Accepted:** 20 March 2026

Keywords: ChatGPT; flood susceptibility mapping; large language models; machine learning; retrieval-augmented generation

Abstract

This paper evaluates the performance of baseline and domain-augmented ChatGPT models for literature-based knowledge support in flood susceptibility mapping (FSM) using machine Learning approaches. To assess this, we designed five key questions related to FSM, with benchmark responses derived from our comprehensive review article (Pourzangbar et al., *Journal of Flood Risk Management* **18**, e70042), which analyzed 100 studies on ML applications in FSM. The same questions were posed (i) to standard ChatGPT-4 and ChatGPT-4o models without additional contextual material, and (ii) to a domain-augmented GPT-4 configuration (Chat-FSM) equipped with retrieval access to the 100 reviewed articles. The comparison highlights that GPT-based models can reasonably reproduce frequently reported machine learning models and conditioning factors from the reviewed literature, but show weaker consistency in feature selection methods, often suggesting less relevant techniques. Among the models, ChatGPT-4o demonstrated the weakest alignment with benchmark data, while Chat-FSM demonstrated the highest agreement with the benchmark dataset across most evaluated questions. In terms of application-level efficiency, GPT models required substantially less time and computational effort compared to manual literature synthesis under the defined experimental setup. While ChatGPT-based systems can support literature-informed exploration in FSM, human expertise remains essential for critical reasoning, methodological design, and application to novel or context-specific scenarios.

Impact Statement

This work underscores the potential of domain-augmented large language models to reshape literature synthesis in flood modeling, enabling faster and more informed decision-making in risk assessment.

1. Introduction

Floods affect around 200 million people annually and cause over half of all disaster-related damages globally (Ritchie and Rosado, 2024). Flood susceptibility mapping (FSM) plays a vital role in identifying at-risk areas to support planning and emergency response. While traditional FSM methods, such as physical and numerical models, are often costly and complex, machine learning (ML) models offer accurate and efficient alternatives. Despite their growing use, inconsistencies in data preprocessing,

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

feature selection, model setup, and validation hinder the reproducibility and generalization of ML-based FSM approaches (Pourzangbar et al., 2024).

Emerging large language models (LLMs), including ChatGPT, have attracted significant attention across scientific disciplines. ChatGPT, developed by OpenAI (2023), is a generative matural language processing (NLP) model capable of producing context-aware text responses and assisting with tasks such as programming support, content drafting, and literature exploration (Ray, 2023; Surameery and Shakor, 2023). Prior studies have highlighted its usefulness in summarization and interactive refinement when guided by structured prompts (Halloran et al., 2023), although questions remain regarding the consistency and domain-specific reliability of its outputs in specialized scientific contexts. Rapid academic adoption, reflected in thousands of publications since 2023, underscores both its potential and the need for systematic evaluation in domain-specific applications.

Integrating ChatGPT into water science and disaster management has opened new research opportunities, though challenges persist (Hosseini and Pourzangbar, 2026). In disaster management, it supports knowledge dissemination and data analysis but faces limitations in real-time responsiveness, data quality, and interpretability, requiring rigorous validation (Xue et al., 2023; Haider et al., 2024). In water science, ChatGPT is useful for qualitative tasks and has shown value in hydrology and Earth sciences, although it may occasionally produce errors in quantitative analyses (Foroumandi et al., 2023). In specialized fields like water treatment and desalination, it simplifies technical tasks but often needs human oversight (Ray et al., 2024). The environmental footprint, particularly high freshwater consumption during model training, raises sustainability concerns (Egbemhenghe et al., 2023). While ChatGPT shows potential for supporting scientific workflows, to the best of the authors' knowledge, its consistency and domain-specific performance in FSM—particularly in literature synthesis and technical ranking tasks—remain insufficiently evaluated. We believe that sooner rather than later, LLMs will become increasingly integrated into scientific workflows, and thus understanding their strengths and limitations in domain-specific applications becomes essential. This study evaluates and compares the performance of different ChatGPT configurations in responding to literature-informed questions related to ML-based FSM. ChatGPT was selected due to its accessibility, configurable knowledge integration features, and widespread use in academic environments. Rather than positioning ChatGPT as a replacement for modeling tools, we examine its role as a literature-support system within FSM research workflows. This design enables explicit examination of how domain-specific retrieval augmentation influences literature-consistency in structured FSM queries, while not constituting validation of performance in novel or out-of-corpus applications. Our focus is on assessing response agreement, ranking consistency, and computational efficiency within a controlled benchmarking framework. The novelty of this work lies in systematically benchmarking AI-generated outputs against a curated review dataset, providing empirical insight into the behavior of general-purpose and domain-augmented GPT systems in FSM-related knowledge tasks. Furthermore, this study contributes to ongoing discussions on how LLM-based systems may be systematically evaluated for use in hydrological research contexts.

2. Material and methods

2.1. Benchmark data

Pourzangbar et al. (2025) reviewed over 100 studies on ML-based FSM from 2013 to 2023. This review forms the basis for the benchmark dataset used in this study, serving as a structured reference for evaluating model-generated responses. The review proposed a comprehensive framework covering data preprocessing, model development, validation, and post-processing. The study identified key aspects such as data considerations, algorithm selection, and modeling procedures. Through comparative analysis, the review synthesized common modeling practices and performance trends in ML-based FSM studies. Emphasizing data preprocessing, feature engineering, and model configuration, the study also proposed innovations to enhance modeling quality. The structured findings of this review were mapped onto the five key evaluation questions and used as the benchmark dataset. Notably, the same corpus of reviewed articles was later provided as reference material to the domain-augmented GPT

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

configuration (hereafter referred to as "Chat-FSM"), enabling controlled assessment of retrieval consistency against the synthesized literature outcomes.

2.2. ChatGPT models

Three ChatGPT-based configurations were employed in this study: ChatGPT-4o, ChatGPT-4, and a domain-augmented GPT-4 configuration incorporating selected literature (Chat-FSM). ChatGPT-4 and ChatGPT-4o are advanced OpenAI models that outperform GPT-3 in contextual understanding, coherence, and response generation (OpenAI, 2023). Both are trained using text prediction and reinforcement learning from human feedback (RLHF) and can process text, images, and third-party data (Abdullah et al., 2022). GPT-4 is a large multimodal model reported to perform strongly across a range of academic and professional benchmark tasks (Oxford Analytica, 2023). ChatGPT-4o extends GPT-4 by adding audio input, faster responses, support for over 50 languages, and improved literary analysis, while being more cost-effective. Both models feature a 128,000-token context window, with knowledge cutoffs in December 2023 (GPT-4) and October 2023 (ChatGPT-4o) (Montañés, 2024).

Chat-FSM was developed using OpenAI's Custom GPT environment with integrated document upload capabilities, allowing retrieval-augmented generation over a curated document collection. The system does not involve model fine-tuning or custom API-level pipeline modifications. Instead, it operates by combining document retrieval from the uploaded corpus with the pretrained capabilities of the base GPT-4 model. For this setup, Chat-FSM was granted retrieval access to a collection of 100 scientific publications—the same set of articles used to construct the benchmark dataset (Section 2.1)—focused on ML-based FSM. Using this identical corpus as the retrieval source enables a controlled evaluation of how effectively the model reproduces the synthesized findings from the literature. Chat-FSM lacks web browsing and external data analysis capabilities, relying solely on uploaded PDFs for knowledge. Its scope is limited to ML applications in FSM, with predefined conversation starters to guide exploration.

While the primary evaluation focuses on literature-consistency benchmarking, an additional out-of-corpus test was conducted to address concerns regarding potential circular validation and to provide a preliminary assessment of generalizability beyond the configured knowledge base. Specifically, Chat-FSM was evaluated against an independent benchmark FSM study (Zhuang et al., 2026) that was not included among the 100 articles used for knowledge augmentation. This external reference enabled examination of whether the model could synthesize relevant concepts without direct retrieval access to the source material, thereby offering an initial indication of its behavior in out-of-corpus scenarios.

2.3. Methodology

To systematically evaluate the performance of the selected ChatGPT configurations, five key literature-informed questions were designed, focusing on ML models, conditioning factors, feature selection methods, and major challenges in FSM (Table 1). These questions target frequently reported themes in ML-based FSM literature to enable structured benchmarking of model-generated responses. The fixed question set ensures comparability across models and facilitates quantitative assessment of content overlap and ranking consistency. Model responses were compared against the benchmark dataset (Section 2.1) using content-overlap and rank-correlation metrics to quantify agreement in both item presence and ordering. The evaluated questions primarily assess literature retrieval and structured ranking consistency rather than context-dependent analytical reasoning. Queries such as identifying commonly used models or frequently reported conditioning factors examine alignment with published trends. In contrast, tasks requiring model selection for specific geographic, hydrological, or data-scarce contexts involve synthesis and expert judgment that extend beyond the scope of the present benchmarking framework. Accordingly, the findings should be interpreted as assessing literature-consistency rather than full analytical reasoning capability. Figure 1 summarizes the comparative evaluation framework implemented in this study.

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

Table 1. *Five key questions designed in this study to assess the agreement of different ChatGPT configurations with the benchmark dataset*

Queries	Theme	Question posed to the ChatGPT models (ChatGPT-4o, ChatGPT-4, Chat-FSM)
Q1	Commonly used ML models for FSM	Rank the top 10 machine learning models from most to least frequently used and best-performing in flood susceptibility modeling. Provide the list in descending order of their combined frequency and performance.
Q2	Key conditioning factors influencing flood susceptibility	Identify and rank the top 10 most important input parameters for flood susceptibility mapping based on their reported impact on model performance. List them from most to least important.
Q3	Most frequently utilized parameters given FSM	Rank the top 10 most frequently used input parameters in machine learning models for flood susceptibility mapping, based solely on how often they appear across various studies. Provide a numbered list in descending order of frequency.
Q4	Most popular methods for feature selection in FSM	Identify and rank the top 10 most frequently utilized feature selection methods for determining the importance of conditioning factors in flood susceptibility modeling. List them in descending order of frequency.
Q5	Significant gaps and challenges in FSM	What are the top five most significant gaps and challenges in developing machine learning models for flood susceptibility mapping, considering aspects like data availability, model generalizability, interpretability, and integration with hydrological knowledge? Rank them from most to least critical based on their impact on model performance and applicability.

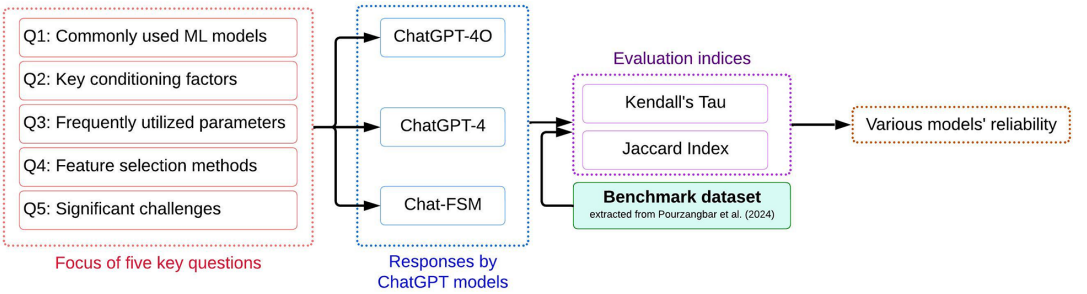

Figure 1. *The procedure used in this study to evaluate the agreement of different ChatGPT configurations with the FSM benchmark dataset.*

2.3.1. Out-of-corpus generalization

In addition to the primary literature-consistency evaluation, an out-of-corpus assessment was conducted using an independent FSM benchmark study not included in the retrieval corpus, allowing preliminary examination of generalization behavior beyond the configured knowledge base. For the out-of-corpus evaluation, we selected the study by Zhuang et al. (2026), which combines social media-derived

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

Table 2. *Three out-of-corpus questions designed to evaluate the generalization behavior of Chat-FSM through comparison with an independent benchmark study (Zhuang et al., 2026)*

Queries	Theme	Question posed to the Chat-FSM
QOC–1	Integration of emerging data source	Describe the integration of social media-derived event data into flash flood susceptibility modeling by listing: (i) the core data-processing steps and (ii) the main comparative advantages over traditional inventory sources.
QOC–2	Limitations and bias of social media-derived inventories	What are the main limitations and potential biases when using social media-derived flash flood event inventories for susceptibility modeling, particularly in regions with uneven population density and heterogeneous reporting activity?
QOC–3	Incorporation of typhoon frequency as a regional driver	In coastal regions frequently affected by tropical cyclones, why might incorporating typhoon frequency as a predictor improve flash flood susceptibility modeling, and what impact would you expect it to have on model performance?

Abbreviation: QOC = Question-Out-of-Corpus.

flash flood inventories with machine learning-based modeling in a typhoon-affected coastal environment. As this study was not included in the 100-article corpus used to configure Chat-FSM, it was employed as an independent external benchmark to assess the model's generalizability and robustness beyond the training domain. Three questions were designed to evaluate conceptual transfer across three dimensions (Table 2). First, the selected study employs social media–derived flash flood inventories, representing an emerging data source in susceptibility modeling and providing a suitable test of whether the model can generalize to new forms of observational input. Second, the study explicitly addresses methodological limitations and reporting biases associated with social media data, enabling assessment of whether Chat-FSM can recognize uncertainties and data-quality constraints rather than merely describing advantages. Third, the study incorporates typhoon frequency as a region-specific hydro-meteorological driver, requiring compound hazard reasoning and evaluation of its expected influence on model performance. Together, these dimensions allow a structured examination of generalization beyond the configured literature corpus.

2.4. Inference statistics

The evaluation followed a structured approach. First, the ChatGPT models were prompted to generate ranked lists of machine learning models and parameters relevant to flood susceptibility mapping. Next, responses were reviewed by the authors to ensure semantic alignment between model outputs and benchmark components prior to quantitative comparison. Finally, the degree of agreement in content and ranking was quantified using two statistical measures: the Jaccard Index for content match and Kendall Tau for rank correlation (Mulekar and Brown, 2017).

Each of the five key questions was posed to the three configurations, and their responses were analyzed to quantify agreement with the benchmark dataset. For each question (except Q5), the output was a list of 10 components. Responses were compared to the Benchmark data using two criteria: Match and Ranking. Match refers to the presence of a component in both the model's response and the Benchmark list, regardless of order, indicating content overlap. Ranking assesses the degree of ordinal alignment between the model's ordering and the benchmark ordering, reflecting consistency in relative prioritization. For instance, if a model ranks "RF" first while the Benchmark ranks it fourth, the component matches, but the

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

ranking differs. This dual-criteria approach enables a comprehensive assessment of both content and structure.

To quantitatively compare model agreement in terms of Match and Ranking, Kendall's Tau (τ), and the Jaccard Index (J) were calculated (Eqs. 1–2). Kendall's Tau measures ordinal correlation between the model-generated ranking and the benchmark ranking, ranging from 1 (complete discordance) to 1 (perfect concordance), and was computed on overlapping components (Kendall, 1938). The Jaccard Index quantifies set similarity as the ratio of the intersection to the union of elements in the two lists, ranging from 0 (no overlap) to 1 (perfect overlap) (Jaccard, 1901).

$$\tau = (C - D)/(C + D) \tag{1}$$

$$J(A,B) = |A \cap B|/|A \cup B| \tag{2}$$

where C and D denote the number of concordant and discordant pairs, respectively. A pair is concordant if the relative order of two elements is consistent across both lists, and discordant otherwise.

3. Results

3.1. Literature-consistency benchmark results

The comparative agreement between the ChatGPT configurations and the benchmark dataset is summarized in Supplementary Appendix B (Tables B1–B5 and Figures B1–B4). The following sections analyze the degree of content overlap and ranking consistency across the five evaluation questions. These questions address commonly reported ML models for FSM, key conditioning factors, frequently utilized input parameters, feature selection methods, and documented gaps and challenges identified in the reviewed literature.

Question 1: *"Rank the top 10 machine learning models from most to least frequently used and best-performing in flood susceptibility modeling. Provide the list in descending order of their combined frequency and performance."*

The benchmark dataset and corresponding model responses for this question are presented in Supplementary Appendix B (Table B1 and Figure B1). Based on the data in Table B1, the Jaccard Index and Kendall's Tau were calculated for each configuration, as summarized in Table 3. Although its ranking correlation (Kendall's Tau = 0.205) is lower, Chat-FSM shows the highest level of content overlap with the benchmark (Jaccard Index = 0.636) among the evaluated configurations. ChatGPT-4o exhibits moderate content overlap (Jaccard Index = 0.428) but achieves the highest rank correlation (Kendall's Tau = 0.244), indicating stronger ordinal alignment with the benchmark ordering among overlapping components. In contrast, ChatGPT-4 exhibits lower agreement levels under both metrics (Jaccard Index = 0.385; Kendall's Tau = 0.022). Notably, none of the evaluated configurations included several models highlighted in the benchmark dataset, such as Bagging ensembles, long short-term memory (LSTM), and deep belief networks. Additionally, some configurations proposed models such as logistic regression, K-nearest neighbors, and Naïve Bayes, which were not ranked among the top 10 most frequently reported models in the benchmark dataset. The benchmark dataset indicates an increasing emphasis in recent literature on ensemble and hybrid modeling strategies, such as random forest combined with optimization algorithms. This trend was not prominently reflected in the responses generated by the evaluated configurations.

Table 3. *Inference statistics of different ChatGPT models considering Q1*

Criterion	Statistical index	ChatGPT-4o	ChatGPT-4	Chat-FSM
Match	Jaccard Index	0.428	0.385	0.636
Ranking	Kendall Tau	0.244	0.022	0.205

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

Table 4. *Inference statistics of different ChatGPT models considering Q2*

Criterion	Statistical index	ChatGPT-4o	ChatGPT-4	Chat-FSM
Match	Jaccard Index	0.429	0.538	0.667
Ranking	Kendall Tau	−0.600	−0.156	−0.022

Question 2: *"Identify and rank the top 10 most important input parameters for flood susceptibility mapping based on their reported impact on model performance. List them from most to least important."*

The benchmark data and corresponding model responses shown in Supplementary Table B2 and Figure B2 were used to compute the Jaccard Index and Kendall Tau, as summarized in Table 4. For consistency in the comparison, "geology" was treated as a composite category encompassing rocks (lithology), soils, minerals, tectonic features, and geological history. As such, both "soil type" and "lithology" are considered subcategories of "geology" and were included under this classification when calculating the match and ranking indices. Inspection of Table 3 indicates limited ordinal agreement between model-generated rankings and the benchmark ordering of input parameters, as reflected by negative Kendall's Tau coefficients across all configurations; however, Chat-FSM exhibited comparatively higher agreement levels in both content overlap and rank correlation relative to ChatGPT-4o and ChatGPT-4.

Question 3: *"Rank the top 10 most frequently used input parameters in machine learning models for flood susceptibility mapping, based solely on how often they appear across various studies. Provide a numbered list in descending order of frequency."*

The benchmark dataset and corresponding model responses—shown in Supplementary Table B3 and Figure B3—were used to compute the Jaccard Index and Kendall's Tau, with results summarized in Table 5. In terms of content overlap (Jaccard Index), ChatGPT-4 achieved complete set agreement with the benchmark list, whereas ChatGPT-4o and Chat-FSM exhibited moderate levels of overlap. However, in terms of ordinal alignment, ChatGPT-4 exhibits the lowest Kendall's Tau coefficient (−0.822), followed by ChatGPT-4o (−0.244), indicating inverse correlation with the benchmark ordering. Chat-FSM, although exhibiting a comparatively higher Tau value, still demonstrates limited rank correlation with the benchmark ordering.

Table 5. *Inference statistics of different ChatGPT models considering Q3*

Criterion	Statistical index	ChatGPT-4o	ChatGPT-4	Chat-FSM
Match	Jaccard Index	0.666	1.00	0.666
Ranking	Kendall Tau	−0.244	−0.822	0.289

Table 6. *Inference statistics of different ChatGPT models considering Q4*

Criterion	Statistical index	ChatGPT-4o	ChatGPT-4	Chat-FSM
Match	Jaccard Index	0.176	0.176	0.176
Ranking	Kendall Tau	0.378	−0.111	−0.333

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

Question 4: *"Identify and rank the top 10 most frequently utilized feature selection methods for determining the importance of conditioning factors in flood susceptibility modeling. List them in descending order of frequency."*

The benchmark dataset and corresponding model responses shown in Supplementary Table B4 and Figure B4 were used to compute the Jaccard Index and Kendall's Tau, as summarized in Table 6. Inspection of Table 6 indicates limited agreement between the model-generated lists and the benchmark, reflected in low content overlap and modest or negative rank correlation values. Based on these agreement metrics, the generated responses show limited consistency with the benchmark ordering for feature selection methods under the defined evaluation framework.

Question 5: *"What are the top five most significant gaps and challenges in developing machine learning models for flood susceptibility mapping, considering aspects like data availability, model generalizability, interpretability, and integration with hydrological knowledge? Rank them from most to least critical based on their impact on model performance and applicability."*

Both the benchmark dataset and the evaluated ChatGPT configurations identify data availability and quality as central challenges in ML-based FSM (Table B5). The benchmark dataset emphasizes strategies reported in the reviewed literature, such as expanding data inputs through augmentation and fusion, integrating climate change indicators, and incorporating social media data in data-scarce regions. It also highlights the importance of physical interpretation and trade-off analysis in flood risk assessment. In contrast, the ChatGPT-generated responses tend to emphasize technical aspects related to model robustness and adaptability. ChatGPT-4o and ChatGPT-4 frequently highlight model generalizability across regions, interpretability, and integration with physical or hydrological knowledge as important considerations in FSM applications. Chat-FSM places comparatively greater emphasis on mitigating overfitting, improving feature selection, and enhancing interpretability.

The distinction lies in emphasis: the benchmark dataset reflects literature trends that prioritize expansion of data inputs and practical data-gap mitigation, whereas the ChatGPT-generated responses place greater weight on methodological aspects such as generalization, explainability, and multi-scale integration. These differences illustrate variation in thematic prioritization rather than categorical correctness.

3.2. Out-of-corpus generalization results

The out-of-corpus generalization assessment was conducted using the Chat-FSM configuration. As the primary objective of this experiment was to examine whether domain-configured retrieval augmentation enables conceptual transfer beyond the configured literature corpus, the evaluation focused on the custom retrieval-augmented system. The previously tested standard ChatGPT configurations were not included in this additional experiment. Accordingly, this assessment isolates the behavior of the retrieval-augmented system when exposed to an independent benchmark study not represented in its knowledge base. Three evaluation questions (QOC-1 to QOC-3; c.f. Table 2) were formulated based on the content of this external study. For each question, the benchmark answer was derived directly from the selected paper and compared with the response generated by Chat-FSM (Supplementary Appendix A contains the Chat-FSM responses to these questions, and Supplementary Appendix B includes a comparison between the benchmark and Chat-FSM responses).

For QOC-1 and QOC-2, which focus on the integration of social media–derived flash-flood inventories, the model responded that the information was not available in the provided documents. This behavior indicates that the system correctly recognized the absence of relevant knowledge within its retrieval corpus rather than generating unsupported or hallucinated explanations. While this conservative response avoids misinformation, it also highlights a limitation: when the uploaded literature does not cover the queried topic, the model cannot generalize beyond the knowledge contained in its indexed documents. In contrast, for QOC-3, which examines the role of typhoon

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

frequency as a predictor in flash-flood susceptibility modeling, the model produced a detailed explanation that largely aligned with the benchmark response extracted from the selected paper. The generated answer correctly identified key mechanisms discussed in the literature, including the relationship between tropical cyclones and intense short-duration rainfall, the representation of regional hazard exposure in cyclone-prone coastal areas, and the expected improvement in predictive performance metrics such as model accuracy. Additionally, the model introduced further explanatory elements such as cyclone pathways and spatial differentiation of flood susceptibility which, although not explicitly discussed in the benchmark paper, remain conceptually consistent with established hydrometeorological reasoning.

4. Discussion

4.1. Variability across ChatGPT versions

Each ChatGPT configuration was included for a specific reason: ChatGPT-4o represents a more recent publicly accessible model (in comparison to others); ChatGPT-4 is widely used in academic workflows; and Chat-FSM, configured with retrieval access to FSM literature, represents a domain-augmented configuration. Their inclusion enables comparative analysis across general-purpose and literature-augmented systems.

The results reveal measurable differences in content overlap and rank correlation across configurations. For ML model rankings (Q1), Chat-FSM exhibited the highest content overlap with the benchmark dataset, whereas ChatGPT-4o and ChatGPT-4 showed lower levels of agreement, including omission of several models highlighted in the benchmark synthesis. For conditioning factors (Q2), all configurations demonstrated limited ordinal agreement with the benchmark ordering, although Chat-FSM showed comparatively higher content overlap. For feature selection methods (Q4), agreement levels were low across all configurations, indicating substantial divergence from the benchmark list under both metrics. The out-of-corpus generalization results suggest that Chat-FSM demonstrates strong reasoning and explanatory capability when the queried concepts are represented within its retrieval knowledge base. However, its performance remains dependent on the coverage of the provided literature, and it is unable to supply substantive answers when relevant information is absent from the indexed documents. This behavior reflects the inherent characteristics of retrieval-augmented systems, where generalization is constrained by the scope of the underlying document corpus. Observed differences likely stem from variations in training data exposure, knowledge cut-off dates, and architectural configurations. ChatGPT-4o and ChatGPT-4 rely on general pretraining and reinforcement learning from human feedback, whereas Chat-FSM operates under a retrieval-augmented setup constrained to a defined corpus of FSM literature. These configuration differences influence response patterns and levels of agreement under the defined benchmarking framework. Notably, higher content overlap in the domain-augmented configuration reflects improved literature-consistency rather than validated superiority in broader analytical reasoning.

Large language models are continuously updated, which may lead to variability in generated outputs over time. The findings indicate that while LLM-based systems can efficiently generate literature-aligned summaries, their outputs should be interpreted within the context of benchmark-based agreement rather than assumed domain reliability, particularly in high-stakes applications such as flood risk management.

4.2. Implications of model variability for literature-based FSM support

Variability across ChatGPT configurations in literature-informed FSM responses may affect confidence in their consistency, particularly if such outputs are interpreted without awareness of their benchmark-based limitations. While ChatGPT-generated outputs may assist with exploratory literature review and conceptual comparison, their use in formal decision-making should be approached cautiously, given the observed variability in agreement with literature-derived benchmarks.

Over-reliance on GPT-generated summaries without independent verification may lead to incomplete representation of literature trends or methodological considerations identified in structured reviews. For example, omission of models emphasized in the benchmark synthesis may influence how methodological

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

options are perceived within exploratory analyses. Additionally, generalized responses may not fully capture location-specific hydrological or socio-environmental nuances that are typically addressed through domain expertise. Accordingly, ChatGPT-based tools are best positioned as supplementary literature-support systems rather than substitutes for domain expertise. Cross-verification with structured reviews and collaboration with domain experts can help contextualize AI-generated outputs, particularly in complex applications such as flood susceptibility assessment.

4.3. Energy efficiency of the process

The energy implications of ChatGPT systems can be considered from two perspectives: (1) model training and development, which are resource-intensive processes external to this study, and (2) application-level use during deployment, which is the focus of the present comparison. Training large language models such as GPT-3 requires substantial computational resources, which have been associated with significant energy and water consumption (George et al., 2023), with reported freshwater usage reaching up to 700,000 liters in certain data centers and potentially higher in other regions (Egbemhenghe et al., 2023). These considerations highlight the broader sustainability challenges associated with large-scale AI development.

Within the defined experimental setup, this study compares the time required for manual literature synthesis with that required for AI-assisted response generation. Reviewing 100 articles to construct the benchmark dataset required approximately 400 hours, whereas ChatGPT-4o generated responses to the five predefined questions in approximately 75 seconds. Under this specific comparison, manual synthesis required approximately 19,200 times more time than response generation using ChatGPT-4o. While this ratio reflects differences in task structure and evaluation scope, it illustrates the potential efficiency gains of AI-assisted literature summarization under controlled benchmarking conditions.

5. Limitations and future research directions

ChatGPT's training and data access are continuously evolving, meaning its responses may vary over time. First, this study does not aim to present definitive answers for FSM modeling but rather evaluates the consistency of ChatGPT-generated outputs through comparison with a literature-derived benchmark dataset. Therefore, the responses should not be interpreted as authoritative references for model development. Second, the five selected questions were designed to represent core themes in ML-based FSM; however, they restrict the scope of evaluation to structured literature-based queries rather than broader real-world modeling applications. In particular, the evaluated questions primarily examine literature retrieval and ranking consistency, whereas context-dependent analytical reasoning—such as recommending models for specific hydrological or data-scarce scenarios—was beyond the scope of the present study. In addition, the out-of-corpus experiment showed that although Chat-FSM demonstrates reliable reasoning when relevant concepts exist within its retrieval corpus, its performance is inherently constrained by the coverage of the indexed literature, and it cannot provide substantive answers for topics that are not represented in the available documents. Future research should expand the range of evaluation scenarios, including additional out-of-corpus benchmarks and context-specific problem settings, to further examine generalization behavior. Finally, prompt engineering plays a significant role, as variations in prompt structure can influence generated outputs. Future studies should systematically investigate how prompt formulation affects content overlap and ranking alignment, thereby clarifying the sensitivity of benchmarking outcomes to prompt design.

6. Conclusions

This study evaluated the agreement of baseline and domain-augmented ChatGPT configurations with a literature-derived benchmark in the context of ML-based Flood Susceptibility Mapping (FSM). By comparing responses from ChatGPT-4o, ChatGPT-4, and a retrieval-augmented GPT-4 configuration

https://doi.org/10.1017/eds.2026.10037 Published online by Cambridge University Press

(Chat-FSM) against a structured synthesis of 100 reviewed articles, measurable differences in content overlap and ranking correlation were identified. Overall, the domain-augmented Chat-FSM configuration exhibited the highest level of content overlap with the benchmark dataset, reflecting improved literature-consistency under the defined evaluation framework. ChatGPT-4 demonstrated comparatively stronger rank correlation in selected cases, whereas ChatGPT-4o generally showed lower agreement levels under the evaluated metrics. Across configurations, feature selection methods displayed limited agreement with benchmark rankings, highlighting variability in how models prioritize methodological components. In addition to agreement analysis, application-level efficiency comparisons indicated substantial time differences between manual literature synthesis and AI-assisted response generation under controlled conditions. However, these efficiency gains should be interpreted within the scope of structured benchmarking rather than as validation of broader analytical reliability.

In conclusion, ChatGPT-based systems demonstrate measurable variation in literature-consistency across configurations, with retrieval-augmented setups yielding higher agreement under controlled benchmarking conditions. However, their performance remains constrained by the coverage of the indexed literature, limiting their ability to address topics absent from the provided documents. Therefore, human expertise remains essential for critical reasoning, methodological design, and context-specific flood susceptibility assessment.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2026.10037.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2026.10037.

Author contribution. Conceptualization: M.J.F., A.P.; Data curation: M.J.F., A.P.; Formal analysis: M.J.F., A.P.; Investigation: M.J.F., A.P.; Methodology: M.J.F., A.P.; Validation: M.J.F., A.P.; Visualization: M.J.F., A.P.; Writing - original draft: M.J.F., A.P.; Writing - review & editing: M.J.F., A.P.; Funding acquisition: M.J.F.; Project administration: M.J.F.; Resources: M.J.F.; Supervision: M.J.F.

Data availability statement. The Chat-FSM GPT, configured with retrieval access to 100 reviewed articles, is available for free access via the following GitHub link: https://chatgpt.com/g/g-j9VsJLbne-chatfsm-1. The details of the benchmark data can be found in Pourzangbar et al. (2025). The responses of ChatGPT models are available in Supplementary Appendix A.

References

Abdullah M, **Madain A and Jararweh Y** (2022) ChatGPT: Fundamentals, applications and social impacts. In *2022 9th International Conference on Social Networks Analysis*, Management and Security (SNAMS), Milan, Italy: IEEEM, pp. 1–8. https://doi.org/10.1109/SNAMS58071.2022.10062688.

Egbemhenghe AU, **Ojeyemi T**, **Iwuozor KO**, **Emenike EC**, **Ogunsanya TI**, **Anidiobi SU and Adeniyi AG** (2023) Revolutionizing water treatment, conservation, and management: Harnessing the power of AI-driven ChatGPT solutions. *Environmental Challenges 13*, 100782. https://doi.org/10.1016/j.envc.2023.100782.

Foroumandi E, **Moradkhani H**, **Sanchez-Vila X**, **Singha K**, **Castelletti A and Destouni G** (2023) ChatGPT in hydrology and earth sciences: Opportunities, prospects, and concerns. *Water Resources Research 59*, e2023WR036288. https://doi.org/10.1029/2023WR036288.

George AS, **George ASH and Martin ASG** (2023) The environmental impact of AI: A case study of water consumption by chat GPT. *Partners Universal International Innovation Journal 1*, 97–104.

Haider S, **Rashid M**, **Tariq MAUR and Nadeem A** (2024) The role of artificial intelligence (AI) and ChatGPT in water resources, including its potential benefits and associated challenges. *Discover Water 4*, 113. https://doi.org/10.1007/s43832-024-00173-y.

Halloran LJS, **Mhanna S and Brunner P** (2023) AI tools such as ChatGPT will disrupt hydrology, too. *Hydrological Processes 37*, e14843. https://doi.org/10.1002/hyp.14843.

Hosseini SH and Pourzangbar A (2026) How well do DeepSeek, ChatGPT, and Gemini respond to water science questions? *Environmental Modelling and Software 196*, 106772. https://doi.org/10.1016/j.envsoft.2025.106772.

Jaccard P (1901) Étude comparative de la distribution florale dans Une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles 37*, 547–579. https://doi.org/10.5169/seals-266450.

Kendall MG (1938) A new measure of rank correlation. *Biometrika 30*, 81–93. https://doi.org/10.2307/2332226.

Montañés A (2024) *ChatGPT-4o Vs ChatGPT-4: ALL the Differences*. Raona. Available at: https://raona.com/en/what-are-the-differences-between-chatgpt-4-and-chatgpt-4o/?utm_source=chatgpt.com (accessed 13 April 2026).

Mulekar MS and Brown CS (2017) Distance and similarity measures. In Alhajj R and Rokne J (eds.), *Encyclopedia of Social Network Analysis and Mining*. New York, NY: Springer, pp. 1–16. https://doi.org/10.1007/978-1-4614-7163-9_141-1.

OpenAI (2023) Models. Available at: https://platform.openai.com/docs/models (accessed 13 April 2026).

Oxford Analytica (2023) *GPT-4 Underlines Mismatch on AI Policy and Innovation*. Emerald Expert Briefings oxan-es.

Pourzangbar A, **Oberle P**, **Kron A and Franca MJ** (2024) On the application of machine learning into flood modeling: Data consideration and modeling algorithm. In Gourbesville P and Caignaert G (eds.), *Advances in Hydroinformatics—SimHydro 2023*. Singapore: Springer Water. Springer. https://doi.org/10.1007/978-981-97-4072-7_11.

Pourzangbar A, **Oberle P**, **Kron A and Franca MJ** (2025) Analysis of the utilization of machine learning to map flood susceptibility. *Journal of Flood Risk Management 18*, e70042. https://doi.org/10.1111/jfr3.70042.

Ray PP (2023) ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003.

Ray SS, **Peddinti PRT**, **Verma RK**, **Puppala H**, **Kim B**, **Singh A and Kwon YN** (2024) Leveraging ChatGPT and bard: What does it convey for water treatment/desalination and harvesting sectors? *Desalination 570*, 117085. https://doi.org/10.1016/j.desal.2023.117085.

Ritchie H and Rosado P (2024) Natural disasters where and from which disasters do people die? What can we do to prevent deaths from natural disasters? [WWW document]. Available at https://ourworldindata.org/natural-disasters (accessed 29 January 2024).

Surameery NMS and Shakor MY (2023) Use chatGPT to solve programming bugs. *International Journal of Information technology and Computer Engineering 31*, 17–22. https://doi.org/10.55529/ijitc.31.17.22.

Xue Z, **Xu C and Xu X** (2023) Application of ChatGPT in natural disaster prevention and reduction. *Natural Hazards Research 3*, 556–562. https://doi.org/10.1016/j.nhres.2023.07.005.

Zhuang Y, **Gong T**, **Fang J**, **Shen D**, **Tang W**, **Lin S**, **Chen X and Zhang Y** (2026) Integrating social media data and machine learning methods for flash flood susceptibility mapping in China. *Journal of Hydrology 664*(38), 134397. https://doi.org/10.1016/j.jhydrol.2025.134397.

Cite this article: Pourzangbar A and Franca MJ (2026). How reliable are retrieval-augmented and standard ChatGPT models to support flood susceptibility mapping?. *Environmental Data Science*, 5: e10. doi:10.1017/eds.2026.10037