Journal of Hydrology X 30 (2026) 100215

Contents lists available at ScienceDirect

Journal of Hydrology X

journal homepage: www.sciencedirect.com/journal/journal-of-hydrology-x

When are AI models ready for deployment? Reassessing Google's global AI flood forecasting system through the lens of responsible modelling

Kailong Li [a,b,*], Saman Razavi [b,c], Holger R. Maier [d], Markus Hrachowitz [e], Ehsan Nabavi [f], Natasha Harvey [g], Khaled Akhtar [b,h], Fisaha Unduche [i]

[a] Desert Research Institute, Las Vegas, Nevada, USA
[b] School of Environment and Sustainability, and Global Institute for Water Security, University of Saskatchewan, Canada
[c] School of Civil and Environmental Engineering, University of New South Wales (UNSW Sydney), Sydney, NSW, Australia
[d] School of Architecture and Civil Engineering, The University of Adelaide, Adelaide, Australia
[e] Faculty of Civil Engineering and Geosciences, Department of Watermanagement, Delft University of Technology, Delft, Netherlands
[f] Responsible Innovation Lab, Australian National Center for Public Awareness of Sciences, Australian National University, Canberra, Australia
[g] Institute for Water Futures, Fenner School of Environment and Society, Australian National University, Canberra, Australia
[h] River Forecast Center, Alberta Environment and Protected Areas, Edmonton, Canada
[i] Hydrologic Forecasting and Water Management, Manitoba Transportation and Infrastructure, Winnipeg, Canada

ARTICLE INFO

Keywords:
AI model
Flood forecasting
Responsible modelling
Model evaluation
Extreme events

ABSTRACT

The development of AI models is increasing at a rapid rate. However, when are they ready to be deployed in real-world operational settings? In this paper, we introduce a framework to support such assessments and apply it to Google's recently released AI-based flood prediction system, which is claimed to achieve "reliability in predicting extreme riverine events" and provide "accurate and timely warnings" that are available "earlier and over larger and more impactful events in ungauged basins". The system has been integrated into an operational early-warning platform producing open, real-time forecasts in more than 80 countries. While this development promises to usher in a new and exciting age in global flood forecasting, the supporting evidence relies heavily on several subjective choices, the implications of which have not been acknowledged or assessed. Here, we evaluate the consequences of these choices on claims of operational deployment readiness across four dimensions: predictive accuracy, forecast timeliness, the characterization of extreme events, and benchmarking against state-of-the-art models. Our assessment reveals that the system's actual predictive accuracy is likely to be substantially lower than reported—particularly for extreme events—raising concerns about *responsible practices across modelling and publicity* in high-stakes applications. The deployment of the Google AI model therefore risks misinforming those who depend on its outputs for evacuation and preparedness decisions, particularly in less-developed countries such as those targeted by the enterprise, given its alarmingly high (>90%) rates of false positives and false negatives. Beyond the immediate operational consequences, if left unaddressed, these outcomes may erode public trust in AI within hydrological sciences. We conclude by calling for greater transparency, accountability, and methodological rigor in the integration of AI into flood forecasting.

1. Introduction

Machine Learning (ML) and Artificial Intelligence (AI) models are unquestionably the 'flavor of the month' in a variety of disciplines, including the earth and environmental sciences (Buchanan, 2024; Maier, 2024). Enthusiasm for these models is fuelled by the widespread perception that they are a panacea for a wide variety of problems, resulting in broad support from government, industry and funding agencies. This has led to an unprecedented increase in research activity in ML and AI that invariably demonstrates the benefits of these models, adding further fuel to the fire (Narayanan and Kapoor, 2024).

Despite the unquestioned potential of ML and AI models, it is vital that we are not swept up in the accompanying excitement at the expense of following the well-trodden, but less exciting, path of applying the scientific method (Buchanan, 2024). This is especially important when these models are used in real-world, operational settings, where there is

* Corresponding author.
E-mail address: kailong.li@dri.edu (K. Li).

https://doi.org/10.1016/j.hydroa.2026.100215
Received 30 October 2025; Received in revised form 9 February 2026; Accepted 23 February 2026
Available online 23 February 2026

a need to (i) adopt good model development practice (Jakeman, 2024; Maier et al., 2023a, b), (ii) evaluate models in a way that ensures they can perform as desired in a robust manner and (iii) test the suitability and reliability of models in their intended operational environment before they can be (iv) deployed in practice (Fig. 1). These steps align with the four stages of assessing the technology readiness levels (TRLs) of ML systems proposed by Lavin et al. (2022), including Research (TRLs 0 to 2), Prototyping (TRLs 3 to 5), Productization (TRLs 6 to 8) and Deployment (TRL 9). It is also critical that the above steps are subject to independent review so that the scientific process can act as "…a guard rail against the frailties of human reason" (Buchanan, 2024).

When developing AI models to support real-world operational decisions, particular attention needs to be paid to the Model Evaluation step. While it is common to evaluate the performance of models, when they are developed for use in real-world settings, evaluation criteria and metrics need to be commensurate with operational requirements (Fig. 1). In addition, the sensitivity of model performance to any assumptions associated with the evaluation criteria and metrics needs to be assessed to ensure the model can be used with confidence in practice (Fig. 1). Finally, the performance and suitability of the developed model need to be compared with an appropriate benchmark, which should consist of state-of-the-art alternative(s) that are currently used in the intended operational setting.

By not assessing the performance of AI models intended to support real-world decision making using all of the steps in Fig. 1, there is a risk of overstating their capabilities (Kapoor et al., 2023; Lavin, 2022; Maier, 2024; Narayanan and Kapoor, 2024, Raj et al. 2022), potentially leading to "algorithmic harm" for affected communities (Lavin et al. 2022, Raj et al. 2022). This harm can be significant, not only in terms of discrediting ML and AI methods unnecessarily, thereby setting the field back many years or even decades, but more importantly also in terms of having a range of negative real-life consequences.

One example of this is the recent paper of Nearing et al. (2024), who introduced Google's global AI system for predicting floods in ungauged basins, in which it is claimed that the system is suitable for issuing operational warnings for extreme floods at multi-day lead times. However, these claims are based on model evaluation criteria and metrics that are calculated using several subjective choices, the impact of which has not been investigated: (i) flood events are defined using return-period thresholds computed separately on the modelled and observed data; (ii) a "hit" is counted when model and observed hydrographs cross their respective thresholds within a ± 2-day window; and (iii) hydrological extremes are emphasized at return periods ≤ 10 years. In addition, benchmarking relies primarily on a single global baseline model, GloFAS, which is an operational global early-warning system intended to provide large-scale, transboundary flood guidance to support preparedness and response, and is often used as a complementary source rather than as a locally calibrated operational forecasting model. In this short paper, we apply the framework in Fig. 1 to provide an independent assessment of the claims of Nearing et at. (2024) that the Google AI model "achieves reliability in predicting extreme riverine events", provides "accurate and timely warnings" and is suitable for real-world deployment across the globe (in fact, according to Nearing et al., the system has already been integrated into an early-warning platform producing open, real-time forecasts in more than 80 countries). This is achieved by re-evaluating the system along four key dimensions used in Nearing et al. (2024): **(i) Accuracy**—the choice of a threshold to classify an event as a flood and how this threshold is calculated, **(ii) Timeliness**—the choice of the acceptable level of forecasting peak timing error, **(iii) Applicability to Extreme Events**—the choice of what constitutes an extreme event, and **(iv) Benchmarking Against the State-of-the-Art**—the choice of alternative model used to evaluate the AI model's performance.

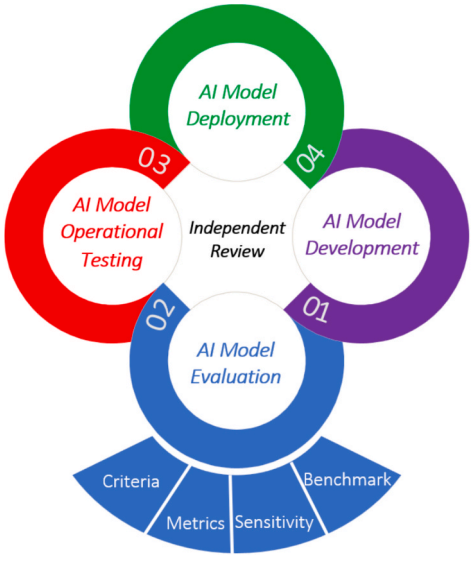

Fig. 1. Proposed framework for developing AI models for use in real-world operational settings.

2. Assessment of suitability of the google AI model for operational deployment

Our assessment proceeds in three steps following the framework in Fig. 1 to evaluate whether the results presented in Nearing et al. (2024) support the claim that the model is suitable for deployment in real-world operational settings. The first two steps assess whether the results obtained using the selected model evaluation criteria and metrics support the statements that the model provides "accurate and timely warnings" and "achieves reliability in predicting extreme riverine events" by analyzing their sensitivity to subjective modelling choices. The third step provides a qualitative assessment of the suitability of the benchmark model.

In the first step, we revisit the model evaluation performances reported by Nearing et al. (2024) for all of the 5,065 gauged basins they considered in terms of the choice of (i) and (ii) above by anchoring flood thresholds to actual real-world observations and requiring same-day concurrence for nowcasts. To do so, we modify the original evaluation code by using observed, rather than modelled streamflow, as done by Nearing et al. (2024), to determine actually observed threshold exceedance. In addition, we have changed the timing tolerance for the occurrence of flood peaks from ± 2 days (window = 2) used by Nearing et al. (2024) to 0 days (same-day concurrence). Next, we assess the model's applicability to extreme events—choice (iii)—by pooling rare (≥50- and 100-year) events across basins and evaluating performance on the combined event set, since metrics for very rare floods are otherwise unstable and easily distorted by the low frequency of positives. Finally, we discuss whether the choice of relying on a single "global" benchmark – choice (iv) – may risk obscuring local hydroclimatic and operational nuances, leading to misleading performance claims.

2.1. Accuracy and timeliness

The subjective choices made in relation to (i) and (ii) can have a significant impact on the perceived performance of the developed model. Nearing et al. (2024) chose to define a flood prediction for a given return period as a 'true positive' "…if the modelled hydrograph and the observed hydrograph both cross their respective return period threshold flow values within two days of each other", with these threshold-flow values "…calculated for observed time series and for modelled time series separately".

The fundamental issues with these definitions lie in the choice of (i) the reliance on modelled time series in gauged basins instead of actual

observations to determine the threshold indicating flood occurrence at a specific point and (ii) deeming two-day lags (instead of zero-day lag) in prediction as acceptable, even in nowcasting. Fig. 2a illustrates the approach by Nearing et al. (2024) in interpreting a 'true positive' for a 10-year flood during a hypothetical flood event.

With respect to choice (i), while computing thresholds separately for modelled and observed flows is an approach sometimes adopted to account for systematic magnitude bias (Hirpa et al., 2016; Thielen et al., 2009), such approaches are not intended to substitute for observation-anchored assessments when retrospectively evaluating real-world flood occurrence and operational warning performance using historical data (Oh and Bartos, 2025). One could argue in support of the Nearing assessment approach that performance assessments based on historical observations are not possible in ungauged basins. However, this argument overlooks an important point: building confidence for applying a model to ungauged basins inevitably requires first evaluating its performance in basins where observed real-world data are available for comparison. If the model is inconsistent with these observations and therefore does not perform well in gauged basins, it is unlikely to perform well in ungauged ones.

Delving deeper into this choice reveals that it can lead to a fundamental misassessment of model performance. Consider for instance the hypothetical event shown in Fig. 2b, which the Google AI model would classify as a 'true positive'. However, the correct assessment should be a 'false negative' because the observed event exceeded the 10-year flood threshold, while the modelled event did not. More importantly, in cases where model performance is *nearly perfect* in replicating observations during events such as those shown in Fig. 2c-d, the model's interpretation algorithm would incorrectly label them as 'false positive' and 'false negative', respectively. In reality, the model *accurately* predicts a 'true negative' and a 'true positive' in Fig. 2c and 2d, respectively.

Fig. 3a illustrates the impact of choice (i) on model performance assessment in isolation, while Fig. 3b demonstrates its combined impact with choice (ii)—a comparison with the results of Nearing et al. (2024) with new results obtained when these subjective choices are removed. This comparison reveals a significant overestimation of forecast accuracy by both the AI model and the benchmark model (GloFAS), with performance dropping notably when benchmark forecasts are replaced by real observations. Here, the event-detection skill was evaluated using F1-score (equation (1), which is the harmonic mean of precision (the fraction of correctly predicted events over all *predicted* events; see equation (2) and recall (the fraction of correctly predicted events over all *actual* events; see equation (3), so low F1 values indicate both frequent misses and frequent false alarms.

$$F1 = \frac{2 \bullet Precision \bullet Recall}{Precision + Recall} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where TP means true positives: predicted positive and actually positive. FP means false positives: predicted positive but actually negative. FN means false negatives: predicted negative but actually positive. Our evaluation of choice (i) in Fig. 3a suggests that switching from model-based to observation-based thresholds generally reduces the F1-score, with one exception at the 1-year return period, where the observation-based threshold for the Google AI model is lower than the model-based threshold.

We do not argue that the use of modeled thresholds is illegitimate in all contexts. Rather, we show that for the specific purpose of evaluating claims about real-world flood warning performance, this choice considerably alters event classification in ways that obscures magnitude and timing errors relative to observations. Sensitivity testing against observation-anchored thresholds is therefore not optional but necessary to support robust interpretation of categorical metrics in high-stakes, operational contexts. This distinction is particularly important given that our results demonstrate poor performance in gauged river basins where observational data are available for validation. In hydrological

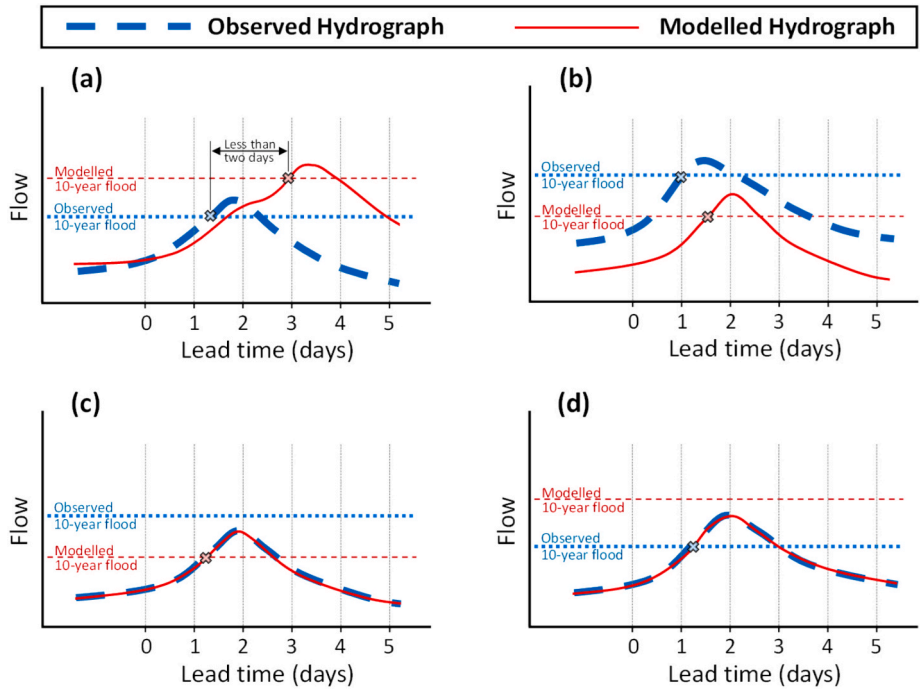

Fig. 2. Hypothetical examples of flood events at a gauge: (a) shows how a 'true positive' is defined in Nearing et al. (2024). Nearing et al. would assess (b) as 'true positive' while it is actually a 'false negative' when compared with observed data. They would assess (c) as a 'false positive' while it is actually a 'true negative' and (d) as a 'false negative' while it is actually a 'true positive.' Note that in the case of (c) and (d), the model predictions are nearly perfectly accurate/true (i.e., observed and modeled data match nearly perfectly).

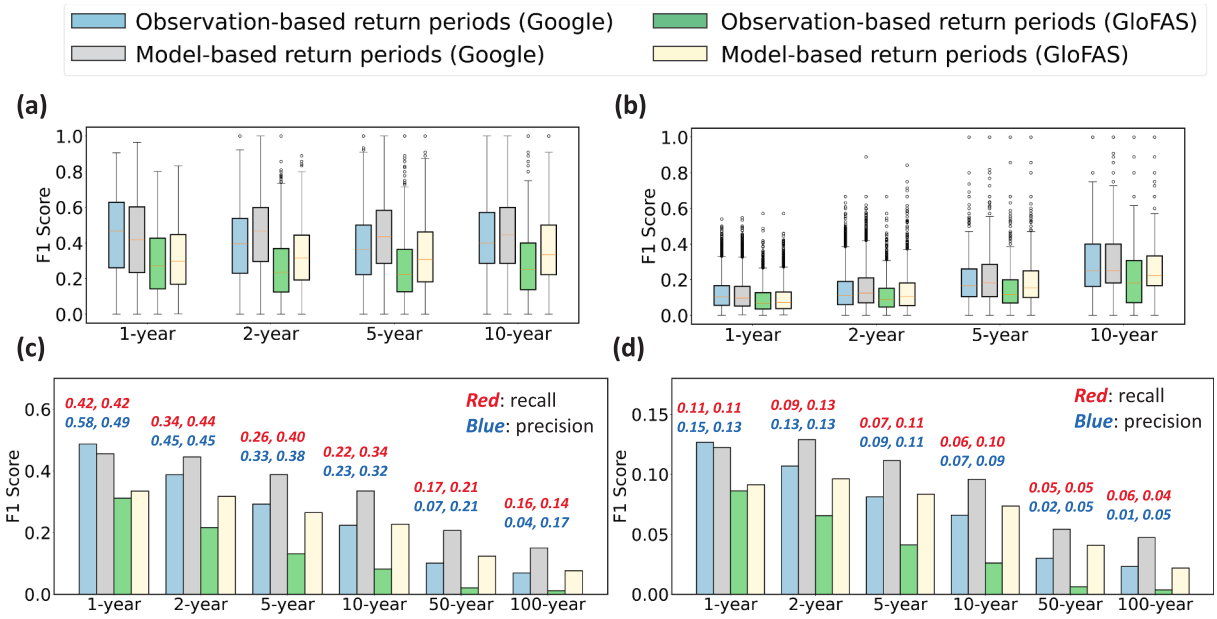

Fig. 3. (a) Performance of models in nowcast (i.e., zero lead time, corresponding to estimating current river flow conditions in response to a storm that has already occurred) when evaluated against model-based return periods, as reported by Nearing et al., compared to their performance when evaluated against observation-based return periods, where a prediction within two days of the actual flood event is considered accurate. (b) Reassessment of the models' performance when a 'true positive' is redefined to require the predicted and actual floods to occur on the same day. (c) Average performance of models globally using a two-day window for event identification, combining data from all basins for calculating metrics rather than conducting a basin-by-basin analysis. The red text in the bottom row represents recall values, while the blue text represents precision values for the first two bars associated with the Google AI model. (d) Average performance of models globally using a same-day criterion for event identification. All results shown correspond to the nowcast setting; forecasts with lead times of one or more days would necessarily exhibit lower predictive skill due to the additional uncertainty introduced with increasing lead time. The results presented are based on the codes and modelling results provided by Nearing et al. (2024), with minor discrepancies due to updates in the historical observation database [https://grdc.bafg.de/] since their original work. The numerical statistics underlying this plot are provided in Table S1 and S2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

modelling, the only rigorous, evidence-based basis for expecting reliable performance in ungauged basins is first to demonstrate reliable performance in gauged basins with observed ground-truth data. When a model cannot deliver meaningful accuracy in locations where observations exist, there is no empirical foundation to support claims of superior performance in ungauged regions. Distinguishing between these contexts is therefore essential for responsible communication and for assessing the suitability of AI-based flood forecasts in high-stakes operational settings.

Regarding (ii), while Nearing et al. (2024) do not provide a justification for this choice, the implied assumption is that errors of up to two days in estimating the arrival time of floods are adequate for real-world flood warnings. However, this is unlikely to be true for many basins worldwide. As shown in Fig. 3b, the models perform very poorly when we remove choice (ii) by redefining a 'true positive' to mean the predicted and actual floods must occur on the same day. For instance, the median F1-score for 5-year floods drops from ~ 0.42 (see Fig. 3a) to under 0.20 (see Fig. 3b) for the Google AI model, even if choice (i) is considered acceptable. This is a crucial consideration because accurate flood timing is critical for effective emergency response, not just the predicted severity of the flood.

For clarity, and to avoid potential misinterpretation by non-specialist readers, we emphasize that a ± 2-day error in flood-timing accuracy should not be conflated with forecast lead time, noting that a 2-day lead time is in fact desirable for enabling actions such as evacuation and resource allocation. Lead time refers to how much advance notice is provided before a flood event occurs, enabling preparedness and response, whereas timing accuracy concerns how precisely the timing of the flood peak is predicted. Our criticism is not related to lead time. Rather, it concerns the acceptance of substantial errors in predicting the timing of flood peaks. We emphasize that the analysis in this paper focuses on the nowcast case (i.e., zero lead time). Under this scenario, a 2-day timing error can be particularly problematic—for example, predicting that a flood is occurring today (current time step) when the event in fact occurred two days earlier leaves no opportunity for preparedness or response, undermining the rationale if such conflation is used in ways that could mislead end users into accepting these timing tolerances.

2.2. *Applicability to extreme events*

The claim that the Google AI model can reliably predict extreme events is not well supported by the results presented, as they are for return periods of 10 years or less—choice (iii)—which can hardly be considered extreme. When we analyzed the results for actual extreme events with return periods of 50 and 100 years, as shown in Fig. 3c and d, we found a significant reduction in the AI model's prediction accuracy, with F1-scores of less than 0.03 and 0.02 for 50- and 100-year floods, respectively, assessed globally across all basins.

It is important to note that training and evaluating performance of data-driven models for extreme events is challenging due to the inherent infrequency of such events. For example, in the dataset used by Nearing et al. (2024), only about 260 out of the 5,065 basins contain 100-year flood events. Therefore, our new results are based on a global evaluation across all basins instead of a basin-by-basin analysis, aggregating all identified events for each return period and calculating the metrics based on the combined data. Even under a global aggregation, the model exhibits very low recall with fewer than 6%, 5%, and 6% of actual 10-, 50-, and 100-year flood events correctly predicted under nowcast conditions by the AI model. Similarly, precision stands at just 7%, 2%, and 1%. Moreover, these figures underscore substantial errors, including 'false negatives,' where missing a high-flow event can lead to disaster, and 'false positives,' where frequent false alarms erode public trust, causing people to ignore warnings and remain unprepared when real danger strikes. False negative rates even in nowcast account for 94%,

95%, and 94% of prediction cases for 10-, 50-, and 100-year flood events, respectively, while false positive rates occur in 93%, 98%, and 99% of cases. Even if choices (i) and (ii) were deemed acceptable, the rates of false negatives (positives) in nowcast would still be alarmingly high: 66% (68%), 79% (79%), and 86% (83%) for predictions of 10-, 50-, and 100-year flood events, respectively.

One could argue, as reflected in the framing of Nearing et al., that higher return periods need not be considered because Early Action Protocols (EAPs) are triggered at relatively lower flood thresholds, suggesting that distinguishing between, for example, a 20-year and a 50-year event is of limited operational relevance once a warning has been issued. However, in practice, higher return periods are explicitly used in many operational contexts. For instance, 20-year thresholds are adopted in EAPs in regions such as Zambia (Anticipation Hub, 2023) and parts of Europe (Alfieri et al., 2018). In addition, this argument conflates the *trigger threshold* of an early-warning system with the *magnitude of the event being forecasted*. For example, for extreme or catastrophic events, accurate magnitude estimation is critical for prioritizing evacuations, allocating emergency resources, and protecting critical infrastructure. Low predictive skill at these magnitudes therefore cannot be dismissed solely on the basis that a lower trigger threshold may already have been exceeded.

2.3. Benchmarking against the state-of-the-art

The subjective choice (iv) of the benchmark model in Nearing et al. (2024) can also have a considerable impact on the perceived performance of the Google AI model. GloFAS, against which the Google AI model is compared, is the only operational, physics-based, openly available "global" flood forecasting system. Our results show that, under both thresholding approaches, the Google AI model continues to outperform GloFAS in relative terms. However, GloFAS is not intended to represent the available forecasting capability for real-world flood warning in many regions. In operational practice, flood emergency response is often supported by locally or regionally calibrated models that are tailored to basin-specific hydrological processes, human regulation, and local climatic conditions, and which frequently outperform such global models (Dasgupta et al., 2025; Fleischmann et al., 2019). Consequently, relying exclusively on a "global" benchmark risks missing crucial local insights, losing context, and skewing model assessments. Nearing et al. (2024) appear to have exploited the poor performance of GloFAS in many parts of the world to argue that their model is reliable in those same regions. However, outperforming a weak benchmark does not in itself establish a model's practical usefulness. As we demonstrate in this paper, the Google AI model's performance can be extremely poor — with precision and recall rates as low as ~ 1% and ~ 6%, respectively, for 100-year return period events, meaning that, for such extreme floods with a magnitude that can, on average, be expected to occur every 100 years, only 1 out of every 100 flood warnings issued by the model are correct.

Consequently, while benchmarking against the GloFAS model is an important, interesting and valid exercise from a scientific and research point of view, it is insufficient to support claims that the global Google AI is ready for deployment in real-world operational settings. Such claims need to be supported by a multi-tiered benchmarking strategy, combining global, continental, and regionally calibrated models where available. There are emerging alternatives at continental and large-regional scales that may provide more informative points of comparison. For example, systems such as GEOGLOWS and Hillslope Link Model (HLM), along with other continental-scale hydrological forecasting frameworks, have demonstrated strong skill in specific regions and represent an intermediate class between global models and fully local, operational forecasting systems (Michalek et al., 2024; Qiao et al., 2019; Sikder et al., 2019; Souffront Alcantara et al., 2019). Where feasible, comparing AI-based forecasts with such systems would help determine whether improvements over GloFAS translate into gains relative to the models that are actually used in practice. This would offer a more robust and decision-relevant assessment of whether AI-based flood forecasting systems meaningfully advance flood warning and risk management practice and are ready and suitable for operational use.

3. A responsible modelling lens

Our analysis does not reject the promise of AI-based flood forecasting, but it highlights how the global scaling of such systems also amplifies the epistemic and ethical stakes of modelling choices. The rapid deployment of AI models, without following the steps outlined in Fig. 1, can overlook critical aspects of model functionality and essential technology transition stages, creating the risk of "algorithmic harm" for affected communities (Lavin et al., 2022; Raji et al., 2022), as mentioned in the Introduction. Such harm can be substantial given the catastrophic impacts of floods, which are exacerbated by climate change, land-use change, and population growth (Hamers et al., 2024; Razavi et al., 2020), particularly in less-developed countries such as those targeted by the Google model (Hamers et al., 2024; Razavi et al., 2020).

Errors in model outputs can have profound consequences. For users, this may mean that faulty flood warnings negatively affect their everyday lives and livelihoods, and potentially undermine their trust in science, or exacerbate existing distrust. For developers, such errors can damage credibility and lead to long-term setbacks for the adoption and advancement of technology itself. In this context, "responsible modelling" offers a valuable lens for understanding these challenges and identifying ways forward (Nabavi, 2022). Research in this area emphasizes that hydrological models are not merely neutral artefacts (Saltelli and Di Fiore, 2023), but act as "intervention technologies", with activities and responsibilities that are socio-technical in nature (Nabavi, 2025). Applying a responsible modelling perspective to AI-based flood forecasting encourages us to carefully examine the assumptions, choices, and boundary judgments that underpin data and method selections. These choices influence not only the performance and reliability of the model but also how responsibility is conceptualized and practically distributed between developers and end-users (Nabavi et al., 2024). At its core, responsible modelling seeks to ensure that the modelling process remains attentive to these considerations and, ultimately, that models serve society (Saltelli et al., 2020).

The modelling community increasingly frames "good modelling practice" as a professional norm and research priority, promoting transparency, reproducibility, and documentation to strengthen the quality, credibility, and societal legitimacy of models (Jakeman et al., 2024; Maier et al., 2024; Maier et al., 2023). Critical social science research also emphasises the need for reflexive engagement with modelling as a *situated* practice, shaped by social and institutional contexts. It recognises that models have power and can shape futures of how, and for whom, water is governed, and with which implications for justice and sustainability (Alba et al., 2025; Klein et al., 2024; Melsen, 2022).

Thus, we do not argue against the potential of AI-based models for hydrological forecasting, as numerous studies have already shown the superiority of AI-based models over traditional hydrological models in terms of accuracy and precision in nowcasting and short-lead time forecasting (Arsenault et al., 2023; Coulibaly et al., 2000; Kratzert et al., 2019; Rezaeianzadeh et al., 2013). However, extending these successes to global flood forecasting, as well as to claims of operational readiness, requires greater caution. Specifically, this paper is not intended to identify flaws in the modelling approach itself (i.e., conventional Long Short-Term Memory (LSTM) network modelling); rather, it examines the subjective choices made in model assessment that may have led to inflated performance metrics and a premature push toward large-scale deployment.

Relevant issues to consider have been identified in several studies that have examined the performance of LSTMs in rainfall–runoff modelling. For example, Bayati et al. (2026) identified deficiencies in

the functional realism of LSTM models using catchment hydrology principles, demonstrating how such limitations can compromise the robustness of LSTM-based forecasts under hydrological extremes and both short- and long-term climatic shifts. Along similar lines, Gupta (2025) argues that evaluations of geo-scientific AI needs to move beyond a narrow focus on accuracy and precision—which corresponds to first-order generalization (Ji et al., 2025)—and instead incorporate explicit out-of-distribution testing and higher-order generalization abilities, which Gupta (2025) conceptualizes as extending up to fourth order.

Furthermore, significant challenges remain in improving data quality, model structures, and our overall understanding of hydrological processes (Rudlang et al., 2025; Thyer et al., 2024). We therefore suggest that future innovation should focus on developing the next generation of models that harness AI's unique features, such as flexibility and scalability across diverse data sizes and types, while staying true to local knowledge and nuances in both natural processes and human-driven elements (e.g., reservoirs and diversions) and principles of model explainability and falsifiability (Razavi et al., 2022).

Taken together, these considerations frame a responsible modelling lens through which the results of our analysis should be read. From this perspective, it is important to place our assessment on the scholarly record—not only to advance open scientific discourse and correct the record, but also to help safeguard societies that rely on accurate and trustworthy flood forecasts, which is an important component of the assessment of the real-world operational readiness of AI models (see Fig. 1). The issues at stake extend beyond a purely technical discussion; they relate to standards of research and publication in an era where AI-driven models increasingly inform public policy and safety–critical decisions. We recognize that such discussions may involve conflicts of interest, including corporate and editorial considerations, yet transparency and independence remain central to scientific credibility. Ultimately, our aim is to encourage a culture of responsible modelling, innovation, and publicity—one in which rigorous evidence, open dialogue, and scientific integrity guide technological advancement.

Code availability

The code to reproduce the results are available via Zenodo at https://zenodo.org/records/15023022.

Contributions:

SR initiated the idea and led the discussions among the authorship team. KL obtained all the data and codes, conducted the numerical analysis, and generated the results. All authors assessed the results and contributed to the discussion. SR, HRM, and KL wrote the manuscript with contributions from all other co-authors. HRM created Fig. 1. SR created Fig. 2. KL and SR created Fig. 3.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Nearing et al. for providing feedback on an earlier version of this manuscript, which helped us better understand their perspective and improve the clarity and balance of the final paper. We also thank the editor and reviewers for their critical and insightful comments, which substantially strengthened the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.hydroa.2026.100215.

Data availability

I have shared the link to my data

Google AI model and GloFAS model simulations are available from Nearing et al. (2024). Streamflow data is available from Global River Data Center (https://grdc.bafg.de/).

References

Alba, R., et al., 2025. Situating hydrological modeling: a proposal for engaging with the power of models. Wiley Interdiscip. Rev. Water 12 (4), e70030.
Alfieri, L., et al., 2018. A global network for operational flood risk reduction. Environ. Sci. Policy 84, 149–158.
Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. Hydrol. Earth Syst. Sci. 27 (1), 139–157.
Bayati, A., Ameli, A.A., Razavi, S., 2026. Evaluating the functional realism of deep learning rainfall-runoff models using catchment hydrology principles. Water Resour. Res. 62 (1), e2025WR040076.
Buchanan, M., 2024. Don't flock to faulty AI fashion. Nature Physics 20, 1220. https://doi.org/10.1038/s41567-024-02604-y.
Coulibaly, P., Anctil, F., Bobée, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. J. Hydrol. 230 (3–4), 244–257.
Dasgupta, A., et al., 2025. Connecting hydrological modelling and forecasting from global to local scales: perspectives from an international joint virtual workshop. J. Flood Risk Manage. 18 (1), e12880.
Fleischmann, A., Paiva, R., Collischonn, W., 2019. Can regional to continental river hydrodynamic models be locally relevant? a cross-scale comparison. Journal of Hydrology X 3, 100027.
Gupta, H.V., 2025. On generalization, language, interpretability and the future of geo-scientific machine learning. Environ. Model. Software.
Hamers, E.M., Maier, H.R., Zecchin, A.C., van Delden, H., 2024. Framework for considering the interactions between climate change, socio-economic development and land use planning in the assessment of future flood risk. Environ. Model. Software 171, 105886.
Hirpa, F.A., et al., 2016. The effect of reference climatology on global flood forecasting. J. Hydrometeorol. 17 (4), 1131–1145.
Anticipation Hub, 2023. Zambia activates its Early Action Protocol for Floods. Anticipation Hub, https://www.anticipation-hub.org/news/zambia-activates-its-early-action-protocol-for-floods.
Jakeman, A., et al., 2024. Towards normalizing good practice across the whole modeling cycle: its instrumentation and future research topics. Socio-Environmental Systems Modelling 6, 18755.
Ji, Y., et al., 2025. An R package to partition observation data used for model development and evaluation to achieve model generalizability. Environ. Model. Software 183, 106238.
Kapoor, S., Cantrell, E., Peng, K., Pham, T.H., Bail, C.A., Gunderson, O.E., Hofman, J.M., Hullman, J., Lones, M.A., Malik, M.M., Nanayakkara, P., Poldrack, R.A., Raji, I.D., Robers, M., Salganik, M.J., Serra-Garcia, M., Stewart, B.M., Vandewiele, G., Narayanan, A., 2023. REFORMS: Reporting Standards for Machine Learning Based Science. https://doi.org/10.48550/arXiv.2308.07832 arXiv:2308.07832.
Klein, A., et al., 2024. From situated knowledges to situated modelling: a relational framework for simulation modelling. Ecosyst. People 20 (1), 2361706.
Kratzert, F., et al., 2019. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55 (12), 11344–11354.
Lavin, A., et al., 2022. Technology readiness levels for machine learning systems. Nat. Commun. 13 (1), 6039.
Maier, H.R., Galelli, S., Razavi, S., Castelletti, A., Rizzoli, A., Athanasiadis, I.A., Sànchez-Marrè, M., Acutis, M., Wu, W., Humphrey, G.B., 2023a. Exploding the myths: an introduction to artificial neural networks for prediction and forecasting. Environmental Modelling and Software 167, 105776. https://doi.org/10.1016/j.envsoft.2023.105776.
Maier, H.R., Zheng, F., Gupta, H., Chen, J., Mai, J., Savic, D., Loritz, R., Wu, W., Guo, D., Bennett, A., Jakeman, A., Razavi, S., Zhao, J., 2023b. On how data are partitioned in model development and evaluation: confronting the elephant in the room to enhance model generalization. Environmental Modelling and Software 167, 105779. https://doi.org/10.1016/j.envsoft.2023.105779.
Maier, H., et al., 2024. How much X is in XAI: responsible use of "Explainable" artificial intelligence in hydrology and water resources. Journal of Hydrology X 100185.
Melsen, L.A., 2022. It takes a village to run a model—the social practices of hydrological modeling. Water Resour. Res. 58 (2), e2021WR030600.
Michalek, A.T., Quintero, F., Villarini, G., 2024. Contiguous United States hydrologic modeling using the hillslope link model TETIS. JAWRA Journal of the American Water Resources Association 60 (6), 1058–1079.
Nabavi, E., 2022. Computing and modeling after COVID-19: more responsible, less technical. IEEE Trans. Technol. Soc. 3 (4), 252–261.
Nabavi, E., 2025. Modelling as intervention technology: Science, politics, and water conflicts. Water Altern. 18 (2), 330–354.
Nabavi, E., Nicholls, R., Roussos, G., 2024. Locating responsibility in the future of human–AI interactions. IEEE Trans. Technol. Soc. 5 (1), 58–60.

Narayanan, A., Kapoor, S., 2024. Scientists should use AI as a tool, not an oracle: how AI hype leads to flawed research that fuels more hype. AI Snake Oil, June. https://www.aisnakeoil.com/p/scientists-should-use-ai-as-a-tool.
Nearing, G., et al., 2024. Global prediction of extreme floods in ungauged watersheds. Nature 627 (8004), 559–563.
Oh, J., Bartos, M., 2025. Flood early warning system with data assimilation enables site-level forecasting of bridge impacts. NPJ Nat. Hazards 2 (1), 64.
Qiao, X., et al., 2019. A systems approach to routing global gridded runoff through local high-resolution stream networks for flood early warning systems. Environ. Model. Software 120, 104501.
Raji, I.D., Kumar, I.E., Horowitz, A., Selbst, A., 2022. The fallacy of AI functionality, Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 959-972.
Razavi, S., Gober, P., Maier, H.R., Brouwer, R., Wheater, H., 2020. Anthropocene flooding: challenges for science and society. Hydrol. Process. 34 (8), 1996–2000.
Razavi, S., et al., 2022. Coevolution of machine learning and process-based modelling to revolutionize Earth and environmental sciences: a perspective. Hydrol. Process. 36 (6), e14596.
Rezaeianzadeh, M., et al., 2013. Assessment of a conceptual hydrological model and artificial neural networks for daily outflows forecasting. Int. J. Environ. Sci. Technol. 10, 1181–1192.
Rudlang, J.M., do Nascimento, T.V., van der Ent, R., Fenicia, F., Hrachowitz, M., 2025. Climate and landscape jointly control Europe's hydrology. Egusphere 2025, 1–42.
Saltelli, A., Di Fiore, M., 2023. The politics of modelling: numbers between science and policy. Oxford University Press.
Saltelli, A., et al., 2020. Five ways to ensure that models serve society: a manifesto. Nat. Publ. Group.
Sikder, M.S., et al., 2019. Evaluation of available global runoff datasets through a river model in support of transboundary water management in South and Southeast Asia. Front. Environ. Sci. 7, 171.
Souffront Alcantara, M.A., et al., 2019. Hydrologic modeling as a service (HMaaS): a new approach to address hydroinformatic challenges in developing countries. Front. Environ. Sci. 7, 158.
Thielen, J., Bartholmes, J., Ramos, M.-H., De Roo, A., 2009. The European flood alert system–part 1: concept and development. Hydrol. Earth Syst. Sci. 13 (2), 125–140.
Thyer, M., et al., 2024. Virtual Hydrological Laboratories: developing the next generation of conceptual models to support decision making under change. Water Resour. Res. 60 (4), e2022WR034234.