Overall comments: This manuscript presents a comparative evaluation of the Google Global Flood Forecasting system Version 2 (v2), Google Global Flood Forecasting system Version 1 (v1), and related benchmark models across 1223 basins worldwide. Using NSE as the main evaluation metric, the authors show that, relative to the v1 0-day lead time forecast, the v2 system extends the reliable predictive horizon by 6 days in gauged basins and by 1 day in ungauged basins. The manuscript is also accompanied by an open-source codebase that enables training of both the v1 and v2 models using the open-source Caravan dataset.

Specific comments:

1. The abstract is too concise. It does not sufficiently reflect the methodological innovations of the v2 model, and it also lacks adequate research background.
2. In the title, abstract, and elsewhere, the authors emphasize that this is a flood forecasting system. However, based on the definition of the model target variable and the official open-source code, the main evaluation in this paper appears to focus on daily streamflow simulation using metrics such as NSE. Daily streamflow simulation is an important component of flood forecasting, but in my view it is not the whole task. For example, when reading the abstract, I would expect hourly-scale flood forecasting, or that the forecasting system would include warning information, water level, and even inundation information.

Therefore, I have a question: is the object of this study the "core runoff/streamflow forecasting component within a flood forecasting system", or the "complete flood forecasting system itself"?

Previous work by Nevo et al. (2022) explicitly stated that Google's operational flood forecasting system consists of four subsystems: data validation, stage forecasting, inundation modeling, and alert distribution. In contrast, the target variable in the present paper is daily streamflow, and the main metrics are NSE and KGE. Therefore, more precisely, this paper evaluates the hydrological prediction core of a flood forecasting system, rather than the complete flood forecasting system itself.

If the term "flood forecasting system" is used, I suggest that the authors at least add event-level flood metrics or provide a clearer discussion in the Supplement. If the paper is only discussing the core support for a "flood forecasting system", then the title and related wording should not directly state "global flood forecasting system". The authors should make the terminology consistent throughout the paper, or address this issue in the outlook or discussion.

3. Following the previous comment, I think the authors should provide additional supporting metrics and justification related to the terms "operational system" and "operational flood". This paper provides substantial support for future flood forecasting systems and operational forecasting, but most of the evaluation focuses on model performance for daily streamflow prediction, rather than improvement of a global flood system.

Using "daily streamflow prediction metrics" to support claims about an "operational system" lacks solid support and evidence. Actual flood forecasting, especially for small and medium-sized basins, usually requires hourly-scale results, whereas daily-scale forecasts are not sufficiently fine. In operational flood control and emergency response, people often care about metrics such as peak flow and time of peak flow, rather than NSE, a goodness-of-fit metric that strongly favors long-term average behavior. We already know the coherent limitations of NSE, KGE, and similar metrics. The authors should discuss why daily streamflow prediction metrics such as NSE are sufficient to support an operational system in the context of a global flood forecasting system, or alternatively define the current system's limitations more clearly. Otherwise, the declaration in the paper may appear overstated.

4. The Introduction is very concise. However, as a research paper, it should clearly present the key research gap, the necessity of the study, and the need to upgrade existing technologies or solutions. For example, the main focus of this paper is the v2 system, but the current first paragraph mainly discusses the status of machine learning in streamflow simulation, model development, interpretability, and uncertainty quantification. These topics are only briefly mentioned, without specific literature citations, which makes the Introduction too brief.

In addition, the second paragraph directly turns to "using operational machine learning hydrology models for global-scale riverine flood forecasting", but it does not discuss the innovation of the v2 system or the improvements over v1. Since a substantial part of the v2 improvement comes from the introduction of GraphCast, I suggest that the authors at least add discussion of how meteorological data can improve flood forecasting models.

5. In Section 2.1.2, the authors state that HRES and GraphCast forecast archives begin in approximately 2012 and 2016, respectively. To use the full historical streamflow record from 1980 to 2024, the authors substitute ERA5-Land reanalysis data for HRES/GraphCast forecast inputs in earlier years when such forecasts are unavailable, and assume that these reanalysis data serve as an "effective proxy" for the forecast inputs. The justification given in the paper is that "HRES shares the same underlying physical model as ERA5, and GraphCast is trained on ERA5". However, this assumption is not supported or demonstrated. I suggest that the authors conduct a comparison for years when HRES/GraphCast and ERA5-Land are both available, and show whether their precipitation, temperature, and other distributions are similar. Alternatively, the authors could compare whether NSE/KGE differs substantially when ERA5-Land is used as input versus when the actual GraphCast forecast inputs are used.

6. In Section 3, the authors state that "For the ungauged setting, the v1 system used random k-fold (k=10) cross-validation, whereas the v2 system used a single holdout test set". I suggest that the authors explain why different spatial evaluation protocols were used for v1 and v2, how the v2 holdout basins were selected, and why a single spatial split is sufficient to evaluate ungauged generalization. This clarification is important because the spatial split strategy may affect the comparability of ungauged performance between v1 and v2.

7. In Section 4.2, the authors state that "Figure 6 disaggregates the improvements provided by the ME-LSTM architecture and expanded training data from the predictive skill injected by the GraphCast meteorological forcings". However, the authors also state that "Blue boxes represent the Delta NSE gained by transitioning

from the v1 to v2 model architecture and expanded Caravan training data. Green boxes represent the additional Delta NSE gained by incorporating GraphCast". The paper has already demonstrated the contribution of GraphCast, but it has not separated the contribution of the ME-LSTM architecture change from the contribution of the expanded Caravan training data. I think additional experiments and evidence could be added in the Supplement.

8. The paper states that "We take the mean of the predicted distribution to be the deterministic model prediction that we evaluate in this Paper". Since both v1 and v2 produce probabilistic forecasts using "a countable mixture of asymmetric Laplacians (CMAL) distribution", I think it is necessary to explain why only the mean of the predicted distribution is evaluated. Deterministic NSE/KGE metrics can indicate predictive performance, but they cannot evaluate the quality of probabilistic forecasts. For flood forecasting, probabilistic forecast results themselves are important. I suggest adding metrics such as prediction interval coverage.

9. In Section 4.5 and Figure 11, the authors state that the v2 system "yields higher absolute performance globally", but is also "proportionally more sensitive to the absence of local streamflow data for training". Specifically, at a 0-day lead time, the v2 system without GraphCast has a median gauged NSE of 0.78 and an absolute median penalty of 0.07, meaning the difference between gauged NSE and ungauged NSE, corresponding to a 10.6% relative decrease. With GraphCast, the v2 system has a median gauged NSE of 0.83, but the absolute median penalty increases to 0.12, corresponding to a 19.8% relative decrease. I suggest that the authors explicitly state in the abstract or conclusion that GraphCast improves overall absolute performance, but also increases the penalty between the gauged and ungauged settings. This does not mean that GraphCast brings the same magnitude of improvement under both gauged and ungauged settings. The paper should not only state that the lead time is extended by one day in ungauged basins.

10.In Section 4.2, the authors state that "GraphCast forcings improve correlation but lower forecast variance", and further mention possible "spatial and temporal smoothing" and "underprediction of variance" at longer lead times. However, this

paper studies flood forecasting, while the NSE and KGE metrics used in the paper cannot accurately demonstrate high-flow prediction performance. Flood peaks and high flows are very important for flood forecasting. From a mathematical perspective, lower forecast variance caused by GraphCast may indicate that the model underestimates high flows. Therefore, I suggest that the authors supplement the analysis with relevant flood peak or high-flow simulation results or metrics.

11. At the beginning of the paper, the authors state that v2 improves upon v1. The paper has explained that v1 is an ED-LSTM, while v2 introduces ME-LSTM, expanded Caravan training data, and new meteorological inputs such as GraphCast. The authors also acknowledge that the performance difference between v1 and v2 is "a compound effect", and that the two systems use different spatial split strategies in the ungauged evaluation. However, the exact differences between v1 and v2 are not clearly compared. I think the paper lacks a table, namely a v1-v2 comparison table, listing differences in model architecture, training data sources, dynamic meteorological inputs, whether GraphCast is included, temporal and spatial splitting strategies, and related aspects.

12. The paper uses the terms Google Global Flood Forecasting system and operational system in multiple places, but the methods, evaluation metrics, and open-source code mainly correspond to the runoff/streamflow model forecasting component of the Google FloodHub flood forecasting platform, rather than an end-to-end operational flood warning system. Section 2.1 clearly states that the model training target is daily streamflow at the basin outlet. The open-source code also mainly provides model training, evaluation, and related workflows; the target variable in the configuration file is streamflow, and the main evaluation metrics are NSE/KGE. Meanwhile, the public configuration file also states that this open-source pipeline differs from the operational pipeline.

I suggest that the authors more clearly distinguish whether what is evaluated and open-sourced in this paper is the "runoff/streamflow model component" or the complete "operational flood warning system". If the paper claims to evaluate the complete system, it should explain whether operational components such as real-time

data validation, flood-threshold determination, inundation mapping, and alert distribution are included in the evaluation and open-source code. If they are not included, I suggest revising the wording in the title, abstract, methods, or Code Availability section, and clearly specifying which results can be reproduced using the released code. The provided code and the paper need to be explicitly aligned.