<Overall Comments>

This manuscript presents an evaluation of version 2 of the Google Global Flood Forecasting system against the previous v1 system and established third-party benchmarks. The study is valuable because it provides transparency about an operational global flood forecasting system, introduces important technical updates such as the ME-LSTM architecture and GraphCast meteorological forcings, and contributes open-source resources for the hydrological community. I consider the topic important and suitable for publication after revision.

My main concerns are related to the interpretation of the reported performance improvements. First, the comparison between v1 and v2 may be affected by changes in the gauged/ungauged status of evaluation basins, because the v2 system uses expanded training data. If some basins were ungauged in v1 but gauged in v2, part of the reported improvement may reflect increased spatial coverage of local streamflow training data rather than improvements in the model architecture or meteorological forcings. This issue should be clarified and, if relevant, quantified.

Second, the manuscript mainly evaluates the system using NSE and KGE components. These metrics are useful for assessing overall hydrograph prediction skill, but the manuscript is framed as a flood forecasting study. The authors should therefore discuss more clearly what the reported improvements imply for practical flood prediction, especially with respect to the improved correlation component, reduced forecast variability, flood peak timing, peak magnitude, and other aspects of flood warning performance. If event-based verification is beyond the scope of the paper, its absence should be acknowledged as a limitation or future direction.

Overall, I think the paper has strong potential, but the above issues should be addressed to make the central conclusions more robust and easier to interpret.

<Major Concerns>

[1] Possible confounding due to changes in gauged/ungauged status between v1 and v2

My most important concern is that it is not clear whether each evaluation basin has the same gauged/ungauged status in both v1 and v2. Since the v2 system uses expanded training data, including Caravan, some basins may have been ungauged in v1 but gauged in v2.

If such basins are included in the current “gauged” evaluation, the reported improvement from v1 to v2 may reflect not only improvements in the model architecture or the use of GraphCast forcings, but also the effect of newly including local streamflow observations from those basins in the training data. In other words, the improvement may partly reflect an ungauged-to-gauged transition. In that case, interpreting the v2 improvement mainly as an effect of the upgraded model structure would be potentially misleading.

I therefore ask the authors to clarify whether the gauged/ungauged status of each evaluation basin is consistent between v1 and v2. It would also be useful to separate the evaluation into at least the following groups:

- basins that are gauged in both v1 and v2;
- basins that are ungauged in v1 but gauged in v2;
- basins that are ungauged in both v1 and v2.

This decomposition would help distinguish the effects of model and input-data improvements from the effect of increased spatial coverage of the training data. In particular, if basins that changed from ungauged in v1 to gauged in v2 show large improvements, the interpretation of the current aggregate v1-v2 comparison may change substantially.

[2] Limitations of NSE/KGE and the practical meaning of the improvements for flood forecasting

The manuscript demonstrates improved hydrograph prediction skill of the v2 system using NSE and KGE components. This evaluation is useful. However, because the manuscript focuses on a flood forecasting system, I think the authors should discuss more clearly what these improvements mean from a practical flood forecasting perspective.

In particular, the improvement in the correlation component of KGE is important. It may indicate better prediction of hydrograph phase, rising limbs, and flood peak timing, which are highly relevant for early warning. On the other hand, the reduction in forecast variability may imply possible underestimation of peak discharge. I therefore suggest that the authors interpret the meaning of both improved correlation and reduced variability more carefully in the context of flood forecasting. Although the Conclusions identify the KGE decomposition result as one of the main improvements, the main text currently contains relatively little discussion of why the correlation improvement is especially important.

If possible, it would also be helpful to include one or a few representative hydrograph examples, such as a basin where v1 missed the timing of a flood peak but v2, or v2 with GraphCast, captured it better. Such examples would help readers understand how the improvement in statistical metrics appears in actual forecast time series.

Finally, NSE and KGE alone do not directly evaluate several important aspects of flood disaster prediction, such as peak discharge, threshold exceedance, false alarms, and missed events. This limitation should at least be clearly acknowledged in the Conclusion or in a Limitations/Future Directions section.

<Specific Comments>

Abstract:

In the Abstract, the system is described only as an “updated flood forecasting system”, but the name of the Google Global Flood Forecasting system is not explicitly stated. Although this is already included in the title, I think it would be useful to name the system explicitly in the Abstract, since the Abstract is often read independently.

In addition, the current Abstract does not clearly explain what technical changes were introduced in v2. I recommend adding one concise sentence summarizing the main technical updates, such as the replacement with the ME-LSTM architecture, improved integration of multiple meteorological input products and robustness to missing inputs, expanded training data through Caravan, and the inclusion of GraphCast meteorological forcings. This would help readers understand the technical basis for the reported improvement, rather than only seeing the performance outcome.

P2 L15: Alignment between the Introduction and the Results

The limitations of the previous system and the improvements introduced in the new system should be presented in a way that is more clearly aligned with the analyses in the Results section.

In the current Introduction, the v1-to-v2 upgrade is described as addressing three data-related challenges: training data availability, temporally limited data records, and input data distribution shifts. These are relevant points, but the Results section mainly discusses the improvements in terms of two components: improvements on the hydrological model side, including ME-LSTM and expanded training data, and improvements on the meteorological forcing side through the use of GraphCast.

I think the Introduction would be clearer if it first described the main limitations of v1 and then explained how v2 was designed to address them through both an improved model architecture and improved meteorological forecast inputs. This would make the narrative from motivation to methods and results more consistent.

If you include analysis on “ungauged to gauged” impact in the result, please arrange this part to align with the analysis in the updated result section.

P2 L18: The study objectives should explicitly include performance evaluation.

At the end of the Introduction, the authors state that the two main objectives of the paper are to provide transparency about the progress and challenges of the operational flood forecasting system, and to facilitate research on ML-based flood forecasting by providing open-source resources. However, the main focus of the manuscript is the performance evaluation and benchmarking of the v2 system. I therefore suggest that the stated objectives should explicitly include evaluating the predictive performance of the v2 operational system against v1 and

third-party benchmarks. This would make the objectives better aligned with the structure and conclusions of the manuscript.

Table 1:

Table 1 is useful for reproducibility because it provides the full list of static catchment attributes. However, the table is very long and mostly consists of an enumeration of input variables, which substantially interrupts the flow of Section 2.1.1. I suggest keeping a concise summary in the main text, including the number of attributes, data sources, major categories, and representative examples, and moving the full attribute list to the Supporting Information or an Appendix. This would improve readability without reducing reproducibility.

P6 L9:

The descriptions of the meteorological input data in Section 2.1.2 and the training settings in Section 2.3 are presented mainly as bullet lists. The use of bullet lists itself is not a problem. However, for a model description paper, it is important not only to state what was used, but also to explain why those design choices were made and what data-availability or operational constraints motivated them. I suggest adding more explanation of the rationale behind choices such as feature unioning, input feature dropout, noise injection, batch size, number of epochs, and batch limits. This would make the model design and training strategy easier to understand and reproduce conceptually.

Figure 3:

Figure 3 is important for explaining the forecast initialization artifact in the ED-LSTM, but in its current form it is not easy to identify where the unnatural behavior appears. I suggest that the authors indicate the transition point from the hindcast period to the forecast period more clearly, for example using arrows, annotations, or highlighting, and explicitly show which part of the predicted hydrograph corresponds to the artifact. It would also help readers if the authors showed, next to the problematic example, a case without a strong artifact or a corresponding ME-LSTM example where the issue is reduced.

More generally, figures with multiple panels should include panel labels such as (a), (b), and (c). This would make it easier to refer to specific panels in the text and captions.

P10 L9 ME-LSTM

The ME-LSTM is one of the central technical improvements in this manuscript, but the roles of the two LSTM layers are not sufficiently clear from the current text and Figure 4. My understanding is that the first LSTM layer represents the evolving hydrological state derived

from the hindcast sequence, while the second LSTM layer combines this state information with forecast embeddings to predict future streamflow. However, the current description does not make clear whether the first layer is only used as an initialization mechanism, or whether it continues to update state information during the forecast period.

It is also unclear how the training loss is applied across the hindcast and forecast periods, and whether the forecast layer is specifically optimized for future lead-time predictions. These points are important for understanding how the ME-LSTM differs from the ED-LSTM handoff approach.

I therefore suggest that the authors explain more clearly, both in the text and in Figure 4, the different roles of the hindcast and forecast models in the ME-LSTM, the flow of information between the first and second LSTM layers, how state information is updated during the forecast period, and how the loss function is applied.

Figure 4:

In Figure 4, it is not clear which LSTM block corresponds to the first layer and which corresponds to the second layer. Since the text describes the ME-LSTM as a two-layer stacked LSTM, the first and second LSTM layers should be explicitly labelled in both the figure and the caption.

The meaning of "Output" in the figure should also be clarified. It is currently unclear whether this refers to the predicted streamflow, the parameters of the CMAL predictive distribution, or the deterministic mean discharge used for evaluation.

In addition, the handling of missing inputs, which is an important advantage of the ME-LSTM architecture, is not easy to understand from the current figure. I suggest making NaNs or missing input products more visually prominent, and clearly indicating which inputs are included in the masked mean operation and which inputs are excluded. This should also be explained explicitly in the figure caption.

Figure 5:

Figure 5 is one of the key figures for the global performance comparison. However, the current CDF panels show many lines corresponding to multiple model configurations and multiple lead times at the same time, making the figure difficult to interpret. I suggest reorganizing this figure, for example by separating the comparison among models from the comparison across lead times, either into different figures or different panels. Another option would be to show CDFs only for selected representative lead times, while presenting the full lead-time dependence using boxplots or median performance curves.

The legend font is also too small and should be enlarged. In addition, each panel should be

labelled clearly, for example as (a), (b), (c), and (d), so that the text and caption can refer to the individual panels more easily.

Figure 8:

Figure 8 supports one of the central conclusions of the manuscript, but the upper panels, especially the upper-left panel, are difficult to interpret because the legend is insufficient. It is not clear from the figure alone what is being compared. The authors should more clearly indicate the correspondence among the v1 nowcast, the v2 forecasts at different lead times, and the gauged/ungauged settings, both in the figure and in the caption.

Section 4.4 Effect of Hydrological Characteristics

In Section 4.4, the authors analyze the relationship between hydrological characteristics and model performance. In addition to the current attribute-based analysis, it would be useful to show the spatial distribution of the v1-to-v2 skill improvement on a world map. This would help readers understand where the updated system improves most, and whether the improvements are concentrated in particular regions or hydroclimatic settings. If the effects of gauged and ungauged evaluation are mixed, it may be better to show separate maps for gauged and ungauged basins. This would also help clarify whether the spatial pattern of improvement is related to model generalization, local training data availability, or meteorological forcing improvements.

If such a spatial map is added to the main text, Figure 10 could potentially be moved to the Supplementary Information, since the map may provide a more direct and intuitive view of where the model improvement occurs globally.

Section 4.5 and Figure 11

Section 4.5 and Figure 11 provide a useful comparison between gauged and ungauged performance. However, the discussion could be expanded to better explain what this performance gap implies for the reliability of the system in ungauged basins. Since global flood forecasting often targets regions where local streamflow observations are limited or unavailable, the relative performance of ungauged predictions compared with gauged predictions is highly important. I suggest that the authors discuss more explicitly how large the ungauged penalty is, whether it varies by region or hydrological characteristics, and what this means for operational confidence in ungauged basins.

Figure 11

In the lower panel of Figure 11, the line corresponding to the improvement ratio or relative

percentage change appears to be shown as a dashed line. However, this dashed line style is not represented in the legend. The authors should revise the legend so that the line styles and colors are consistent with the plotted data.