Reviewer 2:

While the authors raise several important and timely questions regarding training strategies in hydrological machine learning, I believe the current version of the manuscript is affected by some methodological issues that limit the strength of its conclusions. After reviewing both the manuscript and Reviewer 1's detailed comments, I respectfully offer the following major and minor suggestions for improvement.

Response: We thank the reviewer in recognizing the merit of this work and providing constructive comments. lease find our responses below along with our responses to Reviewer 1.

Major Comments

1. Model Comparison Framework and Associated Limitations

I share Reviewer 1's concern regarding the comparability of the models evaluated in this study. The current experimental design involves comparisons among models trained on datasets with different resolutions, quality levels, and real-time availability. In particular, the contrast between:

- Model G (a global operational model using coarse, real-time data),
- Regional models (using high-quality reanalysis data), and
- Single-basin models with lagged streamflow inputs (S-6),

raises concerns about fairness and interpretability. Because these models operate under different data assumptions, it becomes challenging to isolate the effect of training strategies alone. As such, the conclusions drawn about the relative performance of global, regional, and single-basin approaches may be difficult to support in their current form.

Response: We understand the reviewer's concerns about model comparability. We have discussed the philosophy of comparison of models with different types and quality of inputs in detail in the Discussion section.
If this were simply a model development study, we would agree that using the same dataset would ensure fairer comparison – something that Kratzer et al. (2024) already carried out. However, the focus here is on the ultimate utility of models - specifically, which model provides better prediction results in real-world scenarios. Considering this, there is no reason not to (1) train the model with heterogeneous data (different quality, resolution, different hydroclimatic condition) or (2) train with different model architectures and optimization techniques to select the *best and most robust model*. If we keep training models using data of the same quality – dictated by their global availability, the model architecture, optimization techniques, etc., we might be trapped in a "local minimum" of the model performance. We will miss a chance to have another model that might perform better.

We have discussed this in the Discussion section and outlined the role of data quality in model training. The key question is: *what is the purpose of training a model with multiple basins using less reliable data when that model performs poorly compared to a model trained with high-quality (local) data?* It is worth noting that many studies that Kratzert et al. (2024) claim are

"mistakes" actually use high-quality observational data from gauge stations rather than data from large sample datasets.

Regarding model S-6, which uses observed streamflow as input, it is clearly stated that this represents model training for gauged regions in both our study and Kratzert et al. (2024). Therefore, training models with observed streamflow is not problematic when it delivers effective and high-quality forecasting. We question why one would not utilize available data to train a better model, instead ignoring observed streamflow and ultimately obtaining a model with inferior performance. In practice, when observed streamflow data are available, it is preferred to use it for model training rather than training without it, and we find no issue with this approach.

1. Alignment Between Research Question and Study Design

The manuscript appears to address two related but distinct questions:

- Whether models trained on multi-basin data outperform those trained on single-basin data (as stated in the title), and
- Whether locally optimized models using basin-specific data can outperform globally trained models using only limited local information (the implicit research question).

These are both important questions, but they require different analytical frameworks and modeling assumptions. To strengthen the manuscript, it may be helpful for the authors to clarify the primary research question and ensure that the experimental design is tailored to directly address it. Additionally, the first question has already been explored in earlier studies.

Response: Regarding the research questions, we have stated the main research questions in L76-81 in the manuscript and provided comprehensive approaches to answer them. Specifically, the two questions are *whether the studies that Kratzert et al. (2024) listed are mistakes when training with single basins*, and second, *whether training with multiple basins results in better model performance*.
To answer the first question, our approach was to review studies using single basin training approaches to examine the performance of those models as well as the types of data those models used. For the second question, we performed direct comparisons with a model that the same research group developed with published results as well as models that we trained. Note that our trained models still used ERA5-land forcing rather than local data (such as NLDAS-2/Daymet for CONUS or CEH-GEAR + CHESS-PE + CHESS-met for GB). We believe that when using local data sources, the performance of single models would be even better. This further highlights the importance of local/high-quality data in training ML models.
We emphasize that the criteria for evaluating whether a model is good or not should be based on the model's performance and its intended use but rather than on how the model is configured. From an application perspective, as long as the model performs well, that is sufficient. We have clarified these criteria for evaluating model quality in the revised manuscript.
We also concurred with the Reviewer #1 comment that the previous title of our manuscript was confusing. Therefore, we revised the title to "*Is Multiple Basin Training the Best Practice for Machine Learning Streamflow Prediction in Gauged Basins?*" to avoid misunderstanding, particularly regarding the first question, as the reviewer pointed out.

1. Constructive Outlook on Global Model Development

Despite the challenges highlighted in this study, I remain optimistic about the ongoing development of global hydrological models. As data availability and computational methods continue to improve, I believe there is great potential for globally trained models to better incorporate local information.

Rather than viewing the current findings as a critique of global approaches, I suggest framing them as an opportunity to guide future research toward:

- Better integration of local knowledge within global models,
- Scalable methods for collecting and assimilating basin-specific data, and
- Hybrid modeling strategies that combine global generality with local specificity.

This perspective may help position the work within a more forward-looking and solution-oriented context.

Response: This is an excellent suggestion. We have added some related content to the outlook regarding the development of global hydrological models:

"*We acknowledge that as data availability and computational methods continue to improve, there is a great potential for globally trained models to better incorporate local information, particularly given the current strong development trend of global hydrological models (Kraft et al., 2021; Müller Schmied et al., 2021; Emerton et al., 2016). However, it is important to recognize that hydrological forecasting results are only meaningful when applied at the local scale, where water resource management and disaster (flood/drought) decisions are made using results at specific locations and specific watersheds, rather than using general global performance. The analysis results in this study demonstrate that state-of-the-art global models exhibit poor performance, especially in flood forecasting with peak errors reaching ~50%, indicating that substantial work remains to be done to improve modeling systems capable of accurate forecasting at the global scale. In this study, we highlight the need for and importance of high-quality data and integration of local knowledge within global models. For forecasting in gauged basins, methods such as data assimilation or data integration are particularly necessary to provide better forecasting results. This suggests the potential for developing hybrid models that can operate at the global scale, while being configured and customized at the local scale.*"

Minor Comments

1. Overlap Criterion for Gauges

The criterion used to identify overlapping gauges ("with distances not exceeding 1 km") would benefit from further justification. Given potential uncertainties in station locations and historical relocations, a brief explanation or reference supporting this threshold would help strengthen the methodological transparency.

Response: Thank you. In the revision, we have improved and added additional criteria for detecting overlapping gauges. Specifically, beyond the distance between gauges, we also

evaluate the correlation between observed data from the datasets. If R^2>0.99, we identify them as overlapped gauges. This results in changes showing 31 and 124 overlapped gauges for CONUS and GB areas, respectively. References for these criteria have also been added in the revised manuscript.

1. Peak Flow Threshold Consistency

The thresholds for identifying peak flow events differ between figures (e.g., 95th percentile in Figures 1 and 2, versus 99th percentile in Figure 3). Clarifying the rationale behind these choices would improve the consistency and comparability of the analyses.

Response: We have explained the use of different thresholds in the revised manuscript. Specifically, the results in Figs. 1 and 2 are to demonstrate the model performance in simulating high flow. Meanwhile, in Fig. 3, we wanted to evaluate the model performance specifically in simulating extreme events (here we use the 99th percentile to identify these events). We have revised the figure caption to clarify that.