Response to Referee 2. Comments are in black, responses are in blue.

Main points

Pg3, bullet iv: Which version of GloFAS is used in the paper? Version 2 and previous were forced by runoff from the ECMWF land surface model ECLand (formally, HTESSEL) with river discharge produced by the LISFLOOD river routing scheme, whereas from version 3, the full LISFLOOD hydrological model was used. This is quite important given the premise of the benchmarking here. If using version 2 (forced with HTESSEL + LISFLOOD river routing) then this work benchmarks only the river routing part, but if GloFAS version 3 or newer, then GloFAS is driven by a full LISFLOOD hydrological model (forced with e.g. Precipitation and temperature from ERA5, rather than runoff). See the GloFAS version system changes and associated documentation for details: https://confluence.ecmwf.int/display/CEMS/GloFAS+versioning+system
Again on Pg 10, line 5-6: If GloFAS v3 onwards, then it uses the fully physical-based LISFLOOD model, if using GloFAS v2 it uses runoff from ECMWF ECLand + the LISFLOOD river runoff scheme.
In our work we used GloFAS version 4 and we have now made that explicit. While it's true that in this version the full LISFLOOD is used to generate runoff and route river water, we think that it is still a good benchmarking because the runoff is calculated explicitly from a physics-based model, as is HTESSEL. We have tailored the text to make this distinction as clear as possible.

Pg 4, 2nd line from bottom: GloFAS is forced with ERA5 not ERA5-Land, while there is not strong differences, ERA5-Land does show better performance for hydrological modelling (see Munoz-Sabater et al. (2021) for a hydrological benchmark on GloFAS with both). These differences impact the benchmarking in Lima et al. (2024) and the details and differences in experimental set-up must be qualified.
We have made this distinction clear in point "(i)" of Section 3.2.1.

Pg 8, last para, line 5-6: This seems to be where the detail of the ML model as used in this paper is outlined. It's constrained to the Appendix B with details found in code uploaded to Zenodo. While I strongly support and compliment the authors for uploading their code, I still think there is insufficient detail and considerations of limitations explained within the main manuscript. The method is very difficult to repeat without more detail explained to the reader. Figure 3 outlines the RNN, but where are the assumptions/limitations for this particular river routing application, where are the detail on the model training method/time periods/temporal resolution used here etc.
We agree with the referee and we have changed appendix B to include a detailed explanation of the LSTM architecture. Section 2.2 now describes all the hyperparameters used to both setup the RNN architecture and to train it. The temporal resolution of the dynamical inputs are stated in this same section. We hope that with these modifications it's now easier to repeat the method.

Pg 10; line 7: what do you mean by "hindcasts"; hindcast is used in the forecast literature to mean running forecasts for past dates. However, I do not believe you are running forecasts here.
We took the term out of the text and adjusted where needed.

Pg 16; line 6: In the performance of models for different geographical regions in the world, it's important to mention that the LSM within ERA5/Land is forced with precipitation from the Numerical Weather Prediction (NWP) model used within ERA5. NWP models fundamentally struggle to capture precipitation in the tropics, and this will impact the results here. See for example Lavers et al. (2022).
We thank the referee for the reference and the point raised. We have added that to Section 4 on the third paragraph as: "Moreover, the ERA5-Land reanalysis is driven with precipitation from ERA5, known to have biases in the tropics Lavers et al. (2022), which could lead to biases in the runoff of ERA5-Land in these regions. As a consequence, our river model could be learning to correct these biases in addition to routing water. This is a possible pitfall for machine learning models trained with model output and not with observed data."

Table A1: Key details missing. Which datasets do each variable come from? What time period, temporal resolution is used?
We added the source for the data presented in the table in Appendix's A text. It's important to notice that all these variables are static attributes and are supposed not to have time dependence.

Minor points

Pg2, para 1, line 5: please change "regions is" to "regions are".
We corrected this.

Pg2, para2, line 5: suggest changing "forming the streamflow" to "forming streamflow".
We rephrased this.

Pg 3, bullet iv: Change "run operationally by the European Copernicus program" to something like, "GloFAS, the European Union Copernicus Emergency Management Service (CEMS) global flood forecasting system run operationally at the European Centre for Medium-Range Weather Forecasts (ECMWF)".
We corrected that after suggestions made by the GloFAS team in private conversations.

Pg8, line 2: missing the description of "Xs"
We have added an explicit description of x^d_t and x^s.

Pg 9, line 7-9 (I think, there are no line numbers included!): An NSE > 0 or KGE > ~-0.41 is not "good". The interpretation is that the model is performing better than a mean flow benchmark. This is what NSE=0 or KGE=~-0.41 means. It shows your model provides some level of skill beyond a very naïve mean flow benchmark. This should not be confused as "good". But great that you use the ~-0.41 threshold for the KGE, I agree with this!
We apologize for the line numbers. This happened because we posted our preprint both via EGU and via arXiv. EGU recommends having a single preprint, but arXiv doesn't allow papers with line numbers. We agree with the referee concerning the use of the word "good" and we have rephrased the text to avoid it.

Pg 14, Sect. 3.3, line 1: please change “present some simulated” to “present simulated”.
We corrected this

Pg 18; last line: please change “doesn’t” to “does not”.
We corrected this.

Pag 19; line 2: You say routing in LSM component in climate models – but it’s much wider than that. NWP models have a LSM, so this work is also relevant for short, medium and longer range hydrological forecasting using runoff from land surface models within weather models. It has a much wider impact that just the climate models!
We agree with the referee. While our main goal is to integrate this model to a climate model, it’s true that it can also be used within the LSM of any weather forecasting model, as long as it has an LSM component that explicitly accounts for runoff. We have modified the text throughout to include this.