This manuscript by Confidence Duku presents results of a model comparison on flood predictions in various rivers around the world.

I had a very hard time understanding parts of the paper that made reviewing the study in every detail impossible for me. Therefore, I will only concentrate on some main points before eventually looking more into details in a revised version.

1. **Bakaano-Hydro modelling framework**

What is the Bakaano-Hydro modelling framework? Almost no details are given in this manuscript and not even a reference to the seemingly connected preprint https://egusphere.copernicus.org/preprints/2025/egusphere-2025-1633/ is provided. Since the Bakaano-Hydro modelling framework is not an established thing, I expect this paper to at least contain a reference to the modeling paper but still list some details here. E.g. what are the input data (and e.g. for the weather data, from which product/provider)? From what I understand, the framework consists of three steps: 1) VegET, 2) routing, and 3) the TCN based neural network. Since I assume also part 1 and part 2 contain some model parameters, how are they calibrated? Since e.g. part 1 is still without routing, against which observations is this part calibrated? What periods are used for calibrating part 1 and part 2? Since part two is already a routed discharge, how well does this routed discharge perform against the observations from GRDC? How much additional benefit is gained by post-processing the routed discharge with the TCN?
In different places of the manuscript it sounds like the neural network part of the model is responsible for simulating streamflow. E.g. P7 L10: "*The final phase of Bakaano-Hydro involves the application of a deep learning model to simulate streamflow.*" Isn't part two already outputting routed streamflow? From my limited understanding, the TCN layer gets as input routed streamflow (converted to mm/d by dividing through the upstream area) and depths-to-water-index. How can this part be responsible for simulating streamflow? Isn't the neural network part simply a post-processor for the VegET+routing model?
What are the 3 dynamic inputs that are passed to the neural network, as mentioned on P7 L 27. The following sentences only list the routed streamflow and the depth-to-water-index.
Why has the neural network part two different branches, once where streamflow is normalized by the upstream area and once where it is normalized by max simulation? What was the reason behind building this architecture or was it just trial-and-error and in the end what-works-best? Which leads me to my last point for the modeling section: I think a lot of the decisions that have been made are not well explained and could require ablation studies to understand where exactly model performance improvement comes from.

2. **Data**

I strongly recommend splitting Section 2.1 into a dedicated data section and a dedicated model section. For the data part further, I recommend adding a table to the manuscript that clearly shows which input features are being used with columns for: name, units, source, time periods. For the meteorological data, what was the reason for choosing to use a (bias adjusted) climate

model rather than a meteorological dataset (either reanalysis, hindcast, or historical forecast model output)? Since this paper compares two different operational forecasting models and in various places suggests that the presented framework improves upon these operational systems, it is probably worth noting that the model, with the data from this study, could not run operationally.
Another point is that by choosing yet another weather dataset, it is hard to know how much of the performance differences can be attributed to the modeling framework or the the quality of the model input data. Since the presented setup is not suitable for an operational setting (see above) it begs the question what the main focus of this study is? When focussing on the comparison of models, it is usually advised to use the same input data sources for all models to eliminate the impact of data quality on the results. If the focus is comparing flood forecast systems (as in Nearing et al., 2024), then obviously each model can use anything, with the caveat that it should be available in a real operational context.

3. Experimental settings and evaluation protocol

There are actually multiple issues I see with the experimental setting or the evaluation protocol.

First: Operational flood forecasting models don't alert if simulations surpass thresholds computed from *observations* but rather if simulations surpass thresholds based on *simulations*, as it also has been done in Nearing et al. (2024) for GloFAS and the Google Flood Forecasting model. This removes the bias problems of simulations vs. observations and could lead to perfect flood alert rates, even if the model has a constant bias problem. Why is this even more important here? In this study, the author presents a model that performs essentially per-gauge bias correction (by inputting the routed streamflow into the TCN layer and fitting the TCN layer against observations) and then compares this to two globally calibrated operational models.

Second: P12 L 12 "*CSI, POD and FAR values for flood events with 1-, 2-, 5-, and 10-year return periods and 0-, 1- and 2-day timing tolerance were calculated using simulated data from Bakaano-Hydro for the period 1982 to 2016*" This sentence suggests that the majority of the evaluation period for which you computed metrics are the training period of the Baakano-Hydro model (P8 L9 "*The training period covered 1989–2016*"). For the Google Flood Forecasting model, even for the full-run, all simulations are the result of a temporal k-fold cross validation, as stated in Nearing et al. (2024), i.e. all simulations are "unseen test data". If my understanding is correct, this renders large parts of the evaluation/conclusion unfair, as per-gauge bias corrected training data is compared to test period data from a globally calibrated model with the additional caveat from the point above.

Third: The metrics are not well explained or referenced. Since I wouldn't consider CSI, POD, FAR super common metrics in modeling papers of hydrology, I would recommend adding an explanation of these metrics, probably the equations, and also mention the range of these metrics and their optimal values.

4. **Presentation of the results**

In my opinion, the presentation of the results, especially the figures, need some re-work.

Figure 2:
The diverging color scale, centered around 0 and with range (-1,1) does not make sense for all metrics, E.g. Alpha-NSE and Beta-KGE have their optimal value at 1. Also, please indicate what are the optimal values for each metric, which facilitates interpreting the results.

Figure 3/4/5:
If the whiskers show the full range, why are there points outside of the whiskers? Are the whiskers maybe just indicating a certain percentile?

Figure 6/7:
They feel a bit too much and/or not very well presented, as indicating the median of model performance over all stations in a river basin only by color makes it hard for me to really get any details out of these figures. I understand the reason behind these figures (showing model differences across spatial patterns) but maybe you find a better way to visualize the results. Maybe on a map? But two pages of colored rectangles feels like an overkill in the main paper.

On a more general note: Figure 3/4/5/6/7 all have three columns for different flood timing tolerance. I wonder if all three of these columns for each figure are needed in the main paper or if some of the plots could be moved to the appendix. In my opinion, there are not dramatically different patterns between the columns that make it necessary to have all three columns in the main paper.

Line by line comments:

- P12 L9 "*...using the annual maximum series (AMS) method*" This sounds like it is missing a reference.
- P12 L10f As far as I know, the paper by Nearing et al. compares GloFAS and the Google Flood forecasting model on their true forecast skill under operational settings (i.e. using only data that is available in real time) and evaluating the skill at different lead times.
- P22 L4 "*Importantly, Bakaano-Hydro advances the state of data-driven hydrological modeling by embedding physically meaningful processes within a neural network architecture.*" I think I disagree with this conclusion. How was the state of data-driven hydrological modelling advanced by the findings of this paper? The presented framework does not embed physically meaningful processes in a neural network. The neural network gets the output of a physically inspired model as input. This is something fundamentally different than "*embedding physically meaningful processes within a neural network*"