I thank the reviewer for their careful reading of my manuscript and for their constructive comments and suggestions. In the following, I respond to each comment point by point. Reviewer comments are reproduced in full (RC), followed by my author responses (AC). All revisions indicated will be incorporated in the revised manuscript.

RC1: The research compares three forecasting models in regions with low gauge density and usually low representation in traditional approaches. The comparison between models is based on return periods and their success or failure rates in capturing them. In all the metrics presented, the hybrid model outperforms the other models. These results are highly valuable in supporting the use of such models in forecasting frameworks. However, a couple of main points need to be addressed to fully validate the results.

The first point concerns the period used to define the number of successes of the hybrid model. The author used the training period to count these successes, which creates a biased metric that cannot be reliably used for comparison. Using the entire period to define return periods is acceptable, as it provides a good representation of them, but the success rate should be evaluated over an independent period to accurately estimate how the model performs on unseen data in operational settings.

AC1: This is a valid point. In my analysis I used the entire period, including the training and validation periods, for computing the metrics CSI, POD, and FAR. This was in line with the approach adopted by GloFAS. While this comparison is valid for the GloFAS model, I acknowledge that it can provide a biased and unfair comparison against the Google AI model. Hence, I will revise the setup following a similar cross-validation approach to that used in the Google AI model, ensuring that all predicted stations used for metric computations are strictly out-of-sample.

RC2: The second point relates to the number of gauges used in training for each model. This can be problematic because it may lead to unfair comparisons. For example, it is not fair to compare a model trained with 100 gauges in a region to another trained with only 10 gauges. The author only mentions the number of gauges common to all the models, used for evaluation, but does not specify the number of gauges used during training or the training period used for each model.

AC2: On page 4, line 21, I mentioned that 643 stations were used to train the model. It would be better to repeat this explicitly in the methodology section, and I will revise accordingly. I acknowledge that training sizes vary because this paper covers two continents, with the explicit aim of increasing forecasting reliability in Africa and South America—regions that previous studies such as Nearing et al. (2024) show to have the

lowest reliability. The results and conclusions from this paper apply only to these continents and are not global. I will make this explicit in the discussion. Importantly, the difference in training sample sizes actually biases against Bakaano-Hydro: the competing models had access to far more stations, including many from data-rich regions.

RC3: Be more specific about the catastrophic events in Nigeria, Sudan, etc., because they are unknown in other countries or continents.
AC3: I will revise this statement and add concrete examples with dates and impacts.

RC4: P3 L30–32. I am not sure if Fredrick Kratzert supports this statement in his last manuscripts. I am fairly certain that Mass Conservation LSTM demonstrated the opposite.
AC4: I refer the reviewer to Kratzert et al. (2019) (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019WR026065), where the authors concluded: "We found evidence that adding physical constraints to the LSTM models might improve simulations, which we suggest motivates future research related to physics-guided machine learning."

RC5: P4 L2–3. The previous statement and reference do not support your statement. It only mentions the need for higher resolution, which is not necessarily associated with the use of a process-based model.
AC5: I will revise the statement and include the appropriate references.

RC6: P4 L21. Is 643 gauges good enough for the characterization of two continents?
AC6: I acknowledge that 643 gauges may appear modest compared to global datasets. However, this represents the full set of openly available and quality-controlled records from GRDC for Africa and South America within the selected basins. Importantly, these stations span two continents and capture a wide range of hydroclimatic regimes—including humid tropics, semi-arid, arid, and subtropical systems—providing a representative basis for model training and evaluation. I also recognize that gauge density remains much lower than in regions such as Europe or North America, and this very limitation underscores the importance of developing frameworks like Bakaano-Hydro that can extract skill from sparse and heterogeneous observational networks.

RC7: P5 L3–15. A more detailed characterization is needed to fully describe the variability. Area sizes, aridity, slopes, total annual precipitation, etc. Are these areas poorly trained for the ML and process-based models?
AC7: I agree with the reviewer and will revise this section accordingly.

RC8: P5 L16. Add reference to VegET method.
AC8: I agree with the reviewer and will revise this section accordingly.

RC9: P7 L7–9. How much distortion does the 1 km resolution bring for a river with much lower width networks (100 m)?
AC9: I acknowledge that representing river networks at 1 km resolution inevitably smooths narrower channels, which can attenuate flood peaks and introduce timing lags, particularly in smaller catchments. To mitigate this effect, I snapped gauges to the river network and applied a minimum catchment size threshold of 1,000 km^2, ensuring that the evaluation focused on basins where 1 km routing is a reasonable approximation. I will make these points more explicit in the revised version.

RC10: P7 L19–23. This is not a fair statement because CNN is used to convert the routed streamflow to the actual streamflow, which means it was never exposed to the vanishing of memory issue. After all, the process-based model is dealing with that.
AC10: I agree with the reviewer that the TCN component in my framework is not directly exposed to vanishing gradient issues, since runoff generation and routing are handled by the process-based model. My intention, however, was to highlight that the overall hybrid architecture avoids the vanishing gradient limitations faced by purely data-driven approaches when applied directly to long hydrological sequences. By combining a process-based model to carry long-term memory with a TCN for efficient local pattern extraction, the architecture as a whole circumvents these issues while retaining hydrological consistency. I will revise this section accordingly.

RC11: P7 L28. Add a figure with the architecture and the input of the TCN.

AC11: I will revise this section accordingly and include a figure of the architecture, explicitly showing the TCN inputs and processing branches.

RC12: P8 L1. Add more details about spatial periodicity.

AC12: I will revise this section accordingly and provide a clearer explanation of how spatial periodicity was handled in the model.

RC13: P8 L23. How many gauges from the training of each model are present in the areas studied? What if some of the models used gauges and others did not? That is super important to have a fair comparison.

AC13: I thank the reviewer for raising this point. As stated, I used a common subset of 470 stations across all three models to compute evaluation metrics. In doing so, my benchmarking approach (using the full 1982–2016 period) is directly comparable to GloFAS, which similarly reports skill over its entire hindcast period. However, I recognize that this setup may be less fair to the Google AI model, since its simulated streamflow is produced entirely out-of-sample. To address this, I will revise my experimental setup using a cross-validation strategy to ensure that, for Bakaano-Hydro, all discharge data used for computing skill metrics are strictly out-of-sample. This adjustment will provide a fairer basis for comparison across models while preserving methodological consistency.

RC14: P9 L19–20. Why was only one distribution used when each catchment can have a very different distribution?

AC14: I agree with the reviewer that different catchments may be better represented by different extreme value distributions. However, for the purposes of this benchmarking exercise I adopted a consistent approach, following Nearing et al. (2024), who also applied a single distribution across basins to ensure comparability. This avoids introducing additional variability due to distributional choices and ensures that observed differences in performance are attributable to the models rather than the fitting method.

RC15: P10 L4. Why did you use the training period for the metric? This is a biased estimation. You must use the validation period.

AC15: I thank the reviewer for raising this important point. My decision to compute flood detection metrics over the full 1982–2016 period (including both training and validation subsets) was motivated by three considerations. First, this approach is consistent with the evaluation strategy used in GloFAS (e.g. Alfieri et al., 2013; Harrigan et al., 2023), where accuracy is reported over the entire hindcast record rather than split validation subsets.

Second, the independent validation window for Bakaano-Hydro (1982–1988) is too short to support robust estimation of 5- and 10-year return periods, which require multi-decadal records. Third, using the full period provides a fair basis for comparison with GloFAS, which does not provide a distinct validation set.

I acknowledge, however, that this setup is less fair to the Google AI model, since its simulated streamflow is generated entirely out-of-sample. To address this, I will revise the experimental setup to use a cross-validation strategy, ensuring that all stations and discharge records used in the computation of skill metrics for Bakaano-Hydro are strictly out-of-sample.

RC16: P10 L27. Be careful with being overconfident with your results. They are only valid for those catchments studied; this does not mean a true generalization.

AC16: I will revise this statement accordingly to avoid overgeneralization and ensure that claims are limited to the specific basins and conditions studied.

RC17: P11 L6. Comparing with the testing period generates an overconfident metric. This is a biased analysis.

AC17: Figures A1 and A2 were intended to compare the cumulative distribution functions and probability density functions of seven metrics across the three models. I recognize, however, that this comparison is unfair to the Google AI model, since its simulated streamflows are generated entirely out-of-sample, whereas the Bakaano-Hydro results included both training and validation periods. To address this asymmetry, I will revise and reproduce these figures using results from Bakaano-Hydro's cross-validation setup, ensuring that streamflow for each gauging station is generated strictly out-of-sample across the full evaluation period. This will provide a fairer and more balanced comparison across all three models.

RC18: Figure 2. The color scheme is misleading the reader. Values near zero are already a bad performance. Please plot on a scale 0–1.

AC18: I will revise the figure accordingly to use a 0–1 scale and ensure that the color scheme better reflects performance quality.

RC19: P16 L12. Please, place this information in context. Add an aridity index or some descriptor to define clearly how arid those regions are.

AC19: I will revise this statement and incorporate an aridity index or similar descriptors to clearly contextualize the results.

RC20: Figure 6. "For each plot only basins for which differences among the three models were statistically significant are shown." What does it mean? Significant differences for what model? Does it mean you are plotting only the catchments when your model was significantly better than the others?

AC20: I thank the reviewer for pointing out the lack of clarity. To determine statistical significance, I carried out Wilcoxon signed-rank tests between model pairs, focusing in particular on Bakaano-Hydro and the Google AI model, which were the two best-performing frameworks. Basins with a p-value < 0.05 were then selected as those showing statistically significant differences in performance. Figure 6 therefore highlights only those basins where the performance differences between models were statistically significant according to this test. I will revise the text and caption to make this procedure explicit.

RC21 (P19 L10–11): From your results it is clear that your model is better than others; however, it is not clear where this improvement is coming, PB or ML. It would be valuable if the same analysis is added between your PB (without the CNN model) and the GloFas, given that both are PB.
AC21: I thank the reviewer for this suggestion. The improvement of Bakaano-Hydro stems from the combination of both the process-based (PB) component and the machine learning (ML) component. The PB part (VegET + routing) provides structured, physically meaningful signals, while the ML part (TCN) learns to map those signals to realistic streamflow. Importantly, the PB component in my framework is not calibrated against discharge; it is used as an uncalibrated generator of physically consistent runoff and routing signals. The added value therefore comes from how the ML component leverages these structured signals to reproduce observed flows. While a direct PB-only vs. GloFAS comparison would be interesting, it is outside the scope of the present benchmarking study, which is focused on evaluating the hybrid system as a whole against existing operational models.

RC22 (P20 L2–5): From my point of view, both approaches are valuable depending on the purpose. Observations are good for models that are implemented or pretend to be

implemented as an official tool. Simulations are good to present the chance of applying this model in an operational model after a bias correction (fine-tuning). Therefore, there is no approach better than others, just different purposes.
AC22: I acknowledge the reviewer's perspective and agree that both observation-based and simulation-based approaches are valuable, depending on their purpose. My intention was not to claim superiority of one approach over the other, but to demonstrate that using observed thresholds enables a fairer basis for benchmarking across models in the context of this study. I will revise the text to make this distinction clearer.

RC23 (P20 L6–7): I am not sure we can call this a new paradigm. Research applying hybrid models is abundant in the literature. Moreover, the idea of multi-representation approaches is mentioned already in the literature.
AC23: I agree with the reviewer that my wording overstated the contribution. I will revise the manuscript to avoid calling this a "new paradigm" and instead emphasize the novelty of applying a hydrology-guided hybrid approach at continental scale in data-scarce regions.

RC24 (P20 L11–12): Careful with overselling your research; it is not clear that the diversity studies in the models would allow you to make this statement.
AC24: I accept this point and will revise the conclusion to avoid overselling. My statements will be limited to the basins and contexts actually studied.

RC25 (P20 L21): How interpretable is a hybrid model? From where are the good results, PB or ML?
AC25: The interpretability of the hybrid model comes from the structured role of each component. The PB part constrains the inputs with physical realism (runoff and routing dynamics), while the ML part captures the non-linear transformations required to map these signals to observed discharge. The good results cannot be attributed to one part alone but arise from the integration of both. I will revise the text to make this explanation more explicit.

RC26 (P21 L1): The generalization to an ungauged basin is not well supported or explained in the manuscript. For example, how good does the model perform in regions with more extreme climates (Northern Chile or southern Argentina)?

AC26: I agree with the reviewer and will revise the manuscript to better qualify the statements on generalization.

RC27 (P21 L3): From the point of view of a global analysis. Is the Bakaano PB model running in an operational framework? This way, different countries could easily train a CNN model to fit the local information in each country. If this model and the data used to run it are not freely available, it will be very hard to implement in an operational framework.
AC27: At present, Bakaano-Hydro is not implemented as an operational system. However, the framework has been designed with operational scalability in mind. Both the PB component (VegET runoff generation and routing) and the ML component can run with publicly available forcing datasets, such as reanalysis or bias-corrected climate model outputs. I will add a discussion note to make this operational pathway explicit, including the need for open access data and code availability for broader adoption.