RC#2

General comments:

- This is a nice paper which demonstrates the benefits of implementing reservoir operation schemes designed specifically for hydropower reservoirs and an improved version of the Xanthos model. I have a couple of minor comments but otherwise think this will make a nice contribution to the wider literature.

→ We deeply appreciate your thoughtful review and your recognition of the merits of our research. Your positive feedback regarding the implementation of reservoir operation schemes specifically designed for hydropower reservoirs, as well as the improvements we have made to the Xanthos model, is highly encouraging. We have carefully reviewed and addressed your minor comments, and we believe they have further enhanced the quality of our manuscript. Thank you again for your time and contribution to the refinement of our work.

Specific comments:

- L35 – the text here switches between using km3 and m3 to describe the reservoir capacity, it might be easier to compare the numbers if you use consistent units.

→ Thank you for drawing our attention to this inconsistency. We appreciate your suggestion and have revised the manuscript to express all measurements of reservoir capacity in that paragraph in km^3 for easier comparison and consistency. See, revised paragraph 1 of the introduction below

"Reservoirs are pivotal in fulfilling various societal needs, including irrigation, hydropower production, flood control, domestic water supply, and navigation, to list a few (Belletti et al., 2020; Biemans et al., 2011; Grill et al., 2019). There are 6,862 large reservoirs ($\geq 0.1km^3$) globally, with a cumulative storage capacity of 6,197 km^3 in the Global Reservoir and Dam (GRanD) dataset (Lehner et al., 2011). Many of these reservoirs serve multiple purposes. However, if we partition reservoirs into categories based on their primary purposes, 1,789 are irrigation reservoirs with a total storage capacity of ~1,100 km^3; 1,541 are hydropower reservoirs with a total storage capacity of ~3,880 km^3; 542 are flood control reservoirs with a total storage capacity of ~509 km^3; and the rest are water supply, navigation, or recreation reservoirs. Water storage and releases in any given reservoir are managed based on the reservoir's purposes. It is, therefore, important in Global Hydrological Models (GHMs) to represent how management strategies differ across reservoirs with different purposes in order to more accurately simulate water balances and explore the implications of alternative water management strategies. It is particularly important to distinguish the behavior of hydropower reservoirs from others because hydropower production represents the primary purpose for nearly 63% (based on GRanD) of total global reservoir storage capacity. "

- Have the authors looked into whether the model performance is much worse downstream of lumped reservoirs compared to single reservoirs? I'd be interested to know whether combining multiple reservoirs (and reservoirs of multiple types) has a significant impact on the simulated flow.

→ Thank you for your insightful suggestion. We agree that exploring the performance of the model in the context of lumped reservoirs compared to single reservoirs could yield significant information.

However, our current study did not make this distinction. We do recognize the merits of such comparison, but it is important to note that the lumping of reservoirs is a common practice in GHMs, as detailed in the study by Telteu et al., (2021) (https://doi.org/10.5194/gmd-14-3843-2021). This is primarily a result of the general model structure, which uses a gridded representation. From a scientific perspective, we believe it is beneficial to maintain this lumping approach, in line with other GHMs, particularly as we are introducing hydropower reservoirs as a separate group into this form of representation. Moreover, to individually represent reservoirs, we would need to shift from a grid-to-grid routing framework to a vector-based flow routing framework, an option not currently supported by Xanthos. We anticipate exploring into this area of research in future endeavors. We appreciate your perspective and will consider exploring this aspect in our future work.

- L212- the new hydropower scheme is described as being trained with naturalized inflow generated by MRTM without the water management option, and I wondered if you might be able to give some context to the quality of these reservoir inflow simulations? Often bias in model inflows makes running online reservoir simulations tricky, is that something the authors found impacted this work?

→ Thank you for your insightful question. While we agree that bias in model inflows can potentially complicate online reservoir simulations, our approach has been developed to lessen this issue to a certain extent. It is indeed a complex issue. Irrespective of the quality of the input flow, the model is designed to generate an optimal release scenario based on that input, and this can add complexities to the understanding of the quality of the simulated results as there is no direct validation. In this work, we utilized a two-stage calibration procedure. The primary aim of the first stage was to achieve an accurate representation of the annual water balance (please refer to L270-285 for more detail, shown below here in red). This methodology in the first-stage also offers the benefit of generating a reasonable approximation of naturalized flow, serving as a solid foundation for our hydropower scheme. Notably, the input runoff to the second stage (where the reservoir simulation takes place) has already undergone a filtering process at the first-stage to control the annual water balance and is considered of sufficient quality. We plan to investigate further into the uncertainties surrounding these online reservoir inflow simulations in our future studies.

"In the first stage, we determine the optimal values for the five parameters in the runoff generation module (see Table 2) in four steps. 1) We generate one million runoff parameter combinations using a Latin Hypercube Sampling (LHS) scheme (McKay et al., 1979)(Fig. S1). LHS is a statistical method for multidimensional parameter space sampling. The stratified sampling strategy employed by LHS ensures that all portions of the sampling space are represented (McKay et al., 1979). The user decides on the required number of parameter combinations and individual parameters' upper and lower bounds. Based on that, LHS simultaneously stratifies all input dimensions. 2) For each runoff parameter combination, we execute the runoff module to produce the simulated monthly total runoff time series at each grid cell in the study period. In this study, we uniformly apply the same parameter values to all the grid cells in a basin to generate monthly runoff time series at each grid cell. Parameter values vary among basins, just not across grid cells within a basin. 3) We calculate the simulated annual runoff depth at each grid cell. We then take the spatial average across the grid cells within the upstream drainage area of a gauge station where observed streamflow data is available, denoted as Q_{sim_annual} (mm/year). 4) At the river gauge station, we take the long-term mean of observed streamflow and divide it by the drainage area, Q_{obs_annual} (mm/year). We then select the top 100 runoff parameter

combinations that produce the smallest normalized root mean square error (NRMSE) between (Q_{sim_annual} – annual water consumption) and Q_{obs_annual}. "

- L371- Do you know when all the reservoirs represented in this study were built? Presumably, pre 1971 but if not the calibration period could be calibrating the model to an unimpacted time series.

→ We greatly appreciate your insightful question regarding the construction years of the reservoirs represented in our study. We investigated this within the GRanD database, specifically for dams in the 91 calibrated basins. We found that approximately 69.5% of GRanD dams located within these basins were constructed before 1971. Furthermore, around 17.2% were built between 1971 and 1981. Given that our calibration period spans 1971-1981 and our objective was to reasonably capture the streamflow simulation during the validation period of 1981-1991 with the calibrated parameters, we considered it appropriate to include all dams (i.e., 100% of them) in the calibration period regardless of construction year. Although applying them incrementally over time might provide a more nuanced perspective, this approach significantly complicates the modeling process, as it would require creating new aggregated reservoirs each year. One of the major issues we encounter when considering starting the calibration from a period where 100% of the dams were built, was the scarcity of sites with continuous observations far downstream of the basin. We concur with your point that the construction timeline of the reservoirs is a critical aspect. In future research, we aspire to incrementally incorporate the dams, as opposed to applying them all from a uniform starting period. Based on your comments, we added a discussion to the manuscript to clarify this on Line 394-400 (see below).

"The construction years of these dams pose a critical factor in deciding the calibration-starting year. Here, we found that ~ 69.5% of these dams were constructed before 1971 and additional ~17.2% were built between 1971 and 1981. We considered it reasonable to include all dams, regardless of their construction year, in the calibration starting from 1971 mainly for two reasons: i) incrementally aggregating dams built during this period over time, in addition to the dams built before 1971, would significantly complicate the modeling process, and ii) dams constructed before 1981 account for approximately 84% of the total storage within these basins."

- L332 – this seems like a really useful metric, but could you give some example units for X_m and ? Initially it wasn't clear to me that would be cumulative.

→ We appreciate your detailed feedback and agree that clarity regarding the units for Xm and $\bar{X}$ is essential for understanding this metric. As the equation takes into account the annual mean, which is computed as a sum over the 12 months rather than an average, it is appropriate for the unit to represent depth or volume over a specific duration, as opposed to a rate such as m^3/s. To address this point, we have introduced an additional sentence in our manuscript to provide further clarification, see L330 copied below. Thank you for pointing out this area for improvement.

"Where X_m is the mean monthly value for the month m and $\bar{X}$ is the annual mean value. SI ranges between 0 and 1.833, indicating uniform distribution over the 12 months and a single-month occurrence, respectively. When applying this equation, it is recommended to avoid using flow rate units such as m^3/s , instead use units that represent a measure of water quantity over a period of time, such as depth (e.g., mm/month) or volume (e.g., m^3/month)."

- It's quite hard to distinguish between flood control and other reservoirs in figure 3a, perhaps you could consider using a different color.

→ Thank you for your feedback regarding the color differentiation in figure 3a. We agree that clear distinction is crucial for the interpretation of the data. We have therefore revised the figure and enhanced the color contrast to more clearly differentiate between flood control and other reservoirs.

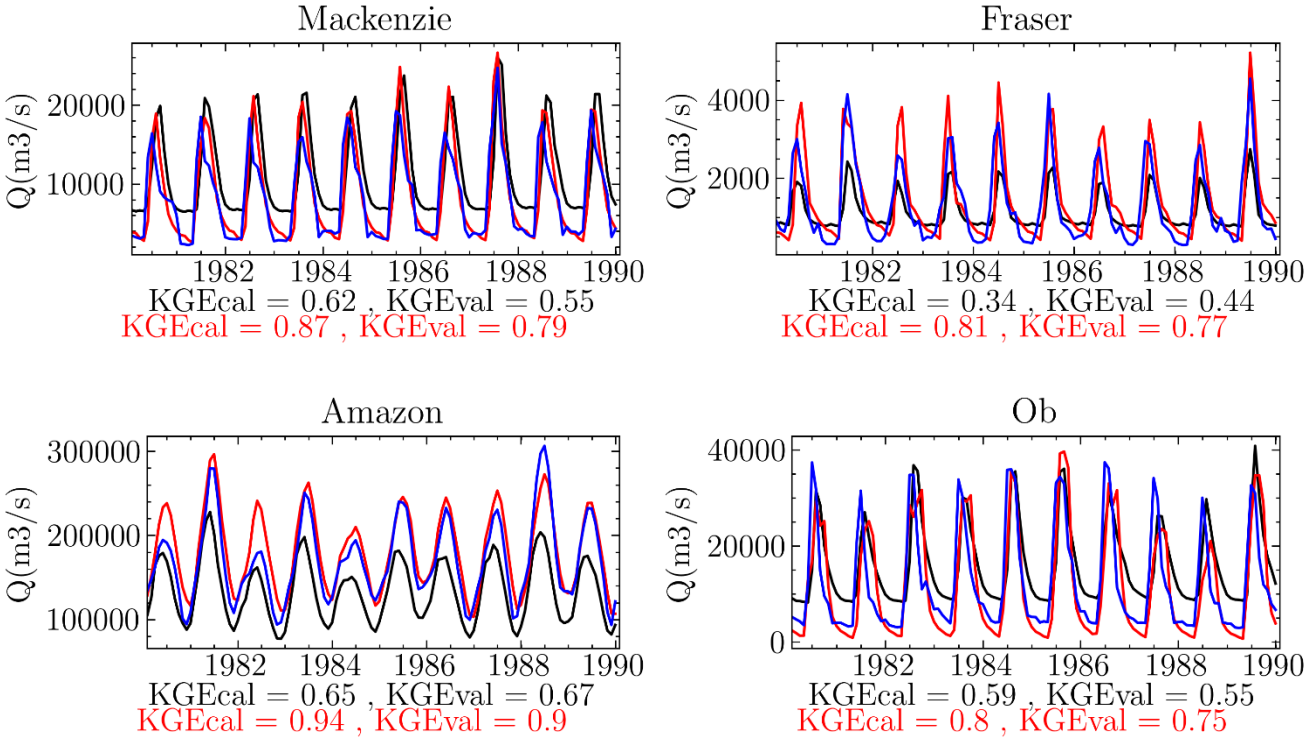
- L384- Does selecting the most downstream gauge in each basin not mean that you exclude a lot of the closer, more impacted gauges from your evaluation? It would be interesting to see whether your gains in model performance are larger at the closer, more impacted gauges.

→ Thank you for your thoughtful comment. Indeed, our selection of the most downstream gauge in each basin may bypass some of the closer, more directly impacted gauges from our evaluation. However, in order to maintain a comprehensive understanding of the basin-wide hydrological dynamics, we calibrate the model parameters considering the entire basin. While focusing on nearer, more influenced gauges could potentially reveal more significant improvements in model performance, such a localized approach may not fully encapsulate the cumulative impacts of all hydrological and management processes throughout the basin. Our decision to focus on downstream gauges was to capture a holistic representation of the basin-wide hydrological dynamics, as our model adopts uniform parameters across all grids within a basin, making downstream sites a practical and logical choice for calibration. We do agree that a more detailed analysis on the closer, more affected gauges could yield further insights. As such, we plan to incorporate this aspect into our future work, where we could calibrate at multiple locations within a single basin. This would involve a multi-objective strategy, wherein instead of solely relying on the downstream-most gage, we would seek an optimal set of parameters at the basin level that yields an annual water balance close to the observed data at each site in the basin. With that, we can evaluate our simulation against observation anywhere in the basin. In light of your comment, we added one paragraph discussing the future direction to section 4, lines 608-621 (see below).

"While our current two-stage calibration framework provides substantial insights, we anticipate its evolution towards a more comprehensive multi-gauge calibration approach. The existing framework, which relies on a single gauge per basin (typically the most downstream one), could potentially be expanded to multi-gauge calibration. Theoretically, this process would calibrate the model parameters using multiple gauges scattered throughout the basin, accommodating the spatial variability inherent in these parameters. Such an expansion could incorporate methods like the hierarchical approach and multi-objective optimization into the present two-stage framework. The hierarchical approach initiates calibration with the smaller, upstream sub-basins. The parameters determined at these stages subsequently inform the calibration of the larger, downstream basins, continuing in this fashion until the calibration of the most downstream gauge. This method capitalizes on the detailed information accessible at smaller scales, thereby assuring the consistency of large scale simulations with those on smaller scales. Incorporating multi-objective optimization, with objectives set at multiple gauges, is another approach that could augment the fidelity of the simulation within the two-stage calibration framework. This approach could mitigate discrepancies between simulated and observed discharges at multiple gauges simultaneously. Consequently, the model could represent a comprehensive array of hydrological behaviors across space, especially in large and heterogeneous basins where significant spatial variability in hydrological processes is common."

- L447-448- is the improvement in high and low flow periods one that has only been observed visually or have you quantified this?

→ Thank you for your question. The improvements in high and low flow periods that we referenced were primarily based on visual observation from simulated flows across all the basins we analyzed. While we have not quantitatively measured these enhancements, the patterns were consistently apparent in our visual data assessments. Here are examples from some of the well-known river systems from different regions (black is without and red with water management , blue is observation)



- L517- could the authors clarify how the RII has been used here? My understanding is that here the metric is only accounting for the largest hydropower reservoir in the basin, but if there were several large reservoirs, would it not be important to account for the cumulative capacity? And equally, to contextualize the RII associated with the hydropower reservoirs with the RII calculated from any reservoirs of another type.

→ We appreciate your thoughtful query. While it's true that RII is conventionally calculated based on cumulative storage, our focus in this context was to identify basins where a single, largest hydropower reservoir significantly contributes to the overall reservoir impact on flow at the GRDC site. If the RII is small, it suggests that the influence of this particular hydropower reservoir on the GRDC site flow is likely to be minimal, regardless of whether it is classified as a hydropower or flood control reservoir. Conversely, when the RII corresponding to the reservoir is substantial, how we represent its type becomes increasingly critical, especially at downstream locations.

- There isn't much discussion of the impacts of the surface water extraction on the model performance. Did this have any noticeable effect or did most of the improvements come from the reservoir operations?

→ Thank you for your thoughtful query. As per our analysis, the impact of surface water extraction on model performance was not significantly noticeable. This observation could potentially be because, for the most part, total extraction magnitudes are relatively small compared to the runoff. Thus, it appears that most of the improvements in our model performance stemmed primarily from the reservoir operations.

Technical corrections:

- L215- this sentence is accidentally repeated

→ Thank you for pointing out this oversight. The repetition of the sentence on line 215 has been corrected in the revised manuscript.