General Response to Reviewer 2

This paper describes the evaluation of five different global hydrologic models using remotely sensed reservoir storage data. To evaluate this, the authors use monthly reservoir data for 424 major dams across the globe and calculate the KGE between monthly modelled storage and monthly remotely sensed storage. They also conduct a sensitivity analysis to the parameters within H08 to address rigidity of parameter values. Ultimately, the authors find that the generic reservoir operating schemes perform poorly when evaluating the KGEs. This is due in part to the dampened intra annual variability. This demonstrates that parameters are a key limiting factor in global reservoir modelling and that satellite observations can be used to calibrate and validate reservoir schemes.

We thank the reviewer for their careful reading of our manuscript and their encouraging outlook. It pleases us that the reviewer recognised the core message of our work: that parametric rigidity is a key limiting factor in global reservoir modelling, and that satellite observations offer a viable pathway to calibrate and validate these schemes. Below, we address each of the major and minor comments in detail.

Part A: Major Comments

Reviewer Comment (Maj C 1): *My first main concern after reading the abstract is that the study only evaluated 424 dams. As these are all global hydrologic models, I would assume there is sufficient overlap between the almost 7,000 large dams in GRanD and remotely sensed reservoir storage.*

Response:

We thank the reviewer for raising this important point. The reduction from ~7000 dams to 424 dams resulted from a series of stringent quality-control filters designed to ensure robust and meaningful model evaluation. Our objective was to compare like-for-like across all five *global hydrological models* (GHMs) and the satellite-derived observation dataset.

Specifically, dams were excluded if: (i) the simulated upstream catchment area differed substantially from the reference (‘true’) value (>30% deviation); (ii) multiple dams were represented within the same 0.5° grid cell and therefore had to be aggregated, as is the norm in GHMs; (iii) they corresponded to natural lakes whose storage is only partially regulated by a dam; (iv) they were not represented consistently across all five GHMs; (v) insufficient overlap existed between simulated and satellite-derived storage records for meaningful performance assessment; or (vi) performance metrics were mathematically undefined because simulated storage exhibited no temporal variability. These filtering rules are mentioned in Sect. 2.3.3 of the original manuscript.

Following these filtering steps, 424 dams remained for further analysis. While smaller than the GRanD inventory of dams, this subset is globally distributed (see Fig. 1 in the main manuscript) and represents the largest common set of dams for which a robust and consistent evaluation could be conducted across all selected GHMs and satellite-derived observations. We have revised Sect. 2.3.3 of the manuscript to provide a clearer description of this screening process.

Action taken:

- Revised Sect. 2.3.3 to provide a detailed description of the filtering and quality-control procedures that reduced the initial inventory of 7319 dams to the final evaluation set of 424 dams.

For transparency, the filtering process can be summarised as follows:

- We started with 7319 dams from the Inter-Sectoral Impact Model Inter Comparison Project Phase 3a (ISIMIP3a).
- Reservoirs were then subjected to a catchment-area quality-control check, in which the catchment areas delineated from GHM global river network maps were compared against reference ('true') catchment areas reported in the Global Dam Watch (GDW) dataset (Lehner et al. 2024). Reservoirs with catchment-area errors exceeding 30% were excluded. This reduced the sample size to 1255 reservoirs. This filtering loss is largely a consequence of the coarse 0.5° spatial resolution used by the ISIMIP3a GHMs: at the equator, a 0.5° grid cell covers ~2500 km^2. Consequently, catchments smaller than 5000 km^2 are represented by fewer than two grid cells, making their areas poorly resolved. Fig. R1 shows that the reservoirs removed during quality control are overwhelmingly associated with small catchment areas, i.e. local reservoirs.
- 1106 dams after aggregation of multiple dams within the same 0.5° grid cell.
- 1071 dams after excluding reservoirs that do not primarily impound river flow.
- 555 dams common to all five GHMs.
- 484 dams after removing cases with insufficient overlap between simulated and satellite-derived observed storage records.
- 424 dams after excluding cases where performance metrics were undefined due to zero storage variability in at least one simulation (Fig. R2).

Consequently, the final set of 424 dams represents the largest robust and consistently comparable sample across all five GHMs and the satellite-derived observational dataset.

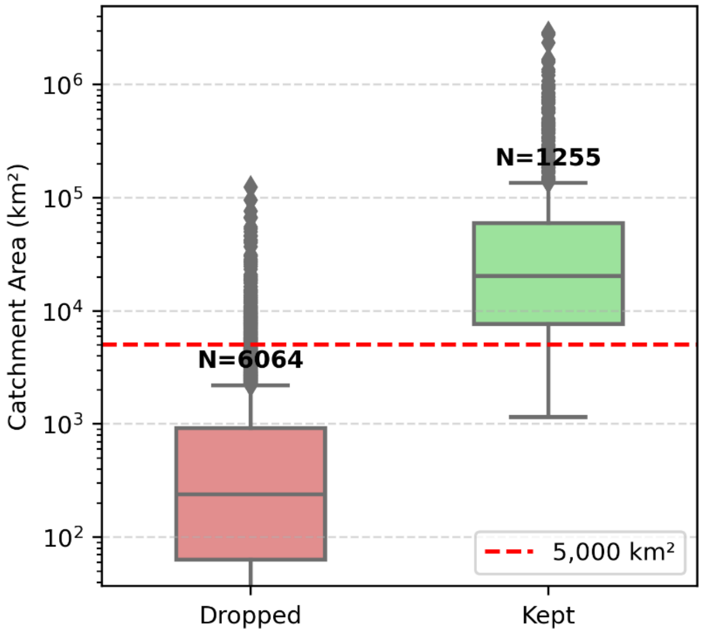

Figure R1: Distribution of GDW reported catchment areas for reservoirs excluded (6064 reservoirs) and retained (1255 reservoirs) following catchment-area quality control. The red dashed line indicates a catchment area of 5000 km², approximately equivalent to two 0.5° grid cells at the equator. Reservoirs excluded from the analysis are predominantly associated with small catchment areas below this threshold, reflecting the difficulty of accurately representing dam catchment areas at the coarse half degree spatial resolution. In contrast, retained reservoirs generally have substantially larger catchment areas and are therefore more reliably resolved at half degree grid size river network map. Boxes show the interquartile range (IQR), horizontal lines denote median catchment area, and whiskers extend to 1.5 × IQR.

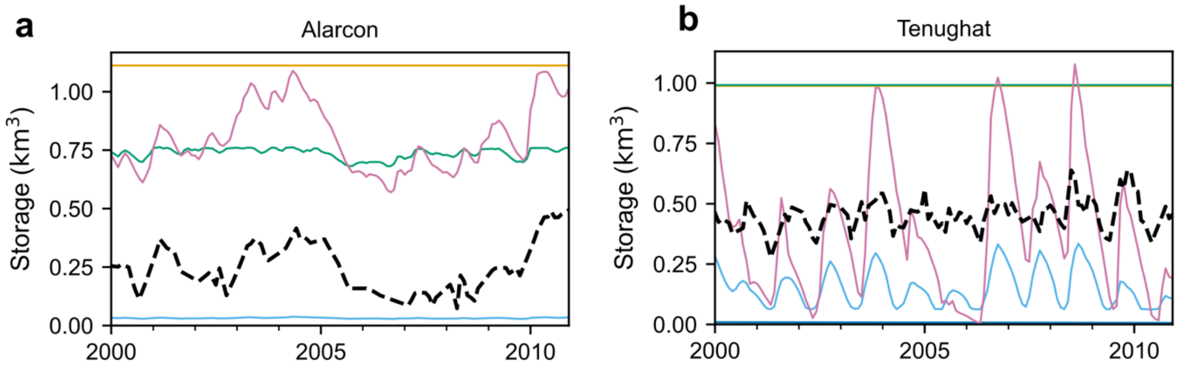

Figure R2: Examples of dams that were dropped due to flatlining simulated storage time series. (a) shows a case of flatlining at Alarcon dam for H08 (orange line), while panel (b) shows a case of flatlining at Tenughat Dam for H08 (orange line) and MIROC-INTEG-LAND (green line)

Reviewer Comment (Maj C 2): *Line 119-120: Averaging across the different forcings is totally valid, however, I would be interested as to which forcings the authors chose to average across and why these were chosen over running a new simulation with forcings that better represent the current climatic conditions (i.e W5E5).*

Response:

We thank the reviewer for this question and for the opportunity to clarify our choice of climate forcings. In this study, we evaluate reservoir storage simulations from GHMs participating in the ISIMIP3a framework. Consequently, we analyse the official ISIMIP3a simulation outputs produced by the respective modelling teams under the three ISIMIP3a standard forcings that were available across all the five GHMs that submitted their simulated storage time series to ISIMIP3a: GSWP3-W5E5, 20CRv3-W5E5, and 20CRv3-ERA5. We recognise that Sect. 2.1 did not communicate this clearly.

Action taken:

- Improved the wording in Sect. 2.1 (Lines 111–113) to clarify that this study evaluates official ISIMIP3a simulations generated and published by the participating GHM teams following the ISIMIP protocol.

Reviewer Comment (Maj C 3): *Line 179: As Cooley et al 2025 already compared remotely sensed reservoir storage datasets against each other. I would be curious to know what the added benefit is of creating the GDS dataset. This could be better explained in this section by stating some of the limitations (such as temporal or spatial coverage) to the other datasets used.*

Response:

We thank the reviewer for raising this important point regarding the GDS dataset and the overall satellite-derived storage product inter-comparison we performed. Because Reviewer 1 raised a similar concern (see letter to Reviewer 1, Major Comment 1), and to avoid unnecessary repetition, we respectfully refer the reviewer to our detailed response to Reviewer 1, Major Comment 1. There, we state that we have opted to entirely remove the development of the GDS dataset and the satellite intercomparison from the revised manuscript to keep the focus strictly on GHM evaluation.

Reviewer Comment (Maj C 4): *Line 185 – 195: The selection of four dams in each continent in order to do a detailed timeseries analysis works, however, I would be more interested in a larger grouping of these dams. 4 dams out of the 7,000 to 22,000 that exist globally is quite small. I would suggest repeating this sensitivity with all the dams present in the study and then grouping based on other characteristics such as continent, main purpose, size, etc.*

Response:

We thank the reviewer for this constructive suggestion and the opportunity to clarify the original intent of this section. First, we clarify that the four dams discussed in Lines 185–195 pertained to the evaluation of satellite-derived reservoir storage against in situ storage, rather than a GHM parameter sensitivity analysis. The four time series were provided solely as illustrative examples to help readers visualise the temporal dynamics that underpinned the large-sample statistical evaluation of 638 dams (formerly presented in Fig. 2a–b).

However, as noted in our response to Major Comment 3, we have since removed the intercomparison and validation of satellite-derived storage products from the revised manuscript. Consequently, the text and figures featuring these four specific dams have been deleted to maintain a strict focus on evaluating the GHMs.

Importantly, we fully agree with the underlying spirit of your suggestion: that analyses should be conducted across a much larger sample and grouped by reservoir characteristics. As detailed in our response to Reviewer 1, Major Comment 2, we have now conducted a global parameter calibration for all 814 reservoirs explicitly represented in the H08 model. Furthermore, we have analysed the resulting optimal parameters by grouping them according to primary purpose (e.g., irrigation, hydropower) and climate zone (e.g., arid, temperate, cold). We believe this new global analysis directly fulfils the reviewer's request for a large-sample, characteristic-based evaluation within the core focus of the manuscript.

Reviewer Comment (Maj C 5): *Line 285 – 290: Perhaps this is my reading, but I would assume that based on the equations in line 285 that dams would not be removed unless they did not have observed or simulated storage. I am still a bit unsure how the authors went from 7000 dams in each model to only a little over 400. It would be good to include in Table 1, how many dams each model has as maybe I am assuming there are more per model.*

Response:

We thank the reviewer for highlighting this and allowing us to clarify. As described in our response to Major Comment 1, the reduction to 424 dams resulted from a comprehensive, multi-step quality-control pipeline rather than a single equation constraint. However, regarding the specific removal of dams based on mathematical constraints (Lines 285–290): a dam is excluded when one or more GHMs simulates a completely flatlined storage time series. In these instances, the simulated standard deviation (σ_{sim}) is zero. Consequently, applying Equation 7 in the original manuscript leads to a division by zero, yielding an infinite storage value. Furthermore, when calculating the Kling-Gupta Efficiency (KGE), the Pearson correlation coefficient (R) mathematically requires dividing by the standard deviations of both the simulated and observed datasets, as shown in Equation R1 below. If σ_{sim} were zero, R becomes undefined, which in turn renders the entire KGE undefined.

$$R = \frac{\sum_i^n (S_{obs} - \mu_{obs})(S_{sim} - \mu_{sim})}{\sqrt{\sum_i^n (S_{obs} - \mu_{obs})^2} \sqrt{\sum_i^n (S_{sim} - \mu_{sim})^2}} \quad \text{Eq. R1}$$

We have revised the text to explicitly state this mathematical constraint, so readers understand exactly why these specific dams were removed. Additionally, we agree that providing the baseline numbers per model provides helpful context, and we have updated Table 1 to display the initial number of dams represented by each GHM.

Action taken:

- Expanded the text in Sect. 2.4.1 and 2.4.3 to explicitly explain why reservoirs with zero simulated storage variability produce undefined KGE values and were therefore excluded.
- Updated Table 1 to include the initial number of dams represented by each GHM prior to filtering.

- Added a clearer description of the complete dam-filtering workflow (detailed in response to Major Comment 1).

Reviewer Comment (Maj C 6): *Line 365: Now it is stated that the GRS dataset has the same coverage as GRanD. I would highly suggest the authors explain how they went from 7000 dams in the GRS dataset to only evaluating 400 globally and then only 4-7 for the sensitivity analysis. One suggestion could be to implement the same reservoir scheme for all the global and local dams in the selected models in order to have a larger subset to work with.*

Response:

We thank the reviewer for these comments. Regarding the reduction from approximately 7000 dams represented in GRS to the 424 dams used in our evaluation, we refer the reviewer to our response to Major Comment 1, where we provide a detailed description of the quality-control and filtering pipeline that led to the final evaluation dataset. Briefly, this reduction was necessary to ensure consistent dam representation across all GHMs, reliable catchment delineation, and valid performance metric calculations.

Regarding the use of 4 dams for visualisation, we respond as follows. We respectfully note that the four dams introduced in Lines 228–236 were not intended to constitute a sensitivity analysis or a statistically comprehensive evaluation. Rather, the entire-sample assessment is already provided through the global performance distributions and composite skill score analyses based on the full 424-dam dataset (e.g., Figs. 3a and 4 in the original manuscript). The purpose of the four selected dams is different: as typical dams, they serve as illustrative case studies that allow readers to visualise how the median model behaviours identified in the global statistics manifest in individual reservoir time series. These reservoirs were selected to provide geographic diversity while exhibiting performance metrics broadly consistent with the median behaviour observed across the global evaluation set. For example, they allow readers to directly observe the muted storage variability simulated by MIL and the exaggerated seasonal drawdowns produced by H08, WGP, LPJ, and CWT.

Regarding the 7-dam sensitivity analysis and the reviewer's suggestion to apply the reservoir schemes to a larger subset, we fully agree. Reviewer 1 raised a similar concern regarding the limited sample size of the sensitivity analysis. Prompted by these similar critiques, we have expanded this analysis. Because the GRS dataset provides complete spatial coverage, we have now conducted a comprehensive global parameter calibration for all 814 reservoirs explicitly represented in the H08 model, directly fulfilling the reviewer's suggestion to work with a much larger subset. To avoid redundancy in this rebuttal, we respectfully direct the reviewer to our detailed response to Reviewer 1, Major Comment 2. There, we present the methodology, and a new discussion on how these optimal parameters relate to reservoir characteristics across the globe.

Finally, regarding the suggestion to expand the sample by applying reservoir operation schemes to local dams, we respectfully note that this is constrained by the coarse spatial resolution of the GHMs. At 0.5° resolution (the minimum catchment area is ~2500 km^2 at the Equator), local dams are associated with very small and often poorly resolved upstream catchment areas. Reliable river routing generally requires catchments spanning multiple grid cells, and the representation of local dams in current GHMs is therefore subject to substantial uncertainty. Applying sophisticated reservoir operation algorithms to such poorly resolved systems would likely introduce additional routing errors and reduce the physical interpretability of the results. Consequently, current GHM frameworks typically prioritise the explicit representation of major reservoirs with more sophisticated reservoir operation schemes while treating smaller water reservoirs (so-called local reservoirs) more simply. We have added an explanation in Section 2.3.3.

Action taken:

- Added a clarification to Section 2.3.3 discussing the spatial-resolution constraints (Oki et al., 1999; Hanasaki et al., 2018) that limit the meaningful application of complex reservoir operation schemes to small local dams within coarse 0.5° GHM spatial resolutions.
- Expanded the description of the dam-filtering workflow.
- Conducted a comprehensive global calibration for all dams represented in the H08 model (H06 and LIS schemes) using the universally available GRS dataset.

Reviewer Comment (Maj C 7): *Line 605 – 609. I agree with the findings that the parameter rigidity could be one of the key limitations in a global reservoir scheme, however I would be interested to know what the authors suggest in order to remedy this issue as it seems highly unlikely that global schemes will end up calibrating their operations to each reservoir due to data limitations. In going off of this, I would suggest combining 4.2 and 4.3.*

Response:

We thank the reviewer for this thoughtful comment and agree that the remedy for parameter rigidity is a central question for the future development of global reservoir modelling. Historically, the prevailing view has been that site-specific calibration is impractical at the global scale because suitable observations are unavailable for most reservoirs. However, one of the key implications of our study is that this assumption is becoming increasingly outdated. Satellite-derived storage products such as GRS and GDS now provide observations for all GRanD-referenced dams. Consequently, the data availability barrier that traditionally prevented large-scale, reservoir-specific calibration has been substantially reduced. Indeed, motivated in part by the reviewer's earlier comments, we have now demonstrated the feasibility of this approach by conducting an in-sample global calibration for 814 dams that are well captured in the H08 model: details of this are in our response to Reviewer 1, Major comment 2.

We also agree that Sect. 4.2 and 4.3 share a common theme regarding limitations and future directions. Accordingly, we have merged these sections to create a more streamlined discussion.

Action taken:

- Revised the discussion (around Lines 684–687) to explicitly highlight how universal coverage of all GRanD-referenced dams by GRS enables global reservoir parameter calibration and creates new opportunities for addressing parameter rigidity in GHM reservoir operation schemes.
- Combined Sect. 4.2 and 4.3 into a single section discussing model limitations and future development pathways.

Part B: Minor Comments

Reviewer Comment (Min C 1): *Abstract line 36 – 37: A KGE of -0.43 is not necessarily satisfactory but instead should be described as no change as the model is not performing better or worse than before.*

Response:

We thank the reviewer for this insightful comment. Reviewer 1 raised the same, issue regarding our interpretation and phrasing of the KGE threshold. To avoid repetition in this letter, we respectfully direct the reviewer to our detailed response to Reviewer 1, Minor Comment 1. There, we fully acknowledge this correction, and outline our transition to using the more neutral term 'informative.'

Reviewer Comment (Min C 2): *Line 118: Figure 1 is referenced a few pages before it shows up. I would suggest keeping figures close to where they show up in the manuscript as it's easier for the reader to see.*

Response:

We agree with the reviewer's suggestion to improve the reader's experience. To avoid the issue of referencing figures before they appear, we have removed this early reference, which we also believe improves the overall flow of the text.

Action taken:

- Removed the premature reference to Fig. 1 in the text to ensure figures appear closer to where they are first mentioned.

Reviewer Comment (Min C 3): *Figure 3 and 4: I really like the colors and the graphics that are shown in Figures 2 , 3 and 4. That said, I would be interested to know if the dashed lines in Figure 4 mean the same thing as the dashed lines in the other figures. If they are also thresholds, it could be interesting to explain how the thresholds are calculated or chosen.*

Response:

We thank the reviewer for appreciating our figures and for requesting clarification on the meaning of the dashed lines in the box plots. Reviewer 1 raised the same question regarding these thresholds. To avoid unnecessary repetition, we respectfully direct the reviewer to our response to Reviewer 1, Minor Comment 6. Briefly, we realised that these lines—which were initially intended to serve as informal visual markers for a 'very good' simulation—lacked a rigorous mathematical derivation for the individual KGE components (alpha, beta, and R). To prevent any future confusion, we have completely removed the orange dashed lines from the component box plots in Fig. 4 and all other similar figures.

Reviewer Comment (Min C 4): *Line 410: This result is a key takeaway and I like the explanation here surrounding the methodology but also the explanation of what it means.*

Response:

We thank the reviewer for their positive feedback. We are pleased that the explanation of these results and their methodological implications resonated well.

Reviewer Comment (Min C 5): *Figure 6,7,8: I think these figures also tell a nice story, but it could be good to try and combine them into one as then we can compare across dams. I also wonder if a heat map is a good way to look at the KGE scores per parameter value. Perhaps it could be nice to look at a line plot or a box plot in order to see the ranges?*

Response:

We appreciate the reviewer's suggestion to explore alternative visualisations for the parameter sensitivity analysis. We carefully considered using line or box plots. However, we concluded that, so far, heatmaps remain the most appropriate format for this specific analysis. This is because the objective is to evaluate the combined influence of the Target Storage Level (TSL) and Degree of Regulation Threshold (DORT) parameters. Changes in one parameter affect the response to the other, creating a two-dimensional parameter space. Heatmaps allow this interaction to be visualised directly, whereas line or box plots would require collapsing one dimension and would therefore obscure important parameter trade-offs.

Regarding the consolidation of Fig. 6, 7 and 8, we kindly refer the reviewer to our response to Minor comment 8 by Reviewer 1. It describes the consolidations we adopted for Fig. 8 to 14

Reviewer Comment (Min C 6): *Figures 9, 10, 11. I would also suggest combining these into one figure in order to allow for a better comparison. Ultimately, I would suggest combining all figures related to different analyses if the individual dams into one figure per analysis type. This way it is easier for the reader to compare across the different dams and evaluations.*

Response:

We appreciate the reviewer's suggestion to consolidate the figures to facilitate cross-dam comparison. Our response to this comment is mostly consistent with the response to the comment prior. While we agree that grouping figures by analysis type would make comparison among reservoirs more direct, we believe that the current structure—grouping inflow, storage, and release information for each individual dam—is more effective for the objectives of this analysis. Our primary aim is to show how inflow conditions propagate through the reservoir operation scheme to influence storage dynamics and release behaviour. Presenting these variables together allows readers to directly follow this sequence for each reservoir and better understand the relationships among the simulated inflow, storage, and release responses. Reorganising the figures by analysis type would separate these interconnected variables and make it more difficult to interpret the reservoir-specific responses that are central to the discussion.

However, a compromise must be made for the sake of compactness and efficiency. Again, we kindly refer the reviewer to our response to Minor comment 8 by Reviewer 1. It describes the consolidations we adopted for Fig. 8 to 14.

References

Cooley, Sarah W., Jida Wang, Huilin Gao, et al. 2025. “Global Intercomparison of Satellite-Derived Variability in Reservoir Storage.” Environmental Research Letters 20 (8): 084035. https://doi.org/10.1088/1748-9326/ade903.

Harrigan, Shaun, Ervin Zsoter, Hannah Cloke, Peter Salamon, and Christel Prudhomme. 2023. “Daily Ensemble River Discharge Reforecasts and Real-Time Forecasts from the Operational Global Flood Awareness System.” Hydrology and Earth System Sciences 27 (1): 1–19. https://doi.org/10.5194/hess-27-1-2023.

Knoben, Wouter J. M., Jim E. Freer, and Ross A. Woods. 2019. “Technical Note: Inherent Benchmark or Not? Comparing Nash–Sutcliffe and Kling–Gupta Efficiency Scores.” Hydrology and Earth System Sciences 23 (10): 4323–31. https://doi.org/10.5194/hess-23-4323-2019.

Lehner, Bernhard, Penny Beames, Mark Mulligan, et al. 2024. “The Global Dam Watch Database of River Barrier and Reservoir Information for Large-Scale Applications.” Scientific Data 11 (1): 1069. https://doi.org/10.1038/s41597-024-03752-9.

Li, Yao, Gang Zhao, George H. Allen, and Huilin Gao. 2023. “Diminishing Storage Returns of Reservoir Construction.” Nature Communications 14 (1): 3203. https://doi.org/10.1038/s41467-023-38843-5.

Yoshida, T., N. Hanasaki, K. Nishina, J. Boulange, M. Okada, and P. A. Troch. 2022. “Inference of Parameters for a Global Hydrological Model: Identifiability and Predictive Uncertainties of Climate-Based Parameters.” Water Resources Research 58 (2): e2021WR030660. https://doi.org/10.1029/2021WR030660.