General Response to Reviewer 1

This manuscript examines two generic reservoir operation schemes implemented in five global hydrological models (GHMs). The authors first evaluate the storage simulation performance of five GHMs against satellite-derived reservoir storage products, using the Global Reservoir Storage (GRS) dataset as the primary benchmark across 424 dams globally. They then test two simple, post-hoc approaches: variance-matching and linear bias correction, to determine whether satellite-derived storage data can recover simulation skill, and conduct a parameter sensitivity analysis on a small number of dams. The study demonstrates that satellite-derived products can be used both for statistical correction and for parameter tuning of generic reservoir operation schemes, improving GHM reservoir simulation performance.

While the topic is relevant to this journal, I have serious concerns about the scientific contribution of the study in its current form. Specifically, three major issues critically affect the quality of the work:

We thank the reviewer for their thorough, critical and constructive assessment. We are encouraged by their recognition that our study addresses a highly relevant topic for the global hydrological modelling community. Before addressing the specific comments, we wish to clarify the original intent of our manuscript, while also highlighting a major new analysis added in response to the review.

Historically, the global hydrological modelling community has lacked the observed reservoir storage data to systematically diagnose how these models handle it globally. Consequently, the original objective of our manuscript was inherently diagnostic: to identify, quantify, and understand the widespread, systematic errors present in current generic reservoir operation schemes (GROS). We acknowledge that our original framing may have inadvertently set the expectation of a fully scalable calibration solution, rather than the diagnostic baseline we intended to provide.

However, we fully agree with the reviewer's assessment that demonstrating a broader, global-scale calibration substantially elevates the scientific contribution of the study. Therefore, rather than merely clarifying our original intent, we have acted directly on the reviewer's suggestion. Because the Global Reservoir Storage (GRS; Li et al., 2023) dataset provides observations for all dams represented in the ISIMIP3a GHMs, we have now conducted a global parameter calibration for all reservoirs explicitly represented in the H08 model (expanding from the three illustrative reservoirs in the original manuscript). We believe this increase transforms the manuscript from a purely diagnostic study, directly addressing the reviewer's overarching concern.

We note one important caveat regarding this new global calibration: because corresponding in situ inflow and release time series are scarce globally, this global calibration optimises solely for reservoir storage. Consequently, we cannot systematically evaluate the trade-offs between optimising state (storage) and flux (release) variables at the global scale—a dynamic we explore in the focused 3-dam sensitivity analysis (Sect. 3.5 and 4.3; Fig. 6 to 14 in the original manuscript). We have explicitly noted this limitation in the revised text.

Below, we address the reviewer's specific concerns point-by-point.

Part A: Major Comments

Reviewer Comment (Maj C 1): *Too much unrelated and unnecessary content. The two central research questions of this study are: (1) How accurate are generic reservoir operation schemes in GHMs globally? and (2) Can reservoir simulation performance be improved by correcting against satellite products and by replacing globally uniform parameters? Substantial content does not serve these questions and could be removed, including the intercomparison of satellite-based storage products (Sect. 2.3.1 and 3.1) and the development of the new GDS dataset (Sect. 2.3.2). Previous research, such as Cooley et al. (2025), already performed a comprehensive global intercomparison of five satellite-derived reservoir storage products... Moreover, the newly developed GDS dataset performs worse than the existing GRS product (Sect. 3.1), so introducing it here does not appear necessary.*

Response:

We thank the reviewer for this sharp and constructive observation. After careful reflection, we now fully agree that the intercomparison of satellite-derived storage products and the derivation of the Global Dam Storage (GDS) dataset might indeed distract the reader from the core objectives of the manuscript: to evaluate and find ways to improve the simulation of reservoir storage in *global hydrological models* (GHMs).

Following the reviewer's guidance, we have entirely removed the satellite product intercomparison and the GDS derivation from the manuscript. Instead, we now directly introduce the Global Reservoir Storage (GRS) dataset as our observational benchmark. We explicitly cite the comprehensive intercomparison by Cooley et al. (2025) to justify the selection of GRS, as it has already been established as a highly robust product for global applications.

We believe that this structural change has streamlined the manuscript, allowing us to maintain focus on the performance of *generic reservoir operation schemes* (GROS). Furthermore, the space freed up by this removal will be filled by the new global parameter calibration, which we believe directly elevates the scientific contribution of the paper.

Actions taken:

- Removed the satellite-derived storage product validation (Sect. 3.1). We have also excluded the derivation of GDS (Sect. 2.3.2) from the manuscript. Section 2.3.1 is now focussed on GRS and its characteristics.
- Revised the Methodology section (Sect. 2.3.1) to directly introduce GRS as the sole observational benchmark.
- Removed all supplementary figures and text related to the GDS dataset and GloLakes.
- Cited Cooley et al. (2025) to provide a strong, justification for selecting GRS over other available satellite-derived datasets.

Reviewer Comment (Maj C 2): *The attempt to use satellite-derived products to improve GHM performance is too simple and lacks further exploration. Testing simple correction methods with satellite data is a reasonable first step, but doing so introduces a satellite-data requirement for each reservoir... I would suggest the authors calibrate the key parameters against satellite products across a larger sample of reservoirs, and then examine whether the resulting optimal parameter values relate systematically to reservoir characteristics. Such relationships could then be transferred to reservoirs lacking satellite coverage...*

Response:

We thank the reviewer for this forward-looking comment. We agree that there is room for improvement in the extent to which we used GRS to improve reservoir storage simulations in our manuscript as highlighted in the original manuscript (Lines 684–685). We shall attempt to tackle the main issues raised by the reviewer one-by-one.

First, on the attempt to use satellite-derived products to improve GHM performance being too simple. We shall start by clarifying the distinct roles of the 'simple bias correction' (hereafter bias-correction, as defined in our manuscript) and sensitivity-analysis experiments presented in the original manuscript as initially intended. The global bias-correction analysis was not intended as a practical modelling solution. Rather, it was designed as a diagnostic experiment to quantify the extent to which errors in mean bias and variability contribute to reservoir storage simulation errors (Lines 278 to 281 in the original manuscript). Technically, this is possible for all dams in ISIMIP3a GHMs since GRS observations are available for all of them. However, we note that such corrections are not a viable final solution within GHMs because they may collapse the water balance. Next, we clarify the parameter sensitivity experiments. Rather than establishing globally applicable parameter values, the 3-dam detailed sensitivity analysis was designed to diagnose model behaviour under contrasting inflow conditions (Sect. 2.4.4). The selected reservoirs represent cases with relatively accurate inflows (John H. Kerr Dam), substantial inflow overestimation (Glen Canyon Dam), and substantial inflow underestimation (Libby Dam). By examining these contrasting situations, we sought to understand how parameter adjustments influence release-storage dynamics and the extent to which parameter tuning compensates for errors originating from the hydrological forcing. This distinction is important because improvements obtained through parameter tuning are not necessarily indicative of improved reservoir operation realism. In generic reservoir operation schemes (GROS), parameter optimisation often compensates for upstream inflow biases, creating trade-offs between the simulation of storage states and release fluxes (Lines 613 to 618 in Sect. 4.3). However, on the other hand, when inflow was simulated fairly accurately, we were able to see that H06 had some physical meaning: for example, at John H. Kerr Dam, the optimal *degree of regulation threshold* (DORT) parameter was the one that characterised the dam as leaning more towards a run-of-the-river system, which is consistent with observation. To ensure this distinction is explicitly clear to readers, the following statement (in quotation marks) will be added to Sect. 2.4.4, starting from Line 331: "The

purpose of the analysis was therefore not to demonstrate that calibration improves performance—an outcome that is indeed expected—but rather to examine the mechanisms through which those improvements arise and the limitations of using parameter tuning to compensate for biased inflows."

Second, regarding the need for further exploration across a larger sample size, we fully agree that a broader, global-scale calibration of GROS is necessary. Prompted by this valuable suggestion, we have now conducted a global calibration of key reservoir operation parameters for 814 global dams represented in H08. Because the GRS dataset provides complete spatial coverage for all dams represented in ISIMIP3a GHMs, we were able to calibrate every global reservoir in H08.

As shown in the newly added figures (Fig. R1 and R2), the global calibration yields substantial improvements in in-sample calibration performance, providing strong evidence that parametric rigidity—the reliance on globally uniform parameters—is a major contributor to storage simulation errors in GHMs . For the H06 scheme, the median KGE improved from an uninformative value of -1.28 to an informative -0.37, exceeding our threshold for informative simulations (KGE = -0.41). In contrast, the LIS scheme exhibited a more modest improvement, with the median KGE increasing from -0.86 to -0.57. While we acknowledge that these improvements represent in-sample storage simulation fit rather than out-of-sample predictive skill, the accompanying spatial analysis demonstrates that these gains are widespread rather than being confined to a few reservoirs. In particular, calibration substantially increases the number of reservoirs exhibiting informative KGE values across many regions of the world (improvement from 280 to 414 dams for H08-H06 and from 333 to 377 for H08-LIS), including notable improvements in parts of North America, Europe and Africa, particularly in the Nile Basin. Nevertheless, a subset of reservoirs continues to exhibit uninformative performance even after calibration, highlighting that parameter optimisation alone may not fully eliminate simulation errors across all reservoirs.

To better understand the source of these improvements, we decomposed KGE into its constituent components. We found that optimisation of parameters such as DORT and TSL (for H06) and LN (for LIS) primarily improved performance by reducing excessive storage variability (α) and correcting systematic volume bias (β; see Fig. R1). In contrast, the temporal correlation component (R) was largely insensitive to calibration, with the median value remaining approximately 0.21 across all model configurations. This behaviour is consistent with the fundamental reservoir water balance. Although parameters such as DORT can modify the strength of the release response to inflow variability, the temporal pattern/evolution of reservoir storage remains strongly influenced by the timing of inflows. Consequently, parameter calibration can effectively damp excessive storage fluctuations and correct systematic volumetric biases, but it has limited ability to rectify timing errors inherited from inflow simulations. This may explain the substantial improvements in α and β relative to the near-constant values of R before and after calibration. This result suggests that further improvements in storage correlation are likely to require improvements in simulated inflows and/or reservoir operating rules rather than parameter optimisation alone.

Third, regarding physical interpretation and systematic relationships, we did not stop at direct calibration, rather, we subsequently examined whether the resulting optimal parameter values relate systematically to reservoir purpose and climate (Fig. R3, R4 and R5). The analysis yielded contrasting results across parameters:

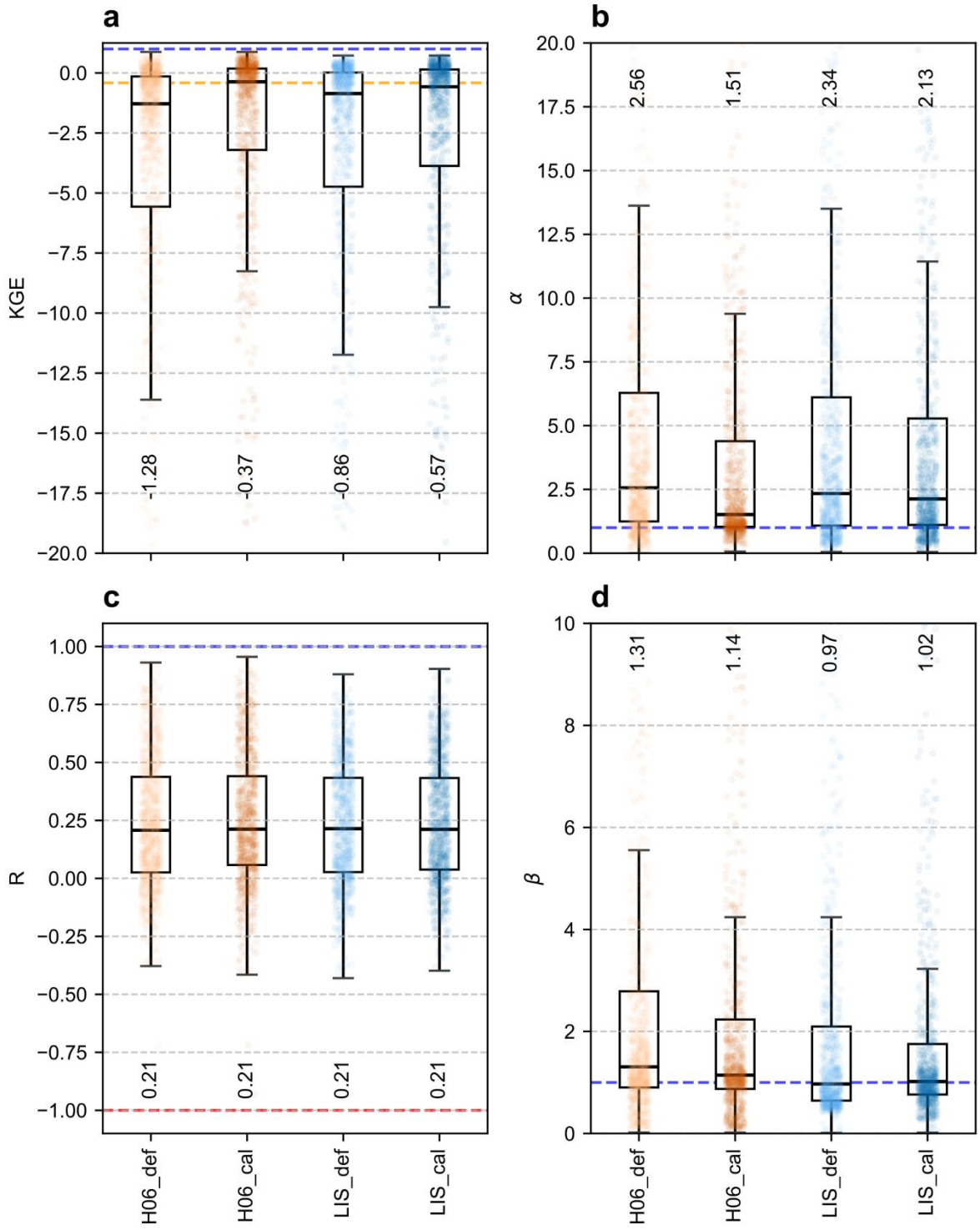
- Across all climate zones, the optimal TSL parameter commonly clustered at the calibration bounds (0.65 and 0.95), away from the default value of 0.85, with a substantially larger proportion of

reservoirs converging to the lower bound of 0.65. The predominance of low optimal TSL values suggests that the default H06 reservoir scheme (TSL = 0.85) tends to simulate reservoirs as operating too full.

- In contrast, the calibrated DORT (H06) and LN (LIS) parameters did not reveal robust global relationships with reservoir purpose or Köppen–Geiger climate class. Parameter distributions were highly heterogeneous and showed substantial overlap among categories. This lack of a strong systematic pattern suggests that globally transferable parameter values for DORT and LN may be difficult to infer from reservoir purpose and climate alone.

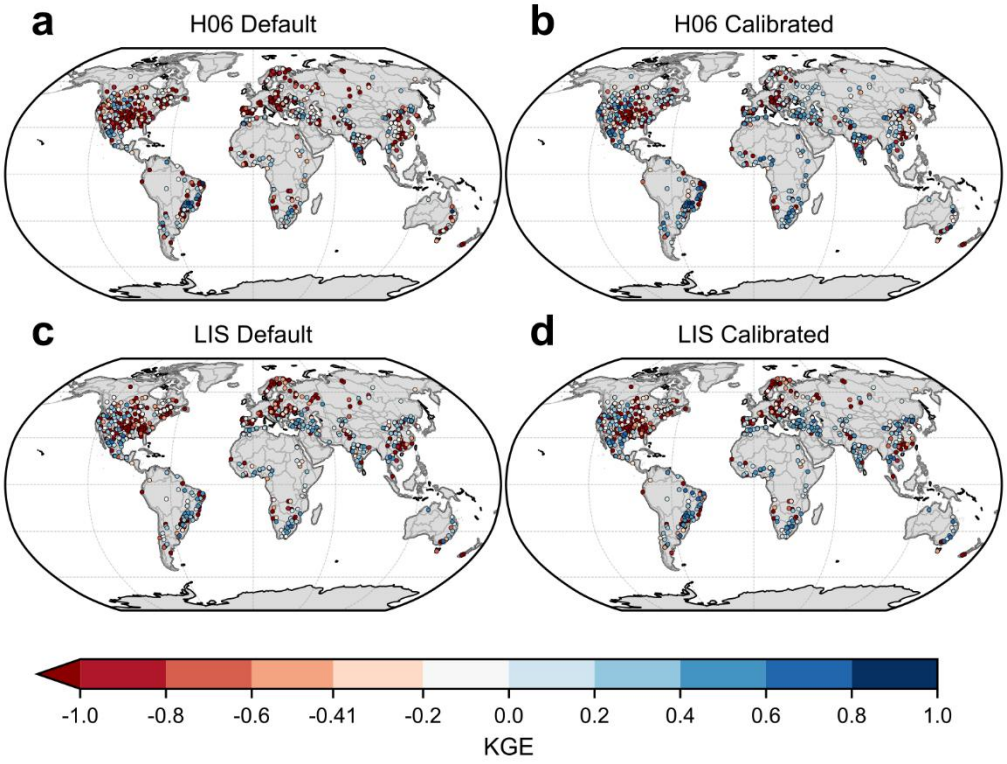
Taken together, these findings indicate that while systematic adjustment of TSL may offer a promising pathway for improving generic reservoir operation schemes, broader regionalisation of DORT and LN based solely on climate and reservoir purpose appears challenging.

We believe these new analyses substantially strengthen the manuscript by moving beyond the original proof-of-concept demonstrations and providing a global assessment of how satellite-derived reservoir storage observations can be used to improve reservoir simulations within GHMs. We have added a comprehensive presentation and discussion of these results to the revised manuscript.



"Figure R1: Performance evaluation of reservoir storage simulations from H08 against GRS for 814 dams. Panels show the distributions of (a) KGE, (b) variability ratio (α), (c) Pearson's correlation coefficient (R) and (d) bias ratio (β) for four model configurations: H08-H06 with default global parameters (H06_def; DORT = 0.5, TSL = 0.85), H08-H06 with calibrated parameters (H06_cal), H08-LIS with default global parameters (LIS_def; LN = 0.5), and H08-LIS with calibrated parameters (LIS_cal). Boxplots summarise the distribution across all reservoirs, with overlaid markers representing individual dams. Median values are indicated by black horizontal lines and annotated numerically. Boxes represent the interquartile range (IQR), and whiskers extend to 1.5 × IQR. In panel (a), the

orange dashed line denotes the threshold for informative simulations (KGE = −0.41), while the blue dashed line represents the optimum value (KGE = 1). In panels (b) and (d), the blue dashed line indicates the ideal value of α = 1 and β = 1, respectively. In panel (c), the blue dashed line indicates perfect positive correlation (R = 1), and the red dashed line indicates perfect negative correlation (R = −1).”



“Figure R2: Spatial distribution of KGE for reservoir storage simulations from H08 compared with the GRS at 814 dams worldwide. Results are shown for (a) H08–H06 with default global parameters (DORT = 0.5, TSL = 0.85), (b) H08–H06 with calibrated parameters, (c) H08–LIS with default global parameters (LN = 0.5), and (d) H08–LIS with calibrated parameters. Markers denote reservoir locations and are coloured by KGE, ranging from −∞ (dark red) to 1 (dark blue). The threshold for informative simulations (KGE = −0.41) is indicated in the colour scale.”

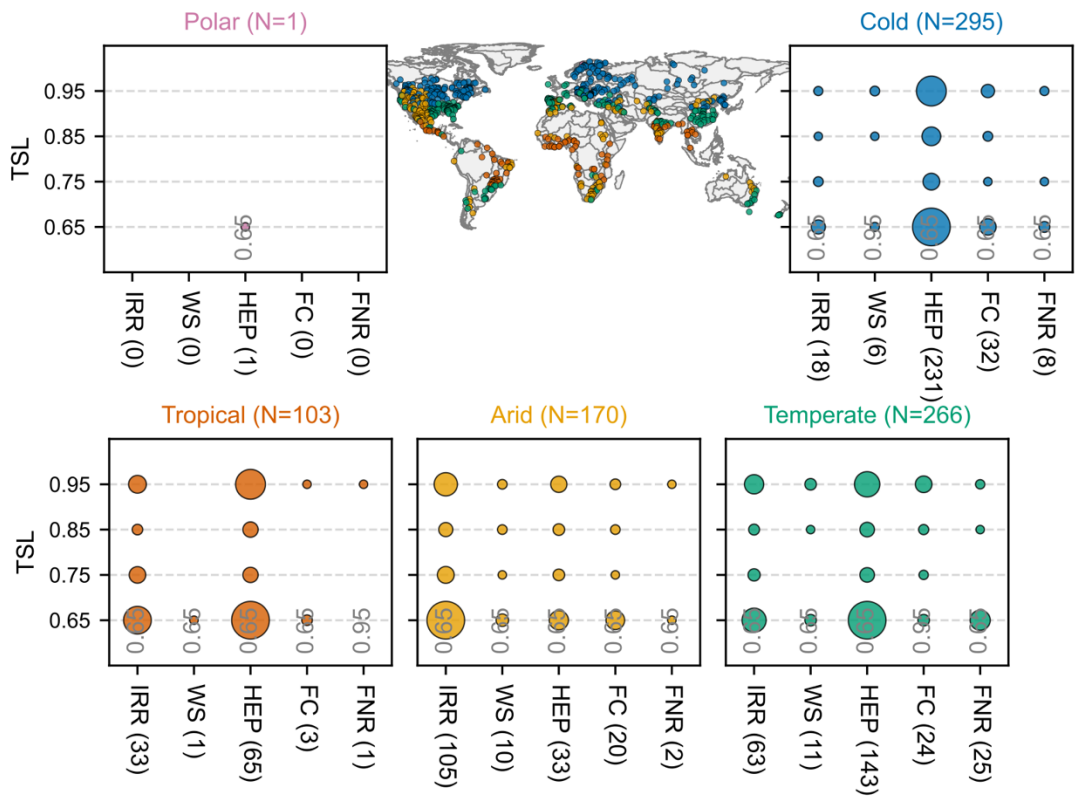
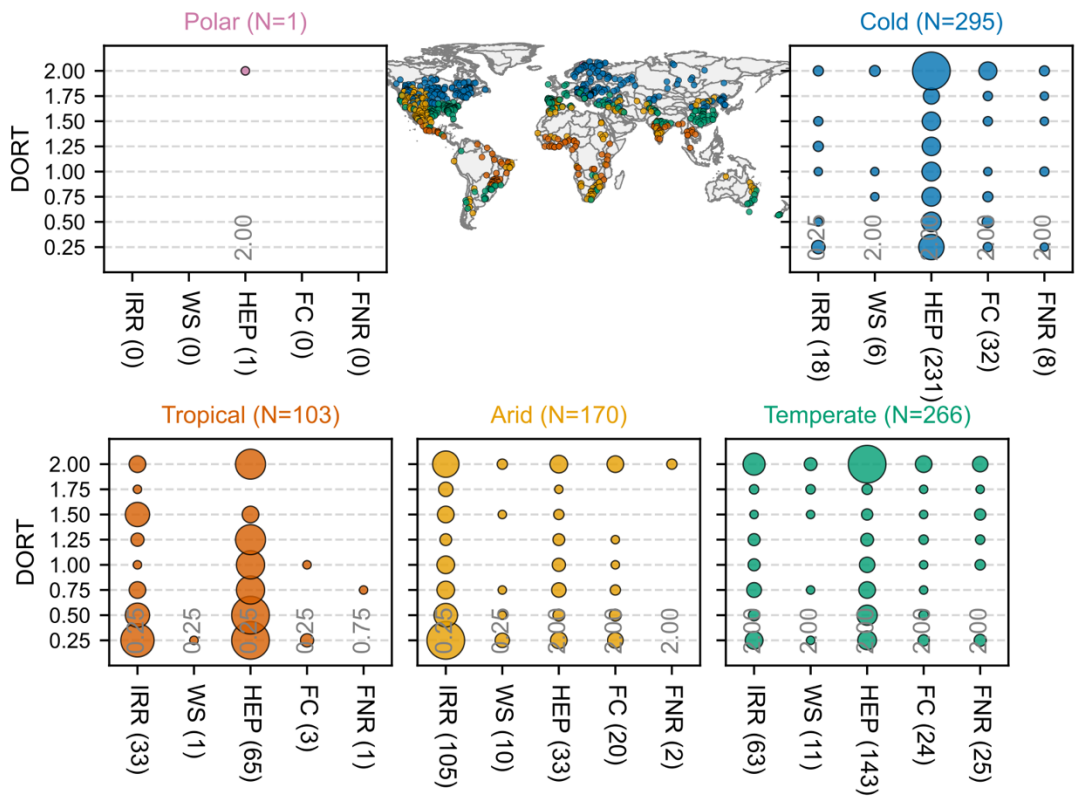


Figure R3: **Spatial and reservoir operation distribution of calibrated TSL parameters. The central map illustrates the global locations of the evaluated dams, colour-coded according to Tier 1 Köppen-Geiger climate zones (Beck et al., 2023). The surrounding bubble plots detail the distribution of calibrated TSL values, disaggregated by climate zone and the primary purpose of the reservoir: Irrigation (IRR), Water Supply (WS), Hydropower (HEP), Flood Control (FC), and Fishing/Navigation/Recreation (FNR). The area of each bubble is proportional to the number of dams sharing that specific calibrated parameter value. The total number of reservoirs (N) evaluated is indicated for each overarching climate zone and within parentheses for each specific purpose category.**



195

Figure R4: **Same as Fig. 1 but for DORT.**

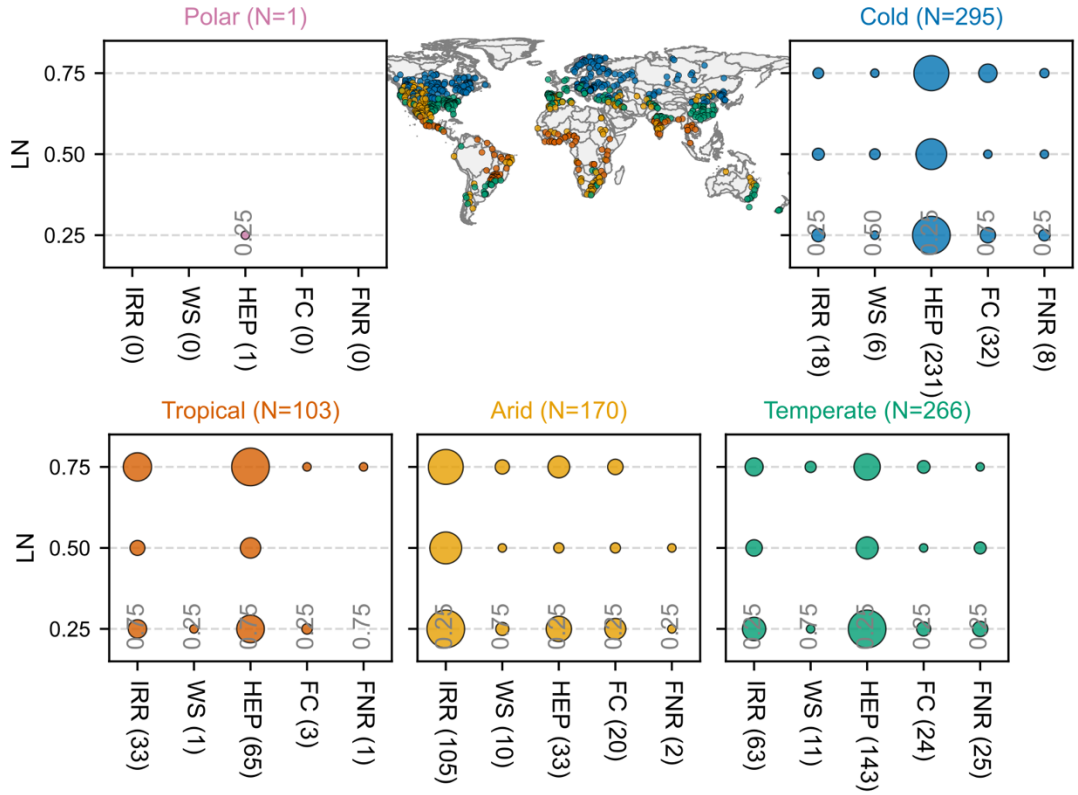


Figure R5: **Same as Fig. 1 but for LN parameter.**

Action taken:

- Left as is:
  - Sect. 2.4.3 already frames the global bias-correction experiment as a diagnostic assessment of the contribution of mean bias and variability errors to reservoir storage simulations.
  - Sect. 2.4.4 also already states that the 7-dam sensitivity analysis was a targeted investigation of how parameter tuning influences release-storage dynamics under contrasting inflow error.
- On the sample size for sensitivity analysis being too small:
  - Revised Sect. 2.4.4 to explicitly mention "The purpose of the analysis was therefore not to demonstrate that calibration improves performance—an outcome that is indeed expected—but rather to examine the mechanisms through which those improvements arise and the limitations of using parameter tuning to compensate for biased inflows."
  - Conducted a comprehensive global calibration for all 814 dams represented in the H08 model (incorporating both the H06 and LIS schemes) using the universally available GRS dataset. We have since added comprehensive global calibration results for all dams represented in the H08 model (H06 and LIS schemes) in Fig. R1 to R5, expanding the parameter tuning analysis from a small set of illustrative case studies to the full available reservoir subset in H08.

Reviewer Comment (MC 3): *Too few reservoir samples for the parameter tuning tests. The detailed sensitivity analysis is conducted on only three dams, with four more dams added only as a supplementary check... Even combined, seven dams out of 424 remain far too few to support general conclusions about suggested parameter ranges. It is easy to imagine that allowing adaptive, reservoir-specific parameters would improve simulation performance even without running these tests. So the tuning experiments on three dams mostly confirm an already-expected result.*

Response:

We thank the reviewer for raising this important point concerning the sample size of dams for the detailed sensitivity test. Because Major Comment 2 raised a similar concern, and to avoid unnecessary repetition, we respectfully refer the reviewer to our detailed response to Major Comment 2. There, we state that: "The purpose of the analysis was therefore not to demonstrate that calibration improves performance—an outcome that is indeed expected—but rather to examine the mechanisms through which those improvements arise and the limitations of using parameter tuning to compensate for biased inflows."

Part B: Minor Comments

Reviewer Comment (Min C 1): *Abstract, Lines 36–37: A median KGE>-0.41 should not be described as "satisfactory" performance. Knoben et al. (2019) defined KGE = -0.41 as the point at which a model performs no worse than a naive long-term mean-flow benchmark (analogous to NSE = 0), not as an indicator of genuinely good or skilful performance. The repeated use of "skilful" and "satisfactory" around this threshold throughout the manuscript should be reconsidered.*

Response:

We thank the reviewer for this insightful comment. We fully understand that KGE > -0.41 indicates the point at which a model provides more explanatory power than a long-term observed mean benchmark (Knoben et al., 2019). We also now agree that terminology like 'skilful' or 'satisfactory' may carry connotations of high accuracy that are too strong for this baseline. Accordingly, we have revised the manuscript to replace these terms with 'informative' when referring to performance above this threshold.

While we fully embrace this change in terminology, we also felt it was important to contextualise for the reader why exceeding this naïve mean benchmark remains a meaningful milestone in global-scale reservoir modelling. As GHMs are constrained by coarse resolutions, reliance on generic operational rules, and uncertainties in global climate forcings, outperforming a naïve mean benchmark is recognised as a valuable achievement in global hydrology (e.g., Harrigan et al., 2023; Yoshida et al., 2022). By adopting the more neutral term 'informative', our goal is to align with this reality of global hydrology without overstating the models' performance.

Action taken:

- Replaced the terms 'skilful' and 'satisfactory' with 'informative' throughout the manuscript when referring to KGE > -0.41.
- Clarified the rationale for choosing KGE > -0.41 as 'informative' in Sect. 2.4.1 as:
  "We adopted KGE = -0.41 as the threshold for an informative model, representing the point at which a simulation provides more explanatory power than a naïve long-term observed mean benchmark (Knoben et al., 2019). Exceeding this baseline is a meaningful milestone in global-scale reservoir modelling, as GHMs must navigate coarse spatial resolutions, rely on generic operational rules applied uniformly across diverse global contexts and must account for considerable uncertainties in global climate forcings. The significance of outperforming a mean benchmark is well-recognised in recent global hydrology literature, including Harrigan et al. (2023) and Yoshida et al. (2022). In this study, we elect to use the term 'informative' to denote simulations that successfully out-predict the naïve mean without overstating accuracy achieved."

Reviewer Comment (Min C 2): *Lines 167–193: Consider exchanging the order of Sect. 2.3.1 (Satellite-based storage observations) and Sect. 2.3.2 (Development of the GDS dataset).*

Response:

We thank the reviewer for this suggestion, which would have indeed improved the flow of the original draft. However, as detailed in our response to Major Comment 1, we have entirely removed the intercomparison of satellite-based storage products and the development of the GDS dataset to streamline the manuscript and maintain a strict focus on GHM evaluation. Consequently, the original sections have been deleted, and the new Sect. 2.3.1 has been rewritten to directly introduce the GRS dataset as our sole observational benchmark, rendering the suggested reordering no longer applicable.

Action taken:

- Removed the original Sect. 2.3.2, as part of the broader restructuring to address Major Comment 1.
- Drafted a new, streamlined Sect. 2.3.1 dedicated exclusively to introducing the GRS dataset and justifying its use.

Reviewer Comment (Min C 3): *Table 2: This table largely duplicates information already given in Sects. 2.3.1–2.3.3 and can be removed.*

Response:

We agree with the reviewer that the information presented in Table 2 is redundant given the detailed descriptions provided in Sect. 2.3.1 through 2.3.3. To streamline the manuscript and reduce unnecessary duplication, we have removed Table 2.

Action taken:

- Removed Table 2 from the manuscript.

Reviewer Comment (Min C 4): *Lines 286–288: Clarify the threshold here for the defined KGE.*

Response:

We thank the reviewer for highlighting this ambiguity and apologise for the confusion. Our phrasing might have inadvertently implied a performance threshold, which was not our intention. By 'defined KGE,' we were strictly referring to the mathematical calculability of the metric itself, i.e., whether the output is a valid number as opposed to a 'NaN.' Specifically, if a simulated time series is flatlining (Fig. R2 in our response to Major comment 1 from Reviewer 2), yielding a standard deviation of zero, the Pearson correlation coefficient (R) cannot be computed. This mathematically renders KGE undefined.

We have revised the text at Lines 286–288 to explicitly state that we are referring to the mathematical calculability of the metric, ensuring there is no confusion with performance thresholds.

Action taken:

- Rephrased Lines 286–288 to clarify that 'defined' refers to the mathematical calculability of the KGE metric—due to non-zero standard deviation—rather than a performance benchmark.

Reviewer Comment (Min C 5): *2.4.4 (Sensitivity test): Add an overview statement clarifying why TSL and DORT (for H06) and LN (for LIS) were identified as the key parameters and selected for parameter tuning, rather than other parameters within these schemes.*

Response:

We thank the reviewer for pointing out the need for this context. We agree that explicitly justifying our parameter selection improves the transparency of the sensitivity analysis. We have added a brief statement to Sect. 2.4.4 detailing our rationale. "For the H06 scheme, TSL and DORT were selected because they are the only tuneable parameters available, and they directly control the mean bias and define the operational regime, i.e., ranging from run-of-river to heavy regulation. For the LIS scheme, LN was chosen due to its role in setting the primary operational storage targets, as well as storage variability. In contrast, LIS parameters such as LC and LF were excluded from tuning because they merely act as static boundary conditions defining dead storage and maximum safe storage limits, rather than dynamic operational drivers."

Action taken:

- Added a paragraph to Sect. 2.4.4 explicitly justifying the selection of TSL and DORT (for H06) and LN (for LIS) for parameter tuning, while explaining the exclusion of other parameters like LC and LF.

Reviewer Comment (Min C 6): *Figure 4: What do the orange dashed lines represent in this figure? If they have the same meaning as in Fig. 3a, do they indicate the threshold for skillful simulations applied to the three KGE components?*

Response:

We thank the reviewer for bringing this to our attention. First, we apologise for not mentioning this in the original manuscript. The orange dashed lines in KGE component boxplots were initially intended to serve as an informal visual demarcation denoting a 'very good' simulation. However, we fully acknowledge that these markers lack a rigorous mathematical derivation for the individual KGE components—alpha, beta, and R—and risk causing unnecessary confusion. To maintain clarity, we have removed the dashed orange lines from the respective box plots in Fig. 4 among other similar figures.

Action taken:

- Removed the orange dashed lines from the alpha, beta, and R component box plots in Fig. 4 and other similar figures to prevent any misinterpretation of performance thresholds.

Reviewer Comment (Min C 7): *Line 416: How is the conclusion that the variability term is the dominant error source derived from Fig. 4? Is this because the median variability term falls outside the skillful threshold for most GHMs?*

Response:

We thank the reviewer for urging us on to be more precise about this. Originally, our conclusion was based primarily on the fact that the median values of both the variability (alpha) and correlation (R) components were consistently further away from their ideal value of one (1) than those of the bias (beta) component (Fig. 4 in the original manuscript). However, upon further reflection, we sense that comparing the population medians of the components does not formally quantify their proportional penalty to the KGE score at the individual reservoir level, due to the squared nature of the KGE formulation.

To definitively resolve this and objectively quantify the contribution of each KGE component to the overall error, we have implemented a rigorous variance decomposition approach: for each reservoir and each GHM, we calculated the squared-distance contribution for each of correlation, variability, and bias—i.e., $(R-1)^2$, $(\alpha-1)^2$, and $(\beta-1)^2$—and normalised them by their sum (e.g. . Variability % $= \frac{(\alpha-1)^2}{(R-1)^2+(\alpha-1)^2+(\beta-1)^2} \times 100$)

This rigorous variance decomposition confirms our initial hypothesis: variability (α) is indeed the primary source of KGE degradation. From Table R1, we see that averaged across all dams, variability accounts for 49.1% of the total KGE error penalty, followed by correlation (30.9%) and bias (20.0%). More precisely, variability was the largest source of KGE degradation for four of the five GHMs (H08, WGP, CWT, and particularly LPJ, where it accounted for 72.2% of the error), with only MIL showing a higher penalty from correlation. We have updated the manuscript to replace our visual interpretation with these definitive quantitative results.

Table R1: Percentage contribution of KGE components to KGE degradation. Values denote the average normalised squared-distance contribution of correlation (R), variability (α), and bias (β) error to the overall KGE error penalty across all evaluated reservoirs. The dominant source indicates the component responsible for the largest proportion of performance degradation for each respective GHM.

GHM	R (%)	Alpha (%)	Beta (%)	Dominant Source
H08	34.8	43.1	22.1	Variability
WGP	29.3	52.3	18.3	Variability
MIL	37.6	33	29.5	Correlation
CWT	35.4	44.9	19.7	Variability
LPJ	17.3	72.2	10.5	Variability
Average	30.9	49.1	20	Variability

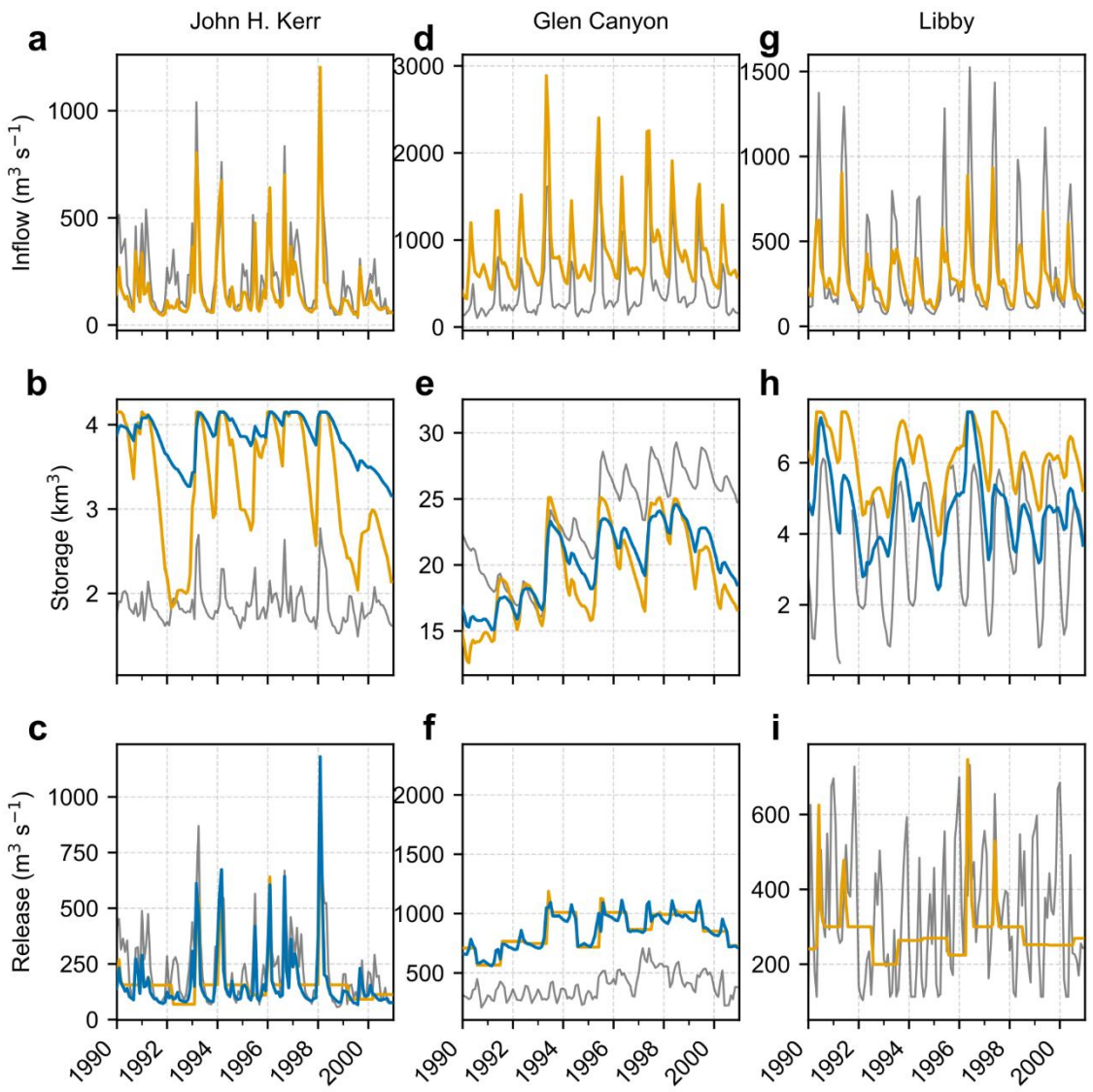
Action taken:

- Revised Sect. 3.3 to replace the visual median comparison with a formal variance decomposition of the KGE components.
- Added the following text to Sect. 3.3: "To formally quantify the contribution of each KGE component to the overall performance degradation, we calculated the normalised squared-distance contribution of correlation, variability, and bias for each reservoir (e.g., $(R-1)^2$ normalised by the sum of all three squared components). This analysis confirmed variability as the dominant source of error, accounting for 49.1% of the total KGE penalty on average across all models, followed by correlation (30.9%) and bias (20.0%)."

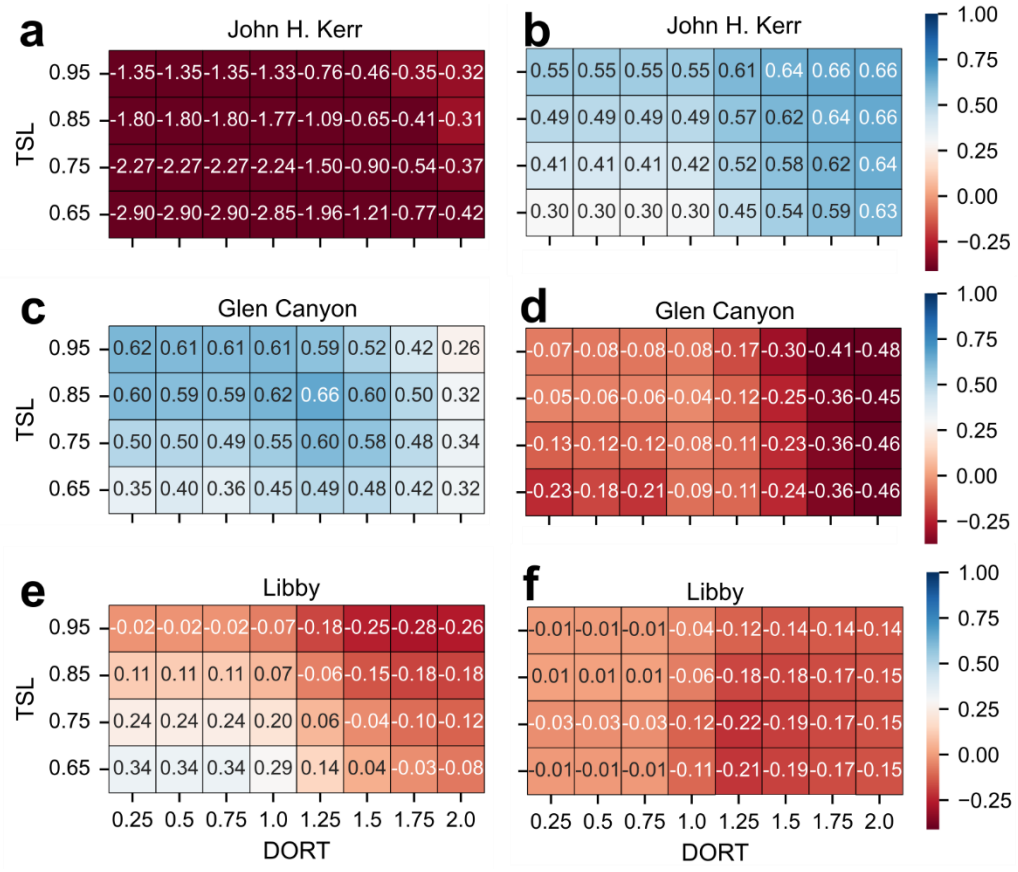
Reviewer Comment (Min C 8): *Figures 6, 8 and 10: Please merge these three figures for better visualization. The time series of inflow, storage and release for the three dams could be combined into one figure, and the KGE heat maps for the parameter tuning experiments could be merged into a single plot.*

Response:

We thank the reviewer for the suggestion to condense the visual information. We have now merged Fig. 6, 8 and 10 into one figure (Fig. R6 below). We also have merged the KGE heat maps for the parameter tuning experiments into a single plot (Fig. R7).



390 **"Figure R6: Sensitivity analysis of H06 parameters at the John H. Kerr, Glen Canyon and Libby dams. Panels (a–c) show results for John H. Kerr Dam, (d–f) for Glen Canyon Dam, and (g–i) for Libby Dam. The first column (a, d, g) compares long-term mean monthly observed (in situ) inflow with H08-simulated inflow obtained using the default global parameter values (TSL = 0.85, DORT = 0.5). The second column (b, e, h) compares observed (in situ) storage with H08-simulated storage using the default global parameters (vermilion curve) and the best-performing tuned parameter set (blue curve). The third column (c, f, i) shows the corresponding comparison for reservoir releases. "**



"Figure R7: Sensitivity analysis of the H06 parameters at John H. Kerr, Glen Canyon and Libby dams. The left column (a, c, e) shows KGE values for simulated reservoir storage while the right column (b, d, f) shows KGE values for simulated reservoir releases. Each heatmap cell represents a unique TSL–DORT parameter combination, with colours and annotations indicating the resulting KGE score."

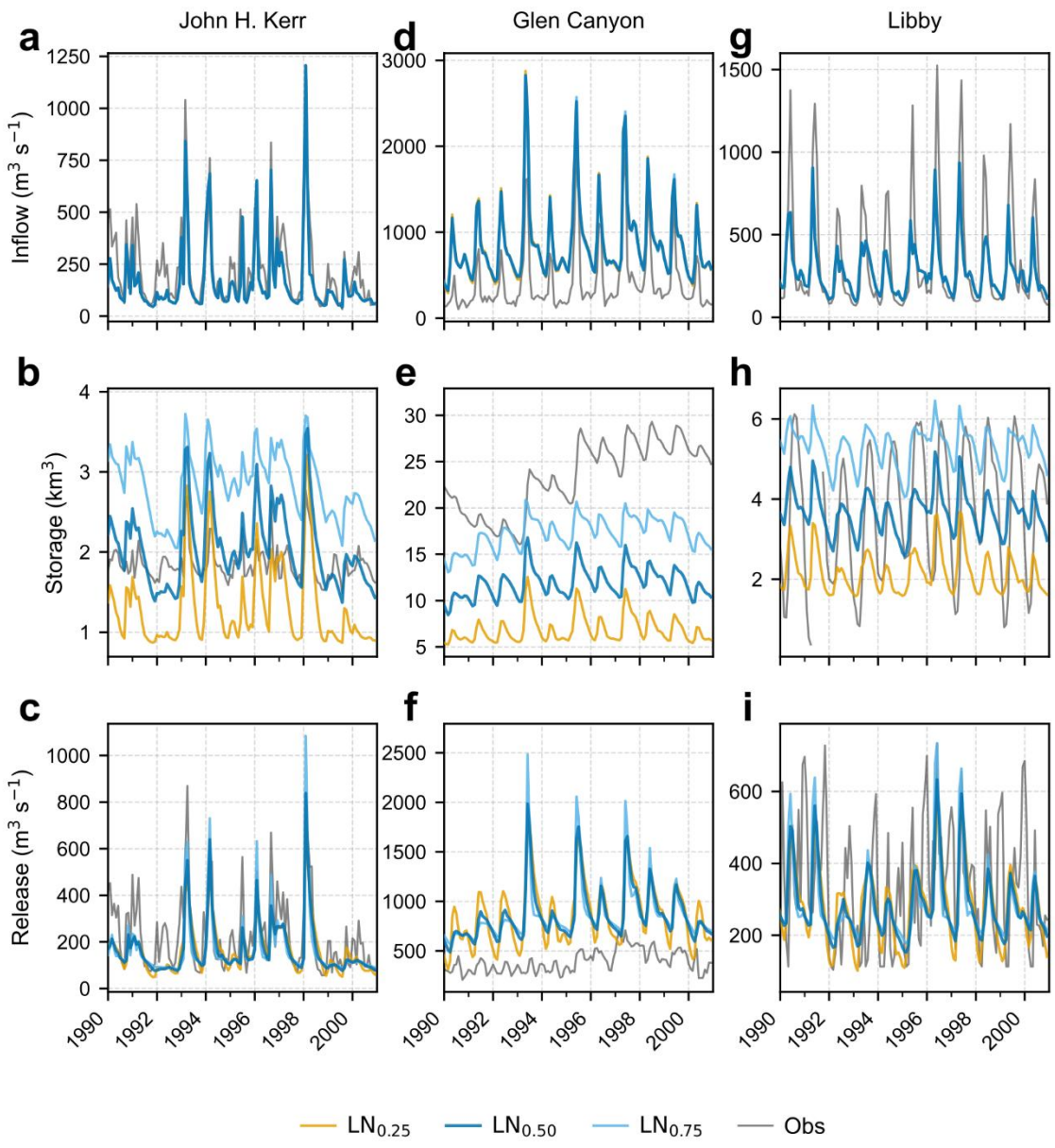
Action taken:

- Merged Fig. 6, 8, and 10 into one figure (Fig. R6)
- For each dam, merged all KGE heatmap panels into one figure (Fig. R7).

Reviewer Comment (Min C 9): *Figures 12–14: Please merge these three figures. The inflow, storage, and release plots at three dams can be merged into one figure, and the metrics for the parameter tuning experiment can be merged into one table.*

Response:

We thank the reviewer for the suggestion to consolidate these visualisations. We have now merged Fig. 12–*14* into one figure (Fig. R8). Moreover, the skill scores from the parameter tuning experiment have now been put into a single table (Table R2).



"Figure R8: Sensitivity analysis of LIS' LN parameter at John H. Kerr, Glen Canyon and Libby dams. Panels (a–c), (d–f), and (g–i) correspond to John H. Kerr Dam, Glen Canyon Dam and Libby Dam, respectively. The first column (a, d, g) compares long-term mean monthly observed (in situ) inflows with LIS simulations obtained using LN = 0.50. The second column (b, e, h) compares observed (in situ) storage with LIS-simulated storage for LN = 0.25, 0.50, and 0.75. The third column (c, f, i) presents the corresponding comparison for reservoir releases."

"Table R2: Sensitivity of LIS to variations in the LN parameter at John H. Kerr, Glen Canyon and Libby dams. Performance metrics (KGE, Pearson's correlation coefficient (*R*), variability ratio (*α*), and bias ratio (*β*)) are reported for simulated inflow, storage and release using three LN values (0.25, 0.50 and 0.75)."

GRanD ID	Dam	Variable	LN	KGE	R	α	β
1761	John H. Kerr	Inflow	0.25	0.71	0.86	0.94	0.75
			0.50	0.71	0.86	0.94	0.75
			0.75	0.71	0.86	0.94	0.75

		Storage	0.25	-0.31	0.74	2.25	0.71
			0.50	-0.27	0.72	2.24	1.10
			0.75	-0.27	0.59	2.07	1.53
		Release	0.25	0.56	0.78	0.72	0.76
			0.50	0.59	0.78	0.75	0.76
			0.75	0.64	0.77	0.87	0.76
597	Glen Canyon	Inflow	0.25	0.05	0.88	1.15	1.93
			0.50	0.05	0.88	1.13	1.93
			0.75	0.05	0.89	1.14	1.93
		Storage	0.25	-0.06	0.44	0.47	0.28
			0.50	0.13	0.52	0.49	0.48
			0.75	0.28	0.60	0.49	0.70
		Release	0.25	-0.26	0.58	1.60	2.03
			0.50	-0.17	0.66	1.44	2.03
			0.75	-0.23	0.67	1.59	2.03
297	Libby	Inflow	0.25	0.41	0.77	0.48	0.85
			0.50	0.41	0.77	0.48	0.85
			0.75	0.41	0.77	0.48	0.85
		Storage	0.25	-0.05	0.39	0.27	0.57
			0.50	0.16	0.53	0.31	0.98
			0.75	0.08	0.56	0.32	1.42
		Release	0.25	0.06	0.17	0.59	0.85
			0.50	0.10	0.26	0.51	0.85
			0.75	0.08	0.22	0.54	0.84

Action taken:

- Merged Fig. 12–14 into one figure (Fig. R8).
- Transferred the skill scores from the parameter tuning experiment into a single table (Table R2).

References

Beck, Hylke E., Tim R. McVicar, Noemi Vergopolan, et al. 2023. 'High-Resolution (1 Km) Köppen-Geiger Maps for 1901–2099 Based on Constrained CMIP6 Projections'. Scientific Data 10 (1): 724. https://doi.org/10.1038/s41597-023-02549-6.

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.

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.