This study benchmarks the performance of four reservoir operation schemes across the US in a large-scale hydrological model. The manuscript compares four different calibration strategies, finding that calibrating to reservoir storage is more informative than calibrating to reservoir outflow. I think that the results of this study are very important for the modelling community and the take-home messages should be carefully considered by anyone incorporating reservoirs into large-scale hydrological modelling. In fact, I have been hoping someone would publish a study similar to this for a while so thank you! In general the manuscript was very clear and well-written, but below I have left a few suggestions for how I think it could be improved.

Comments

It would be interesting to hear more about your model calibration strategy (as introduced on **L60**). It is not super clear to me how the specific details of the calibration worked. Am I right in thinking that you used the 'default' reservoir parameters from the literature listed in Tables 2, 3 and 4 and then calibrated the non-reservoir parameters around these? If so, considering that the results using the default reservoir parameters often failed to capture the storage dynamics well, do you think that the non-reservoir parameters were calibrated in a way which means they overcompensate for poorly represented reservoir processes? Did you compare the selected non-reservoir parameters to values used in natural catchments or the literature to see whether they were physically realistic? Perhaps you can elaborate on this a bit.

In this study, we isolated the reservoir model from the rest of the catchment hydrology. To do so, we selected reservoirs for which inflow, outflow and storage records are available, so we can model solely the reservoir management. Therefore, the only model parameters are those from the specific reservoir routine; there aren't any non-reservoir parameters. In each of those routines, we ran different simulations: one using the reservoir model parameters reported in the literature (named *default*), and 3 other calibrations targeting different variables.

We will revise Line 60 in the future version to ensure that this calibration process is clear.

Finding that calibrating the model to reservoir storage is more informative than calibrating to outflow is a really interesting (and useful!) result. I am pleased that your results suggest we may be able to utilize satellite data for model calibration but wonder whether you should demonstrate this in the manuscript. Did you try integrating satellite data (e.g. the data you discuss in Appendix A) into some of your reservoir storage calibration experiments? If not, I think this would be a very valuable addition to **Appendix A** or the manuscript. If this paper is going to advocate for this possibility it would be nice to showcase this, particularly because in many places storage data like in ResOpsUS is not available. It would be interesting to know how the differences in satellite derived storage impact the results.

We appreciate this suggestion. We agree that evaluating the suitability of these routines for calibration against satellite products is a valuable extension of our study. Unlike continuous daily records, satellite observations are temporally sparse; this makes STARFIT particularly well-suited for such data due to its specific design (Steyaert et al., 2025). We agree that testing satellite data within our benchmark analysis is highly relevant, and we will include this analysis in Appendix A.

Line 100. How were the thresholds for DOR and DOD selected to define a significantly altered natural flow regime?

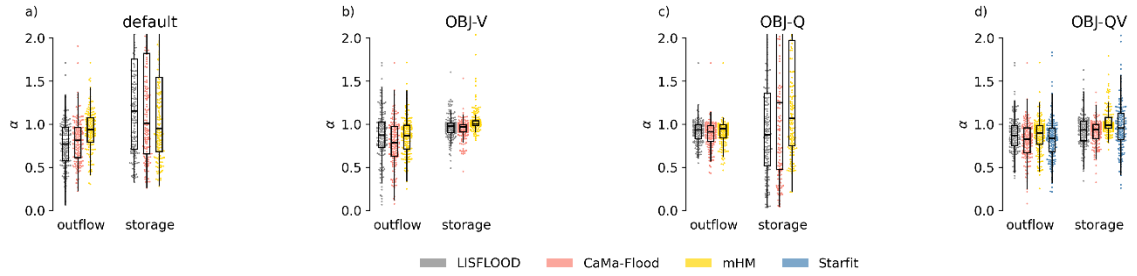
We took those thresholds from Shrestha et al. (2024). They computed the degree of regulation (DOR), degree of disruptivity (DOD) and a third disruptivity index called amended annual proportional flow deviation (AAPFD), which requires more data as it needs both natural and regulated flows. By using K-means clustering, they came out with the DOD and DOR thresholds (indices easy to compute) that identify disruptive reservoirs.

Out of the 284 reservoirs in ResOpsUS with records for the three reservoir variables (inflow, storage and outflow), only 12 reservoirs were removed from the sample based on the regulation conditions. The most limiting conditions are:

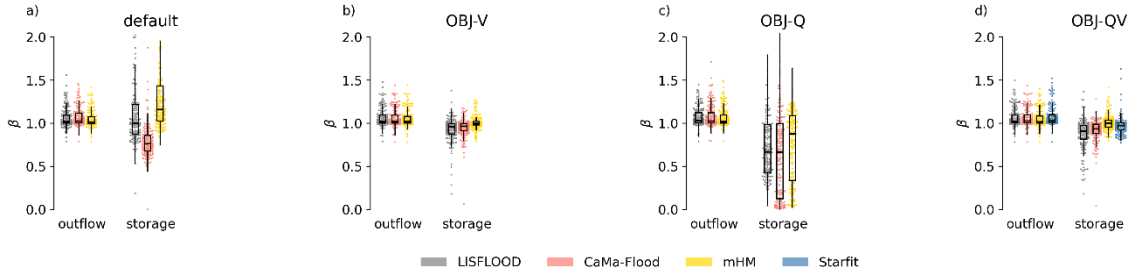
- The minimum length of the records. 210 out of 284 reservoirs have at least 4 years of data.
- The water balance integrity. 27 out of those 210 reservoirs with long enough records were removed from the sample as the absolute bias between average inflow and outflow is larger than 30%.

At some point (even if in an appendix) I would be interested to hear about the breakdown of the individual KGE components. Did all aspects of the metric perform similarly or were there some that were always high or always low? How did this vary across reservoirs of different types?

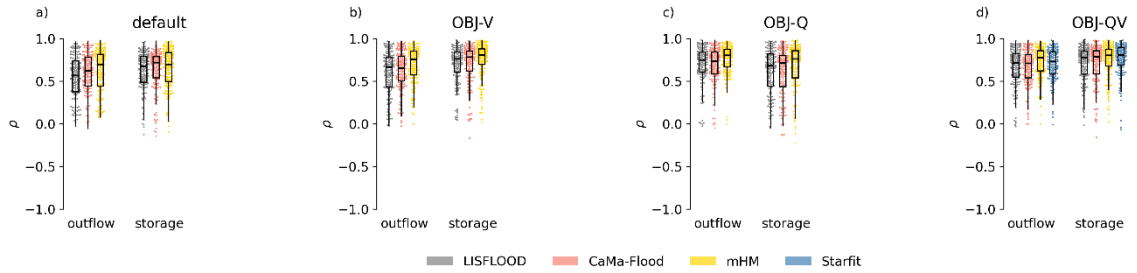
We thank the Reviewer for this suggestion. While we omitted the individual components of the Kling-Gupta Efficiency (KGE) for brevity in the original manuscript, we agree that this breakdown provides valuable insight into model behaviour. We will add plots such as those below in a new Appendix B.



*Comparison of the model performance in terms of **variability** (ratio of the coefficients of variation).*



*Comparison of the model performance in terms of **bias** (ratio of the means).*



*Comparison of the model performance in terms of **Pearson correlation coefficient**.*

A summary of the KGE components for the bivariate calibration (**OBJ-QV**) reveals the following:

- The superior performance of the mHM routine is driven by its more accurate representation of outflow variability and correlation, as well as storage variability and bias.
- Overall, errors in the storage simulation are slightly smaller than those in outflow across all models. For storage, the correlation coefficient is the weakest metric across all models; for outflow, both variability and correlation show lower performance compared to bias.

We observe that the univariate calibration of outflow (**OBJ-Q**) significantly degrades both the variability and bias of simulated storage. In contrast, the univariate calibration of storage (**OBJ-V**) only marginally affects the correlation of simulated outflow and has a negligible impact on outflow variability and bias.

Could Figure1 also show the primary purpose of the reservoirs? This seems important for the operations. Is it possible to add some analysis somewhere which describes how the KGE performance varied across reservoirs of different types? This could link nicely to the discussion in **section 5.2**. I think it would be useful for readers to understand whether the results of this study would apply to other locations where perhaps there is a different distribution of reservoir types.

We thank the reviewer for this suggestion. We agree that reservoir purpose is a critical driver of the operational logic and, consequently, model performance. We have explored the results grouped by reservoir main use (as reported by GRanD), and we propose the following changes:

1. Updated Figure 1 in which the main reservoir use is presented. To accommodate this new information without cluttering the map, we now use colour to represent reservoir use and proportional symbols (dot size) to represent the Degree of Regulation (DOR)—the ratio between storage capacity and total annual inflow.
2. New Appendix B that explores the model performance in relation to the main reservoir use. As shown in the table below, our results indicate that mHM remains the highest-performing routine across most categories. However, CaMa-Flood reproduces storage slightly better in flood control and water supply reservoirs. We also note that the lower performance in Navigation and Recreation categories is likely due to the very small sample size.

Performance (in terms of median KGE') of the bivariate calibration disaggregated by target variable (storage or outflow), model and the main reservoir use.

	Storage				Outflow				
Main use	**LIS-FLOOD**	**CaMa-Flood**	**mHM**	**Starfit**	**LIS-FLOOD**	**CaMa-Flood**	**mHM**	**Starfit**	**Count**

Flood control	0.63	**0.74**	0.70	0.73	0.57	0.50	**0.56**	0.46	66
Hydropower	0.72	0.75	**0.77**	0.69	0.69	0.66	**0.72**	0.68	22
Irrigation	0.65	0.67	**0.82**	0.73	0.64	0.60	**0.79**	0.76	54
Navigation	0.24	0.28	**0.35**	0.33	**0.87**	0.86	0.86	0.84	4
Recreation	0.00	0.10	-0.34	**0.48**	0.62	0.60	**0.66**	0.55	1
Water supply	0.55	**0.74**	0.73	0.56	0.67	0.65	**0.71**	0.60	17

I think one of the most interesting results in this paper is on **L356** where you state that STARFIT was not markedly superior to CaMa-Flood which is far simpler. There is often an assumption in our field that more data/ complexity will always lead to better results and so I think it is important that we highlight that this is not always the case. Could you consider mentioning this in the abstract?

We thank the reviewer for this observation. We agree that this is an important finding of this study, so we propose rewriting **Lines 7-9** in the Abstract as:

"Our results indicate that the mHM routine consistently achieves the highest performance; however, its dependence on site-specific demand data limits its applicability at the global scale. In contrast, the CaMa-Flood routine provides a robust compromise, significantly outperforming the linear logic of LISFLOOD and matching the performance of the more complex STARFIT routine. This suggests that increased model complexity does not always yield superior results."

You mention several times that STARFIT still has distinct advantages over the other schemes (e.g. on **L357** and **L445**) but I cannot see how your results evidence this? It seems to have been outperformed by simpler methods. Can you make it clearer why you think this?

We acknowledge that based strictly on performance, STARFIT was matched or outperformed by simpler routines. However, its distinct advantage refers to its structural flexibility in data-sparse scenarios, rather than its performance in data-rich environments.

STARFIT uses harmonic functions to compute reference values. This allows the model to be fitted at coarse or irregular temporal resolutions (e.g., weekly or monthly) and still be executed at daily time step by simply adjusting the frequency term (ω).

This characteristic makes STARFIT uniquely suited for calibration against temporally sparce satellite-derived products. As noted in the discussion, this potential has already been demonstrated by Steyaert et al. (2025).

I think it is really nice that you have published the ResOpsUS+CARS dataset!

We thank the reviewer for the comment. We have also published in Zenodo a similar dataset for Brazil (ResOpsBR+CARS), and soon there will be another one for Spain.