

In this study, the authors evaluate four reservoir operational schemes: LISFLOOD, CaMa-Flood, mHM and STARFIT, in order to evaluate which scheme to identify which scheme is the most robust for large scale hydrologic models. To do this, they evaluate 160 dams in the CONUS domain in order to determine which parameters are the most informative for model calibration. The results demonstrate that mHM performs the best but requires site specific data which is not always accessible on the global scale. CaMa-Flood ultimately provides the best tradeoff with lower data requirements but more accurate reservoir storage levels. The authors also find that reservoir models should be calibrated against reservoir storage over reservoir outflow.

Major Comments

The cut offs for DOR and DOD are not well defined. I would be interested to know more about why these exact values were used. I would also be interested to know if these values cut dams with main purposes that are not labeled as Hydropower as I would assume any dam focused primarily on providing hydropower regulation would have a lower DOR. Based on **Figure 1**, it looks like the majority of the dams that were cut correspond to hydropower dams in the eastern US. Inclusion of the dam purpose would be nice to have a broader idea of the purposes the study includes.

We thank the reviewer for this comment regarding the selection criteria.

The Degree of Regulation (DOR) and Degree of Disruptivity (DOD) thresholds were adopted from [Shrestha et al. \(2024\)](#). In their study, K-means clustering was used to correlate these easily estimated indices with the more complex Amended Annual Proportional Flow Deviation (AAPFD). The thresholds we used represent the validated clusters that identify disruptive reservoirs—those with a non-negligible impact on downstream flow regimes.

These thresholds are applied in large-scale hydrological models to reduce complexity without compromising the streamflow simulation. In operational systems such as the Global Flood Awareness System (GloFAS) or European Flood Awareness System (EFAS), the hydrological model includes around 2000 reservoirs. While we have relaxed certain capacity conditions in upcoming versions, our objective is to maintain model parsimony by excluding reservoirs that do not significantly alter downstream streamflow, which remains our primary forecasting target.

Only 12 reservoirs were removed from the sample based on the regulation. According to GRanD, the main use of these 12 reservoirs is irrigation (4), hydropower (3), navigation (3), other (1) and flood control (1). They are geographically distributed across the CONUS.

To provide a better idea of the reservoirs used in this study, we will update Figure 1 to display reservoir “main use” via colour coding. Besides, Appendix B will explore model performance grouped by reservoir use. For context, our final study sample is dominated by flood control (66) and irrigation (54), followed by hydropower (22), water supply (17), navigation (4) and recreation (1).

I really like that you published your updated version of ResOpsUS +CARS. I imagine it will have great use in further evaluations of reservoir models.

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

The calibration of storage and outflow is mentioned in **L60** as well as **Table 1** and **L165**, however I would be curious to know what this de-coupled calibration looks like. It seems like it is calibrating the reservoir model independent of the hydrologic model, but without the explicit link I am unsure. I would add a brief description in the article and then perhaps a flow chart in Appendix. I would also be interested in how the actual calibration for the desired parameters were done per model.

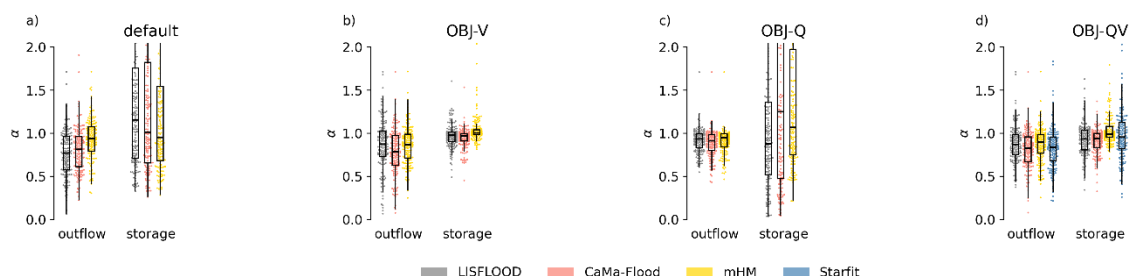
We thank the reviewer for this request for clarification. The "de-coupled" calibration was a deliberate methodological choice to isolate the structural performance of the reservoir routines from the uncertainties of the broader catchment hydrology. To do so, we selected reservoirs for which inflow, outflow and storage records are available, so we can model solely the reservoir management.

The calibration of the reservoir parameters was done using the genetic algorithm Shuffle Complex Evolution-University of Arizona (SCE-UA). The details are explained in Line 273. For each reservoir routine, we calibrated different setups: one using the values of the model parameters reported in the literature (named *default*), and 3 other calibrations targeting different variables.

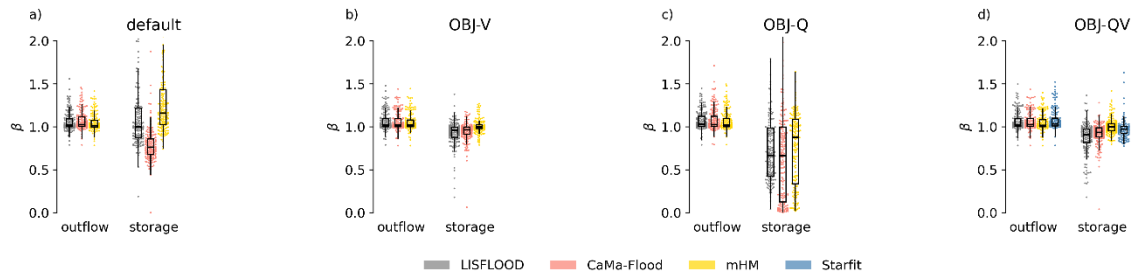
For the final implementation of the reservoirs in GloFAS, we trained a machine learning model (gradient boosted regression tree) that estimates model parameters based on reservoir attributes available in any reservoir in the world. In this way, we were able to generate model parameters for all the reservoirs in GloFAS, so the reservoir parameters were removed from the calibration of the hydrological model. We discarded this from the paper as it is out of the scope of the benchmarking, and it used datasets from Brazil, Mexico and Spain, not mentioned here.

The inclusion of KGE for outflow and storage is an informative addition. I would be interested to know how the outflow and storage KGE was calculated. Is this just the average of the outflow and storage KGEs? Additionally, the **KGE components** could be interesting to show as they might inform a bit more about the dynamics with regard to variation and biases in these schemes. It would also be nice to look at the storage and streamflow KGE components to see if storage on average fits better with respect to one of the components. Perhaps this could strengthen the argument that storage calibration is more important than streamflow calibration.

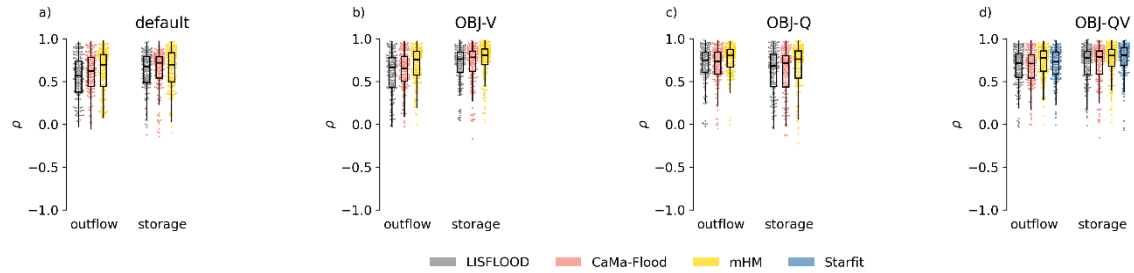
We thank the reviewer for this comment. We will add the analysis of the KGE' components for the different variables in a new Appendix B. The plots below show a first analysis of the individual performance of those components.



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.

The bivariate KGE is shown in equation 15b. Its calculation follows the same logic used in the KGE metric, i.e., it is the Euclidean distance from the ideal value that would be a KGE of 1 in both storage and outflow.

I really like your conclusions on **L312 -315** with respect to incorporating remotely sensed data as a form of calibration, however, remotely sensed storage time series also contain errors due to overestimation of low storage values. It could be interesting to know if the authors looked at model calibration with remotely sensed data as well (perhaps using the data discussed in Appendix A). If so, I would be interested in potential biases that they found when comparing this calibration to calibrations done on direct observations of reservoir storage.

We thank the reviewer for this suggestion. We didn't calibrate the reservoir routines to satellite products in the original version of the manuscript, but we agree that evaluating the feasibility of satellite-based calibrations is of interest for this study. We will include this analysis in Appendix A.

The authors mention that mHM performs the best yet is greatly limited by the reliance on demand time series. The discussion continues to include how in data rich regions demand can be inferred

using machine learning. I would be interested to know the authors opinions on **using modelled demand from other large scale hydrologic models** and if that would be a suitable replacement in data scarce regions. If so, would this be a recommendation to try and include demand in more generic reservoir schemes?

We thank the reviewer for this comment.

It is important to distinguish how different routines "see" demand. In LISFLOOD or CaMa-Flood, demand is often treated "passively"—it is subtracted from storage, but the release logic remains driven by physical rules (e.g., storage levels). In contrast, the mHM routine uses demand "actively" as the primary driver of the release function. All routines in our provided library (https://github.com/casadoj/reservoirs-LSHM/tree/main/src/reservoirs_lshM/models) support demand time series as input, but only mHM uses them to define the operational target.

While using simulated demands from other LSHMs is possible, it introduces significant cascading uncertainty. For example, in OS LISFLOOD, water demands are aggregated at the "water region" level based on national-scale statistics (https://github.com/ec-jrc/lisflood-code/blob/feature/docs/docs/4_Static-Maps_water-use/index.md). Demands are not allocated to specific reservoir locations and do not account for the complex partitioning between surface water and groundwater at the local scale.

Minor Comments

Figure 4: The dashed blue line in panel b is hard to see. I would recommend making it wider or perhaps another color.

To be improved in the manuscript.

In **Line 400** there is a floating comma.

To be corrected in the manuscript.

Appendix A: I believe the title should read Evaluation of GWW Storage Estimates in lieu of Evaluatin of GWW storage estimates.

To be corrected in the manuscript.