Response to the Editor

Comments

Consider adding a table to summarize the parameter of the STARFIT model (similar to the other models).

The new Table 5 indicates the 19 parameters in the STARFIT routine.

Please increase the size of Fig. 5 and 6 as well as the fontsize of the text in these two figures to improve readability.

Figure 5 and 6 have been revised to improve readability.

p13 L250: “minimum” instead of “minium”.

It has been corrected.

P21 L400: a word is missing.

The word “resolution” was missing.

Response to Saskia Salwey

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 have revised **Lines 61-65** 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 have conducted a fifth experiment in which we calibrate the four reservoir routines to the storage estimates from Global Water Watch. Results are included 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%.

We have rephrased this paragraph (**Lines 100-107**) to clarify the use of those two thresholds. We don't explain how those thresholds were derived, as we cite the publication.

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 have included the analysis of the KGE components in a new **Appendix B**.

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.

1. We have updated **Figure 1** to indicate the reservoir use and the Degree of Regulation (DOR).
2. We have added a new **Appendix C** that explores the model performance in relation to the main reservoir use.

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 have rewritten **Lines 7-11** 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).

These ideas are included in **Lines 362-369** and **Lines 454-457**.

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.

Response to Anonymous reviewer

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 note 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.

We have rephrased **Lines 100-107** to clarify the use of those two thresholds. We don't explain how those thresholds were derived, as we cite the publication. To provide a better idea of the reservoirs used in this study, we have updated **Figure 1** to display reservoir "main use" via colour coding. Besides, the new **Appendix C** explores 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 276**. 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. 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.

We have included the analysis of the KGE' components for the different variables in the new **Appendix B**.

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.

In the new version, we have conducted a fifth experiment in which we calibrated the four reservoir routines using the reservoir storage estimates from Global Water Watch. The results are included 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.

It has been changed in the new version.

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

The word "resolution" was missing. It has been corrected in the new version.

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

It has been corrected in the new version.