Dear Editor,

We have revised the manuscript based on the insightful comments you provided. All recommendations have been addressed in the updated manuscript. We would like to thank you for your thorough consideration and feedback, which helped us enhance our manuscript. Below, we have included a point-by-point response (in blue) to your comments. The suggested grammatical modifications are not listed here.

Editor's comments

Editor: Dear Authors, thank you for the careful revision of the manuscript following the reviewers' comments. The reviewers have reconsidered the manuscript and found that the concerns raised have been adequately addressed. I would like to propose the manuscript for publication. However, before proceeding, after a careful read of this final manuscript I believe there are some minor revisions that are required. Mostly these refer to the clarity of writing as well as some editorial corrections. I am attaching an annotated version of the manuscript. In this final revision, I would also ask to carefully review the results and discussion sections. Overal the results as presented and discussed are interesting and relevant, as also noted by the reviewers, but the writing it extremely dense and in many places difficult to follow. I would like to ask the authors to carefully review these sections and if the logic of message they aim to convey emerges clearly from the narrative. In the annotated attachment some examples are provided (as in the discussion), but the "dense" writing style pervades these sections. In many cases simplifying sentences with fewer sub-clauses can help a great deal.

Response: Thank you for your thoughtful feedback and for considering our manuscript for publication. We truly appreciate your effort to thoroughly review the manuscript and offer detailed comments to enhance its clarity and readability. We have carefully reviewed the annotated version of the manuscript and made revisions to address the editorial corrections and suggestions for enhancing the clarity of the Results and Discussion sections. Specifically:

- *Clarity and Simplification*: We have revisited the Results and Discussion sections to address your concerns.. Sentences with multiple sub-clauses have been simplified, and we have worked to ensure the logic and flow of the narrative are clear and concise.
- *Annotated attachment*: The specific examples highlighted in the annotated attachment have been addressed, and similar revisions have been applied consistently throughout the manuscript.
- *Final Review*: In addition to responding to the provided examples, we thoroughly reviewed the entire manuscript to ensure that the central messages emerge clearly and that the sections are accessible to a broad audience.

We hope these revisions meet your expectations and contribute to a more readable and compelling manuscript. If you require any further changes or clarifications, we will be happy to address them promptly.

Thank you again for your valuable feedback and for guiding our manuscript toward publication.

Editor: Line 330: This is not so clear. Maybe write this out a bit more clearly.

Response: We have revised this part as follows:

"In the case of the SA and WA approaches for considering downstream demand, the process involves using the provisional release R'_d instead of $\overline{T}$ in Eqs. 7 and 8. Therefore, similar to the DH algorithm, Eq. 4 was used with the default value for the parameter a_2 to estimate R'_d. Please note that, since the WA and SA approaches work with $\overline{I}$ and not $\overline{I'}$, $\overline{I}$ was applied in Eq. 4 instead of $\overline{I'}$ in the SA and WA approaches."

Editor: Line 460: sharp decline of what?

Response: It referred to the simulated storage. In the revised version, we have rephrased this section and removed the statement.

Editor: Line 464: So does this mean that the storage reported by GRAND is incorrect? If this is the case it is obvious the performance would poor! Somewhat trivial. Should that not have been corrected?

Response: Unfortunately, discrepancies exist in several cases between the reservoir storage capacities reported by GRanD and the maximum observed values, as also noted by Steyaert et al. (2022). We acknowledge that such discrepancies should ideally be addressed at the local scale. However, the primary goal of this paper is to provide guidance for large-scale models and their calibration against remotely sensed storage anomalies in regions where in-situ data is unavailable. Given that observed storage data is largely unavailable globally, and large-scale models heavily depend on GRanD data, we have opted not to modify the GRanD data. Instead, we aim to highlight one potential cause for the poor performance of reservoir operation algorithms.

Editor: Figure 4: The figure is already very busy. However, it may be useful to also indicate the relative storage lines (e.g. Srel = 0.4, 0.7 and 1.0) as these are of significance in the algorithms

Response: Thank you for your suggestion. The revisions have been made.

Editor: Table 3: This table is very difficult to read and understand if the results are significant. This is in part due to the improvements only being identified for skillful simulations, which means the total number considered varies (I assume this is the sum of the numbers in an outside the parenthesis). Numbers are often small so could all just be chance. Should a significance test be done?

Response: Indeed, the evaluated numbers are the same and equal to 21. However, to determine how often one approach yields better results, we first filter the 21 KGE values of that approach using a threshold of -0.73. Then, we compare whether considering or not considering water demand leads to a better outcome. Situations where considering or not considering water demand results in non-skillful simulations are disregarded and compared against the alternative variant (not considering or considering water demand, respectively). We believe in selecting models and parameters based on overall performance instead of relying on significance tests. We choose the model that typically delivers the highest performance, even if the second-best model shows similar results according to a significance test. When two models demonstrate the same performance, we prefer the simpler approach. In our case, since factoring in downstream water demand does not produce better results than ignoring it, we opt for the simpler model—i.e., the algorithm that does not account for water demand. Furthermore, we believe that the small sample size hinders our ability to conduct a reliable significance test.

We have revised the table caption for clarity as follows:

"**Table 3.** Comparison of reservoir simulation performance using different algorithms, both with and without considering downstream water demand for 21 irrigation and supply reservoirs. Numbers outside parentheses indicate the number of reservoirs (out of 21) where performance improves when downstream demand is taken into account.

In contrast, values inside parentheses represent reservoirs where ignoring downstream demand leads to higher KGE values. Improvements are noted only for skillful simulations achieving a KGE value greater than -0.73. All algorithms are calibrated against outflow, storage, storage anomaly, and estimated storage using KGE as the objective function. The inflow data is sourced from the WaterGAP model."

Editor: Line 630: It is not so clear that this sentence provides a summary of this sub-section. Please check. It comes across as somewhat confusing

Response: We have revised the summary sentence for this subsection to enhance clarity and better align with the key findings:

"In summary, our results suggest that enhancing the quality of inflow data is more crucial than calibrating reservoir operation algorithms, particularly when the objective is to achieve accurate outflow simulation. Only calibrating against storage anomalies does not ensure better outflow predictions."

Editor: Line 682: This is because for DH some defaults are provided. Can this not be done for WA and SA also?

Response: Both the WA and SA algorithms are not calibration-free; their parameters can be regionalized or adapted to other reservoirs using specific techniques that are beyond the scope of the current study. However, the key point is that calibration-free algorithms are generally not ideal for reservoir operation because each reservoir behaves differently. Nowadays, with storage anomaly data for reservoirs accessible through remote sensing, it is reasonable to use this information to calibrate reservoir operation algorithms for each reservoir individually. However, if such data is not available, traditional calibration-free algorithms remain the best available option.

Editor: Some of the comments in this final paragraph are more befitting of the discussion. For example, the discussion does not raise the issue of the grid search method - so whether that would improve results or change the outcomes is somewhat speculative. I would suggest a final paragraph that is better aligned to the scope of the study as a sort of overall conclusion.

Response: We have revised the final paragraph of the conclusions as follows:

"Although the algorithms introduced in this study outperform the conventional DH algorithm, there remains scope for improvement. For example, integrating knowledge-based equations with deep learning in hybrid machine learning methods could be beneficial for simulating reservoir dynamics. However, improving the accuracy of inflow simulations and validating reservoir-related characteristics is very likely more important for achieving better reservoir outflow and storage simulations than refining the algorithm itself."

Editor: Line 718: Is this intended to be a URL external to the article. If not then the supplementary material will be published with the article. The URL will be provided. Otherwise please provide the URL. The preferred approach is to publish this with the article on the HESS repository and not externally.

Response: The supplementary material will be published alongside the article.

References

Steyaert, J. C., Condon, L. E., Turner, S. W. D., and Voisin, N.: ResOpsUS, a dataset of historical reservoir operations in the contiguous United States, Sci. Data, 9 (1), 34, doi: 10.1038/s41597-022-01134-7, 2022.