Dear Reviewer 2,

Thank you for taking the time to review our paper. We are grateful for your thoughtful comments. We trust that the revisions presented below address the issues you raise.

Yours faithfully,
Naota Hanasaki (on behalf of authors)

The aim of the work of Otta et al. is to use satellite-based storage estimation to improve reservoir representation in Global Hydrological Models. This is an exciting topic because a number of GHMs do not include reservoir storage (Telteu et al. 2022) or generalized reservoir algorithms, which can be quite different from real reservoir operation rules. On-ground observation for most parts of the world is limited for different reasons; therefore, satellite-based data might be a feasible way to bypass the lack of data. The authors tested satellite-based storage estimation against gauge observation and two GHM model results.

We are pleased to know that the reviewer is interested in our research topic.

The first part – Satellite vs. gauge is already discussed in a more extensive way in other papers e.g. Sorkhabi et al. 2002, Gourgouletis et al. 2022, Hou et al. 2022, and others, either for single reservoirs or on global level. Therefore chapter 3.1 does not add much.

Thank you for sharing your observations. We agree with you that Subsection 3.1 includes some points that have been discussed elsewhere. We believe, however, that the contents are helpful because they disclose the data used in detail; otherwise, the interpretation of the latter part of the paper would become rather challenging. Reviewer 1 also made similar comments and we have therefore removed Research Question 1 and contextualized Subsection 3.1 as a data introduction subsection. We have added the following text to the beginning of Subsection 3.1: "Satellite data have been cross-referenced with the growth of observations from various studies (e.g., Hou et al. 2022). The following section presents the critical details of the data and methods used in this study to ensure the comprehensiveness of our approach."

Further, we've moved section *3.1.2 Decomposed monthly reservoir storage from satellite and ground observation* to supplementary information in order to tone down section 3.1.2

The second part (chapter 3.2.) is the interesting part, but it is a bit small to stand alone. An in deep discussion, why the 2 GHMs representation of monthly and annual average storage can be improved (I think an average 0.5 correlation is far from good representation). Maybe some variation of the reservoir operations (some parameter change) would be nice, especially as the title of the paper is "… to validate reservoir operations.". The paper has potential , but it needs major changes.

Thank you for this comment, which indicated to us that we need to clarify further the objective and scope of this work. The primary objective of the study was to assess the opportunities for, and challenges of, introducing satellite-based reservoir storage estimation products for use in global hydrological model intercomparison projects. A detailed model intercomparison will follow this work, along with an expansion of the study domain (i.e., from CONUS to the whole globe), which will involve the participation of more GHMs. This point was further clarified in the Introduction section. We also clarified the topics not covered by this study at the end of the Introduction section. "The following topics are beyond the scope of this study: 1) the improvement of the generic reservoir operation algorithms embedded in GHMs; 2) the generation of new satellite-based reservoir data; and 3) the production of case studies at a handful of reservoirs (this is, rather, a path to a future global study)."

We have made several changes in the manuscript to make our point clear and incorporate the concerns raised by the reviewers.

L 40f: Telteu et al. 2022 give a good overview which GHM uses reservoirs and how

Thank you. We have cited Telteu et al. (2021) in the Introduction section.

L 56f: This lines a repeating L. 51-55. Instead you can cite Zajac et al. 2017 who did a global (real global) analysis on the effect of dams and lakes on streamflow.

Thank you. To reduce the redundant text, we have removed the following part:
"such as the Green-Colorado and the Missouri-Mississippi Rivers (Masaki et al. 2017)."
Our intention here was to discuss the past multi-model intercomparisons of reservoir operation estimations. Because the main topic of Zajac et al. (2017) (i.e.,

the importance of reservoir parameterization in global flood forecasting) was slightly different, we retained the original content.

L 60f: There are quite a number of studies that evaluate satellite-derived altimetry and the use for reservoir studies. As mentioned before, some of them are more elaborated than this paper on comparing satellite vs. gauge. But the authors are right: "… there remains a need to establish a method... for GHM validation" and I would say the paper shows a need for improvement of reservoir representation in GHM.

Thank you. Because Reviewer 1 also made a similar comment, we have removed Research Question 1 (see our response to Reviewer 1). We clearly state at the end of the Introduction that a comparison between satellite and gauge data is NOT the focus of our study.

L 144: really year 286?

Yes, the GRanD dataset includes a dam built in A.D. 286. Please see our response to Reviewer 1. We've mentioned:
GRanD includes Lake Qattinah (also called Lake Homs), which was a reservoir constructed by the ancient Romans. We have therefore not changed this text.

L 145　if the write approx. than 7000 km3 without digits is ok.

Thank you. We have removed the digits.

L220 Table 1: Maybe add a column for Database GRBD

Please note that GRBD adopts an identical ID to GRanD. We note this in the caption to the table.

L224: You can write a bit more about this assumption. There are databases e.g. Hou et al 2022 (GloLakes).　You used Hoover dam, there is even a paper looking in detail for this lake mead Li et al. 2022 (Constructing Reservoir Area–Volume–Elevation Curve from TanDEM-X DEM Data)

Thank you. To address the first part of your comment, we have added the following sentence: "Note that more advanced data are being produced by developers, including Hou et al. (2022)." Regarding the latter part of your comment, although we fully acknowledge that much interesting research has been conducted to develop new databases, we omitted all studies dealing with a single or limited number of reservoirs for the reasons stated in the Introduction.

L 230: Figure 1: This fig looks like fig 3 in Busker et al. 2019 which I think is ok. But you have to explain the dots and a ΔS in the middle would be nice.

Thank you. We have revised the figure and caption as follows.

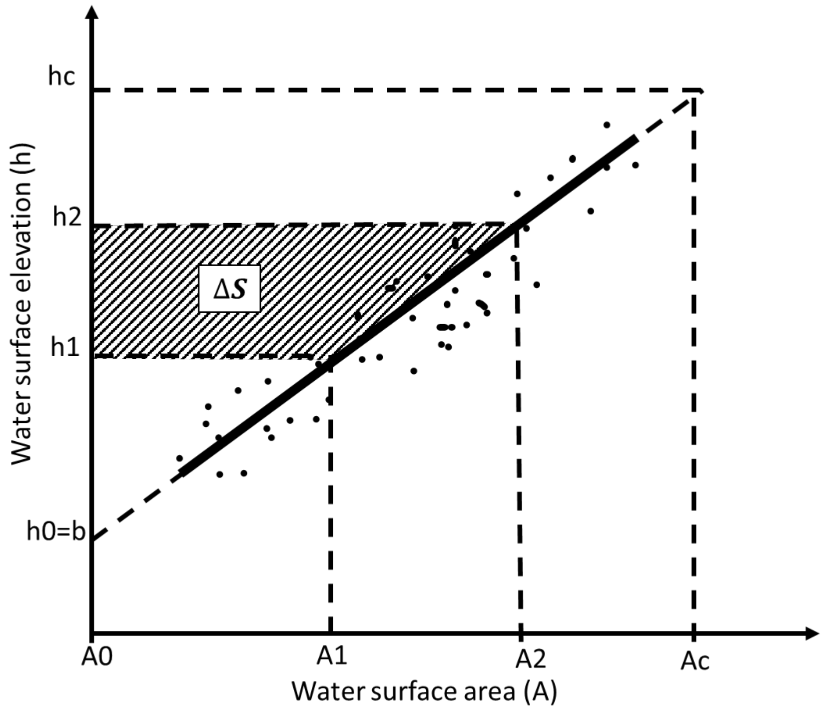

Figure 1: Relationship between the observed water surface area (A), water surface elevation (h), and the change in storage volume (Δ*S*, shaded region). The black dots depict the values of A and h for a hypothetical reservoir, while the thick solid black line represents the correlation within the observed range. This solid line can be extended in both directions to illustrate a hypothetical state ranging from an empty (h0 and A0) to a full reservoir (hc and Ac).

L 238: Please explain Sc (I assume maximum storage). Maybe explain here the difficulties to get Ac and Hc from digital elevation models. Here you used the GRBD database which is based on SRTM data?

First, Sc is the maximum storage and is defined in Subsection 2.2. Second, we have added the following text: “hc records are, in most cases, missing in the reservoir specification inventories.” Third, we utilized not only the GRBD but also the ISIMIP database. This is clearly shown in Table 2.

L 254 later on, you also use GRSAD_ISIMIP. Please explain here.

As shown above, we used two inventories to refer to Sc. GRSAD_ISIMIP indicate GRSAD data in which Sc is adopted in the ISIMIP inventory. This is shown in Tables 2 and S2.

L266: Some of the equations are very basic and fits more into a book of statistics. Equ 7 is not really needed.

We understand the point, but data normalization is key to this study. Because it accounts for only a small part of the paper we have retained this text to ensure that it is available to users.

L285: Also basic statistics. Maybe interesting if you replace fig 2 with some real examples from your 7 reservoirs. Maybe even 3 reservoirs to show the variability of seasonal variability in your 7 reservoirs example (e.g. Hover,Fort Peck, Wesley)

We agree the contents are basic, but we believe that they help readers to understand fully the following sections. We have therefore retained Figure 2, but please note that the “real examples” are shown in Figures 3－7.

L 286. It should be Raw (a-d) and … (e-h)

Thank you for pointing out the typo. We have corrected the labels in the caption.

L 290-305: I do not think correlation and NSE has to be explained in this detail. I think that is done 75724 times before. No equations needed.

Thank you. We have moved the equations for r and NSE to the supplemental material.

L 320: Fig 3 After normalization r should stay the same anyway. For f) the comparison to Dahiti does not make much sense. After normalization the NSE for F) becomes worse?

Thank you for pointing this out. This text has been revised as follows: "As clearly seen in the panels for the Wesley and Coolidge Dams, the temporal coverage of DAHITI is rather limited (Figure 3f, g, m, n). Because the overlap between DAHITI and ground observation is too short, the normalization does not account for the temporal variations over the long ground observation period. Therefore, the NSE score deteriorates on normalization."

L 326 What is GRSAD_ISIMIP?

This means the GRSAD data were converted using the capacity data provided by the ISIMIP project. This is explained in Table S2. To make this clear, we have added "(see Table S2)" to this line of text.

L 360: Fig4 What would be interesting to see is the variance of the seasonal variability. Later on in fig 7 we see that the GCMs have a hard time to represent the seasonal variability in some cases. Is the seasonal part very different each year or is it quite stable even in year of drought (e.g. Hoover and Glen from year 2000 onward)?

Thank you for this interesting question. It is difficult to distinguish each year's seasonal variability and residual. As Reviewer 1 mentioned, it might be possible to compare seasonality between two periods, but we would need to set a year to split the study period into two arbitrary parts. Alternatively, we can always choose particularly dry years to get a dry "seasonality" and make a comparison. However, we believe such a wet-dry study on validation should be another independent research. We really appreciate this suggestion and in response, we've added the following line in our Conclusion section:
"Even several aspects of validation, such as distinction of GHMs' performance between wet and dry period, can be incorporated in future studies."

L415: Figure 5ff: The graphics of S2 is a bit too much spaghetti plot, but still find the supplement figures much more interesting than the average values.

Thank you for this interesting comment. There are two different groups of potential readers; one wants simplicity and the other requires detail. Because both aggregated (Figure 5) and individual (Figure S2) data are presented, the figures have been retained in their original forms.

L460f: I would not say that the model simulations are good. The standard for good should be higher than a correlation of 0.5. The authors do not have to make the results better.

Thank you for sharing your concern. We believe the description of r and NSE values in qualitative terms is rather subjective, and people will have slightly different interpretations of whether the values are good or bad. The standard can also vary due to data accuracy. The differences between the gauge records, the satellite products, and the simulation outputs are very large. Moreover, the GHMs simulate these reservoirs with general reservoir operation rules, which is not specific to any particular reservoir. For a specific reservoir, with its operation rules set, 0.5 would still be considered "fair". Therefore, for the generalized approach with large data uncertainties, we see 0.5 to be a good correlation.

L499: As mentioned before there are studies and databases looking at this in detail.

Thank you. We have added the following text: "although recent studies may have done this differently (e.g., Hou et al. 2022)"

L502: Not currently available. I think there are at least some efforts toward this e.g. GloLakes and others. Referring to a publication from 2012 is not enough.

What we refer to here is that the water level at the reservoir's total capacity is seldom given in historic global inventories of dams. We added the following text: "Monitoring Sc and hc from space is challenging because the water level of operational reservoirs has very little opportunity to reach the exact full capacity. The bathymetry approach (e.g., Messager, 2016) also does not provide information about the water level at full capacity unless the highest shoreline is specified by other methods."

L 508: I cannot find a detailed discussion of Glen Canyon dam in 3.1.2. But this involves also a discussion on digital elevation products like using SRTM, Aster, Tandem.

In our understanding, the published studies utilizing SRTM, ASTER, and TanDEM-X are mostly site-specific (Li et al., 2021) and do not cover the continental scale. We fully understand that better results could be achieved if the number of reservoirs studied were to be limited to sites where advanced satellite data are available, but this was not our aim. This standpoint is clearly stated in the Introduction.

L496 Uncertainties: the GHMs uncertainty is missing. This is about satellite vs. Gauge. But the uncertainty of reservoir representation in GHMs is not in yet. But you showed already some e.g. uncertainty of climate forcing . Why is the climate forcing so relevant for Wesley but not for Toledo? What is the uncertainty from reservoir operation in the models (parameter uncertainty and operation uncertainty). Some variation of reservoir control could be done here.

Thank you for this comment. As clearly shown in the revised Introduction, what we originally wanted to list in this subsection was the uncertainty in utilizing satellite products to validate GHMs. The uncertainties of GHMs will be elaborated on in future work after the methodology has been established and the intercomparison is initiated (this work). At this stage, we have just two GHMs, which we believe are too small to provide a comprehensive description of model uncertainties in simulating reservoir operation. We would like to wait for more models under the ISIMIP project to submit their data for such research work.
Regarding the climate forcing question, we speculate that the runoff estimation of the two models produced the difference in the sensitivity of climate forcing, but the intercomparison of the hydrological models was not the primary focus of this study.

L 533: For 2 out of 7 it is not satisfactory (at least a correlation < 0.5 is not satisfactory for me)

Please kindly note that we did not claim 2 of 7. We specifically wrote, “for seven reservoirs in CONUS, the two ISIMIP3a models, H08 and WGP, demonstrated satisfactory performance in terms of normalized annual average storage and seasonal variabilities.” Actually, the correlation coefficient exceeded 0.5 for the majority of cases. The correlation coefficient of annual average storage exceeded 0.5 for 6 of 7 for GRSAD, 4 of 6 for ground observations for H08, and 7 of 7 and 5 of 6 for WGP.

The correlation coefficient of seasonal variabilities exceeded 0.5 for 4 of 7 and 5 of 7 for H08 and 5 of 7 and 5 of 7 for WGP.

L540: Here it would be nice to answer the question . Why is climate forcing more important for some reservoirs than for others?

Thank you. We have added the following text: “Generally, the discrepancies among forcing data are relatively minor at a continental and monthly scale, but exhibit significant variations at a catchment and daily scale. It is therefore not unexpected that climate data play a substantial role in influencing the variability of GHMs (Müller-Schmied et al. 2016).

L540: Maybe to mention: You cannot compare the effect of forcing by absolute values but only by correlation. Selecting the best meteo forcing by discharge seems still the better method. A multi-objective method would be even better -e.g. snow, evaporation, reservoirs, discharge.

Thank you. Inspired by this comment, we added the following text: “Furthermore, it is crucial to emphasize that singling out the performance of the reservoir operation scheme is difficult because the simulation performance is dependent on the overall hydrological simulation (i.e., the daily water balance calculation carried out by the land surface hydrology and the river routing sub-models in GHMs.)”

L544: It is not clear if the Dahiti data are general have a low temporal resolution for a majority of global reservoirs or only for your selected 7 reservoirs. If it has a low temporal for a majority of reservoirs this conclusion can be drawn from the very beginning and another satellite product should be used for testing against GRSAD.

No, this is not what we meant. We observed that DAHITI includes more missing data than GRSAD and have referred to this in the text. Please note that detailed follow-up information has also been added here in response to a comment by Reviewer 1.

L554: It seems that the quality of the elevation model really matters here. There are a number of DGMs e.g. SRTM, Aster, Tandem and hydrological composites of them e.g. Merit.

Thank you. What you have pointed to is the importance of the Area-Elevation curve. We have added the following: “It includes a better understanding of the Area-Elevation relationship.”

L 562: For me the paper shows exactly this. The error between satellite and gauge, between different climate forcing (with some exceptions) is quite small compared to the error of GHM vs. satellite or gauge. There is quite a room for improvement in reservoir representation in GHMs. But simply mentioning that refining models, better data and additional factors is way to short for this paper. Why are some reservoirs are very well represented by the models and some not e.g. why is seasonality at Hoover dam so bad and why does it fit so well for Toledo. You might say that was not the purpose of your paper, but chapter 3.1 is not really new, so you have to put something into chapter 3.2.

If GHM developers could access time series of volumetric reservoir storage data for hundreds of major reservoirs in the world, they would have already partially answered the question. In reality, as described in detail in the Introduction, we experienced difficulties in preparing ready-to-compare global reservoir storage data. This paper attempts to demonstrate the scale of this challenge and highlights a way forward. This is because we do not think that the lack of data is a reason for not doing anything. We, therefore, show that DAHITI and GRSAD have good value and can be used as part of the solution to the challenge. The challenge will also be addressed in the near future, as higher-res satellite data becomes available for longer time periods (as discussed earlier in the response). We noted that further promising datasets are emerging, (e.g., Li et al., 2023; Cooley et al., 2021; Hou et al., 2022).

Finally, reproducing the seasonality at Hoover Dam is not an easy task under the current simulation framework. First, the inflow to the Hoover Dam needs to be well simulated. It is challenging for GHMs to simulate the inflow of individual reservoirs because their parameters have been optimized with certain limitations (see Yoshida et al. 2022 for further details). Second, because the Hoover Dam (Lake Mead) is located downstream of Lake Powell (the Glen Canyon Dam), the reservoir operation of Lake Powell also needs to be well simulated. Third, human water withdrawal needs to be accurately reproduced and incorporated in simulations because water withdrawal substantially alters the streamflow of the Colorado River beyond the Hoover Dam (US Department of Interior Bureau of Reclamation, 2012)

References

Busker, T., De Roo, A., Gelati, E., Schwatke, C., Adamovic, M., Bisselink, B., Pekel, J. F., and Cottam, A.: A global lake and reservoir volume analysis using a surface water dataset and satellite altimetry, Hydrol. Earth Syst. Sci., 23, 669–690, https://doi.org/10.5194/hess-23-669-2019, 2019.

Cooley, S. W., Ryan, J. C., and Smith, L. C.: Human alteration of global surface water storage variability, Nature, 591, 78–81, https://doi.org/10.1038/s41586-021-03262-3, 2021.

Hou, J., van Dijk, A. I. J. M., Beck, H. E., Renzullo, L. J., and Wada, Y.: Remotely sensed reservoir water storage dynamics (1984–2015) and the influence of climate variability and management at a global scale, Hydrol. Earth Syst. Sci., 26, 3785–3803, https://doi.org/10.5194/hess-26-3785-2022, 2022.

Li, Y., Gao, H., Allen, G. H., and Zhang, Z.: Constructing Reservoir Area–Volume–Elevation Curve from TanDEM-X DEM Data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 14, 2249–2257, https://doi.org/10.1109/JSTARS.2021.3051103, 2021.

Li, Y., Zhao, G., Allen, G. H., and Gao, H.: Diminishing storage returns of reservoir construction, Nat. Commun., 14, https://doi.org/10.1038/s41467-023-38843-5, 2023.

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O.: Estimating the volume and age of water stored in global lakes using a geo-statistical approach, Nature Communications, 7, 13603, 10.1038/ncomms13603, 2016.

Müller Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., Oki, T., Portmann, F. T., Reinecke, R., Riedel, C., Song, Q., Zhang, J., and Döll, P.: Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use, Hydrol. Earth Syst. Sci., 20, 2877-2898, 10.5194/hess-20-2877-2016, 2016.

Telteu, C. E., Müller Schmied, H., Thiery, W., Leng, G., Burek, P., Liu, X., Boulange, J. E. S., Andersen, L. S., Grillakis, M., Gosling, S. N., Satoh, Y., Rakovec, O., Stacke, T., Chang, J., Wanders, N., Shah, H. L., Trautmann, T., Mao, G., Hanasaki, N., Koutroulis, A., Pokhrel, Y., Samaniego, L., Wada, Y., Mishra, V., Liu, J., Döll, P., Zhao, F., Gädeke, A., Rabin, S. S., and Herz, F.: Understanding each other's models: an introduction and a standard representation of 16 global water models to support intercomparison, improvement, and communication, Geosci. Model Dev., 14, 3843-3878, 10.5194/gmd-14-3843-2021, 2021.

U.S. Department of the Interior Bureau of Reclamation: Colorado River Basin Water Supply and Demand Study Executive Summary, U.S. Department of the Interior Bureau of Reclamation, 28, 2012.

Verma, K., Nair, A. S., Jayaluxmi, I., Karmakar, S., and Calmant, S.: Satellite altimetry for Indian reservoirs, Water Sci. Eng., 14, 277–285, https://doi.org/10.1016/j.wse.2021.09.001, 2021.

Yoshida, T., Hanasaki, N., Nishina, K., Boulange, J., Okada, M., and Troch, P. A.: Inference of Parameters for a Global Hydrological Model: Identifiability and Predictive Uncertainties of Climate-Based Parameters, Water Resources Research, 58, e2021WR030660, https://doi.org/10.1029/2021WR030660, 2022.

Zajac, Z., Revilla-Romero, B., Salamon, P., Burek, P., Hirpa, F. A., and Beck, H.: The impact of lake and reservoir parameterization on global streamflow simulation, Journal of Hydrology, 548, 552-568, https://doi.org/10.1016/j.jhydrol.2017.03.022, 2017.