**Response to Reviewers**

**''Remotely sensed reservoir water storage dynamics (1984-2015) and the influence of climate variability and management at global scale'' by Jiawei Hou et al.**

**We thank the editor and the second reviewer for the thoughtful comments and constructive suggestions, which helped us improve the manuscript. We have thoroughly considered all comments and suggestions, and made revisions as outlined below (reviewer comments in blue, our response in black bold font).**

**Editor Comments:**

EC1) Your interesting work received two more reviews of which #1 is satisfied ans has no remarks but #2 has still fundamental reservations on several aspects of the methodology of lake volume time series which seem valid points to me. The aspect of the influence of image quality should be addressed and some kind of assessment of the accuracy needs to be given. I also agree that a Pearson coeff R >0.7 is not that a strong correlation. This requires some more argumentation then "Messager et al 2016 used this as well" , i.e. why the authors put the bar on R2~0.5 and why that is acceptable for this global study. These aspects need some more critical discussion in the ms both in methodology and discussion section.

**Thank you so much for your time and effort in handling this manuscript.**

**We have added further analysis regarding image quality. Because the MODIS 8-day composites provide temporally more dense and reliable information, we used the MODIS-derived lake product (Tortini et al., 2020) to investigate the influence of Landsat image quality on the volume time series estimation. Zhao and Gao (2018) regarded images with a contamination ratio ranging from 5% to 95% as 'poor-quality' images and developed an automated fill-gaps method to restore these 'poor-quality' images. We divided reservoir surface water time series (Zhao and Gao, 2018) into four groups (Table S3) according to contamination percentages and validated them against the MODIS-derived water product (Tortini et al., 2020) for 100 lakes. The results (Table S3) show that images with 5-35% contaminated ratio do not affect the temporal accuracy (R=0.87) of lake volume estimation after applying the gap-filling method. The performance drops slightly to R=0.80 when contaminated ratio increases to 35-65% but does not decrease further between 65% and 95%. Overall, the performance of lake volume estimation using gap-filled images (contaminated ratio ranging from 5% to 95%) is the same as using good-quality images, thanks to the gap-filling method.**

**Table S3** The average performance of lake volume estimation using either good quality Landsat images or gap-filling images (original ones with different contamination percentages).

| Contamination Percentages | <5% | 5% - 35% | 35% -65% | 65% - 95% |
|---|---|---|---|---|
| R | 0.86 | 0.87 | 0.80 | 0.80 |
| | 0.86 | | 0.87 | |

**We added this table into the Supplementary and included this analysis in L251-255:**

*"We investigated the influence of Landsat image quality on the volume time series estimation by comparing time series derived from images with different contamination ratios (0~95%) against the MODIS-derived lake product (Tortini et al., 2020). The temporal accuracy slightly decreases as the contamination ratio increases (Table S3). However, the overall performance of lake volume estimation using images with contaminated ratio ranging from 5% to 95% is commensurate to using only good-quality images, thanks to the gap-filling method."*

**We agree that the higher hypsometric correlation we used, the less uncertainties they have. But the trade-off is that using stricter filtering would lead to less data available, especially for a global analysis. The less information used, therefore the less reliable the conclusion. Among the 58 reservoirs with correlations above 0.7, 29 reservoirs have R≥0.9 while 13 have R between 0.7 and 0.8. These data are mainly used for trend analysis, in which we used the annual mean value and total volume in a basin. The uncertainties from the hypsometry will therefore decrease to a much lower level in this temporal and spatial aggregation.**

**There is no strict rule to determine which Pearson correlation threshold between A-L or A-V should be considered before calculating lake volume. Tortini et al. (2020) used R≥0.85 and Busker et al. (2019) used R≥0.9, while Gao et al. (2012) used R≥0.5 (including 0.76 for Port Peck, 0.72 for Sakakawea, 0.83 for Mead and Oahe, and 0.66 for Powell). However, note that the sample size (N) varied between these studies, which is the other variable in addition to the correlation coefficient (R) that affects the interpreted statistical significance via the p value. In our study, the average number of samples (monthly A-L pairs) for each lake is around 166. We calculated significant Pearson correlation threshold using:**

$$t = R\frac{\sqrt{N-2}}{\sqrt{1-R^2}} \quad \textbf{(S1)}$$

**The result suggests the linear relationship is significant (p<0.01) when R is above 0.19. We also calculated p values using satellite-derived water extents and heights for each lake individually. In all cases, p < 0.01 when R was above 0.18. Therefore, we argue 0.7 is a relatively conservative high correlation in our cases.**

**We added sentences in L131-136 in the Data and Methods Section to explain why we chosen R≥0.7:**

*"The interpretation of Pearson correlation depends on p value and the number (N) of samples. Among these 132 reservoirs, the average number (N) of sample (i.e., the monthly pairs of extent and height) is around 166. We used a significance level of p<0.01 to determine the corresponding Pearson correlation threshold with the t-test (Eq. S1). The result suggested that the linear relationship is significant when R is above 0.19. In this context, we conservatively considered R≥0.7 as evidence of strong correlation. Such a strong correlation between extent and height was found for 58 reservoirs (Group A; Fig. 1)."*

**We add sentences in L426-432 in the Discussion to the selection R≥0.7:**

*"The higher hypsometric correlation we used, the less uncertainties volume estimations would have (Crétaux et al., 2016). We selected correlation threshold of 0.7 in this study, which is lower than Tortini et al. (2020) (R≥0.85) and Busker et al. (2019) (R≥0.9), but higher than Gao et al. (2012) (R≥0.5). The selection of an appropriate correlation threshold can also depend on the purpose of the study. Tortini et al. (2020), Busker et al. (2019) and Gao et al. (2012) aimed to provide accurate measurements for an individual reservoir. Here, our priority is to understand the 32-year volume trend at basin scale. The uncertainties from the individual hypsometry (0.9≥R≥0.7; total 29 reservoirs) therefore average out by temporal (i.e., annual) and spatial (i.e., basin) aggregation."*

**[1] Crétaux, J.-F., Abarca-del-Río, R., Berge-Nguyen, M., Arsen, A., Drolon, V., Clos, G., & Maisongrande, P. (2016). Lake volume monitoring from space. Surveys in Geophysics, 37, 269-305**

EC2) Then I personally found it hard to always follow when it is about numbers of reservoirs, total stored volume or percentages of those. I think it would help the reader if you would report that in a consequent way (e.g. always give number and stored volume) and the percentages of those.

**Thank you for your suggestion. We modified the corresponding sentences and reported number/volume and its percentage in a consistent way in L182-194 in the revised manuscript:**

*"There are 6,862 reservoirs reported in the GRanD database (Lehner et al. 2011), with the total 6,196 km$^3$ reported storage capacity. In this study, we were able to estimate monthly storage dynamics for 6,695 or 97.6% of the total number of reservoirs, with 3,941 km$^3$ or 63.6%of cumulative capacity (Fig. 1). There were only 58 (0.8%) reservoirs for which storage dynamics could be estimated most directly, by a combination of satellite extent and water level observations (Group A), but together they already represent up to 1,394 km$^3$ (22.5%) storage capacity (Fig. 1). The total capacity of the 172 (2.5%) reservoirs not measured constitutes 2,255 km$^3$ (36.4%) of storage capacity. There were 6,637 (96.7%) reservoirs in Group B for which by the geo-statistical approach could be applied, and their total capacity is 2,547 km$^3$ (41.1%). To ensure consistency in the 1984-2015 time series used for long-term trend analysis, we ignored reservoirs with less than 360 months (i.e., 30 years) of Landsat-derived observations or for which more than five years of water extent observations were inter- or extrapolated by Zhao and Gao (2018). Our focus was on interactions between precipitation, streamflow, evaporation and storage in existing reservoirs, rather than the consequences of new impoundments. Therefore, we excluded from consideration all reservoirs that were destroyed, modified, planned, replaced, removed, subsumed or constructed after 1984. This left 4,573 (66.6%) reservoirs available for with combined storage capacity of 2,583 km$^3$ (41.7%) (Fig. 1)."*

**Reviewer #2 Comments:**

I appreciate the authors' efforts on addressing my comments. The revised manuscript reduces parts of my previous concerns. But I still hold my major concerns as some of the responses are not very solid. Please see my comments below.

R2C1) The authors claim that for 5,917 reservoirs that have Landsat observations every month from 1984 – 2015. However, most of the monthly estimates were generated from poor-quality images. I know Zhao and Gao did a comparison between estimated areas and directly observed areas for all reservoirs. They did not provide a comprehensive assessment on the fidelity of the time series for each individual reservoir. This is less a concern for their study as they were interested in aggregate global reservoir areas. I agree that using Zhao and Gao's approach, you can get monthly (or near-monthly) area time series. Here, you attributed storage change in each reservoir case by case. Do the estimates from poor-quality images accurately capture the monthly variability for each reservoir? I think it would be helpful to compare the accuracy of areas from good quality images vs bad. It is seemingly less confident for areas estimated from only a fraction (e.g., 5-10%) of their ROI. Including additional estimates from poor quality images may be fine but also needs to consider the accuracy from these "poor" estimates.

**Thank you for your suggestion. In line with Zhao and Gao (2018), we aggregated reservoir volumes at basin scale before carrying out trending analysis, which reduce uncertainties from individual reservoir. In addition, we did analysis on the influence of Landsat image quality on the volume time series estimation - please see our response to comment EC1.**

R2C2) I am not sure why the author chose to use R instead of R2. A perfect A-V relationship should have a R2 of 1. Busker et al studied 137 lakes and found 58 water bodies having a hypsometric relationship with a R2 > 0.8. I suggested the authors to compare their results with Busker et al. I do not understand why they chose to use R rather than R2 and a lower threshold (0.7). In your studied 132 large reservoirs, 58 water bodies have a hypsometric relationship with a R2 > ~0.5. It seems that your hypsometries have significantly larger uncertainties compared with theirs. As uncertainty in hypsometry critically matters the accuracy of the volume estimate (Cretaux et al. 2016). I am not convinced that their reservoir volume estimates are based on the state-of-art approaches.

Crétaux, J. F., Abarca-del-Río, R., Berge-Nguyen, M., Arsen, A., Drolon, V., Clos, G., & Maisongrande, P. (2016). Lake volume monitoring from space. Surveys in Geophysics, 37(2), 269-305.

**Thank you for this comment. We discussed on why we chosen R≥0.7 in response to the editor, please see our response to comment EC1.**

R2C3) Only a fraction (<3%) of studied reservoirs have observed water levels. I would recommend the authors pay a better attention to the concepts here. A reference volume is a static value, rather than the observed volumes (e.g., by a gauging station). The reported values do not make sense as they cannot justify the fidelity of the volume time series for each reservoir. Why not using observed storages from in-situ stations to validate your estimates? I know at least hundreds of U.S. reservoirs are gauged. As this empirical approach has been applied to the vast majority (>97%) of reservoirs and this approach is

not well validated, I still concern about the quality of the generated datasets. Is your approach remote-sensed based or an empirical based or a mix of them?

**Indeed, the volume is estimated based on a geo-statistical model for the vast majority (in terms of number) of reservoirs (Group B). Specifically, lake depth can be extrapolated from surrounding topography if we have satellite-derived surface water extent observation. The validation results mainly cover lakes from Group B. To validate more reservoirs more comprehensively, we contacted Australian Bureau of Meteorology again and got the updated in-situ stations covering more reservoirs in Australia. We also accessed more in situ reservoir storage records from US Bureau of Reclamation via https://www.usbr.gov/uc/water/hydrodata/. We included these updated validations in L239-245:**

> *"In situ monthly storage records from the US Army Corps of Engineers, US Bureau of Reclamation and Australian Bureau of Meteorology were used for error assessment. There are totally 131 reservoirs with at least 20-year overlapped time series between in situ data and satellite-derived data. We did validation for all these 131 reservoirs (5 for Group A and 126 for Group B). The averaged correlation between observed and estimated volumes is 0.82 (R>0.7 for 82% of the 131 reservoirs). Messager et al. (2016) reported that the symmetric mean absolute percent error (SMAPE) of the geo-statistical model is 48.8% globally. In our study, the average SMAPE between predicted and reference volumes was 32.13%, lower mainly because we adjusted reservoir storage estimates by reported reservoir capacity."*

R2C4) Do the 65 reservoirs cover a fraction of reservoirs with 20-year in-situ data or a subset? I mean you validated on ~1% of reservoirs and recommend justify they are representative. I am surprised that the authors used r to present the accuracy. A perfect r can still be associated with under- or over-estimation. How many of the 65 reservoirs cover the water bodies without water levels?

**In the revised, expanded validation results, we validated overall 131 reservoirs with at least 20-year overlapped time series between in situ data and satellite-derived data (5 for Group A and 126 for Group B). Please see our response to comment R2C3. We provided both R and SMAPE to assess the accuracy of lake volume estimation. However, we indeed more forced on correlation coefficient (temporal accuracy) because this study aims to understand reservoir storage trends, rather than absolute values (L177-179).**

R2C5) I do not understand why you need this assumption: Assuming the estimation method for Group A is more accurate than that for Group B, the latter can also be evaluated against the former. Why not grasp available in-situ data for the validation on Group B?

**Apologies for the confusion. We did use in-situ data to validate Group B. Please see our response to comment R2C3. And this part means we did cross-validation between Group A and Group B. We modified this sentence to clarify this point in L248-249 in the revised manuscript**

> *"In addition, we did cross-validation between Group A and Group B. The results show that 25 of the total 33 overlapping estimated reservoirs show strong agreement (R≥0.9) between the two*

*methods, and the average SMAPE between them is 13.1%. This implies good consistency of reservoir storage estimates from Group A and B. Some cross-validation examples are shown in Fig. 3."*

R2C6) The authors claimed that they explicitly considered the difference between coincidence and causation in our study, which does not seem to be the case. How confident can you attribute the reservoir storage change based on the trending consistency of in-situ river flow and storage? How about in-situ flow only explaining <50% of the storage trend? Additionally, in-situ river flow can be affected by human activities. I am not convinced that the authors examined the causation confidently.

**We agreed that trend consistency is not enough to examine the causation confidently. In this study, we provided four lines of evidence (including analysis of net evaporation and global water withdrawal data) to explore causation (L81-94), rather than only looking at trend consistency. To emphasize this, we summarized the causation analysis in L486-494 in the Conclusion.**

*"We provided four lines of evidence to explore which factor (precipitation, net evaporation, or dam (demand-related) water releases) drives the global reservoir storage trends. First, we found trend consistency between precipitation, streamflow and reservoir storage. Second, we found robust temporal correlation between precipitation, streamflow and reservoir storage. Third, we inferred the role of human activity based on the reservoir water balance equation: because we found changes in net evaporation only accounted for a small fraction of reservoir volume changes, together with the first two lines of evidence, we can infer that dam (demand-related) water releases are less likely to be the main driver of storage changes. Fourth, we examined water use data and did not find that increasing water use corresponded to deceasing reservoir storage, or vice versa, in the majority of basins. Therefore, we conclude that reservoir volume changes are dominated by (multi-decadal) precipitation changes."*

R2C7) As stated earlier, large reservoirs are highly regulated by humans. The proposed approaches have limited capacity to capture the human reservoir regulations, which could falsely sign the dominant importance of natural climate variability.

**We agreed that this study has limited capacity to understand the influence of human reservoir regulations as there is no available global dam release data. However, we used an indirect approach to infer the role of human activity (L356-370). Please see our response to comment R2C6. We also argued that human interventions can affect seasonal variability in reservoirs (Cooley et al., 2021) but not necessarily multi-decadal trends (L444-448).**

R2C8) Decreased evaporation could be a result of decreased water surface. Comparing the trending directions between these two does not tell whether E is the driver or not.

**In terms of net evaporation, we did not compare trend consistency. We can calculate how much of storage <u>changes</u> can be explained by changes in net evaporation. The results showed that changes in net evaporation accounted for well below 10% of the overall trends in storage (Fig. 10). This means there is little influence of net evaporation on reservoir storage changes.**

R2C9) Why not using a water mass balance approach? Isn't more robust compared with the seemingly simplified comparisons of trending directions?

**We cannot use the water mass balance approach directly as there is no in situ or simulated inflow and water release data for all individual reservoirs. Instead, we analyzed the influence of climate variability and human activity on the basin-scale reservoir storage changes based on reservoir water balance (Eq. 9) in an indirect way (L356-370). Please see our response to comment R2C6.**

R2C10) "There were only 58 reservoirs for which storage dynamics could be estimated most directly, by a combination of satellite extent and water level observations (Group A), but together they already account for 25.5% of combined global reservoir capacity (Fig. 1). The total capacity of the 193 reservoirs not measured constitutes 36.4% of global capacity. There were 6,611 reservoirs in Group B for which by the geo-statistical approach could be applied, and these contribute 41.1% to total global capacity."

Aren't your conclusions based on the large number of reservoirs which storage estimates have the largest uncertainty due to the empirical method?

**We have performed validation for both Group A and Group B and added more validation results for Group B in the revised manuscript. The result indicated that the generated dataset has a robust temporal accuracy, which is sufficiently reliable for trend analysis. Please see our response to comment R2C3. We also performed cross-validation between Group A and Group B. The result showed that they have a comparable level of temporal accuracy. Please see our response to comment R2C5.**

R2C11) "Our focus was on interactions between precipitation, streamflow, evaporation and storage in existing reservoirs, rather than the consequences of new impoundments. Therefore, we excluded from consideration all reservoirs that were destroyed, modified, planned, replaced, removed, subsumed or constructed after 1984."

As you focus on old reservoirs, does sedimentation affect the estimated storage? It seems to be a significant issue for old reservoirs.

**We agree that sedimentation can also contribute to the decrease of reservoir water storage but the effect of sedimentation on our global 32-year analysis will be small, according to Wisser et al. (2013) (L437-440).**