<u>Response to Reviewers</u>

''GloLakes: water storage dynamics for 27,000 lakes globally from 1984 to present derived from satellite altimetry and optical imaging'' by Jiawei Hou et al.

We thank the two reviewers for their thoughtful comments and constructive suggestions, which helped us improve the manuscript significantly. We have thoroughly considered all comments and suggestions, and made modifications accordingly below (review comments in blue, our responses in black bold font).

In summary, we added several validation analyses in the revision below:

(1) We extended our water storage validation analysis from 238 lakes to 494 lakes (now including in situ data from USA, Australia, South Africa, India, and Spain), and showed correlation and bias results for all 494 lakes.

(2) We performed uncertainty analysis of the geo-statistical model in the water storage validation analysis for 21 lakes as examples. Uncertainties sources include observed lake area (the omission and commission errors of surface water mapping from GSWD used in this study are 5% and 1%, respectively; Pekel et al. (2016)) and DEM (±10 m; Robinson et al. (2014)).

(3) We compared lake mean volume from our product against in situ data for 494 lakes to examine any systematic bias in our data.

(4) We validated near real-time (NRT) storage estimates for 9 lakes where in situ data include NRT information and analysed the change in the performance of storage estimation between NRT and historical methods.

(5) We cross-validated NRT storage estimates between our different approaches and showed the performance of NRT storage estimation for lakes with different sizes.

(6) We compared lake water area time series from our product against Zhao and Gao (2018) and Donchyts et al. (2022) for 5318 and 11101 lakes, respectively, in terms of correlation and bias, and showed the comparisons of lake mean area between our product and these two published studies in the 1:1 relationship to examine the uncertainties of our derived lake area time series.

(7) We provided a summary as to where and for how many lakes NRT storage can be estimated by remote sensing in each basin around the world. This will provide key information for the newly launched Surface Water and Ocean Topography (SWOT) to monitor storage changes in global lakes as it will be able to measure both extent and level on a single satellite platform.

(8) We showed examples of comparisons between our relative storage product and Tortini et al. (2020) and correlation results for all evaluated lakes.

We rephrased several paragraphs in Abstract, Introduction, Data and method and Conclusion sections, and reconstructed Results and Discussion Section to have the following sections:

(1) Lake area estimation validation

(2) Feasibility of near real-time lake monitoring

(3) Validation of historical lake storage estimates

(4) Validation of near real-time lake storage estimates

(5) Future opportunities

Following the reviews, we also made some changes to our dataset:

(1) **We implemented the ICESat-2 quick look data into our GloLakes product to improve monitoring abilities of ICESat-2 based storage estimations**

(2) **We released a new product to derive lake storage changes directly using satellite-derived extents (Sentinel-2) and levels (ICESat-2)**

(3) **The GREALM has now included Sentinel-6 data. This was included in our product as well**

Response to Reviewer #1 Comments:

Hou et al present a dataset of global water storage variations. This paper is unique in that it attempts to construct absolute storage variability time series, which are challenging to produce. It also aims to fuse together multiple freely available datasets in a novel approach. However, I find it to overall be a flawed manuscript and dataset which needs many necessary improvements (see major comments below). In brief, I am concerned that the authors have not accurately described the dataset, both in regards to the long term and NRT storage dynamics and have not performed sufficient validation analyses. The paper is also poorly written in places, with numerous typos and grammatical errors as well as paragraphs that are poorly structured, and the figures are weak and do not sufficiently illustrate the dataset. Without substantial changes to the manuscript, presentation and perhaps the dataset itself, I'm not sure this paper and dataset would be of value to the broader community.

We thank the reviewer for the detailed and valuable comments and suggestions, which enabled us to greatly improve the quality of our manuscript. Below please find our response to reviewer's comments in detail.

In the revised manuscript, we carefully and accurately distinguished between "estimate" and "measure" to describe the different approaches to derive lake water storage time series. We also added the results of several more validation analyses to strengthen the manuscript.

We apologise for any typos and grammatical errors. We carefully examined and corrected them in the revised manuscript. We also added several more figures and restructured paragraphs in several sections to better illustrate our product.

Major Comments

R1C1) Estimation of long term storage dynamics. Throughout the text, the authors state that they 'measured' long term absolute storage dynamics and/or 'produced' absolute storage time series. To be clear, what the authors did was apply a geostatistical model to estimate water depth and then use this to estimate a absolute storage time series when combined with a Landsat-derived dataset of lake extent. There was thus no 'measurement of storage' here – what the authors did was 'estimate' storage based on statistical relationships. While there is nothing wrong with estimating using these geostatistical relationships, it is imperative that this is explained correctly and consistently throughout the paper so as not to cause confusion with other methods which actually calculate volume change based on water level observations.

Agreed. We carefully revised the manuscript and make sure that we consistently use "estimate" where a geostatistical model was used while retaining "measure" where storage is calculated from water extent and level observations, to indicate the lesser uncertainty in the latter. Please see the changes in the revised manuscript.

R1C2) ICESat-2 data is not NRT. The authors state the importance of Near Real Time (NRT) lake monitoring with a latency of ~1-10 days. They then include ICESat-2 as one of the potential datasets to use for NRT monitoring. However, this indicates a fundamental misunderstanding of ICESat-2 and how it is processed. First of all, ICESat-2 has a repeat time of 91 days, so on average you get an observation ~once every three months (though this does vary based on the size of the lake). Second of all, unlike with say MODIS, Landsat, or Sentinel-2, ICESat-2 data is not immediately released. Currently (as of Oct 9, 2022), the most recently available ICESat-2 data is through June 8th, 2022, and this has been fairly consistent over the past few years (ICESat-2 releases data about every ~6 months). While the NSIDC ICESat-2 website is perhaps a little misleading that it says data is available up to the present, so I understand some of the confusion, simple playing around with the data will quickly reveal the extremely long latency of ICESat-2 products. It is thus very much inaccurate to use ICESat-2 as a potential NRT water volume estimator in the method described here.

We agree that ICESat-2 cannot provide sufficient near real-time (NRT) information if we used the "ATLAS/ICESat-2 L3A Along Track Inland Surface Water Data, Version 5 (ATL13)" product because their data were not updated regularly when new observations obtained. In using the term NRT, we referred to the time between the ICESat-2 data becoming available and our use of it to produce storage estimates, but we appreciate this can be a misleading use of words. We also thank the reviewer for referring us to the "ATLAS/ICESat-2 L3A Along Track Inland Surface Water Data Quick Look, Version 5 (ATL13QL)" product which can provide NRT water height observations (R1C21). In the revision, we implemented the ICESat-2 quick look data into our GloLakes product to improve monitoring abilities of ICESat-2 based storage estimations and made changes in the revised manuscript accordingly. We also included how we process ATL13 and ATL13QL data and combine them together to derive NRT water height observations and discussed the advantages and disadvantages of using these two products in the revised manuscript.

We were aware that ICESat-2 has a low temporal frequency of around 91 days. However, this kind of temporal frequency can still provide updated and useful seasonal variations of global lakes. In revising, we clarified that NRT monitoring for ICESat-2 based storage estimations has latency to update each of lake storage but that we produce a rapid update global lake data *once* there are new observations.

Another important aspect of using ICESat-2 in this study was to investigate where lake storages can be measured by satellite-derived extent and level simultaneously, as it has very dense coverage of global lakes, thus overlapping a large number of lakes with Landsat observations. These will provide key information for the newly launched Surface Water and Ocean Topography (SWOT) to monitor storage

changes in global lakes as it achieves to measure both extent and level on a single satellite platform. Please see detailed response to R1C5.

Based on all above, we modified the paragraph in L134-154 in Section 2.1.2 (Surface water height) in the revised manuscript:

> *"We used the ATLAS/ICESat-2 L3A Along-Track Inland Surface Water Data, Version 5 (ATL13) (Jasinski and Ondrusek, 2021). Unfortunately, updated ATL13 data are not released until 30-45 days after new observations are obtained, which does not suit NRT monitoring. Therefore, we also used ATLAS/ICESat-2 L3A Along Track Inland Surface Water Data Quick Look, Version 5 (ATL13QL), available within three days of new observations. ATL13 and ATL13QL apply the same algorithms to derive surface water height measurements for inland water bodies, including rivers, lakes, reservoirs and coastal water. The approach is described by Jasinski and Ondrusek (2021), but in brief, the procedure: 1) identifies ATLAS beams that intersect inland water body shape masks; 2) collects photons in short segments (~100 m) for calculating water height and in longer segments (1~3 km) for estimating and removing subsurface backscatter; 3) applies the physical and statistical modelling to derive inland water heights. Cooley et al. (2021) used ATLAS/ICESat-2 L3A Land and Vegetation Height (ATL08) to derive lake height time series and found the mean absolute error between USGS gauge data and ICESat-2 height measurements is 0.14 m. Unlike ATL08, the ATL13 used here was designed to measure water height variations in rivers and lakes as it considers physical processes of light propagation in open water bodies. An error of 6.1 cm per 100 inland water photons has been reported (Jasinski and Ondrusek, 2021). We collected both ATL13 and ATL13QL water surface height data from any of the six beams within individual lake boundaries from HydroLAKES and derived water height for each lake observed by ICESat-2 between 2018 and the present. ATL13QL has larger uncertainties (~100 m) in geolocation than ATL13 (~5 m). As a result, segment heights from ATL13QL are 2.7 m, with a standard deviation of ~ 7 m, lower than those from ATL13. According to the user guide (Jasinski and Ondrusek, 2021), 2.7 m can be added to ATL13QL before merging these two products, and the differences between them have little impact on measuring relative heights. Each lake will be revisited by ICESat-2 around every 91 days. This means our ICESat-2-based storage product is limited to providing information on seasonal variation rather than short-term dynamics. However, it has the benefit of providing observations with a rapid turnaround."*

R1C3) No validation of NRT data. While the authors do perform validation for 238 lakes for what appears to be the geostatistically estimated historical time series, it is not explicitly stated whether they perform any validation of the NRT time series (both the absolute and relative). More detailed information on the accuracy of the NRT time series, and how it varies between using V-H vs. V-A relationships (or how it varies by lake size, if possible), is required to be able to evaluate this dataset.

We thank the reviewer for this comment. It is a challenge to find publicly accessible in situ data especially with NRT information. Of course, this is one of the reasons we are using remote sensing in the first place.

We would like to first emphasize that NRT storages were only estimated if V-H or V-A relationships were significantly correlated (please see the detailed response to R1C5). After a search, we found NRT data for nine sites and used these to validate our NRT product. In addition, we performed cross-validation between NRT estimates from our different storage products (ICESat-2, Sentinel-2, and GREALM), and summarized results for lakes with different sizes (please note this result is a bit different from the response in the discussion phase as there are updates and changes in the datasets). We included these new validation results and added a new section 3.4 (Validation of near real-time lake storage estimates) in the revised manuscript as follows:

> *"To our knowledge, we produced the first four decades-long data set of relative and absolute lake storage dynamics from 1984-present for all HydroLAKES registered lakes with an area exceeding 1 km^2. Previous studies have focused on measuring relative lake storage dynamics using remote sensing. For example, Busker et al. (2019) used Landsat imagery and radar altimetry to estimate lake volume variations for 137 lakes. Similarly, Tortini et al. (2020) produced a storage change dataset for 347 lakes but using MODIS instead of Landsat imagery. Neither study offered an NRT monitoring capacity. They were also limited to a few hundred lakes due to the sparse track of radar altimetry. With denser coverage, the ICESat-2 laser altimetry improves the number of measured lakes by a factor of more than hundred. Based on this, we were able to monitor a wide range of lakes at global scale.*
>
> *We validated our ICESat-2 and Sentinel-2 derived water storage time series from 1984 to present against in situ data that can help to assess the accuracy of both our historical and NRT methods (Fig. 9). We compared the performance of water storage estimates between the respective methods used to derive historical and NRT estimates. The average R and SMAPE for the method, evaluated for 1984-2020 with storage estimated using a geostatistical model, were 0.94 and 22%. The performance for the NRT method, evaluated for 2020-present with storage estimated using a V-H relationship or V-V merging, decreased slightly to R=0.88 and SMAPE=31%, respectively. For most lakes, the agreement between predicted and observed storages shows similar accuracy (in terms of correlation and bias, Table 2) for the historical and NRT methods, as NRT estimates were only produced if the V-H or V-A relationships were significantly correlated. For example, both ICESat-2 and Sentinel-2 derived products record sharply decreasing in water storage in Big Sandy Reservoir, Vega Reservoir, and El Vado Lake in the NRT method period (Fig. 9). However, the performance of the NRT method decreases for Lake Summer, Horsetooth Reservoir and El Vado Lake as there was one erroneous observation (Fig. 9).*
>
> *We compared NRT estimates from our different storage products (ICESat-2, Sentinel-2, and GREALM) and summarised the results for lakes of different sizes (Table 3). The mean R and SMAPE between ICESat-2 and Sentinel-2 products were 0.83 and 7%, respectively. The ICESat-2 and GREALM products had a similar correlation (R=0.81) and bias (SMAPE=7%), although they are both derived from altimetry. The evaluation results generally did not vary with lake size, except that ICESat-2 and Sentinel-2 products show better correlation (R=0.92) and less bias (SMAPE=2%) for lakes with an area greater than 100 km^2."*

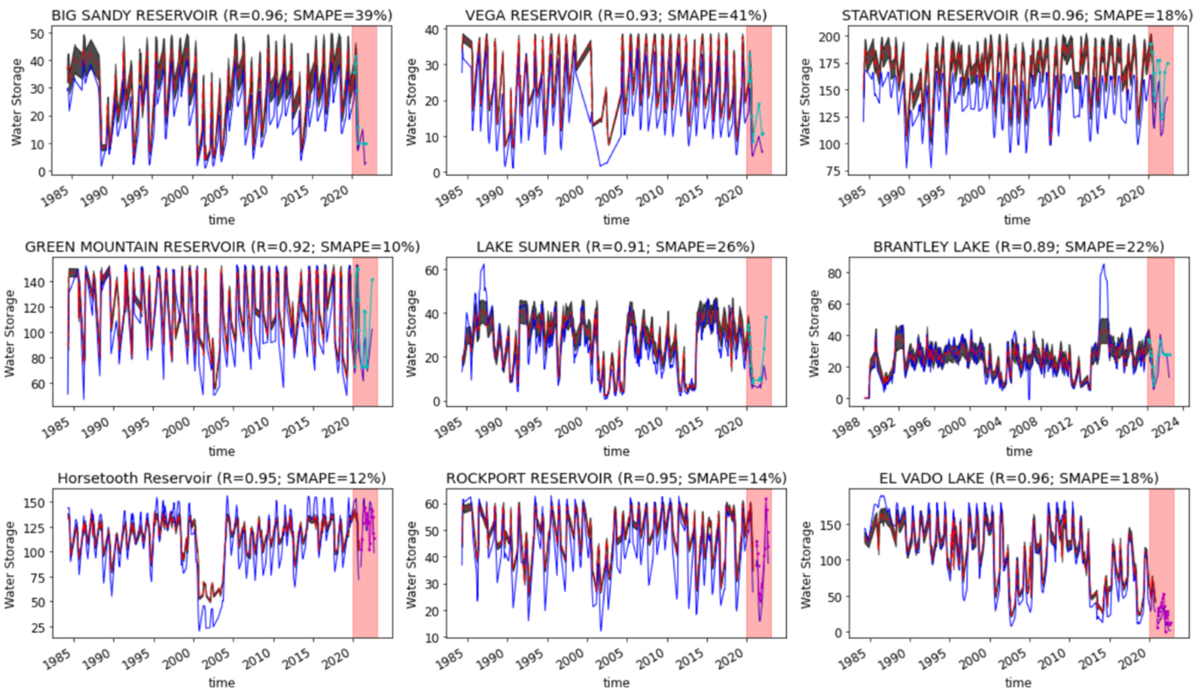

Figure 9 Comparisons of historical and NRT lake water storage time series from GloLakes against in situ (observed) data (blue line: observed data; red line: historical storage estimates; black shade: error bars of historical storage estimates; light pink shade: NRT monitoring period; cyan line with dots (first two rows): NRT storage estimates from ICESat-2; magenta line with dots (third row): NRT storage estimates from Sentinel-2).

Table 2 Comparisons of correlations and biases between historical and NRT methods (results from Fig. 9).

Lake Name	NRT Monitoring Approach	Historical (1984-2020)		NRT (2020-present)	
		R	SMAPE	R	SMAPE
Big Sandy	L-H; ICESat-2	0.96	39%	0.94	53%
Vega		0.93	41%	0.98	52%
Starvation		0.96	18%	0.97	18%
Green Mountain		0.92	10%	0.93	13%
Sumner		0.91	26%	0.81	43%
Brantley		0.89	22%	0.89	21%
Horsetooth	L-A; Sentinel-2	0.95	12%	0.56	8%
Rockport		0.95	14%	0.99	19%
El Vado		0.96	18%	0.81	48%

Table 3 Cross-validations of NRT method between our different storage products for lakes of different sizes.

	Landsat + ICESat-2							
	Landsat + Sentinel-2				Landsat + GREALM			
	<10 km^2	10-100 km^2	>100 km^2	All	<10 km^2	10-100 km^2	>100 km^2	All
Total number of overlapped lakes	1247	612	92	1951	7	22	63	92
R	0.83	0.83	0.92	0.83	0.84	0.84	0.80	0.81
SMAPE	8%	7%	2%	7%	7%	11%	5%	7%

R1C4) Need for large error bars and more details about the approach. The paper glosses over the specifics of the statistical method (pointing towards another paper) used to contruct the historical time series, but more details on this method should be provided to enable the reader to better understand the method and its potential accuracy. The dataset should also include significant error bars on the resulting time series, or at least clear information that these are all estimates with on average ~50% error. This is even more important for the NRT estimations, as given the use of look-up tables to approximate V-H/V-A relationships, these should have even greater error than the original volume time series (see comment about need for validation above)!

Thank you for these suggestions. We included error bars considering the uncertainties from the geo-statistical model (Fig. 5 and 9). And we modified Section 2.1.3 (Geostatistical model) to better explain the method and how we assess its uncertainties:

> *"The HydroLAKES database provides the boundary outlines of more than 1.4 million individual lakes globally with a surface area above 0.1 km^2 (Messager et al., 2016). This database is compiled from several global and regional lake and reservoir datasets derived using topographic maps, optical remote sensing imagery composites and radar instruments. HydroLAKES also comprises a geostatistical model with parameters to predict average water depths and volumes based on surrounding topography information for lakes with a surface extent between 0.1 km^2 and 500 km^2. The geostatistical model (Messager et al., 2016) is described as follows:*
>
> $$Log_{10}(D) = C_1 + C_2 \times Log_{10}(A) + C_3 \times Log_{10}(S_{100}) + s^2 \qquad (1)$$
>
> *where D is the predicted mean depth (m), A the observed surface area of the lake (km^2), S_{100} the average slope (derived from DEM) within a 100-m buffer around the lake, C_1, C_2 and C_3 coefficients estimated from fitting the model using global lake data for different lakes sizes (i.e., 0.1-1 km^2, 1-10 km^2, 10-100 km^2, 100-500 km^2), and s^2 the residual variance. The fundamental assumption is that one can extrapolate the slope around the lake towards the centre of the lake to estimate lake depth. The traditional approach uses a DEM within the lake to derive A-H curves but cannot retrieve true land surface elevation below water. In comparison, the model used here considers the DEM outside lake rather than inside it and hence is not affected by water inundation at the time of DEM acquisition. Messager et al. (2016) reported that the symmetric mean absolute percent error between predicted and reference volume is 48.8% without significant bias in volumes for the majority of lakes around the world, with the exception of Finland, Sweden and northwestern Russia, the European Alps, and the Andes. We compared this approach with alternative approaches. Khazaei et al. (2022) developed a statistical model, GLOBahty, relating lake depth with surface water area, elevation, volume, shoreline length, and watershed area. They demonstrated that the performance of this statistical model is better than previous approaches to calculating lake depth by assuming the lake shape fits one of four geometries: box, cone, triangular prism, and ellipsoid (e.g., Yigzaw et al., 2018). We compared both models to in situ data. The resulting SMAPE error for the GLOBahty method was 70.5%, and therefore not better than the geostatistical method. This result informed our choice to use the simpler geostatistical method in favour of the GLOBathy dataset in this study. The uncertainties*

from the geostatistical model are mainly from the observed surface area and the DEM used to derive slope. The omission and commission errors of surface water mapping from GSWD used in this study are 5% and 1%, respectively. EarthEnv-DEM90 was used in the geostatistical model to derive slope and its vertical accuracy is around 10 m (Robinson et al., 2014). We assessed how these errors propagate into water storage estimation in the validation analysis."

We extended our validation analysis from 238 lakes to 494 lakes. In the revised manuscript, we had water storage validation results across USA, Australia, South Africa, India, and Spain. The additional in-situ water storage data were found in a recent paper by Donchyts et al. (2022). We compared predicted mean water storage against observed data (compared to a 1:1 relationship) for 494 lakes to illustrate a ~50% error (Fig. 6). We also displayed all validations results (R and SMAPE) for 496 lakes (Fig. 4), to help readers better understand the accuracy of our product. Please see the detailed changes in the Section 3.3 (Validation of historical lake storage estimates) in the revised manuscript.

[1] Robinson, N., Regetz, J., & Guralnick, R. P. (2014). EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. ISPRS Journal of Photogrammetry and Remote Sensing, 87, 57-67.
[2] Donchyts, G., Winsemius, H., Baart, F., Dahm, R., Schellekens, J., Gorelick, N., ... & Schmeier, S. (2022). High-resolution surface water dynamics in Earth's small and medium-sized reservoirs. Scientific reports, 12(1), 1-13.

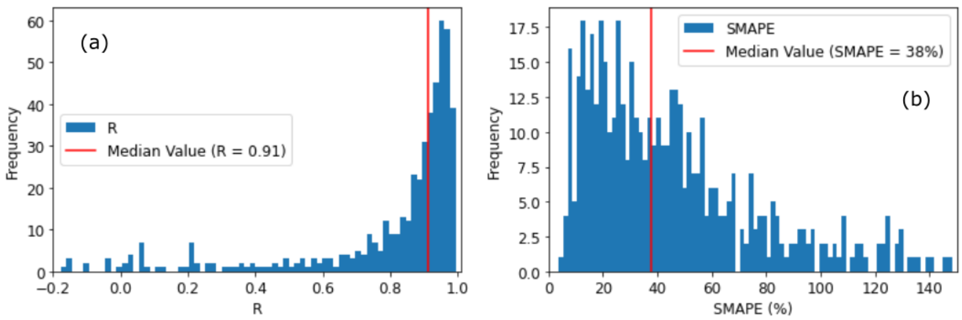

Figure 4 The distribution of R and SMAPE of absolute water storage validation results against in situ data for 494 lakes (vertical red line: the median value).

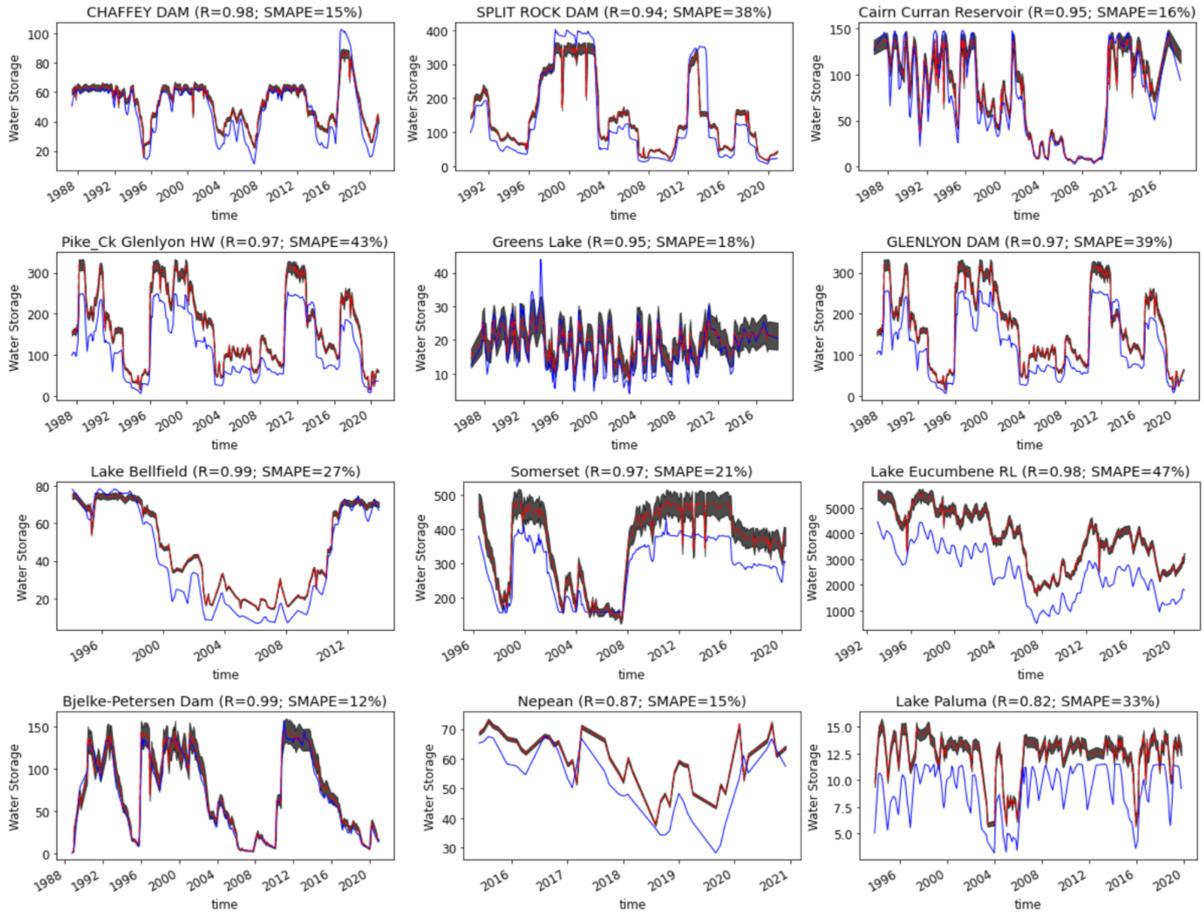

Figure 5 Absolute lake water storage time series from GloLakes against in situ (observed) data for selected lakes (blue line: observed data; red line: historical storage estimates; black shade: error bars of historical storage estimates).

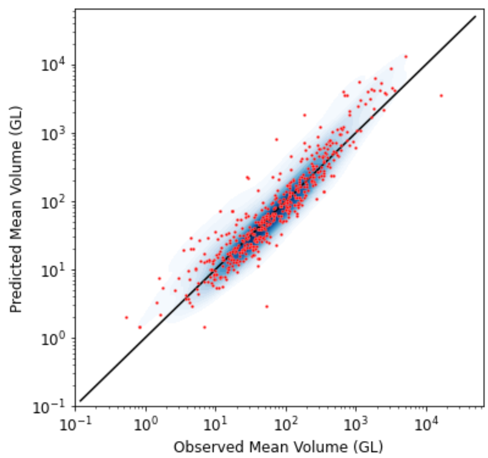

Figure 6 Scatterplots of predicted mean lake volume vs. observed mean volume (red dot: mean lake volume; black line: 1:1 relationship; background blue colours indicate data density).

R1C5) Poor quality figures. The figures included are weak and do not provide sufficient info about the dataset. Figure 3 is very difficult to interpret (what is the difference between panel (b) and panel (c))? Figures 1 and 2 are useful but are very large and all 4 examples for each may not be needed. Figure 4 is useful and depicts the accuracy analysis well. I'd suggest adding additional figures illustrating the accuracy of the geostatistical estimation of water depth (and perhaps the resulting volume time series). It also would be useful to provide more information about where is was possible to build NRT and relative time series, and where it was not possible (perhaps in a figure? Or table?).

We apologise for the confusion. Fig. 3b and 3c showed the percentage of total basin water volume for which water storage dynamics were estimated in this study. The percentage is calculated in each basin by Equation R1:

$$p = \frac{\sum_{i=0}^{m} V_i}{\sum_{i=0}^{n} V_i} \qquad \text{(R1)}$$

where *p* is the percentage of total lake water volume in each basin, *m* the number of lakes for which water storage dynamics were estimated in a basin, *n* the total number of lakes from HydroLAKES in a basin, and V_i the water volume of individual lake from HydroLAKES. The difference between these two figures is that Fig. 3b counted the lakes that have been observed by either optical remote sensing or altimetry instruments while Fig. 3c calculated those whose historical time series were estimated.

In the revised manuscript, we deleted Fig. 3b and 3c and added Fig. 3b-d to emphasize one of highlights in this study. To make it easier to interpreted, we calculated the percentage of total number of lakes rather than total volume explained above in the new Fig.R3b-d. Fig. R3b-d shows where and for how much of the lakes NRT storages can be measured by remote sensing in each basin around the world. These will provide valuable information for the newly launched Surface Water and Ocean Topography (SWOT) to monitor storage changes in global lakes as it achieves to measure both extent and level on a single satellite platform. We added a new section 3.2 (Feasibility of near real-time lake monitoring) to explain this analysis in the revised manuscript:

> *"We estimated monthly lake volume dynamics for 170,611 lakes during 1984–2020 (Fig. 3a). Most lakes are in the northern hemisphere's high latitudes, especially in North America. We only considered lakes with an area greater than 1 km^2 given the resolution of the Landsat imagery. There are more than a million lakes with areas between 0.1–1 km^2 that remain unmeasured, but the combined storage of these small lakes only accounts for 1% of global total lake water storage, according to the HydroLAKES dataset. Overall, a large portion of global lake volume for the period 1984-2020 was measured. Here we focus on estimating both historical and NRT lake dynamics. NRT lake water storage was only estimated if A-H-V relationships were sufficiently strong, even though we produced historical water storage change for more than 170,000 lakes. This also ensures that satellite-derived extent and level observations are consistent in characterising lake change before producing lake storage time series from 1984 to the present.*
>
> *As the Landsat, Sentinel-2 and ICESat-2 have comprehensive coverage of global lakes, we investigated if any two of them (optical + altimetry or two optical) can be used to monitor global lakes and how much of the lakes in each basin storage can be measured in each case. This should provide valuable information for the newly launched Surface Water and Ocean Topography (SWOT) mission that aims to monitor storage changes in global lakes by measuring both extent and level on a single satellite platform. The*

HydroBASINS database delineates watershed boundaries worldwide at different basin or catchments scales (Lehner and Grill, 2013). The catchment boundaries in the Pfafstetter level-3 product from HydroBASINS were used as the spatial basin units for analysis. If remote sensing measurements from two sources (derived extent and level or derived storages) are significantly correlated for a particular lake, we considered that this lake can be monitored by Earth observation. We followed three complementary approaches: (1) geostatistical model (Landsat + Sentinel-2), (2) H-V relationships (Landsat + ICESat-2), and (3) extent and level observations (Sentinel-2 + ICESat-2). Approach (3) would likely be the most reliable as storage is directly measured by extent and level from remote sensing. However, approaches (1) and (2) make it possible to estimate absolute storage changes. Landsat and ICESat-2 together could measure lake water storage in nearly all (i.e., 234 out of 292) river basins worldwide (Fig. 3b). Satellite-derived extents and levels are significantly correlated for more than one-fourth of lakes in each of the 145 basins. This feature is evenly distributed across the continents, except for Antarctica and high northern latitudes, due to the influence of frozen water surfaces. Sentinel-2 and ICESat-2 cover 122 basins globally (Fig. 3d). There are 58 basins where over half of the lakes can be monitored by that method, mainly located in the USA, southeastern South America, the Mediterranean, southern Africa, southern Asia, and Australia. Landsat and Sentinel-2 both measure surface water extent and show strong time series consistency in 63 out of 124 basins. Three-quarters of lakes have a significant A-H relationship (Fig. 3c) with a distribution pattern similar to that of Sentinel-2/ICESat-2."

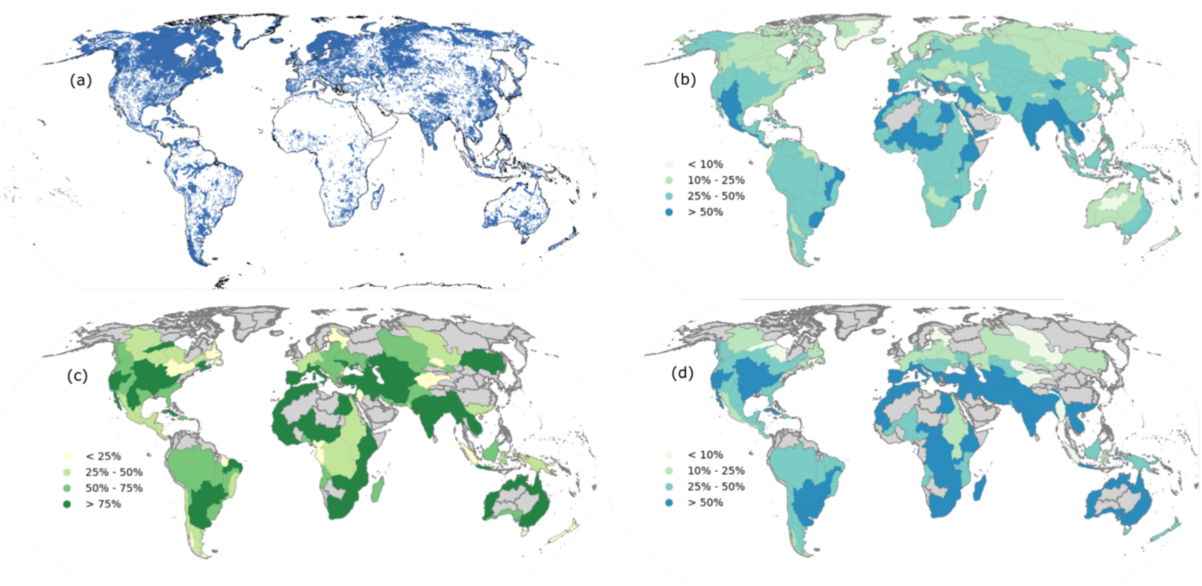

Figure 3 The locations of 170,611 lakes whose storage dynamics for the period of 1984-2020 were estimated in this study (a), the percentages of lakes (in terms of number) whose NRT water storage dynamics can be derived using Landsat and ICESat-2 (b), Landsat and Sentinel-2 (c), and Sentinel-2 and ICESat-2 (d) in each basin.

On the other hand, Fig.1 and Fig.2 showed how our gap-filling algorithm addressed missing Landsat data in different scenarios (cloud cover, SLC failure, mix issues, different contamination ratios, different shapes of lake). Following the reviewer's comment, we clarified this in L280-283 in the revised manuscript and moved Fig. 2 to the Supplemental Material.

Besides we showed examples (Fig. 8) of comparisons between our relative storage product and Tortini et al. (2020) and correlation results (Fig. 7) for all evaluated lakes in L382-387 in the revised manuscript as follows:

> *"We compared our relative lake storage dynamics data against the global surface water storage change time series for 1992–2018 produced by Tortini et al. (2020) for 104 common lakes. The data was downloaded from the Physical Oceanography Distributed Active Archive Center (PODAAC) of NASA's Jet Propulsion Laboratory. The two show strong agreement (Fig. 7) with a median R of 0.94 and 95% of R values above 0.7. Selected time series are compared in Fig. 8 and show that both datasets are most limited by the temporal frequency of image acquisition."*

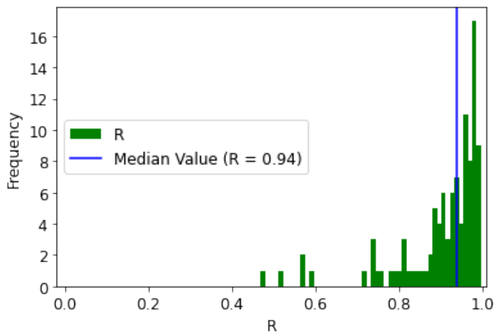

Figure 7 The distribution of R of relative water storage validation results against Tortini et al. (2022) for 104 lakes (vertical blue line: the median value).

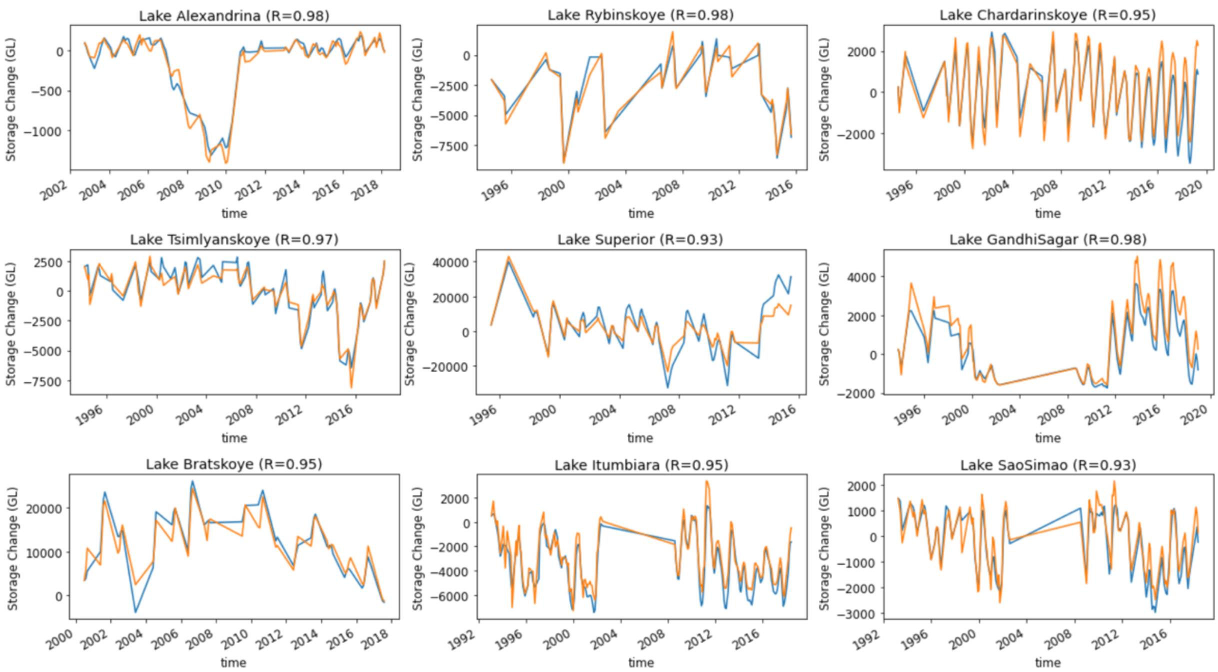

Figure 8 Relative lake water storage time series from GloLakes (blue line) against Tortini et al. (2022) (brown line) for selected lakes.

R1C6) Why not include Landsat and Sentinel-2 (not BLUEDOT) as NRT? Given that the authors calculate NRT storage variations simply by building V-H or V-A relationships with the results of their Landsat and geostatistical depth model volume time series, it should be possible to construct NRT time series for thousands more lakes globally by just classifying water in Landsat-8/9 and Sentinel-2 (and this would actually be NRT, unlike ICESat-2) While I understand that part of the point of this paper is to use existing datasets, classifying water in Landsat-8/9 and Sentinel-2 is pretty darn standard and straightforward at this point, particularly with the existence of cloud platforms like GEE. I'm not suggesting the authors do this globally, more just pointing out that this approach (while likely quite inaccurate) could in theory be used for thousands more lakes. Also, why (for the Sentinel-2 data in particular) do you need to use a lookup table to estimate volume via a V-A relationship – surely you could use the same geostatistical model to calculate volume from the Sentinel-2 (or Landsat 8/9) area observation?

Thank you for this suggestion. Regards to GEE, we agree that we could directly use Sentinel-2 to derive water extent for thousands more lakes, but it was indeed well beyond the scope of this study. In the revised manuscript, we discussed the limitation of using BLUEDOT data and how this might be addressed using Sentinel-2 and GEE in future in L452-456 in the revised manuscript as follows:

> *"Sentinel-2-based lake area data are currently already available from BLUEDOT for several thousand lakes worldwide, and there is no fundamental limitation that would prevent a similar approach from measuring all 170,957 lakes measured by Landsat in this study. The advantage of Sentinel-2 would be that it can provide NRT lake observations with low latency. The high computation and storage demands can potentially be met by cloud platforms like Google Earth Engine (GEE). In future research, we hope to consider such approaches to improve our data set."*

We would also like to clarify that we did in fact use the geostatistical model to estimate water storage for both of Landsat and Sentinel-2 and then merged them together to derive historical and NRT water storage time series. We modified Section 2.2.2 (Global near real-time lake volume estimation) to clarify this point and all the other NRT storage estimation approach in the revised manuscript as follows:

> *"To provide routine and low-latency measurements of lake water storage, we obtained NRT satellite-derived water heights and extents from different monitoring satellite sources, including ICESat-2), US Department of Agriculture (USDA) GREALM and BLUEDOT (Sentinel-2) water observatory (Table 1). We selected all estimates for lakes whose volume dynamics from 1984–2020 were derived from GSWD with the geostatistical model. We examined pairwise relationships (i.e., V-H or V-A) of overlapping monthly time series between historical volume (V) and NRT height (H) or area (A). We only extended historical volume estimation to NRT monitoring when there was a statistically significant ($p<0.05$) correlation between the predictor and predictand. Specifically, we calculated the significant Pearson correlation threshold (R_t) using a t-test allowing for sample size N (i.e., the number of data pairs). If the Pearson correlation (R) between V-H or V-A exceeded R_t we used H or A, respectively, to estimate lake storage dynamics after 2020.*

Depending on some of the features of the satellite data sources (Sentinel-2, ICESat-2, GREALM), we used different approaches to estimate absolute NRT storage. If both historical and NRT sources (i.e., Landsat and Sentinel-2) only observed lake area changes, we converted lake area to storage using the geostatistical model (Eq. 1) and merged them to derive storage time series from 1984-present. If NRT satellite sources (i.e., ICESat-2 or GREALM) observed lake water height changes, we developed the relationship between historical volume (V) and NRT height (H) for the overlapping months and used that V-H to extend historical volume estimation to NRT monitoring using only satellite-derived H. As the relationship is not necessarily linear, NRT lake storage was estimated using cumulative distribution function (CDF) matching. A look-up table was developed to rank all historical H and V, allowing one to be estimated from the other based on the ranking. Overall, we derived absolute water storage dynamics from 1984 onwards for 24,865 lakes using ICESat-2, 129 lakes using GREALM, and 4,054 lakes using Sentinel-2.

Our main objective was to estimate NRT absolute water storage dynamics for lakes worldwide as much as possible. However, we also measured relative storage where radar or laser-derived surface water heights, as well as Landsat-derived surface water extent, were available (i.e., Landsat with ICESat-2 or Landsat with GREALM). Unlike absolute storage products, relative storage products are unaffected by any bias from the geostatistical model. We produced alternative storage datasets to provide options to users requiring different bias tolerances or different applications. For the relative storage products, we calculated historical storage changes for lakes where the A-H relationship was significantly correlated as follows (Crétaux et al., 2016):

$$\Delta V = \frac{(H_t - H_{t-1}) \times (A_t + A_{t-1} + \sqrt{A_t \times A_{t-1}})}{3} \qquad (2)$$

where ΔV is storage change between two consecutive measurements; H_t and H_{t-1} are altimetry-derived surface water heights at time t and t-1, respectively; A_t and A_{t-1} are optical remote sensing derived surface water extents at time t and t-1, respectively. Where there are only H data available in the NRT period, we estimated corresponding A using CDF matching and calculated relative storage change time series using Eq (2). As Sentinel-2 and ICESat-2 can measure NRT surface water extent and level, respectively, we used them to derive relative storage changes from 2018-present. This product is free from any errors in the geostatistical model and in the A-H or V-H relationships used in the other products. Overall, we measured relative lake storage dynamics from 2018 onwards for 24,990 lakes using ICESat-2 and Landsat, for 2,740 lakes using ICESat-2 and Sentinel-2, and from 1993 onwards for 227 lakes using GREALM and Landsat."

Considering that Sentinel-2 and ICESat-2 can measure surface water extent and level, respectively, simultaneously, we produced an additional data set as part of this collection to derive lake storage changes directly rather than using the A-H or V-H relationships. Please see the details above. Below are some examples (Fig. R1) of H-A (Sentinel-2 + ICESat-2) relationships for the overlapping period (2016-present).

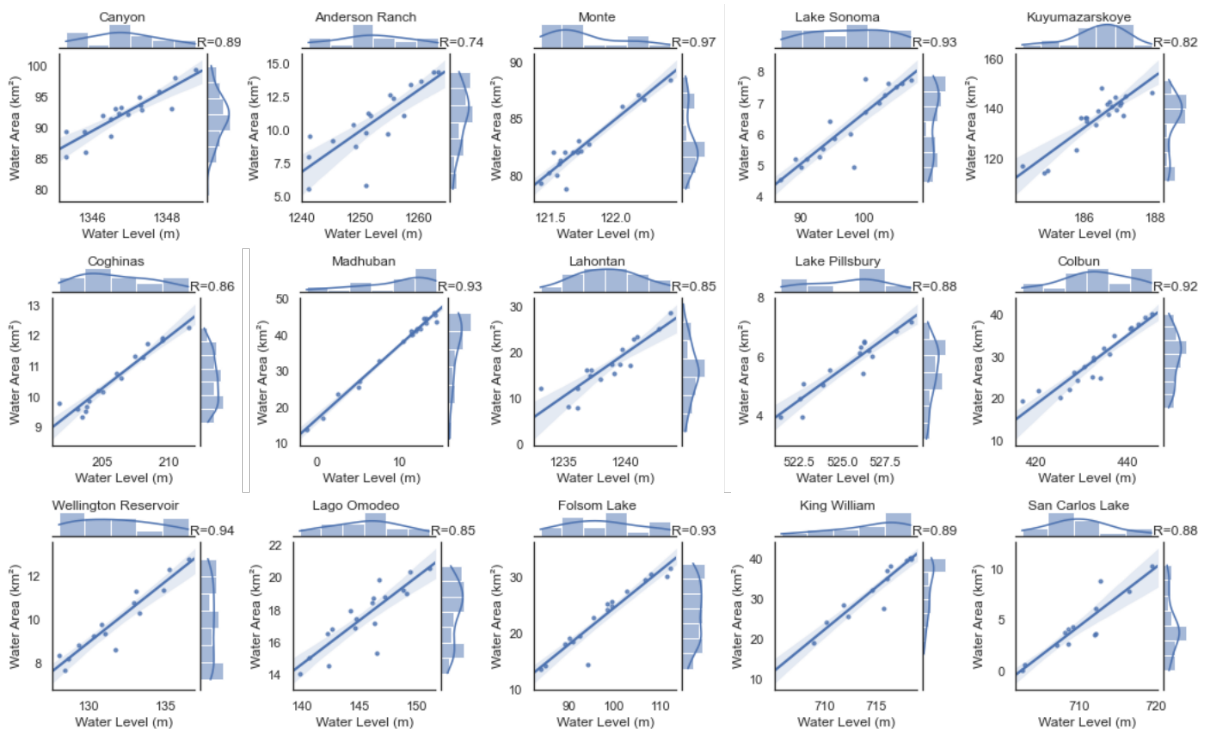

Figure R1 Some examples of lake *H-A* (Sentinel-2 and ICESat-2) relationships from 2016-present (blue line: fitting relationships; blue shade: 95% confidence interval; top axis: the distribution of water level; right axis: the distribution of water area).

Specific Comments

R1C7) Lines 6-7: The first sentence of the abstract is unnecessarily long and wordy and contain a typo. Please rephrase.

Thank you. We shortened this sentence in L6-7 in the revised manuscript:

> *"Measuring the spatiotemporal dynamics of lake and reservoir water storage is fundamental for assessing the influence of climate variability and anthropogenic activities on water quantity and quality."*

R1C8) Lines 56-74: Nice review of all of the different global water datasets, but this paragraph could use some restructuring as in its current form, this transition from the statement about ICESat-2 laser altimetry towards stating that "the spatial resolution of global satellite-dervied surface water dynamics projects…" doesn't make much sense, as the paragraph then moves to talking about measurements of surface water extent, not water level. I would suggest reorganizing this paragraph and being clearer about developments in observations of water level vs. water extent.

Apologies for the confusion. These are two separate paragraphs and we modified them in the revised manuscript.

R1C9) Line 85: Add "cannot measure lake depth and therefore are unable to measure absolute water volume without the use of bathymetric data"

Thank you. We added this in L86 in the revised manuscript.

R1C10) Lines 93-105: This paragraph is confusingly worded. The authors state that "Relative storage changes were estimated … while absolute storage changes were…" which is immediately followed by a statement that this was possible for more than 27,000 lakes worldwide. According to my understanding of the paper, the absolutely storage changes estimated using a geostatistical model were possible for all ~170,000 lakes, whereas the relative changes were possible for ~23,000 and NRT volumes (using the statistical model plus V-H or V-A relationships) were estimated for ~27,000. As currently written, this paragraph thus inaccurately describes the results.

Thank you. We rephrased this paragraph in L95-112 in the revised manuscript:

> *"The objectives of this study were to (1) derive nearly four decades of data on relative and absolute water volume measurements for lakes worldwide and (2) enable a global NRT lake monitoring capability. To develop this, we first estimated water body extents between 1984-2020 from Landsat-derived surface water maps for 170,957 lakes. We applied the gap-filling algorithm (Hou et al., 2022) in contaminated Landsat images to restore missing data, thus improving the total number of usable images to derive lake area time series. Second, we estimated absolute water storage dynamics from 1984-2020 for each lake whose water area and its surrounding slope measurements are available using a geostatistical model (Messager et al., 2016). Third, as Landsat does not provide NRT observations, we considered a range of alternative satellite data sources (including Sentinel-2, Jason-3, Sentinel-3 and -6, and ICESat-2) with monitoring abilities to derive NRT absolute lake storage. We examined where and for how many lakes NRT storage can be estimated using different combinations of these remote sensing data in each basin worldwide. Fourth, we extended historical absolute water storage estimates to NRT monitoring using the volume-height relationship if radar or lidar altimetry data available after 2020. Where only NRT lake water area observations (e.g., from Sentinel-2) were available, we converted lake area to storage estimates using a geostatistical model. As these absolute lake storage products are affected by the bias errors of lake depth estimates from the geostatistical model, we also provided relative lake storage estimates for lakes observed simultaneously by optical imagers and altimetry. Overall, this made it possible to monitor more than 27,000 lakes and reservoirs worldwide. The underlying strategy in developing this global lake monitoring system was to consider all readily available, validated, and frequently updated satellite data sources, explore the relative advantage of each source, and combine them to complement their respective weaknesses."*

R1C11) Line 97: Remove "Furthermore"

Thank you. We removed "Furthermore" as this sentence has been changed in the revised manuscript.

R1C12) Line 153: What is meant by "future GSWD water bodies"?

Apologies for the confusion. We can simply remove "future" in L195 in the revised manuscript

R1C13) Line 195: While NSIDC is where the ICESat-2 data is hosted, it is incorrect to call it a monitoring platform. Just call it ICESat-2 data.

Thank you. We called it ICESat-2 data consistently in the revised manuscript

R1C14) Table 1: It is incorrect to call NSIDC the platform for ICESat-2. I'm not sure exactly what term would be useful here, but I'd suggest simply calling the datasets something like ICESat-2, USDA G-REALM and BLUEDOT (Sentinel-2).

Agreed. We used ICESat-2, USDA G-REALM and BLUEDOT (Sentinel-2) as their names in the revised manuscript.

R1C15) Figure 3 caption (and throughout the manuscript). It is incorrect to state that "storage dynamics for the period of 1984-2020 were measured in this study". Given that storage was estimated based purely on statistical relationships that are likely to be highly inaccurate in many places (and with an overall error of ~50%), 'estimated' is the only appropriate word to describe this approach.

Thank you. We carefully corrected and used either "estimate" or "measure" to describe different approaches (see response to R1C1).

R1C16) Line 277: What is meant by "Overall, the relative volume dynamics are generally more reliable, as indicated by the correlation values"? Does this refer to the NRT time series? Or specifically the NRT time series calculated from A-H relationships? And correlation with what? The results on this accuracy are not reported in the paper (see major comment above).

Apologies. By this we mean that the validation results (predicted volume vs. observed volume) showed strong correlation so our product should be more reliable to indicate lake volume changes than absolute depth changes. We also produced relative storage products that avoid using the geo-statistical model and focus on providing accurate relative changes in lake storage. In addition, we added validations of the NRT estimates in the revised manuscript (see response to R1C3).

R1C17) Line 291: Again, please be clear here that these results are 'estimated'.

Agreed, see R1C15.

R1C18) Line 299: It is not necessary to state that Busker's dataset was not publicly available (you can always just email an author and ask for it, as technically every dataset should be publicly available on some level). I suggest removing this part of the sentence.

Agreed. We removed this in the revised manuscript.

R1C19) Line 343: SWOT will measure water height every 11 days (its orbit has a return period of 21 days, but since it is an off-nadir satellite, it will measure every water body every 11 days – see https://swot.jpl.nasa.gov/

Thank you. We changed it to "11 days" in the revised manuscript.

R1C20) Line 356: I explored the online Global Water Monitor linked in the paper. This is a cool way of displaying and communicating the data, and I commend the authors for putting it together, though it needs some cleaning up a bit (for example, what is the unit "GL" on the y axis for the annual time series of water volume)? If possible, it also would benefit from including error bars.

Thank you for this suggestion. The first version of the Global Water Monitor (GWM) has been released in January 2023 and we have updated the access link to web-based data explorer (www.globalwater.online) in the revised manuscript. Beyond this manuscript, we will consider all comments and suggestions and improve GWM in the next version in future.

Additional Comments:

R1C21) I wanted to lightly addend my review, specifically about the latency of ICESat-2, as I just learned about the existence of ICESat-2 Quick Look, low-latency products (https://nsidc.org/data/user-resources/help-center/faqs-icesat-2-quick-looks). My comments about the limited value of ICESat-2 for NRT data due to its 91 day repeat cycle still apply, but the existence of these quick look products does mean that it is technically possible to use the data within a few days after collection. I apologize for not acknowledging this before.

If the authors are to continue to use ICESat-2 data for their NRT dataset, I would advise explicitly discussing the quick look data as well as its advantages/disadvantages relative to the final product. Similarly, while the other two datasets mentioned in the NRT section (G-REALM and BLUEDOT) are relatively easy to download and process into lake height/area (since they already come processed to individual lake height/area) doing so with ICESat-2 is significantly more complicated as it requires choices around how to aggregate different tracks, filter out poor quality or outlier water level observations, etc. This additional difficulty should be described in the manuscript with additional details on how the authors process the ICESat-2 data automatically.

Thank you for these addition comments, see response to R1C2.

<u>Response to Reviewer #2 Comments:</u>

Hou et al. presented a nice study on creating a new time-varying dataset on global lake water storage. Understanding the lake storage variablity is critical for securing freshwater supplies. The authors leveraged multiple datasets to produce a global dataset with hundreds of thousands of lakes included, which seems to be an impressive work. However, I have a few major comments on the method and data quality.

We thank the reviewer for the thoughtful comments and constructive suggestions, which helped us to improve the quality of the manuscript. Below please find our response to reviewer's comments in detail.

R2C1) Pekel et al. used a global model to classify water and land. The authors depended on the Pekel et al.'s data to track water area changes in each lake locally. It is not clear to me whether the global model is suitable for studying each individual water body, for example, how can lake area change be accurately captured by the global model in each lake?

The GSWD dataset developed by Pekel et al. (2016) has been widely used to derive surface water area dynamics for individual lake in many studies (for example, Zhao and Gao (2018); Busker et al. (2019)). These previous studies have demonstrated the validity of using the GSWD dataset to study lake and reservoir changes. Pekel et al. (2016) also validated their product. The results showed that the omission errors of water mapping are less than 5%, while the commission errors are less than 1%. In the revised manuscript, we included these uncertainties from GSWD in the storage validation analysis. Please refer to the response to R1C4. In revision, we also included comparisons between our lake area change estimation and other published lake area datasets (i.e., Zhao and Gao (2018) and Donchyts et al. (2022)). Please see response to R2C2.

Most importantly, this study focuses on estimating both historical and near real-time (NRT) lake water storage dynamics. NRT lake water storage time series were only estimated if *A-H-V* relationships were significantly correlated, although we produced historical water storage change for more than 170,000 lakes (Fig. R5). This also means we make sure that satellite-derived extent and level observations are consistent in characterising lake change before producing storage estimates. We added a new section 3.2 (Feasibility of near real-time lake monitoring) to highlight this point in the revised manuscript. Please also see the detailed changes in R1C5.

R2C2) The monthly lake areas were generated from recovering water areas from contaminated images as in a previous publication by the authors (Hou et al., 2022). This is not new as quite a few recent studies have done a similar thing. As monthly lake areas are critical for generating monthly storage given monthly level data is pretty rare, the uncertainty of the recovered water areas seems to have non-negligible impact on the derived storage change. Had the authors assessed the uncertainty of areas from contaminated images? How did the generated time series compare with other existing approaches?

We thank the reviewer for this suggestion. In the revised manuscript, we included comparisons between our lake area change estimates and other published lake area datasets (i.e., Zhao and Gao (2018) and Donchyts et al. (2022)) (Fig. R7) in section 3.1 (Lake area estimation validation) as below:

"Our lake area data have 5318 lakes in common with the data of Zhao and Gao (2018) and 11,101 lakes with Donchyts et al. (2022) with average areas from 0.1 km^2 to 1000 km^2. Zhao and Gao (2018) used the same Landsat data source (GSWD) and the same lake boundary delineation (GRanD included in HydroLAKES) but a different gap-filling approach to derive lake area. Nonetheless, the mean lake area between the two products scatters closely around a 1:1 relationship (Fig. 2a). We also calculated correlation and bias in lake area time series for the common period 1984–2018 for each lake. The median R and SMAPE were 0.91 and 3.6%, respectively. This suggests that our gap-filling algorithm produces results overall similar to those of Zhao and Gao (2018). Donchyts et al. (2022) used a different Landsat data source, lake boundary delineation, and gap-filling algorithm to derive lake area time series. Despite these differences, mean lake area values still cluster fairly closely around the 1:1 relationship, especially for lakes greater than 1 km^2 (Fig. 2b). Larger biases exist for some lakes smaller than 1 km^2. This was caused mainly by different definitions of lake boundaries between HydroLAKES and Donchyts et al. (2022); we found that only one-tenth of lakes had boundary area differences within 20%. The median R and SMAPE in lake area time series from 1984 to 2020 for 11,101 lakes between our product and Donchyts et al. (2022) were 0.76 and 9.7%, once again lower mainly due to the differences for small lakes."

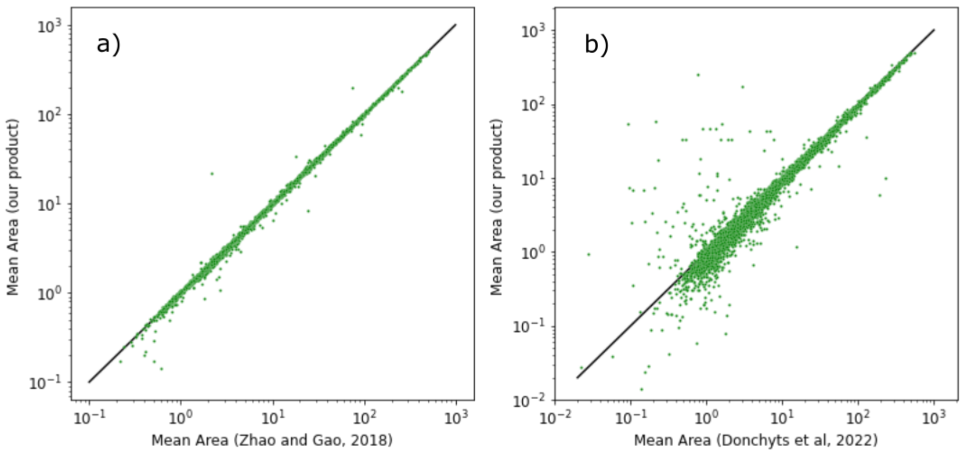

Figure 2 Scatterplots of mean lake area from our product vs. two other published datasets (black line: 1:1 relationship).

R2C3) I do not believe the Geo-statistical model used by Messager et al., 2016 to predict the total volume of a lake can be used to derive actual lake bathymetry here. The Geo-statistical model was based on global DEM products which have an uncertainty of several meters on average. Additionally, the water level conditions at the DEM acquisition time vary. As the used DEM data in Messager et al., 2016 cannot retrieve the true land surface elevation underneath water, I think this would introduce an even larger uncertainty (e.g., dozens of meters) when the authors extrapolated water levels beneath the level at the DEM acquisition date.

We agree that lake volume cannot be estimated very precisely unless detailed lake bathymetry data is available, but this is beyond current satellite remote sensing abilities. We emphasized this point in L378-379 in the manuscript. However, we respectfully disagree that the approach used in this study is affected by water level condition at DEM acquisition time, as the approach is different from the traditional DEM approach. The traditional DEM approach indeed is affected by water level condition at DEM acquisition

time as it used DEM within the lake to derive A-H curves. In comparison, the model used here considers DEM outside lake (i.e., extrapolating slope around the lake towards the centre of the lake to estimate lake depth) rather than inside, therefore it is not affected by water inundation at the DEM acquisition time. We highlighted this point in Section 2.1.3 (Geostatistical model) in the revised manuscript.

We did compare different approaches to estimate lake depth before choosing the geo-statistical model, but did not include the full analysis in this manuscript. Khazaei et al. (2022) develop a statistical model relating lake depth with surface water area, elevation, volume, shoreline length, and watershed area. They demonstrated that the performance of this statistical model is better than the previous approaches to calculate lake depth by simply assuming lake shape in four geometries: box, cone, triangular prism, and ellipsoid (e.g., Yigzaw et al. 2018). In our study, we compared this statistical model (Khazaei et al., 2022) with the geostatistical model (Messager et al., 2016) and compared them against in situ data. The resulting SMAPE error for the GLOBahty method was 70.5%, and therefore not better than the geo-statistical method. This led our choice to use the geo-statistical method in favour of the GLOBathy dataset in this study. We emphasised this analysis in the Section 2.1.3 (Geo-statistical model) in the revised manuscript.

[3] Yigzaw, W., Li, H. Y., Demissie, Y., Hejazi, M. I., Leung, L. R., Voisin, N., & Payn, R. (2018). A new global storage-area-depth data set for Modeling reservoirs in land surface and earth system models. Water Resources Research, 54(12), 10-372.

The geo-statistical model used in this study has uncertainties sources from DEM and observed lake area as demonstrated in Equation 1. The GSWD data used in this study to derive lake area has the omission and commission errors of 5% and 1%, respectively. The DEM used in geo-statistical model to calculate slope has an error of around 10 m. We have considered all these uncertainties in our storage validation results. In the revised manuscript, we modified Section 2.1.3 (Geostatistical model) to better describe the geo-statistical model and included uncertainties analysis in Fig. 5 and 9. Please also see the detailed changes in R1C4.

R2C4) The validation appears to be insufficient. How did the authors select the 238 lakes for the validation and why this is a comprehensive evaluation? Did the authors consider the performance of the method on cold-region lakes (e.g., in Canada). What the accuracy for smaller lakes given this method is only significant on small lakes as existing studies did a fairly good estimate on large lakes. Why the authors use relative metric R given R only gives a correlation estimate? For example, if the storage was scaled from area, no matter how large the level error is, the R value remains the same. Without a comprehensive validation, it is hard to foresee that the produced datasets would be useful for scientific inquiries.

Thank you for this comment. In the revised manuscript, we added several more validations analyses. Please see the details in the revision summary at the beginning of the author’s response.

Unfortunately, we could not find any in situ data in Canada or other cold regions, although we have extended our storage validation results to cover more regions around the world. However, we assessed where and how much of lakes in each basin whose storage can be measured by remote sensing. The

results showed that only a modest percentage of lakes can be monitored by our approach in northern high latitude regions due to the influence of frozen water surfaces in the cold season (see response to R1C5).

We assessed the accuracy of our products for lakes of a wide range of sizes. The evaluations not only include validation of lake area estimates (see response to R2C2), but also lake storage (see Fig. 6 in R1C4). The lakes for which area estimates were validated had average areas from 0.1 km^2 to 1000 km^2 while those for which storage was validated had volumes from 1 GL to 10000 GL. We emphasised these in L293 and L358-359 in the lake area and storage validation sections in the revised manuscript.

We calculated not only R (correlation) but also SMAPE (bias) to assess the accuracy of our product. Storage was not just scaled from area, in fact, monitoring approaches include three types: (1) geo-statistical model (e.g., Landsat + Sentinel-2); (2) H-V relationships (e.g., Landsat + ICESat-2); (3) direct extent and level (e.g., Sentinel-2 + ICESat-2). We clarified this in L328-331 in the revised manuscript.

Specific comments:

R2C5) Line 64: I would suggest replacing “a larger number” with the actual number.

Thank you. The actual number was 227,386. We changed this in the revised manuscript.

R2C6) Line 86: It seems that Avisse et al. is not the only study on estimating lake storage change based on DEM data. Maybe also highlight other relevant studies here.

Thank you. We included the additional studies below in the revised manuscript.

[4] Bonnema, M., Sikder, S., Miao, Y., Chen, X., Hossain, F., Ara Pervin, I., Mahbubur Rahman S.M., & Lee, H. (2016). Understanding satellite-based monthly-to-seasonal reservoir outflow estimation as a function of hydrologic controls. Water Resources Research, 52(5), 4095-4115.

[5] Vu, D. T., Dang, T. D., Galelli, S., & Hossain, F. (2022). Satellite observations reveal 13 years of reservoir filling strategies, operating rules, and hydrological alterations in the Upper Mekong River basin. Hydrology and Earth System Sciences, 26(9), 2345-2364.

R2C7) Line 280: the authors only show a few case studies for the validation. It would be better to have a figure or table to show all data used for the validation.

Thank you, agreed. We showed R and SMAPE validation results for all 494 lakes (Fig. 4; see response to R1C2) in the revised manuscript.