<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 invaluable time and dedicated effort in reviewing our manuscript. We have carefully considered all comments and suggestions provided by the reviewers and have made modifications accordingly (review comments in blue, our response in black bold font). In summary, the major changes include:

(1) We carefully addressed all terminology issues that might lead to misinterpretation throughout the manuscript.
(2) We added quality labels in the GloLakes dataset to clarify the NRT methods. Data users can interpret from the meta-data how the data were generated. For example, one can readily choose the gap-filled GSWD-geostatistical model dataset by simply opting Q1 label or a time range spanning from 1984 to 2020.
(3) We enhanced the discussion on the revisit time of different datasets used in this study.
(4) In order to enhance the validation analysis of our lake extent estimates, we used Google Earth Engine to manually delineate lake extents based on Sentinel-2 data and compared to our lake extent estimates derived from GSWD (Landsat).
(5) We collected in situ lake level data from Australian Bureau of Meteorology and Environment Canada and used them to validate altimetry data used in this study. We paid more attention to the cold region in both lake extent and height validation analysis in the revised manuscript.

<u>**Response to Reviewer #1 Comments:**</u>

The authors have responded to all reviewer comments in detail and clearly have taken care to improve the manuscript, and I commend them for their hard work and attention to detail. The manuscript has become clearer in places, including the description of the datasets and the greater emphasis on validation. However, the manuscript has a few areas where it still needs significant improvement.

First, there are important terminology issues that need to be addressed. As the manuscript is currently written, it is extremely misleading, as it does things like claim that Landsat is not NRT but ICESat-2 is, and also claim that ICESat-2 observes way more lakes than Sentinel-2. These things are true in the context of the datasets that are used in this paper (i.e. GSWD for Landsat, BLUEDOT for Sentinel-2, QL data for ICESat-2) but are not true generally, and so this language needs to be updated throughout the paper to prevent the paper from reading like a fundamental misunderstanding of remote sensing data. There's more detail in the comments below, but I'd suggest referencing the specific dataset, NOT the sensor, throughout the paper to clear up some of this confusion. Similarly, the authors refer to the GSWD-geostatistical model volume time series as the 'Landsat' time series in multiple places (including in Figure 3), which is also highly misleading and confusing. I also am still unsure about the value of the NRT analysis in this paper, and feel it may weaken the paper compared to the

GSWD-geostatistical model dataset. Lastly, the paper is missing some much needed discussion/results around frequency of observation and revisit times.

Thank you so much for pointing out the weakness of this manuscript. We tried our best to mitigate these issues in the revised manuscript and please see our detailed responses below.

Major Comments

R1C1) 1. Problematic use of data language and descriptions
There are significant issues with how the datasets are described throughout the paper. I understand that the reason so few lakes can be measured with Sentinel-2 is because the authors are not actually performing any classification themselves, but rather using the BLUEDOT dataset. I am not asking the authors to perform this classification, but I do think that every time the authors use the term 'Sentinel-2' in this paper, they should replace it with 'BLUEDOT'. It just seems very misleading to claim that ~24,000 lakes can be estimated in NRT with ICESat-2 vs. only ~4,000 lakes with Sentinel-2.

For one, this is really comparing apples to oranges – ICESat-2 has a revisit time of ~91 days, whereas Sentinel-2's is every 5 days, so time series estimated from Sentinel-2 would be much denser and more valuable. Not acknowledging this distinction throughout the paper is disingenuous. The 'science' value of a with a 91-day NRT revisit vs. a 5-day NRT revisit is wildly different, so this should be explicitly stated.

And secondly, it isn't the actual Sentinel-2 dataset that is being analyzed, rather a derived product. If the authors had actually classified all Sentinel-2 imagery over the HydroLAKES dataset, they would be able to build NRT time series for all ~170,000 lakes for which they estimated volume in the historical time series. Therefore, the results described here are quite misleading as they way understate the value of Sentinel-2 and way overstate the utility of ICESat-2 for NRT monitoring. At the very least, along with changing the terminology, these two issues (i.e. the difference in revisit time and the fact that the Sentinel-2 actually observes all lakes) should be discussed thoroughly in the discussion section.

Apologies for the potentially misleading wording. In the revised manuscript, we replaced "Sentinel-2" with "BLUEDOT (Sentinel-2)" to clarify both data sources. We also clarified the limitations of BLUEDOT data used in this study and the abilities of Sentinel-2 to observe all lakes around the world in L517-523 in the revised manuscript. In addition, we modified Section 3.5 about the differences in revisit time of different datasets used in this study and their original remote sensing observations as well (Please see the detailed response in R1C4).

We agree that we could classify all Sentinel-2 imagery and combine it with GSWD (Landsat) to produce both historical and NRT volume dynamics for all ~170,000 lakes. Thank you for this suggestion. We believe this could be a future study and implemented in a next version of GloLakes. We have highlighted this in L522-528 in the revised manuscript.

R1C2) Additionally, as noted in a specific comment below, the authors refer to the practice of deriving H-V relationships for the NRT data as 'Landsat-ICESat-2', but this is not a good way to describe what they are doing. They are (1) building a volume time series by combining the gap-filled JRC-GSWO (Landsat) dataset with the geostatistical model and then (2) correlating this with ICESat-2 observations to produce volume estimates from each ICESat-2 observation. This needs to be explicitly stated, and calling it a V-H correlation a 'Landsat-ICESat-2' correlation is highly misleading, because *Landsat does not measure volume*! This is also extra confusing when the authors describe V-A correlations as 'Landsat-Sentinel-2' since technically those both measure area…

Thank you for pointing this out. In the revised manuscript, we clarified that V represents the GSWD or BLUEDOT-geostatistical model volume time series and avoided mentioning GSWD (Landsat) and BLUEDOT (Sentinel-2) when describing V-H-A relationships or water volume.

R1C3) 2. Concerns about the NRT approach
I still feel very conflicted about the NRT method. If I'm understanding it correctly, the first part of the NRT method works by assessing the correlation between the historical volume data (which itself is based on a geostatistical model) and either NRT height or area, and then simply using a lookup table to convert the NRT data into volume time series. Most studies trying to use area and height data to calculate volume are specifically building curves between area and height and then extrapolating those (which is also done in this paper), whereas the geostatistical lookup table approach (i.e. correlated area OR height with estimated volume which itself is based on the combination of a geostatistical model and area observations), while applicable to large numbers of lakes, just seems to be overly complicated, especially as it is currently explained.

Also, I'm not sure what the value of this NRT analysis is (especially the ICESat-2 part), as it feels like it would be easier to just classify lake extent from any number of NRT optical sensors and then relate these to the geostatistical model to calculate volume if *actual* NRT data were needed, rather than trying to Frankenstein all these other datasets together, especially in way that significantly overstates the value of ICESat-2 and understates the value of Sentinel-2. If the goal is just to produce a global dataset of consistent estimates of lake volume, why not just wait until the GSWD data are updated? What value does having an NRT observation every ~91 days actually provide? (especially when you could just classify Sentinel-2 data and then have an observation for EVERY lake every 5 days…).

Overall, I understand and appreciate the authors' approach here – take a bunch of publicly available datasets, try putting them all together – but the end result as currently presented is misleading in places and I do question its scientific value (i.e. what will the NRT data actually be useful for). I personally think the paper and dataset would be stronger if the NRT section were removed entirely and this paper/dataset focused more on the fusion of the gap-filled GSWD data with the geostatistical model (which is a valuable contribution to the community), but I will of course leave that to the authors' and editor's discretion.

Thank you for your comments and suggestions. The main purpose of this study is to develop full time series of water storage estimates from 1984-present and to cover as many global lakes as possible by combining different freely accessible and regularly updated satellite data sources. We believe that the GloLakes dataset provides valuable long-term and updated information to understand past and current influences of climate change and variability and water management on natural lakes and reservoirs. We understand the reviewer's concern about the use of the term NRT for ~91 days old ICESat-2 observations and therefore replaced "NRT" with "up-to-date" for the ICESat-2 relevant estimates. We do argue that up-to-date (semi-seasonal, i.e. ~91 days) lake information from ICESat-2 is still valuable, however. We added a paragraph to highlight the value of the NRT (or up-to-date) lake information in L341-347 in the revised manuscript:

> *"The primary objective behind generating the NRT lake data is to evaluate present-day conditions in lakes and dams in their historical context. This becomes takes on ever greater importance in an era of rapid climate change. Of particular significance is the storage volume, as it enables the aggregation of numerous water bodies within a single system. Such an aggregation offers a more comprehensive and insightful perspective on the hydrological status of catchments or river systems, such as provided by initiatives like the Global Water Monitor (www.globalwater.online). Knowledge of combined stored volumes are also needed to understand concepts such as catchment water equilibrium and to interpret GRACE terrestrial water storage estimates for example, enhancing our understanding of global water availability."*

We appreciate your comment that the gap-filled GSWD-geostatistical model volume time series can be a valuable contribution to the community. We concur that our objective (namely, the estimation of NRT volume time series encompassing a significant portion of global lakes) can be accomplished through the classification of Landsat or Sentinel-2 imagery to delineate lake extents, followed by the application of a geo-statistical model to convert these areas into volumes. We hope to address this in our future research and include it in the next version of the GloLakes dataset.

We apologise for the confusion around the NRT methods, especially using the "lookup table". We argue that cumulative distribution function (CDF) matching as used in this study serves a similar purpose as building curves for V-A-H relationships. However, CDF matching is simpler compared to the traditional approach that involves selecting and fitting empirical equations (such as linear or nth degree polynomial equations). We modified the sentences in L256-263 to clarify CDF matching approach:

> *"Unlike the conventional approach to build V-A-H curves, which typically requires the selection and fitting of empirical equations (e.g., linear or higher-degree polynomial equations), CDF matching takes a simpler route. It entails the development of a monotonic cumulative distribution function for historical height (H) and volume (V) data. The distribution allows for the ranking of values, showing their relative positions in the dataset. When a new height value (H) is to be retrieved, CDF matching uses the pre-*

established cumulative distribution functions and their associated rankings. By comparing the rank of the new value with those in the distributions, the corresponding volume value (V) can be determined."

We decided not to entirely remove NRT estimates, as data users already have the flexibility to focus on the 1984-2020 timeframe in which period lake volume estimates were exclusively generated through the gap-filled GSWD (Landsat) and geostatistical model (Table 4). In addition, there is one product whose both historical and NRT lake volume dynamics were entirely estimated based on the geo-statistical model (Table 4). However, to clarify the different NRT methods and their corresponding uncertainties, we have included quality labels in the time series of all the GloLakes products in revision. This helps a user to interpret rapidly from the meta-data how the data were generated. The quality labels are:

"Q1: absolute volume estimated using geostatistical model and satellite-derived lake extents

Q2: absolute volume estimated based on the V-H relationship

Q3: relative volume estimated based on both satellite-derived heights and extents.

Q4: relative volume estimated based on heights obtained from satellite measurements, combined with the extents of the lake derived from the area-height (A-H) relationship."

We also marked these for different GloLakes products in the Table 4 in the revised manuscript below:

Table 4 Overview of GloLakes Product Descriptions

Filename	Type of volume	Satellite sources	Historical method	NRT method	The number of measured lakes	Time Coverage
Global_Lake_Absolute_Storage_LandsatPlusGREALM (1984-present).nc	Absolute	GSWD (Landsat) + G-REALM	Q1	Q2	129	1984-current
Global_Lake_Absolute_Storage_LandsatPlusICESat2 (1984-present).nc	Absolute	GSWD (Landsat) + ICESat-2	Q1	Q2	24,865	1984-current
Global_Lake_Absolute_Storage_LandsatPlusSentinel2 (1984-present).nc	Absolute	GSWD (Landsat) + BLUEDOT (Sentinel-2)	Q1	Q1	4,054	1984-current
Global_Lake_Relative_Storage_LandsatPlusGREALM (1993-present).nc	Relative	GSWD (Landsat) + G-REALM	Q3	Q4	227	1993-current
Global_Lake_Relative_Storage_LandsatPlusICESat2 (2018-present).nc	Relative	GSWD (Landsat) + ICESat-2	Q3	Q4	24,990	2018-current
Global_Lake_Relative_Storage_Sentinel2PlusICESat2 (2018-present).nc	Relative	BLUEDOT (Sentinel-2) + ICESat-2	Q3	Q3	2740	2018-current

R1C4) 3. Lack of discussion of revisit times

My apologies if I've missed it, but it seems to me this paper is entirely missing discussion about the frequency of observations, particularly for the NRT data. Thorough discussion and presentation of results about revisit times and how often you actually get an NRT observation of each lake is absolutely vital for a reader of this paper to determine the usability/relevance of the resulting dataset. As noted above, this lack of discussion of revisit times in the results/discussion section is especially important given the huge difference between ICESat-2 and Sentinel-2 revisits. For example, Figure 3 and the section about 'how many lakes can be monitored' using this approach is not useful without some discussion around frequency of observation.

We have highlighted the revisit times in Table 1 and discussed it in the Section 3.5 Current limitations and future opportunities. However, to further clarify the difference in 'revisit' time of different datasets used in this study and their remote sensing sources and to address the reviewer's concern, we modified the paragraph in L503-528 in the revised manuscript as follows:

"3.5 Current limitations and future opportunities

Remote sensing provides an opportunity to measure water in most lakes worldwide and provide NRT or up-to-date information, which is impossible with the current in situ network. Topex/Poseidon (1992-2002), Jason 1/2/3 (2002-present), and Sentinel-6 (a.k.a. Jason-CS; 2020-present) are all able to measure lake height every ten days, which should be adequate to monitor dynamics in many lakes. Sentinel-3A/B (2016-present) has a revisit time of 27 days, still providing valuable, quasi- monthly updates on changes on water level. The dynamic estimates of lake height from these radar altimeters were seamlessly processed and derived from G-REALM. However, radar altimeters cannot detect many smaller lakes in between the sparse ground tracks. The ICESat-2 laser altimeter covers many more lakes globally, benefiting from its dense reference tracks enhanced by the six laser beams onboard. The trade-off is its temporal resolution of ~91 days, but this is still sufficient to observe seasonal changes in many lakes worldwide, and more frequent water extent mapping can be used to interpolate between these observations. This study used the Landsat-derived GSWD to estimate lake surface water extents. The main limitation of using GSWD is that it only produced monthly observations from 1984-2020 and cannot provide NRT or up-to-date information, despite its use of Landsat observations with a temporal frequency of 16 days. The GSWD are expected to be updated annually by the Joint Research Centre of the European Commission (https://global-surface-water.appspot.com/download), at which point lake extent estimates can be extended beyond 2020. In this study, we harnessed the capabilities of BLUEDOT (Sentinel-2) to expand lake surface water extent estimates beyond 2020, benefitting from the 5-day revisit interval of Sentinel-2 within the BLUEDOT framework. However, lake area data from BLUEDOT are presently available only for a subset of several thousand lakes across the globe. This issue could be overcome using MODIS, VIIRS or Sentinel-2 data directly, for example. The daily or 8-day composite MODIS and VIIRS products have a better chance to provide valid observations, but the hundred-meter range resolution is often not sufficient to accurately detect lake area changes. 5-day, 10-m resolution water extent derived from Sentinel-2 should be a promising candidate for NRT global lake monitoring. The sheer volume of data presented us with a challenge for data storage and processing, but there is no fundamental limitation that would prevent a similar approach from measuring all 170,957 lakes measured by GSWD (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 could 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."

Specific Comments:

R1C5) Line 49-53: You might considering referencing/discussing GeoDAR here, a new global reservoir and dam database that is arguably better than the databases discussed here, in this paragraph: https://essd.copernicus.org/articles/14/1869/2022/

Thank you. We included this reference and added one sentence (L52-54) about it in the revised manuscript.

> "The more recent Georeferenced global Dams And Reservoirs (GeoDAR) dataset (Wang et al., 2022) not only provides the locations of 22,560 dams but also delineates the boundaries of 21,515 reservoirs around the world"

[1] Wang, J., Walter, B. A., Yao, F., Song, C., Ding, M., Maroof, A. S., Zhu, J., Fan, C., McAlister, J. M., Sikder, S., Sheng, Y., Allen, G. H., Crétaux, J.-F., and Wada, Y.: GeoDAR: georeferenced global dams and reservoirs dataset for bridging attributes and geolocations, Earth Syst. Sci. Data, 14, 1869–1899, https://doi.org/10.5194/essd-14-1869-2022, 2022.

R1C6) Line 74: There's a mistake here – the sentence just reads "Compared to…" but presumably there should be more text there

Apologies for this mistake. We deleted "Compared to" here in the revised manuscript.

R1C7) Line 81: The phrasing "whose number soars exponentially as smaller lake sizes are considered" doesn't make sense here, maybe replace with something like "which are exponentially more numerous than large lakes"

Thank you for your suggestion. We rephased it as "which are exponentially more numerous than large lakes" in L82-83 in the revised manuscript.

R1C8) Appreciate the adding of discussion around the different ICESat-2 products.

Thank you again for referring us to the ICESat-2 quick look data.

R1C9) Figure 3 (and paragraph above): When discussing what percent of lakes in a given basin that can be observed using these techniques, it is imperative to note the percent of what (i.e. what is the denominator. There are millions upon millions of lakes globally, so I assume the denominator here is the 170,000 lakes whose storage dynamics were tracked, but you should be explicit about this.

Yes, the denominator used to calculate the percentages in Fig3b-d is the total number (170,611) of lakes whose storage dynamics have been estimated in this study (Fig3a). We clarified this in the figure caption and corresponding sentences in the revised manuscript.

R1C10) Line 332: It’s meaningless to state that ‘Landsat and ICESat-2 together could measure lake water storage in nearly all rivers basins worldwide’. If ICESat-2 is only measuring a handful of lakes in a basin, is that really estimating its storage? Perhaps rephrase this sentence.

Apologies for the confusion. We rephased this sentence to clarify what we mean here:

> *“GSWD (Landsat) and ICESat-2 together could measure lake change in nearly all rivers basins (i.e., 234 out of 292) worldwide (Fig. 3b).”*

R1C11) Line 332-333: When you state that extents and levels are correlated for ~1/4 of all lakes, is that because ICESat-2 only observes 1/4 of all lakes, or are there only good correlations for 1/4?

Following the change in R1C10, we modified this sentence in L372-373 in the revised manuscript to clarify this:

> *“Extents derived from GSWD (Landsat) data and levels derived from ICESat-2 exhibit a substantial correlation in over 25% of lakes within 145 out of the total 224 river basins.”*

R1C12) Line 325-330: I am confused here by the three complementary approaches – I think this should read (1) geostatistical model + extent (Landsat + geostatistical model + Sentinel-2), (2) geostatistical model + height (Landsat + geostatistical model + ICESat-2) and (3) extent + height (Sentinel-2 + Landsat). As written, it is confusing because you are essentially equating Landsat with the geostatistical model. Throughout the paper (for example, in Figure 3), please fix this and make it clear that when you say ‘Landsat’ you mean the GSWO dataset + the geostatistical model.

Thank you. We changed this sentence following your suggestions:

> *“(1) geostatistical model + extent (GSWD (Landsat) + geostatistical model + BLUEDOT (Sentinel-2)), (2) geostatistical model + height (GSWD (Landsat) + geostatistical model + V-H relationship + ICESat-2) and (3) extent + height (BLUEDOT (Sentinel-2) + ICESat-2)”*

R1C13) Line 359: What is GL?

GL is gigalitre (10^6 m^3) and numerically equal to millions of cubic meters (MCM). We would argue GL and MCM are both commonly used units for surface water volume but generally S.I. units are preferred by scientific journals. To clarify, we added "(V in gigalitre (GL), millions of cubic meters (MCM) or 10^6 m^3)" in L233-234 in the revised manuscript.

R1C14) Figure 5. What is the unit on the y axis?

The unit used throughout this manuscript is GL. We added the unit (GL) in this figure caption in the revised manuscript.

R1C15) Figure 6. What is GL?

Please see our response to R1C13.

R1C16) Line 375: I'm not sure I agree that the relative agreement is more important than the absolute error. I can see what the authors mean here, but when discussing the comparison with the Tortini data, the authors should include the SMAPE/MAE to provide better validation context.

This discussion pertains to the outcomes presented in Figure 4, 5 and 6, where we provided both R and SMAPE validation results, but our primary emphasis is on R values due to their importance. In Figure 7 and 8, We compared our relative change estimates in lake volume with the Tortini data, making the R value the key metric for reflecting relative agreement.

R1C17) Figure 8: All lakes chosen to display in Figure 8 have very high R values, but there clearly are lakes with far worse agreement. This combined with the fact that the authors do not report any absolute error and only the R values (and then don't show any of the worse R values) makes me somewhat question the accuracy of the results.

The validation results mostly show strong R values (Fig. 7), but we have included three instances of worsened results in Figure 8 in the revised manuscript.

R1C18) Line 447: This is wrong to state that 'Landsat cannot provide NRT observations'. Landsat data IS available NRT (at 16-day revisit), but rather the JRC-GSWD product is not NRT.

We corrected this sentence in the revised manuscript:

> *"The main limitation of using GSWD is that it only produced monthly observations from 1984-2020 and cannot provide NRT or up-to-date information, despite its use of Landsat observations with a temporal frequency of 16 days."*

R1C19) Line 466: I'm not sure it's correct to state that the spatial resolution of SWOT is worse than Landsat and Sentinel-2. SWOT doesn't have a spatial resolution in the same way that Landsat and Sentinel-2 do (i.e. it requires spatial averaging to produce height observations). SWOT will be able to measure lake height and area ~6 million lakes globally, which is far more than the number of lakes observed here, so perhaps just remove this part about the spatial resolution of SWOT.

Thank you for your suggestion. To avoid the confusion, we removed this sentence in the revised manuscript.

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

I comment that the authors paid attention to my comments. In particular, they added additional details and validation results to improve the clarity and quality of this manuscript. I still have a few major concerns that some statements seem to be inaccurate or potentially biased and that some validation analysis is not convincing or completed.

We want to thank the reviewer again for their comments and suggestions. We tried our best and added further validation of satellite-derived lake extent and height estimates in the revised manuscript. We hope that these address the reiewer's concerns. Please see our detailed responses below.

R2C1) I have a major concern on the validation of estimated areas. Previously, I commented the uncertainty of areas due to using a global model and the global model may not be sufficient to track water area changes in each lake locally due to differences in water quality, surrounding land cover types, and atmospheric conditions (e.g., aerosol). Because of this challenge, there are multiple water indices that have been developed to track water area changes under different scenarios in recent literature. The authors' validation has little contribution in reducing my concern. First, Zhao and Gao (2018) used the same data source (Pekel et al). Comparing with Zhao and Gao is barely a comparison rather than a validation. Donchyts et al. also solely depend on one water index NDWI, while the limitations of NDWI have been acknowledged by many studies. I recall at least a couple of regional-to-global-scale studies on generating lake area time series using more sophisticated approaches (e.g., multiple water indices) rather than a global classification model. I expect the authors to be familiar with recent literature when producing a global dataset in this area. Second, validating against the two studies shows a remarkable difference. This also emphasizes my surmise of the uncertainty of estimated areas. Third, the authors did not response to my concern on the fidelity of water area changes for each lake individually. This is particularly important as a dataset paper, the users may more likely choose to use one lake for a case study than applying the dataset globally. What's the mean bias of each generated lake area time series? In sum, I strongly recommend the authors consider more quality-assured data sources and methods for the validation rather than seemingly flawed validation analysis.

We agree that there are limitations and challenges when using remote sensing to precisely estimate area change for each lake, especially in global studies. However, it is important to clarify that our study does not focus on enhancing the current classification algorithm for estimating lake extent. Instead, our primary objective is to ensure that our lake extent estimation attains a level of proficiency comparable to other published datasets (e.g., Zhao and Gao (2018) and Pekel et al. (2016)). Then, we can gain the confidence to convert extent to volume using a geo-statistical model. Second, we argue that the GSWD data used in this study is a reliable global surface water classification product as it used non-parametric classifiers (expert systems), rather than single or multiple indices, to account for uncertainties in the data. The thorough validation of this classification algorithm can be found in Pekel et al. (2016). Third, relatively larger bias metrics does not necessarily imply that the estimates contain significant errors, as different datasets may have different definitions of lake boundaries, e.g. one can include broader upstream channels and ephemeral connected water bodies as part of lake area. For example, lake boundaries from HydroLAKES and Donchyts et al. (2022) are different to some extent in many lakes but

conceptually and technically equally justifiable. Lastly, we do not use all lake area time series derived from GSWD. Instead, we only used GSWD-derived lake area time series, which are significantly correlated with other data sources, i.e., Bluedot (Sentinel-2) lake area time series, or GREALM/ICESat-2 lake height time series (L240-243 and Figure 3). We assumed that if two satellites source show consistency in lake changes, the derived lake area and height time series are more likely to be reliable.

Incorporating your suggestions, we expanded our analysis beyond the area time series comparison of 16,419 lakes against Zhao and Gao (2018) and Donchyts et al. (2022). To further validate our lake area estimates, we conducted a comparison with Sentinel-2 data, utilizing the capabilities of Google Earth Engine. The revised manuscript includes an additional paragraph, a figure, and a table to thoroughly discuss these new validation results:

> *"To further assess our lake extent estimates, we manually derived lake area based on bands 8 (near infrared), 4 (red) and 3 (green) from high-resolution Sentinel-2 imagery using Google Earth Engine and compared it against lake area derived in this study (Figure S2 and Table S1). We selected two Sentinel-2 images without cloud cover - one with a relatively small extent and another with a relatively large extent - for each of 20 lakes chosen to represent different continents and climates and with varying surface water color, lake shape and size and surrounding land cover. We delineated the lake area using Sentinel-2 surface reflectances in the three bands and visually validated it with a false-color image. The Sentinel-2 derived lake area was then compared to GSWD (Landsat) derived lake area. The results show that the average difference of lake area estimates derived between GSWD (Landsat) and Sentinel-2 for all 40 pairs is 2.62% (Table S1). The estimated lake area differences do not vary systematically with geographical location, surrounding land cover or lake shape and size. Most importantly, the area estimates for both small and large extents exhibited very strong consistency between the two satellite sources (Figure S2). This confirms GSWD (Landsat) can successfully capture changes in lake area. Comparatively large biases (from 3-10%) are observed exclusively in lakes with surrounding ephemeral disconnected surface water bodies or emerging islands within the lake during dry periods, such as Lake Pozuelos in Argentina, Lake Baia Grande in Brazil, and Lake Aksehir in Turkey (Figure S2 and Table S1)."*

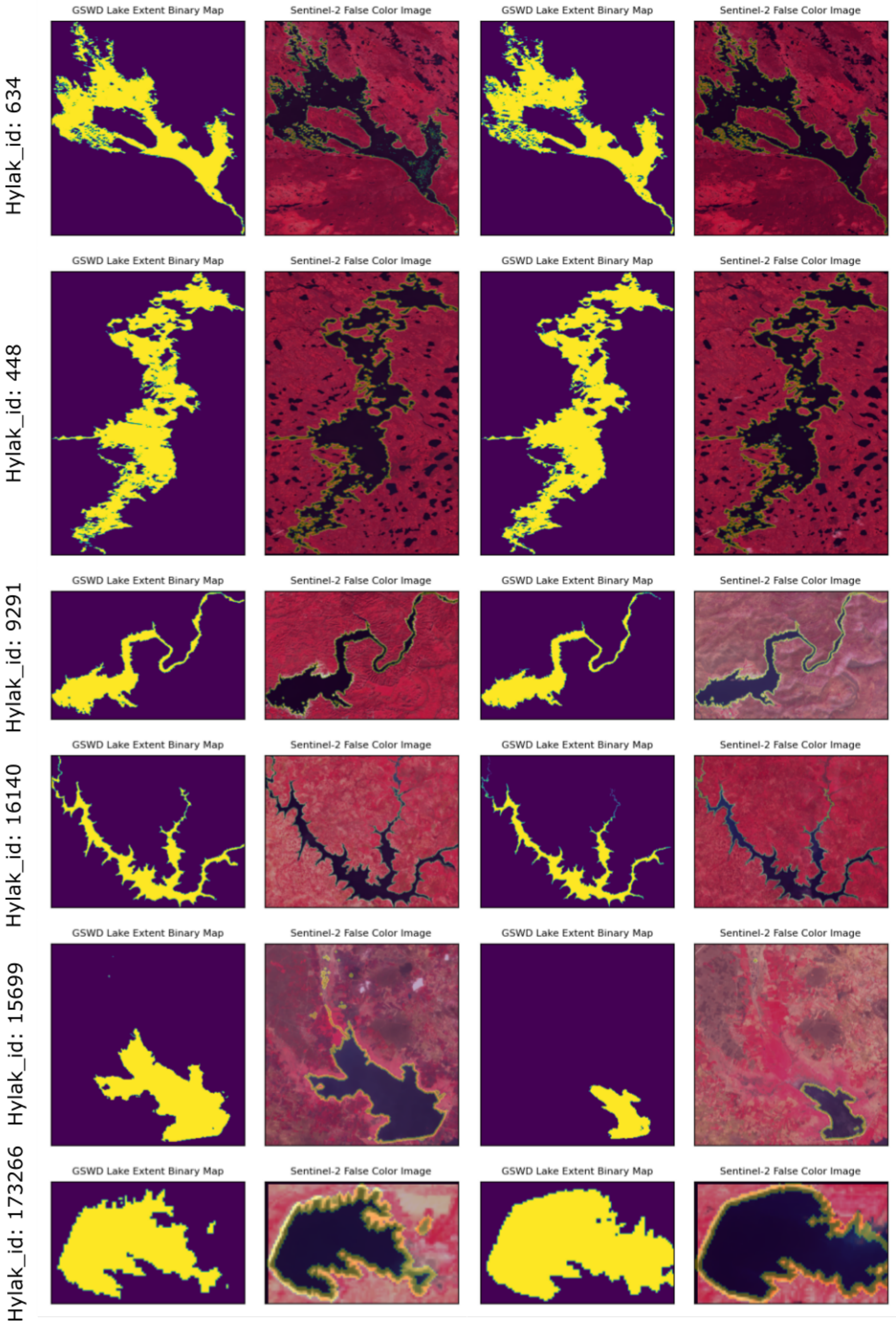

Figure S2. Comparisons between GSWD lake extent binary map (yellow: wet pixels; blue: dry pixels) and Sentinle-2 false color imagery (yellow line: derived lake extent boundaries) at different time for selected 20 lakes around the world.

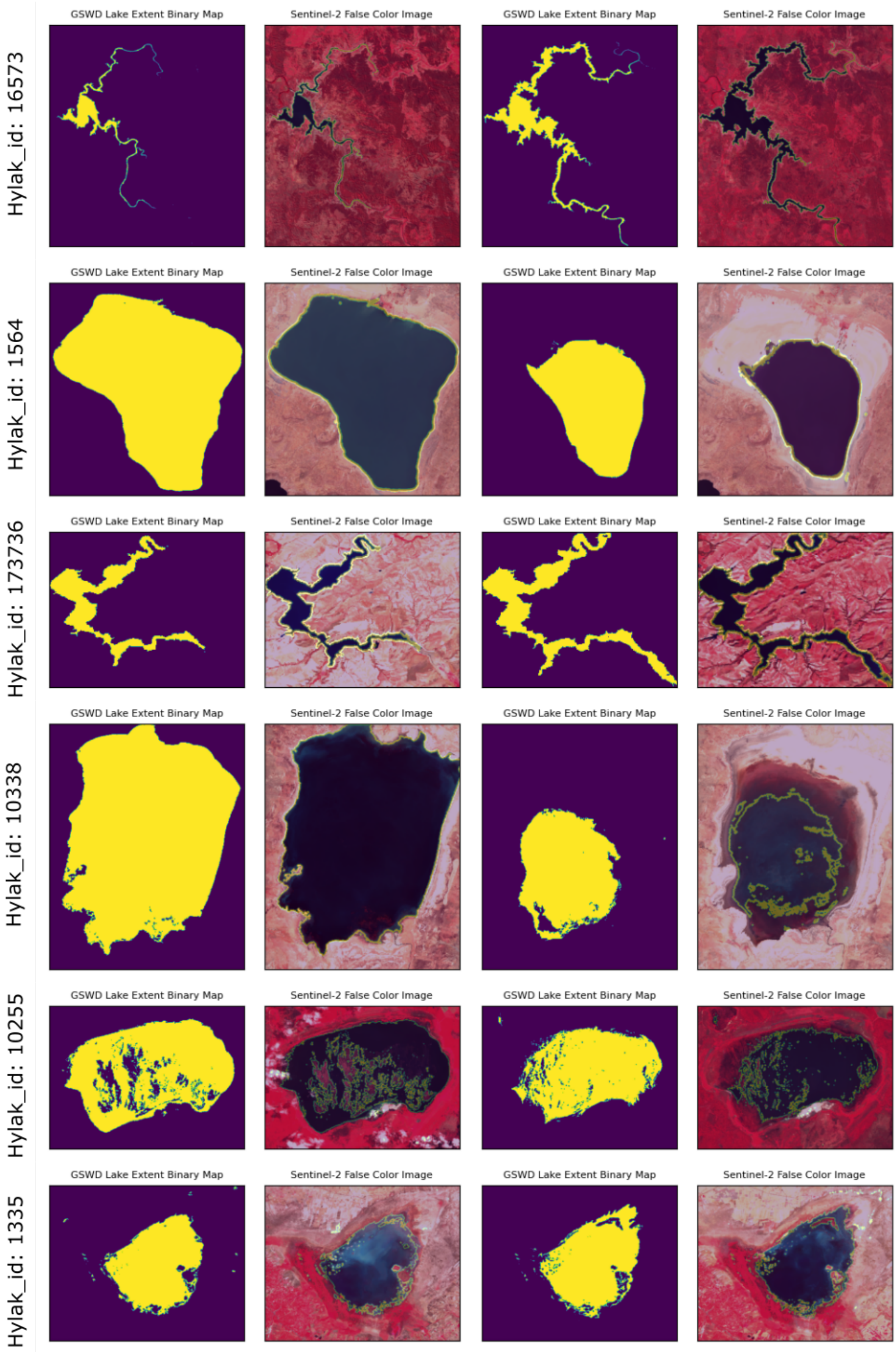

Figure S2 (continued). Comparisons between GSWD lake extent binary map (yellow: wet pixels; blue: dry pixels) and Sentinel-2 false color imagery (yellow line: derived lake extent boundaries) at different time for selected 20 lakes around the world.

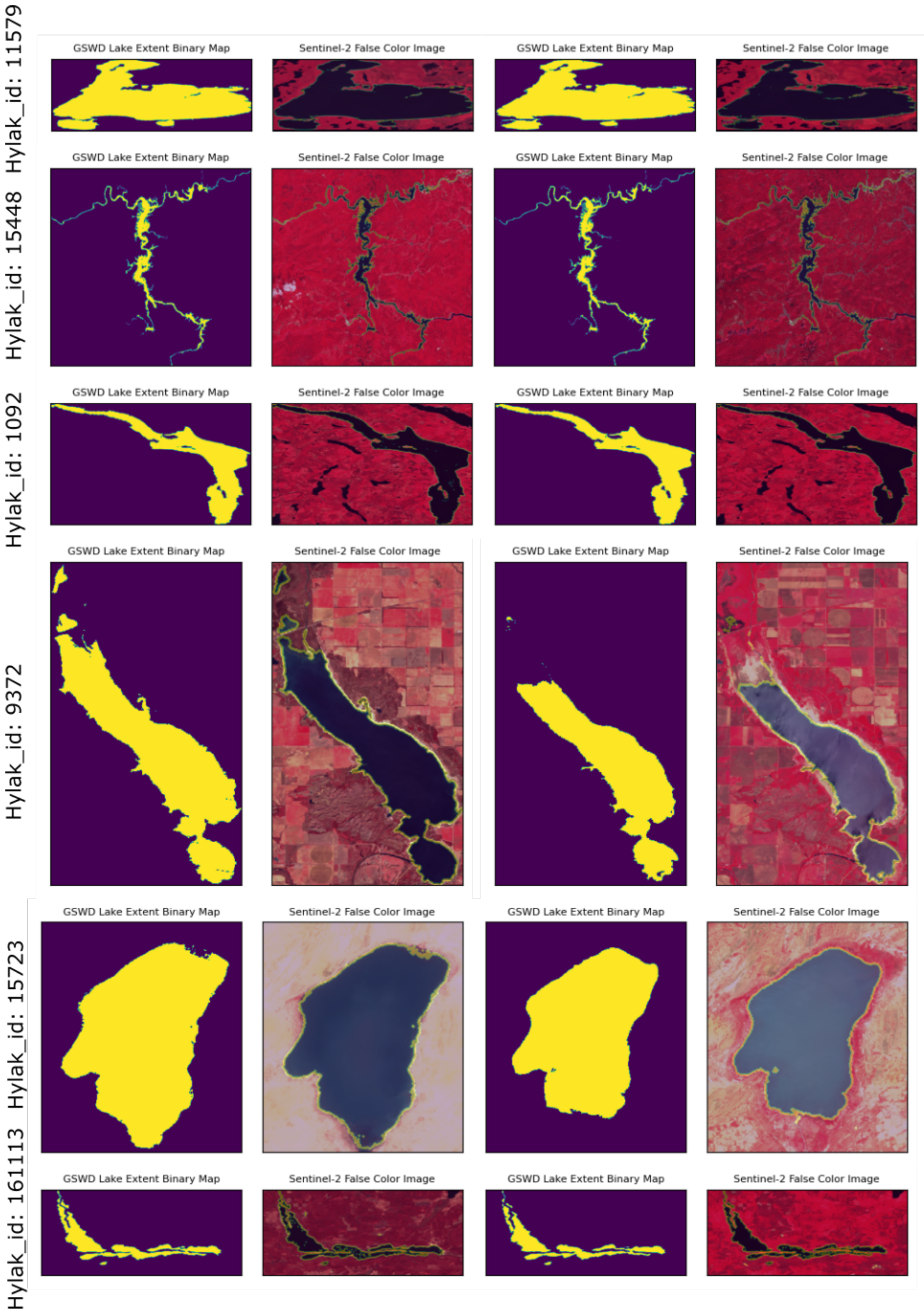

Figure S2 (continued). Comparisons between GSWD lake extent binary map (yellow: wet pixels; blue: dry pixels) and Sentinel-2 false color imagery (yellow line: derived lake extent boundaries) at different time for selected 20 lakes around the world.

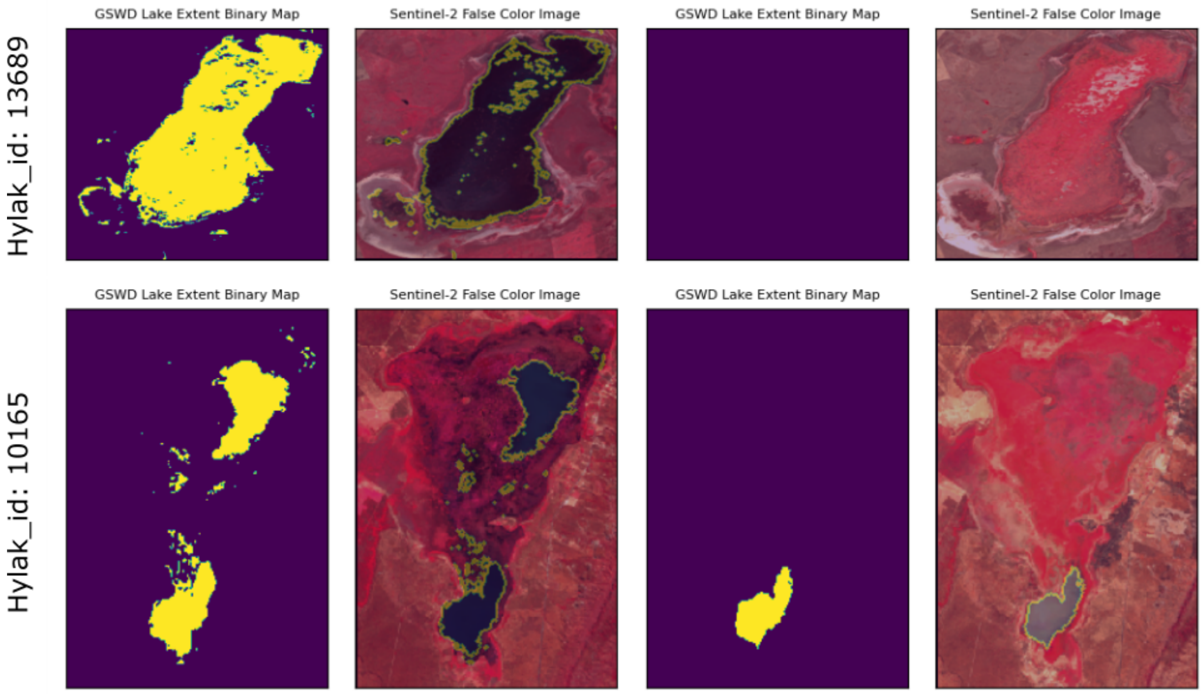

Figure S2 (continued). Comparisons between GSWD lake extent binary map (yellow: wet pixels; blue: dry pixels) and Sentinel-2 false color imagery (yellow line: derived lake extent boundaries) at different time for selected 20 lakes around the world.

Table S1 Comparison results of lake extent estimates derived between Landsat-GSWD and Sentinel-2 (statistics results from Figure S2)

ID	Lake Name	Country	Latitude	Longitude	Date	Landsat-GSWD lake extent (km^2)	Sentinel-2 lake extent (km^2)	Difference (%)
634	Petit Manicouagan	Canada	51.82	-67.80	2020-06	312	313	0.32
634	Petit Manicouagan	Canada	51.82	-67.80	2019-07	305	320	4.69
448		Canada	58.28	-96.99	2019-07	226	227	0.44
448		Canada	58.28	-96.99	2021-08	229	224	2.23
9291	Millerton Lake	United States of America	37.00	-119.70	2020-02	17.3	17.5	1.14
9291	Millerton Lake	United States of America	37.00	-119.70	2020-09	15.2	14.9	2.01
16140	Gove	Angola	-13.45	15.87	2021-02	101	102	0.98
16140	Gove	Angola	-13.45	15.87	2019-04	76	78	2.56
15699		India	17.23	79.52	2019-02	10.7	11.3	5.31
15699		India	17.23	79.52	2019-04	2.92	3	2.67
173266		Uzbekistan	38.87	66.42	2019-09	1.84	1.83	0.55
173266		Uzbekistan	38.87	66.42	2019-05	2.7	2.8	3.57
16573	Burrendong	Australia	-32.67	149.11	2019-04	16.8	17	1.18
16573	Burrendong	Australia	-32.67	149.11	2021-02	45.4	45.2	0.44
1564	Abayata	Ethiopia	7.61	38.61	2021-01	150	149	0.67
1564	Abayata	Ethiopia	7.61	38.61	2019-01	71.2	70.5	0.99
173736	Vadomojon	Spain	37.64	-4.23	2019-08	4.97	5.06	1.78
173736	Vadomojon	Spain	37.64	-4.23	2019-02	6.93	7.19	3.62
10338	Pozuelos	Argentina	-22.32	-65.99	2019-05	95	98.6	3.65
10338	Pozuelos	Argentina	-22.32	-65.99	2019-11	31.4	33.3	5.71
10255	Baia Grande	Brazil	-15.53	-60.19	2020-03	56.6	51.7	9.48
10255	Baia Grande	Brazil	-15.53	-60.19	2021-06	43.7	40.3	8.44
1335	Aksehir	Turkey	38.51	31.42	2019-07	91	85.5	6.43
1335	Aksehir	Turkey	38.51	31.42	2020-07	96	92	4.35
11579		Russia	67.76	124.22	2021-06	191.8	189	1.48
11579		Russia	67.76	124.22	2020-08	186	183	1.64
15448		China	26.71	117.12	2019-09	31.8	33.1	3.93
15448		China	26.71	117.12	2021-01	31.6	32	1.25
1092	Flasjon	Sweden	64.14	15.91	2020-05	254	252	0.79
1092	Flasjon	Sweden	64.14	15.91	2021-07	254	253	0.40
9372	Lake Altus	United States of America	34.89	-99.29	2019-03	22.9	22.1	3.62
9372	Lake Altus	United States of America	34.89	-99.29	2021-06	15.9	16	0.62
15723		Mali	15.78	-4.53	2019-03	36.4	36.1	0.83
15723		Mali	15.78	-4.53	2019-09	28.9	29.1	0.69
161113		Russia	56.00	28.71	2019-04	10.5	10.6	0.94
161113		Russia	56.00	28.71	2021-06	9.3	9.5	2.11
13689		Kazakhstan	51.37	61.86	2021-05	30.6	29.2	4.79
13689		Kazakhstan	51.37	61.86	2021-09	0	0	0.00

10165	Brazil	-10.23	-44.68	2021-07	3.4	3.2	6.25
10165	Brazil	-10.23	-44.68	2019-09	0.93	0.91	2.20

R2C2) I am still concerned about the accuracy of the produced datasets in cold regions given cold regions have a large share of Earth's lakes. The authors claimed that they could not find any in-situ data in Canada. I am not sure if the authors really tried the best to find the in-situ data in Canada. Here is in-situ data for Canadian lakes: https://wateroffice.ec.gc.ca/. Some countries in Northern Europe may have in-situ data as well. In many lakes, storage changes are dominated by level changes (Cooley et al. 2021 Nature). As the authors used different approaches to derive levels in order to estimate the storage time series, what's the mean level error of each generated time series? How does the level error vary among different approaches applied? I would like to remind the authors that in-situ level data has a much larger spatial coverage than 438 lakes used in the current validation, particularly for small lakes. How many natural lakes are included in the validation?.

Thank you for your suggestions and sharing the link to these in situ data. In the revised manuscript, we included validation results on lake area estimates for 10 lakes in cold regions, including two lakes in Canada, one lake in Sweden, and two lakes in Russia. The validation results show strong agreement in lake area estimates between our study and GEE Sentinel-2 manually derived data in cold regions. Please see the detailed response to R2C1. Secondly, while we have successfully estimated lake extents and heights for the majority of lakes in cold regions, our analysis does not encompass all lakes in these regions. Rather, we have specifically focused on providing data for lakes that exhibit consistent satellite-derived extents and heights over time, ensuring the reliability of our published estimates through significant correlation. This rigorous selection process has led us to present data for less than 10% of the lakes in cold regions (refer to Figure 3b-d), as we prioritize data accuracy and precision. Lastly, we agree that there are more in situ level data available from different countries, but obtaining data from most of these agencies poses a significant challenge due to the lack of easily accessible means for discovering and downloading the information. This has been a challenge to many global studies. For example, the global lake study of Cooley et al. (2021) only used in-situ level data from the USA for validating lake height estimates (Figure S3).

We tried our best to collect more in situ level data and used them to validate our lake height estimates. We were unable to locate any in situ lake level data from Donchyts et al. (2022) that could be utilized to validate our lake level estimates, but we did use their in situ lake extent data to validate our extent estimates for several hundred lakes (most of their data have in situ lake extent and volume data, but not necessarily level data). We invested time and effort in understanding the API for accessing in situ data from Environment Canada. Additionally, we submitted a request to the Australian Bureau of Meteorology to obtain their in situ data. Overall, in addition to the validation results in the USA from Cooley et al. (2021), we included lake height validation results for two more countries (one in a cold region) in the revised manuscript (Figure 9):

> *"To validate satellite-derived lake heights used to estimate NRT storages in this study, we obtained in situ lake level measurements from the Australian Bureau of Meteorology (http://www.bom.gov.au/water/index.shtml) and Environment Canada (https://wateroffice.ec.gc.ca/). In*

total, we identified 96 lakes in Australia (including 83 reservoirs) and 54 lakes in Canada (including 23 reservoirs) with in situ level data corresponding to our study. Following the same validation approach as Cooley et al. (2021), we compared temporal changes in lake heights between in situ data and altimetry data. The results show strong agreement between ICESat-2 and in situ measurements, with a mean absolute error (MAE) of 0.23 m and a mean Pearson correlation (R) of 0.99 in Australia and 0.19 m, and 0.97 in Canada (Figure 9).”

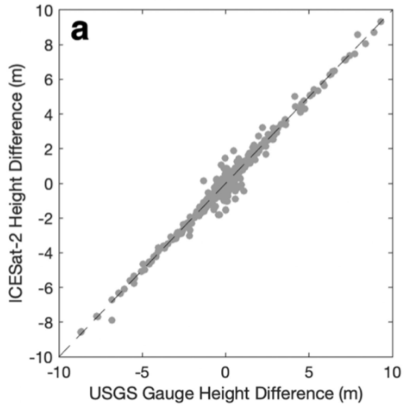

Figure S3 Evaluation of ICESat-2 derived lake heights against in situ gauge measurements from United States Geological Survey (Cooley et al., 2021).

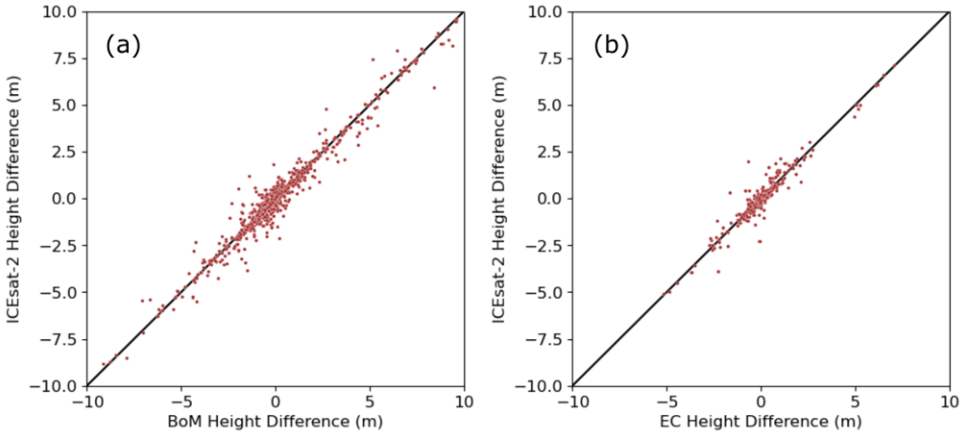

Figure 9 Evaluation of ICESat-2 derived lake heights against in situ gauge measurements from Australian Bureau of Meteorology (BoM) and Environment Canada (EC).