**Our Response to Anonymous Referee #2**

The paper evaluates water levels based on almost all historic and current altimetry missions and their standard retrackers over 12 lakes of different sizes. Here, especially, the results of the older missions are interesting. The main issue with this paper is the small sample size. 12 lakes are too small to provide any solid recommendations. Having a larger and more representative sample size would make this paper much more valuable. The Paper is well written and organized. The paper can be accepted if the review comment is addressed. Here, especially a discussion of low sample size is needed and the conclusions should be modified accordingly.
**RESPONSE:** We thank the reviewer for the positive comments on the value, writing and organization of our manuscript. We will address the sample size of case study lakes and other issues in the following item-by-item responses.

General comments: To make solid statements and recommendations about the rekrak-ing performance, 12 lakes are too small a sample size. This should at least be mentioned in the discussions section. However, the results in the paper support similar results in the literature.
**RESPONSE:** The selection of case study samples lake for our evaluation must meet the two requirements: 1) the sample lakes must be overpassed by all the satellite missions; and 2) Simultaneous in situ gauge data are available for the sample lakes. After our thorough search, we have identified 12 sample lakes that satisfied these two conditions. In most of the previous similar evaluations (in the introduction section), usually less than 5 sample lakes were used in their evaluations, and 12 sample lakes for our evaluation study still represents the largest sample size in the literature. More importantly, the twelve lakes in our study are located in different continents, latitudes and geographical environments. They include both natural lakes and reservoirs. They have different sizes, and winter ice conditions. We believe that this group of sample case study lakes are representative for the majority of inland lakes around the world and therefore we are confident that evaluation results for the historical and operational satellite altimetry missions through these sample lakes are valid. Nevertheless, we agree that it is even better if we have a much larger sample size that satisfy the above conditions. In response to the reviewer's comments, we have added a brief discussion on the lake sample size in Section 2.1 in the revised manuscript, and we hope that we can include more sample lakes in our future research when their in-situ gauge data become available.

The method section is vague and must be extended so it at least summarizes the methods from the mentioned reference studies. Hence, The MAD is estimated but what is the threshold to reject an observation.
**RESPONSE:** In response to the reviewer's comments, we have added more information about the robust MAD statistical method in the revised manuscript. We have also clarified the threshold value of the MAD statistic score used to exclude an observation from the subsequent calculation.

A main point in the paper is to construct consistent long-term time series and one of the issues is the intermission/retracking bias. In section 5.2 the gauge is used to estimate the biases. However, as discussed in the Discussion Section a gauge is not always available and therefore the bias should be estimated relative to a reference(s) mission. Why did the authors not test this approach?
**RESPONSE:** We appreciate the reviewer's comment and suggestion. This primary purpose of this study is to evaluate the historic and operational missions, to identify the reference mission, and then to develop a general strategy for estimating the biases. The estimation of the biases and the construction of long-term records need to carefully consider the temporal and spatial overlapping between these missions, particularly the overlapping with the consistent Topex/Poseidon-Jason series. This entails much more work on data processing, result analysis and discussion. Based on the current work, we plan to construct consistent long-term time series at regional or global scale in the future, relative to a reference mission as the reviewer suggested.

Why do the authors select evaluation targets in ice-covered regions when measurements during ice-covered periods are removed anyway?
**RESPONSE:** Lakes located in high latitude are more frequently overpassed by satellite missions, but the ice cover in the winter season may introduce significant errors to the elevation measurements of satellite altimetry missions. Since the official retrackers of all the satellite altimetry missions are not designed to handle the ice-cover on lakes, we identified and excluded the measurements obtained in the ice-covered condition in order to have a fair comparison between different altimetry missions. To clarify this confusion, we have added a brief discussion in Section 4.3 in the revised manuscript. We also noted that a non-official retracker developed by Shu et al. (2020) is able to accurately retrieve the water-equivalent lake level in the ice-covered condition for constructing a seasonally consistent lake water level time series from Sentinel-3 altimetry observations.
Why only use 1 track from each mission in the time series if more are available, this would improve the temporal resolution and the statistical foundation. Anyway, some of the missions are in different orbits anyway. For this reason, C2 could also have been included. Several authors have successfully applied C2 for lake level estimation.
**RESPONSE:** We appreciate the reviewer's comment.

Indeed, including more ground tracks will increase the temporal resolution of the time series. This is recommended when the analysis of the temporal variation of lake water levels is the primary goal in the practical applications.

Nevertheless, the primary focus of this study is to compare and evaluate the performances of multiple satellite radar altimetry missions. Our strategy is to minimize the influences of distant ground tracks from the gauge stations and ensure the objective comparison and evaluation of these missions.

For a large lake (e.g. the Great Lakes), strong wind, big wave, diurnal tide, geoid undulation, and other factors may significantly influence lake water level at different locations in the lake. The in-situ water level measurements from a gauge station may not reflect the actual water level of those ground tracks far away from the gauge station. Thus, the overall RMSE of the altimetry-derived estimates will increase when altimetry observations from distant ground tracks are included for evaluation (Birkett, 1995). To minimize the possible influence of wind, waves, tide and other environmental factors for an objective comparison between different satellite missions, we thus select the ground track nearest to the gauge station and exclude distant ground tracks in the performance evaluation. As sufficient footprints along the nearest ground track are available, we are able to derive statistically reliable RMSE and r values for objective comparisons between different missions.

Currently, CryoSat-2 uses a geodetic orbit (long-term repeat orbit). It is difficult to form a time series of co-located water level estimates for the evaluation. Although a time series of water level estimates from CryoSat-2 observations can be derived for a large lake by including many different ground tracks separated by a large distance, this will inevitably introduce the evaluation uncertainties due to the wind, wave, tide, and other environmental factors as explained above. This is why we did not include the Cryosat-2 data in the evaluation. For the same reason, we did not include the data collected by other satellite missions during their geodetic phases or drifting phases. In this revision, we have added a brief explanation why we only include the nearest ground track in our evaluation, and included the following reference:

Birkett, C.M. (1995). The contribution of TOPEX/POSEIDON to the global monitoring of climatically sensitive lakes. Journal of Geophysical Research, 100, 25,179-"125,204"

Specific Comments

L296: Shu et al, 2020 is not the reference of the standard S3A retrackers.

**RESPONSE:** In response to the reviewer's suggestion, the citation of (Shu et al, 2020) has been removed and the correct citation has been added.

L306: why only use such a small time period of S3 and Jason-3 in the evaluation?

**RESPONSE:** This research involves a large volume of data processing for all eleven missions over the twelve lakes. We started the data processing with Sentinel-3, then Jason-3. At the time we wrote this manuscript, the available data for Sentinel-3 and Jason-3 was from 2016 to 2018. Since the time period is longer than a full year (winter and summer), which cover a full hydrologic cycle of lakes. So, we believe that it is sufficient to support the performance evaluation of these two missions, although more data become available now.

L331: add a reference to EGM2008

**RESPONSE:** Added as the reviewer suggested.

L361: which criterion is used to remove outliers

**RESPONSE:** In the revised manuscript, we have added that the measurements with MAD statistic score larger than and equal to 3 are excluded.

L364: "through" -> over

**RESPONSE:** Corrected.

L393: The r indicates -> the Pearson correlation r ...

**RESPONSE:** Changed according to the reviewer's suggestion.

L420: When you calculate the data loss rate is that based on the "valid" measurements or all measurements

**RESPONSE:** It is based on the valid measurements after filtering the spurious measurements.
In this study, the "data loss rate" refers to the data loss rate of lake level estimates, instead of data loss rate of elevation measurements.

We have added two sentences in the result section to clarify this confusion and modified the abstract and conclusion accordingly.

L440: This only makes sense to state if the gauge and altimetry has the same vertical reference

**RESPONSE:** We agree. We modified the sentence to clarify this issue.

L448: is the bias calculated w.r.t the gauge? then add this
**RESPONSE:** Yes. We added this information in the revised manuscript.

L495-503: Put all the numbers in a table
**RESPONSE:** These numbers are summarized in Table 9. The numbers are cited here for the description purpose.

L510: Such conclusions are difficult to state based on just a few lakes
**RESPONSE:**
As responded above, the selection of case study samples lake is limited by two requirements: 1) the sample lakes must be overpassed by all the satellite missions; and 2) Simultaneous in situ gauge data are available for the sample lakes. In most of the previous similar evaluations, usually less than 5 sample lakes were used in their evaluations. The 12 sample lakes in this study still represents the largest sample size in the literature. More importantly, the twelve lakes in our study are located in different continents, latitudes and geographical environments. They include both natural lakes and reservoirs. They have different sizes, and winter ice conditions. We believe that this group of sample case study lakes can represent the majority of inland lakes around the world and therefore we are confident that evaluation results for the historical and operational satellite altimetry missions through these sample lakes are valid.

Nevertheless, we agree that it is even better if we have a much larger sample size that satisfy the above conditions. In response to the reviewer's comments, we have added a brief discussion on the lake sample size in Section 2.1 in the revised manuscript, and we hope that we can include more sample lakes in our future research when their in-situ gauge data become available.

L582: How would you determine which mission provides the best measurement?
**RESPONSE:** In the revised manuscript, we added that the water level estimates from the satellite mission with higher r and lower RMSE should be used.