Response to Editor

I am pleased to see that the authors have been able to incorporate most of the reviewers' recommendations. In addition to those comments, I would like to raise a major concern regarding the error assessment. Specifically, assuming a fixed error of 0.77 km^2 for all lakes—which appears to be half the absolute difference between two mapping approaches—seems somewhat arbitrary. For instance, for half of all lakes (<0.27 km^2), this would imply an error estimate nearly three times larger than their average area.

Ideally, error estimates should be specific to each lake and relative to its 'true' area. Prior research has shown that the relative error in lake area decreases with lake size (Wang et al., 2020; https://doi.org/10.5194/essd-12-2169-2020); larger lakes tend to have smaller relative errors. Implementing this would require a reference or 'master' dataset against which to compare the current dataset. Do the authors see any possibility to derive better estimates of individual lake errors?

Thank you for your feedback regarding the error estimation of classified lake areas. We have revised both the error analysis and the manuscript to address the concern in adopting a fixed absolute error, particularly for smaller lakes.

To better capture the variability in uncertainty, we now report the error estimate primarily in terms of relative error (as percentages), rather than as a fixed absolute value. This provides a more meaningful representation of uncertainty across the wide range of lake sizes in our dataset. Specifically, we now present both the median and average area differences between the SAR and multi-spectral classifications, calculated per lake where both methods successfully classified the same lake in the same year. This revision reflects a more statistically robust approach and allows the error to scale appropriately with lake size.

Additionally, we calculate error at both the individual lake level and at the aggregate (total lake area) level. This dual-level reporting is detailed in Section 5.3 (Lake size error estimation), and we have added Table 4 to summarise the results. Contrary to findings reported by Wang et al. (2020), who observed decreasing relative error with increasing lake size, our results show that relative error remains relatively consistent across different size classes in our dataset, ranging between 6–11%.

Regarding the use of SAR vs. multi-spectral classifications for error estimation, we agree that a reference or "master" dataset would be ideal. However, such a dataset does not exist for most ice-marginal lakes in Greenland—many of which have not been previously mapped or studied. We adopt a widely used approach that estimates uncertainty by comparing multiple, independently derived classification methods. This strategy has been

applied in numerous glacial lake studies and offers a reasonable proxy for error in lake extent (e.g., Leeson et al., 2017; Williamson et al., 2018; Moussavi et al., 2020; Lesi et al., 2022; Tom et al., 2022). In our case, the SAR and multi-spectral classifications provide two independent estimates of lake extent, which form the basis for error calculation.

This approach yields a more nuanced and transparent assessment of classification uncertainty, and better reflects both the limitations and strengths of our dataset. All relevant changes have been incorporated into the manuscript and are clearly documented in the revised version.

In this context, it would also be helpful to understand which of the three methods performed best in (1) detecting lakes and (2) estimating their area. A summary or guideline showing how the different methods compare, and how users might leverage them depending on their goal (e.g. accurate area of a single lake vs. total lake area in a region), would be highly valuable.

Both the SAR backscatter and multi-spectral classification approaches (from Sentinel-1 and Sentinel-2 imagery, respectively) are direct water-detection methods, and therefore provided consistent, spatially explicit measurements of water extent. As such, lakes identified using these approaches are suitable for both individual lake area analysis and regional total lake area assessments.

By contrast, the DEM-based topographic sink approach is an indirect method that identifies depressions likely to retain water. As this does not confirm the presence of surface water, each classification is manually validated. Consequently, this approach is not appropriate for calculating absolute lake area. However, it is highly valuable for identifying potential or theoretical lake extents, tracking the presence or absence of lakes over time, and supplementing analysis in regions where SAR and optical classifications are limited; such as northern Greenland, where persistent snow or ice cover and data gaps reduce the effectiveness of direct classification methods.

We have provided a guideline for the strengths of the three methods used in this study, and how each should be used in subsequent studies and analysis. In effect, this is what has been summarised above. This information is included in each of the methodology subsections associated with each method (Sections 3.1.1, 3.1.2, and 3.1.3).

More specific comments

Abstract: it is important to note from the very beginning what you consider as an ice-marginal lake (fully in contact with ice) in your study.

Noted. This has now been changed, with the beginning of the abstract firstly defining what an ice-marginal lake is, followed by their importance (and the motivation for producing this dataset):

"Ice-marginal lakes form at the edge of the Greenland Ice Sheet and its surrounding peripheral glaciers and ice caps (PGIC), where outflowing glacial meltwater is trapped by a moraine, or by the ice itself, and create a reservoir that is in contact with the adjacent ice. While glacial meltwater is typically assumed to flow directly into the ocean..." (Line 1-2).

L9: This refers only to lakes that are present throughout the study period, i.e. in all years? What about newly emerging and completely vanishing lakes?

The reported statistic refers to a surface extent change for at least the two end members of the study period (2016 and 2023) so it does not include emerging, detaching or vanishing lakes. However, newly emerging and vanishing lakes are captured in the statistics from the preceding sentence on lake abundance.

L10: These average sizes are difficult to judge, if there is no other estimate such as the mean or minimum lake size. In general, I would suggest to avoid using the mean (see a similar comment later) and rather suggest using the median as a diagnostic for size distributions that can be prone to outliers (in your case very large lakes).

Median statistics are now provided in this passage of the abstract:

"The northeast region contained the largest lakes, with a median size of 0.40 km^2 at the ice sheet margin and 0.24 km^2 at PGIC margins." (Line 10)

Median statistics are also provided in the corresponding results section (Section 4.2: Lake surface extent). And statistics presented in Figure 3 refer to change in median lake area across the inventory years.

L25: Please mention here what an ice-marginal lake is according to your definition (I notice that it is mentioned later in the manuscript, but it would be clear about this definition very early in the manuscript)

Done. The definition of an ice-marginal lake is elaborated on, and introduced earlier in this section.

"...along the ice edge of the Greenland Ice Sheet and in front of, and beside, surrounding PGICs. The delayed release of..." >> "...along the ice edge of the Greenland Ice Sheet and in front of, and beside, surrounding PGICs. An ice-marginal lake is a reservoir of meltwater that is dammed by a moraine or by the ice itself, therefore the trapped meltwater is in contact with the ice margin. The delayed release of..." (Line 24-25).

L32: There is probably little evidence for (recent) megafloods in Greenland, so I suggest to remove this hint.

This passage has now been removed.

L72: This buffer is a bit confusing: why is there a 1-km buffer, if lakes need to be in direct contact to glaciers according to your definition above?

One of the limitations of our dataset production is the lack of a dynamic, time-resolved ice margin dataset. We rely on the GIMP ice margin, which was produced in 2019 and does not reflect changes over time. If we applied this ice margin without a buffer, we would effectively be assuming a static ice margin across all inventory years, which is not realistic. The 1-km buffer is therefore used as an initial spatial filter to account for glacier retreat and to remove lakes that are very unlikely to have been in contact with the ice margin at any point. After this automated step, we manually inspect and remove lakes that are clearly detached from the ice margin. This is now detailed in Line 162-164, Section 3.1.4 (Inventory compilation).

L89: Please clarify in the manuscript that several lake polygons within the same year can share the same ID. In addition, specify the criteria used for assigning new IDs: Is there a minimum distance threshold between polygons that triggers the assignment of a new ID, or is a new ID assigned whenever two polygons do not spatially overlap?

Multiple lake polygons within the same year can share the same ID. This occurs when polygons have been connected at some point in time in the inventory series, even if they appear disconnected in a specific inventory year. Assigning the same ID in such cases preserves temporal continuity and reflects the dynamic nature of lake evolution, including drainage and partial refilling events. ID assignment is performed manually based on visual inspection across years; therefore, there is no fixed minimum distance threshold or

automated rule. Automated approaches were tested but did not reliably capture the complexity of lake connectivity over time. We have clarified this in the manuscript (Line 174, Section 3.1.4: Inventory compilation).

L93: In checking the GIMP GeoTiffs, I wondered how this dataset allows you distinguishing whether a lake is touching an ice cap, ice sheet or peripheral glacier?

The GIMP GeoTiffs are merged and transformed into a vectorised (.shp) format, from which we can distinguish lakes that are in contact with the GrIS and/or a PGIC. This information has now been added to the manuscript, along with the acronym definitions for MEaSUREs and GIMP:

"This margin information originates from the MEaSUREs (NASA Making Earth System Data Records for Use in Research Environments) GIMP (Greenland Ice Mapping Project) 15 m ice mask, where the provided GeoTiff files were merged and transformed into a vector format (Howat et al., 2014; Howat, 2017)."

L106: This point shapefile should ideally indicate which of the lakes were not detected by the algorithm.

This is a good idea and has now been added to the dataset point geopackage (.gpkg) file. The new version is now released on the GEUS Dataverse (https://doi.org/10.22008/FK2/MBKW9N), with files denoting a new version (fv3).

3.1.1 – 3.1.3: We understand that your multi-temporal analysis is based on the algorithm proposed by How et al. (2021), and that you refer extensively to the methods outlined in that study. However, for a data publication in ESSD, traceability and reproducibility are essential. We therefore kindly ask you to expand the methodological sections in your manuscript to make the workflow fully understandable without requiring the reader to consult external sources.

We have added methodology information where possible to the subsections of Section 3.1 (Lake classification) in order to improve the traceability and reproducibility of the classification methods. In addition, we have added a subsection detailing how the inventory was compiled (Section 3.1.4: Inventory compilation), which includes filtering approaches and the subsequent manual curation. Metadata generation is referred to here, but the majority of the metadata production information remains in Section 2.3 (Data format and structure).

For clarification, the SAR backscatter classification approach is simplified because of its migration to Google Earth Engine and therefore does not require as much detail (compared to How et al., 2021). The multi-spectral indices classification approach is summarised in Table 3 (Summary of multi-spectral indices for ice-marginal lake classifications from Sentinel-2 Level 1C scenes), including thresholds and indices targets. As a result of these aspects, we felt that the subsections regarding the SAR backscatter and multi-spectral indices classification approaches were adequately refined whilst also retaining all necessary information for the reader to understand the methodology.

L163: Where can we see that the correction factor appears to agree well with the limited datasets available?

This passage is eluding to results presented in the manuscript, therefore we have moved it to the appropriate place in Section 5.4 (Lake surface temperature error estimation) at Lines 307-210.

L190: Romer Soe needs a coordinate.

Coordinates (in degrees, minutes and seconds; DMS) have now been added:

"The largest lake in the inventory is Romer Sø, located in northeast Greenland..." >> "The largest lake in the inventory is Romer Sø (80°59'54"N, 19°09'21"W), located in northeast Greenland..." (Line 190).

L190-195: Most studies focusing on lake area change report the total (summed) lake area within a given region, including the errors in total lake area, and we would appreciate if you could provide similar summary statistics. Given the potential influence of outliers (very large lakes such as Romer Sø), we suggest reporting the median lake area as a more robust diagnostic for comparing different study regions.

Total (summed) lake area and median lake area now included in Figure 3, where the total ice sheet lake area and total PGIC lake area are provided alongside lake abundance and median lake area. In addition, the associated text has been updated in Section 4.2 (Lake surface extent) describing changes in total lake area and median lake area across regions and through time (Line 224-237).

Additionally, we would like more clarity on how the different methods contributed to the estimated lake surface area. If we understand correctly, you obtain between one and three separate estimates of lake area per lake. How are these individual estimates combined to form the reported mean? For instance, are the estimates averaged per lake, or are the lake areas first dissolved across methods before statistics are derived?

As you describe, lake areas are first dissolved across methods and then the statistics are derived for comparing changes in lake area through the inventory years. This is now outlined in Section 4.2:

"The inventory series also holds information on the change in lake area over time, by comparing corresponding lake extents from one of the direct classification methods (i.e. from SAR and/or multi-spectral imagery) (Figure 3). Lake areas are first dissolved across the SAR and multi-spectral classifications, and then statistics are derived by comparing lake area through the inventory years. Change in average lake area over..." (Line 217-220)

L253: Could you please clarify in more detail how you ensure that the algorithm performs robustly across the entire time series, given that the static DEM—which represents only a single point in time—appears to contribute significantly to the derived lake areas? Specifically, how do you account for potential changes in topography over time, e.g. in situations when glacial retreat creates a larger basin, that may affect the accuracy of lake area delineation when relying on a fixed elevation reference?

We outline that the ArcticDEM does not directly detect water (Section 3.1.1, Line 147-149) and therefore lakes classified using the DEM sink classification approach are not used in the lake area analysis presented in this manuscript (as stated in Section 4.2, Line 226-232). Manual verification of each detected lake is performed for each inventory year in order to remove any topographic depressions that are not filled and to remove lakes that are detached from the ice margin (as stated in Section 3.1.4, Line 167-176).

L254-256: Please revise this sentence for clarity.

We have changed this sentence accordingly to better convey this:

"This is likely because the classification methods have been extensively applied and developed in the SW region compared to others..." >> "This is likely because the classification methods have been extensively applied and developed in the SW region, making them particularly well-suited for use there compared to other regions" (Line 270-271)

5.2./ 5.3 We would appreciate it if you could focus this paragraph more explicitly on the error estimates presented in the current study, rather discussing the method/ results from the earlier study. As stated above, we encourage you to provide error estimates at the level of individual lake areas, rather than reporting only a single aggregated or gross estimate across all lakes.

For Section 5.2, we have removed information regarding the error estimation from How et al. (2021).

For Section 5.3, the error estimation for lake area has been updated to better quantify error based on total lake area (rather than an average individual lake area error). Our analysis shows that the total absolute error (i.e. the total difference between SAR lake area and multi-spectral lake area classifications) is 774.94 km^2 , equating to a central percentage error of ± 5%. We have included a summary of these error statistics (Table 4) showing this reported total lake area error, along with lake area error based on lake size groupings. We included this to show that, whilst the absolute error propagates with the size of a lake, this does not affect the relative (%) error. The text in this section has been updated accordingly (Lines 289-301).

Additionally, we are unclear about your error estimate for lake abundance. If we understand correctly, your results suggest a negative bias—that is, an underestimation of lake numbers—yet your reported error range includes both negative and positive values. Could you please clarify the rationale behind this, and explain whether this range represents uncertainty around a central estimate, or some other form of error characterization?

We form the lake abundance error estimation around a central estimate. We have clarified this in the text and included a passage to account for why the dataset under-estimates ice-marginal lakes in Greenland:

"Across all inventory years, 4543 ice-marginal lakes were manually identified in total, of which 2915 (64%) are captured by the automated classification approach. This forms a central error estimation of ± 809 (36%), and demonstrates that the automated classifications in the ice-marginal lake inventory series underestimate the number of ice-marginal lakes across Greenland. However, manually classified lakes include those under the size threshold (i.e. <0.05 km^2) adopted in the automated classification approach. The under-estimation of ice-marginal lakes within the inventory series therefore, in part, reflects smaller lakes that are removed from the dataset automatically due to the minimum area filtering. To summarise, lake abundance in the ice-marginal lake inventory series

should be adopted as a conservative estimate as it does not account for lakes with a surface extent below 0.05 km^2." (Line 280-285)