Review of "Sentinel-1 Detection of Ice Slabs on the Greenland Ice Sheet"
Culberg et al.

This manuscript describes a new algorithm which uses Sentinel-1 backscatter observations to detect ice slabs across the Greenland Ice Sheet. The work is new and interesting and will help fill important scientific gaps by providing a method to detect ice slabs at higher spatial and temporal resolution. Overall, the paper is clear and well written and has high-quality figures. Once my concerns are addressed, I believe this paper will be an excellent contribution.
Thank you! We appreciate your thoughtful and constructive comments that will help us improve this manuscript.

Major comments
[R1] *I understand that the authors take into account different incidence angles for the Sentinel-1 data by applying a linear fit to incidence angle and backscatter; however, I feel like the impact of incidence angle on this work needs to be more fully understood before this algorithm can be applied. How many different incidence angles are available for each pixel? If one pixel has substantially more incidence angles available, how does this impact the cross-pol ratio and therefore the delineation of ice slabs? It would be interesting to investigate how the defined ice slab boundaries change if only some of the available incidence angles in a given region were used. This would shed some insight into how sensitive this algorithm is to various incidence angles.*

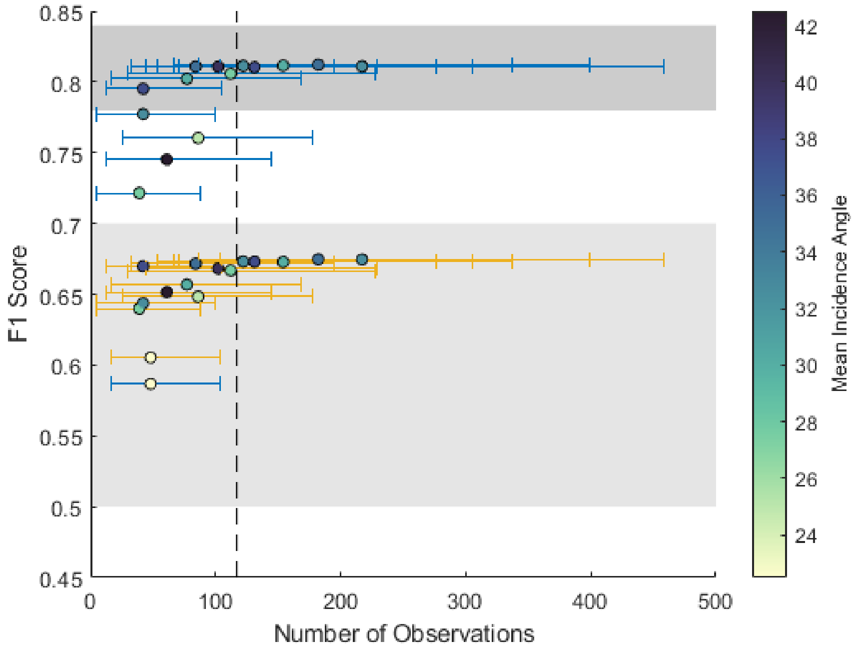
Thanks for this interesting suggestion! We have conducted this sensitivity analysis and describe the results below.

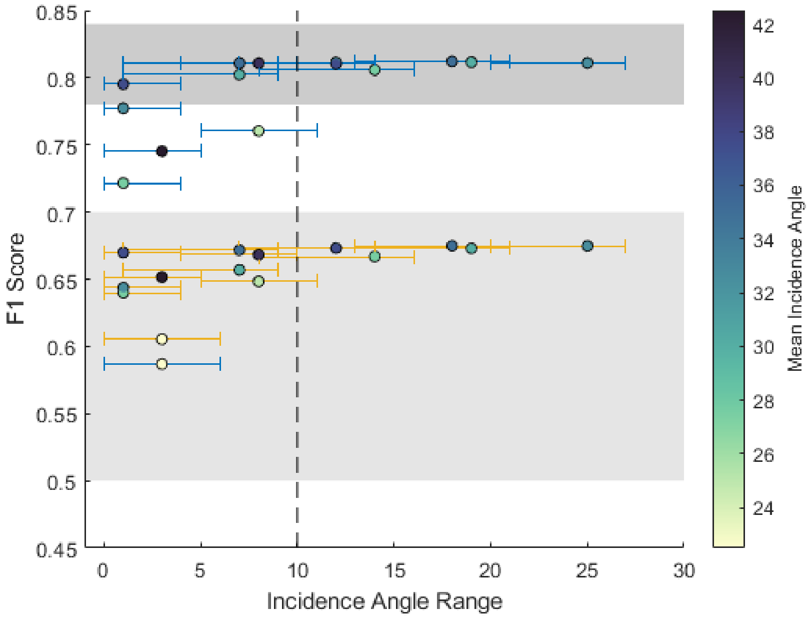
In EW mode, Sentinel-1 incidence angles from 18.9-47 degrees. We generated 15 σ_{HV} and σ_{HH} backscatter mosaics of the Greenland Ice Sheet using subsets of these incidence angles. The full list of mosaics is given in the table below. This has the two-fold effect of limiting the angular diversity per pixel, as well as the total number of observations per pixel.

TABLE I

Mosaic #	Angles	Mosaic #	Angles	Mosaic #	Angles
1	18.9-25	6	25-30	11	30-40
2	18.9-30	7	25-35	12	30-37
3	18.9-35	8	25-40	13	35-40
4	18.9-40	9	25-47	14	35-47
5	18.9-47	10	30-35	15	40-47

For each of these fifteen mosaics, we optimized the ice slab detection thresholds ($\alpha, \beta,$ and ϕ) using the full 2017 OIB dataset and calculated the F1 score quantifying the agreement between the OIB observations and the S-1 ice slab detections. We only considered pixels with observations from more than one incidence angle. The figures below show how the agreement between the OIB observations and S-1 detections changes as a function of the number of observations per pixel and the angular diversity of those observations over our region of interest. The median number of observations or angular range (difference between true minimum and maximum incidence angles in each pixel) for each mosaic are show in dots. The dots are colored by the mean incidence angle in the range. (For example, mosaics 6 and 10 both have an incidence angle range of no more than 5°, but mean incidence angles of 27.5° and 32.5° respectively). Error bars indicate the 5th and 95th percentiles of those values. Blue bars show how the F1 score for the upper elevation limit of the ice slabs changes, while yellow bars show the F1 score for the lower limit of the ice slabs. Grey patches show the range of F1 values reported from the 10-fold cross-validation scheme, and therefore quantify the variability in agreement between OIB and S-1 detections that we observe when optimizing α, β, and ϕ using subsets of the OIB data from different regions of the ice sheet.

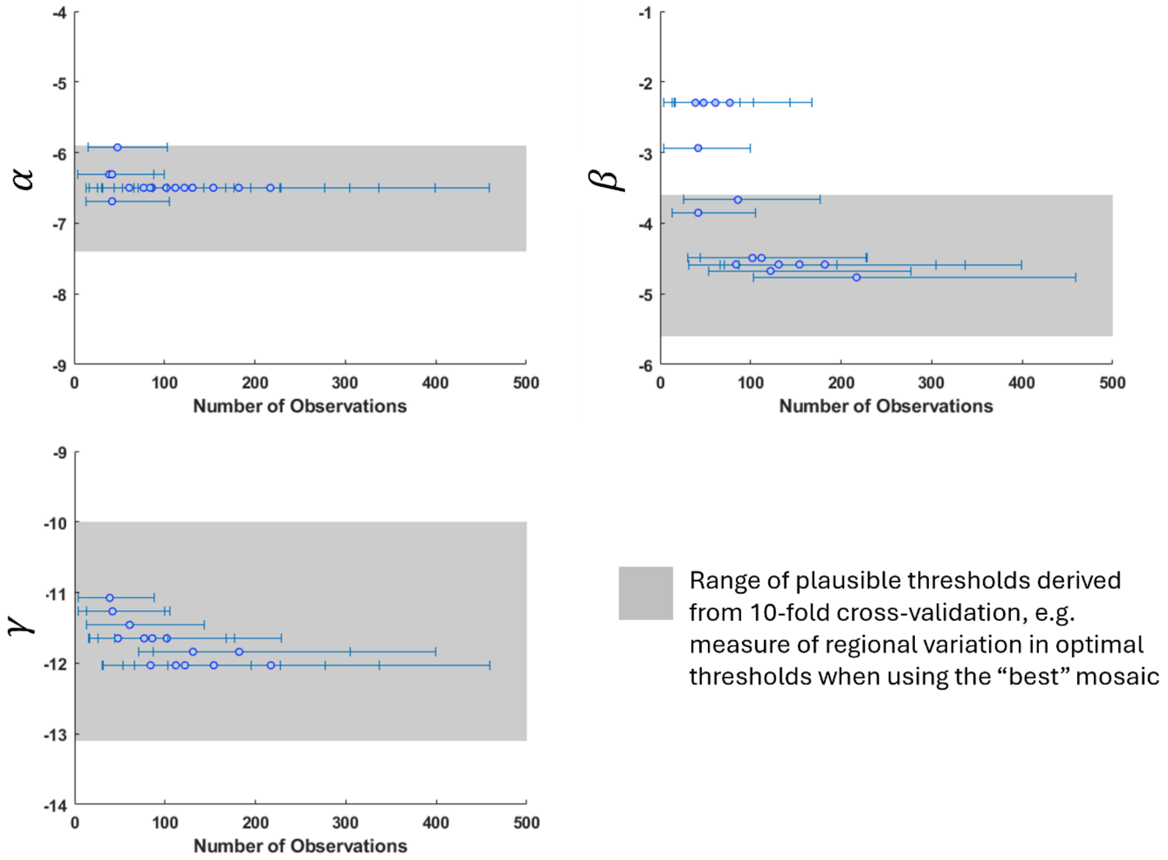




From these results, we see that the agreement between the OIB and S-1 observations converges once we have a median of ~117 observations per pixel and a median angular diversity of ~10 degrees per pixel. We also find that past these thresholds, any uncertainty in the final ice slab extent due to variations in the number of observations or angular diversity of observations in each pixel is well within the inherent uncertainty from regional variations in ice slab structure and backscatter response as quantified by the range of F1 scores from the 10-fold cross-validation. Since the original manuscript uses the maximum number of observations and incidence angles (equivalent to mosaic 5, or the furthest right data point on both plots), we conclude that the results are robust, since similar results are achieved with mosaics that use fewer observations or incidence angles.

We also considered the convergence of the detection thresholds α, β, and ϕ

themselves. Plots for these results are shown in the figures below as a function of number of observations (the plots for angular diversity are essentially the same). Similarly, we see that the thresholds converge fairly quickly and fall comfortably within the range of plausible thresholds inferred from the 10-fold cross-validation. We conclude that these thresholds are robust when all available data is used, and that the uncertainty introduced by spatial variations in the number of observations or range of incidence angles is less than the uncertainty from spatial variations in ice slab structure. We think that the minimum and maximum ice slab extent shown in the original manuscript adequately quantifies the uncertainty in ice slab extent, since it is derived from the minimum and maximum thresholds that define the grey boxes in the plots, which fully encompass the variability from observational geometry.



In the final manuscript, we will add these sensitivity analysis results and conclusions to a supplementary information section.

[R2] *I am also a bit concerned (or maybe just confused) about the testing of the algorithm. In the abstract, the authors state "The S-1 inferred ice slab extent is in excellent agreement with ice penetreting radar ice slab detections from spring 2017". However, I find this to be misleading since the training dataset was from spring 2017. Of course the S1 ice slab extent is in good agreement since the algorithm seems to be empirically derived from this data. Was there a completely independent dataset used to test the algorithm? Can it be tested with OIB data from a different year? It seems that the F1 scores given in lines 205-208 and Figure 7 were from the training dataset.*

You are correct that we do not use independent training and validation sets to estimate the optimal backscatter thresholds for mapping ice slab extent. When we say that the S-1 mapping and OIB detections are excellent agreement, we simply mean to highlight that we have shown that the S-1 backscatter has a strong relationship with near-surface ice content and therefore can be used to map the presence of ice slabs. Prior to this study, we did not know that this would necessarily be the case. Certainly, we can see from the issues with detecting the lower elevation limit of the ice slabs that even when using all available data to choose our empirical thresholds, there is still the potential for significant discrepancies between the airborne ice-penetrating radar detections and the best-fit S-1 mapping.

We considered using the smaller but independent dataset of airborne ice-penetrating radar detections from spring 2018 for validation. However, we felt that this would add significant complications to the manuscript, since we would need to address the interannual radiometric stability of S-1 and environmental conditions other than ice slabs that might drive interannual variability in mean winter backscatter to fully interpret the validation results. While we think this can be done, and it the obvious next step, we think it is beyond the scope of this paper which is focused on investigating whether S-1 is capable of or appropriate for mapping ice slabs at the ice-sheet scale in the first place.

Since we do not introduce independent data from a different year, we use the 10-fold validation scheme used to assess uncertainty in the optimal thresholds (and the resulting most likely ice slab extent), which is why we also provide the confusion matrices and F1 score for the minimum and maximum ice slab extent. In the 10-fold cross-validation, we only use ~10% of the data to select the thresholds and then evaluate the results on their agreement with the remaining ~90% of the data that was not used for estimating the thresholds. So, in this process, we have independent training and validation sets.

We will make some key changes throughout the manuscript to clarify our approach and avoid overstating our results in the final paper. Key changes are detailed below:

[1] The abstract will read as below, where amended text is noted in bold.
"Ice slabs are multi-meter thick layers of refrozen ice that limit meltwater storage in firn, leading to enhanced surface runoff and ice sheet mass loss. To date, ice slabs have largely been mapped using airborne ice-penetrating radar, which has limited spatial and temporal coverage. This makes it difficult to fully assess the current extent and continuity of ice slabs or to validate predictive models of ice slab evolution that are key to understanding their impact on Greenland's surface mass balance. Here, for the first time, we map the extent of ice slabs and similar superimposed ice facies across the entire Greenland Ice Sheet at 500 m resolution using dual-polarization Sentinel-1 (S-1) synthetic aperture radar data collected in winter 2016-2017. ***We do this by selecting empirical thresholds of the cross-polarized backscatter ratio and HV backscattered power that optimize agreement between airborne ice-penetrating radar data detections of ice slabs and the S-1 estimates of ice slab extent. Our results show that there is a sufficiently strong relation between C-band backscatter and the ice content of the upper ~7 meters of the firn column to enable ice slab mapping with S-1. We find that ice slabs are nearly continuous around the entire margin of the ice sheet. This includes regions in Southwest Greenland where ice slabs have not been previously identified, but where the S-1 inferred ice slab extent is in excellent agreement with the extent of visible runoff mapped from optical imagery.*** *The algorithm developed here lays the groundwork for long-term monitoring of ice slab expansion with current and future C-band satellite systems and highlights the added value of future L-band missions for near-surface studies in Greenland."*

[2] In Section 4.1, when introducing our results, we will explicitly discuss how the F1 score can be interpreted. Bolded text below will be added in this paragraph.
"Figure 6 shows the S-1 estimated ice slab extent in winter 2016-2017, compared with the OIB ice slab detections. We find ***strong*** *agreement between the upper limit of the ice slabs as identified by OIB and the S-1 estimated upper limit. Figure 7 shows the confusion matrices, F1 scores, and Cohen's κ for the minimum, most likely, and maximum S-1 estimated ice slab extent that quantify*

this agreement. The most likely ice slab extent has an F1 score of 0.811 with a true positive rate of 94% when detecting the upper limit of the ice slabs. ***However, it is important to keep in mind that the optimal values of α, β, and ϕ are derived from all available ice-penetrating radar detections. Therefore, the high F1 score quantifying the agreement between the OIB detections and most likely ice slab extent mapped by S-1 simply indicates that there is a sufficiently unique relation between S-1 backscatter and firn shallow ice content that S-1 backscatter can reasonably be used as a proxy to map ice slabs. The high F1 score does not provide information on whether α, β, and γ generalize to data collected in other places or at other times. However, the 10-fold cross validation scheme estimates α, β, and γ using only ~10% of the OIB data and validates the applicability of that threshold to the rest of the ice sheet using the withheld ~ 90% of the data. Therefore, the minimum and maximum ice slab extents, derived from this cross-validation scheme, show how well thresholds estimated in one region of the ice sheet can be generalized to the ice sheet as whole****.”*

[3] In the final manuscript, we will add an additional subsection in the methods that discusses regional variations that could lead to the variations in α, β, and ϕ that we see when the thresholds are selected based on data from a particular region (see response to R3 below).

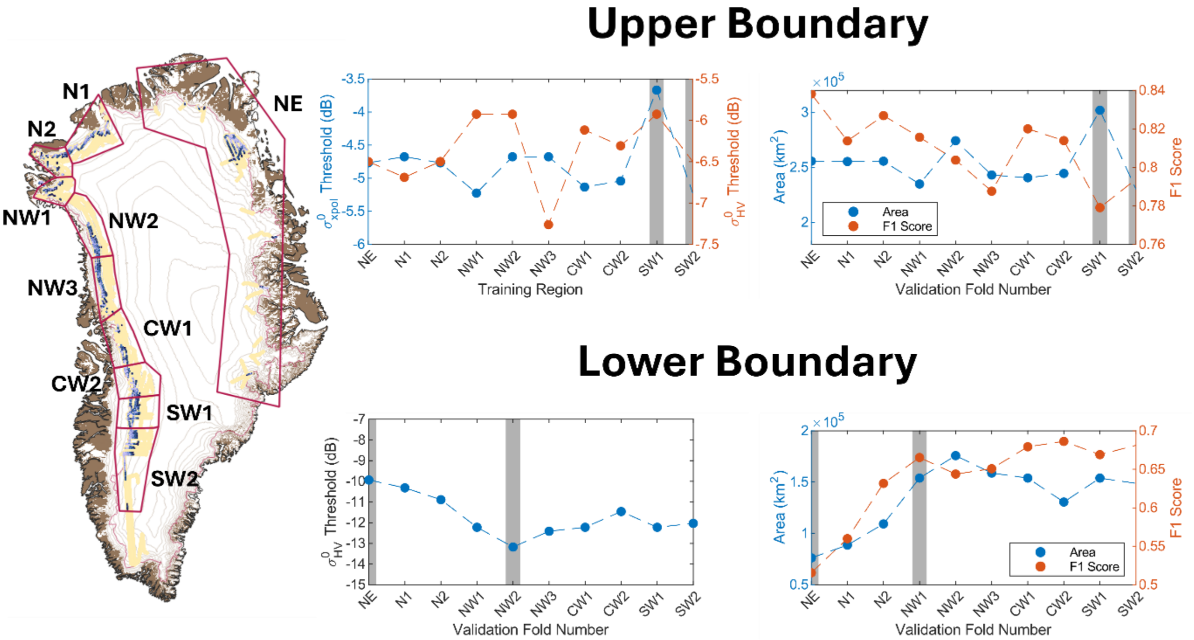
[4] We will clarify our explanation of the 10-fold cross-validation scheme to highlight why it is implemented and how we use it in place of an independent validation data set to quantify uncertainty in our results. Please see the response to Reviewer 2 or 3 for the full text of our revisions in Section 3.3.

[R3] *I am also a bit confused with how the folds were created. Were these folds selected completely randomly or separated by specific regions of the ice sheet. If the latter, this could provide insight into the spatial robustness of this algorithm. For example, in Figure 4, which region corresponds with fold 2 and why is the F1 score for the lower elevation limit so much worse in this region? The authors state in the caption of Figure 4 that “we discard the iteration marked with the red bar due to anomalously poor F1 score...” but it seems like this anomalously low F1 score should be important as it says something about the robustness of the algorithm? This should be further explored and discussed.*

Thank you – this is a very constructive suggestion on designing the folds! We now split the training data set into 10 distinct regions of the ice sheet. This leads to small variations in the size of each validation fold (e.g. total number slab and no slab observations in each fold), but allows us, as you suggested, to interrogate the spatial robustness. Results are shown in the figure below.

We find that the algorithm is quite spatial robust when delineating the upper elevation limit of the ice slabs. The thresholds vary by only ~± 1 dB depending on the training region. The F1 score on the withheld portion of the data set (e.g. the measure of how well the threshold generalizes to the rest of the ice sheet) varies between 0.78 and 0.84. Given that the F1 score when using the entire dataset for training is 0.81, this suggests very good generalizability for thresholds derived from only a subset of the data.

Not surprisingly, there is more variation in the results for delineating the lower elevation limit of the ice slabs, with distinct region variations. In particular, thresholds derived only from data in regions NE and N1 do not generalize well to the rest of the ice sheet, whereas the rest of the training regions generalize well. This might be for several reasons. First, the northern regions have the least number of ice slab detections, so thresholds derived from data in those regions may be overfit to conditions that are not representative of larger areas. Second, we see steeper gradients in backscatter as a function of elevation in the northern regions compared to the Northwest and Southwest. This suggests that small variations in threshold values would lead to large changes in ice slab area in the Northwest and Southwest, but small changes in ice slab area in the North and Northeast. As a result, the agreement between the OIB observations and S-1 detections is much more sensitive to the threshold values in the Northwest and Southwest than in the North and Northeast, so thresholds derived from NE and N1 do not generalize well. Finally, these regional variations might represent how well the lower boundary of refrozen ice in each region actually agrees with the modeled long-term equilibrium line used to cutoff the OIB detections, and therefore reflect errors or uncertainties in that model. We will update the final manuscript to include this new division of the folds and include a section discussing the regional differences we observe.



Minor comments

[R4] *In paragraph 1 of the introduction please also mention that mass is also lost due to dynamical processes, and it would be helpful to briefly compare this mass loss to that from surface processes.*

We will edit the opening of the first paragraph to read: “Over the last two decades, around half of mass loss from the Greenland Ice Sheet (GrIS) has come from the runoff of surface meltwater, with the remaining 45-50% attributable to ice dynamical processes and ice-ocean interactions in marine terminating sectors (Van Den Broeke et al., 2009; Enderlin et al., 2014; Mouginot et al., 2019). However, surface processes are projected to remain the dominant contributor to Greenland’s sea level contribution over the next century, particularly as the ice margin retreats onto land above sea level (Fox-Kemper et al., 2021). By extension, much of the uncertainty in future mass loss from the ice sheet can also be ascribed to uncertainty in surface processes (Fox-Kemper et al., 2021).”

[R5] *L28: “preferences the formation of perennial firn aquifers” is a bit awkward wording.*

We will edit this sentence to read:

“The southeast basin is the only major region where no ice slabs have been detected, due to the high snow accumulation rate that insulates subsurface liquid water from refreezing and leads to the formation of perennial firn aquifers (Forster et al., 2014; Munneke et al., 2014).”

[R6] *L42: “… including the first high elevation rain event, such as 2019, 2021, and 2023.” This wording makes it sound like the rain event occurred in 2019, 2021, and 2023.*

We will edit this sentence to read:
"With the end of the OIB mission in 2019, there is no new ice-penetrating radar data to improve these time series or assess the impact of more recent heavy melt seasons, such as 2019, 2021, and 2023, which included a significant high elevation rain event in August 2021 (Tedesco and Fettweis, 2020; Box et al., 2022, 2023)."

[R7] *L111: From Fig. 1, it looks like the HV backscatter is closer to -4 dB in the percolation zone.*
Corrected in the text as suggested.

[R8] *L112: "…eventually reaching a plateau around -11 dB". This is a bit misleading, I think. There is still substantial variation around this new plateau as the HV backscatter changes from -8 dB in the upper part of the ablation zone to -13 dB in the wet snow zone.*
We will edit this sentence to read:
"The percolation zone HV backscatter (σ_{HV}^0) is consistently about -4 dB, but decays at lower elevations as ice slabs begin to form and thicken, eventually plateauing around an average of -11 dB across the upper ablation and wet snow zones. However, there is significant local variability in the upper ablation and wet snow zones, with the HV backscatter varying from -13 dB to -6 dB around the mean."

[R9] *L191: "… we optimize independent backscatter thresholds…" What are the thresholds independent from?*
The backscatter thresholds are independent from one another – the choice of backscatter threshold to delineate the upper elevation limit of the ice slabs has no bearing on the threshold for the lower limit and vice versus (e.g. no joint optimization). To clarify this point, we will edit this sentence to read: "…we optimize ***separate*** backscatter thresholds…".

[R10] *L196: What is meant by "high-end estimate"?*
Jullien et al. (2023) developed a semi-automated routine for delineating ice slabs in ice-penetrating radar data, tuned to in-situ measurements from firn cores and GPR measurements at KAN-U the produced both a minimum and maximum likely ice presence. This is based on sensitivity tests that account for the fact that there is an overlap in the signal strength distributions between refrozen ice and porous firn. In their published data set accompanying their paper, they refer to the maximum likely ice presence estimate as the "high-end estimate" of ice slab extent, and we follow this terminology here. To clarify this point for readers who may be less familiar with their dataset, we will add the sentence:
"This high-end estimate corresponds to the maximum likely refrozen ice content given the observed ice-penetrating radar signal strength."

[R11] *L202: what step size did you use to test α and β within these ranges?*
We used step sizes of 500 in digital number space for both α and β (since data were original exported at 16-bit unsigned integers). In dB, this corresponds to a step size of 0.2 dB for α and 0.08 dB for β. Increasing the dB resolution for α to be the same at β leads to no meaningful change in the results. The optimal of α is reduced by 0.04 dB leading to a 0.002 improvement in the F1 score (from 0.8114 to 0.8116). Similarly, when delineating the lower boundary, changing the dB resolution for γ from 0.2 dB to 0.08 dB leads to a 0.08 dB change in γ and no change in the F1 score.

[R12] *After the 10-fold validation, how were the optimal empirical parameters chosen?*
The optimal empirical parameters are based on the optimization using the entire OIB dataset from the whole ice sheet. We use the 10-fold validation to assess uncertainty in that optimal estimate by considering how the estimated ice slab extent might change if only part of that data were used

for the threshold optimization. From amongst the ten different sets of thresholds produced by this cross-validation scheme, we pick the two sets of thresholds that produce the largest and smallest total ice slab extent to conservatively represent this uncertainty range. In response to this comment, as well as comments from both R2 and R3, we will edit the portion of the manuscript describing the 10-fold cross-validation scheme to improve clarity.

[R13] *L243: Please add Dunmire et al 2021 with Koenig et al 2015 citation.*
Added.

[R14] *Lines 285-290: I find this section confusing. Isn't "ice formed be refreezing" (L286) the same as an "ice slab"? The distinction between ice slabs and other refrozen ice is unclear throughout this section. Also, it seems that "ice formed by refreezing induces significant volume scattering due to trapped air bubbles…" (L286) contradicts the introduction "with relatively little volume scattering since heterogeneities such as air bubbles are significantly smaller than the C-band wavelength" (L95).*
We agree that this section can be somewhat confusing, in part because of imprecision in the language we have to describe ice sheet facies that form in the equilibrium zone. Ice slabs are generally understood to be multi-meter thick layer of refrozen ice that are perched over any otherwise porous and permeable relict firn layer. In this technical sense, areas where refrozen ice sits directly on top of meteoric ice would not be considered ice slabs. In some places, this ice might meet the definition of superimposed ice, which is ice that forms by refreezing within the annual snowpack on top of an otherwise solid ice column. However, in other places, this layer of refrozen ice over meteoric ice might form where water drained through crevasses into deep firn, completely filling it before refreezing, or where older ice slabs have exhumed through advection and ablation, or through other modes that are not yet well characterized. We will clarify this point in the revised manuscript by explicitly listing the definitions given above for readers who may not be familiar with technicalities.

Good point on L286! What we wanted to express here is that refrozen ice may contain remanent interstitial firn or other void space of sizes closer to the radar wavelength due to heterogeneous infiltration and refreezing, not air bubbles as might be found in ice formed via compaction. We will revise this sentence to read: "We hypothesize that any ice formed by refreezing induces significant volume scattering due to trapped interstitial firn pockets, void space, or other heterogeneities in density…"

Technical corrections
[R15] *L22: Please add "meltwater" before "retention and runoff"*
Adjusted in the text as suggested.

[R16] *L32: Please add "elevation" in "upper elevation limit" (also for L194).*
Adjusted in the text as suggested.

[R17] *L150: Delete "to" before "correction" at the start of this line.*
Adjusted in the text as suggested.