Response to Referee #1 (Alek Petty)

In this document, we outline our responses to comments from the referee, including modifications that we intend to make to the manuscript where necessary.

Referee comments are shown in black and our responses in blue with intended changes to the revised version of the manuscript given in italics.

Review of "Year-Round High-Resolution Sea Ice Freeboard Retrieval Using ICESat-2 ATL03 Photon Data" by Liu et al.,

Review by Alek Petty

This manuscript presents a new high-resolution photon-based lead classification and height/freeboard retrieval framework using ATL03, with the goal of improving sea surface height and derived freeboard estimates relative to the current ATL07 and ATL10 products. The work builds on recent machine-learning lead classification efforts (by the primary author and others) that showed strong promise compared to the pre-launch empirical ATL07 classification approach, and I think this is the logical next step. They combine this new ML trained classification scheme with a much higher resolution processing approach.

In general, the study clearly involved a considerable amount of analysis and description, and I really commend the effort. The paper was also very clearly written and the figures were all of high quality. The study is also very appropriate for this journal. That said, I have a number of questions about the robustness of the approach and assessments presented. In a few places it feels like the algorithm works nicely for the examples shown, but I am less convinced about how broadly applicable it is beyond those cases and I think there is a potential need for more filtering and at the very least clear documentation of the processing steps involved to ensure transparency and reproducibility. My comments are below.

Thank you for taking the time to read and check our manuscript, providing detailed and valuable comments that improve the clarity of the work.

Major comments

Height precision:

I didn't see any mention in the paper of the fact ATL07 uses a ~150 aggregation (pulses are aggregated towards this number, but the actual photons aggregated is variable) motivated by a desire for something like a 2-3 cm precision over flat surfaces for lead retrieval considering the theoretical ~10-20 cm precision of ATL03 heights (Markus et al., 2017, Neumann et al., 2019, Kwok et al., 2019). In reality the ATL03 precision is probably a bit than that (Brunt et al., 2019), but this is hard to

interpret. That motivation needs to be discussed more. The clear downside of the ATL07 focus on lead precision is the reduced along-track resolution and issues with very long segment lengths.

We agree that the motivation for the ATL07 aggregation strategy was not sufficiently discussed in the original manuscript. In the revised manuscript, the following will be added in the introduction to clarify that the approximately 150-signal-photon aggregation used in ATL07 is motivated by the need to achieve high precision for lead-height retrieval:

Second, ATL07 uses variable segment lengths to accumulate sufficient signal photons (~150) for precise surface-height retrieval, particularly over flat lead surfaces used for sea-surface referencing. This precision-oriented design is well suited for stable lead-height retrieval, but it can also smooth fine-scale ice morphology when segment lengths become long (10-200 m for strong beams; 40-800 m for weak beams), thereby reducing the ability of ATL07 to resolve ridge-related height variations.

We can see by your height SD plots that your use of a much finer fixed 5 m along-track photon aggregation is resulting in roughly 10 cm height standard deviations over open water leads, which is quite high! This all needs to be discussed way more in terms of impacts on freeboards.

As defined in Kwok et al., (2019), **precision** refers to the repeatability of height measurements over a given spatial scale, typically assessed as the standard deviation of height estimates over flat surfaces at kilometer scales (e.g., 3–10 km). In that study, precisions of 1.5–1.9 cm were obtained by aggregating some segments (Fig R1). As shown in Fig.8 of the original manuscript, our method achieves comparable centimeter-level precision when evaluated over kilometer-scale flat leads as well. In contrast, the standard deviations shown in Fig. 4 of the original manuscript represents the within-segment photon height variability. This metric is different from the ATL07 precision definition used in Kwok et al., (2019) and is instead analogous to the Gaussian width parameter reported in ATL07. This is supported by the strong consistency between the HRFM within-segment height standard deviation (STD) and the ATL07 Gaussian width parameter, as shown in Fig. R2.

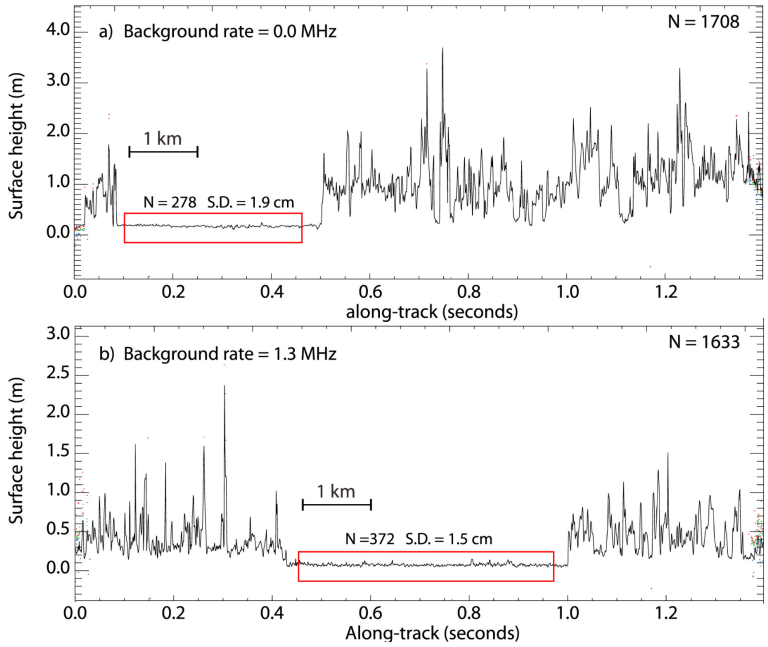

Fig R1 Kilometer-scale height precision over flat leads reported for ATL07 (Kwok et al., 2019).

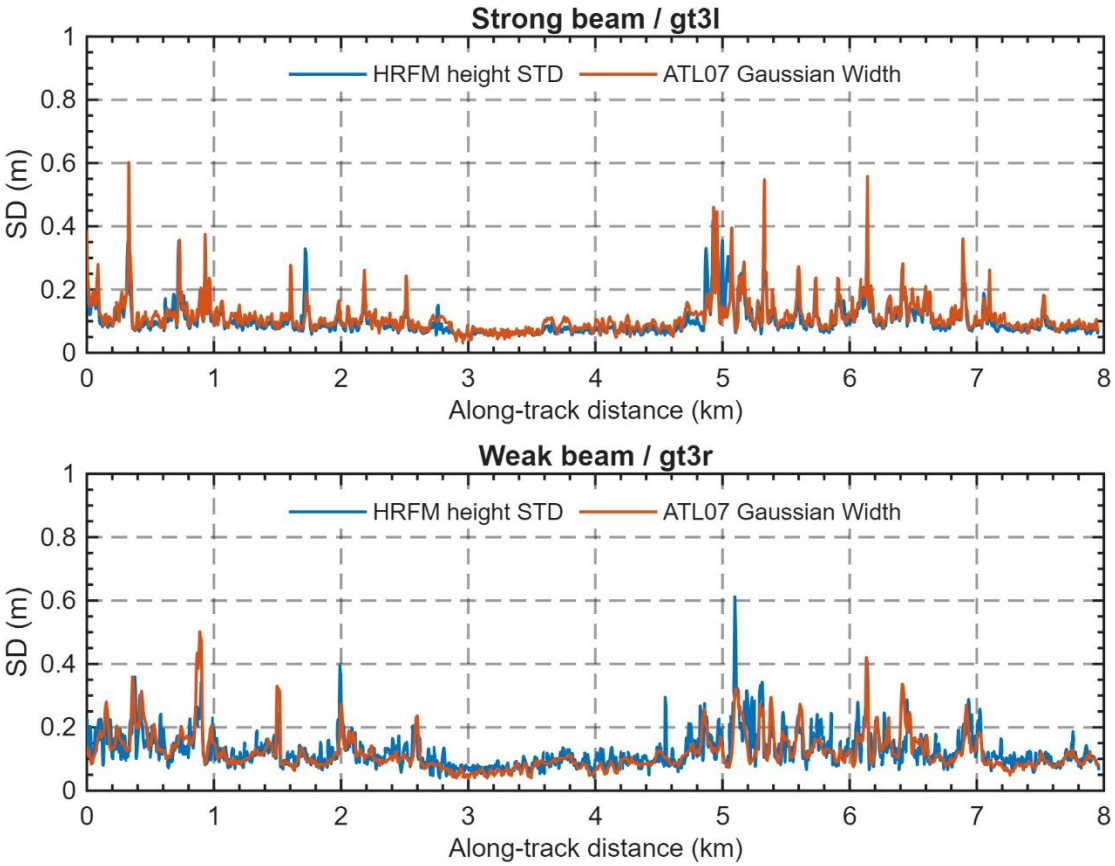

Fig R2 Comparison between HRFM within-segment height STD and ATL07 Gaussian width.

Related – I find some of the discussion around ATL07 height biases over ridges to be

over stated, a lot of the difference seems to be resolution/sampling differences and not a bias.

You are right. We will revise the relevant text to avoid implying a general ATL07 bias over ridges and describe these features more cautiously as resolution- and sampling-related height differences.

ATM comparisons:

The ATM comparison section needs more clarity. For one, it is not clear whether you implemented the cross-correlation maximization procedure used in Kwok (2019). In that study, raw ATM was aggregated into 17 m by (segment length plus 17 m) blocks with Gaussian weighting, and the ATM data were shifted to maximize correlation with ICESat 2. What exactly have you done here? Are you using the same aggregation? Are you shifting the profiles? Which beams are included in the comparisons? Note that the current thinking is ~11 m footprints (Magruder et al., 2020), but maybe 17 m would keep things consistent with Kwok.

Your reported correlations and RMSE values between ATM and ATL07 are notably worse than the very high values reported in Kwok (2019). The lack of details regarding the processing and comparison approach needs a lot more explaining to make such a big claim (unless I am missing something!). It would also be helpful to repeat the April 8 and 12 flights shown in Figure 1 of Kwok (2019) as a sanity check. At the moment it is difficult to reconcile and trust the differences presented.

We will add a clearer description of the ATM comparison procedure in the revised manuscript. In our original analysis, we did not fully implement the cross-correlation maximization procedure used in Kwok et al. (2019). Instead, ATM and ICESat-2 profiles were compared directly without applying an optimal along-track shift. The raw ATM data were aggregated within an approximately 11 m footprint (segment length plus 11 m) for ATL07. The beams included in the comparison were gt2l (strong beam) and gt2r (weak beam). We note that gt2r was not included in the corresponding analysis of Kwok et al. (2019).

We will clarify that our ATM comparison is therefore not a strict replication of the Kwok et al. (2019) validation procedure. Because HRFM and ATL07 are both derived from the same ATL03 ground tracks, their relative performance can still be compared under the same non-shifted collocation framework. However, the resulting RMSE and correlation values should not be directly compared with the cross-correlation-optimized results reported by Kwok et al. (2019). We will revise the text accordingly and avoid over-interpreting the differences from that study.

Following the referee's suggestion, we re-examined the ATM underflights acquired on 8, 12, 19, and 22 April 2019. In the revised manuscript, the 22 April case is used as the main validation example because it provides the clearest overlap with both the strong

and weak ICESat-2 beams. The 8 April and 12 April (Fig R3) flights are now included as additional sanity checks in the Appendix. These additional comparisons confirm that the validation statistics are sensitive to the degree of ATM–ICESat-2 spatial overlap and to the use of non-shifted collocation. In particular, the 12 April case shows degraded statistics for both HRFM and ATL07 under poorer overlap. The 19 April underflight was also examined, but it was not used for quantitative statistics because the spatial overlap with the corresponding ICESat-2 ground track was insufficient for a reliable segment-wise comparison.

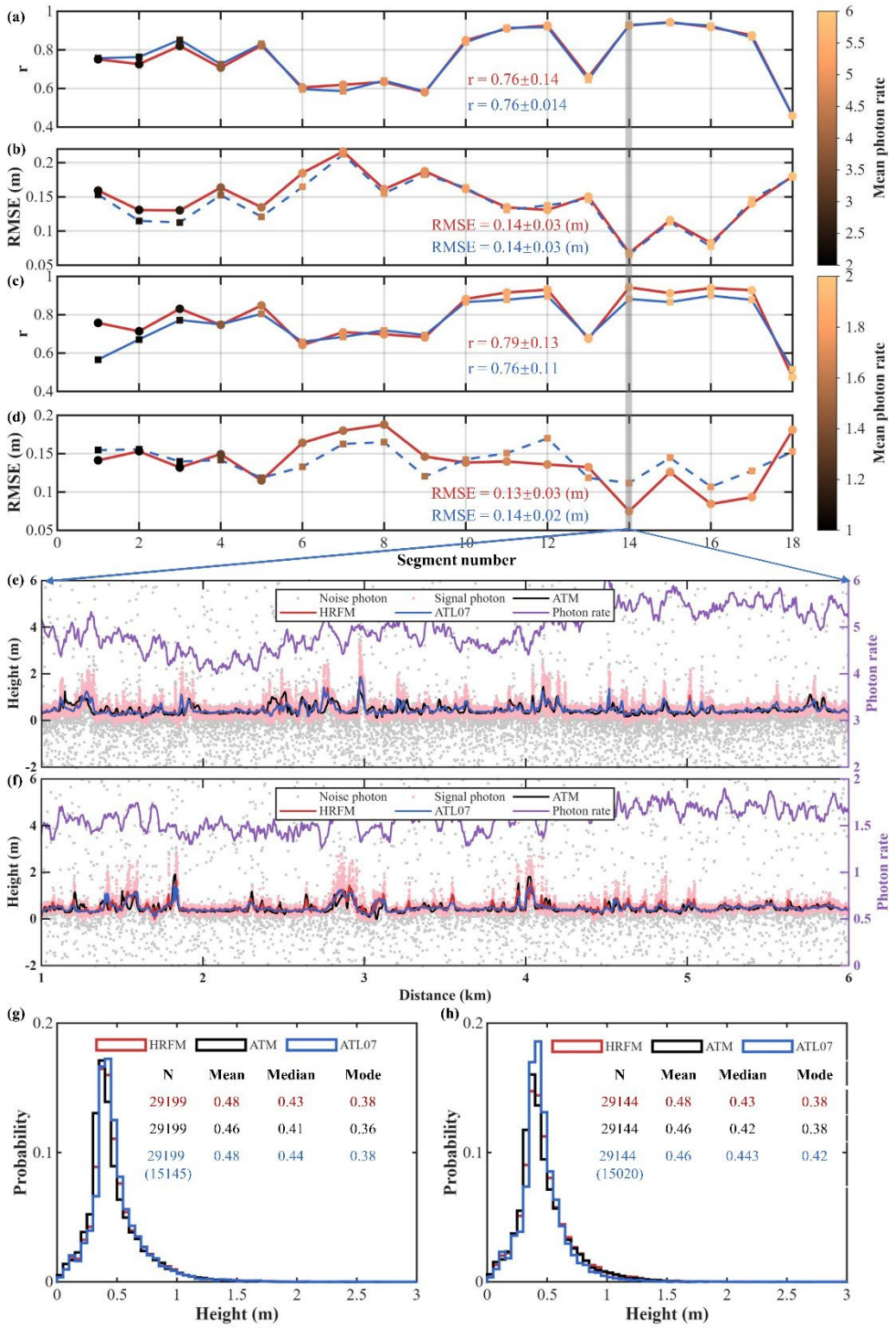

Fig R3 Evaluation of HRFM and ATL07 surface-height retrievals against ATM observations obtained on 19 April 2019. The comparison was performed using direct, non-shifted collocation, with HRFM, ATL07, and ATM heights referenced to the DTU21 mean sea surface. (a, b) Correlation coefficient (r) and root-mean-square error (RMSE) between ATM and ICESat-2 surface heights from HRFM (red) and ATL07 (blue) for the strong beam. (c, d) Same as (a, b), but for the weak beam. The color scale indicates the mean photon rate for each segment. (e, f) Surface-height profiles of segment 14 for strong and weak beams, respectively, including ATM heights (black), HRFM heights (red), ATL07 heights (blue), HRFM-identified signal (pink) and noise (gray) photons, and the corresponding photon rate (purple). (g, h) Distributions of surface-height estimate for all 25 segments for strong and weak beams, respectively, with the number of observations (N), mean, median, and mode indicated for each dataset. Numbers in parentheses denote the original number of ATL07 samples before resampling.

Geophysical corrections:

The treatment of geophysical corrections is glossed over very quickly. What corrections are applied, and are they the exact same as ATL07? I think you are using a different MSS for one. These details matter, especially since you are doing height comparisons with ATM (unless I'm missing that you did something else to reconcile the heights). In previous SSH comparison work, getting the corrections consistent took quite a lot of effort (Bagnardi et al., 2021). I would encourage a more careful and explicit description here.

In the revised manuscript, we will provide a more explicit description of the corrections used in HRFM. Specifically, the HRFM heights inherit the geophysical corrections from ATL03, including ocean tide correction, the long-period equilibrium tide correction, and the dynamic atmospheric correction. For the comparison with ATM, HRFM, ATL07, and ATM elevations are all referenced consistently to the DTU21 mean sea surface. We will list these correction terms explicitly in the method section.

First photon bias:

I also saw no mention of the first photon bias correction. Even if your ultimate goal is relative freeboard, once you start comparing heights directly (again as you do here with ATM data) this correction should be included and documented, and it could still impact the freeboard results too as this is variable (I do confess I'm not sure exactly how variable it is over sea ice/leads…).

The first photon bias correction was applied in our processing but was not sufficiently documented in the original manuscript. We will clarify this in the revised manuscript. Specifically, we used the first-photon-bias correction tables provided in ATL03. The correction was estimated for each segment using the height-distribution width, signal strength, and average detector dead time, and was applied before the ATM height

comparison and freeboard retrieval.

Filtering:

The photon filtering strategy needs more explanation. ATL07 uses windowing filters to separate signal from background photons, I was surprised you do not do any windowing. I wondered what would happen in cloudy conditions with your approach as I don't think you are implementing any type of cloud filter. The coarse filtering does not appear to work very consistently, but the second step does. I am a bit worried you're showing the best cases here, which is why I would like to see the along-track data included with the release (see comment later).

We will clarify this point in the revised manuscript. Although the original text did not emphasize it, HRFM includes a window-based constraint during the coarse denoising stage. Specifically, after the initial density-based coarse signal extraction, candidate signal photons are further checked against a local height window defined from the modal height in the 3 km along-track window. Candidate photons are retained only within the range from 2 m below to 6 m above the modal height. This step is designed to retain ridge returns while reducing the risk of selecting erroneous returns under cloud-attenuated or noisy conditions.

Fig R4 shows a cloudy-condition example from 22 April 2019, where the raw ATL03 photons do not show a clear surface return because of cloud-related signal attenuation. HRFM still identifies the residual surface signal, and the retrieved height remains broadly consistent with the ATM profile. The release of along-track data is discussed in the Data availability response below.

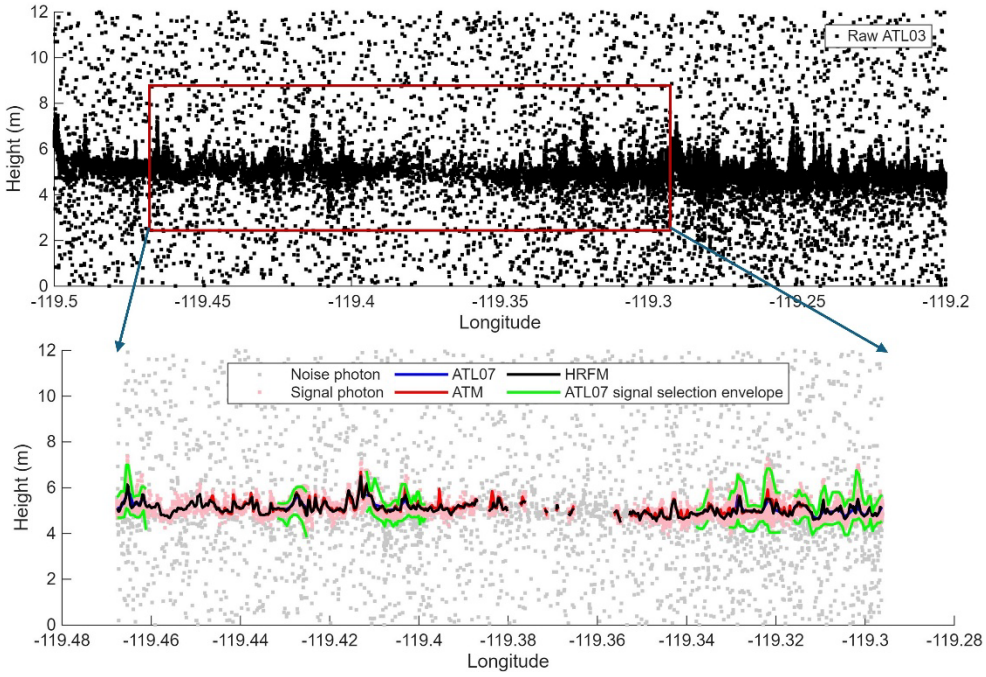

Fig R4 Performance of HRFM photon filtering under cloudy conditions.

Small related point - Why not use the new Yet Another Photon Classifier (YAPC) variables from ATL03? A lot of effort has gone into developing those precisely for photon level signal/background classification like this! Fine if you want to do your own thing, but I think worth at least acknowledging as that data is now all on ATL03 (and ATL07).

The following text is added to section 3.1.2:

We note that the Yet Another Photon Classifier (YAPC) variables provided in recent ATL03 and ATL07 releases offer an additional density-like metric, which could be incorporated as an additional filtering variable in future versions of HRFM.

Extra processing concerns:

ATL07 reduces the energy of the middle beam by a factor of 0.82. Is this applied here? You report 2 to 3 cm freeboard differences between strong and weak beams. I would not call that slight. It would be clearer to show this as a difference plot rather than as two separate seasonal curves.

How do you treat saturated or highly specular returns?

Release 005 introduced filtering using the PODPPD flag and beam angle in ATL03 to identify pointing issues. I think if you want to introduce this as a dataset you need to consider all these important edge cases and filter them out (or decide what to do with them).

We noticed that the ATL07 classification applies the relative gain of each beam (beam_gain, [1.0, 1.0, 0.82, 1.0, 1.0, 1.0]) for each beam and did the same thing.

We agree that a 2–3 cm strong–weak beam freeboard difference should not be described as “slight”. We will revise this wording and discuss the remaining inter-beam difference more cautiously. To make the comparison clearer, we provide an additional diagnostic in Fig R5, which shows the monthly freeboard distributions for strong and weak beams under the same HRFM processing framework. The strong- and weak-beam distributions largely overlap in most months, indicating broadly consistent retrieval behavior, although small centimeter-scale differences remain.

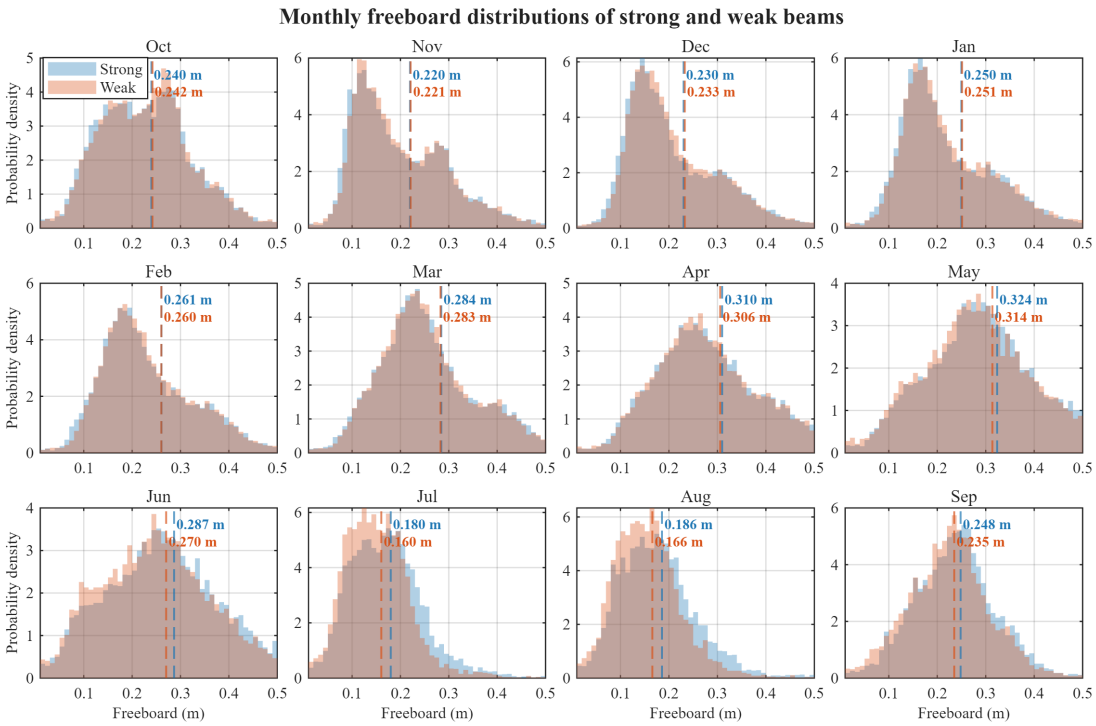

Fig R5 Monthly freeboard distributions derived from strong and weak beams using HRFM.

Saturated or highly specular returns were retained for height estimation and surface classification (although they may be affected by first photon bias heavily), but they were not used in the final freeboard estimation.

We also appreciate the referee's point about the PODPPD flag. It's will be applied in future dataset construction.

Classification training:

I am a bit confused about the training framework. You mention leave one out cross validation which is good, but it appears that all scenes are then used to train the final classifier that is evaluated and shown in Figure 9? Or maybe I got confused about that. Please clarify, as if so, that is not a fully independent test and should be clarified?

Related – I didn't quite get/see if the models were trained for weak and strong beams separately? I think you need to apply the middle beam gain correction if not trained by beam?

As in the above, I wasn't sure what you are doing about very specular/saturated pulses.

The classification framework will be clarified in the revised manuscript. For model evaluation, we used a leave-one-scene-out cross-validation strategy. In Fig. 9, each example was classified using a model trained without that scene; that is, the training data came from the other 24 scenes. Therefore, the classification examples shown in Fig. 9 are independent of the corresponding test scenes.

After the cross-validation indicated that the classifier was robust, the final classifier used for the full-year HRFM production was trained using all available training scenes.

Strong and weak beams were trained separately. Please see the response above for beam-gain correction and saturated pulses.

The following text is added to section 4.2:

After this independent evaluation, the final binary RF classifier was trained using all 25 labelled scenes and applied to the full-year HRFM data.

Data availability:

I strongly encourage the authors to make the along track height estimates and classification outputs publicly available. Having access to these intermediate products is essential for assessment of these new SSH and freeboard approaches. Much of what we have learned about the strengths and deficiencies of ATL07 has come from the community being able to interrogate the along track heights and classifications directly independently. Given the standards and expectations of The Cryosphere, I believe these along track outputs should be made available as part of the publication.

The gridded freeboard data have already been uploaded to a public repository. Because the full along-track ATL03-derived product is very large, we will prioritize the release of the along-track height estimates and classification outputs used directly in this study, including at least the data used for model construction and validation. Other data is available upon request.

Seasonal analysis:

You only have one winter of data, so the seasonal discussion felt quite premature and over confident. I would suggest tempering that interpretation considerably until more years are processed.

We will temper the seasonal interpretation in the revised manuscript. The aim of this study is not to provide a definitive multi-year assessment of seasonal sea-ice freeboard variability, but to demonstrate the applicability of HRFM under both winter and summer conditions. We are currently processing more years of ATL03 data (the full ATL03 data volume is too large to process quickly).

Figure 9c:

I include in major comments as I find this quite confusing/worrying (?). But in the middle profile for the bottom lead, *height_segment_type* shows sea ice while *ssh_flag* is set to sea surface on one of the beams (see yellow circles). How is that possible?

We checked Fig. 9c carefully and found that the issue was caused by a plotting error. The ATL07 classification in this case contains a -1 value, corresponding to an invalid

segment. This category had not appeared in the other examples, and the original plotting code did not explicitly handle it. As a result, the -1 category was assigned to the first color in the colormap, which shifted the color mapping of the remaining ATL07 classes. We have corrected the plotting code by explicitly assigning colors to all ATL07 class values, including the -1 invalid category. The corrected version of Fig. 9c is shown in Fig R6. We will revise the text related to the figure.

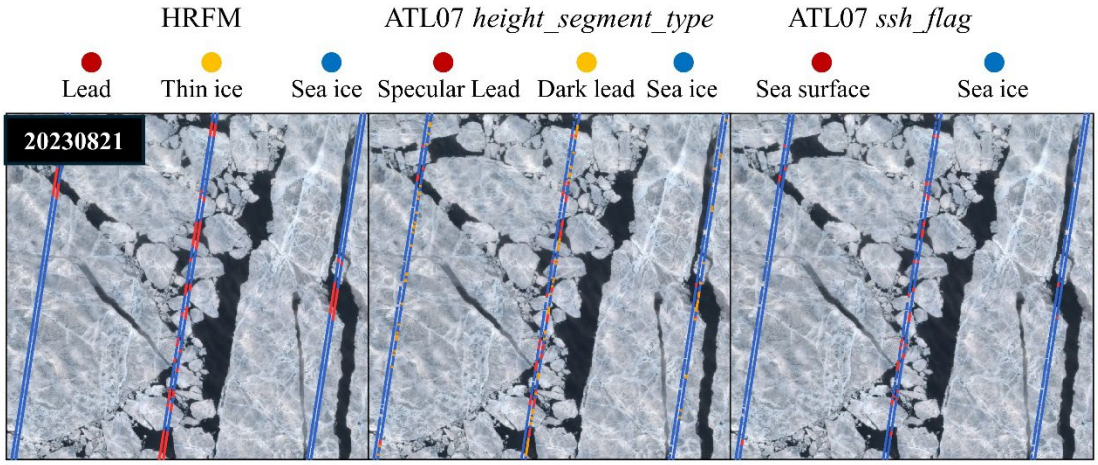

Fig R6 Corrected plot of Fig. 9c

Minor comments

On the multi beam discussion, much of the concern in the literature relates to using sea surface points across beams to generate a swath like SSH estimate. Your approach seems to combine independently derived freeboards across beams, which is conceptually simpler and something ATL20 could easily do now anyway. I would make that distinction clearer.

We will revise the multi-beam discussion to make clear that this is conceptually different from cross-beam SSH interpolation or swath-like sea-surface construction. The following is added:

In the present HRFM framework, this means combining freeboard estimates that are independently retrieved for each beam, thereby increasing effective sampling density, improving spatial coverage, and reducing gridded uncertainty, particularly in regions and seasons where observations are otherwise sparse (Ricker et al., 2023). A related but distinct approach is to construct a swath-like sea-surface reference to retrieve sea-ice freeboard by using sea-surface points across beams, which has also been explored in ATL07/ATL10 algorithm development (Kwok et al., 2026).

It would also be useful to test the algorithm on the cloudy but usable scene discussed in Kwok (2021), just to see how it behaves under less ideal conditions. This is a pretty big issue considering the dropping of the dark lead classification from ssh/freeboard in ATL07/10.

We agree that testing HRFM on the cloudy but usable case discussed in Kwok (2021) would be valuable. We will attempt to process this case and include the result in the revised manuscript or supplement if it can be completed within the revision period.

I would use the term background rather than noise, since noise suggests a sensor issue rather than environmental background photons being reliably detected.

We will replace "noise" with "background" where appropriate.

L71: I would not characterize this as a clear underestimation. These often seem like smoothing/resolution differences to me that are over-characterized as height biases.

We agree. We will rephrase the statement to describe these differences more cautiously as smoothing- or resolution-related differences rather than clear height underestimation. The text added in the revised manuscript has been replied above.

L126: Please clarify the version number.

We will clarify the product version number in the revised manuscript. The following text is added in the section 2.1:

In this study, our HRFM utilizes ATL03 (Version 6) as the primary input to retrieve surface height, surface type, and freeboard (Sect. 3). ATL07 (Version 6) and ATL20 (Version 5) are used for comparison and evaluation purposes (Sect. 4). Although the newer release ATL07 (Version 7) introduces fixed 10 m segment lengths, it was experimental and limited to strong-beam observations during the study period. Therefore, to ensure a consistent and comprehensive evaluation across strong and weak beams, we use ATL07 (Version 6) as the comparison reference unless otherwise specified.

L149: Also worth noting that there is an additional ssh_flag = 2 flag in ATL10 for the segments actually used.

Good suggestion. Will take a look.

L214: Clarify whether a different MSS is being used and what corrections are inherited from ATL03.

We will clarify that the validation was performed after placing HRFM, ATL07, and ATM on the same DTU21 MSS reference. A new table is added in the appendix to document the source of corrections.

L260: Please list the geophysical corrections and their source and comparisons with ATL07. ICESat-2 has a geophysical corrections document that lists this all out.

A new table is added in the appendix to list the geophysical corrections used in HRFM and ATL07, including which correction terms are inherited from ATL03 and whether the same MSS is used.

L265: OK fine, but ATL07 provides both and I think you just focus on norm, so that gets a little confusing later.

Yes, we used the normalized background rate. The raw background rate can be preserved as well. Will make it clearer.

L266: Make clearer that ATL07 aggregates in order to beat down the noise and achieve a precision of 2 cm over flat surfaces, as driven by mission requirements.

Please see the reply to **Height precision** in major comments.

L316: Do you really expect 10 cm surface roughness over a level ice lead?

This point has been explained in 'Height precision' above.

L443: These could still be small leads. If possible, show a coincident Sentinel 2 scene and compare classifications.

Coincident Sentinel-2 imagery has been added for the two selected cases, together with the ICESat-2 ground tracks and HRFM surface-type classifications.

Replace ‘dark ice’ with ‘dark lead’ in the ATL07 figure labels.

It will be corrected.

References

Kwok, R., Markus, T., Kurtz, N. T., Petty, A. A., Neumann, T. A., Farrell, S. L., Cunningham, G. F., Hancock, D. W., Ivanoff, A., and Wimert, J. T.: Surface Height and Sea Ice Freeboard of the Arctic Ocean From ICESat-2: Characteristics and Early Results, Journal of Geophysical Research: Oceans, 124, 6942–6959, https://doi.org/10.1029/2019JC015486, 2019.

Kwok, R., Kacimi, S., Markus, T., Kurtz, N. T., Studinger, M., Sonntag, J. G., Manizade, S. S., Boisvert, L. N., and Harbeck, J. P.: ICESat-2 Surface Height and Sea Ice Freeboard Assessed With ATM Lidar Acquisitions From Operation IceBridge, Geophysical Research Letters, 46, 11228–11236, https://doi.org/10.1029/2019GL084976, 2019.

Kwok, R., Kurtz, N., Wimert, J., Petty, A., Cunningham, G., Markus, T., Hancock, D., Ivanoff, A., Bagnardi, M., Herzfeld, U., Trantow, T., and ICESat-2 Science Team: Ice, Cloud, and Land Elevation Satellite (ICESat-2) Project Algorithm Theoretical Basis Document (ATBD) for Sea Ice Products, version 7, https://doi.org/10.5067/KPMXUOH7TNIY, 2026.

Response to Referee #2

In this document, we outline our responses to comments from the referee, including modifications that we intend to make to the manuscript where necessary.

Referee comments are shown in black and our responses in blue with intended changes to the revised version of the manuscript given in italics.

The manuscript is well written and clearly structured, and it presents a promising high-resolution freeboard retrieval method using ICESat-2 ATL03 photon data. The results show improved performance relative to existing ICESat-2 products, and the method may also increase the usability of weak-beam observations for sea-ice freeboard retrieval. However, some methodological details require further clarification. In particular, the machine-learning workflow should be described in more detail, especially the link between GMM clustering, Sentinel-2-based human interpretation, and RF classification. In addition, the diversity of the Sentinel-2/ICESat-2 coincident cases should be better documented, including their seasonal coverage and time offsets, to support the claimed year-round applicability of the method.

We appreciate the reviewer's positive assessment and constructive suggestions that improve the clarity of the work.

Specific comments:

P6, L168: Because both the GMM clustering and the Sentinel-2-based manual interpretation rely on the 25 coincident scenes, the authors should show how these scenes are distributed across months/seasons and surface conditions. A table summarizing the acquisition dates, Sentinel-2–ICESat-2 time offsets, beam types, and assigned surface classes would help assess whether the training data adequately support the claimed year-round applicability of the classifier.

We have added a table in the Appendix to document the 25 Sentinel-2/ICESat-2 coincident scenes, including the Sentinel-2 and ICESat-2 file names, acquisition dates, and the time offset between the two observations. All six ICESat-2 beams, including three strong beams and three weak beams, were used in the training data set.

Figure 2: As I understand the workflow, the authors first grouped HRFM segments into 20 GMM clusters and then manually assigned these clusters to three surface classes using Sentinel-2 imagery. However, this important step is not easy to follow from Fig. 2. I suggest revising the flowchart so that the intermediate GMM clustering step and the subsequent human interpretation/manual class assignment based on Sentinel-2 imagery are shown more explicitly. This would make the training-data generation procedure clearer to readers.

We will make the training-data generation workflow clearer, as Figure 1 shown.

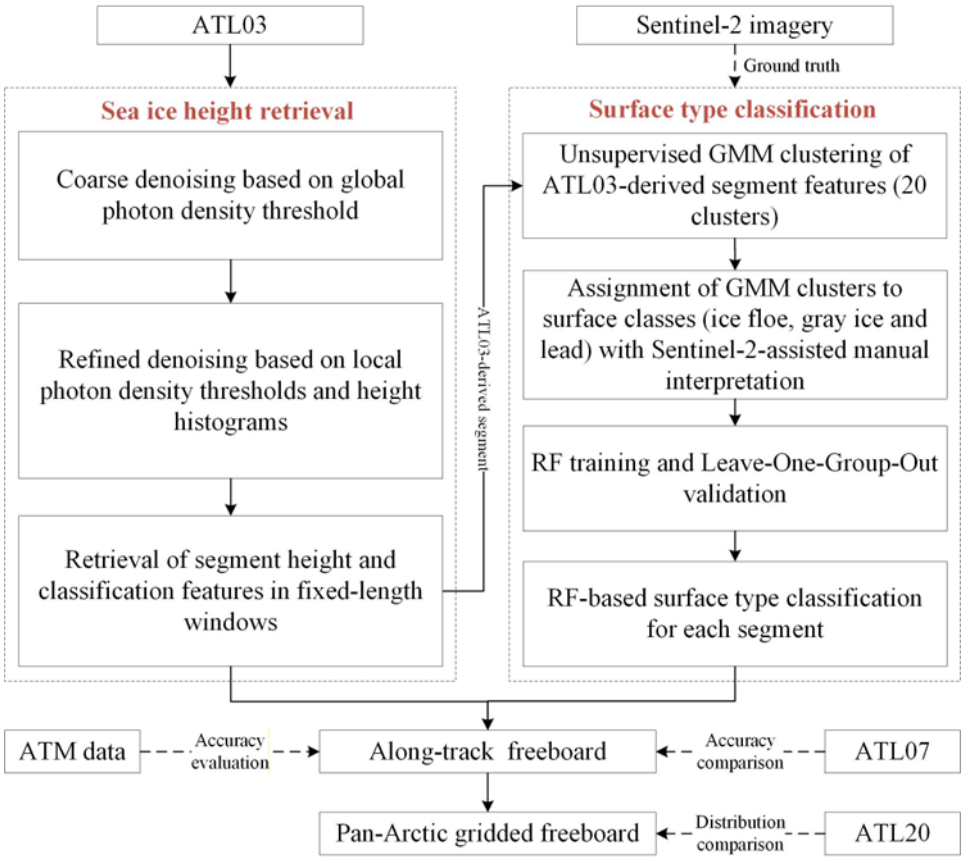

Figure 1 The flowchart of HRFM

P11, L280: More detail is needed on how transitional or ambiguous classes were handled during the Sentinel-2-based manual labeling, especially for thin ice, dark leads, and melt-pond-covered ice. This subjectivity may affect the RF training labels and should be discussed as a source of uncertainty.

We will add the following text to the section 3.2.2:

For freeboard retrieval, misclassifying non-lead surfaces as lead segments is more problematic than the opposite case, because false lead detections can bias the local sea-surface reference and subsequently propagate into the freeboard estimates. Therefore, clusters were assigned to the lead class only when they appeared as clearly dark, open-water-like features in Sentinel-2 imagery and were consistent with the local ATL03 photon characteristics. Ambiguous surfaces were not labelled as leads unless they could be confidently interpreted as open-water leads.

Several transitional surface types remain difficult to separate using the available optical information. Therefore, the thin-ice class is used as a broad category that includes visually similar transitional ice types, such as nilas, frazil ice, and slush ice. In addition, melt-pond-covered ice clusters was conservatively treated as sea ice class, even when the ponds appeared dark in Sentinel-2 imagery. This choice was made to avoid misclassifying melt ponds as leads and incorrectly using them as sea-surface reference points in the freeboard retrieval. It should be noted that the manual labeling inevitably introduces some subjectivity and classification

uncertainty, especially for transitional or ambiguous surface types.

Figure 10 shows systematic seasonal differences between HRFM and ATL20, with HRFM generally lower during the cold season and higher during the melt season. The manuscript attributes these differences to surface-height retrieval, lead detection, and reference sea-level construction, but the explanation would be stronger if supported by more quantitative evidence. For example, monthly differences in the number of detected lead segments, reference sea-level estimates.

We have discussed the potential sources of the freeboard differences in Sect. 5.2 using along-track data, including the differences between HRFM and ATL07, which provides the input data for ATL20, in surface-height retrieval, surface classification, and reference sea-level construction, as well as their combined impacts on freeboard estimates.

However, a full decomposition of the HRFM–ATL20 differences is limited by product availability, because ATL20 does not provide key intermediate variables such as detected lead segments and reference sea-level estimates.

Table 2 shows that a non-negligible number of thin-ice segments are classified as leads, particularly for weak beams. Since HRFM uses identified lead segments to estimate the local sea-surface reference, misclassified thin ice could bias the reference sea level and propagate into the freeboard estimates. This issue is only indirectly discussed through the low thin-ice classification performance, but its potential impact on freeboard retrieval should be addressed more explicitly.

The thin-ice classification shown in Table 2 was used to evaluate the potential of ICESat-2 photon data to distinguish transitional ice surfaces, but this three-class classifier was not used to generate the final year-round HRFM freeboard product. The reason is that the thin-ice class has relatively high classification uncertainty, especially for weak beams and for surfaces near the lead boundary.

For the final freeboard retrieval, we therefore used a more conservative binary sea-ice/lead classifier. In this binary classification, thin ice was merged into the sea-ice class rather than treated as an independent class. This design was chosen specifically to reduce the risk of using thin-ice segments as lead segments for local sea-surface reference estimation. The binary classification performance was reported in Table A1 of the original manuscript. For the strong beam, the precision/recall values are 0.99/0.99 for sea ice and 0.96/0.95 for leads. For the weak beam, the corresponding values are 0.99/0.99 for sea ice and 0.95/0.91 for leads.

Therefore, the thin-ice-to-lead confusion shown in the three-class experiment does not directly propagate into the final year-round freeboard product in the same form. However, we agree that any remaining misclassification of thin ice or other sea-ice surfaces as leads in the binary classifier could bias the local sea-surface reference and subsequently affect the freeboard estimates. We will revise the manuscript to make the

role of the binary classifier clearer.

Minor comments:

Figure 2: Please show the RF classifier more explicitly in the workflow after the GMM clustering and manual labeling steps.

This has been addressed together with the related comment above on Fig. 2.

Figure 4 caption: (d) height standard deviation (STD) à (e) height standard deviation.

Will be corrected.

Please clarify whether “thin ice” and “gray ice” refer to the same class and use the terminology consistently throughout the manuscript.

“Thin ice” and “gray ice” refer to the same class in our manuscript. To avoid confusion, we will use “thin ice” consistently throughout the revised manuscript.

Response to Referee #3

In this document, we outline our responses to comments from the referee, including modifications that we intend to make to the manuscript where necessary.

Referee comments are shown in black and our responses in blue with intended changes to the revised version of the manuscript given in italics.

Review report on "Year-Round High-Resolution Sea Ice Freeboard Retrieval Using ICESat-2 ATL03 Photon Data" by Liu et al.,

Sea ice freeboard is a critical parameter not only a proxy for snow depth estimation, but also critical for sea ice thickness extraction. This manuscript introduces the High-Resolution Freeboard Method (HRFM), a novel framework that retrieves sea ice freeboard directly from ICESat-2 ATL03 photon data at 5 m along-track resolution. The approach combines a two-stage denoising strategy (Kernel Density Estimation with adaptive histogram-based filtering) and a machine-learning surface-type classifier trained on 25 coincident Sentinel-2 scenes covering winter and summer conditions across diverse ice types. The authors validate surface heights against ATM (Airborne Topographic Mapper) data, compare their freeboard estimates with the operational ATL20 product, and demonstrate improved performance in preserving ridge heights, mitigating after-pulse effects, handling melt-season complexities, and enabling consistent weak-beam utilisation.

The study is highly relevant to the scope of TC Journal and timely to the cryosphere society, addressing a critical need for year-round, high-resolution sea ice thickness monitoring. The manuscript is well-structured, the methodology is clearly described, and the results represent a substantial advance over existing ICESat2 sea ice products (ATL07/ATL20), which convinced me that this study warrants publication in TC Journal. However, before final acceptance. There are numerical issues that either need further clarification or significant improvement. Please see my major and minor comments below that I hope the authors can consider during the revision.

We sincerely thank the reviewer for the positive evaluation of our manuscript. We're grateful to the reviewer for taking the time to check our manuscript, providing valuable comments that improve the clarity of the work.

Major Comments / Required Revisions

1. The authors should specify explicitly what annual cycle was investigated either in the title or in the abstract. Why was this annual cycle selected in this study?

The primary objective of this manuscript is to develop and evaluate the HRFM method, rather than to investigate the interannual variability of sea-ice freeboard in a specific

year. The annual cycle used in this study was therefore selected as a demonstration period for algorithm development and year-round applicability assessment, not for any specific climatological reason.

We do not think it is necessary to emphasize the specific year in the title, because the focus of the manuscript is the retrieval method rather than the characteristics of a particular annual cycle.

We are currently extending HRFM processing to all available ICESat-2 years to generate a comprehensive multi-year high-resolution sea-ice freeboard dataset. Such a multi-year product will allow us to identify and interpret the importance of specific years based on the data, which is beyond the scope of the present method-development manuscript.

2. The key algorithm parameters are mentioned throughout the text (e.g., horizontal scaling a=25/100, histogram bin size 0.1 m, after-pulse offsets 0.45/0.9 m, 10-km window for sea-level reference, 3σ outlier removal, 10th percentile density threshold for refined denoising), a single table summarising all critical parameters with their values and justifications would greatly enhance reproducibility. Please add such a table (e.g., in Section 3 or as an appendix) in the revised manuscript.

We thank the reviewer for this valuable suggestion. We will add a new table (as shown below) in the revised manuscript as an appendix to summarize the key HRFM algorithm parameters and their values. For parameters that differ between strong and weak beams, the values will be reported in the form strong beam / weak beam.

3. In the abstract and results (Sect. 4.1), the authors state that weak-beam retrievals achieve "comparable accuracy" to strong beams. How "comparable" it was? Please provide a concrete assessment. Please consider adding a statistical test to show whether the remaining difference is meaningful or not.

In general, weak beams are expected to perform less favorably than strong beams because of their lower transmitted energy and consequently lower photon return rates. In the original manuscript, this statement that weak-beam retrievals achieve comparable accuracy was intended to indicate that the weak-beam results show similar performance to the strong-beam results when evaluated against independent reference data, including both the surface-height comparison with ATM measurements and the cross-validation of surface-type classification.

We will remove the phrase "comparable accuracy" because it may imply a formal equivalence test that was not intended in the original manuscript. Instead, we will state that weak-beam retrievals showed similar validation performance to strong-beam retrievals when evaluated against the same independent data.

The following will be added in the abstract:

Weak-beam retrievals also show encouraging validation performance under the same

evaluation framework

4. Table 2 shows thin-ice recall is only 0.50 (strong beam) and 0.43 (weak beam), with precision ~0.65. The authors merge thin ice into sea ice for two-class validation (Table A1, overall accuracy >0.98). However, thin ice is a distinct surface type with different radiative, thermodynamic, and mechanical properties. Please discuss:

a: How often does misclassified thin ice affect freeboard retrieval (e.g., when thin ice is wrongly labelled as lead and thus included in the reference sea level)?

b: Should users treat the thin-ice class as "experimental" or apply the two-class version for freeboard estimation? A clear recommendation would be helpful.

In this study, the three-class experiment was mainly designed to examine whether ICESat-2 ATL03 photon-based features have the potential to distinguish thin ice from leads and snow-covered sea ice. The results in Table 2 indicate that such potential exists, but they also show that the thin-ice class is still difficult to identify robustly. Thin ice can have photon-rate, density, height-variability, and background characteristics that overlap with both leads and snow-covered sea ice, which explains the relatively low recall and moderate precision reported in Table 2.

Because of this limitation, the operational HRFM freeboard retrieval in this manuscript does not rely on the three-class classifier. Instead, thin ice is merged into the sea-ice category, and a two-class lead/sea-ice classifier is used for freeboard estimation. This choice is made specifically to reduce the risk that thin-ice surfaces are incorrectly treated as leads and then used in the reconstruction of the local reference sea level. Therefore, the low thin-ice recall in the three-class experiment does not directly propagate into the freeboard estimates presented in this study.

Regarding point (a), if the three-class classifier were used directly for Arctic-wide freeboard retrieval, misclassified thin ice could affect the reference sea level when thin ice is incorrectly labelled as lead. Since thin ice generally has a positive freeboard relative to the local sea surface, including such segments in the lead samples would tend to raise the estimated reference sea level and could consequently lead to an underestimation of surrounding sea-ice freeboard. However, for large-scale retrievals it is difficult to quantify this effect rigorously because the true lead/thin-ice identity is not known outside the labelled validation scenes.

Regarding point (b), we will provide a clear recommendation in the revised manuscript. For sea ice freeboard retrieval, we recommend using the two-class lead/sea-ice classifier.

5. The reference sea level is computed by aggregating lead segments within a 10-km along-track window. The choice of 10 km is plausible but not justified. Could the optimal window size vary with ice regime (e.g., compact vs. marginal ice zone, winter vs. summer)? A brief sensitivity test (e.g., 5 km, 10 km, 20 km) on a representative

track would strengthen the method. If such a test already exists, please cite or summarizeit.

We adopted a 10 km window primarily to be consistent with the ICESat-2 ATL07/ATL10 sea ice freeboard algorithm (Kwok et al., 2026) and with the physical rationale discussed by Kwok et al. (2019). The 10-km length is selected to minimize the contribution of residual sea surface slopes to uncertainties in freeboard calculations; the general expectation is that the local sea surface height is relatively constant over a length scale corresponding to the first mode baroclinic Rossby radius of deformation, which is on the order of 10 km for polar latitudes above 60° latitude.

6. Figure 10 shows seasonal mean differences between HRFM and ATL20 up to ±0.04 m, with sign reversals between winter and summer. The authors attributed these to differences in surface-height retrieval, lead detection, and reference sea-level construction. A rough quantitative decomposition (e.g., how much of the winter difference is due to after-pulse removal vs. lead sampling density) would greatly strengthen the discussion. Even an approximate breakdown based on the examples in Figs. 8 and 12 would be valuable.

We agree that a quantitative decomposition of the seasonal differences between HRFM and ATL20 would strengthen the interpretation of Fig. 10. However, a rigorous component-by-component decomposition is challenging because sea ice freeboard retrieval is a sequential and highly coupled procedure. Differences introduced at the initial signal-photon identification stage can propagate into the segment height estimates and also affect the derived features used for surface-type classification. The estimated surface heights further influence lead identification and the reconstruction of the local sea-surface reference. Therefore, the effects of after-pulse removal, surface height retrieval, lead detection, and sea-level reference construction are not fully independent and cannot be simply separated in an additive manner.

In addition, HRFM and ATL07/ATL20 use substantially different processing strategies, and ATL20 does not provide all intermediate variables required for a strict attribution analysis, such as the photon-level filtering results, intermediate lead-selection decisions, and the exact local sea-surface reference construction at each processing step. This further limits the possibility of performing a fully or partly quantitative decomposition of the HRFM–ATL20 differences.

Nevertheless, we agree that even an approximate analysis would be useful. In the revised manuscript, we will attempt to provide a rough quantitative assessment based on representative cases. If the resulting estimates are physically meaningful and robust, we will include them in the revised discussion. Where a strict decomposition is not possible, we will explicitly state this limitation.

7. The classifier was trained on scenes south of 82°N because of Sentinel-2's orbital limit. How confident are the authors that the classifier works north of 82°N (e.g., the

central Arctic) where no coincident optical imagery is available for validation? Please add a discussion of potential extrapolation risks and whether the physical feature pace (photon rate, density, height STD, background rate) is expected to remain valid poleward.

Although direct Sentinel-2-based validation is not possible north of 82°N, the monthly mean freeboard fields retrieved by HRFM show spatial patterns that are broadly consistent with the ATL20 products over the entire Arctic, including regions poleward of 82°N. This agreement provides indirect evidence that the classifier and the subsequent freeboard retrieval framework remain effective in the central Arctic. Based on the current experiments, background rate is the feature most likely to become less reliable under low solar elevation conditions. We have described the limitation in Section 5.3 of original manuscript.

Minor comments

Except for the points that are addressed separately below, we will revise the manuscript following the reviewer's suggestions.

-Line 70: "detector after-pulses" should be introduced earlier (it is defined in Sect. 3.1.2 but referenced here). Consider adding a brief definition at first mention.

Thank you for the suggestion, the following will be added in the Line 70:

First, ATL07 signal extraction can be affected by detector after-pulses, i.e., delayed artificial photon events that occur after strong surface returns and appear at predictable height offsets below the primary signal. These after-pulses are particularly relevant during saturation events over specular leads and can occasionally introduce biases in surface-height retrieval

-Line 125: Use consistent version notation (e.g., "Version 6" rather than "V06", or define "V06" at first use).

Will be corrected to " Version 6".

-Figure 8 caption: "The green dashed lines represent the ATL07 signal selection envelope" : please describe this explicitly in the caption.

The following will be added in the caption:

The green dashed lines represent the ATL07 signal selection envelope.

-Section 5.1, line 585: "lower transmitted energy level (~80% of outer beams)" – please verify and provide a citation for the energy difference between middle and outer beams.

This is reported by Kwok et al. (2019): The photon rate distribution for the weak beams (GT1L, GT2L, GT3L—dashed lines) are consistently at ~1.5 photons/shot. The strong beam photon rate distributions (in GT1R, GT2R, GT3R—solid lines), however,

indicate that GT2R (Beam 3) has consistently weaker surface returns (~0.81 of GT1R and GT3R): The mean rates from GT1R and GT3R are ~6.8 while those from GT2R are ~5.5. This is attributed to the expected variations in the custom construction of the optical component used to split the laser energy into the six beams. These values are consistent with pre-launch expectations.

-Figure 5 caption: “It’s a 10-km profile” should read “It is a 10-km profile”

Will be corrected.

-Language: The manuscript is generally well written, yet there are a few grammatical issues that remain (e.g., “It’s” in figure captions, occasional missing articles). I see you have English co-authors, and please ask them to proofread the language.

We will carefully proofread the revised manuscript to correct the remaining grammatical issues.

- The text fonts in almost all figures are quite small. Please consider enlarging them to improve the readability of the figures

We will redraw the figures with small text labels using larger font sizes to improve readability in the revised manuscript.

- The authors provided this data availability statement: The Sentinel-2 imagery was derived from Google Earth Engine at https://developers.google.com/earth-engine/datasets/catalog/sentinel-2. I am not sure if this is adequately enough by TC, or should authors provide the Sentinel-2 imagery data somewhere for readers to explore. I leave this for the TC handling editor to decide.

We will improve the data availability by providing a detailed list of the Sentinel-2 scenes and the corresponding ICESat-2 tracks used for classifier training and validation, as shown in Table 1 in our response to Referee #2.

References

Kwok, R., Markus, T., Kurtz, N. T., Petty, A. A., Neumann, T. A., Farrell, S. L., Cunningham, G. F., Hancock, D. W., Ivanoff, A., and Wimert, J. T.: Surface Height and Sea Ice Freeboard of the Arctic Ocean From ICESat-2: Characteristics and Early Results, JGR Oceans, 124, 6942–6959, https://doi.org/10.1029/2019JC015486, 2019.

Kwok, R., Kurtz, N., Wimert, J., Petty, A., Cunningham, G., Markus, T., Hancock, D., Ivanoff, A., Bagnardi, M., Herzfeld, U., Trantow, T., and ICESat-2 Science Team: Ice, Cloud, and Land Elevation Satellite (ICESat-2) Project Algorithm Theoretical Basis Document (ATBD) for Sea Ice Products, version 7, https://doi.org/10.5067/KPMXUOH7TNIY, 2026.