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.

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, we will add a more detailed description in Sect. 2.1.2 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.

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.

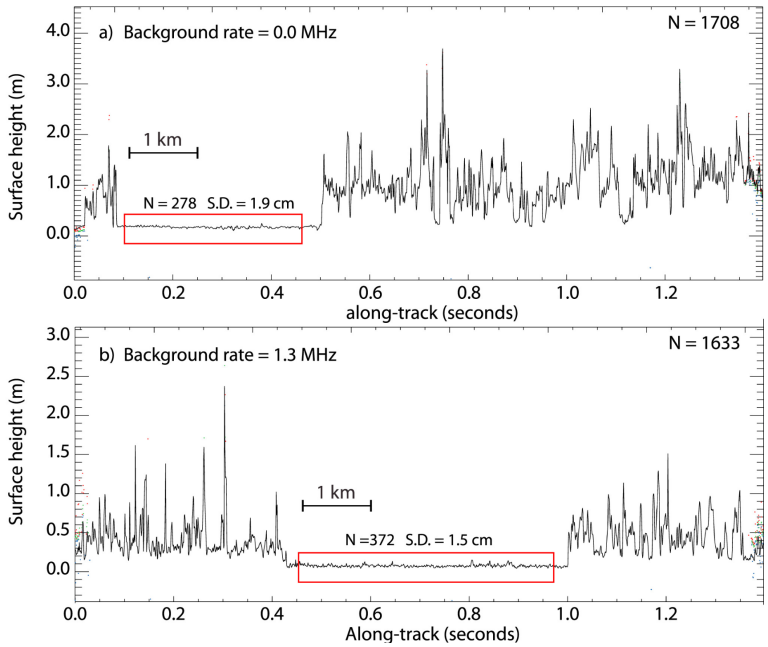

Fig R1 Kilometer-scale height precision over flat leads reported for ATL07 (Kwok et al., 2019).

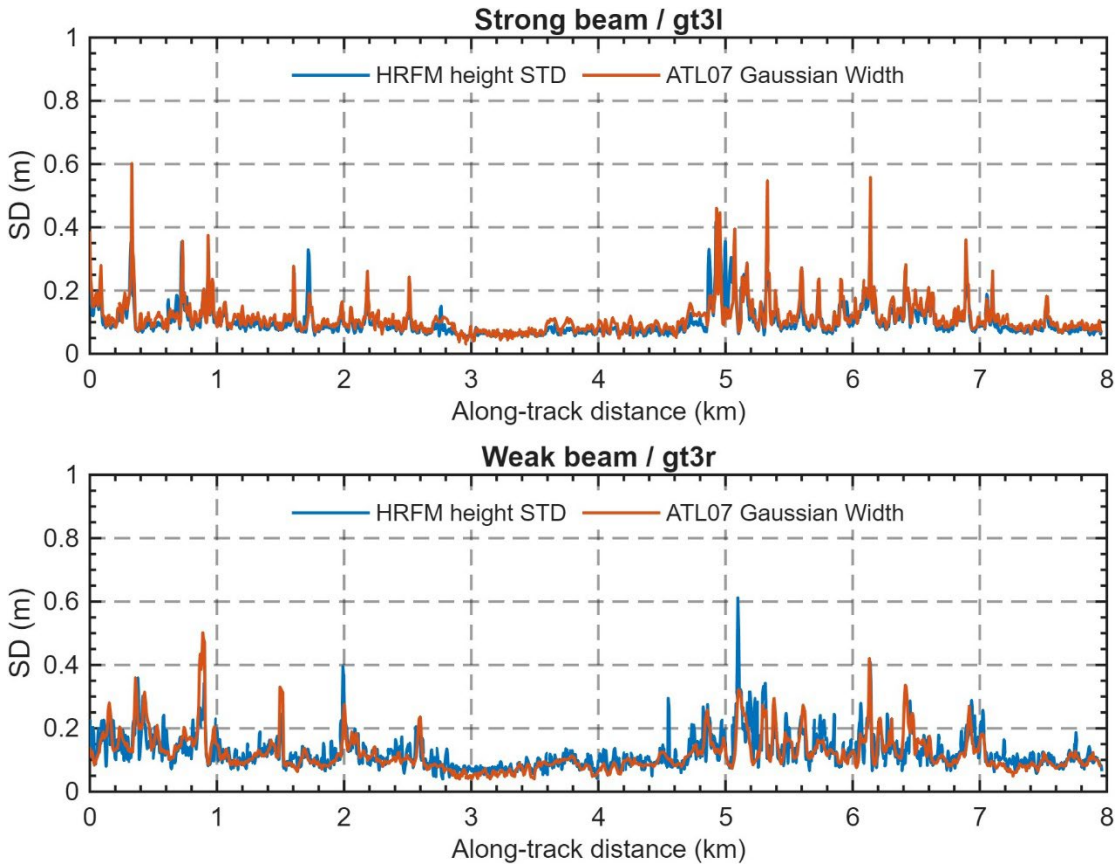

Fig R2 Comparison between HRFM within-segment height STD and ATL07 Gaussian width.

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.

In our preliminary tests, the 8 April and 22 April 2019 ATM flights showed better spatial overlap with ICESat-2, whereas the 12 April case showed poorer overlap. For this reason, the original manuscript focused on the 8 April and 22 April cases. Following the referee's suggestion, we will add the 8 April and 12 April flights as an additional check. The 12 April result shows that, under poorer overlap and without cross-correlation alignment, the comparison statistics of both HRFM and ATL07 degrade (Fig R3). Nevertheless, HRFM remains broadly consistent with ATL07 under the same comparison framework. This supports the interpretation that part of the discrepancy relative to Kwok et al. (2019) arises from differences in collocation and alignment procedures rather than from the height retrieval alone.

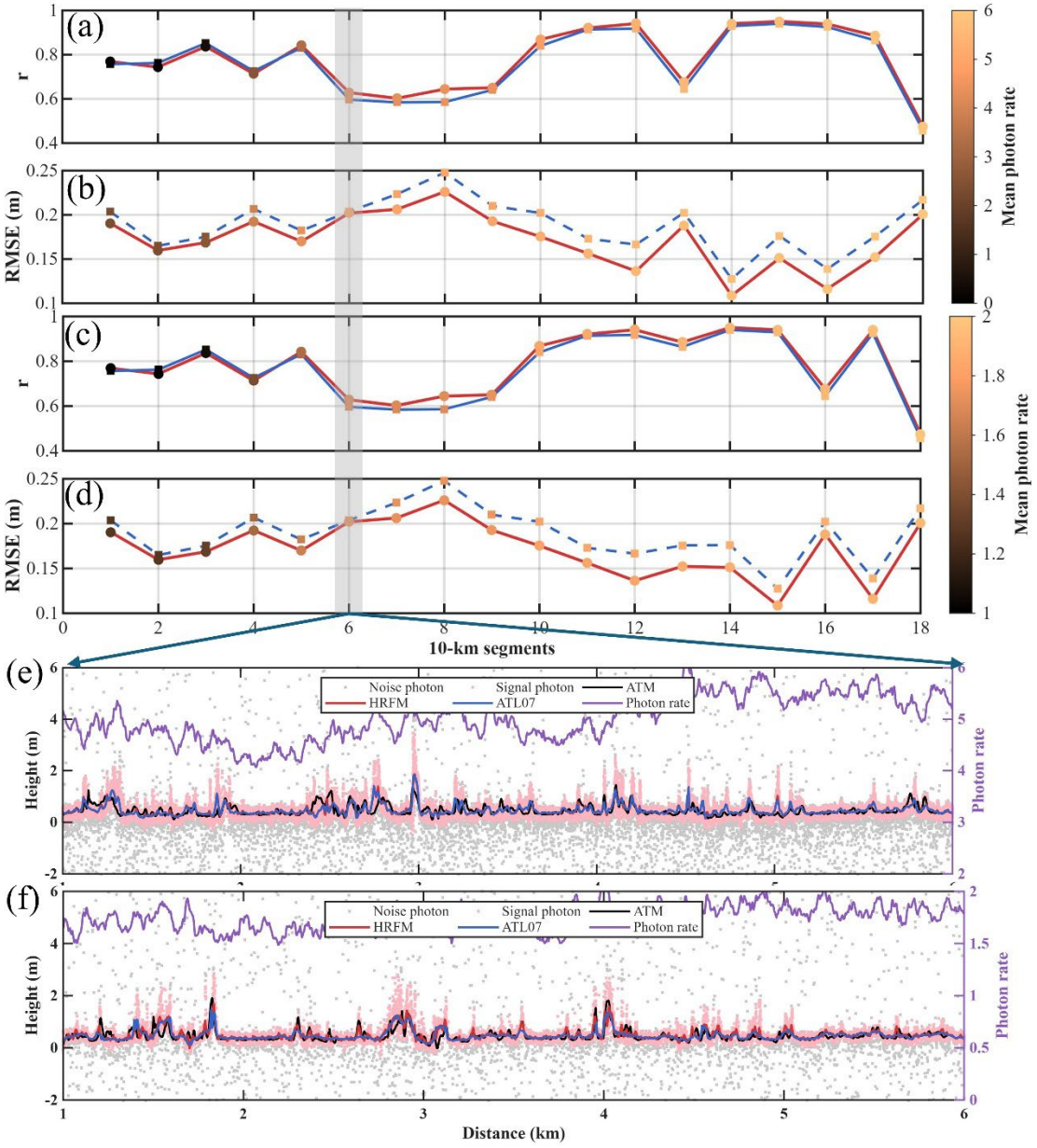

Fig R3 Evaluation of HRFM and ATL07 surface height retrievals against ATM heights obtained on 12 April 2019. (a) and (b) Correlation coefficient (r) and root-mean-square error (RMSE) between HRFM (red) and ATL07 (blue) against ATM for strong beam (gt2l). (c) and (d) The same statistics for weak beam (gt2r). The color scale indicates the mean photon rate for each segment. (e) and (f) Surface height profiles of segment 6 for strong and weak beam, 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).

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 8 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.

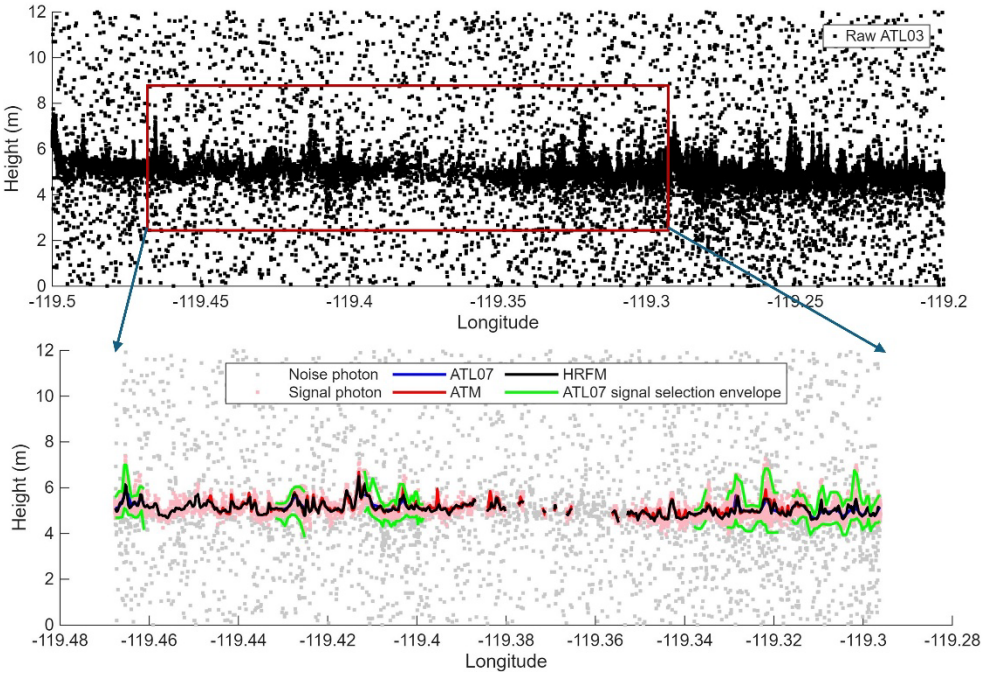

Fig R4 Performance of HRFM photon filtering under cloudy conditions.

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).

We will add the availability and purpose of YAPC in the revised manuscript, and will discuss YAPC as a useful alternative signal filter variable for future HRFM versions.

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) for each beam and did the same thing. We will add a direct strong-minus-weak beam difference plot, together with a histogram of the beam-dependent freeboard differences, to make the comparison clearer.

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.

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.

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 R 5. We will revise the text related to the figure.

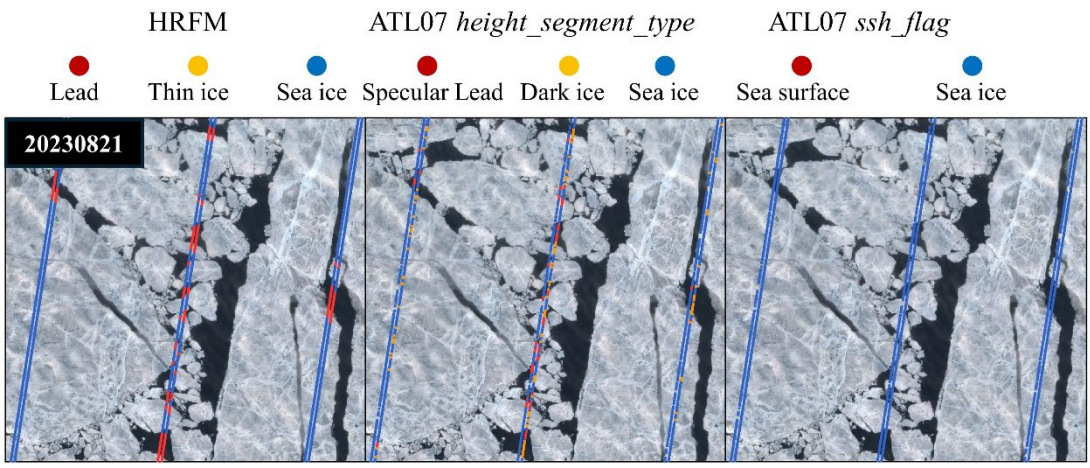

Fig R 5 Corrected plot of Fig. 9c

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.

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.

L126: Please clarify the version number.

We will clarify the product version number in the revised manuscript.

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.

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 will be added 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 in each processing step.

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.

It will be made clearer.

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.

We will try to identify a coincident Sentinel-2 scene and, if available, use it to compare the classification.

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.