We would like to thank the referee Tom Armitage for his timely review as well as for his the time and effort he spent on it. In the following, we would like to respond to all comments made by the referee and our reasoning behind all our corresponding changes.

General Comments/Suggestions:

1) Sections 2.3.2 and 2.3.3: I am not familiar with these statistical techniques, and I don't think the wider polar altimetry community will be either, unless there have been other publications on these techniques within the context of polar altimetry? There is a lot of jargon (specialist terms e.g., "k-means", "clustering", "decision trees", "trained forests", etc.) and sentences that are nonsense without specialist knowledge, for example, how does a "decision tree" in a "trained forest" cast a "vote"?! I am imagining woodland creatures and some kind of election! While I appreciate that I could follow the references you have provided to textbooks etc, I think the manuscript would be greatly improved if you could provide a more intuitive explanation of these procedures, as well as some relevant equations and figures if applicable, including specific examples of how this applies to altimeter waveform classification. In particular, I think this is important because this seems to be one of the major developments in the paper, so I think readers should be able to gain an intuitive picture of what is happening in your processing in order that they might develop/reproduce what you have done. The schematic in Figure 1 is sort of useful but there is still a lot of jargon, and it's not clear to me what is happening at each stage.

We agree with reviewer that this is an essential part of the manuscript and of course, we would like to have this section as intuitive for potential readers as possible. We agree that relatively new, and complex statistical approaches such as the use random forests could most likely be explained in even more detail as it already is (We think one has to appreciate the effort we already put into this instead of a simple mention + reference). However, an additional thoroughly explanation of widely and commonly accepted and proven techniques such as a k-means clustering or simple decision trees might be out of the scope of this manuscript, as it is not intended to give a broad statistical methods review. For the further interested reader we provided all necessary references.

Nevertheless, we made several changes in the mentioned sections and added additional information as well as corrected the mentioned rather figuratively paragraph about the giant democratic group trees.

In 2.3.2:

"Next, a subset of 1% is sampled at random without replacement (i.e., each original waveform with corresponding surface backscatter, pulse peakiness, and leading-edge width can only appear once) for each month in the MOP and for each sensor independently. This data sample is then separated into three clusters using unsupervised methodology named k-means clustering (MacQueen, 1967; Hartigan and Wong, 1979). This unsupervised method (i.e., without any a-priori information about the data) is widely used to separate input data of N observations into K clusters of equal variance. In our case, based on the input classifier parameters of surface backscatter, pulse peakiness, and leading-edge width, whereby the within-cluster sum-of-squares are iteratively minimized (MacQueen, 1967; Hartigan and Wong, 1979). The result is a 'labeled' data set where each input waveform with corresponding surface backscatter, pulse peakiness, and leading-edge width is labeled as an either sea-ice-type, lead-type, or ambiguous-type waveform.

Generally, the preselection of the number of clusters can be a problem when utilizing k-means clustering. However, while we tested a higher number of initial clusters with perspective of later reunion of similar clusters, a separation into just three clusters turned out to be sufficient. Overall, lead waveforms account for a smaller fraction of the total measurements than sea-ice waveforms. Because of this and the fact that k-means clustering tends towards generating equal-size clusters (this is a presumption of k-means clustering algorithms), sole use of k-means clustering for the complete data set was not feasible."

In 2.3.3:

"Random forests are based on multiple decision trees. A decision tree is a rather simple statistical tool to predict data categories based thresholds. Over several steps, the input data set is split at each step (called a 'node') based on a threshold of a given parameter until all input data is categorized. When visualized, a decision tree resembles a tree with an increasing numbers of branches, leading to the final categories (Breiman, 2001)."

As well as some smaller changes to the whole sub-section to increase clarity about the random-forest procedure.

2) How are your results affected by just using a fixed retracking threshold for Envisat lead waveforms, rather than including lead and floe waveforms in the tuning/fitting procedure? The reasoning for retracking near (or at) the maximum power for lead waveforms is equally valid for CS-2 and Envisat, i.e., specular scattering from leads reduces the effective footprint to the size of the lead, which in turn gives you a return which is close to the transmitted pulse (convolution with a delta function rather than the flat-surface impulse response). Using a single retracker for all CS2 waveforms and essentiallytwo separate retrackers for Envisat leads and floes represents an inconsistency with your approach that I don't feel is justified.

While physically not correct, the rather empirical solution to use a 50% threshold for the retracking of leads and sea ice for CryoSat-2 using the TFMRA retracker results in the overall best and most plausible results (also compared to validation data, e.g., from EM measurements). While we agree that this is not in any case physically justified, changing our 'reference' without a proper additional validation does not seem justified either.

As the reviewer explains, retracking near the maximum power is the most logical choice for leads, which is also, why we did not aim for an adaptive procedure for Envisat. Additionally, as seen in the paper from Guerreiro et al (2017), using a lead threshold of 50% for leads from Envisat results in unrealistically negative freeboard estimates.

Moreover, using different thresholds for the same retracker algorithm is in our understanding something different than using two completely different retracker algorithms as it was done during SICCI-1 and is a definite improvement by providing a consistent retracking methodology.

3) Related to this, I appreciate that your fitting procedure is essentially try to match the Envisat freeboard to the CS2 freeboard by tuning the Envisat retracking threshold. But couldn't you simply skip a stage here and fit the Envisat waveform parameters (LEW/sigma-0) to the CS2 freeboard directly?

While we appreciate the reviewers' suggestion, we do not agree that the proposed procedure would feature the same intuition. Fitting waveform parameter directly to the freeboard would ignore the fact that sea ice with the same surface features can have different freeboards. Retracking preserves the time delay information, which is not included in the shape parameters but in the position of the waveform in the range window.

4) Section 3.1: You get a better match with the surface type classification than in SICCI-1, but isn't this by construction? Haven't you tried to match the number of waveforms classified as leads and floes for the two satellites, or have I misunderstood something? Further, I wouldn't necessarily expect there to be agreement between the number of leads/floes detected by the two instruments, simply because of the different footprints – from physical/geometric arguments I would expect far more 'ambiguous' waveforms in the Envisat data. More encouraging is the broad spatial agreement between the lead/floe distributions.

As mentioned by the reviewer, we indeed see much more ambiguous-type waveforms in the Envisat data and that especially in the Antarctic. However, we also achieve a much better overall spatial agreement of occurrences. While we of course intended to mirror CryoSat-2 patterns of lead-/and sea-ice-occurrences with Envisat, the achieved results are not 'constructed' in a comparable way, as the adaptive retracker threshold procedure. As it is mentioned in the manuscript, all classification is always done for each sensor separately, i.e., while the used procedure is consistent, the methodology is applied for each sensor independently.

5) Section 3.2: Similarly, isn't the small observed difference in freeboard by construction? Here, I think the manuscript would be greatly improved by comparison of the two satellites with independent radar freeboard measurements, e.g., by combining the IceBridge laser and snow radar.

Here, the reviewer is correct, as this agreement is indeed through the applied tuning mechanisms. However, no procedure is perfect which is why of course we need to stress the overall very good agreement that our methodology can achieve (but also its limitations e.g. in the Antarctic or in the inter-seasonal variability). The procedure still relies on the different waveform parameters to decide on the best threshold to retrack the freeboard height and there are areas and times where it works better or worse (which we also highlight in the manuscript).

Specific Comments:

The title is clunky, I would suggest "Consistent retrievals of Arctic and Antarctic sea ice freeboard from Envisat and CryoSat-2"

We would like to thank the reviewer for the hint. While we did not follow the exact suggestion, we agree that the title might not have been put together elegantly. Furthermore, from the comments we received from all of the three reviewers, we came to the decision that the title was not chosen specifically enough for the purpose of this manuscript. We changed the title to read:

"Empirical Parametrization of Envisat Freeboard Retrieval of Arctic and Antarctic Sea Ice Based on CryoSat-2: Progress in the ESA Climate Change Initiative"

Page 1, Line 2: I suggest "…estimation over recent years, however, precursor…"

We would like to thank the reviewer for his suggestion and changed that accordingly.

Page 1, Line 13-15: “cover” should be “extent”. Also, join the sentences “…Meier et al), while Antarctic sea ice extent…”

We changed that.

Page 1, Line 15-16: I suggest the following “…(Turner et al). Arctic sea ice is also thinning, as observed by…”

We changed that.

Page 1, Line 24: I suggest “…that measurement of sea ice thickness at circumpolar scales in both polar regions…”

We changed that.

Page 2, Line 4: really this type of processing dates back to at least Laxon (1994), “Sea ice processing scheme at the EODC”.

We would like to thank the reviewer for pointing this out to us. We added the provided reference.

Page 2, Line 5: I’m not really familiar with the term “run-time” within the context of altimetry, could you explain or use a more familiar term.

In accordance with a similar remark from reviewer 2/3, we changed that sentence to read:

“In a first step, the echo power waveforms are classified as returns from either sea-ice floes or returns from the sea surface of leads between sea-ice floes.”

Page 2, Line 6-7: “so accurate that…”, this isn’t really the case for individual lead/sea ice measurements due to speckle noise. Also, explain explicitly that this elevation difference is termed the freeboard, otherwise the next sentence might not make sense to people unfamiliar with this term.

Reviewer 2/3 also questioned this part so we decided to change it to read:

“These measurements are then converted into distance measurements that let one calculate the elevation difference of the snow surface or the sea-ice surface relative to the sea surface in the leads.”

Page 2, Line 9-11: “not true” – I’m aware of all the studies on this (including some of my own work!), but I would still argue that such a strong statement on this issue is still not possible. I would suggest “When estimating sea ice thickness from radar altimeters, it is often assumed that…”, and you should provide a more extensive list of studies that might suggest otherwise.

We rephrased this part slightly following the reviewers suggestion:

“When estimating sea-ice freeboard from radar altimeters, it is often assumed that the retrieved distance over sea ice using Ku-band radar always coincides with the snow/ice interface. However, this

assumption is not true, especially for a highly stratified sea-ice snow cover and/or for multi-year sea-ice regimes."

However, we still think the statement made is valid as a formulation with "not always" clearly implies that there are of course cases where the returned distance actually coincides with the snow/ice interface. But that is clearly not always the case.

Page 2, Line 18: I believe the Envisat altimeter was Radar Altimeter 2 (RA-2)?

This is correct. We changed that.

Page 4, Line 32: Is this filtering important? How many waveforms are removed?

This filtering step remains from the processing done in during SICCI-1. While the number of rejected data values is potentially small, flag names suggest that it is better to have them removed nonetheless.

Page 4, Line 33-Page 5, Line 1: Do you also apply a land mask filter? Are the inbuilt land surface type flags accurate enough to catch all land contaminated waveforms?

We currently rely on the built-in surface-type flags in the CryoSat-2 as well as the Envisat product and use as stated in the manuscript all waveforms flagged as 'Ocean'. No further masking is applied in that matter.

Page 5, Line 4-8: I think what you are saying is that distinguishing leads is essential in order to estimate the instantaneous sea level anomaly along track?

In accordance with a comment from reviewer 3 we changed that paragraph to read:

"The surface-type classification is a crucial part in the processing chain, because the detection of leads is essential for determining the instantaneous sea-surface height anomaly with respect to the mean sea-surface height at the ice-floe location. The resulting sea-surface height at the ice-floe location in turn is used as the reference from which the sea-ice freeboard is calculated."

Page 5, Line 9-14: Before this paragraph you should explain why it is possible to distinguish leads and floes to begin with i.e., explain the different surface scattering characteristics. Otherwise, this paragraph is not clear.

We agree with the reviewer and changed that first part to read:

"In general, leads feature a specular reflection due to their rather smooth surface, whereas sea ice features a diffuse reflection due to a higher surface roughness. With smaller instrument footprint sizes, less surface-type mixing occurs and the return signal is easier to classify. However, leads often dominate acquired waveforms due to their specular reflection. Off-nadir leads still represent sources of strong backscatter and therefore result in false range estimates."

Page 5, Line 12: “footprint of 2km”, this is the pulse-limited footprint. “increase up to 10km (Chelton et al.)”, I don’t think Chelton was talking about off-nadir ranging to leads, he was talking about the effect of significant wave height on the pulse limited footprint, which is fundamentally different i.e., strong off-nadir backscatter in the case of leads vs. large surface roughness in the case of high SWH.

We removed the latter part.

Page 5, line 18: “sea ice backscatter”, you mean σˆ0?

Yes. Based on a comment of reviewer 3, we changed ‘sea-ice backscatter’ into ‘surface backscatter’ throughout the manuscript to avoid confusion as the parameter is also used to differentiate between leads and sea ice.

Page 5, Line 29: young, thin ice areas, cause specular reflections, you should add a citation.

We added the following reference:

Zygmuntowska, M., Khvorostovsky, K., Helm, V., and Sandven, S.: Waveform classification of airborne synthetic aperture radar altimeter over Arctic sea ice, The Cryosphere, 7, 1315-1324, https://doi.org/10.5194/tc-7-1315-2013, 2013.

Page 5, Line 10-31: surely the rejection rate could be decreased as well?

It surely can, this is in our cased achieved by the proposed way of using monthly thresholds. Estimating thresholds based on a single month however leads to misclassifications/rejections in other months, whereas using all data together potentially also results in rather soft thresholds. This however might allow rather ambiguous signals to be taken into the freeboard retrieval.

Page 9, Line 24-Page 10, Line 3: This is all rather unclear to me.

In response to some suggestions made by reviewers 2 and 3 we made changes to this paragraph. This should be clearer now.

Page 10, Line 1; Figures 2-5: Show the pulse peakiness maps as well, also show the multi-year ice mask for comparison.

We added the pulse peakiness as well as the resulting Envisat freeboard estimates after application of our adaptive retracker threshold procedure to the Figures. In panel a), showing the freeboard difference between Envisat and Cryosat, we also added the 50% MYI fraction threshold line.

Page 11, Line 2: Why disregard PP in the fitting procedure? Presumably you tried different iterations but discarding PP gave you the best result?

We chose sig0 over pp for its more direct physical relation to surface roughness. However, both measures are correlated quite well. Nevertheless, there are differences between pp and sig0 as shown in the new Figures 2-5. In order to keep things as simple as possible and in order to rely on as few parameters as possible, we decided to use the leading edge width and sig0. As described in the manuscript, the resulting fits are very good based on the given measures.

Page 12, Line 3-11: The fitting procedure is not clear to me at all. What are you fitting to what? What is the "x-y" plane? Perhaps a diagram/Figure could help to illustrate this?

We agree and rephrased the first paragraph to read:

"In the next step, we derive a functional relationship between optimal-threshold values and the waveform parameters of surface backscatter/leading-edge width for our adaptive threshold range-retracking. Therefore, we first average all optimal-threshold values during the mission-overlap period (MOP) for bins of 0.25 dB for the surface backscatter and 0.025 for the leading-edge width, respectively. Here, we use a three dimensional coordinate system with average optimal threshold (z-axis) against leading-edge width (x-axis) and surface backscatter (y-axis)."

We add here the following sketch for further illustration, where the x-axis represents the leading-edge width, y-axis the surface backscatter and z-axis the optimal Envisat retracker threshold.

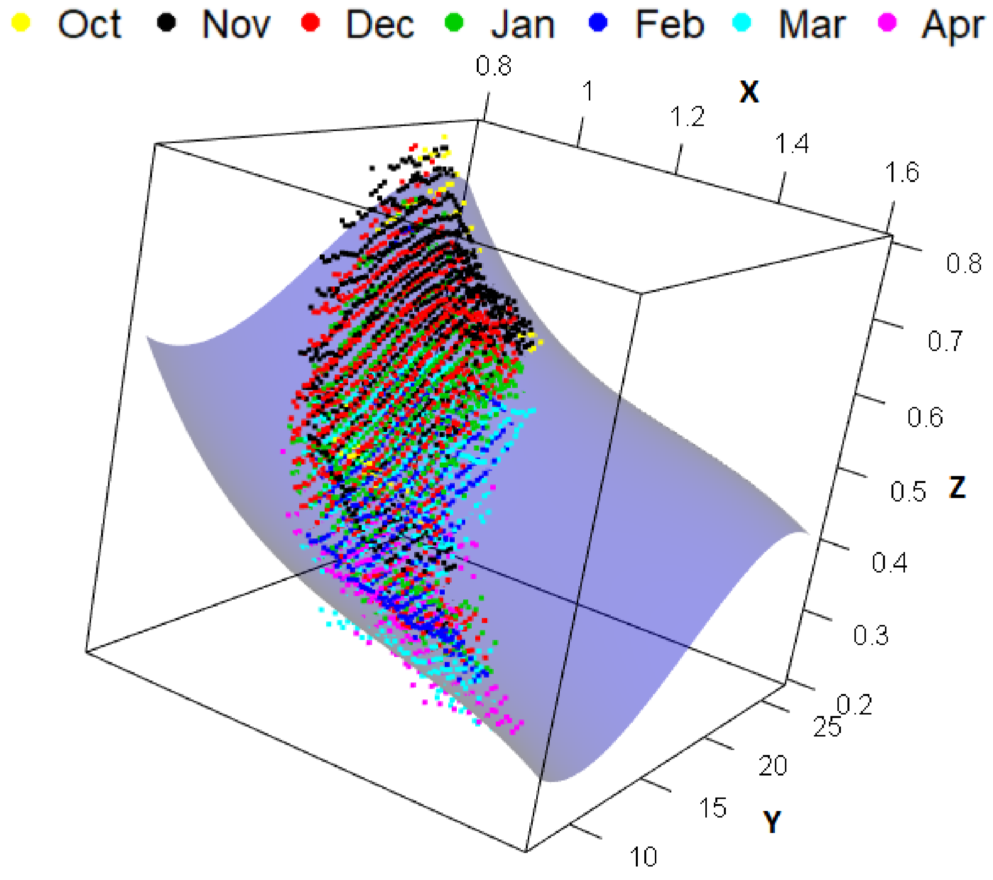

Figure 10: Could you change the x-axis scale to ~ -25cm to ~ 40cm to make the figure clearer

We changed that for Figures 10 and 11.

Page 16, Line 26-34: Could the difference with Guerreiro be explained by the speed of light propagation correction, or do you both apply it in the same way?

In contrast to Guerreiro et al. who use a correction proposed by Kwok & Cunningham 2015 depending on snow depth and density, we apply a speed of light reduction in the snow pack by using a fixed factor of 0.22. However, the resulting difference should be very small and not be able to explain the differences on its own.