We would like to thank referee Nathan Kurtz for his valuable comments and suggestions. We appreciate the feedback by the referee and would like to go through it point by point and highlight our changes accordingly.

General Comments/Suggestion:

In particular I think there are flaws in the CryoSat-2 retracking procedure which need to be addressed prior to publication, these are noted below. Particularly I note the need to better refine the lead tracking procedure and verify this through direct elevation comparisons between the measurements. Independent validation data of the CryoSat-2 data set through comparison with field campaigns such as CryoVEx or IceBridge are also needed. Additionally, only freeboard differences are plotted so it is difficult to evaluate whether the retrieved freeboards themselves are accurate. Some maps and statistics of the actual retrieved freeboard are needed. This is especially important for the Antarctic region where due to the complexity of the surface prior studies with satellite radar altimeters have not demonstrated the capacity for obtaining accurate measurements.

We appreciate the reviewers suggestions and comments. Especially in the case of not shown actual freeboard results, the manuscript had a clear lack of information. We added the resulting Envisat freeboard as well as the gridded Envisat pulse peakiness (suggested by another reviewer) to Figures 2-5.

Referring to the reviewers concern about the lead retracking in the Cryosat product, we would like to mention previous studies that compared the Cryosat product to in-situ data as well as other products (e.g. UCL/CPOM):

Ricker, R., Hendricks, S., Kaleschke, L., Tian-Kunze, X., King, J., and Haas, C.: A weekly Arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data, The Cryosphere, 11, 1607-1623, https://doi.org/10.5194/tc-11-1607-2017, 2017.

Haas, C., Beckers, J., King, J., Silis, A., Stroeve, J., Wilkinson, J., Notenboom, B., Schweiger, A., & Hendricks, S. (2017). Ice and snow thickness variability and change in the high Arctic Ocean observed by in situ measurements. Geophysical Research Letters, 44, 10,462–10,469. https://doi.org/10.1002/2017GL075434

Ricker, R., Hendricks, S., Helm, V., Skourup, H., and Davidson, M.: Sensitivity of CryoSat-2 Arctic sea-ice freeboard and thickness on radar-waveform interpretation, The Cryosphere, 8, 1607-1622, https://doi.org/10.5194/tc-8-1607-2014, 2014.

While we agree that a correct Cryosat reference is important for our procedure (also mentioned by the other reviewers), we think the general plausibility of this data set was shown in other studies. In addition, the framework of this work is the ESA Climate Change Initiative, which advices the use of existing algorithms whenever possible. Thus, the focus of this manuscript lies on matching Envisat freeboard retrievals to those of Cryosat-2 based on Envisat waveform characteristics. While changes in the Cryosat freeboard algorithm would impact the resulting Envisat freeboards, we expect the presented procedure to achieve a comparable fit between the two radar altimeter generations.

We rephrased parts of the Introduction to further stress this approach of using existing algorithms:

"*In this study, we focus on deriving an inter-mission consistent waveform interpretation scheme over sea-ice areas for Envisat and CryoSat-2 in the framework of the second phase of SICCI (SICCI-2).*

Therefore, the focus of this study lies not in a further optimization of the CryoSat-2 freeboard retrieval, but in the application of an evaluated methodology as is (Ricker et al.; 2014). Based on this approach, we want to find an optimal way to match the freeboard retrieval of Envisat to that of CryoSat-2 and build a consistent sea-ice freeboard data record that takes the different sensor configurations and differing footprints between both sensors into account."

We do also acknowledge the limitations of the current empirical CryoSat-2 freeboard retrieval, but the development and validation of an algorithm evolution for SAR sea ice altimetry is beyond the scope of this study. To highlight the ESA CCI approach on relying on existing algorithms, using CryoSat-2 as a reference for Envisat, we changed the title of the manuscript 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"

We of course replied to all suggestions made by the reviewer on this topic in the corresponding specific comments below.

Specific Comments:

P2, second paragraph: The wording choice is a bit awkward in parts...I'm not sure what quasi-nadir run-time measurement means here. "...which are so accurate" could be rephrased better.

The other reviewers also highlighted this paragraph as rather unclear so we changed it 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. 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. Here, one can differentiate between the height difference between the top of the snow surface and the sea surface (i.e., the total freeboard) and the height difference between the sea-ice surface and the sea surface (i.e., the sea-ice freeboard)."

P2 L18: "a the"

We changed that.

P4: If you have daily passive microwave measurements for snow depth retrievals then why is a climatology used for the Antarctic?

It has been shown in a number of publications that the snow depth based on passive microwave data can be substantially biased due to various physical properties of the sea ice and the snow itself, making the retrieved snow depth noisy and unrealiable at times. Using a climatology suppresses this noise. As the focus of this manuscript is on the possibility to match Envisat freeboard retrievals to those of CS-2 ones based on Envisat waveform characteristics (see last paragraph on page 1) we find it justified if not even mandatory to use a consistent snow depth on sea ice data set. We are aware of the fact that using a climatology is not ideal when it comes to the derivation and geophysical interpretation of a sea-ice thickness time series.

P5 L17-19: The use of SAR processing on CryoSat-2 will impact both the leading edge width as well as peakiness, I wouldn't expect these value to be equivalent to a pulse limited radar system for lead discrimination.

This is correct and is also not stated this way in the manuscript. While the overall used scheme is consistent, the resulting thresholds are not. From the provided threshold values in the appendix, one can see that the resulting thresholds are indeed different for the two sensors. However, what is meant here is that instead of relying on classifier parameter such as the stacked standard deviation, which would be available for CryoSat-2, but not for Envisat, we only use parameters that are available to both.

Section 2.3.2: Some further details on the k-means clustering is needed. Were the peakiness, leading edge width, and backscatter used here? What exactly is coming from the three clusters?

We changed several parts in the paragraph as also Reviewer 1 suggested changes here to increase the paragraphs clarity to the potential reader. To answer the reviewers question: Yes the three classifier parameters are used here and the result is a labeled training data set that can be used as training data for the second step in the surface-type classification procedure.

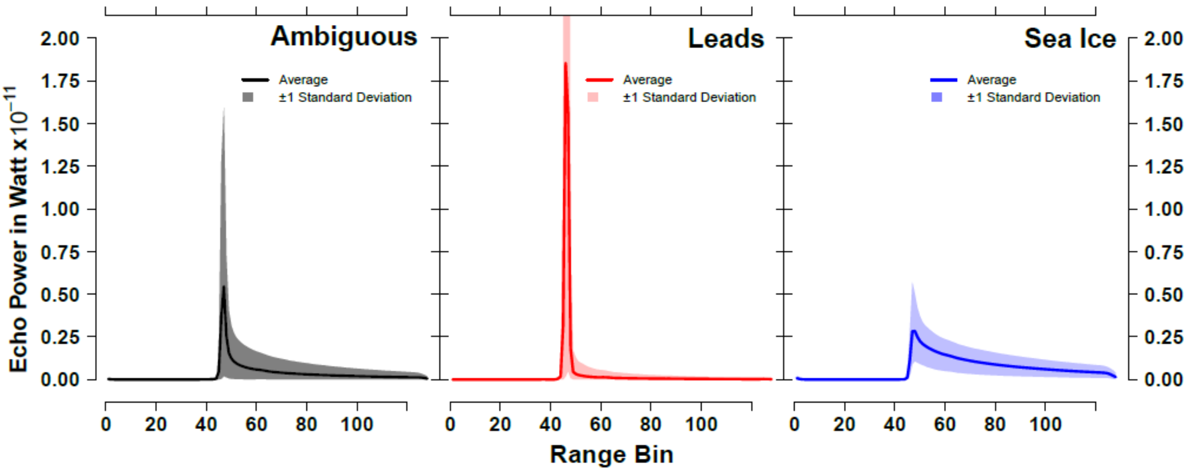
"Next, a subset of 1% is sampled at random without replacement (i.e., each original waveform with corresponding surface backscatter, pulse peakiness, and leading-edge width can only appear once) for each month in the MOP and for each sensor independently. This data sample is then separated into three 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."

Section 2.3.3: Same here for the need for further details. What is the training data set that was used, and how was this selected? How was it clear that the method separated leads and floes other than the fact that they had expected values for peakiness and backscatter? Was any validation done of the results to assess the quality of the classification?

In addition to the changes in the aforementioned paragraph, we also added more information to the subsection about the random forest classifier. We hope that we answered the reviewers question about the training data in the section about the k-means clustering (see above). The selection is rather simple statistics (through grouping of equal variance and increasing cluster homogeneity).

We want to argue that the presented results in the manuscript highlight the benefits and the capabilities of the proposed method. We highlight results from both, orbital as well as gridded data that again can be compared to all waveform parameters (Figures 2-5) as well as the resulting freeboard maps and differences. We think added information about sub steps would only lengthen the manuscript without adding much additional information.

However, we wanted to provide the reviewer with an additional Figure of the randomly picked and averaged Arctic Envisat waveforms in this response letter to persuade him from the resulting data quality. However, no additional validation was conducted.



While the average of the ambiguous waveforms does not look too bad, looking ad individual waveforms reveal the very high amount of noise and inherent surface-type mixture. Nonetheless, an improved future surface-type classification approach might be able to enhance the number of classified valid waveforms.

Section 2.4: The need for different thresholds for sea ice leads and floes from CryoSat-2 was shown in Kurtz et al., 2014. This should apply to Envisat since both operate on the same physical principle: the effective geometrical area of the lead return is very small causing a radar return which is close to the transmit pulse shape. As both satellites have the same bandwidth the transmit pulse shape should be very similar for both CryoSat-2 and Envisat. However, for sea ice floes the pulse-limited footprint size of Envisat should require a different threshold than the unfocused SAR footprint of CryoSat-2. This implies the threshold chosen for CryoSat-2 floe returns needs to be adjusted. No matter the methodology used though, some validation of the choice of thresholds needs to be done and I think that is lacking in the manuscript. Note too that the approach described in this section assumes the threshold used for CryoSat-2 is a control data set to which the Envisat data is tied, this means the threshold selected for the CryoSat-2 data set is of utmost importance. Thus some validation of this to demonstrate it is correct is sorely needed.

In general, we agree with the reviewer. We followed the described principles in this manuscript. We also agree that in that in a future algorithm evolution, it is our goal to have an adaptive threshold procedure for both, Envisat and Cryosat-2 tuned based on a large amount of different validation data. However, as mentioned above, the goal in the current project was also to use as many as established algorithms and procedures as possible and we likely stretched to the limit with our current implementation.

The reviewer is also correct in his statement about the pivotal role of and importance of the used CryoSat-2 data for the described principle. We agree that the 50%/50% choice of retracker thresholds for leads/sea-ice used in our implementation is rather an empirical than a physical choice. Nonetheless, in all validation exercises we conducted so far (see references on top) results appear to be plausible and robust. However, the general method presented in this manuscript evolves around the potential of linking freeboard differences between Envisat and Cryosat to Envisat waveform parameters through an adaptive choice of sea-ice retracker thresholds. This is independent of the absolute precision of the

CryoSat-2 data and rather relative towards it. In the aforementioned future evolution of the procedure, where we will have an adaptive retracker solution for Cryosat as well, the here presented procedure will be still valid, with just small adjustments in the predictors of the optimal threshold formulae. Because of that, we would argue that and intensive validation exercise is out of scope for this methodology manuscript. Aside from that, the ESA CCI project features a large validation exercise part that will be published in itself and will be freely available to all users.

P11 L8-14: This test should be done on the retrieved elevations (not just freeboard values) between Envisat and CryoSat-2, particularly for leads. That would more clearly demonstrate whether the differences in the threshold algorithms are properly handled.

We have deliberately chosen to evaluate only sea ice freeboard since absolute elevations, or biases therein, have no practical impact for the quality of sea ice thickness. In addition, we cannot assume that the elevations are identical due to the different orbits of Envisat and CryoSat. This kind of analysis would be required for sea level estimations, which is not the scope of our work.

P11 L15: How was the optimal value chosen? Was it that which had the smallest mean difference, RMS difference, or something else?

The optimal threshold was chosen based on the smallest absolute difference between Envisat and CryoSat-2 using the gridded monthly data. Here, the monthly Envisat freeboard was calculated for each TFMRA retracker threshold between 5% and 95% of the first maximum in steps of 5%. For example, in case a threshold of 50% results in a absolute difference of 0.2cm of a given pixel, whereas the 55%/45% thresholds result in higher differences, then the optimal threshold is set to 50% for that pixel given these values of sigma0 and leading-edge width.

Figures 10 and 13 seem to not match up visually. In Figure 13 there seems to be a far higher spatial coverage of red, indicating a higher Envisat freeboard whereas the distributions in Figure 10 seem to show only small mean differences and a more symmetric distribution. Some clarification on this is needed.

While the mean difference is indeed small, especially for November 2011 (~2cm; the left panel in Figure 13), the key to this visually appearing 'inconsistency' lies in the histograms (Figure 10, bottom row, second to the left panel). While the CryoSat-2 freeboards peak at a lower range than the Envisat freeboard estimates, we find these larger amount of red in Figure 13a. However, as these histograms are based on all available corresponding freeboard estimates, also all areas where Envisat underestimates CryoSat-2 are taken into account. Nevertheless, a very large amount of differences is centered on zero, which leads to the overall small mean difference.