Answers to Referee 2

Major comments:

The manuscript provides a good overview of the improvements in Cryosat-2 Baseline D product over the past Baseline C. This is an important paper for anyone using the product and deserves to be published. Overall quality of the paper is good. However, in my opinion, the authors should include more of the details of the processing in the manuscript and not only link to ESA technical notes online. Furthermore, there are missing details on some of the comparisons – for example the CS2/BGEP comparison is lacking information on averaging altogether. Also, one of the subsections (3.3.5, lead detection) needs considerable work.

The Section 3.3.5. is the weakest of the manuscript and must be considerably improved before publication. I would expect to see the justification of the inclusion of stack peakiness and the new classification scheme would be better than the old one. However, what is now shown is that the surface type classification of Baseline C and Baseline D are identical and for that only two tracks over two MODIS images are used. What should be (at least) included here is an overall statistical analysis of surface classification results over the whole Arctic (in the style of Figure 11) to see if the two are really same. If that is the case, the authors should further discuss why the SP was used in the classification. If the two classification results are different, the authors should make a solid point why the one in Baseline D is better than the Baseline C version.

Reply: We have entirely re-written section 3.3.5 adopting stack peakiness in the lead classification. This lead classification using SP conservatively returns fewer leads than previous lead classification, including SSD and PP (Tilling et al. 2018). We added a comparison in the monthly lead fraction map in April 2011. While overall spatial patterns are similar, the mean lead fraction in the whole Arctic is different. The lead classification using SP identifies somewhat big and wide leads with over SP 15 (Fig. 1). The threshold of SP should be optimised by evaluating the accuracy of ice freeboard and thickness. Adopting SP might consequently improve ice freeboard and thickness estimation by isolating nadir returns. Although it is hard to draw firm conclusions from this comparison, it is expected that adopting stack peakiness might help isolate nadir returns.

Minor comments:

Section 2 - Please include a short list (or a table) of the main variables in L1b and L2 products and what are their expected uses. Yes, they are in the product handbook, but they deserve to be mentioned in this paper as well.

Reply: Thanks to the referee for the suggestion. The following paragraphs will be added in the revised version of the manuscript in section 1 in correspondence with figure 1 which explains the actual implemented processing steps.

“The CryoSat Ice Processor generates Level 1B and Level 2 Ice products from L0 LRM, SAR and SARIn products. These products are primarily designed for the study of land ice and sea ice, although they are also relevant and useful to a wide range of additional applications.

Level 1B data consist, essentially, of an echo for each point along the ground track of the satellite. In all three modes, the data consists of multi-looked echoes at a rate of approximately 20 Hz.
Level 2 products instead are considered to be most suitable for users, as they contain surface height measurements fully corrected for instrumental effects, propagation delays, measurement geometry and additional geophysical effects such as atmospheric and tidal effects. In the L 2 products, the value of each geophysical correction provided is the value applied to the corrected Surface Height. Sea level anomalies and radar freeboard data are also included in the CryoSat Level 2 data products"

L122-123 Perhaps this (reasonably short) table could be included as an annex to this paper as well?

Reply: the link reported is related to an official ESA document and according to the authors, it is more appropriate to refer to the official document instead to copying the table in the actual manuscript.

Section 2.2. Here I would love to see a statement if there were anomalies or problems with Baseline-C that are still not fixed in Baseline-D. Maybe everything is fixed, but I'd love to know if there are pending improvements left.

Reply: All the foreseen evolutions and fixes of Baseline-C, have been implemented in the current Baseline-D processing chains (L1B + L2). Obviously, there is always room for improvement in operational products such as the CryoSat ones. Any other improvements or evolutions suggested by the scientific community will be analysed and considered by ESA to be potentially implemented in a future version of the ice processing chain.

L172-175 The new retracker should be described in detail. The authors should also present the rationale of choosing the retracker.

Reply: the sentence will be changed to:

"In addition, a new threshold-of-first-maximum retracker is used…"

And after that sentence, the following text will be added:

"Retracking is the process whereby the initial range estimate in the L1B data is corrected for the deviation in the first echo return within the waveform from the reference position."

L182 "Some tuning of the thresholds for the other metrics" - please tell us what kind of tuning and on which metrics!

Reply: Some tuning of the thresholds for the other metrics has also been performed, based on analysis of the test datasets

In addition, the following text will be added:

The discrimination algorithm currently uses sea ice concentration, waveform peakiness, and standard deviation of the stack of waveforms as metrics, in addition to peakiness of the stack. The discrimination thresholds are checked and adjusted whenever the L1 processing is modified to maintain the discrimination results.

L189 Why not just tell us what the surface type mask model now is?

Reply: The Level 2 products contain a flag word, provided at 1 Hz resolution, to classify the surface type at nadir. This classification is derived using a four-state surface identification grid, computed from a static Digital Terrain Model 2000 (DTM2000) file provided by an auxiliary file to the processing chain.

L241 typo – were compute

Reply: the word "were" is related to "L2 type products".

L250 – Where does this mask come from? Which mask is it?

Reply: The Level 2 products contain a flag word, provided at 1 Hz resolution, to classify the surface type at nadir. This classification is derived using a four-state surface identification grid, computed from a static Digital Terrain Model 2000 (DTM2000) file provided by an auxiliary file to the processing chain.

L253-L259 Also the retracker has been changed, has it not? How can we distinguish the effect of new retracker and new slope correction?

Reply: We didn't find any differences in the retracked range for LRM retrackers. Therefore, the effect of slope correction is independent of retracked ranges and can be distinguished.

L381 I would love to see the formula here as well. As well as a detailed description how it is used in the surface classification process. Even the Design Summary document does not include the thresholds used – and they might be beneficial for anyone trying to improve the surface classification in the future.

Reply: Empirical thresholds are found in Passaro et al. 2018. Given the complexity of the analysis and the length of its description, we do not find the scientific value of listing here the procedure which is already described in another peer-reviewed scientific paper.

Section 3.3.4 – In addition to WHB, there are also significant differences in the Kara and Barents seas. Would be good to mention and discuss there in the text. I would reckon this has something to do with relatively thin ice and lot of specular echoes in the area. Maybe include a zoomed version of Figure 11 difference map for these areas as well?

Reply: We agree that there are differences in Kara and Barents seas. We however did not highlight this region, as the observed difference is not related to a change in the IPF1C and IPF1D algorithms but rather to a lower number of orbit data sets in the IPF1D test data set. We have clarified this now in the text:

Average gridded SIT uncertainty in the AWI product for April 2014 is 0.64 m and we therefore conclude that a maximum ΔSIT of -0.015 m in the period of the TDS is insignificant for the stability of sea ice data records. This bias also includes an issue in the Barents and Kara Seas, where the number of orbits in the IPF1D test data set was less than in the IPF1C data and minor thickness differences can be observed in Figure 11 due to this selection bias.

L400 – 413 – Which ice type (density) and snow estimates are used in Baseline D? Are they same as in Baseline C? How is data averaged (both spatially and temporally)? Are all BGEP moorings used? Averaged together? What about OIB -which OIB freeboard is used here: radar or laser? How close to each other CS2 and OIB points need to be in place and time to form a pair? Averaging?

Reply: These are indeed very good questions. Within the CryoSat Baseline D products the freeboard is a radar freeboard.
Therefore, it does not require neither ice and snow densities nor snow depth.

For our validations we need these information to convert the radar freeboard into sea ice thickness or draft. For that purpose, we use the density provided by Warren99 with the official OSI SAF ice type product available on the NSIDC to separate the FYI and the MYI.
The snow depth used to take into account the decrease in radar velocity in the snow pack is the same Warren99 modified climatology for the 2 baselines.
All the BGEP mooring measurements of the 2013-2014 winter are used to perform the comparison (specified in the figure label).
The OIB dataset used is the NSIDC Quicklook version available at https://daacdata.apps.nsidc.org/pub/DATASETS/ICEBRIDGE/Evaluation_Products/IceBridge_Sea_Ice_Freeboard_SnowDepth_and_Thickness_QuickLook
(This point is added L401)

To process the OIB freeboard, we use the difference between the ATM laser total freeboard and the snow depth of the snow radar. The exact methodology will be added into the article (L400).

The same methodology is used for OIB and BGEP.
These in situ data are gridded into monthly EASE2 500*500 grids (the same grid as for the altimetric freeboard product).
Each in situ 'measurement' shown in figure 9 is the average of all data in a 12.5 km x 12.5 km pixel size.
This method removes the small scale variations in OIB and BGEP data that cannot be detected from satellite, therefore making the in situ data more representative of altimeter observations.

In order to clarify these points the following sentences will be added into the article:

L390: "...between the 2 baselines. The freeboard_20_ku parameter (freeboard of the 2 baselines) is a radar freeboard, i.e the raw measurement of the freeboard without corrections (such as the snow depth).

L400: Figure 9 presents scatter comparisons with the Beaufort Gyre Exploration project (BGEP, https;//www.whoi.edu/beaufortgyre) and NSIDC Operation Ice Bridge official product (OIB, https://daacdata.apps.nsidc.org/pub/DATASETS/ICEBRIDGE/Evaluation_Products/IceBridge_Sea_Ice_Freeboard_SnowDepth_and_Thickness_QuickLook) in situ measurements. To compute OIB sea ice freeboard, we calculate the difference between the ATM mean total freeboard and the snow depth estimated from the snow radar. The freeboard radar is derived taking into account the decrease in the radar velocity in the snow pack as follows:

$$FB_{radar} = FB_{ice} - snowdepth \times (1 + 0{,}51 \times \rho_s)^{(-1.5)} (2)$$

with $\rho_s = 0{,}3$

To compare with BGEP data, we compute a CryoSat ice draft from the difference between the gridded sea ice thickness (that integrates the snow load) and ice freeboard data. Note that the ice freeboard is calculated from the radar freeboard taking into account the decrease in radar velocity in the snow pack using the formula specified in Eq 2 with the snow depth provided by the Warren99 modified climatology and the official OSI SAF sea ice type classification available at the NSIDC.

To ensure the consistency between in situ measurements and altimetric observations, all data are projected onto monthly EASE2 500x500 grids identical to the one of the altimetric product. Each in situ measurement presented in Figure 9 is the average of all data in a 12.5 x 12.5 km grid pixel size.

L409-413 – The caption is confusing. What it should say is C and D are BGEP drafts compared to drafts calculated from CS-2 freeboards.

Reply: the comparisons reported in Figure 9 are indeed the Baseline-C and Baseline-D freeboard data (on Y axes) versus the OIB freeboard (X axes) for figures a) and b), while the c) and d) figures report the comparison between the derived drafts from Baseline-C and Baseline-D to BGEP draft.

L559 – which water mask?

Reply: For Sweden: Global Lakes and Wetlands Database

Lehner, B., & Döll, P. (2004). Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology, 296(1), 1–22.

For Tibet: Landsat based water mask

Jiang, L., Nielsen, K., Andersen, O. B., & Bauer-Gottwein, P. (2017). Monitoring recent lake level variations on the Tibetan Plateau using CryoSat-2 SARIn mode data. Journal of Hydrology, 544, 109–124. https://doi.org/10.1016/j.jhydrol.2016.11.024

The above references will be added to the revised version of the manuscript.

L564 – Why one meter? Where does this definition stem from? How would results change of more strict requirement (say 50 cm) would be used?

Reply: The one meter threshold was just chosen as a reference for comparing the two baselines. The point is to quantify the difference in valid observations between the two baselines. As suggested we could also choose a threshold of 0.5 meters as the reference. The results of this threshold are illustrated in the figure below. We will add the following sentence: "The one meter threshold is arbitrary and was simply selected to establish a common reference".

L575 – where did this offset originate from and which correction fixed it?

Reply: The range window extension introduced for SAR/SARIn modes in Baseline-C required that the code account for the change in reference bin position to avoid a 60 m height bias being introduced. For SARIn mode, the code was updated to fix the issue for the target surface types of ocean and continental ice, but not for other regions where the mode mask places the satellite in SARIn mode (i.e. rivers and lakes as in this case). This has been corrected in Baseline-D, removing the 60 m height bias everywhere.

References to this can be found in presentations held at Living Planet symposium of 2016 such as:

Bercher, Nicolas; Fabry, Pierre; Ambrózio, Américo; Restano, Marco; Benveniste:. Jerome: "Validation of CryoSat-2 SAR and SARin modes over rivers and lakes for the SHAPE project",

and

Borsa, Adrian: "Validation of CryoSat-2 LRM and SARIN-mode elevations over the salar de Uyuni, Bolivia"

L590 - "All kinds" however limited here to land ice, sea ice and (marginally) inland water. Rephrase.

Reply: done.