Manuscript ID: egusphere-2022-214

Title: Arctic sea ice radar freeboard retrieval from ERS-2 using altimetry : Toward sea ice thickness observation from 1995 to 2021

Marion Bocquet, Sara Fleury, Fanny Piras, Eero Rinne, Heidi Sallila, Florent Garnier, and Frédérique Rémy

---

We would like to thank again all reviewers as well as the editor for their relevant comments and for improving the quality of the manuscript. To answer to the major concern, we started our answer with a note as well as a new title suggestion.

General note concerning the manuscript objectives and new title suggestion :

The numerous questions raised by reviewer 1, and repeated in this 2nd review, lead us to believe that there has been a misinterpretation of our objectives.

Indeed, in this study, we intend to reconstruct a first homogeneous series of sea ice radar freeboard over the period covered by ERS-2, Envisat and CryoSat-2, by relying on the CS-2 SAR measurements. Indeed, CS-2 radar freeboard is the only one consistent for now and with the present knowledge of LRM over sea ice.

This work is in direct continuity of Guerreiro et al. 2017, Paul et al. 2018 and Tilling et al. 2019, all three of whom have worked in the past to extend the CS2 measurements by calibrating the Envisat LRM measurements on CS2 using one or two auxiliary data (PP for Guerreiro, LEW and Sig0 for Paul, distance between leads and floes for Tilling). They have obtained first series that are references, and Paul's version has even been adopted for SI-CCI.

Our objective was to extend this homogeneous series with CS2 back to ERS-2. The remaining discrepancies between CS2 and Envisat in the previous products, prevented us from considering Envisat as a new CS2 reference to extend the series to its predecessors ERS-21. We have therefore developed a method based on NNs that allows us i) to generalize the number of input parameters and ii) to take into account non-linear correlations between them. This allows us to effectively combine altimetry-bases parameters (FB TFMRA50, LES, PP), but also auxiliary dataset (ice type ratio, ice concentration, month of the year). Our solution is not only uniform with respect to CS-2, but it also improves the agreement with available in-situ measurements. The same approach was then extended to ERS-2, whose results were also validated with in-situ measurements.

Main concerns from the review are not related to the homogeneity of the series obtained, nor their physical representativeness, which are the intended objectives of this work. The reviewer's main concern is whether our results are mainly driven by TFMRA50 radar freeboard or derived altimetry parameters or by auxiliary measurements. The question is indeed interesting, and we attempted to address this concern, but this would be the subject of a proper study and unfortunately not the purpose of this paper.

Our concern come from the fact that same questions were repeated in the 2nd review with new analysis requests. We have once again responded to these questions in this new report, but we claim that the work requested deviate from the main purpose of the paper.

Indeed, we have never claimed that this series is based exclusively on altimetry data, especially since it is obviously impossible to reconstitute SAR measurements using LRM altimeter only (either SAR progresses would have been useless). Nevertheless, it can be noted that: 1. Three of the six used parameters are based on waveforms characteristics, and we had made many preliminary tests to check that each of these 6 parameters improves the solution. Even if these additional parameters would have been sufficient on their own (which would have been surprising), this would not have invalidated our objectives. 2. More importantly, CS-2 radar freeboard product is a pure based altimetry product, and our objective is really to reconstitute this radar-freeboard using Envisat and ERS-2.

Thus the radar freeboard estimation we propose, even for ERS-2 is mainly based on CS-2 one and less on TFMRA50 FBr from ERS2.

Nevertheless, we agree that the title "Arctic sea ice radar freeboard retrieval from ERS-2 using altimetry : Toward sea ice thickness observation from 1995 to 2021" is ambiguous from this point of view. We therefore propose to change it to **«A first homogeneous Arctic sea ice radar freeboard from ERS-2, Envisat and CryoSat-2 : Toward sea ice thickness observation from 1995 to 2021"**. This title better explains our objectives and avoids associating directly and exclusively the notion of radar freeboard to ERS-2. (We have not yet changed the title within the manuscript and in TC to avoid changing it several times if reviewers and editor do not find it relevant).

In this document, the referee's comments are in bold type, the answers are in italic type, and the corrections to the revised manuscript are in normal type.

Answers to report #1

Minor comments

Given my concerns about whether the presented neural network can be reasonably described as a "calibration", I should point out that calibration actually refers to the comparison of one data set to a standard, and does not refer to subsequent adjustment of the initial data set to match the standard. Really what is being done by the NN in prediction mode (after training) is adjustment. I appreciate that calibration is sometimes uses wrongly to refer to adjustment. I've done this, and there are several instances of this in the remote sensing literature. But the BIPM is clear in its most recent glossary 3 that calibration should not be confused or used interchangeably with adjustment. The last paragraph of Section 1 actually uses "adjust" correctly three times, but elsewhere the process of adjustment is referred to as calibration. I think by using the proper terms both here and in the manuscript, we would be able to better discuss the performance of the blur-corrected TFMRA50 Rfbs in the calibration procedure, and the relatively large magnitude of the subsequent adjustment required.

In the first version of the manuscript, the term 'correction' was used, which we replaced with "calibration" according to your suggestion. Your new review suggests to replace 'calibration' by 'adjustment' following the definition of 'International Vocabulary of Metrology'. As you noticed, 'calibration' is not adapted according to this reference, as it doesn't properly characterize the procedure we developed. However, this change has been operated according to you advise. We suggest keeping the initial term 'correction' that seems to be right and keep consistency with former studies (Guerreiro et al 2017, Tilling et al 2019) that deal with Envisat LRM TFMRA issue, even if it is a more general term. As it is a way to correct a systematic error on the radar freeboard evaluated thanks to CS-2 and due to surface roughness and signal penetration. To be clearer in the manuscript, we suggest writing "NN FBr" while dealing with radar freeboards that have been corrected based on the Neural Network.

L377-378 : When quoting the standard deviation bias between data like this, it would be best to see them as scatter plots, with the line superimposed. Especially as these numbers appear in your abstract.

It is not always easy to choose the most adequate illustration, however in this case what we want to show is that the radar freeboard distribution has changed. It also allows to show the distribution of the residual difference between the two missions. As indicated, we still provide the main statistics, thus the scatter would not provide mush more information. We prefer to keep the histograms and not to overload the manuscript with other figure and scatter plots.

L107: The authors have not meaningfully responded to my previous comment about the MYI fraction product they have constructed. Trivially, they have again misidentified the NSIDC ice age dataset as "0061", when it is "0611". More importantly, they have not addressed my other point that 0611 is not a simple map of ice age, but a map of the oldest

ice contained in a given grid cell (see Tschudi et al. 2020 and my previous review). At the moment, they have treated grid cells containing any MYI at all as being made up of 100% MYI (see figure 10), which will undoubtedly introduce a high-bias into the MYI fraction product that they construct. I would urge the authors to go back and engage with my previous comment on this matter.

We thank the referee to have noticed this mistake regarding the NSIDC reference, it has been corrected in the manuscript.

Concerning the "myi fraction", we are aware that it is a map of ice age corresponding to the oldest ice contained in a given grid cell. It is still possible to derive an information concerning the type of ice that occurs within this grid cell. Indeed, Yufang Ye et al. 2023 shows that in most of the case "NSIDC-SIA can generally capture the SITY distribution pattern but exhibits a slight over- or underestimation of MYI, which can be explained by the ice age assignment of the oldest ice and different temporal resolution of NSIDC-SIA compared to SAR. These results agree with previous studies (Korosov et al., 2018; Ye et al., 2019) and once again confirm the use of the NSIDC-SIA product as a cross-validation dataset".

To our knowledge, there are no available datasets that provide 30-years of MYI concentration. However, it is an important information to better estimate thickness from sea ice that can have similar backscatter but from different ice type (e.g rough ice and ice close to the edge). This MYI proportion will be overestimated but as there is nothing better, it gives the information we need, to recover what we are looking for: the radar freeboard.

The major concern here comes from the fact that it's not a proper multi-year ice fraction, but a mean ice type seen by the satellite. We propose to change this paragraph to inform readers on that point by adding these sentences:

The proportion of multi-year ice of a given grid cell refers, in this study, to the mean ice type observed by all the tracks (for each month of each mission) that pass within a 25 km radius of this grid cell. This value is computed during the gridding step. The proportion would consequently be overestimated compared to what can be estimated with ice age tracking algorithms.

Ye, Y., Luo, Y., Sun, Y., Shokr, M., Aaboe, S., Girard-Ardhuin, F., Hui, F., Cheng, X., and Chen, Z.: Inter-comparison and evaluation of Arctic sea ice type products, The Cryosphere, 17, 279–308, https://doi.org/10.5194/tc-17-279-2023, 2023.

Korosov, A. A., Rampal, P., Pedersen, L. T., Saldo, R., Ye, Y., Heygster, G., Lavergne, T., Aaboe, S., and Girard-Ardhuin, F.: A new tracking algorithm for sea ice age distribution estimation, The Cryosphere, 12, 2073–2085, https://doi.org/10.5194/tc-12-2073-2018, 2018.

L80: I don't think the change that you've put in your review responses has been included in the resubmitted manuscript? The text reads the same as before. I'm also still not convinced that SARM is less sensitive to a given roughness than LRM, as implied. I understand that the footprint is smaller, so potentially the footprint contains less variability in surface height (by perhaps including fewer floes of distinct freeboards). But isn't the point of altimetry to characterize some kind of average height? I can't see how a ridge in the footprint is less of a problem for a SAR waveform

The change is missing in the resubmitted manuscript, we apologize for this omission. It has been actualized according to this review.

Coming back to the relative impact of surface roughness between SAR and LRM, the SAR treatment haw two main effects regarding LRM : the first is to reduce the footprint significantly (with a ratio 20-40). and the second is to make this footprint constant. Indeed, the footprint of CS-2 SARM is about $5km^2$. *The reduced footprint allows to focus the measurements on or closer to the nadir. Moreover, as explained in Raney et al 1995., the footprint size is constant by construction in SARM, contrary to LRM. Indeed, the area illuminated by the altimeter depends on the surface roughness (Chelton et al 1989), Envisat and ERS footprint vary from 100 to 200* km^2 *(studies mainly deal with oceanic surface, and the surface roughness over the ocean is characterized by the wave height). Thus, roughness impacts the waveforms, in particular the leading edge, especially in LRM. For oceanic surfaces, Brown's model deal with the roughness and can recover the average surface height and the roughness. During the retracking step, the modeled roughness is taken into account. Such a model, unfortunately, doesn't exist on sea ice. Heuristic approach gives an idea of the surface height impacted by a roughness. SARM sightly less impacted by this phenomenon thanks to its footprint and characteristics, (it is in some way an intermediary between LRM and laser technology). SARM heights are the only radar measurement that is directly usable on sea ice, whereas LRM as to be corrected from surface roughness using SARM. From this point of view, CS-2 was a real revolution.*

Chelton, D. B., Walsh, E. J., and MacArthur, J. L.: Pulse Compression and Sea Level Tracking in Satellite Altimetry, Journal of Atmospheric and Oceanic Technology, 6, 407–438, https://doi.org/10.1175/1520-0426(1989)006<0407:PCASLT>2.0.CO;2, 1989.

Raney, R.: A delay/Doppler radar altimeter for ice sheet monitoring, in: 1995 International Geoscience and Remote Sensing Sym-posium, IGARSS '95. Quantitative Remote Sensing for Science and Applications, vol. 2, pp. 862–864, IEEE, Firenze, Italy,665 https://doi.org/10.1109/IGARSS.1995.521080, 1995

We propose to replace :

Contrary to SARM, LRM altimetry measurements are strongly impacted by the surface roughness of the surface illuminated by the radar, also affecting the freeboard measurement (Raney, 1995) (see Sect 3.4 for more information). Our approach is to make use of these processing mode differences to derive an LRM corrected freeboard. To that end, we compare Envisat and CryoSat-2 datasets during missions-overlap period which runs from November 2010 to March 2012 (see Sect. 3.4).

by:

The surface illuminated by the satellite is significantly larger in LRM than in SARM (by a factor of about 30). Moreover, in LRM the surface roughness conditioned the size of the illuminated footprint, the larger the roughness, the larger the footprint is, whereas it is constant in SARM (Chelton et al., 1989; Raney, 1995). Thus, surface roughness will impact more LRM range retrieval than more nadir-focused measurement, SAR technologies (see Sect. 3.4 for more information). There is no waveform model for sea ice to account for the effect of roughness, and conventional retracking methods don't allow relevant radar freeboard estimation using LRM information alone. Our approach is to exploit these processing mode differences to derive an LRM-corrected freeboard. To this end, we compare Envisat and CryoSat-2 datasets during the mission overlap period from November 2010 to March 2012 (see Sect. 3.4).

L306: "Moreover, in order to avoid over-fitting, an early stopping criterion is used to stop the model training as soon as the score is not improved during 10 consecutive iterations, with a defined tolerance." I don't see how this would stop overfitting? You could easily have overfitted your model by the time that you fail to get an improvement in score, because overfitting can result from spurious improvements to the score that don't reflect improvements in the underlying model. I think overfitting is normally identified through comparison of performance on test sets with training sets. In fact, I would suggest this model may well suffer from overfitting, due to the very low ratio of bias to variance (3mm to 9 cm in the Envisat-CS2 evaluation, 2mm and 3.8 cm for ERS/Envisat). This ratio is a well-known indicator of overfitting, and suggests your model has captured the noise in its training data as well as the underlying signal. 4 You should address this in your discussion.

Early stopping works with a validation dataset that is not used to fit the model during the training step and which is different from the testing set. Early stopping doesn't work with the training set, of course it would not be able to prevent any overfitting. We suggest to develop this part a way more in the manuscript to answer your concerns.

We are not sure to understand where these numbers come from, nevertheless, all the statistics presented in section 4.2.1 and 4.2.2 are corrected radar freeboard, predicted from the NN that have been trained and the whole dataset (2010-2011/ 2011/2012 or 2002/2003), minus the validation set that is randomly chosen. So we think that these numbers won't reveal any overfitting. However, unexpected a light overfitting can obviously occur. As mentioned by Jack Landy, comparisons with validation datasets show good agreement even for ERS-2, so even if we have a light overfitting, ERS-2 thickness remains consistent with validation dataset.

L300: If I understand right, you've specified a range of discrete hyperparameters and then investigated which combination optimizes the score. But this description should be contextualized with the ranges and intervals over which you searched, the dimensionality of the hyperparameter space that was searched, and most importantly the results of the search. You've made choices about the learning rate, regularization term and weights-solver based on your grid search,

but you haven't reported what those choices were. Given you performed what must have been a highly multidimensional and computationally intensive score-sensitivity analysis to get at these specific hyperparameters, it would be great to know what they are. I'm also not totally clear how the activation function was both found through this grid-search, and also motivated by the domain of the TFMRA Rfbs. I guess the domain of possible activation functions was motivated by the domain of Rfbs, and then you ended up selecting the sigmoid specifically through the grid search method ?

In the previous review it was suggested to justify the sigmoid choice compared to RELU, that the reason why we added a possible explanation for why sigmoid works better. However, all available activation functions and solvers were tested (with grid search) before choosing the best combination.

The hyperparameters selection is presented in appendices in the new version of the manuscript, as suggested.

L429: Replace "seems" with "is". It definitely is sensitive to the algorithm used.

Modified

L34 of the manuscript now reads "Sea ice thickness estimation by spatial altimetry was first introduced by Laxon (1994) and Peacock and Laxon (2004) based on the freeboard methodology". I think the first real description of sea ice thickness estimation from radar altimetry derived freeboards was by Stanley et al. (1979). Their paper compared elevations from repeated tracks of the GEOS-3 altimeter in both the presence and absence of sea ice. They explained that the differences between the ocean elevations and the sea ice elevations (which were too large in their investigation due to poor gain control) contained information on the freeboard, which in turn could be used to estimate thickness if better constrained. The breakthrough of Laxon's work with ERS1 was the individual classification of waveforms into lead/floe, allowing effective interpolation of the local sea surface height – this unlocked the method. So I think to make Laxon (2004) the valid citation, you should just add something like "in its modern form" to your sentence.

Taken into account.

I can't see why a non-geophysical parameter such as "date" is being used to adjust the TFMRA waveforms (Figure 5). What physical mechanism could justify its inclusion in the neural network?

It seems to me that the inclusion of a date parameter will only make the neural network behave like a seasonal climatology (further reducing the ratio of bias against variability); this would evaluate very well against in-situ sources in the way that you've constructed your tests (given the dominance of seasonal variability over interannual variability in sea ice thickness), but would potentially be an example of spurious overfitting. At minimum, a bit of justification needs to be given for its inclusion, preferably with some information about the final effect of its inclusion.

This choice comes from the seasonal variability of sea ice properties. Indeed, The physical properties of sea ice and snow on sea ice can be slightly different from one season to another, this will necessarily impact waveforms so the derived TFMRA50 FBr, LES, PP. Given the difference between SARM and LRM antenna gains, this impact will be different depending on the mode, and the correction will be more accurate if it takes into account this seasonality. Figure 1 shows BGEP A and C sit time series with NN FBr for Envisat and the NN FBr with a NN trained without month as a parameter. Envisat draft that has been trained with month information capture better the seasonal variability. Since the time series run before October 2010, it's shows that it doesn't lead to an overfitting as it works similarly for 2010-2012 than for some other winters, except for 2006-2008. For this two winters the draft varies a lot and its evolution differs from other winters, we can imagine that moorings captured floes from

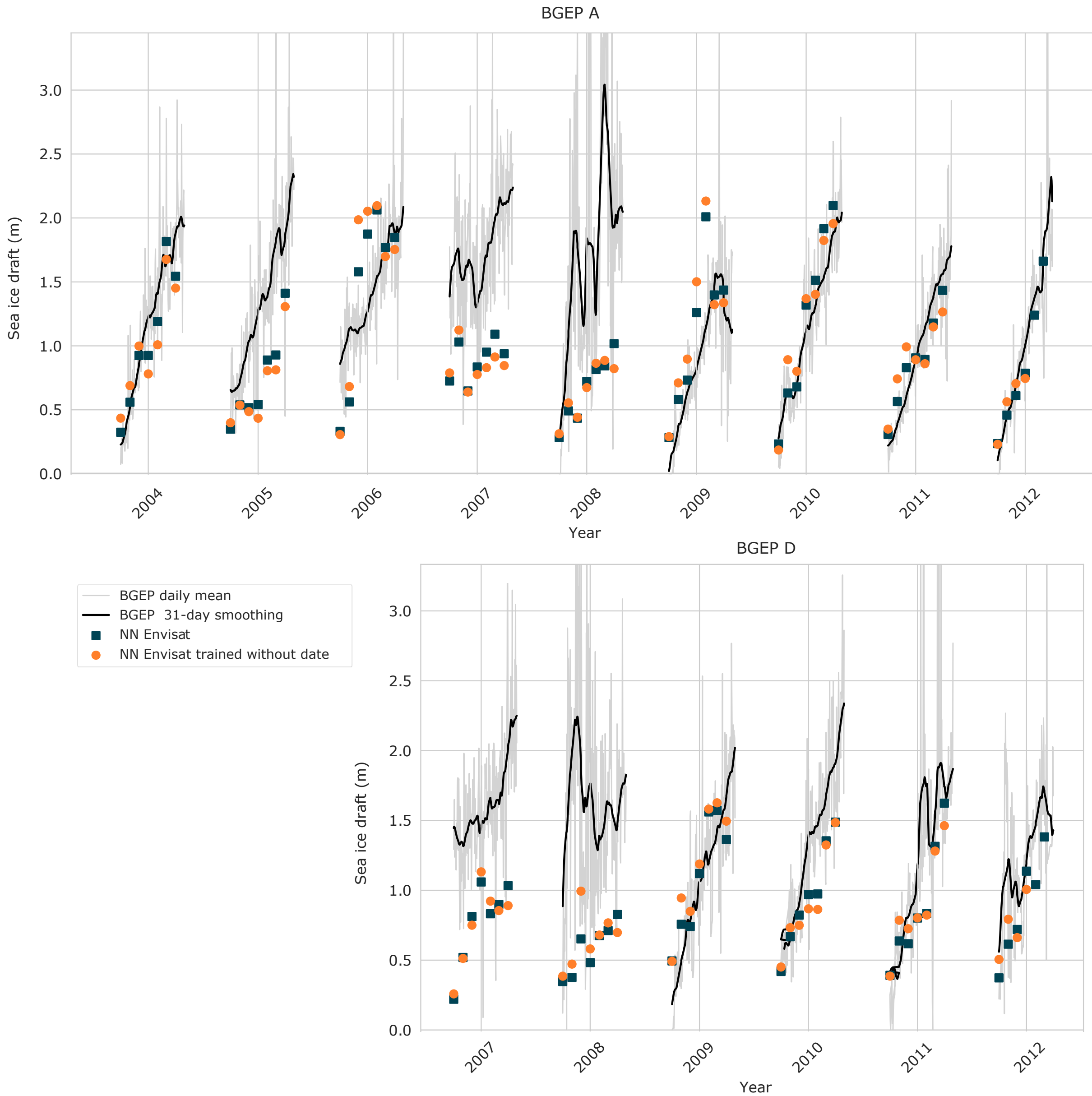

Figure 1. BGEP A (top) and D (bottom) daily and 31-days smoothed draft compared to Envisat NN draft between 2003 and 2012. In blue Envisat NN draft and in orange Envisat NN but trained without the month information.

different types and that MYI edge was at moorings locations.

The following sentence has been added to justify the non-geophysical inputs, the period of the year :

Note that the period of the year is taken to capture better the seasonal variability, as snow on sea ice as well as sea ice physical properties change along the seasons.

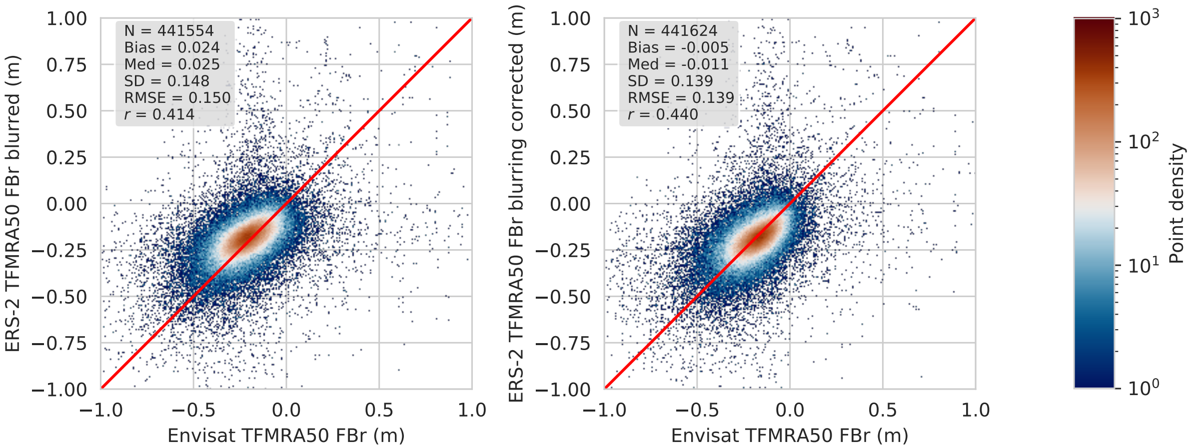

Figure 2. ERS-2 TFMRA50 radar freeboard before and after blurring correction compared to Envisat TFMRA50 radar freeboard for the missions overlap period (winter 2002-2003) with monthly maps.

I think the language surrounding the pulse-blurring correction needs a bit of clarifying. The reader should be in no doubt that a height correction has been applied to the retracked elevations to remove the effects of pulse blurring. The waveforms have not been "deblurred" and then retracked. I think the authors use of "corrected" is good, and I would suggest removing references to "deblurred" and "deblurring".

This remark has been taken into account.

On the pulse-blurring correction – have the authors checked to what degree this actually improves their radar freeboard estimates after the NN-based adjustment? Seems like this would be a fairly significant finding, and would justify both their efforts and their phrasing of it as a "correction".

We don't have trained the neural network on ERS2 blurred data, but we have evaluated the signal-to-noise ratio between ERS and Envisat before and after blurring correction. We showed in the past review as well as in the manuscript that the blurring essentially noise the heights and that this noise is highly asymmetric, which introduces a non-negligible bias. As presented in the manuscript, correcting the blurring largely reduced the radar freeboard variability and after the blurring correction, its variability is similar to Envisat radar freeboard one (A1) and the mean bias between radar freeboards is divided by 5. 2

Nevertheless, since blurring is mainly noise, it is hard to quantify precisely the impact of the correction on monthly means and on another side as RA are sufficiently different to prevent from relevant along track deep comparisons.

L41: I think a better reference for snow penetration is potentially Ricker et al. (2015) 6, which is entirely focussed on the topic.

This reference has been added.

Major comments

The purpose of this study was to present a consistent and homogeneous FBr for the ERS-2, Envisat and CS-2 missions. The challenges for LRM waveform retracking have been explained both in the previous review and in this one. We hope that the additional explanation of LRM/SARM differences will help to address some concerns. To sum up, LRM waveform retracking *process need to take into account the surface properties more than SARM to have meaningfull FBr, so fixed threshold retracking in LRM can not capture the surface height for sea ice. To recover a consistent FBr we need at least information on observe roughness as demonstrated in previous studies (Guerreiro et al 2017 and Paul et al 2018). Readers should not expect TFMRA50 FBr to tell something on the SIT because LRM waveform fixed-threshold retracking is not enough to recover a valuable height in LRM on sea ice. Thus, TFMRA50 FBr and some waveform characteristics e.g. TFMRA50 heights, LES, PP, Sigma0 is expected to be enough to capture enought information to estimate the SIT. Moreover, month, concentration, MYI proportion are also used to give more information to the regression in order to capture better the seasonal variability, the spatial variability and improve the consistency with CS-2. So indisociably, TFMRA50 FBr and roughness will tell you a coherent FBr. Because MYI proportion is in a way linked to roughness, it will be correlated to the NN FBr, but we can still not still that because the correlation of the NN FBr is stronger with age/type than with TFMRA50 FBr, the TFMRA50 contribute relatively little compared to type/age in the regression. As shown in the previous review, regression without MYI proportion, concentration of month information give a meaningful FBr.*

As requested, Figure 4 shows inputs in function of NN FBr. Month and MYI proportion are not represented as it is discrete values, scatters don't represent anything. As the purpose of the study is not to mathematically describe the relationship between parameters and NN FBr, we won't go further into details and propose to put this figure in Appendix. We will unfortunately not go further to understand the size and the nature of the correction so we will not explain more the impact of the different input parameters. We do think that it could be a full study to do it and that the understanding around sea ice roughness with altimetry and especially LRM altimetry does not enable to go further. Concerning the point that for some regions, the size of the correction is equal to TFMRA50, this is a pretty interesting results. This scatter shows that there is very few TFMRA50 FBr that have not been corrected, and that noisy TFMRA50 (close to marginal ice zone for instance) is set to 0, so the correction has to be equal to the TFMRA50. It thus gives interesting information but do not draw conclusion on the correction size for all the point, as it is only a low proportion of the dataset. These same points also correspond to Envisat NN FBr that are equal to zero see 4 .

Concerning Figure 3 of the previous review, we apologize for not responding to your concern. Indeed, Fig. 3 doesn't show whether NN trained with TFMRA50 FBr, show better results than without. Note that without TFMRA50, PP and LES the NN doesn't converge to consistent values.

We assume that theese conclusions (about the NN) are also true for ERS-2 as the RA as well as orbit are even more similar.

Answer to report #2

We would thank the reviewer for his comment, for helping and bringing clarification and supplementary information of the previous and present comments. As suggested, we have added some information within the introduction as well as in the conclusion to reframe the purpose of the study and give information to the readers on the usage of the time series.

Answers to report #3

Regarding your general reply: I appreciate the exhaustive analysis and the work the authors put into this review. There are two points that however still concern me:

1) All results and additional comparisons you show are with ENVISAT instead of ERS2. Wouldn't the authors agree that with all additional dead weight (pulse blurring, degradation of satellite systems, etc.) ERS2 is supposedly the sensor of interest here as it is clearly the NEW thing? Envisat was shown before in several studies that it can be made 'reliable'. I would love to see some comparison with ERS2 and in addition.

2) The authors went to such great lengths to put this analysis together but nothing of it appears to be in the resubmitted manuscript. I think the authors should give themselves credit for their additional work and put this analysis

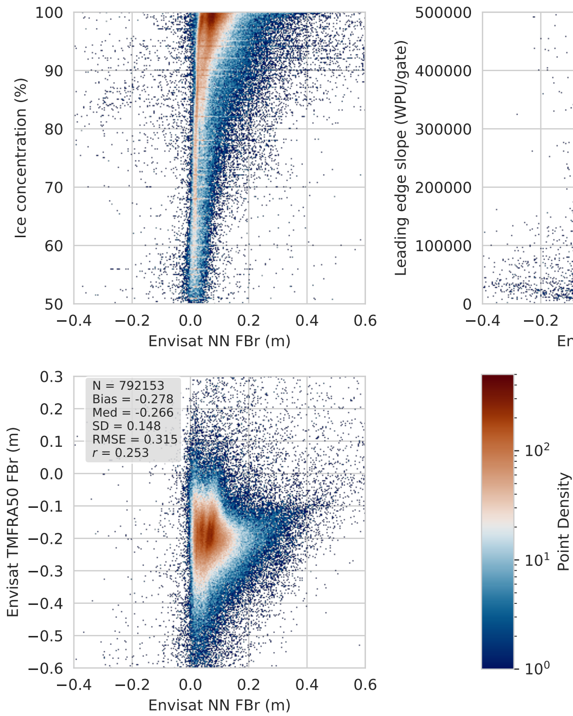

Figure 3. Envisat NN FBr compared to some of the regression input parameter, the sea ice concentration, the Leading Edge Slope, the Pulse Peakiness and TFRM50 FBr.

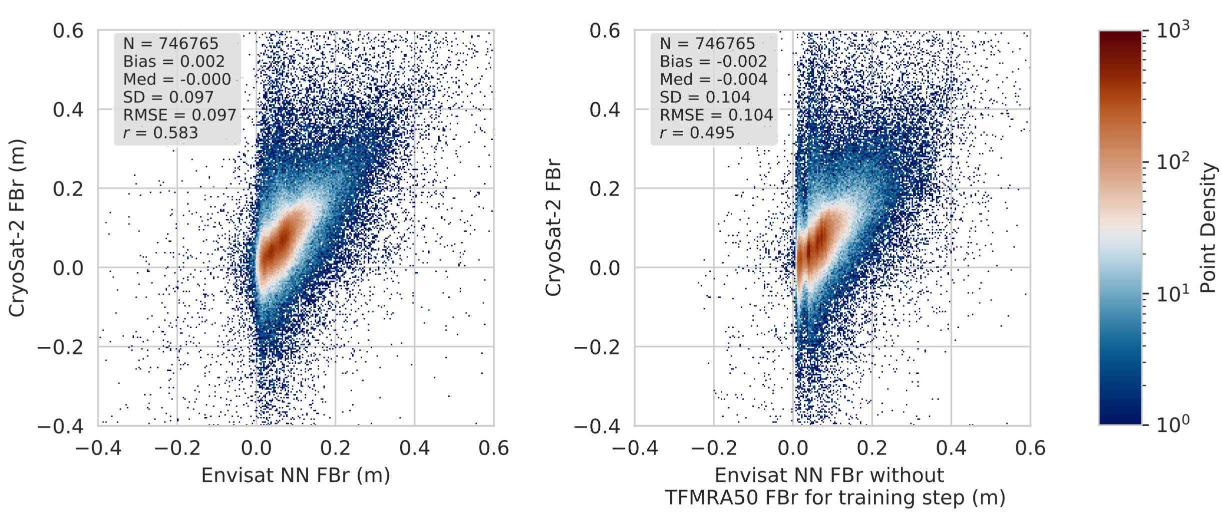

Figure 4. Envisat NN FBr compared to CS-2 with a NN trained with and without TFMRA50 FBr.

into their discussion. Otherwise the raised points of concerns by the reviewers potentially remain also for any future readers.

1) If you refer to figures within the previous review, they were made to explain a bit more the objectif and to answer to some concerns about the methodology, but are not at all part of the purpose of the paper. We added some new comparisons in appendices to show the impact of Blurring correction as requested by Robbie Mallet, but for other it will take a lot of time to do to order and to integrate knowing that it is not the purpose of the study. What is interesting is that the most difficult part of the "LRM correction" using this methodology, is between Envisat and CryoSat-2 as they don't have the same orbit neither the same altimeter, they do not observe the same sea ice floe at the same moment contrary to Envisat and ERS-2 that are few hours delayed with the same orbit. In other words, sea ice properties are much more comparable between Envisat and ERS-2 than CS-2 and Envisat. If you refer to ERS-2 validation, there are several comparisons with independent dataset in the paper.

2) We hope the previous paragraphe will answer to your concern. Moroever, we added some new comparisons between inputs and outputs of the Neural network for Envisat. Similary it is just given as information but it is not a part of the study.

Regarding my general comments: Similarly, to my suggestions above, the comparison to Guerreiro et al, could be added to your manuscript or at least to the appendix.

Thanks for the suggestion of adding Envisat FBr from guerreiro et al 2017, it has been taken into account.

Regarding my specific comments: L183 of the rebuttal: Minus the ‘and’ in front of ADAM.

Taken into account

Regarding the Test/Train split: I found this quite surprising that a random split lead to perfectly equally distributed data sets? And this is valid for each of the 5(?) randomly chosen test runs per model setup?

Yes it is, the number of data that represent 10% of the missions overlap period is large enough to capture the distribution according to the law of large numbers.