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

Jack Landy (referee n°1) - global comment

The authors construct a 25-year record of Arctic sea ice radar freeboard by reconciling the measurements from three radar altimetry missions, one ongoing and two historic. Their primary motivation is to generate the first step towards a long-term sea ice thickness record for the Arctic Ocean. This would be the first observational sea ice thickness record spanning such a long period and would offer valuable comparison to existing proxy sea ice thickness (SIT) records based on ice age and models. In my view, a robust 25+ year time series of Arctic sea ice thickness would represent a major scientific breakthrough with implications for understanding global climate changes in the modern era and validating and improving sea ice models, among other potential applications.

Generally, I find the approach and methods to be scientifically sound. I have some minor comments but nothing that questions the rigour of the generated time series. The validation against existing SIT data from satellites, airborne and in situ sensors is comprehensive and convincing.

Excellent work on a really valuable study – it was a pleasure to read! Feel free to get in touch if you have any questions, Jack Landy

Answer to Jack Landy (referee n°1) - global comment

We would like to thank the reviewer for his careful reading of the manuscript, for this positive feedback and for the relevant and constructive remarks that have helped to improve the quality of the manuscript. In order to fit with your comments, we have made a revision of the manuscript that should have corrected the textual issues and well improved the readability of the document. We hope that these modifications will meet your requirements. Please find below the details on how your specific comments have been taken into account. *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 Jack Landy (referee n°1) : specific comments

Line 2. Sea ice volume's..?

This correction has been done.

L14-15. I would suggest including other statistics of the variability on the bias within the abstract. Given the ML algorithm aims to remove the bias I would argue the stats on variability are more interesting for the reader.

These statistics have been added to the abstract. The following modification has been done :

L 14-15: Comparisons of corrected radar freeboards during overlap periods reveal good consistencies between missions, with a mean bias of 3 mm for Envisat/CryoSat-2 and 2mm for ERS-2/Envisat.
replaced by
Comparisons of corrected radar freeboards during overlap periods reveal good consistencies between missions, with a mean bias of 3 mm and a standard deviation of 9.7 cm for Envisat/CryoSat-2, and respectively 2 mm and a 3.8 cm for ERS-2/Envisat.

L28. Technically past radar altimeters have not allowed basin scale, so altimetry doesn't offer a 'global approach' over the long term. But this is nit-picky.

That's true, we have modified a little the sentence.

L28. A global approach is possible through satellite altimetry, especially with radar altimetry, which is not impacted by the cloud cover and whose missions are continuous since 1991.
replaced by
A quasi-global approach is possible through satellite altimetry, especially with radar altimetry, which is not impacted by the cloud cover and whose missions are continuous since 1991.

L51. Explain 'heuristic retracker TFMRA50'.

Indeed, the sentence was not clear. It has been corrected. The purpose of this sentence was to explain that before any calibration (LRM/SAR) all the waveforms for both missions were processed with the same algorithm and the retracker that has been used is a TFMRA retracker (which is categorized as an empirical retracker based on a heuristic approach).

L51 : The consistency between missions is preserved by using the same processing chain regardless of the mission, with the heuristic retracker TFMRA50 (Helm et al., 2014)
replaced by
The consistency between missions is preserved by using the same processing chain regardless of the mission (before calibration), starting by the retracking algorithm : the empirical threshold first-maximum retracker algorithm (Helm et al., 2014) with a threshold of 50 % (TFMRA50).

L65 : Check Appendix Table 1. Does this tally?

The appendix indicates the RA characteristics. The table is supposed to help the reader to understand the estimation of uncertainties from the speckle noise for Envisat and ERS-2. Wingham et al. (2006) estimates the uncertainty due to speckle noise for CryoSat-2 in SAR, SARIn Mode as well as for LRM. CryoSat-2 LRM mode data are not used (neither pLRM) in this paper, but the CS-2 LRM speckle noise error on range is used to compute ERS-2 and Envisat uncertainties that come from the speckle

noise (L297-302). The legend of table A1 (now A2) has been developed, and a reference has been added in the corresponding section 3.5.

L103-104: How is it aggregated? Bit vague.

The following modification has been done to try to improve the clarity of the manuscript:

L103-104 : This information comes from the NSIDC 0061 sea ice age product (Tschudi et al., 2019) that is aggregated into two classes (MYI and FYI)

replaced by :

The study also requires a sea ice type product, this information is derived from the NSIDC 0061 sea ice age product (Tschudi et al., 2019) that is aggregated into two classes (MYI and FYI) according to the age of the ice (FYI : ice age between 0 and 1 year, MYI : ice age of at least one-year) at a weekly frequency. Data are respectively available as daily and weekly map with a 12,5 km grid resolution. The fraction of MYI is derived from the ice type information during the gridding processing step.

L134-135. Requires citations.

The citations were a few lines after, but for clarity we have added one in the first line.

Figure 1. Could you add here a map of the satellite coverage and the locations of different validation datasets? This would be useful for the reader to understand limits of the record and interpret differences to specific validation data.

Figure 1. has been completed with a sub figure to represent the locations of different validation data sets with satellite coverage limitations.

L166-167. Which version of the IS-1 data was used?

ICESat-1 version used is the one from NASA Goddard not from NASA JPL. This precision has been added in the description paragraph.

Figure 2. I would suggest to add histograms to one side for each of the three elevation profiles, so it is easier to visualize any differences/biases

As suggested a sub figure with probability density function, especially density probability function has been added in the following figure 2:

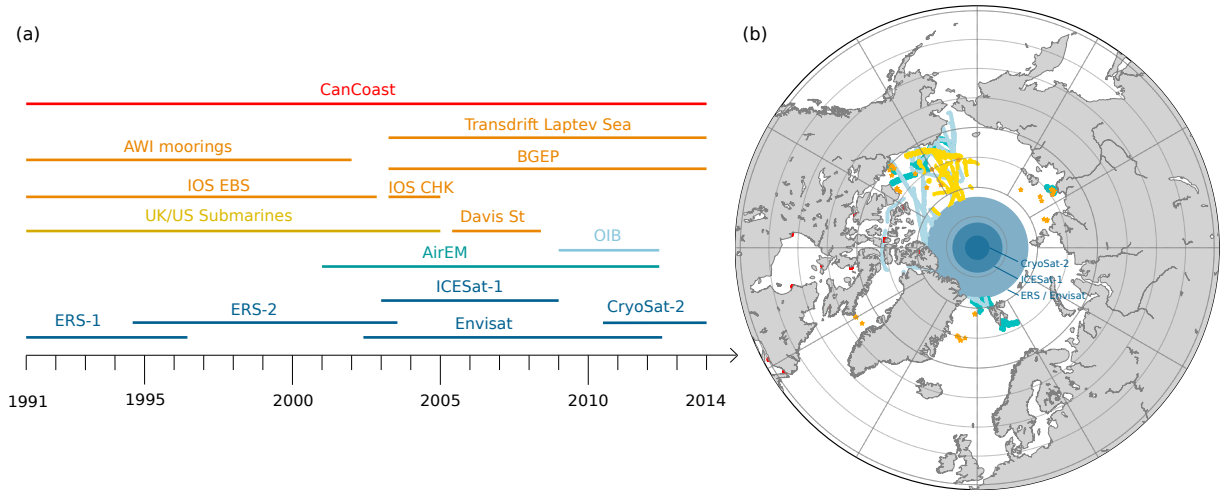


Figure 1. Summary of various available dataset for Envisat and ERS validation. Colors distinguish the different types of data. Dark blue for satellite products, light blue for airborne data, yellow for submarines, orange for anchored moorings, green for buoys and red for direct measurements. (a) Temporal availability (b) Spatial availability and extent of missions data gaps. Blue rounds represent altimeters coverage limitation due to their orbit inclination (81.5°N for Envisat, 86°N for ICESat-1 and 88°N for CryoSat-2)

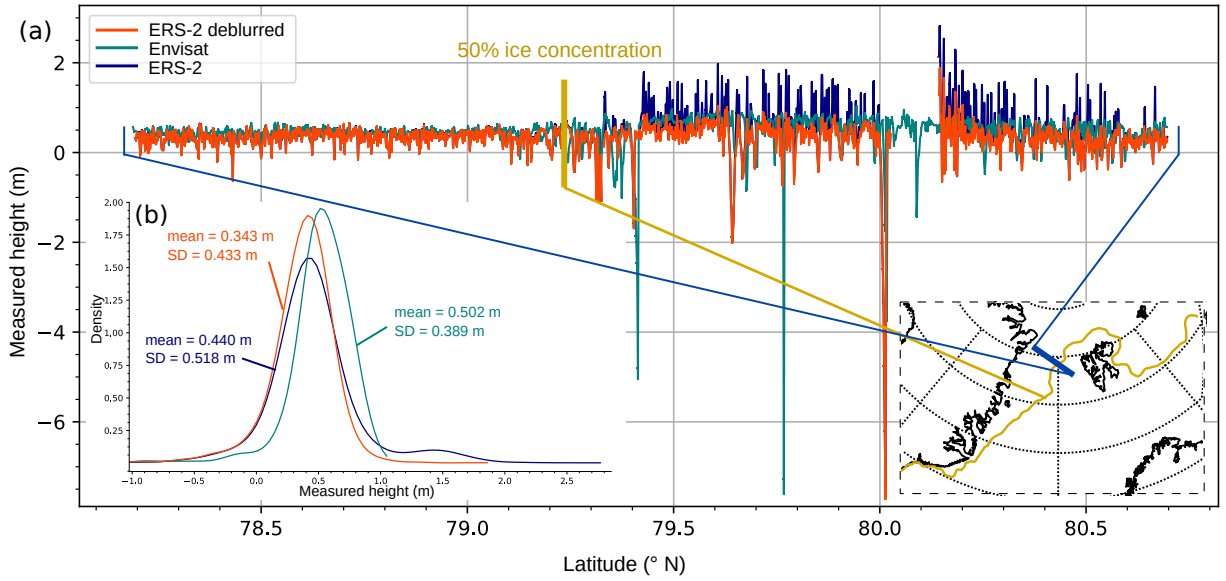


Figure 2. Profiles of surface height anomaly over sea ice and ocean for pass 25 between 78° N and 81° N for Envisat in blue-green (cycle 12), ERS-2 in blue and ERS-2 deblurred in orange (cycle 80). The red line represents the limit of 50 % concentration of sea ice, so as the limit between open ocean and an ice-covered area. The dark blue line shows the location of the pass between Svalbard Island and Greenland. (a) The surface height along the latitude and (b) the probability density function of surface height for the three passes with the associated statistics, the average and the standard deviation (SD). The color legend is identical for both sub-figures.

L215. Sure, but how much are they improved quantitatively if we are using Envisat as the reference ?

Pulse Blurring has to be seen as an asymmetric noise. As the FB or SLA processing are mainly form by statistical operations, succession of smoothing or interpolations, this noise will be reduced, some outliers due to the blurring will be removed. The resulting impact of blurring will be a positive non-constant bias on the SLA or on the FB, nevertheless the comparison with Envisat is not as easy because both missions (even if the Radar altimeter instrument is identical) are biased, so ERS ASA averaged over the basin with or without blurring compared to Envisat won't be so much relevant. However, we can see the impact of the deblurring on the noise of the data. We thus suggest to compare the standard deviation of ASA within each grid cell of a 12.5km resolution grid between Envisat (cycle 12) and ERS (cycle 80) before and after deblurring as an appendice (A1), see Fig.3.

The deblurred surface anomalies of ERS-2 now appear similar to Envisat.

replaced by :

The deblurred surface height anomalies of ERS-2 now appear more similar to Envisat in terms of noise and amplitude of variation. For this particular track, the standard deviation has been reduced by 16 % and get closer to Envisat's one. Figure A.1, shows more results on the impact of deblurring on ASA noise reduction comparing to Envisat during a whole cycle.

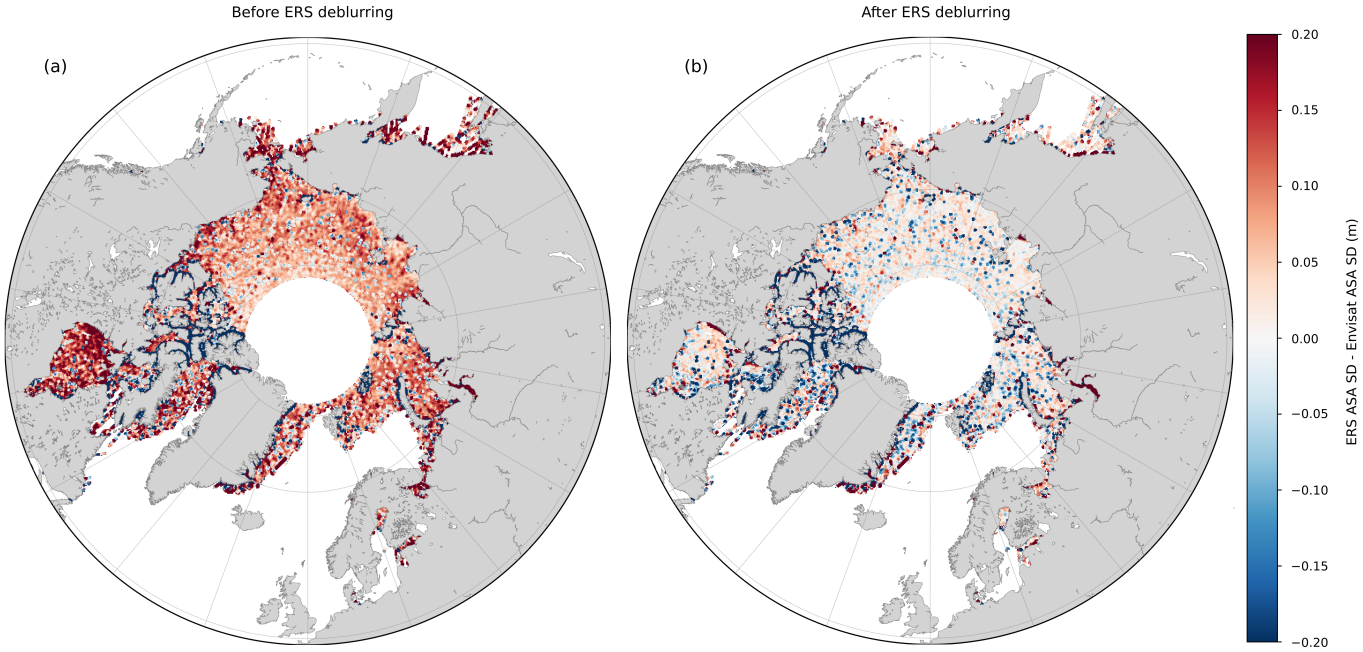


Figure 3. Comparison of the standard deviation of ASA between Envisat (cycle 12) and ERS (cycle 80) before (a) and after deblurring (b) within each grid cell of a 12.5km resolution grid. Median of the standard deviation difference for (a) is 0.075m and 0.008 m

---

L226-227. Can you add a table of the thresholds after they have been calculated to keep lead/floe proportions the same during overlap periods ? This would aid the repeatability of the study.

We recompute the threshold only for ERS-2, considering Envisat as a reference. Thank you for this suggestion, the clarification has been made in section.The following table : Tab.1 has been added in appendices (A1) in the manuscript.

Table 1. Pulse peakiness thresholds for lead/floe classification

Mission (RA mode)	PP lead threshold	PP floe threshold
CryoSat-2 (SAR)	0.3*	0.1*
Envisat (LRM)	0.3*	0.1*
ERS-2 (LRM)	0.2839	0.1328

* Guerreiro et al 2017

L236-237. More information is required on the interpolation method and procedure.

As suggested, we have added some precision to the interpolation procedure.

L236-237 : Outliers are filtered out with a three standard deviation threshold along 25 km sliding windows. ILA and SLA are interpolated respectively over leads and floes and smoothed with a 25 km rolling mean. The difference between the measured height over floes and over leads is finally made to retrieve the radar freeboard. For the remains of the study, we will only use the FBr measurements made above the floes because the characteristics of the waveforms are used.

replaced by:

ILA and SLA outliers are removed by filtering data that are outside the interval : rolling mean $\pm$ 3 rolling Standard Deviation, with a 60 km large sliding window. After filtering, ILA and SLA are smoothed using a rolling mean at 12.5 km, then SLA and ILA are linearly interpolated (including bellows the floes for SLA and above the leads for ILA) and are again smoothed using a rolling mean at 12.5 km. No limit of distance is used to discard radar freeboard, but the interpolation as well as smoothing and filtering is not done with values separated by land. Indeed, the processing is done within ocean segments, separated by land, in order to isolate statistics between segments. In this study, we will only use the FBr measurements that are made over the floes, indeed, the LRM data calibration, explained in section 3.4, is based on floes characteristics.

L237-238. Do you discard rFBs above a max distance to the nearest lead? If so what limit do you use?

No, we don't discard freeboard above a max distance to the nearest lead. This information has been added commonly within the previous comment. To do so, we have applied the usual geophysical corrections in altimetry, taking into consideration the choice of these data so that they are particularly appropriate for polar oceans.

L250. Can you explain a little more about this constant SLA bias in LRM? Why does it appear and what could be done, in theory, to remove it?

The SLA bias comes from the choice to use an empirical retracker with the same fixed threshold for both leads and floes. Poisson et al 2018, explains that over rough surfaces such as ocean, the usual retracked point is close to the position of the half power of the waveforms. Specular waveform that can be found over leads should have retracked point higher, nevertheless, measurements are more stable with lower threshold (Poisson et al 2018 fig.9). This explains why we have a bias for the SLA.

Laforge et al 2020 shows that over leads comparing to physical retracker, the SLA bias is constant for altimeters in SARM, nevertheless this conclusion is also relevant for LRM as peaky waveform are all the same over leads. To remove this bias, another threshold should be use e.g. 80 or 90 or a calibration over another mission such as CS-2. The choice over a higher threshold can introduce noise due to the power sampling of the waveform. It could be noticed that this bias also occurs for mission in SAR mode, but it is much smaller due to the fact that SAR waveforms on leads are more peaky than for LRM.

L248-251. In LRM, most of this error comes from a constant bias on the Sea Level Anomaly

replaced by

Negative radar freeboards are mainly due to the retracker choice. Indeed, a TFMRA50 is used to retrack heights on both leads and floes, this introduces a bias on the height over leads. The TFMRA threshold to retrack heights over leads should be closer to 80% and the use of a 50% threshold corresponds to the position of the retrack point for ocean surfaces, not specular ones (Poisson et al 2018), the surface over leads is measured to be higher than it is and even higher than the surface over floes. The SLA bias (over leads) is evaluated constant for SARM altimeter in the study of Laforge et al 2020, this conclusion is also relevant for LRM altimeters as waveforms over leads are peaky and similar from a lead to an other. This positive constant bias over leads results to a negative bias on the radar freeboard. To avoid this bias, the retracker threshold could be adapted for leads or the SLA could be calibrated on CryoSat-2 one. Nevertheless, a threshold of 50% ensure the stability of the range (Poisson et al 2018, Fig.9) contrary to higher thresholds (80%-95%) that could lead up to 47 cm of random error on the SLA. A TFMRA at 50 % for both leads and floes is preferred in this study as a constant bias is easier to correct than an undetermined random error.

Figure 4. Add the sensing mode to the plot. The CS2 data here is SAR mode right, not calculated from pLRM?

Yes, CS-2 data is in SAR+SARIn mode here. This precision has been added to the plot (cf 4)

L278. Explain these terms.

The activation function is a sigmoid.

replaced by:

The activation function for the hidden layers neurons is a sigmoid, motivated by possible negative radar freeboard values and the optimizer is and ADAM [Kingma et Ba, 2014].

L283-284. What does this mean? Retrained again or just some sort of tuning? Might it be very different from the training with 90-10 split?

Once the hyper-parameter combination is set, the MLP is trained with the whole dataset" means that the chosen model (with the optimal combination) is trained 'again' with 100% of the data (again, but weight are reinitialized). The training on 90% comparing to the one on 100% are not slightly different at all, but it's supposed to be better because with have trained it with a larger dataset (larger the dataset is, better the model is supposed to be). For instance, as we have only one winter for Envisat-ERS-2 mission overlap period, 10% is not negligible as it represents a bit less than a month. We have clarified this part, hope it would help the reader to understand better this part of the methodology.

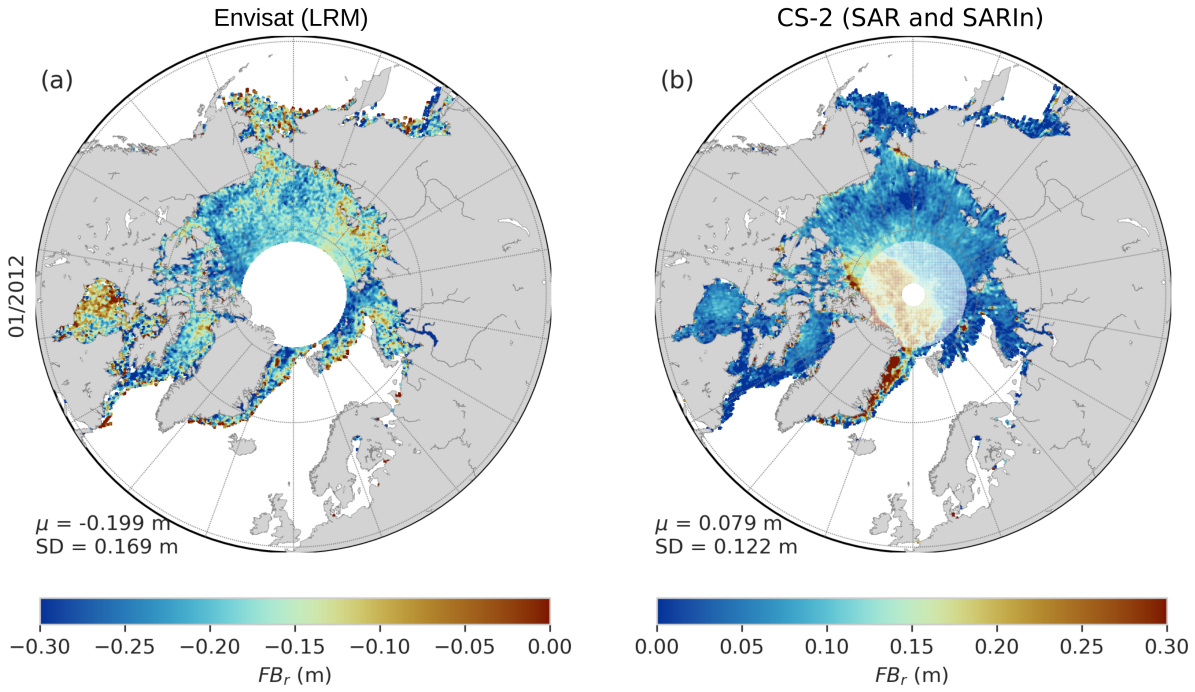


Figure 4. Pan-Artic radar freeboard maps for January 2012 for (a) Envisat uncorrected and (b) CryoSat-2

L283-284 : Hyper-parameters have been tuned by dichotomy by choosing at each step the hyper-parameter combination with the highest mean score (average score made on 5 models) on the test sample. The score used for this regression is the Pearson correlation coefficient. To determine the most suitable hyper-parameter combination, the dataset is randomly split into a training and a testing dataset, corresponding respectively to 90% and 10% of the initial dataset. The activation function used is a sigmoid. Once the hyper-parameter combination is set, the MLP is trained with the whole dataset.

replaced by:

The neural network used is a multilayer perceptron (MLP). Both calibrations have been processed with Scikit learn [Pedregosa et al, 2011]. The MLP is composed of 5 hidden layers, each composed of 100 neurons. The choice of hyperparameters : number of neurons, the learning rate, the regularization term, batch size, activation functions, solver for the weights optimization, have been done using gridding methodology, e.g. testing combinations and take the one that give best score. The evaluation criterion, called the score, is chosen as the determination coefficient. Models are trained on 90% of the dataset and tested on the remaining 10%, the splitting in random. During the tuning step, models are cross validated, it means that they are each trained 5 times with the same combination of hyperparameters but without the same train/test dataset, the 5 scores are then analyzed to determine the best combination. Cross validation give a better idea of the model performance as the dependence to the training dataset is limited. The activation function for the hidden layers neurons is a sigmoid, motivated by possible negative radar freeboard values and the optimizer is and ADAM [Kigma et Ba, 2014]. 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.

Finally, once the hyperparameters combination is set, the MLP is trained on the whole dataset to provide the calibration function. The trained model is then applied to the LRM monthly grids to obtain a monthly LRM-corrected radar freeboard.

L301-303. Needs more info. Why do you calculate uncertainty differently between leads and floes? The uncertainty at floes is governed by the variability in height measurements at proximal leads.

What distance is used to calculate an along-track mean elevation? Is the variability in individual floe height obs around this not just a measure of the topography? It will be higher over MYI but does this realistically mean the uncertainty is higher?

We compute differently the SLA uncertainty for the floes (where we interpolate the SLA). Indeed, the rolling standard deviation of the interpolated SLA (where we do not have leads) does not really make sense to estimate the SLA uncertainty on floes. That's the reason why the uncertainty of the SLA where there are no leads is different and is assumed to be the difference between the estimated (interpolated and smoothed SLA) and the mean SLA, this method is taken from Ricker et al 2014. Concerning the along track mean elevation, no limit of distance was used, nevertheless, such as for the interpolation or the smoothing (section 3.2), they are computed per segments of water (so there is one mean SLA per segment of water). As we consider values on floes, we assume that the main part of the random uncertainty of the ILA (ice level anomayl, above the floes) comes from the speckle noise, even if ILA is smoothed, this could be a limitation of the random uncertainty budget. If we understood well the last part of your comment, this will not be the topography of the ice, but the difference between the SLA we have and the mean SLA, which is supposed to represent the SLA we should have. We suggest the following modification in the manuscript, with more details.

SLA uncertainty (σ_{SLA}) estimation depends on the surface type (leads or floes). For leads, we take the standard deviation of the measured height within a sliding 25 km window. Concerning floes, the uncertainty is estimated as the difference between the height measurement and the mean elevation along the track (Ricker et al., 2014).

replaced by :

SLA uncertainty (σ_{SLA}) is estimated to be the standard deviation of the SLA within a sliding window of 25 km if there are some leads within this window. If not, the SLA uncertainty is taken as the difference between the interpolated and smoothed SLA and the mean SLA computed as the mean values of measured SLA at leads within a segment of ocean (if the pass is over land, statistics are made segment of ocean by segment of ocean).

L307. How are the uncertainties reduced during gridding? Speckle noise should drop as a function of N observations, but SLA uncertainty should only drop as a function of N tracks (because SSH error is highly correlated along track).

The way the random uncertainties are computing during the gridding are slightly more complicated than that. We have done the same choice as in Ricker et al, 2014 and compute the grid-cell resulting radar freeboard with a weighted average, with the radar freeboard uncertainty as weights. For each grid-cell : $FB_r = \sqrt{\frac{\sum_1^N \frac{1}{\sigma_{R_i^2}} \cdot FB_{ri}}{\sum_1^N \frac{1}{\sigma_{R_i^2}}}}$*, so the resulting uncertainty computation should take it into account that weighted average as* $\sqrt{\frac{1}{\sum_1^N \frac{1}{\sigma_{R_i^2}}}}$*. Taken this method, it's difficult to draw a way to make uncertainties for the speckle noise and the SLA dropping differently, even we agree that theoretically the* σ_{SLA} *is correlated along the tracks, thus, the random uncertainty can be a little bit underestimated.*

L312. Systematic uncertainty due to roughness is 20-30% of the freeboard as well as of the thickness.

Thank you for this confirmation. We adapted the sentence to the radar freeboard.

Roughness is estimated to be respectively about 20% and 30% of the sea ice thickness for FYI and MYI (Landy et al., 2020).
replaced by :
Roughness is estimated to be respectively about 20 % and 30 % of the sea ice thickness for FYI and MYI according to Landy et al., 2020, this results is also applicable for the freeboard

---

L316-318. Based on the schematic in figure 6 everything you've done seems fine, but it is still confusing to follow all the steps. What are these 'other inputs'? And which variables do you divide by the sqrt of the number of observations vs the sqrt of the number of tracks when gridding?

We apologize for the confusion raised by this figure. The other inputs are the one in the grey box in the top right of Fig 6, and detailed in Fig.5, (except the date) : the concentration, the Leading edge slope, the Pulse Peakiness, and the Multi-year ice fraction. The uncertainty of these 4 variables is defined as : two times the standard deviation of the values of these variables within the grid cell divided by the sqrt of number of tracks within this grid cell. The uncertainty of the radar freeboard, as said in Sec 3.5, is neither multiplied by two nor divided by the number of tracks. In order to make this steps clearer, we have developed the legend of Fig.5 and Fig.6. and modified the following paragraph :
L316-317 : The uncertainty of the other inputs is considered to be, for each grid cell, the standard deviation of the measurements used to calculate the average value (grid cell value) divided by the number of tracks passing through the corresponding grid cell.

replaced by:
The uncertainty of the other inputs (LES, PP, sea ice concentration, MYI fraction), is considered to be, for each grid cell, two times the standard deviation of the measurements used to calculate the average value (grid cell value) divided by the number of tracks passing through the corresponding grid cell.

As well as :

L325-327 : The method consists of training a number M of NN with noisy inputs (noise has been added to all inputs according to a Gaussian distribution), and then to analyze the distribution of radar freeboard predictions from the M noisy NN applied on noisy N inputs for each considered grids. The whole uncertainty budget process is summarized in Fig. 6.

replaced by :

The method consists in training a number M of NN with noisy inputs. The noise has been added to all inputs (for each grid cell and each month) according to a Gaussian distribution centered on the estimated value and the corresponding uncertainty as standard deviation. The calibration processing (training and prediction) is done for all the noisy inputs/output, then the distribution of MxN radar freeboard predictions (from the M noisy NN models applied on N noisy inputs) has been analyzed of each grid cell and each month. The whole uncertainty budget process is summarized in Fig. 6.

L325-329. How do you estimate the gaussian noise distribution statistics? is this the $sigma = 2 * sigma_{omega}$ in Figure 6? The output from a monte carlo error budget depends closely on the assumptions taken for the error distributions so this is important.

We hope we have clarified this information with the modifications made for the previous comment.

L340. For which months in Fig 8?

This missing information has been added to the manuscript.

Figure 8 presents the same feature for Envisat and ERS-2 radar freeboard during December 2002 and April 2003.

L357. What are these numbers as a % of the mean rFB?

We have added this information in the manuscript.

The higher mean difference is 7 mm and concerns February 2011, the Envisat calibration. For the ERS-2 calibration, the mean freeboard difference with Envisat does not exceed 3 mm. Concerning all the overlap times, the mean difference is 3 mm for Env/CS-2 calibration, and -2 mm for ERS-2/Env one.

has been replaced by :

The highest mean difference reaches 7 mm in February 2011 for Envisat calibration, i.e. 9.5% of the mean Envisat radar freeboard. For ERS-2 calibration, the mean freeboard difference between ERS-2 and Envisat does not exceed 3 mm, 3.3% of ERS-2 mean radar freeboard. Concerning all the overlap times, the mean difference is 3 mm for Env/CS-2 calibration, 4.1% of Envisat mean radar freeboard and -2 mm for ERS-2/Env one, about 2.2% of ERS-2 mean radar freeboard.

L359. Again what are these as a % of the mean rFB?

This information has been added to the previous paragraph (see previous comment).

L363. Can you do the same for the ERS2-Envisat comparison?

Yes, this information has been added as following :

Similarly, for the period 2002/2003, the mean and median uncertainties of ERS-2 are always larger than those of Envisat by about 6 cm over the median radar freeboard (see Tab.A5 and Tab.A3 for more statistics)

has been replaced by :

Concerning the period 2002/2003, the median uncertainty is 8 cm for ERS-2 radar freeboard and 7.3cm for Envisat, similarly, statistics on uncertainties are globally higher for ERS-2 estimates (see Tab.A3 and Tab.A4 for detailed statistics).

Figure 7b. It looks like you may have some spurious tracks in Hudson Bay, Baffin Bay and Bering Strait that could contaminate the comparisons?

Unfortunately, in spite of the different filtering, spurious tracks remains, especially close to the coast. Nevertheless, few validation data are available in these areas, so comparisons with independent data sets would not be contaminated with these spurious tracks.

Figure 7 caption. Emphasize the distributions include CS2 data only for the coinciding region south of 81.5N.

Figure 7 caption has been modified as followed :

Comparison of Envisat calibrated radar freeboard against CryoSat-2 reference for December 2010 in the upper half and April 2011 in the lower half. The maps (a) and (g) refers to Envisat aside with corresponding CryoSat-2 radar freeboard (b) and (h). Maps bellow (d), (e), (j) and (k) are the related uncertainties. The right column presents differences freeboard maps (Env-CS-2) ((c) and (i)). (f) and (l) are the distribution of Envisat F Br in red, CryoSat-2 F Br in blue and in grey.

has been replaced by

Comparison of Envisat calibrated radar freeboard against CryoSat-2 reference for December 2010 in the upper half and April 2011 in the lower half. The maps (a) and (g) refers to Envisat aside with the corresponding CryoSat-2 radar freeboard (b) and (h). Maps bellow (d), (e), (j) and (k) are the related uncertainties. The right column presents freeboard difference maps (Env-CS-2) ((c) and (i)). (f) and (l) are the distribution of Envisat FBr in red, CryoSat-2 FBr in blue and ΔFBr in grey. Histograms only include common data between Envisat and CryoSat-2, data north or 81.5 °N are excluded. μ refers to the average and SD to the Standard Deviation.

L384. 'static data'?

'Static data' refers to moorings, data set have fixed longitude and latitude for a given time. We have clarified this point as following:

Static data are monthly averaged to get one value per month.

replaced by :

Concerning moorings or coastal measurement stations, data are averaged to get one value per month.

L392. I think it is reasonable to discount IMBs because they represent only the single floe they are deployed on (usually a thicker floe) and not their surrounding 12x12 km grid cell area. The authors could remove these comparisons so they don't draw reader's attention and they come to the wrong conclusions about the satellite data validity; but that is up to the authors.

Thank you for this suggestion. We agree with this point of view. This comparison was a part of the validation for completeness, but also to get feedback from the community as we were struggling while using this data. To remove the confusion, we decided not to compares Envisat data with IMB as advised.

L397. Could be attributed to, but not definitely.

We have changed this sentence to stay more hypothetical.

The bias between OIB and Envisat estimation can also be attributed to the OIB snow depth estimation that remains slightly different from one algorithm to another (Kwok et al., 2017).

replaced by :

The bias between OIB and Envisat estimation could also be partly attributed to the OIB snow depth which estimation seems sensitive to the algorithm used (Kwok et al., 2017).

Figure 9 and elsewhere. Define the acronyms of statistical tests in the caption.

Captions of Figure 7,8,9,10,11,12 have been modified to add the statistic acronyms definitions.

L406-407. Can this say anything about the calibration? Are the BGEP ice conditions more representative of average sea ice conditions in the Arctic and the other ULS datasets more of thin ice conditions? Was the calibration not slightly overestimating thin ice thickness for Envisat?

This could tell something about the calibration, but it's hard to conclude knowing that BGEP comparisons and other moorings comparisons draw different conclusions. Considering BGEP, calibration seems to overestimate thin ice which is not the case comparing to moorings within the Laptev sea.

L415-416. How do these numbers compare to your estimated uncertainties for the same regions?

We don't really know how to compare properly the Standard deviation and biases with our uncertainties as the question of the uncertainties on the SIT is another issue and for this paper the uncertainties are limited to the radar freeboard. We propose the Figures 10,11,12 and 13 that present the 95% confidence interval of the SIT but without taking into account uncertainties on snow depth, densities etc for the FBr to SIT conversion step.

The plots have been updated with the variable snow density as asked by Robbie Mallet so they are not similar to those in the previous version of the manuscript. For esthetical reason, bounds are not represented for comparisons with other satellite-based SIT estimation.

L420-422. What are the statistics like for CS-2 data processed with this method? You don't necessarily need to show a plot, but some idea of biases would be useful. Do you also see generally negative biases for CS2? Especially over FYI?

Just for information, Figure 9 show comparisons for CS-2 thickness with OIB, BGEP and Transdrift Laptev Sea. For BGEP bias is higher, CS-2 FYI thickness seems to be more overestimated than Envisat one but with find this same negative bias for Transdrift Laptev Sea moorings especially for thick ice.

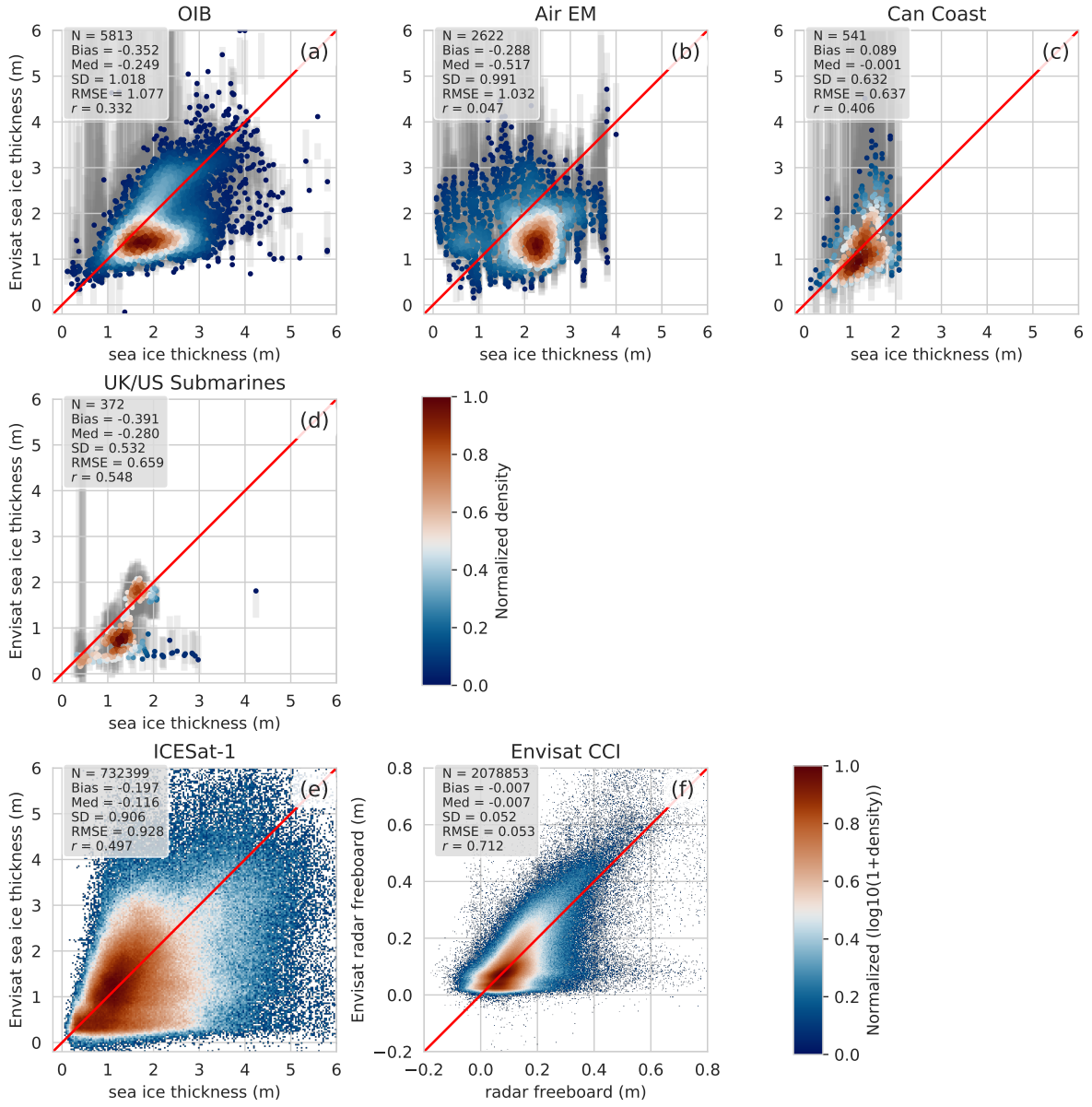


Figure 5. Comparative scatter-plots between Envisat sea ice thickness or radar freeboard estimations and other data sets. The x-axis indicates the sea ice thickness from (a) OIB total ice freeboard, (b) Air EM snow plus ice thickness, (c) Can Coast ice thickness, (d) UK/US submarines draft and (e) ICESat-1 total freeboard. (f) compares our Envisat radar freeboard with SI-CCI Envisat solution. Colorbars represent the normalized density. A log_{10} has been applied before the normalization for (e) and (f) due to the large number of data. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

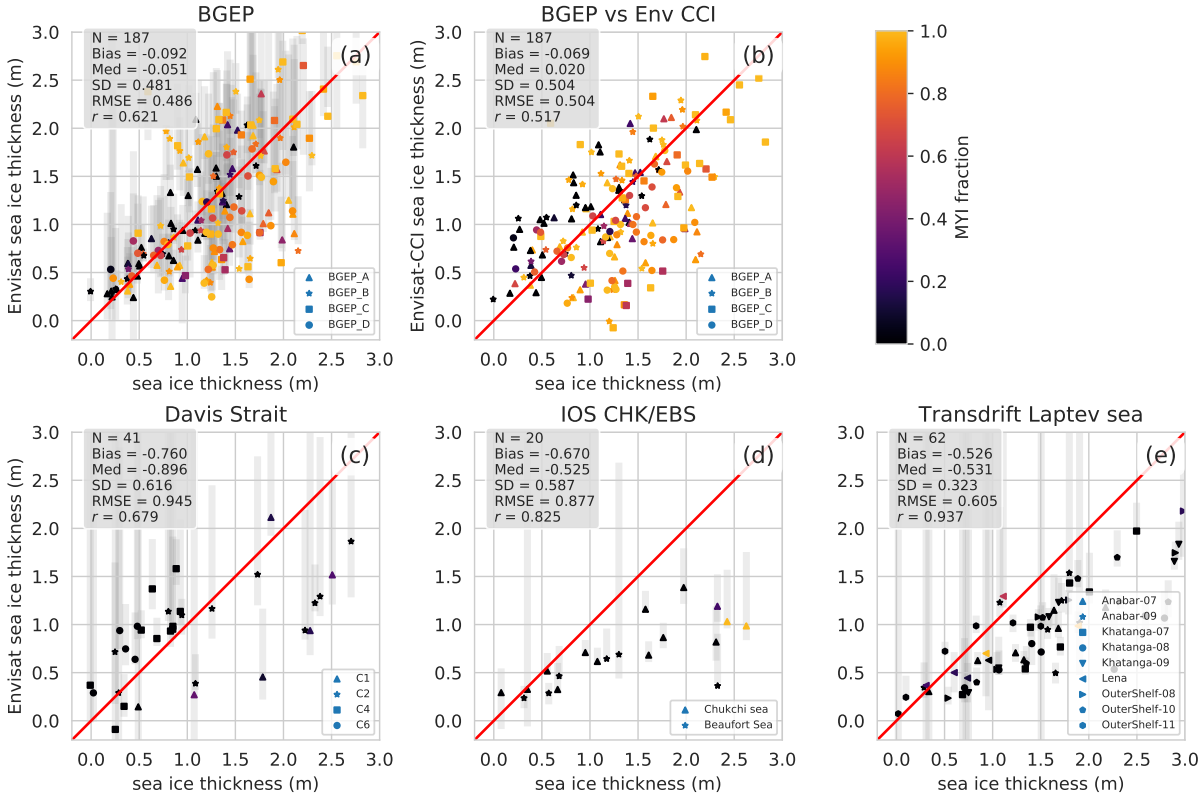


Figure 6. Comparative scatter-plots between Envisat sea ice thickness estimations and anchored moorings data sets. Each dot corresponds to a monthly averaged value. The x-axis indicates the sea ice thickness from (a) BGEP, (b) BGEP vs Env CCI, (c) Davis Strait, (d) IOS CHK/EBS and (e) Transdrift Laptev Sea ice draft. The colorbar shows the MYI fraction. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

In order to help the reader, with have added the following sentences:

As a comparison, the bias between CryoSat-2 and OIB between 2010 and 2019 is about 16 cm and the RMSE 77 cm. Concerning BGEP (2010-2021) comparisons, the bias is 21 cm with the same overestimation of FYI thickness for CS-2 and with Transdrift Laptev Sea (2010-2016) comparisons show a negative bias of -38 cm.

L425-426. Could you try them also with the adapted warren climatology and see if biases get any smaller? Would help to clarify the impact of snow loading.

The same plots with Warren 99 modified climatology are presented in Fig.10, Fig. 11, Fig. 12 and Fig. 13. Note that the scatters for OIB, CanCoast and EnvisatCCI are unchanged because no additional snow products are used. Biases with Air EM is reduced as well as for the submarines. Nevertheless, as explained in the manuscript, a bias is expected when comparing to Submarines of about 30 cm. The dispersion is augmented for the comparisons with moorings data. Thus, the snow load has an important impact, we suggest showing this figure in appendice, as it could impact the clarity of validation if it is shown in the validation section.

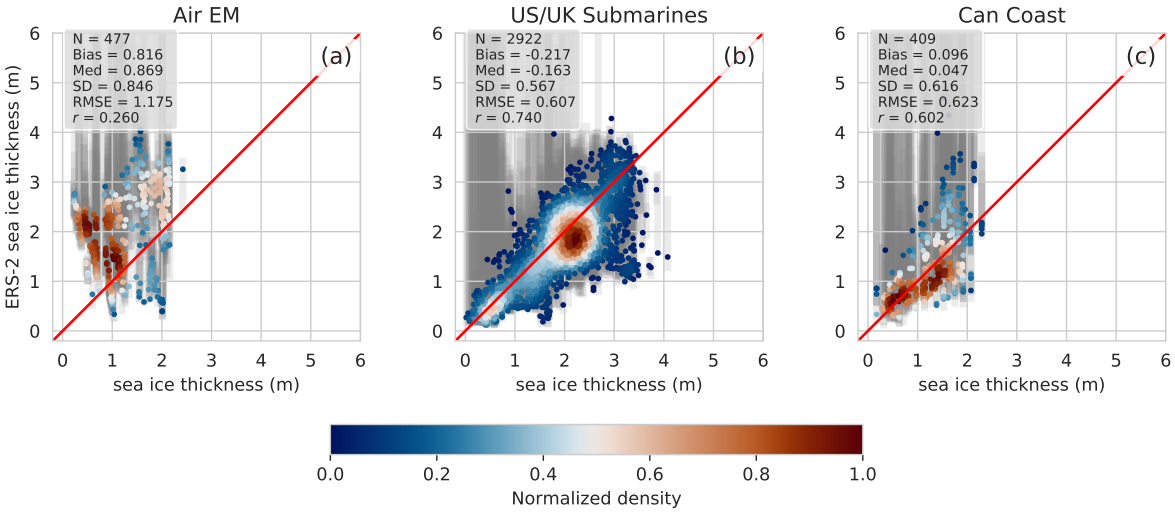


Figure 7. Comparative scatter-plots between ERS-2 sea ice thickness estimations and 3 in-situ data sets. The x-axis indicates the sea ice thickness from (a) AirEM total thickness, (b) UK/US Submarines draft and (c) Can Coast sea ice thickness. Colorbar indicates the normalized density. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

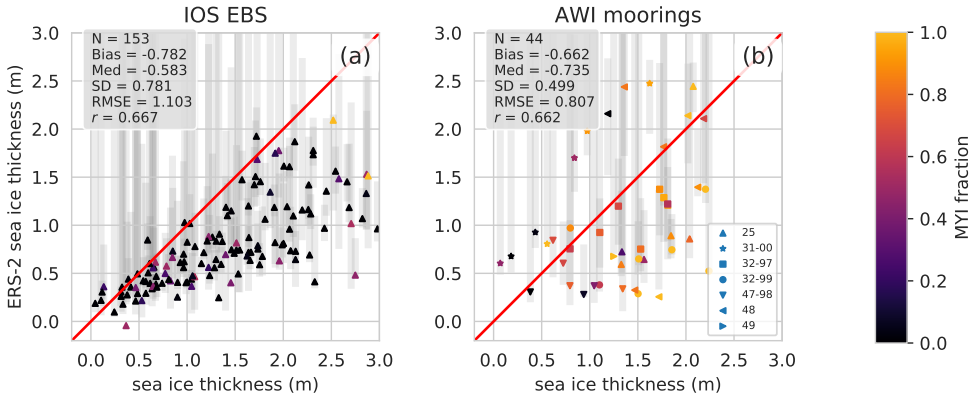
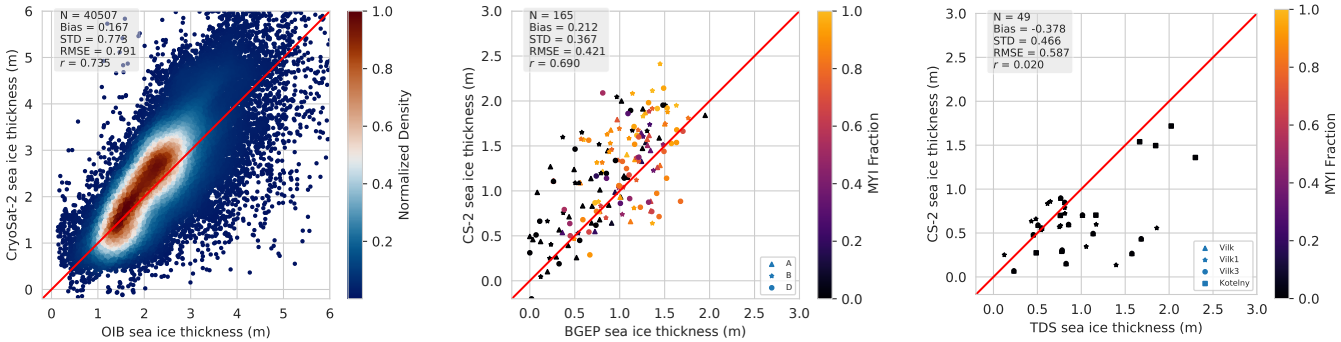


Figure 8. Comparative scatter-plots between ERS-2 sea ice thickness estimations and 2 anchored moorings data sets. The x-axis shows sea ice thickness estimations from (a) IOS Beaufort Sea and (b) AWI moorings sea ice draft. The color bar indicates the respective MYI fraction. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.



(a) OIB (b) BGEP (c) Transdrift

Figure 9. Comparative scatter-plots between CryoSat-2 sea ice thickness estimations and OIB/BGEP. The x-axis indicates the sea ice thickness from (a) OIB total thickness, (b) BGEP draft and (c) Transdrift Laptev Sea draft. Colorbar indicates the normalized density and MYI fraction. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

L427. Is Section 2.3 correct?

Sorry for this typo mistake, the reference was indeed "section 4.2". This has been corrected.

L441. It is worth making it a bit clearer on Fig 13 and throughout this section that these volumes miss out everything >81.5N.

The caption has been clarified as following :

Fig 13 caption : Time series representing radar freeboard volume up to 81.5°N for each winter month for ERS-2 in orange, Envisat in teal and CS-2 in dark red. Blue triangles are winter mean volumes. Red lines are linear regressions of winter mean volume until 2002/2003 for dashed line and 1995/1996 for solid line.

replaced by:

Fig 13 caption : Time series representing radar freeboard volume up to 81.5°N for each winter month (no data between 81.5 °N and 90 °N, even for CS-2 for consistency). ERS-2 in orange, Envisat in teal and CS-2 in dark red. Blue triangles are winter mean volumes. Red lines are linear regressions of winter mean volume until 2002/2003 for dashed line and 1995/1996 for solid line.

addition of : It's important to note that the volumes presented in Fig. 13 are only considering values up to 81.5 °N.

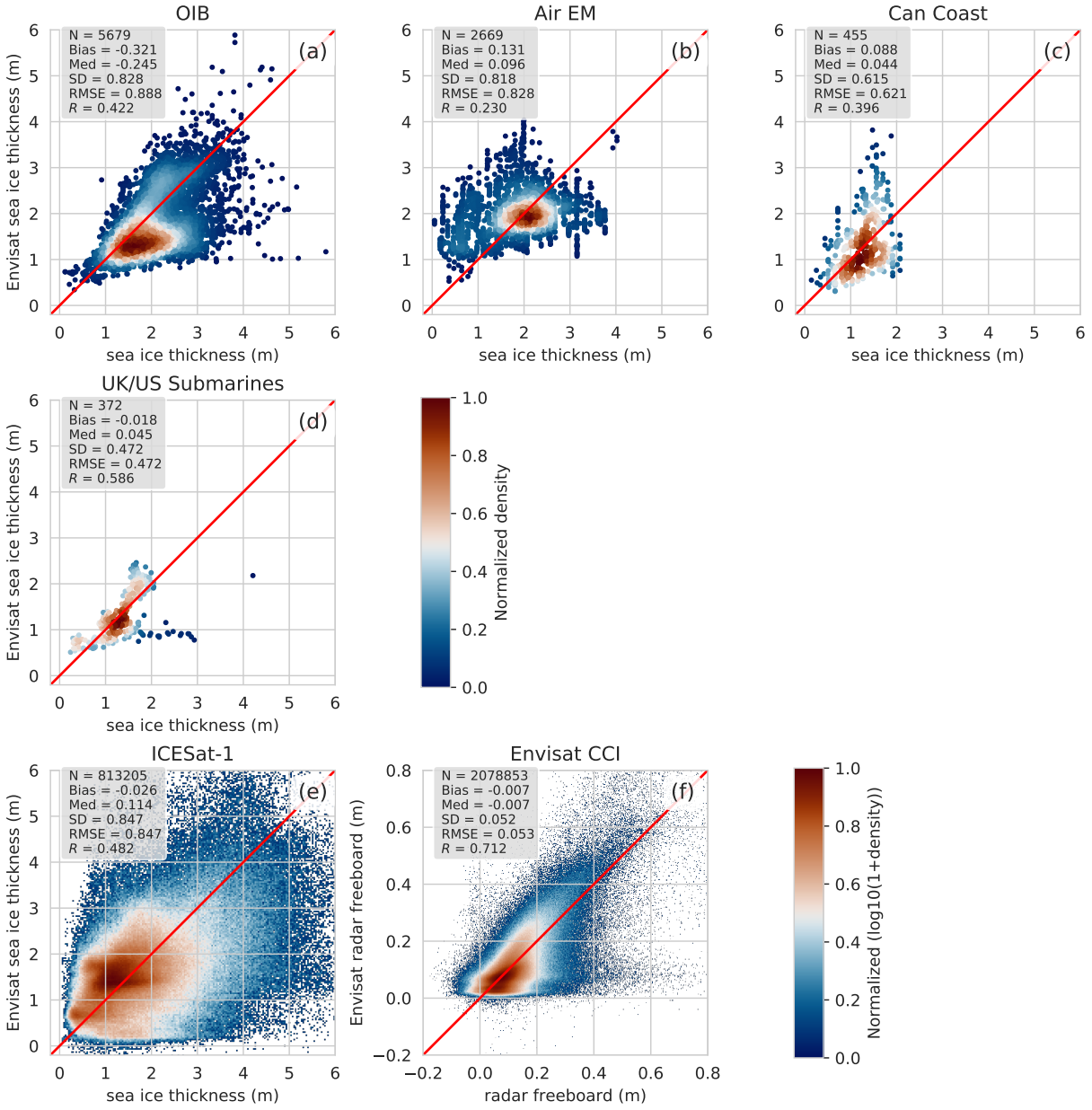


Figure 10. Comparative scatter-plots between Envisat sea ice thickness or radar freeboard estimations and other data sets. The x-axis indicates the sea ice thickness from (a) OIB total ice freeboard, (b) Air EM snow plus ice thickness, (c) Can Coast ice thickness, (d) UK/US submarines draft and (e) ICESat-1 total freeboard. (f) compares our Envisat radar freeboard with SI-CCI Envisat solution. Colorbars represent the normalized density. A log_{10} has been applied before the normalization for (e) and (f) due to the large number of data. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and R the correlation coefficient.

Figure 13. I think readers would find it interesting to see more of your rFB dataset. I'd suggest an additional figure showing trends in rFB as a map for the overlap region, highlighting where the trends are significant or not.

We apologize to show so limited data set, we feel that the paper is already rather long as it is and this will be the subject of another study.

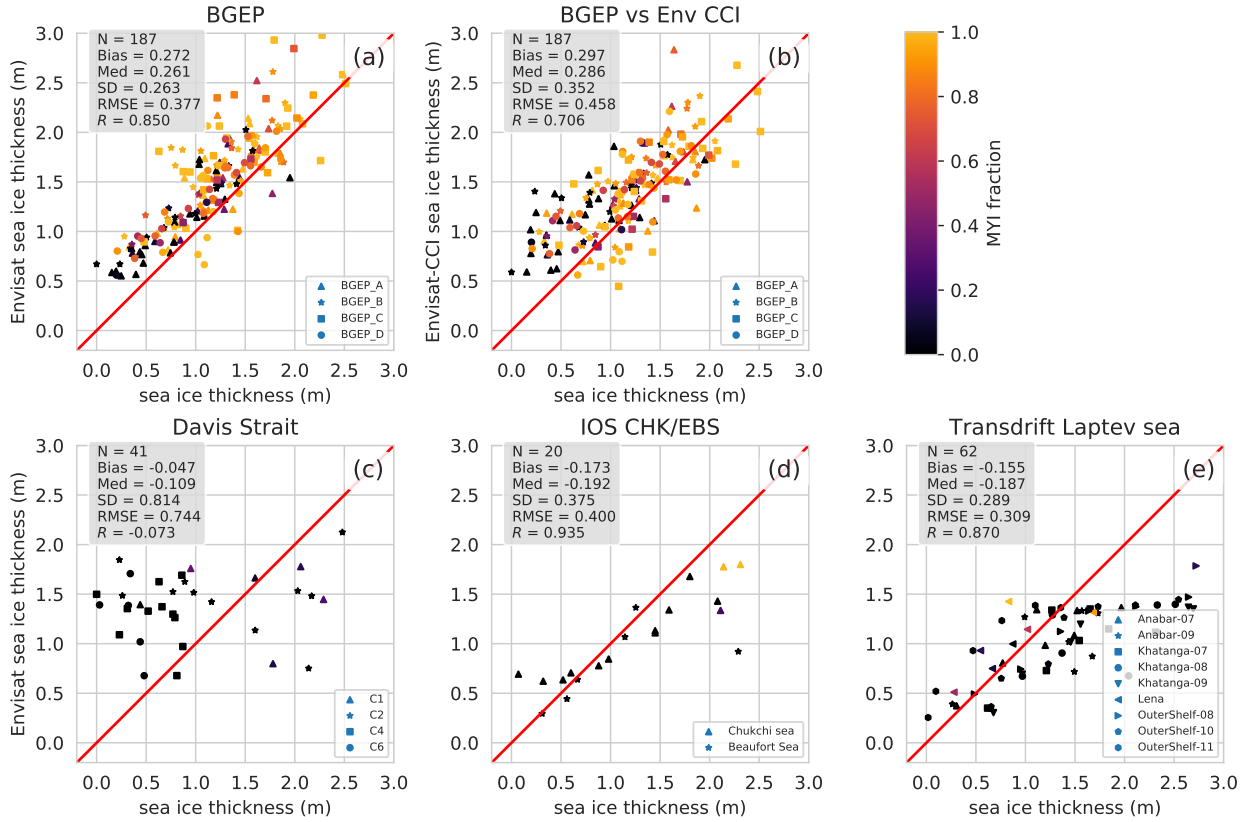


Figure 11. Comparative scatter-plots between Envisat sea ice thickness estimations and anchored moorings data sets. Each dot corresponds to a monthly averaged value. The x-axis indicates the sea ice thickness from (a) BGEP, (b) Davis Strait, (c) IOS CHK/EBS and (d) Transdrift Laptev Sea ice draft. The colorbar shows the MYI fraction. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and R the correlation coefficient.

Table A1. SAR you mean? or is this actually the CS2 LRM mode? I think SAR was used here right so state SAR parameters?

It's actually the pLRM mode, this has been explained in a previous answer, we hope that the use of this table is now clearer.

Guerreiro, K., Fleury, S., Zakharova, E., Kouraev, A., Rémy, F., and Maisongrande, P.: Comparison of CryoSat-2 and ENVISAT radar freeboard over Arctic sea ice: toward an improved Envisat freeboard retrieval, The Cryosphere, 11, 2059–2073, https://doi.org/10.5194/tc-11-2059-2017, 2017.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, CoRR, 2014.

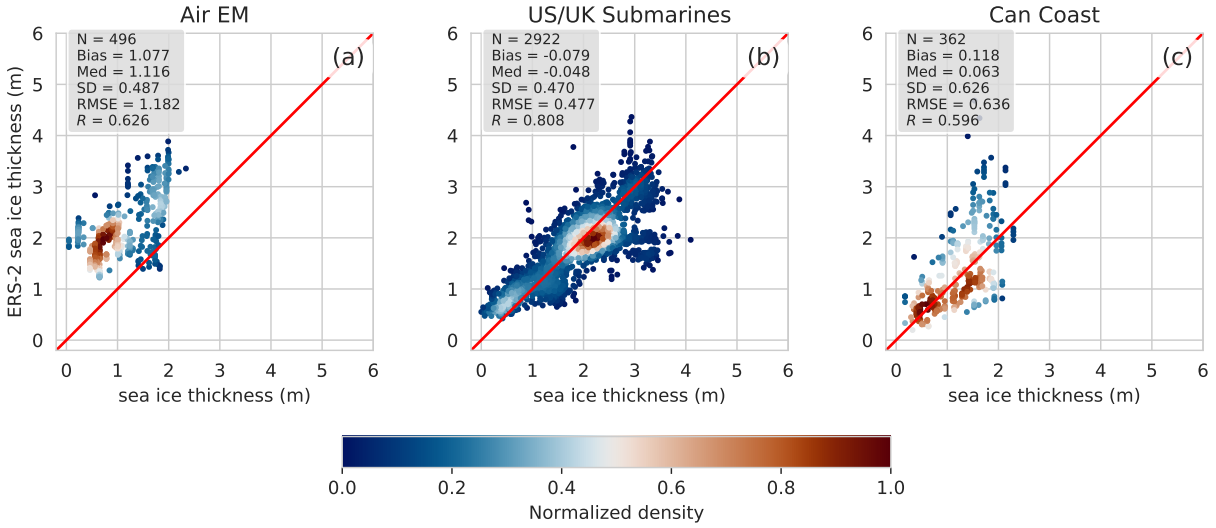


Figure 12. Comparative scatter-plots between ERS2 sea ice thickness estimations and 3 in-situ data sets. The x-axis indicates the sea ice thickness from (a) AirEM total thickness, (b) UK/US Submarines draft and (c) Can Coast sea ice thickness. Colorbar indicates the normalized density. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and R the correlation coefficient.

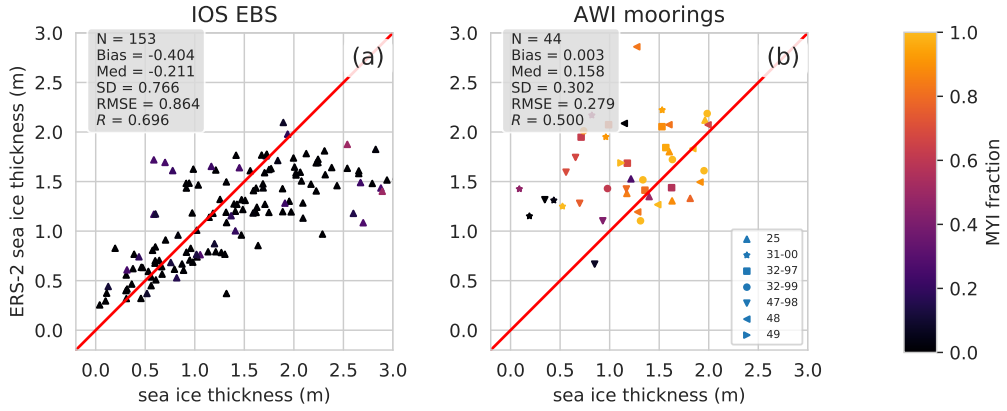


Figure 13. Comparative scatter-plots between ERS2 sea ice thickness estimations and 2 anchored moorings data sets. The x-axis shows the sea ice thickness measurements from (a) IOS Beaufort Sea and (b) AWI moorings sea ice draft. The color bar indicates the respective MYI fraction. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and R the correlation coefficient.

Laforge, A., Fleury, S., Dinardo, S., Garnier, F., Remy, F., Benveniste, J., Bouffard, J., and Verley, J.: Toward improved sea ice freeboard observation with SAR altimetry using the physical retracker SAMOSA+, Advances in Space Research, p. S0273117720300776, https://doi.org/10.1016/j.asr.2020.02.001, 2020.

Landy, J. C., Petty, A. A., Tsamados, M., and Stroeve, J. C.: Sea Ice Roughness Overlooked as a Key Source of Uncertainty in CryoSat-2 Ice Freeboard Retrievals, Journal of Geophysical Research: Oceans, 125, https://doi.org/10.1029/2019JC015820, 2020.

Landy, J. C., Dawson, G. J., Tsamados, M., Bushuk, M., Stroeve, J. C., Howell, S. E. L., Krumpen, T., Babb, D. G., Komarov, A. S., Heorton, H. D. B. S., Belter, H. J., and Aksenov, Y.: A year-round satellite sea-ice thickness record from CryoSat-2, Nature, 609, 517–522, https://doi.org/10.1038/s41586-022-05058-5, number: 7927 Publisher: Nature Publishing Group, 2022.

Nandan, V., Geldsetzer, T., Yackel, J., Mahmud, M., Scharien, R., Howell, S., King, J., Ricker, R., and Else, B.: Effect of Snow Salinity on CryoSat-2 Arctic First-Year Sea Ice Freeboard Measurements: Sea Ice Brine-Snow Effect on CryoSat-2, Geophysical Research Letters, 44, 10,419–10,426, https://doi.org/10.1002/2017GL074506, 2017.

Paul, S., Hendricks, S., Ricker, R., Kern, S., and Rinne, E.: Empirical parametrization of Envisat freeboard retrieval of Arctic and Antarctic sea ice based on CryoSat-2: progress in the ESA Climate Change Initiative, The Cryosphere, 12, 2437–2460, https://doi.org/10.5194/tc-12-2437-2018, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Poisson, J.-C., Quartly, G. D., Kurekin, A. A., Thibaut, P., Hoang, D., and Nencioli, F.: Development of an ENVISAT Altimetry Processor Providing Sea Level Continuity Between Open Ocean and Arctic Leads, IEEE Transactions on Geoscience and Remote Sensing, 56, 5299–5319, https://doi.org/10.1109/TGRS.2018.2813061, 2018.

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

Rheinlænder, J. W., Davy, R., Ólason, E., Rampal, P., Spensberger, C., Williams, T. D., Korosov, A., and Spengler, T.: Driving Mechanisms of an Extreme Winter Sea Ice Breakup Event in the Beaufort Sea, Geophysical Research Letters, 49, https://doi.org/10.1029/2022GL099024, 2022.

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, 8, 1607–1622, https://doi.org/10.5194/tc-8-1607-2014, 2014.

Stammer, D.: Satellite altimetry over oceans and land surfaces, Earth observation of global changes, 2018.

Stroeve, J. and Notz, D.: Changing state of Arctic sea ice across all seasons, Environmental Research Letters, 13, 103 001, https://doi.org/10.1088/1748-9326/aade56, publisher: IOP Publishing, 2018.

Tilling, R., Ridout, A., and Shepherd, A.: Assessing the Impact of Lead and Floe Sampling on Arctic Sea Ice Thickness Estimates from Envisat and CryoSat-2, Journal of Geophysical Research: Oceans, 124, 7473–7485, https://doi.org/https://doi.org/10.1029/2019 2019.

Tschudi, M., Meier, W. N., Stewart, J. S., Fowler, C., and Maslanikand, J.: EASE-Grid Sea Ice Age, https://doi.org/10.5067/UTAV7490FE type: dataset, 2019.

Wingham, D. J., Francis, C. R., Baker, S., Bouzinac, C., Brockley, D., Cullen, R., de Chateau-Thierry, P., Laxon, S. W., Mallow, U., Mavrocordatos, C., Phalippou, L., Ratier, G., Rey, L., Rostan, F., Viau, P., and Wallis, D. W.: CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields, 37, 841–871, https://doi.org/10.1016/j.asr.2005.07.027, 2006.

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

Robbie Mallet (referee n°2) - global comment

Global comment as a community comment :

I was really pleased to see this paper come up in TCD. I think that generating radar freeboard data from ERS1/2 is one of the most pressing tasks for the sea ice community, and so I agree with Jack Landy's review. Overall I think the paper is well written and addresses what is a very significant gap in our knowledge of the Arctic Ocean. In particular I think the figures are well-designed. I do have a couple of concerns, questions and suggestions over wordings, citations etc. I hope the authors will take these in the spirit of discussion, rather than as negative criticism. I really do think that this research is high-quality and useful.

Global comment as a referee :

I left a community comment on this manuscript (https://doi.org/10.5194/egusphere-2022-214-CC1) before being nominated as a referee. I have therefore read and considered the manuscript again. As part of this, I investigated the data that was made available to me as a nominated reviewer. I wanted to see the size of the correction/calibration applied by the neural network presented in this paper. This has led me to question the nature of the 'correction' being applied, and whether it is reasonable to present this data product as a series of 'corrected' radar freeboard values at all. I would like to review this manuscript again once the queries raised here have been addressed.

Answer to Robbie Mallet (referee n°2) - global comment

We would like to thank the reviewer for his careful reading of the manuscript and for the relevant remarks that have helped to improve the quality of the manuscript. In order to fit with your comments, we have made a revision of the manuscript that should have corrected the textual issues and well improved the readability of the document. We hope that these modifications will meet your requirements.

In our understanding, the main concern of the reviewer is : "To what extent can we claim that the resulting product is a corrected or calibrated retrieval when it doesn't reflect the variability in the raw, retracked values ? " Expressed in other word, the referee states that "nothing of the original radar freeboard measurement remains in the corrected value" and that is an issue.

First, we would like to indicate that your detailed analysis on the correlations between raw freeboard and calibrated radar freeboard have pointed out difficulties that have motivated the calibrations (past studies) and thereafter the use of a neural network.
Using the exact same processing chain as for CryoSat-2 (with a TFMRA-50 retracker), the Envisat, and ERS radar freeboard estimates are very different to what we are supposed to observe in terms of magnitude and spatial patterns. Indeed, LRM waveforms are strongly impacted by the size of the footprint which is much larger in LRM than in SAR Mode ($\sim$ 180 km^2 to $\sim$ 5 km^2 [Stammer et al, 2018]). This is the main cause for the misfit between Envisat and CS-2 radar freeboards. In order to

deal with this issue, differences between CS-2 and Envisat have been analyzed, please see Guerreiro et al. 2017, Paul et al. 2018 and Tilling et al. 2019 for a complete overview. As mentioned in the manuscript, the first two studies point out that the radar freeboard differences between the two altimeters are correlated (not especially linearly correlated) to the sea ice roughness, characterized by the waveform backscatter, the leading edge width or the pulse peakiness. The third study identified a link between the misfit and the distance between floes and leads.

The optimal solution would be to find a theoretical model, such as the Brown's model over open ocean, to represent the radar response over sea ice in order to correctly retrack the waveform. Despite significant progress in SAR mode (SAMOSA+, LARM, etc.), these models are not yet able to represent all the complexity of this response even in SARM. For instance, they are not able to represent snow penetration effects (i.e. volume backscatter effects). Moreover, in LRM, no study reports relevant retracked height over sea ice floes with a physical retracker and the complexity of the response is still poorly understood. The objective of this paper is neither to model any effect of the ice surface condition, nor to understand its influence on the FBr but rather to reconstruct the best possible ERS-2 radar freeboard with our actual knowledge consistently with Envisat and CS-2 ones. To do so, roughness or more globally sea ice surface state proxies are used to post-correct the estimated radar freeboard using as a reference Envisat previously calibrated on CS-2. Our study is based on the principle that the radar freeboard computed with a TFMRA50 from LRM waveforms is strongly polluted by the surface roughness. Then, we propose to calibrate LRM radar freeboard on CS-2 using some parameters characterizing the sea ice surface roughness. The same methodology is applied to calibrate ERS-2 radar freeboard on a CryoSat-2 like radar freeboard from Envisat. Thereafter, some other parameters such as the ice concentration or the sea ice age were added to improve and consolidate the learning of the NN so to reach a better match with CS-2 (in the case of Envisat and Envisat calibrated for ERS-2).

Unlike the review suggests, we would like to specify that the sea ice age is not directly used in the regression, we use a MYI fraction. The way this fraction is calculated has been developed in the manuscript, but it is not discrete values, as it is considered by the reviewer. Also, we would like to specify that correlations calculated with a variable that takes only two values can not be relevant.

To illustrate that the calibration is based on the PP and the LES, Figure 1 shows radar freeboard for April 2011 for CS-2, for Envisat with the calibration presented in the manuscript and one with another model trained only with the raw freeboard, the Pulse Peakiness and the Leading edge slope. It shows that these three parameters are sufficient to represent the magnitude and the patterns we are supposed to see in the Arctic. The other parameters help the calibration to get closer to CS-2 radar freeboard and bring more spatial variability.

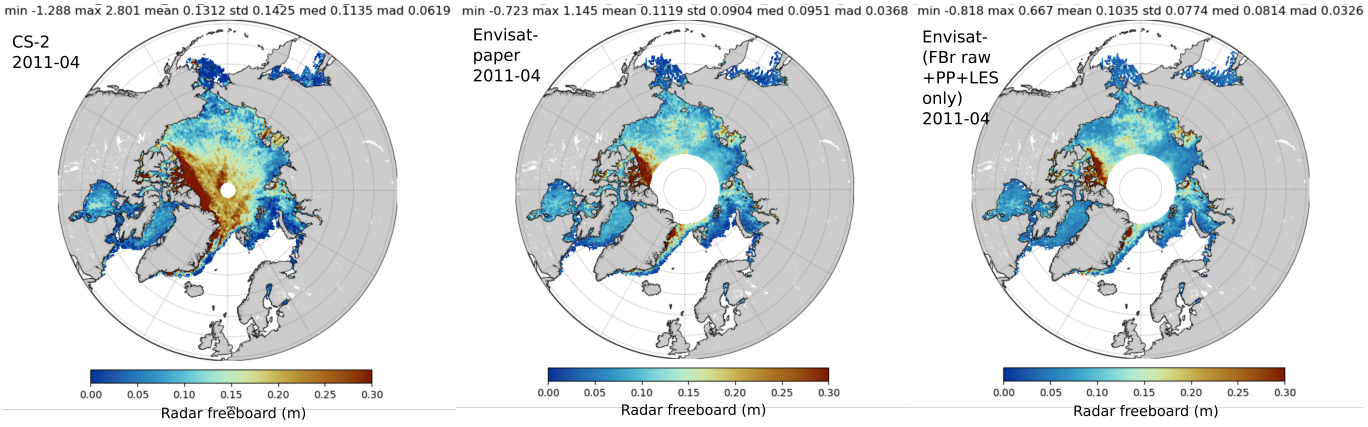


Figure 1. Radar freeboard from CS-2 (left), Envisat calibrated presented in the paper (middle) and Envisat using only the raw freeboard+LES+PP (right)

The correction we have to process is strongly non-linear, so this is the reason why we have chosen a neural network approach, which has the specificity to handle well with non linearities. Then, correlation between parameters based on linear approximation are not representative of the dependencies between parameters (inputs/output) in the neural network. Indeed, it is much more complicated to estimate the relative importance of each parameter in a regression, and it is not given by the correlations between the inputs and the predicted value. As it has been already mentioned, the main reason is that the relations that have been established by the neural network are not linear, while the correlation only evaluate whether the variables are related by a linear relation. Figure 2 shows the "partial dependencies" which refers to an illustration (statistically computed) of the relations between each input parameters (x-axis) and the predicted value (y-axis). It also illustrates the relative importance of each parameter: a parameter with no influence would have a horizontal curve as a mean state but it is not a quantitative approach. Partial dependency plots should be interpreted with caution, it refers to the mean state of statistical computation, depends on a discretization choice and values of input parameters have been standardized (mean = 0 and Standard deviation=1). The Figure 2 presents 2 panels, one for Envisat calibration (left) and one for ERS-2 calibration (right). Nevertheless, we can say that curves are not linear, no input is unused or with a very low influence, we can also note that LES and PP have the largest influence on the predicted radar freeboard.

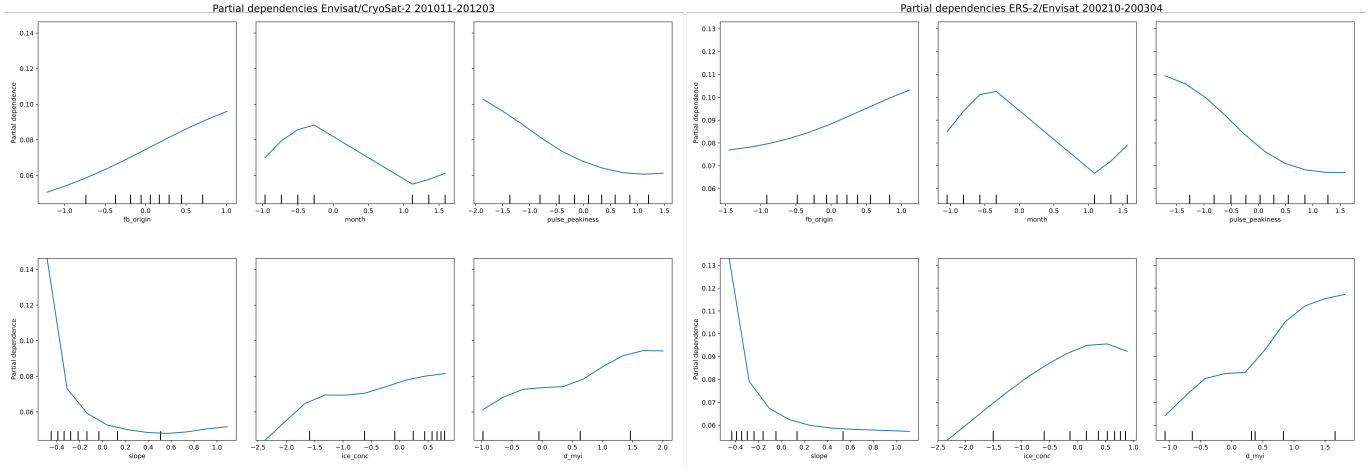


Figure 2. Partial dependencies plots for Envisat and ERS-2 calibration, from top left to top right, inputs are, the raw radar freeboard, the month, the pulse peakiness, the Leading edge slope, the concentration and the f_{MYI}.

The question of the non-linearity is central in our study but also in your analysis. But, since the correction is not linear at all, the largest raw radar freeboard is not necessarily the largest corrected freeboard, just as it is not the largest raw freeboard that will benefit from the largest correction. The raw radar freeboard, even noisy, still gives information on how the altimeter perceives the surface and how much it should be corrected, which remains an important information. We expect that the raw freeboard define the space and time variability of the calibrated radar freeboard over the whole period but this is hard to show since we don't have any reference of the expected variability of the SIT/FBr/FB during 1995-2010. To enhance the fact that the raw radar freeboard impacts the corrected radar freeboard, figure 3 shows the relative difference between the predicted FBr of the NN presented in the paper and one from a NN trained without the raw FBr for April 2011. It shows that for a large part of the basin, the difference of FBr is up to 25% of the predicted radar freeboard.

Finally, it's important to keep in mind that we have trained the neural network to reach the best score i.e. the best coefficient of determination (compared to CS-2 for Envisat and to Envisat corrected for ERS-2). Choosing the best NN, means choosing the combination of hyperparameters and even the choice of input parameters that gives the best scores. This means that the fraction of MYI allows to better fit CS-2 radar freeboard, that's why we keep it. However, it's even expected to find a good correlation between the sea ice age and the sea ice freeboard because, in average, older ice will be thicker.

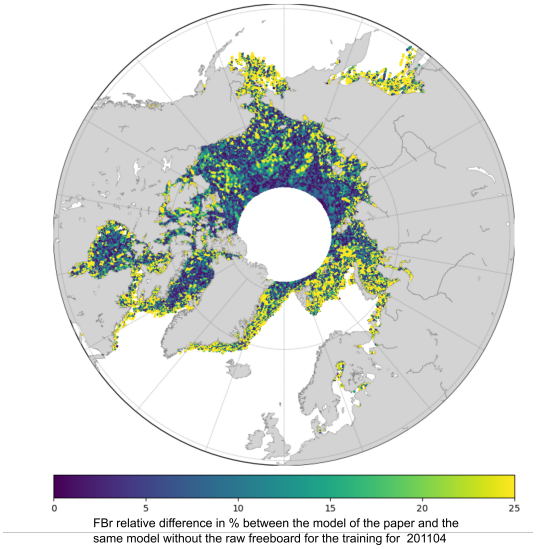


Figure 3. Relative difference in % between the corrected freeboard of our study and one with the same NN trained without the raw freeboard

To sum up, the purpose of this paper is to retrieve a consistent radar freeboard estimation for ERS-2 using the current knowledge on LRM waveforms over sea ice. Because LRM waveforms are highly impacted by the surface state and poorly understood over sea ice, raw freeboard have to be calibrated. Two calibrations need to be implemented to get consistent ERS-2 radar freeboard, first Envisat against CS-2 and then ERS-2 against Envisat calibrated radar freeboard. The calibration is first based on surface roughness proxy because evident link have been emphasized with the size of the correction (previous studies) and secondly on auxiliary data that were used to reach better fit with CS-2. The calibrated radar freeboard is partly driven by the raw radar freeboard, both parameters are not linear correlated as it would say that the calibration did not perform well. The "age" or in our case the MYI fraction is not the key input for the NN training. Furthermore, 'ice with a higher-than-average raw FBr in a given month" can not necessarily "end up with a higher than average corrected FBr value" as the calibration is not linear.

Please find below the details on how your specific comments have been taken into account. We have split in two part the specific comments, as the referee gives two detailed comments, one as a community and one as a referee. *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 Robbie Mallet (referee n°2) : specific comments

Specific comments - Referee comment

L280: I think you should by convention use the coefficient of determination rather than Pearson-r as a test score. Otherwise you'll end up with highly correlated relationships that have the wrong slope?

This is an error in the manuscript, we do use the coefficient of determination as the score for our regression. The correction has been made in the manuscript.

You need to explain quite a lot more about what's going on in Figure 6. The manuscript should not feature undefined letters and symbols, and there are many in this figure.

Caption has been largely developped as following :

Summary diagram of the uncertainty budget from along track to the propagation by the neural network.

replaced by:

Summary diagram of uncertainty budget during along track, gridding and calibration steps. Top left panel corresponds to the along track to grid uncertainty budget. Top right panel defines the notations, for the Monte Carlo procedure : Ω for the Neural Network input parameters, Γ for the Neural Network output parameter (radar freeboard) with σ_Ω and σ_Γ, the corresponding uncertainties. The middle panel corresponds to the training of M models with noisy inputs and outputs. Bottom panel show the predictions of the N noisy input with the M neural network trained. γ is the predicted radar freeboard estimation for one pixel of the MxN predictions. M=100, N=200.

Similar to above, you should explain much more about what's going on between lines 277 285. Papers in The Cryosphere should be accessible to scientists without extensive experience in machine learning. Don't be afraid to use the supplement for this, as I appreciate it's wordy. For instance, why did you choose 5 hidden layers and 100 neurons, and what are the implications of your choice? Why a sigmoid? There are noticeably no references to support your choices, and there's no element of later discussion about the impacts.

We have chosen a MLP because it has a very simple architecture but can deal with non-linear problem, which is the case of this study. The choice of the architecture (number of layer and neurons per layer) as explained in the manuscript as been fixed by testing a large amount of setup (called gridding) and choosing ones that have give the best score on the validation sample, with a reasonable time of learning. Concerning the activation function, the sigmoid was chosen to allow negative FBr as it is for target FBr and not to drop the value and bias the statistics of the predicted values. The sigmoid activation function was chosen so that it could allow negative FBr values in order not to artificially drop the negative predicted FBr values.

Machine Learning is largely used for various application even for geosciences but settings architecture and hyperparameters resides in testing testing and testing to get the best model with the best score. Citing study that use a MLP for geosciences will not be relevant as the hyperparameters highly depends on the issue we want to deal with so it could even be wrong. The paragraph you refer in your comment have been detailed to make it clearer. The implication of all choice is that by choosing the Neural Network type, MLP, with trained the best model possible so to have the best prediction possible comparing to a reference (CS-2 or Envisat calibrated).

The neural network is a multilayer perceptron regressor (MLP) composed of 5 hidden layers, each composed of 100 neurons. The activation function used is a sigmoid. Hyper-parameters have been tuned by dichotomy by choosing at each step the hyper-parameter combination with the highest mean score (average score made on 5 models) on the test sample. The score used for this regression is the Pearson correlation coefficient. To determine the most suitable hyper-parameter combination, the dataset is randomly split into a training and a testing dataset, corresponding respectively to 90% and 10% of the initial dataset. To avoid overfitting, we use early stopping to interrupt the training when the score is not improving anymore. Once the hyper-parameter combination is set, the MLP is trained with the whole dataset. The NN trained is then applied to the LRM monthly grids to obtain a monthly LRM-corrected radar freeboard.

replaced by:

The neural network used is a multilayer perceptron (MLP). Both calibrations have been processed with Scikit learn [Pedregosa et al, 2011]. The MLP is composed of 5 hidden layers, each composed of 100 neurons. The choice of hyperparameters : number of neurons, the learning rate, the regularization term, batch size, activation functions, solver for the weights optimization, have been done using gridding methodology, e.g. testing combinations and take the one that give best score. The evaluation criterion, called the score, is chosen as the determination coefficient. Models are trained on 90% of the dataset and

tested on the remaining 10%, the splitting in random. During the tuning step, models are cross validated, it means that they are each trained 5 times with the same combination of hyperparameters but without the same train/test dataset, the 5 scores are then analyzed to determine the best combination. Cross validation give a better idea of the model performance as the dependence to the training dataset is limited. The activation function for the hidden layers neurons is a sigmoid, motivated by possible negative radar freeboard values and the optimizer is and ADAM [Kigma et Ba, 2014]. 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.

Finally, once the hyperparameters combination is set, the MLP is trained on the whole dataset to provide the calibration function. The trained model is then applied to the LRM monthly grids to obtain a monthly LRM-corrected radar freeboard.

I also have the view that 'radar freeboard' is not a geophysical quantity to be measured with an uncertainty. Instead it is precisely the retracked elevation of a waveform returning from sea ice, and is specific to a given radar's geometry and the chosen retracking algorithm. See the original definition in the supplement of Armitage Ridout 2015, and Tilling et al. 2019 for how different radars will generate different Rfbs even if they could 'look at' the same ice. Similarly, different retrackers will generate different Rfbs when 'looking' at the same waveform, all of them valid and precise.

So I think you should change the phrase 'radar freeboard correction' to 'radar freeboard calibration', as you're not correcting some uncertain value. Instead you're calibrating the Rfb from one instrument so that it's consistent with another instrumental geometry. The same with 'radar freeboard estimation' - you're not estimating it: it's a precise value resulting from the radar geometry and choice of retracker. I have a lot more to add on this issue, but it's quite philosophical/ subjective and I think we need to first focus on the issue concerning the representation of the TFMRA50 Rfbs in the 'corrected' product.

If we understood your comment correctly, we don't have the same point of view. The geophysical quantity we want to estimate is the freeboard. For that purpose, we measure the height over the floes and the height over the leads that are extrapolated below the floes. The so-called radar freeboard is the difference between both heights. These two heights are subjects to uncertainties (like for all measurements) which are propagated to the difference. Among these uncertainties, we have the speckle noise, the interpretation of the retracking to estimate the range e.g. the retracking step. I can't find any definition of radar freeboard on Armitage and Ridout 2015's paper supplement more than 'The radar freeboard is then simply the retrieved elevation of the sea ice floe relative to this interpolated sea level' but it is also the definition we consider with an uncertain SLA and uncertain sea ice floe elevation anomaly or what we call ILA (Ice Level Anomaly) in reference to SLA.

Note that "error" is used in the manuscript while dealing with speckle noise because Wingham et al 2006 used that terminology, but it refers to uncertainties, both words were often mistaken to qualify uncertainties until a few years ago. This has been clarified in the manuscript.

We do agree that the word 'corrected' is a bit confusing, as we also deal with uncertainties. Even though, an uncertain value can not be corrected, at least the uncertainty can be reduced contrary to an error that is known and could be corrected. These two words have a different signification.

In order to clarify the reading, we suggest replacing corrected by calibrated while dealing with the predicted radar freeboard from the NN or the surface state bias corrected radar freeboard.

Specific comments - Community comment

L25) I would question whether "thin ice is more sensitive to climatic hazards". Bitz and Roe (2004) argued the opposite: that thick ice is thinning faster, because areas of thin ice grow more quickly in winter. Age products also show that thicker, older ice is disappearing from the Arctic and being replaced by thinner, seasonal ice (e.g. Nghiem et al., 2007). So I'm not sure it makes sense to say that thin ice is more sensitive to climatic hazards, when thin ice is coming to dominate the Arctic and is more robust to temperature perturbations.

We do agree that because of the global warming, multi-year ice have started to disappear and be replaced (in area) the next winter by FYI. Thin ice thickness will be recovered more easily than thick ice if it melts. Nevetherless, the thinner the ice is, the faster it undergoes melting and breakup when temperature rise in late spring. It will be more supposed to break while occuring climate hazard such as cyclones or strong winds [Rheinlænder et al 2022]. During all seasons, thin ice will ridge, raft, diverge easier than thick ice [Stroeve et al 2018] so can highly affect sea ice area (and of course volume). We suggest the following modification :

Thin ice is indeed more sensitive to climatic hazards than thicker ice but it especially enables to compute the volume.

replaced by:

Thick and old ice is disappearing and being replaced by younger, thin ice that has a higher mechanical sensitivity. Thin ice is more prone to deformation [Stroeve et al 2018] that induce area changes, and is more sensitive to climate hazards such as cyclones or strong winds [Rheinlænder et al 2022]. Thickness is a key parameter for sea ice study, it varies a lot according to the regions and it modulates the sea ice volume evolution in the Arctic ocean [Landy et al 2022].

L32) I don't agree that it's "commonly accepted" that Ku-band radar waves penetrate the snow layer when it is sufficiently cold. I think that assumption is still up for discussion, and I would argue the opposite. I'm not aware of any in-situ or airborne CryoSat evaluation ever done over sea ice that has produced evidence that Ku-band radar waves consistently return from the snow-ice interface. For instance, neither of the airborne CryoVex 2006 and 2008 campaigns (Willatt et al. 2011) indicated that this was consistently the case over FYI. Results from a different radar system in Antarctica (Willatt et al. 2010) also showed that radar waves do not always return from the ice surface. Results from a third radar system deployed on MOSAiC (on SYI) indicate that more Ku- band power comes back from the snow surface than from the ice surface (Stroeve et al., 2020 Fig. 7; Nandan et al., 2022 Fig. 8). Garnier et al. (2022; Figure 9) shows results from CryoVex 2017 where the difference between Ka and Ku band ranging is at times negative, further casting doubt on the assumption. Moving to satellite-based evidence, Armitage and Ridout (2015) calculated CryoSat-2's penetration factor as 82%. Ricker et al. (2015) used buoys to show that snow accumulation caused increases in Rfb, not decreases (implying that the radar waves are not penetrating fully). This agrees with the work of Gregory et al., (2022; Figure 9) that shows that snowfall is correlated (not anti-correlated) with Rfb over both ice types. I would also argue that the often-cited work of Beaven et al. (1995) was not realistic – it featured snow that was shovelled, sifted through a screen, and then artificially smoothed at the surface by the weight of a metal plate before measurement. It is also striking that what the authors identify as the snow-ice interface appears at 20 cm range when it was 21 cm away in free space. Since it was 21 cm away in free space it should have appeared further away, at something like 25 cm in range due to the wave-propagation delay. There's no need to mention all this in your paper, but I wanted to briefly state my evidence before making the point that full Ku-band penetration is not a settled consensus, even for cold, dry snow. I think it would be fair to say that full penetration is "commonly assumed in satellite-based sea ice thickness products". But just because we're forced to assume it in our products doesn't mean the we should actually believe or accept the assumption.

We do agree that the knowledge of how far into the snow layer Ku-band radar waves can penetrate is still under deep discussion in the community. There is no possible consensus on the fact that signal penetration will depend on salinity, temperature, humidity, snow age and other parameters... However, it is important not to confuse the results of studies done in Antarctica with those done in the Arctic, just as it is important not to confuse the SAR results with the results of field studies, since the SAR treatment impacts the waveforms and does not only reflect the behavior of the ku-wave in snow. We suggest the following modification :

It is commonly accepted that the Ku frequency penetrates the snow layer when it is sufficiently cold, in other situations this assumption can be questioned (Ricker et al., 2014; Nandan et al., 2017).

replaced by:

Implementing this method requires the assumption that the Ku-band radar wave completely penetrates the snow layer, which is still widely discussed and is not the subject of a definitive consensus (Ricker et al., 2014; Nandan et al., 2017).

L381: Year of this citation is 1986.

We have taken into account this semantic shade in the manuscript.

L65: I think we're not really measuring sea ice thickness, but instead estimating it based on freeboard measurements (or radar-altimetry measurements). This might seem like a semantic point, but I think users of sea ice thickness products do benefit from this distinction. "estimates" rather than "measurements" is more commonly used by convention (e.g. Tilling et al., 2018, Kurtz et al., 2014; Landy et al., 2017).

Increasing the along-track resolution of the aperture radar has led to considerable advances in the measurement of sea ice thickness.

replaced by:

Increasing the along-track resolution of the aperture radar has led to considerable advances in sea ice thickness estimation.

L75: I think readers like me who aren't expert in roughness would benefit from a citation here. Is LRM definitely more impacted by a given roughness than SARM? I can believe it, but would like to read some evidence.

Kurtz et al 2014 pointed out that ice roughness or more generally, sea ice surface properties impact the waveform of return echoes. Such as a lot of remote sensing instruments, the illuminated area will impact your measurements. Concerning the difference of roughness impact between SARM and LRM range measurement, it's due to the acquisition processing itself. Knowing how both work (see https://www.aviso.altimetry.fr/en/techniques/altimetry.html), the theoretical return power will be the same for both nadir and off-nadir in LRM whereas in SAR most of the return echo power will be concentrated to nadir, which reduces the impact of off-nadir and give the peaky shape to SAR waveforms and a lower impact of surface roughness [Raney et al 1998]. It was a bit more explicated in section 3.4. We propose to add this citation in the sentence you are mentioning and add a reference to the section where it is more explained.

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.

replaced by:

Because LRM altimetry has a larger footprint than SARM altimetry (by a factor 30), LRM range retrieval are significantly more impacted by surfaces roughness of the [Raney et al 1998] than the more nadir-focused measurement (SAR technologies).

L103: I think you mean NSIDC 0611? This product gives the maximum of the ice age distribution in a grid cell at each timestep (see quote below). So I'm not sure how you've used these max values to generate an MYI fraction product? I think it could be done if you had access to the Lagrangian data, which is out there. But if you've used this I think you should state that. (Tschudi et al. (2020) states "This approach does not consider new ice that may form within a grid cell because it retains only the oldest ice in its accounting. Thus, the product is effectively an estimate of the oldest ice in a given grid cell.")

The type is attributed to the 20hz along track measurements from the NSIDC age nested in two categories (whether the age is greater than 1 year). During data gridding, the type is also gridded and gives us an idea of the fraction of MYI by averaging the ice type into cells.

L103-104 : This information comes from the NSIDC 0061 sea ice age product (Tschudi et al., 2019) that is aggregated into two classes (MYI and FYI).

replaced by :

The study also requires a sea ice type product, this information is derived from the NSIDC 0061 sea ice age product (Tschudi et al., 2019) that is aggregated into two classes (MYI and FYI) according to the age of the ice (FYI : ice age between 0 and 1 year, MYI : ice age of at least one-year) at a weekly frequency. Data are respectively available as daily and weekly map with a 12,5 km grid resolution. The fraction of MYI is derived from the ice type information during the gridding processing step.

L115: I think at some point you should direct the reader to Kwok and Haas (2015), which discusses some key issues in the product that you've chosen.

This section aims to present the dataset not to discuss it, however, we added the reference to section results as following:

The bias between OIB and Envisat estimation could also be attributed to the OIB snow depth which estimation seems sensitive to the algorithm used (Kwok and Haas, 2015; Kwok et al., 2017).

L310: "Surface roughness is identified as the largest source of uncertainty" - I didn't really understand how you made it to this conclusion. I think this is specifically a reference to Fig. 8 of Landy et al. (2020). The error in the sea ice roughness over FYI is 4cm, and the error from the snow basal salinity (just part of the "penetration bias") is 7 cm, and the uncertainty due to snow depth is 6 cm. So over FYI the roughness uncertainty is smaller than either the snow depth or the snow salinity. As such I don't think roughness can be reasonably characterised as "the largest source of uncertainty" over FYI based on Landy et al. 2020 Fig. 8. Over MYI the sea ice roughness uncertainty is equal to the snow depth uncertainty, and admittedly larger than "partial snow penetration" uncertainty. So the statement is narrowly true if you only consider MYI and don't factor in the (highly related) uncertainty in snow depth in the comparison. But I think that only considering the largest source of uncertainty and ignoring the other uncertainties is a pretty risky strategy, given the other sources are comparable and perhaps actually larger in magnitude? If you are wedded to this approach, I think you should state that this will induce a pretty serious underestimate in your uncertainty values (which is important info for product end-users).

This sentence is inexact, it has to be shaded to "surface roughness is identified as one of the largest sources of uncertainty". Nevertheless, the other sources of uncertainty while measuring the FBr, as summed up in Landy et al 2020, is the uncertainty due to SLA (off nadir and low density) and the limited Ku-band penetration in the snowpack (caused for instance by snow basal salinity for FYI or metamorphic snow for MYI) not the snow depth. The point that was not explained in the manuscript, incorrectly, is that the uncertainties due to partial signal penetration in the snow are only indirectly taken into account, we don't ignore it. Indeed, it is not so trivial when comparing freeboards from different retrackers, to differentiate between roughness and penetration [Ricker et al 2014]. We believe that a "significant" part of the uncertainty on penetration is included in the uncertainty on roughness presented in [Landy et al 2020]. For this reason, we made the choice not to add values for the undefined limited penetration of the signal in the snowpack in this uncertainty budget. The problem of penetration is not ignored, but the manuscript lack of information on this point. As contribution of sources are not defined, yes, it is possible that the final uncertainties are underestimated. We suggest the following modifications:

Landy et al. (2020) decomposed it in two, the FBr systematic uncertainty budget, on the one hand, the uncertainties due to the penetration of the signal in the snow (depending on its salinity or if it is composed of metamorphic snow, according to the type of ice) and 310 in the other hand, the surface roughness. Surface roughness is identified as the largest source of uncertainty

and we, therefore, choose to consider only this source in our systematic uncertainty evaluation. Roughness is estimated to be respectively about 20 % and 30% of the sea ice thickness for FYI and MYI (Landy et al., 2020). Note that this systematic uncertainty budget only concern CS-2 mission which are afterward propagated to Envisat and ERS-2, indeed other mission will be "corrected" from surface roughness effect during the calibration procedure.

replaced by :

In Landy et al 2020, the FBr systematic uncertainty budget is decomposed in two parts, on the one hand, the uncertainties due to the penetration of the signal in the snow (depending on its salinity or if it is composed of metamorphic snow, according to the type of ice) and in the other hand, the surface roughness. We assume, as in Ricker et al 2014, that the comparison of the freeboard from different retrackers does not enable to separate the contribution of the roughness from the signal partial penetration. We therefore assume to consider both sources as one mixed contribution, estimated to be respectively about 20 % and 30% of the sea ice thickness for FYI and MYI (Landy et al 2020). The systematic uncertainties can be underestimated as the penetration of the radar waves in the snow uncertainty may be poorly handled. Note that this systematic uncertainty budget only concerns CS-2 mission which is afterward propagated to Envisat and ERS-2, indeed other missions will be "calibrated" from surface roughness effect during the calibration procedure.

Fig. 6: I see in the top panel that you've "summed the squares", which has the implicit assumption that uncertainties that you have considered are uncorrelated. It may be that you have good evidence to support this that I'm ignorant of, but it seems, for instance, that speckle noise may well be (anti?)correlated with surface roughness? Just as an example. I think that the omitted snow uncertainties involving penetration & depth are more likely than not to be correlated in some way. I think you should state that you've assumed the uncertainties are uncorrelated in your analysis, and give the reader some information as to what the results of that assumption may be.

Yes, we assumed the uncertainty due to the speckle noise and the SLA uncertainty to be uncorrelated, as it is precised l 304 in the manuscript. The speckle noise will not be correlated to the surface roughness but it is attributed to the surface asperities which are of the order of magnitude of the wavelength of the signal, about 2 cm, that causes interference in the signal. But a surface can have several roughness scales, for instance MYI highly rough can have asperities of about 2cm as well as newly formed sea ice, both will present speckle noise that induce the same uncertainty on the range.

By construction, systematic and random uncertainties are not correlated, so this is not really an assumption. Concerting the uncertainties due to snow penetration, we redirect you to the previous comment (and here the snow depth is not considered, so its uncertainty either).

Figs. 7 & 8: These are really well designed and presented

Thank you for this comment.

L381: 4) Why take snow density as constant? SnowModel-LG outputs depth and density, and includes some physics of densification/settling over time. So I think it's odd to use one of its variables and not the other, since they're so linked in the model. Snow impacts thickness retrievals by weighing the floe down and slowing radar waves: both of these effects are proportional to the mass of overlying snow – not the depth (see Mallett et al., 2021). So I think it makes a lot more sense to use both the depth and density (the SWE) in your thickness retrievals rather than just the depth. Here's a plot of the seasonal densification of SnowModel-LG snow north of 88N for the period 1995-2018. You'll see that as well as being more dense than your assumption, it also evolves over the season.

As suggested, we have done the scatter plot using the snow density from the model, figures have been updated in the manuscript. Nevertheless, it was not possible for CanCoast due to SnowModel-LG output coverage, and as we take OIB snow depth for OIB/Envisat SIT conversion, we have chosen to keep a constant density to keep consistency. It is interesting to see that the comparisons look really similar than with a constant snow density. Comparisons with moorings even give worse results with higher biases than using constant density. You can find the statistics in figure 4, 5, 6 and 7.

Fig 13: I'm a little unclear what the radar freeboard timeseries is supposed to represent. I imagine it mostly reflects the trend and variability in sea ice extent, and I think you should point this out to the reader. A simple correlation with SIE would quantify this relationship and reveal if the quantity is useful. For the part of it that doesn't represent SIE, would decreasing volume reflect a thinning of sea ice? Thinning snow (Webster et al., 2014) will mask the effect of thinning ice on the Rfb. In areas where the snow is really thinning quickly, the Rfb could potentially even increase even if the ice is thinning. I guess I would like to see a little interpretation of this quantity figure 13 rather than being left to do it as the reader.

The time series present the radar freeboard volume we have computed during the ERS-2, Envisat and CryoSat-2 period. We chose to show the evolution of the volume of radar freeboard instead of mean radar freeboard as it is more difficult to interpret as the mean freeboard depend on the number of pixel covered by ice and is not necessary representative of the global ice state because low concentration area have the same weight as compact ice area. The motivation of computing the volume is to represent better this evolution. according to [Landy et al 2022], figure 8, the anomaly in volume are mainly driven by thickness anomalies and not area for the Arctic, so this would not reflect the variability of sea ice area. Concerning the impact of snow on trend, you are right, it would change the trend as snow load have changed during the past 30 years, but it will give the same trend as using a snow depth climatology to derive the volume as it has been done in the majority of the previous studies. We suggest adding the following precision:

The evolution of the snow load is not taken into account in Figure 13, which means that the evolution of the volume is not fully represented, in the same way as if the total volume were derived with a snow depth climatology. Indeed, a decrease in FBr volume may merely indicate that the snow depth is greater and the ice thickness unchanged.

L450: I think you should state the limitations in your uncertainties here. In particular (and I think this is key), do the "observed" thicknesses fall within your uncertainty bounds? If not, then either your uncertainty bounds are wrong or the validation data is wrong. I think uncertainty bounds on retrievals are not useful unless you can show that observed data fall within them.

Thank you for this suggestion, uncertainties of satellite estimates have been added to validation plots for the review. Nevertheless, as explained in the manuscript section results, conclusion are not as simple to draw out, validation or retrieval are not necessarily wrong if validation dataset don't fall within uncertainties bounds. First, because validation data also present uncertainties but also because validation procedure of monthly satellite estimation assume that we observe in average the same sea ice surface than the validation do and that can be questioned for airborne or submarines dataset for instance.

Figures 4,5,6 and 7 present the 95% confidence interval of the SIT but without taking into account uncertainties on snow depth, densities etc for the FBr to SIT conversion step.

The plots have been updated with the variable snow density. For esthetical reason, bounds are not represented for comparisons with other satellite-based SIT estimation.

Guerreiro, K., Fleury, S., Zakharova, E., Kouraev, A., Rémy, F., and Maisongrande, P.: Comparison of CryoSat-2 and ENVISAT radar freeboard over Arctic sea ice: toward an improved Envisat freeboard retrieval, The Cryosphere, 11, 2059–2073, https://doi.org/10.5194/tc-11-2059-2017, 2017.

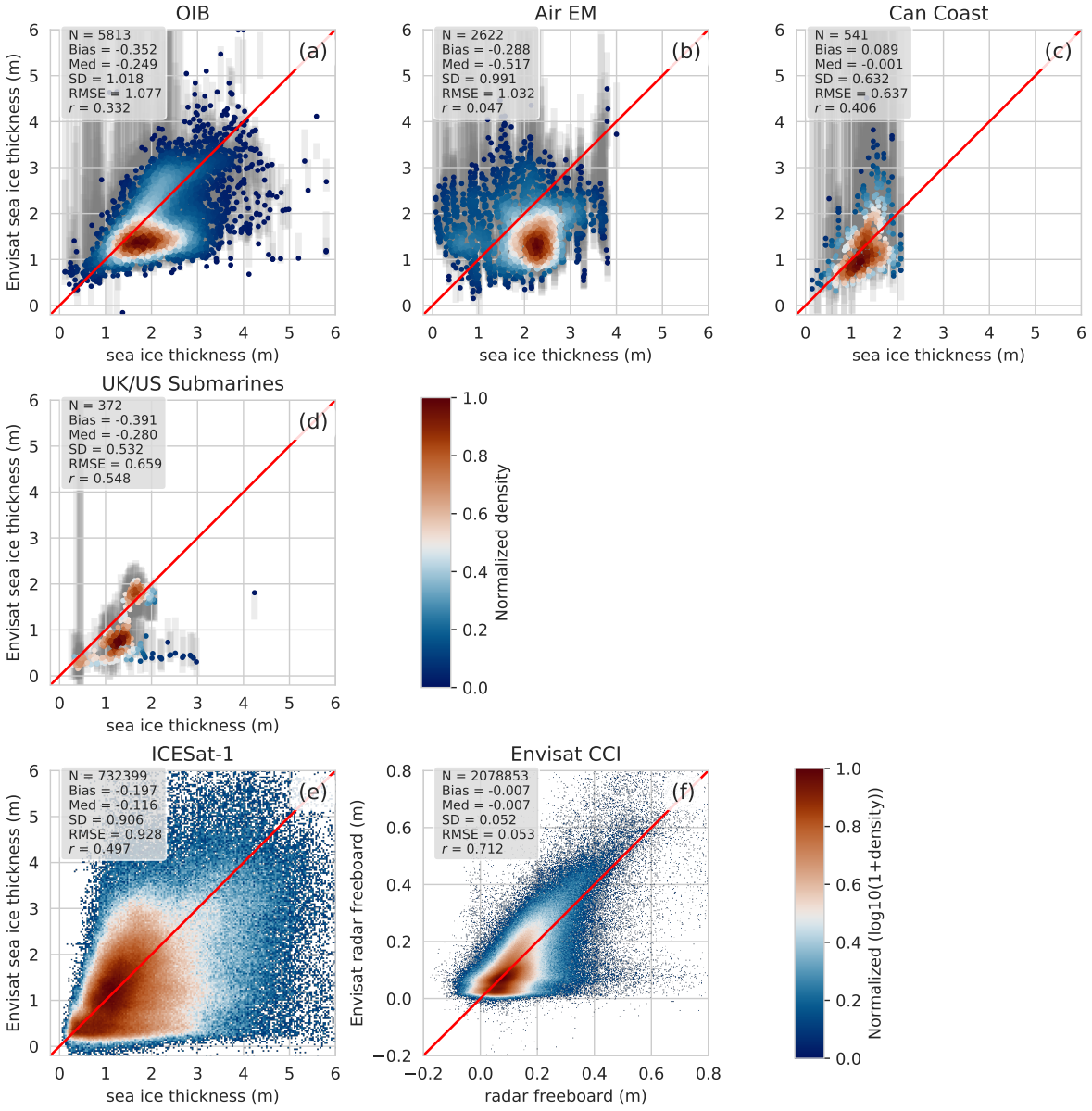


Figure 4. Comparative scatter-plots between Envisat sea ice thickness or radar freeboard estimations and other data sets. The x-axis indicates the sea ice thickness from (a) OIB total ice freeboard, (b) Air EM snow plus ice thickness, (c) Can Coast ice thickness, (d) UK/US submarines draft and (e) ICESat-1 total freeboard. (f) compares our Envisat radar freeboard with SI-CCI Envisat solution. Colorbars represent the normalized density. A log_{10} has been applied before the normalization for (e) and (f) due to the large number of data. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, CoRR, 2014.

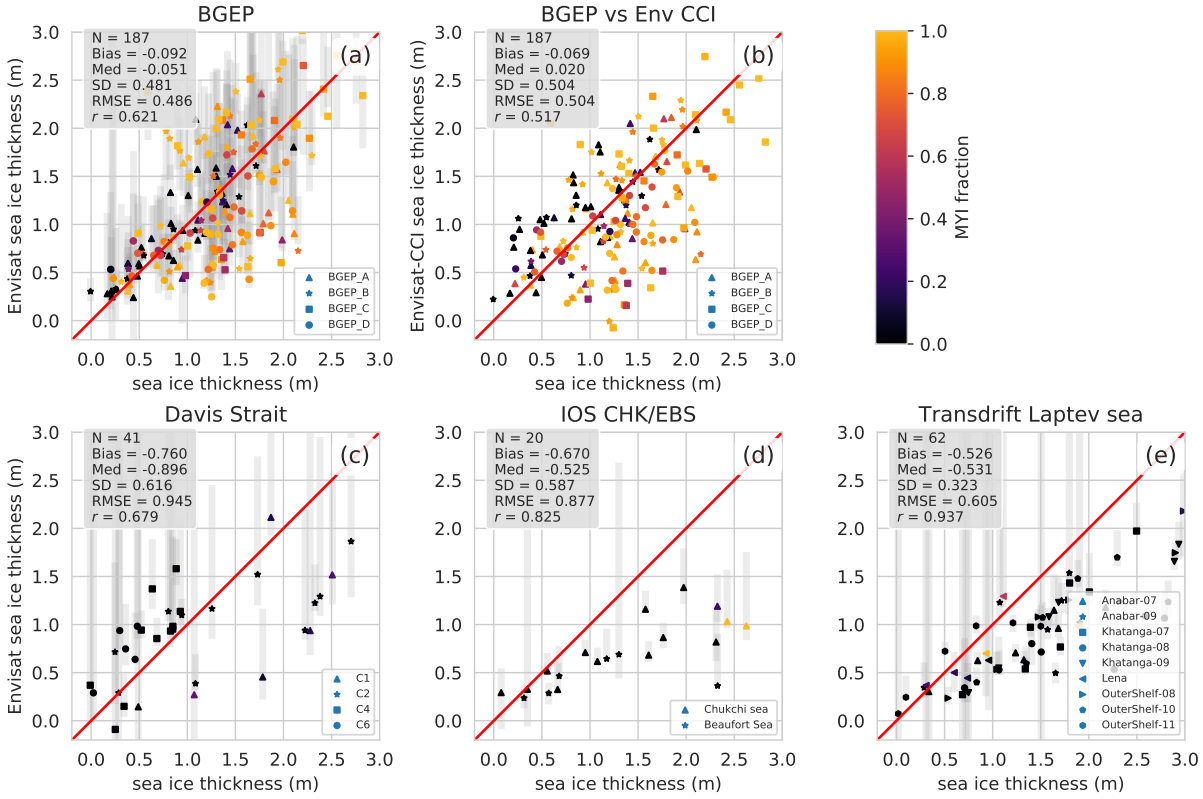


Figure 5. Comparative scatter-plots between Envisat sea ice thickness estimations and anchored moorings data sets. Each dot corresponds to a monthly averaged value. The x-axis indicates the sea ice thickness from (a) BGEP, (b) BGEP vs Env CCI, (c) Davis Strait, (d) IOS CHK/EBS and (e) Transdrift Laptev Sea ice draft. The colorbar shows the MYI fraction. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

Laforge, A., Fleury, S., Dinardo, S., Garnier, F., Remy, F., Benveniste, J., Bouffard, J., and Verley, J.: Toward improved sea ice freeboard observation with SAR altimetry using the physical retracker SAMOSA+, Advances in Space Research, p. S0273117720300776, https://doi.org/10.1016/j.asr.2020.02.001, 2020.

Landy, J. C., Petty, A. A., Tsamados, M., and Stroeve, J. C.: Sea Ice Roughness Overlooked as a Key Source of Uncertainty in CryoSat-2 Ice Freeboard Retrievals, Journal of Geophysical Research: Oceans, 125, https://doi.org/10.1029/2019JC015820, 2020.

Landy, J. C., Dawson, G. J., Tsamados, M., Bushuk, M., Stroeve, J. C., Howell, S. E. L., Krumpen, T., Babb, D. G., Komarov, A. S., Heorton, H. D. B. S., Belter, H. J., and Aksenov, Y.: A year-round satellite sea-ice thickness record from CryoSat-2, Nature, 609, 517–522, https://doi.org/10.1038/s41586-022-05058-5, number: 7927 Publisher: Nature Publishing Group, 2022.

Nandan, V., Geldsetzer, T., Yackel, J., Mahmud, M., Scharien, R., Howell, S., King, J., Ricker, R., and Else, B.: Effect of Snow Salinity on CryoSat-2 Arctic First-Year Sea Ice Freeboard Measurements: Sea Ice Brine-Snow Effect on CryoSat-2, Geophysical Research Letters, 44, 10,419–10,426, https://doi.org/10.1002/2017GL074506, 2017.

Paul, S., Hendricks, S., Ricker, R., Kern, S., and Rinne, E.: Empirical parametrization of Envisat freeboard retrieval of Arctic and Antarctic sea ice based on CryoSat-2: progress in the ESA Climate Change Initiative, The Cryosphere, 12, 2437–2460, https://doi.org/10.5194/tc-12-2437-2018, 2018.

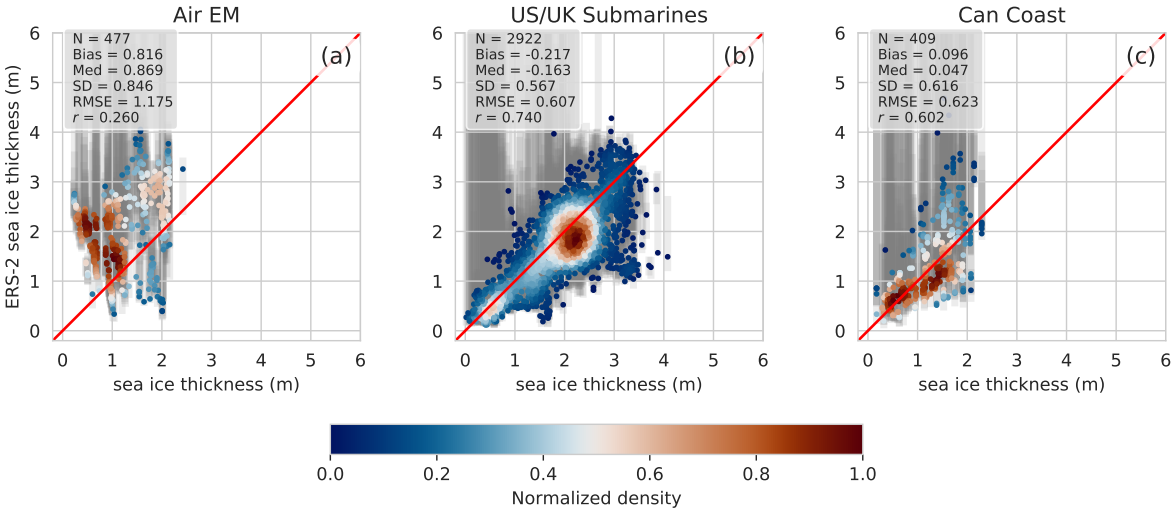


Figure 6. Comparative scatter-plots between ERS-2 sea ice thickness estimations and 3 in-situ data sets. The x-axis indicates the sea ice thickness from (a) AirEM total thickness, (b) UK/US Submarines draft and (c) Can Coast sea ice thickness. Colorbar indicates the normalized density. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

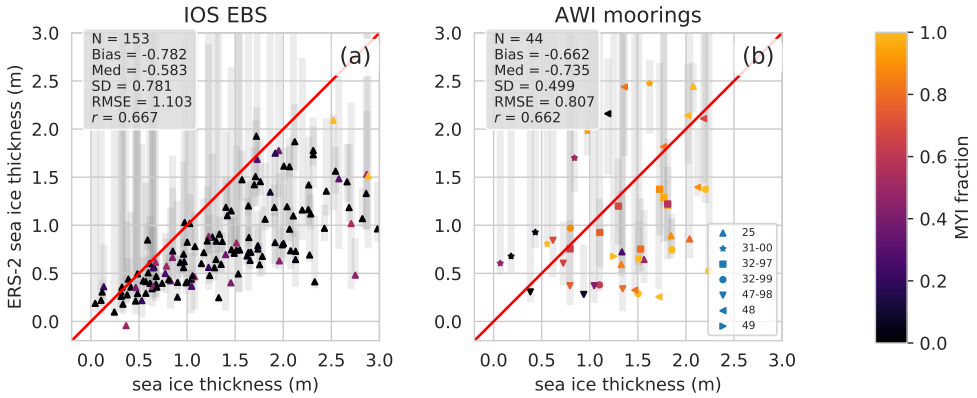


Figure 7. Comparative scatter-plots between ERS-2 sea ice thickness estimations and 2 anchored moorings data sets. The x-axis shows sea ice thickness estimations from (a) IOS Beaufort Sea and (b) AWI moorings sea ice draft. The color bar indicates the respective MYI fraction. N is the number of the couple of values that are compared, Med refers to the Median, SD the Standard deviation, RMSE the Root Mean Square Error and r the correlation coefficient.

Poisson, J.-C., Quartly, G. D., Kurekin, A. A., Thibaut, P., Hoang, D., and Nencioli, F.: Development of an ENVISAT Altimetry Processor Providing Sea Level Continuity Between Open Ocean and Arctic Leads, IEEE Transactions on Geoscience and Remote Sensing, 56, 5299–5319, https://doi.org/10.1109/TGRS.2018.2813061, 2018.

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

Rheinlænder, J. W., Davy, R., Ólason, E., Rampal, P., Spensberger, C., Williams, T. D., Korosov, A., and Spengler, T.: Driving Mechanisms of an Extreme Winter Sea Ice Breakup Event in the Beaufort Sea, Geophysical Research Letters, 49, https://doi.org/10.1029/2022GL099024, 2022.

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, 8, 1607–1622, https://doi.org/10.5194/tc-8-1607-2014, 2014.

Stammer, D.: Satellite altimetry over oceans and land surfaces, Earth observation of global changes, 2018.

Stroeve, J. and Notz, D.: Changing state of Arctic sea ice across all seasons, Environmental Research Letters, 13, 103 001, https://doi.org/10.1088/1748-9326/aade56, publisher: IOP Publishing, 2018.

Tilling, R., Ridout, A., and Shepherd, A.: Assessing the Impact of Lead and Floe Sampling on Arctic Sea Ice Thickness Estimates from Envisat and CryoSat-2, Journal of Geophysical Research: Oceans, 124, 7473–7485, https://doi.org/https://doi.org/10.1029/2019 2019.

Tschudi, M., Meier, W. N., Stewart, J. S., Fowler, C., and Maslanikand, J.: EASE-Grid Sea Ice Age, https://doi.org/10.5067/UTAV7490FE type: dataset, 2019.

Wingham, D. J., Francis, C. R., Baker, S., Bouzinac, C., Brockley, D., Cullen, R., de Chateau-Thierry, P., Laxon, S. W., Mallow, U., Mavrocordatos, C., Phalippou, L., Ratier, G., Rey, L., Rostan, F., Viau, P., and Wallis, D. W.: CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields, 37, 841–871, https://doi.org/10.1016/j.asr.2005.07.027, 2006.

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

Anonymous referee n°3 - global comments

First, I would like to express my apologies to the authors for taking this long to provide my review due to personal reasons. Nonetheless, I was asked to still provide it also in the light of the two already published referee comments. This in mind, I will focus on aspects I do not see covered yet or extend on raised issues as I see fit with a focus on the "calibration" using a neural network. I provide general comments first with some additional specific comments at the end.

The authors present in their study their way of generating a new dataset of altimetry-based freeboard data with ERS-2 data incorporated for the first time. This is a great achievement in itself and definitely justifies publication. Furthermore, the authors put substantial effort in validating their results against several different types of validation data. ERS data in general is a great challenge to work with and there is a reason why not many people are actually working on the task to make use of them over sea ice.

However, as also pointed out in the very detailed review by Robbie Mallet, who went to great lengths to analyze the results and underlying data, it appears the chosen methodology does not really work the way the authors or at least any potential reader would expect it. There appears to be strong evidence that the large mix of input data to the neural network along side the ERS freeboard estimates dominate the outcome. Hence, the NN did not learn what was expected but something else. While this is not necessarily bad, it is a fundamental problem of the presented study, as in my opinion, this is can be seen as grist to the mills of all machine learning or artificial intelligence sceptics. It should clearly be stated what the impact of each dataset is on the resulting product or rather that its apparently not the input raw freeboard. Potentially, the product could even e generated without the raw freeboard? This really should be clarified upfront and likely further investigated by the authors before publication.

Answer to Anonymous referee n°3 - global comments

We would like to thank the reviewer for his careful reading of the manuscript and for the relevant remarks that have helped to improve the quality of the manuscript. In order to fit with your comments, we have made a revision of the manuscript that should have corrected the textual issues and well improved the readability of the document. We hope that these modifications will meet your requirements. Please find below the details on how your specific comments have been taken into account.

As the referee n°3 seems to have the same concerns as referee n°2 with some conclusions taken from the other referee's review, we propose the same global answer as for referee n°2.

In our understanding, the main concern of the reviewer is : "To what extent can we claim that the resulting product is a corrected or calibrated retrieval when it doesn't reflect the variability in the raw, retracked values ? " Expressed in other word, the referee states that "nothing of the original radar freeboard measurement remains in the corrected value" and that is an issue.

First, we would like to indicate that referee n°2's detailed analysis on the correlations between raw freeboard and calibrated radar freeboard have pointed out difficulties that have motivated the calibrations (past studies) and thereafter the use of a neural network.

Using the exact same processing chain as for CryoSat-2 (with a TFMRA-50 retracker), the Envisat, and ERS radar freeboard estimates are very different to what we are supposed to observe in terms of magnitude and spatial patterns. Indeed, LRM waveforms are strongly impacted by the size of the footprint which is much larger in LRM than in SAR Mode ($\sim$ 180 km^2 to $\sim$ 5 km^2 [Stammer et al, 2018]). This is the main cause for the misfit between Envisat and CS-2 radar freeboards. In order to deal with this issue, differences between CS-2 and Envisat have been analyzed, please see Guerreiro et al. 2017, Paul et al. 2018 and Tilling et al. 2019 for a complete overview. As mentioned in the manuscript, the first two studies point out that the radar freeboard differences between the two altimeters are correlated (not especially linearly correlated) to the sea ice roughness, characterized by the waveform backscatter, the leading edge width or the pulse peakiness. The third study identified a link between the misfit and the distance between floes and leads.

The optimal solution would be to find a theoretical model, such as the Brown's model over open ocean, to represent the radar response over sea ice in order to correctly retrack the waveform. Despite significant progress in SAR mode (SAMOSA+, LARM, etc.), these models are not yet able to represent all the complexity of this response even in SARM. For instance, they are not able to represent snow penetration effects (i.e. volume backscatter effects). Moreover, in LRM, no study reports relevant retracked height over sea ice floes with a physical retracker and the complexity of the response is still poorly understood. The objective of this paper is neither to model any effect of the ice surface condition, nor to understand its influence on the FBr but rather to reconstruct the best possible ERS-2 radar freeboard with our actual knowledge consistently with Envisat and CS-2 ones. To do so, roughness or more globally sea ice surface state proxies are used to post-correct the estimated radar freeboard using as a reference Envisat previously calibrated on CS-2. Our study is based on the principle that the radar freeboard computed with a TFMRA50 from LRM waveforms is strongly polluted by the surface roughness. Then, we propose to calibrate LRM radar freeboard on CS-2 using some parameters characterizing the sea ice surface roughness. The same methodology is applied to calibrate ERS-2 radar freeboard on a CryoSat-2 like radar freeboard from Envisat. Thereafter, some other parameters such as the ice concentration or the sea ice age were added to improve and consolidate the learning of the NN so to reach a better match with CS-2 (in the case of Envisat and Envisat calibrated for ERS-2).

Unlike the review suggests, we would like to specify that the sea ice age is not directly used in the regression, we use a MYI fraction. The way this fraction is calculated has been developed in the manuscript, but it is not discrete values, as it is considered by the reviewer. Also, we would like to specify that correlations calculated with a variable that takes only two values can not be relevant.

To illustrate that the calibration is based on the PP and the LES, Figure 1 shows radar freeboard for April 2011 for CS-2, for Envisat with the calibration presented in the manuscript and one with another model trained only with the raw freeboard, the Pulse Peakiness and the Leading edge slope. It shows that these three parameters are sufficient to represent the magnitude and the patterns we are supposed to see in the Arctic. The other parameters help the calibration to get closer to CS-2 radar freeboard and bring more spatial variability.

The correction we have to process is strongly non-linear, so this is the reason why we have chosen a neural network approach, which has the specificity to handle well with non linearities. Then, correlation between parameters based on linear approximation are not representative of the dependencies between parameters (inputs/output) in the neural network. Indeed, it is much more complicated to estimate the relative importance of each parameter in a regression, and it is not given by the correlations between the inputs and the predicted value. As it has been already mentioned, the main reason is that the relations that have been established by the neural network are not linear, while the correlation only evaluate whether the variables are related by a linear relation. Figure 2 shows the "partial dependencies" which refers to an illustration (statistically computed) of the relations between each input parameters (x-axis) and the predicted value (y-axis). It also illustrates the relative importance of each parameter: a parameter with no influence would have a horizontal curve as a mean state but it is not a quantitative approach. Partial dependency plots should be interpreted with caution, it refers to the mean state of statistical computation, depends on a discretization choice and values of input parameters have been standardized (mean = 0 and Standard deviation=1). The Figure 2 presents 2 panels, one for Envisat calibration (left) and one for ERS-2 calibration (right). Nevertheless, we can say that curves are not linear, no input is unused or with a very low influence, we can also note that LES and PP have the largest influence on the predicted radar freeboard.

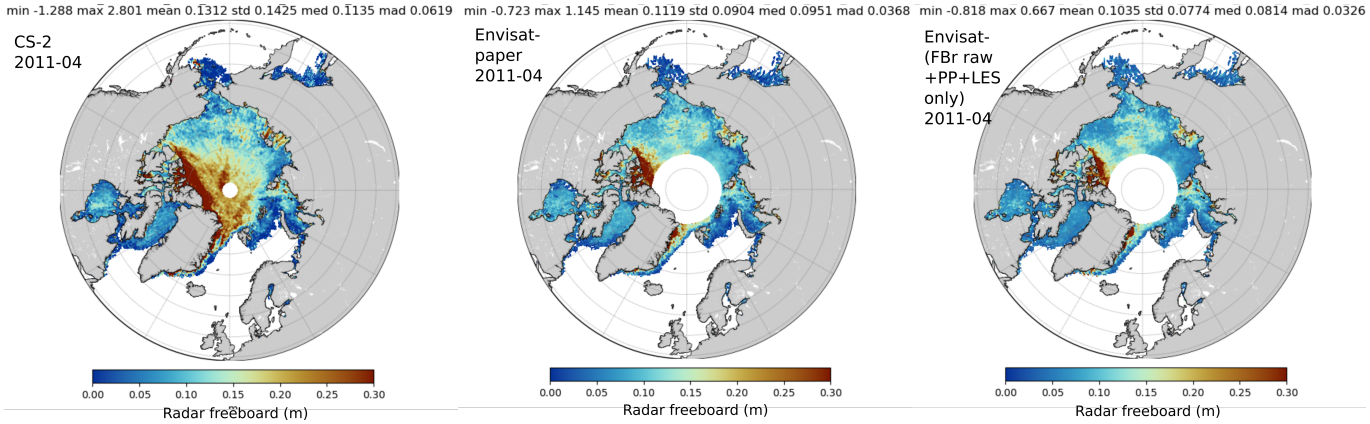


Figure 1. Radar freeboard from CS-2 (left), Envisat calibrated presented in the paper (middle) and Envisat using only the raw freeboard+LES+PP (right)

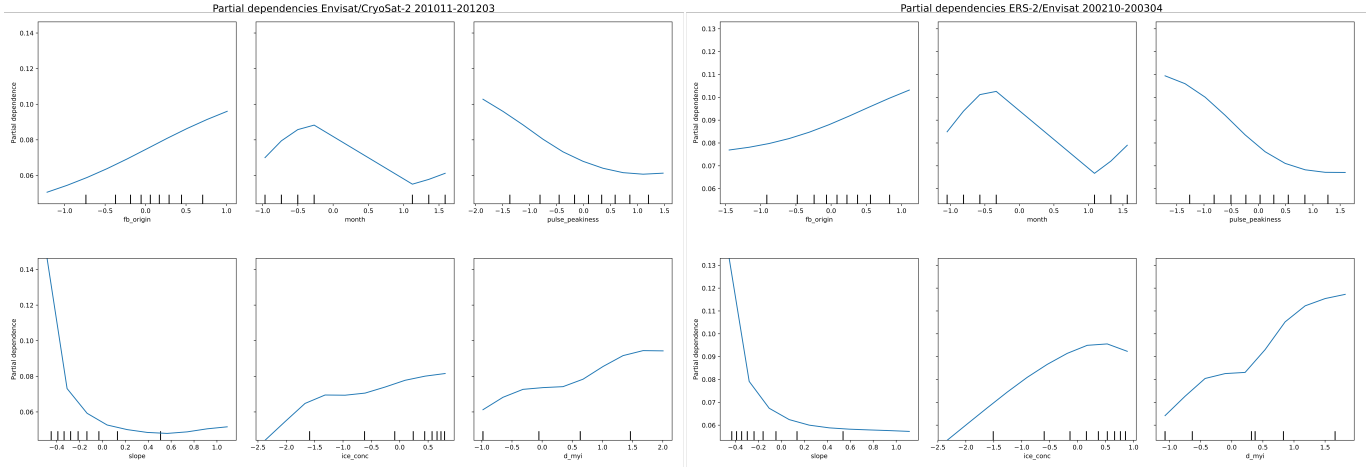


Figure 2. Partial dependencies plots for Envisat and ERS-2 calibration, from top left to top right, inputs are, the raw radar freeboard, the month, the pulse peakiness, the Leading edge slope, the concentration and the f_{MYI}.

The question of the non-linearity is central in our study but also in referee n°2's analysis. But, since the correction is not linear at all, the largest raw radar freeboard is not necessarily the largest corrected freeboard, just as it is not the largest raw freeboard that will benefit from the largest correction. The raw radar freeboard, even noisy, still gives information on how the altimeter perceives the surface and how much it should be corrected, which remains an important information. We expect that the raw freeboard define the space and time variability of the calibrated radar freeboard over the whole period but this is hard to show since we don't have any reference of the expected variability of the SIT/FBr/FB during 1995-2010. To enhance the fact that the raw radar freeboard impacts the corrected radar freeboard, figure 3 shows the relative difference between the predicted FBr of the NN presented in the paper and one from a NN trained without the raw FBr for April 2011. It shows that for a large part of the basin, the difference of FBr is up to 25% of the predicted radar freeboard.

Finally, it's important to keep in mind that we have trained the neural network to reach the best score i.e. the best coefficient of determination (compared to CS-2 for Envisat and to Envisat corrected for ERS-2). Choosing the best NN, means choosing the combination of hyperparameters and even the choice of input parameters that gives the best scores. This means that the

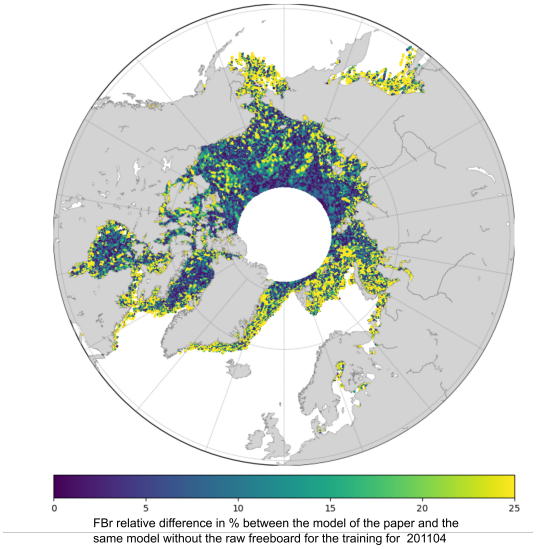


Figure 3. Relative difference in % between the corrected freeboard of our study and one with the same NN trained without the raw freeboard

fraction of MYI allows to better fit CS-2 radar freeboard, that's why we keep it. However, it's even expected to find a good correlation between the sea ice age and the sea ice freeboard because, in average, older ice will be thicker.

To sum up, the purpose of this paper is to retrieve a consistent radar freeboard estimation for ERS-2 using the current knowledge on LRM waveforms over sea ice. Because LRM waveforms are highly impacted by the surface state and poorly understood over sea ice, raw freeboard have to be calibrated. Two calibrations need to be implemented to get consistent ERS-2 radar freeboard, first Envisat against CS-2 and then ERS-2 against Envisat calibrated radar freeboard. The calibration is first based on surface roughness proxy because evident link have been emphasized with the size of the correction (previous studies) and secondly on auxiliary data that were used to reach better fit with CS-2. The calibrated radar freeboard is partly driven by the raw radar freeboard, both parameters are not linear correlated as it would say that the calibration did not perform well. The "age" or in our case the MYI fraction is not the key input for the NN training. Furthermore, 'ice with a higher-than-average raw FBr in a given month" can not necessarily "end up with a higher than average corrected FBr value" as the calibration is not linear.

Unfortunately, the impact of each parameter could be a dedicated study and it would not be the purpose of this study.

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.

L257: One could doubt the idea to use this kind of freeboards as an input in the first place. Wouldn't it make a difference to choose a more appropriate retracker threshold for leads in LRM waveforms like 90/95%? This might not solve the problem with regional patterns but would likely eliminate the negative freeboards and deliver a better initial state.

Empirical retracker with a threshold of 90/95% can be a risky strategy. The error on the range would vary a lot according to the sampling, randomly up to a gate (∼ 47cm). Using a 50% ensure the stability of the range (Poisson et al 2018 fig.9) even if we know we have a bias. Laforge et al 2020 shows that over leads comparing to physical retracker, the SLA bias is constant for altimeters in SARM, nevertheless this conclusion is also relevant for LRM as peaky waveforms are similar over leads. We

prefer to correct a bias than a random error, as we don't usually have a lot of measurement over leads. As suggesting for Jack Landy's review, we propose the following modification:

In LRM, most of this error comes from a constant bias on the Sea Level Anomaly

replaced by

Negative radar freeboards are mainly due to the retracker choice. Indeed, a TFMRA50 is used to retrack heights on both leads and floes, this introduces a bias on the height over leads. The TFMRA threshold to retrack heights over leads should be closer to 80% and the use of a 50% threshold corresponds to the position of the retrack point for ocean surfaces, not specular ones (Poisson et al 2018), the surface over leads is measured to be higher than it is and even higher than the surface over floes. The SLA bias (over leads) is evaluated constant for SARM altimeter in the study of Laforge et al 2020, this conclusion is also relevant for LRM altimeters as waveforms over leads are peaky and similar from a lead to an other. This positive constant bias over leads results to a negative bias on the radar freeboard. To avoid this bias, the retracker threshold could be adapted for leads or the SLA could be calibrated on CryoSat-2 one. Nevertheless, a threshold of 50% ensure the stability of the range (Poisson et al 2018, Fig.9) contrary to higher thresholds (80%-95%) that could lead up to 47 cm of random error on the SLA. A TFMRA at 50 % for both leads and floes is preferred in this study as a constant bias is easier to correct than an undetermined random error.

On a very general note: What are the improvements over Guerreiro et al (2017)? What justifies the use of a neural network instead of simply extending this methodology? As it had a more direct link to the actual measurements of the instrument? (as suggested also by the authors in L258-260)

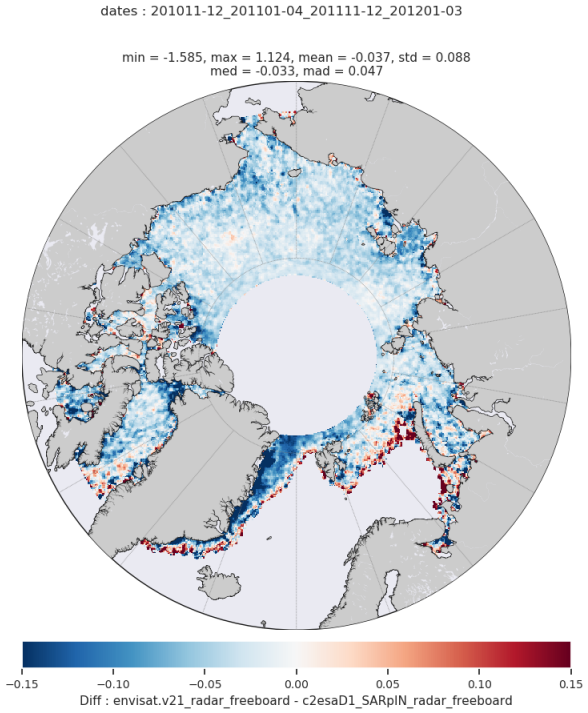
Figure 4 shows comparisons between Envisat-G radar freeboard from Guerreiro et al 2017, Envisat-B from this study compared to CryoSat-2 TFMRA50 radar freeboard. Maps present the mean radar freeboard difference between Envisat and CS-2 for the 12 months of the mission overlap period. These maps reveal that the first radar freeboard estimates were underestimated over the entire basin and especially in thick ice area. The correlation is much higher for Envisat-B radar freeboard. Compared to CryoSat-2, the improvement is evident with the method developed in this study. This study also reveals two modes of radar freeboard that could correspond to MYI/FYI.

As mentioned in the manuscript, the use of a neural network is justified by the non-linearities that exist between inputs and output, especially since it is not necessary to make any assumptions about the nature of the relations. The calibration can be seen as a usual regression in the continuity Guerreiro et al, 2017 with additional inputs to increase the match with SARM radar freeboard.

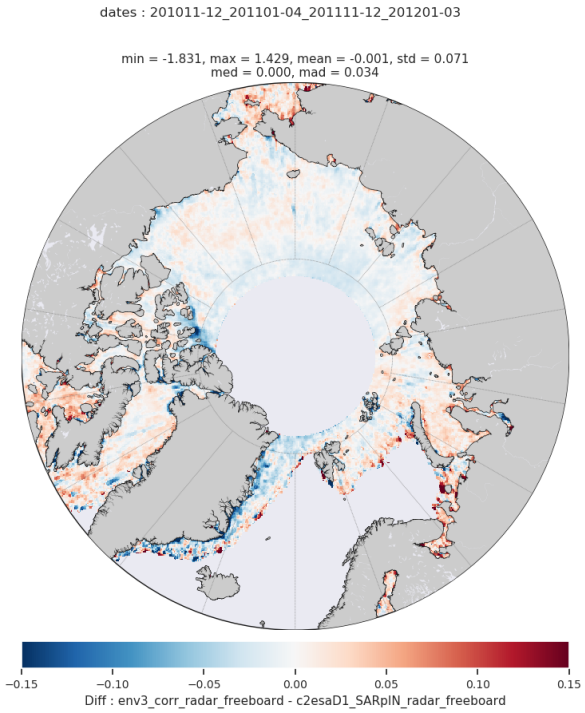
L277: Out of curiosity, did the authors test various setups and this architecture of the NN showed the best results? How was it evaluated and what different setups were used? Things like the number of layers, number of neurons per layer, activation functions etc. come to mind and all the mentioned specifics come without references or justification! For example, there are pretty much no modern studies on ML/AI that do not use some sort of ReLU activation functions, why do the author use a Sigmoid? Some elaboration on this might be informative to the readers as well and also provide a broader background also to non-ML enthusiasts in the sea-ice community.

Yes, a lot of setups has been tested. As already mentioned in the manuscript, the hyperparameters which also included the architecture (number of neurons and layers) have been choosing by gridding, e.g. by testing a large amount of combination. It was evaluated by cross validation, training 5 times the same setup on different training sets (randomly chosen) and test on the remaining 10% testing dataset (each time different). The best statistics on the ensemble of the 5 five scores were used to evaluate a combination. However, models with 102 neurons per layer instead of 100 were not significantly different.

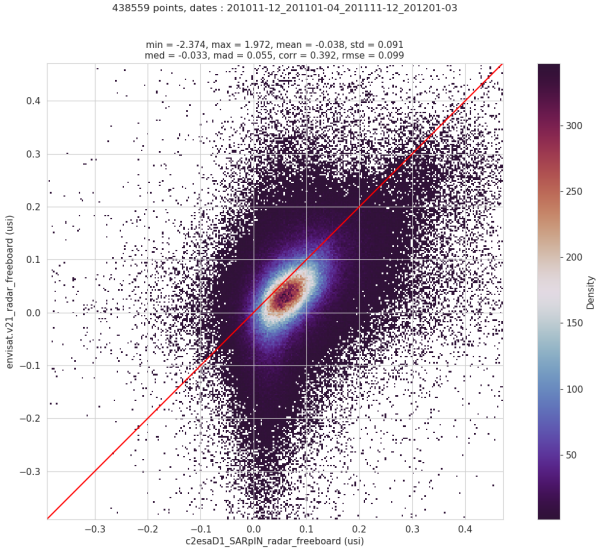
The tuning of hyperparameters has been made by experimenting, that the reason why there are no references, if you mean referencing other geophysical studies, it will have no sens as the tuning is really specific to the study. Activation function relies



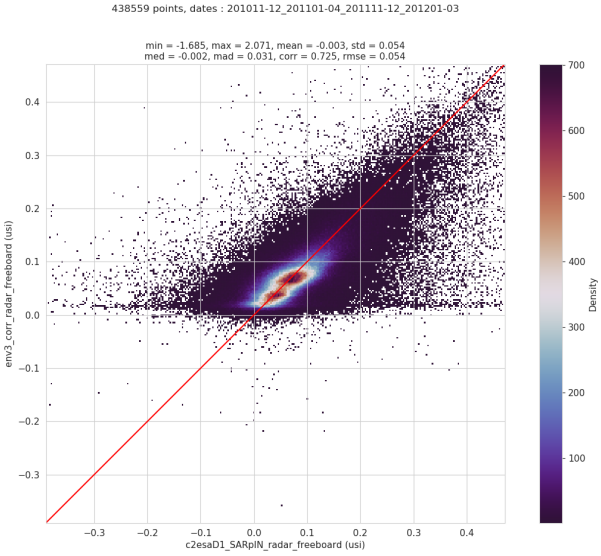
(a) Mean difference Envisat-G - CS-2



(b) Mean difference Envisat-B - CS-2



(c) Envisat-G vs CS-2



(d) Envisat-B vs CS-2

Figure 4. Comparison of Envisat-G (Guerreiro et al, 2017) and Envisat-B (Bocquet et al, 2022) with CryoSat-2 for the missions overlap period

on the type of value of the output data and ReLu function does not allow negative values, which could occur here. We propose the following modifications:

The neural network is a multilayer perceptron regressor (MLP) composed of 5 hidden layers, each composed of 100 neurons. The activation function used is a sigmoid. Hyper-parameters have been tuned by dichotomy by choosing at each step the hyper-parameter combination with the highest mean score (average score made on 5 models) on the test sample. The score used for this regression is the Pearson correlation coefficient. To determine the most suitable hyper-parameter combination, the dataset is randomly split into a training and a testing dataset, corresponding respectively to 90% and 10% of the initial dataset. To avoid overfitting, we use early stopping to interrupt the training when the score is not improving anymore. Once the hyper-parameter combination is set, the MLP is trained with the whole dataset. The NN trained is then applied to the LRM monthly grids to obtain a monthly LRM-corrected radar freeboard.

replaced by:

The neural network used is a multilayer perceptron (MLP). Both calibrations have been processed with Scikit learn [Pedregosa et al, 2011]. The MLP is composed of 5 hidden layers, each composed of 100 neurons. The choice of hyperparameters : number of neurons, the learning rate, the regularization term, batch size, activation functions, solver for the weights optimization, have been done using gridding methodology, e.g. testing combinations and take the one that give best score. The evaluation criterion, called the score, is chosen as the determination coefficient. Models are trained on 90% of the dataset and tested on the remaining 10%, the splitting in random. During the tuning step, models are cross validated, it means that they are each trained 5 times with the same combination of hyperparameters but without the same train/test dataset, the 5 scores are then analyzed to determine the best combination. Cross validation give a better idea of the model performance as the dependence to the training dataset is limited. The activation function for the hidden layers neurons is a sigmoid, motivated by possible negative radar freeboard values and the optimizer is and ADAM [Kigma et Ba, 2014]. 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.

Finally, once the hyperparameters combination is set, the MLP is trained on the whole dataset to provide the calibration function. The trained model is then applied to the LRM monthly grids to obtain a monthly LRM-corrected radar freeboard.

L279: The authors should clarify hyper parameters to the non-AI/ML expert readers. Without any reference I fear this is a lot to ask from potential readers of a non-AI journal. Additionally, what optimizer did the authors use as this can also have a substantial impact on the training process and the model performance and is totally unmentioned in the current version of the manuscript.

We hope that the previous comment will give an element of answer. The optimizer used is an adam, according to your comment, this information has been added.

L280: It is not clear to me how these 5 models are differing from each other? By slightly different choices on the learning rate? Please elaborate!

They are differing from their training dataset as it has been randomly split each of the 5 times. 5 models are used to make the cross validation. We also hope that the modification of the paragraph L277-285 help the reader and has clarified this part of the manuscript.

L282: Common practice would be a split around 80/20% or 75/25%, how do the authors justify such a small test-set size? This could result in a quite non-representative test dataset in the end.

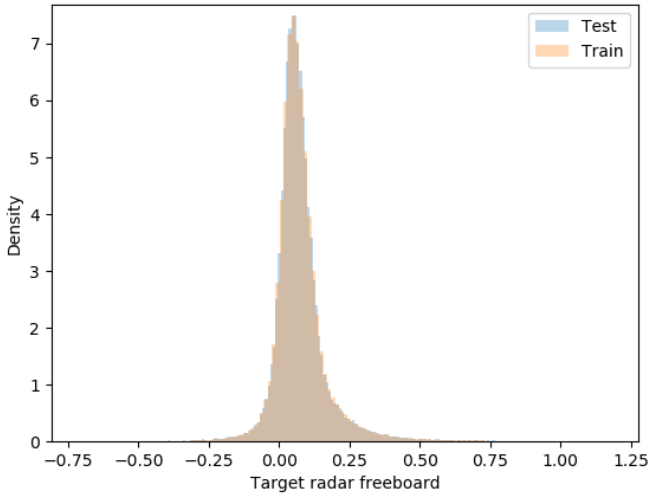


Figure 5. Probability density function of target radar freeboard for Envisat calibration for Test (blue) and Train (orange) dataset

This split can be justified by the fact that the dataset is quite important, with about 600000 values for Envisat calibration and about 300000 values for ERS-2 calibration. Figure 5 shows the probability density of the target radar freeboard for Envisat calibration for both test and train sample. Densities are identical for both datasets, this supports the fact that this splitting, in our case, leads to a representative test dataset.

Answers to referee 3 : specific comments

L118 & 122: the (Lindsay and Schweiger, 2013) reference should not be in parenthesis.

This point has been corrected.

L126: I think these PP thresholds should be mentioned here in a Table or within the text.

The following table has been added in appendix (Table A1).

L284: This should be 'the trained NN' not the 'the NN trained'.

This comment has been taken into account.

Table 1. Pulse peakiness thresholds for lead/floe classification

Mission (RA mode)	PP lead threshold	PP floe threshold
CryoSat-2 (SAR)	0.3*	0.1*
Envisat (LRM)	0.3*	0.1*
ERS-2 (LRM)	0.2839	0.1328

* [Guerreiro et al, 2017]

L286: I might just have missed it (sorry then) but what is the SARM abbreviation?

Indeed, this acronym was missing in the list of acronyms. SARM refers to 'Synthetic Aperture Radar Mode' to be consistent with LRM 'Low resolution Mode'. This point has been taken into account.

Figure 6: This definitely needs a much larger figure caption!

We do agree with that comment, the following modification has been done:

Summary diagram of the uncertainty budget from along track to the propagation by the neural network.

replaced by:

Summary diagram of uncertainty budget during along track, gridding and calibration steps. Top left panel corresponds to the along track to grid uncertainty budget. Top right panel defines the notations, for the Monte Carlo procedure : Ω for the Neural Network input parameters, Γ for the Neural Network output parameter (radar freeboard) with σ_Ω and σ_Γ, the corresponding uncertainties. The middle panel corresponds to the training of M models with noisy inputs and outputs. Bottom panel show the predictions of the N noisy input with the M trained neural network. γ is the predicted radar freeboard estimation for one pixel of the MxN predictions. M=100, N=200.

Guerreiro, K., Fleury, S., Zakharova, E., Kouraev, A., Rémy, F., and Maisongrande, P.: Comparison of CryoSat-2 and ENVISAT radar freeboard over Arctic sea ice: toward an improved Envisat freeboard retrieval, The Cryosphere, 11, 2059–2073, https://doi.org/10.5194/tc-11-2059-2017, 2017.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, CoRR, 2014.

Laforge, A., Fleury, S., Dinardo, S., Garnier, F., Remy, F., Benveniste, J., Bouffard, J., and Verley, J.: Toward improved sea ice freeboard observation with SAR altimetry using the physical retracker SAMOSA+, Advances in Space Research, p. S0273117720300776, https://doi.org/10.1016/j.asr.2020.02.001, 2020.

Landy, J. C., Petty, A. A., Tsamados, M., and Stroeve, J. C.: Sea Ice Roughness Overlooked as a Key Source of Uncertainty in CryoSat-2 Ice Freeboard Retrievals, Journal of Geophysical Research: Oceans, 125, https://doi.org/10.1029/2019JC015820, 2020.

Landy, J. C., Dawson, G. J., Tsamados, M., Bushuk, M., Stroeve, J. C., Howell, S. E. L., Krumpen, T., Babb, D. G., Komarov, A. S., Heorton, H. D. B. S., Belter, H. J., and Aksenov, Y.: A year-round satellite sea-ice thickness record from CryoSat-2, Nature, 609, 517–522, https://doi.org/10.1038/s41586-022-05058-5, number: 7927 Publisher: Nature Publishing Group, 2022.

Nandan, V., Geldsetzer, T., Yackel, J., Mahmud, M., Scharien, R., Howell, S., King, J., Ricker, R., and Else, B.: Effect of Snow Salinity on CryoSat-2 Arctic First-Year Sea Ice Freeboard Measurements: Sea Ice Brine-Snow Effect on CryoSat-2, Geophysical Research Letters, 44, 10,419–10,426, https://doi.org/10.1002/2017GL074506, 2017.

Paul, S., Hendricks, S., Ricker, R., Kern, S., and Rinne, E.: Empirical parametrization of Envisat freeboard retrieval of Arctic and Antarctic sea ice based on CryoSat-2: progress in the ESA Climate Change Initiative, The Cryosphere, 12, 2437–2460, https://doi.org/10.5194/tc-12-2437-2018, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Poisson, J.-C., Quartly, G. D., Kurekin, A. A., Thibaut, P., Hoang, D., and Nencioli, F.: Development of an ENVISAT Altimetry Processor Providing Sea Level Continuity Between Open Ocean and Arctic Leads, IEEE Transactions on Geoscience and Remote Sensing, 56, 5299–5319, https://doi.org/10.1109/TGRS.2018.2813061, 2018.

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

Rheinlænder, J. W., Davy, R., Ólason, E., Rampal, P., Spensberger, C., Williams, T. D., Korosov, A., and Spengler, T.: Driving Mechanisms of an Extreme Winter Sea Ice Breakup Event in the Beaufort Sea, Geophysical Research Letters, 49, https://doi.org/10.1029/2022GL099024, 2022.

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, 8, 1607–1622, https://doi.org/10.5194/tc-8-1607-2014, 2014.

Stammer, D.: Satellite altimetry over oceans and land surfaces, Earth observation of global changes, 2018.

Stroeve, J. and Notz, D.: Changing state of Arctic sea ice across all seasons, Environmental Research Letters, 13, 103 001, https://doi.org/10.1088/1748-9326/aade56, publisher: IOP Publishing, 2018.

Tilling, R., Ridout, A., and Shepherd, A.: Assessing the Impact of Lead and Floe Sampling on Arctic Sea Ice Thickness Estimates from Envisat and CryoSat-2, Journal of Geophysical Research: Oceans, 124, 7473–7485, https://doi.org/https://doi.org/10.1029/2019 2019.

Tschudi, M., Meier, W. N., Stewart, J. S., Fowler, C., and Maslanikand, J.: EASE-Grid Sea Ice Age, https://doi.org/10.5067/UTAV7490FE type: dataset, 2019.

Wingham, D. J., Francis, C. R., Baker, S., Bouzinac, C., Brockley, D., Cullen, R., de Chateau-Thierry, P., Laxon, S. W., Mallow, U., Mavrocordatos, C., Phalippou, L., Ratier, G., Rey, L., Rostan, F., Viau, P., and Wallis, D. W.: CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields, 37, 841–871, https://doi.org/10.1016/j.asr.2005.07.027, 2006.

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

Lars Kaleschke (editor) - comments

Dear author,

thank you very much for your detailed replies. I'd like to ask you to submit a revised version of the manuscript but I also have some questions/comments of my own:

I am not familiar with the term "gridding methodology" in connection with the choice of the MLP hyperparameters. Could you perhaps provide a references for this approach?

Could you please also consider my comment in my initial evaluation report about the data policy? I.e. compile data availability references in one section.

I think the time series analysis lacks an error estimate of the trend. It would also be interesting to know how much of the trend can be attributed to changes of the freeboard and how much to the passive microwave based ice extent. Perhaps you could estimate this by showing the result for a constant mean radar freeboard, e.g. the long term average of radar freeboard?

Answer to Lars Kaleschke (editor) - comments

The author would like to thank the editor for the carefull reading of the manuscript and the comments that helped to improve the manuscript and especially the part concerning the time series analysis.

"Gridding methodology" has been corrected to "grid search" methodology in the manuscript wich the common terminology.

Data availability for each data set have been summed up in the appropriate section as required.

Concerning the time series, we have added as suggested, an uncertainty estimation for each trends. To give a better idea to the reader of how the radar freeboard volume changes could be attributed to the radar freeboard itself we have added, as suggested the evolution of the radar freeboard volume using a climatology of radar freeboard build on the 1995-2021 period. Volume variability can be clearly attributed to radar freeboard evolution.

Figure 13 has been changed as following :

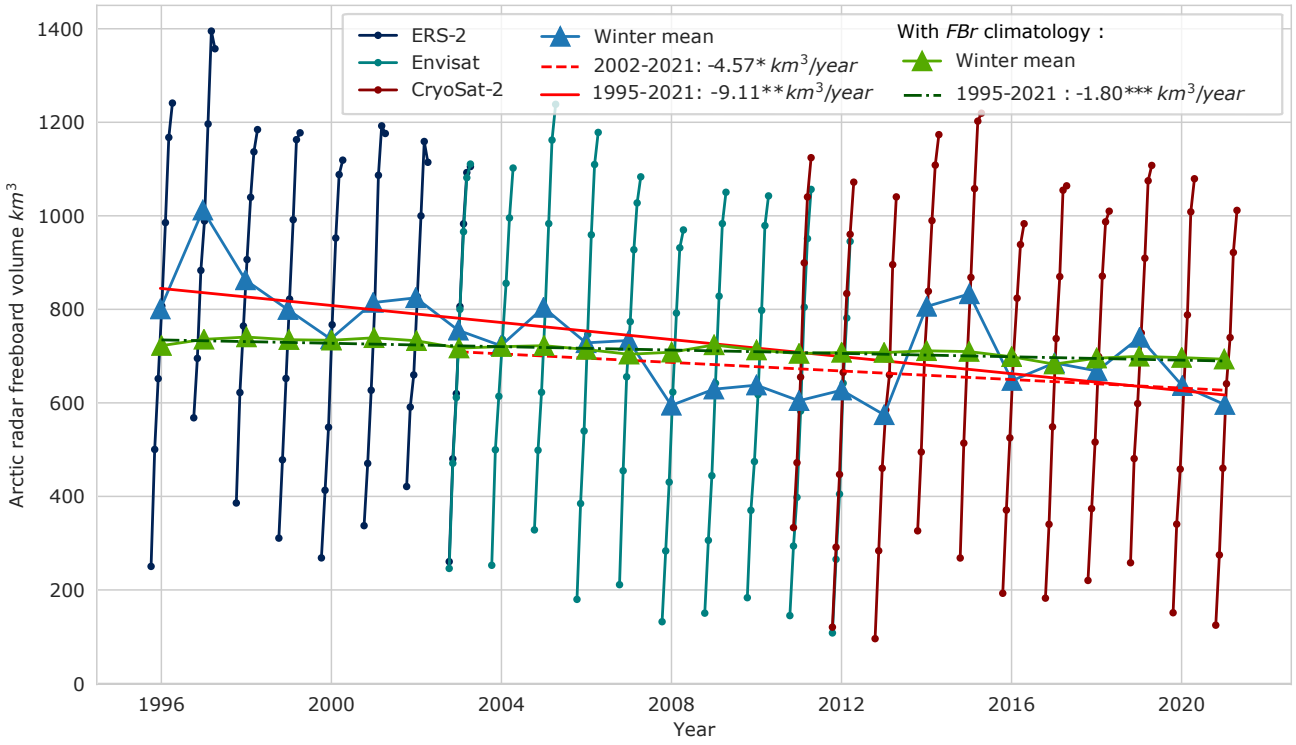


Figure 1. Time series representing radar freeboard volume up to 81.5°N for each winter month for ERS-2 in dark blue, Envisat in teal and CS-2 in dark red. Blue triangles are winter mean radar freeboard volumes. Red lines are linear regressions of winter mean volumes from 2002/2003 for dashed line and 1995/1996 for solid line, estimated trends are respectively $-4.57 \pm 8.73 \mathrm{km}^3/\mathrm{year}$ and $-9.11 \pm 5.16 \mathrm{km}^3/\mathrm{year}$. Green triangles represent winter mean radar freeboard volumes computed with a climatology of radar freeboard between 1995 and 2021, dash-dot green line is the regression for FBr volume with FBr climatology, the estimated trend is $-1.80 \pm 0.42 \mathrm{km}^3/\mathrm{year}$. *(1-p)<0.5, **(1-p)>0.99, ***(1-p)>0.999999, the probability value of the Mann-Kendall test.