This is a review of “Arctic sea ice radar freeboard retrieval from ERS-2 using altimetry: Toward sea ice thickness observation from 1995 to 2021”

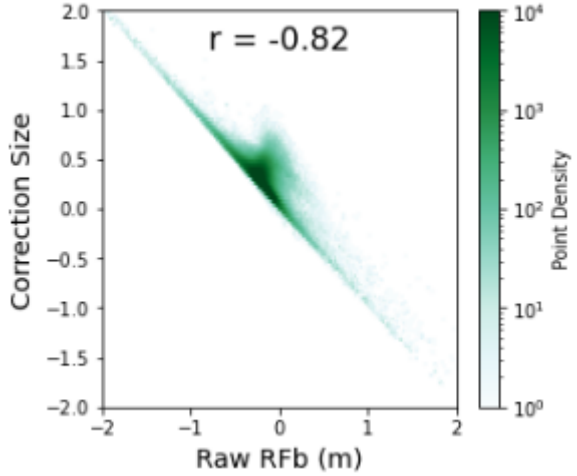
I should start with a quick self-correction: I previously misidentified Armitage and Ridout (2015) as having provided an “original definition” of the term *radar freeboard*. This was a mistake, as it was used prior to that for CryoSat-2 by Ricker et al. (2014), before that in an airborne context (Hendricks et al. 2010[1]), and even before that in non-peer-reviewed work.

Synopsis

While the authors have made a number of useful clarifying additions to their manuscript, I’m afraid to say my fundamental concern about the relative size and nature of their adjustment[2] remains unresolved.

I still don’t agree that the final data truly represent a “radar freeboard retrieval” (per the title), when they are so loosely related to the process of waveform retracking, and so dominated by other signals from ancillary data used in the adjustment process. So I still believe that the final product would be more accurately described as a *modelled* or *proxy* radar freeboard data set, given the second order influence that retracked elevations seem to have in the final result.

In their manuscript, the authors have not addressed or even mentioned my point that most of the work done by the adjustment is to oppose the TFMRA radar freeboard value. This is not how most calibration and adjustment workflows behave in my experience. I am confident that potential TC readers will not appreciate or understand this important feature of the method if the manuscript is published in its current form.



In my view, a calibration-based adjustment should be fairly *small* relative to the original and final values, and not be comparable in magnitude and opposite in sign to the value undergoing adjustment.

At the end of this review I have indicated some changes that I think should be implemented before this manuscript can be published.

All quoted line numbers that follow refer to the revised manuscript, not the “track-changes” manuscript, which I thank the authors for providing.

1 Hendricks, Stefan, et al. "Effects of surface roughness on sea ice freeboard retrieval with an Airborne Ku-Band SAR radar altimeter." 2010 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2010.

2 I’m using “adjustment” rather than “calibration” here, see my point below in **Minor Comments**

Technical Comments

Font sizes of axis labels on figure 3 should be equalised.

Double parentheses on L216

Need to define “ASA” and include it in your list of acronyms.

Minor comments

Given my concerns about whether the presented neural network can be reasonably described as a “calibration”, I should point out that *calibration* actually refers to the comparison of one data set to a standard, and does not refer to subsequent adjustment of the initial data set to match the standard. Really what is being done by the NN in prediction mode (after training) is *adjustment.*

https://calibrationselect.co.uk/general/calibration-or-adjustment/
https://soluzionesolare.com/news/difference-between-calibration-and-adjustment/
https://allometrics.com/what-is-the-difference-between-calibration-adjustment/

I appreciate that *calibration* is sometimes uses wrongly to refer to adjustment. I’ve done this, and there are several instances of this in the remote sensing literature. But the BIPM is clear in its most recent glossary[3] that calibration should not be confused or used interchangeably with adjustment.

The last paragraph of Section 1 actually uses “adjust” correctly three times, but elsewhere the process of adjustment is referred to as calibration.

I think by using the proper terms both here and in the manuscript we would be able to better discuss the performance of the blur-corrected TFMRA50 Rfbs in the calibration procedure, and the relatively large magnitude of the subsequent adjustment required.

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

L107: The authors have not meaningfully responded to my previous comment about the MYI fraction product they have constructed. Trivially, they have again misidentified the NSIDC ice age dataset as “0061”, when it is “0611”. More importantly, they have not addressed my other point that 0611 is not a simple map of ice age, but a map of *the oldest ice contained in a given grid cell* (see Tschudi et al. 2020 and my previous review). At the moment they have treated grid cells containing any MYI at all as being made up of 100% MYI (see figure 10) , which will undoubtedly introduce a high-bias into the MYI fraction product that they construct. I would urge the authors to go back and engage with my previous comment on this matter.

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

3 International Vocabulary of Metrology (2021) International Bureau of Weights and Measures; see p39 for definition of *calibration*. Indeed both the entries for *adjustment* and *calibration* include notes stressing that the two concepts should not be confused or used interchangeably. I note that the definition in this draft publication is essentially unchanged from the previous 2008 publication.

than it is for an LRM waveform. I'm not saying that the authors' claim isn't true – but a clear underlying mechanism and relevant reference should be provided.

L306: "Moreover, in order to avoid over-fitting, an early stopping criterion is used to stop the model training as soon as the score is not improved during 10 consecutive iterations, with a defined tolerance." I don't see how this would stop overfitting? You could easily have overfitted your model by the time that you fail to get an improvement in score, because overfitting can result from spurious improvements to the score that don't reflect improvements in the underlying model. I think overfitting is normally identified through comparison of performance on test sets with training sets.

In fact, I would suggest this model may well suffer from overfitting, due to the *very* low ratio of bias to variance (3mm to 9 cm in the Envisat-CS2 evaluation, 2mm and 3.8 cm for ERS/Envisat). This ratio is a well known indicator of overfitting, and suggests your model has captured the noise in its training data as well as the underlying signal.[4] You should address this in your discussion.

L300: If I understand right, you've specified a range of discrete hyperparameters and then investigated which combination optimises the score. But this description should be contextualised with the ranges and intervals over which you searched, the dimensionality of the hyperparameter space that was searched, and most importantly the results of the search. You've made choices about the learning rate, regularization term and weights-solver based on your grid search, but you haven't reported what those choices were. Given you performed what must have been a highly multidimensional and computationally intensive score-sensitivity analysis to get at these specific hyperparameters, it would be great to know what they are. I'm also not totally clear how the activation function was both found through this grid-search, and also motivated by the domain of the TFMRA Rfbs. I guess the domain of possible activation functions was motivated by the domain of Rfbs, and then you ended up selecting the sigmoid specifically through the grid search method?

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

L34 of the manuscript now reads "Sea ice thickness estimation by spatial altimetry was first introduced by Laxon (1994) and Peacock and Laxon (2004) based on the freeboard methodology".

I think the first real description of sea ice thickness estimation from radar altimetry derived freeboards was by Stanley et al. (1979)[5]. Their paper compared elevations from repeated tracks of the GEOS-3 altimeter in both the presence and absence of sea ice. They explained that the differences between the ocean elevations and the sea ice elevations (which were too large in their investigation due to poor gain control) contained information on the freeboard, which in turn could be used to estimate thickness if better constrained.

The breakthrough of Laxon's work with ERS1 was the individual classification of waveforms into lead/floe, allowing effective interpolation of the local sea surface height – this unlocked the method. So I think to make Laxon (2004) the valid citation, you should just add something like "in its modern form" to your sentence.

I can't see why a non-geophysical parameter such as "date" is being used to adjust the TFMRA waveforms (Figure 5). What physical mechanism could justify its inclusion in the neural network? It seems to me that the inclusion of a date parameter will only make the neural network behave like a seasonal climatology (further reducing the ratio of bias against variability); this would evaluate very well against in-situ sources in the way that you've constructed your tests (given the dominance of seasonal variability over interannual variability in sea ice thickness), but would potentially be an example of spurious overfitting. At minimum, a bit of justification needs to be given for its inclusion, preferably with some information about the final effect of its inclusion.

---

4 Some nice discussion of the relationship between bias-variance tradeoff and overfitting here https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

5 Stanely, R.H., Brooks, R.L and Brown, G.S: Ice Freeboard Determination by Satellite Altimetry (1979) Proceedings of the International Workshop on Remote Estimation of Sea Ice Thickness, St. John's, Newfoundland, Canada.

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

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

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

Major Comments

I think the authors have misinterpreted my point regarding the correlations between sea ice age and their adjusted product. My point was that their adjusted product is more strongly determined by sea ice age than it is by the value actually being adjusted. The TFMRA50 radar freeboards are not the first order component of the adjusted values, and in fact contribute relatively little compared to a well-known proxy variable. To what extent, therefore, can this meaningfully be considered a calibration and adjustment exercise? Instead, it looks like more predictive variables such as PP, MYI fraction and LES are themselves being slightly adjusted using TFMRA50 data. I would have preferred to do my original analysis based on PP or LES, but I didn’t have that data to hand so just used sea ice age which was easily downloadable.

Put another way – if you’re constructing a radar freeboard data set, I think readers and users will expect the TFMRA50 radar freeboard to tell you more about your final result than the MYI fraction does, and indeed to be the first-order determinant. That’s not to say that MYI fraction shouldn’t tell you about the radar freeboard – there are obvious links. It’s just to say that you’d hope the TFMRA radar freeboard would tell you more, *especially* as you’ve spent so long and worked so hard correcting the data for blurring.

Turning to the rebuttal, I still think that the relationships between parameters such as pulse peakiness and the adjustment value can be properly characterised even though they are non-linear. I think the authors should scatter-plot or point-density-plot the adjustment size and final data against (a) input radar freeboard (b) input pulse peakiness (c) MYI fraction (d) LES (e) date, and see which input has the clearest and strongest relationship to the output (regardless of the linearity of that relationship).

If I am interpreting Figure 3 correctly, then I conclude that the radar freeboard goes from having a limited role in the training of the neural network, to having a 25% influence on the training near the ice edge. However my concern is really that the input Rfb seems to have relatively little influence on the output Rfb in the *prediction* phase.

While the authors have given a lengthy rebuttal to my concern about the size and nature of the adjustment, they have made relatively few changes to the manuscript in this regard. I was expecting at least one or two plots to be added to the manuscript (or a supplement) quantifying and explaining how the different input parameters affect the adjustment values and the final results.

---

6 Ricker, Robert, et al. "Impact of snow accumulation on CryoSat-2 range retrievals over Arctic sea ice: An observational approach with buoy data." *Geophysical Research Letters* 42.11 (2015): 4447-4455.

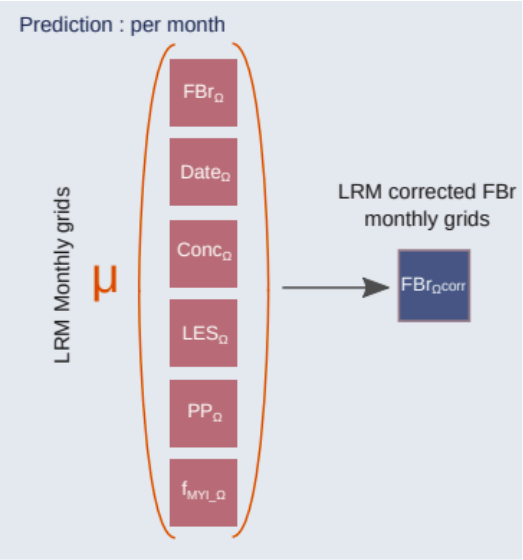
Summary

My opinion is that the adjustments presented here is so large relative to the value being adjusted, that the final value cannot reasonably be presented as a calibrated and adjusted "retrieval". This is particularly the case given the adjustment so often acts to counter the value being adjusted. I see this substantial body of work as a useful and interesting exercise in modelling a CryoSat-2-like radar freeboard value based on date, pulse peakiness, MYI fraction and leading-edge slope data, with a relatively small input from retracking waveforms. In particular, the work to resolve pulse-blurring issues is valuable.

Most importantly, I think *users* of this product (perhaps from the field of modelling or seasonal forecasting) would very likely mistake these data as being primarily derived from retracking radar waveforms to retrieve elevations. When actually they are primarily produced by a neural network which assimilates several other, quite diverse parameters.

If this manuscript is to be published in TC, I believe it must undergo some major revisions and additions as follows:

1) A proper exploration of the impact that each parameter in Figure 5 has on the output. I appreciate that an effort has already been made to do this via the "partial dependence" figure in the reviewer replies, but it's unclear to me how this was constructed and what it represents. I think it would be much clearer to scatter-plot or point-density plot (a) the size of the calibration against the different input variables, and (b) the final result against the input variables. Without this, the machine learning approach is simply the deployment of a "black box" machine.



At absolute minimum, the authors should explicitly quantify and report the apparantly loose correspondence between the TFMRA radar freeboards and the final radar freeboard product.

2) Show that their blur-corrected TFMRA50 Rfb's from ERS2 are related in at least some way to TFMRA50 Rfbs from Envisat in the overlap period.

3) Show that the influence of TFMRA50 Rfbs in the NN's adjustment improves the match between their ERS2 results and the EnviSat radar freeboards. At the moment, it seems that most of the work done by the adjustment is in *counteracting* the influence of the Rfbs. This makes me think you might be better off just straight-up modelling the radar freeboard based on your ancillary data sets alone.

I see that you've already thought about this with the map in Figure 3 of the reviewer replies – but this only shows the difference that's caused by omitting the TFMRA50 radar freeboards in the training. It doesn't show that the incorporation of TFMRA50 radar freeboards in the adjustment actually improves the situation at all (just that it makes a difference). To summarise points 2 & 3, I still have a suspicion that TFMRA50 retracking and blur-correction of ERS2 waveforms doesn't produce meaningful information about the sea ice, and thus doesn't contribute to your neural network's success in matching Envisat data sin the ERS/Envisat overlap period.