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.

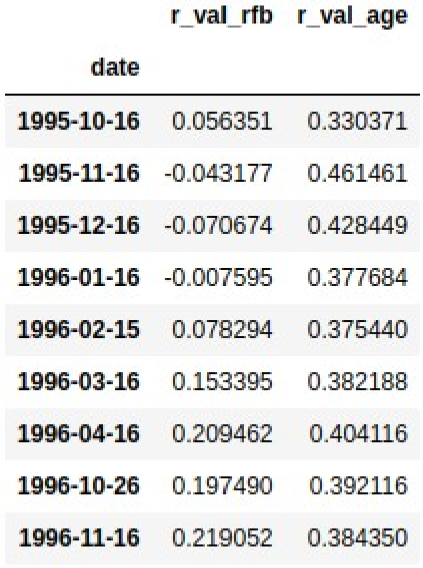
As part of this, I investigated the data that was made available to me as a nominated reviewer. I wanted to see the size of the correction/calibration applied by the neural network presented in this paper. This has led me to question the nature of the 'correction' being applied, and whether it is reasonable 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.

Firstly, it's possible that I have misunderstood how to read the data here, or made a mistake in my analysis. I have used the netcdf 'radar_freeboard_corr' variable as the 'corrected' Rfbs, although I did also run the same code on the 'corr_median' and 'corr_mean' and results were very similar. Please do let me know if I'm mistaken in any way. I've uploaded my code here: https://github.com/robbiemallett/ERS2_rev

In this paper the authors present a method of `correcting' the radar freeboard values generated by the ERS2 satellite. To do this, the authors retrack raw ERS2 waveforms using the TFMRA50 algorithm to generate a radar freeboard value. They then apply a neural network to the retracked heights, which also assimilates sea ice age and concentration data alongside other waveform statistics like pulse peakiness. The neural network takes in all this data and returns a 'corrected radar freeboard', which is theoretically consistent with other missions.

I began by mapping the raw and corrected ERS2 radar freeboards for each month, as presented in the data. It immediately struck me that the 56 pairs of maps did not look visually similar. That is to say, it was not clear that areas with larger TFMRA50 radar freeboards ended up as areas of higher corrected radar freeboards.

I then took a quantitative approach; For each year/month of data, I converted the raw radar freeboard values to anomalies from the month's mean, and did the same for the corrected values. I then performed a linear correlation between the two data sets for each month. This produced a table with 56 rows (one for each year/month combination from 1995 – 2003), here's the first few rows:



| date | r_val_rfb | r_val_age |
|---|---|---|
| 1995-10-16 | 0.056351 | 0.330371 |
| 1995-11-16 | -0.043177 | 0.461461 |
| 1995-12-16 | -0.070674 | 0.428449 |
| 1996-01-16 | -0.007595 | 0.377684 |
| 1996-02-15 | 0.078294 | 0.375440 |
| 1996-03-16 | 0.153395 | 0.382188 |
| 1996-04-16 | 0.209462 | 0.404116 |
| 1996-10-26 | 0.197490 | 0.392116 |
| 1996-11-16 | 0.219052 | 0.384350 |

*Illustration 1: First few rows of correlations between corrected RFb anomaly and TFMRA50 RFb (col 2) and sea ice age (col 1)*

The correlations between raw and corrected freeboard anomalies are very low, the mean value is r=0.07. I then performed the same analysis, but instead correlated the `corrected' radar freeboard anomalies with the sea ice age data (reconstructed per the manuscript). I set the values of MYI to 1, and FYI to 0. The mean r-values for this are 0.33. This means that the sea ice age data is a much stronger determinant of the corrected radar freeboard value than the raw radar freeboard value itself.

I then calculated the correlation between the size of the correction and the corrected anomaly itself. The mean r-value for this is 0.49. This means that the correction itself is a strong determinant of the end product,  by comparison to the raw, retracked radar freeboard value itself.

This is all a bit concerning. If my analysis is correct, it seems like the presented neural network is really modelling the radar freeboard based on the sea ice age and other assimilated variables (e.g. PP), and more or less ignoring the raw retracked elevations.  I therefore don't think it's reasonable to describe the neural network as applying a `correction' or 'calibration' to the raw values, since the resulting values are so unrelated.

To illustrate how the sea ice age is a stronger determinant of the corrected Rfb than the raw Rfb, I've plotted the r-values for each month as a histogram below. As you can see, the r-values between corrected Rfb and age are much larger than the corresponding corrected vs raw values.

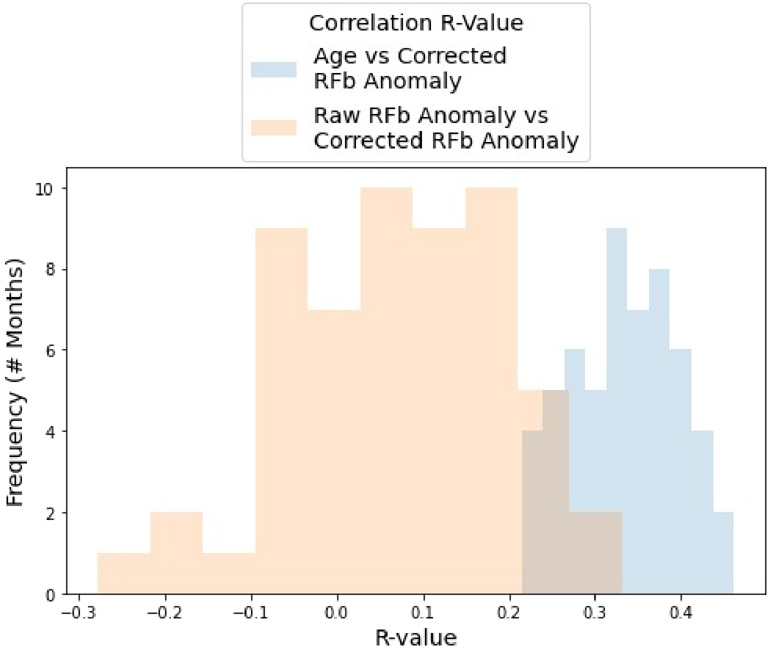


*Illustration 2: Correlations between age and corrected radar freeboard (blue) and correlations between raw radar freeboard and corrected radar freeboard (orange).*

To further investigate, I took all the years of a given month and plotted the raw anomalies against the corrected anomalies. Again, there's no strong relationship. My major concern is therefore this: to what extent can you claim that the resulting product is a corrected or calibrated retrieval when it doesn't reflect the variability in the raw, retracked values?

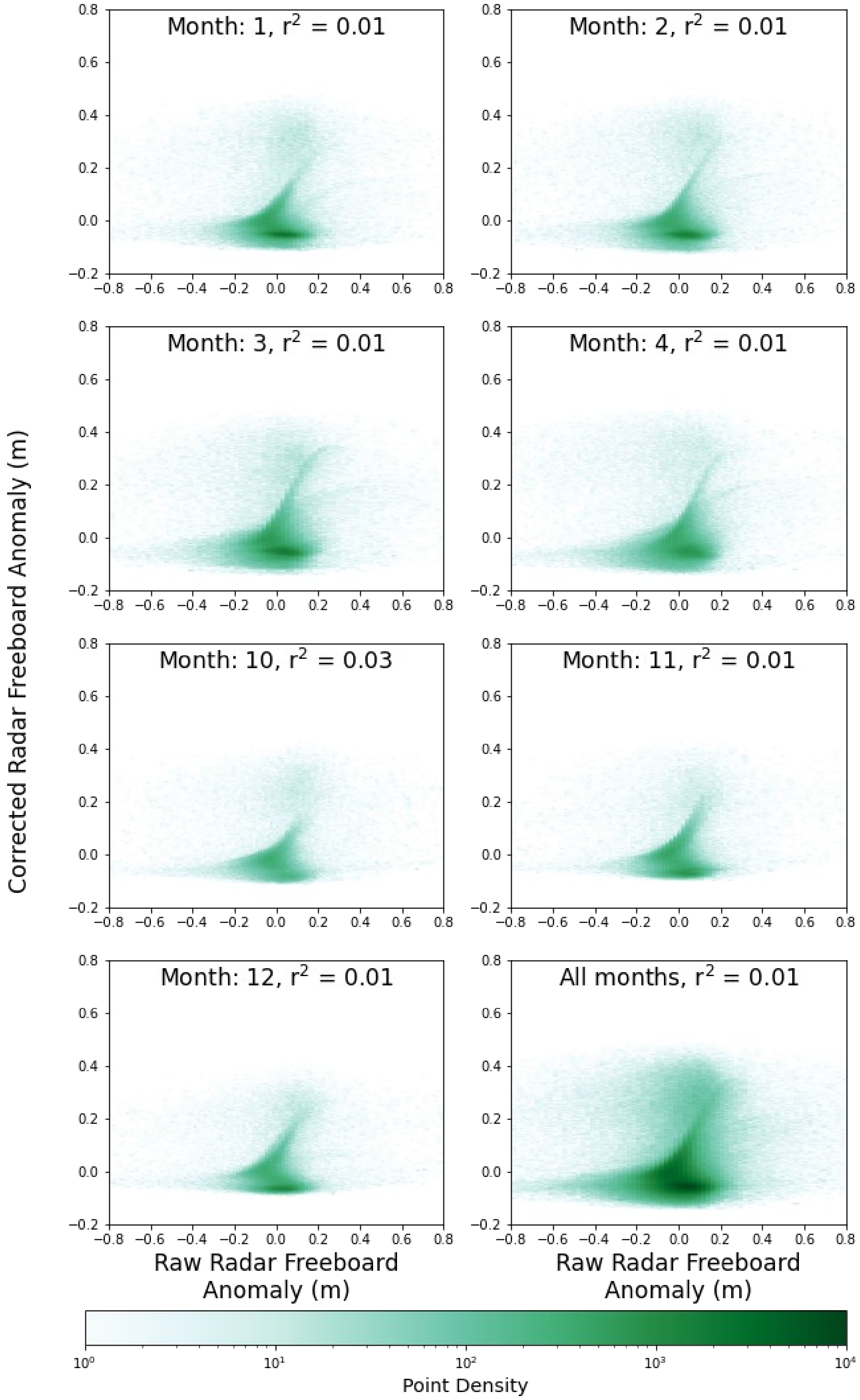


*Illustration 3: Relationships between the monthly raw radar freeboard anomaly and the corrected radar freeboard anomaly. It seems that ice with a higher-than-average raw radar freeboard in a given month does not end up with a higher-than-average corrected radar freeboard value.*

In my above analysis I have considered anomalies – for each month in the netcdf files I calculated anomalies of the raw and corrected radar freeboards, such that I could investigate whether higher radar freeboards in a given month translated into higher corrected values.

I now just consider the absolute values of all 3,054,312 valid measurements presented in the data. To do so I took the raw and corrected RFb value of every grid cell in every month and differenced them to get the size of the correction. Doing this shows there is still essentially no relationship between the raw (input) Rfb and the corrected (output) Rfb, the r-value is 0.094 (not shown). This again is concerning – it implies to me that larger measured radar freeboards are simply not translating to larger corrected freeboards. If this isn’t happening then what is the point of measuring the radar freeboard? The end RFb value seems to be entirely dictated by variables other than the raw Rfb. Once again, the r-value between age and corrected Rfb is much larger (0.33).

The result of all this is that there’s a highly linear, negative relationship between the raw radar freeboard and the correction size (r=-0.82). I fear the neural network has learned to shift the input distribution and then scale down high input values and scale up low values such that destroys any information encoded in the raw Rfbs.

Along these lines, it is telling that the raw radar freeboards are weakly correlated with sea ice age, in fact the r-value is negative (r = -0.087). This suggests to me that the raw radar freeboards do not currently represent a geophysical signal that can be ‘corrected’. Instead, the signal:noise ratio from TFMRA processed ERS2 waveforms may just be too low to be useful.

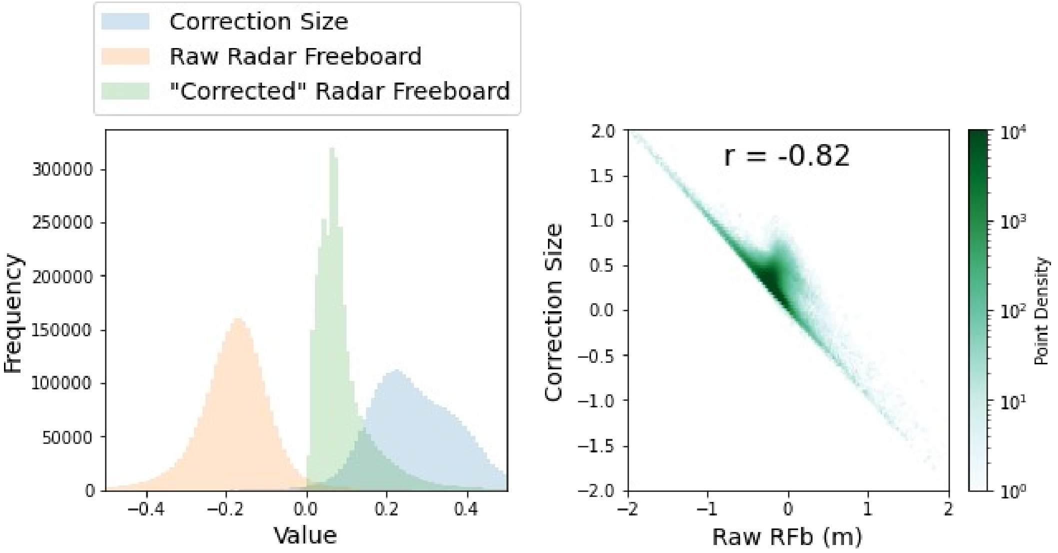


*Illustration 4: Left: input, correction, and output. Right: the highly linear nature of the correction that's applied. Seems to squash all signal out of the TFMRA50 RFbs.*

## **Summary**

Based on the above analysis, it appears that almost nothing of original radar freeboard measurement remains in the ‘corrected value’. If this is the case, I think the method in this paper needs to be fundamentally revisited, as the retracked elevations from ERS2 **must** have an impact on the final product if it is to be honestly presented as a “radar freeboard product”. If the authors feel that a product based primarily on different data (e.g. sea ice age or PP) is more appropriate, then this should be clearly presented and the product renamed to unambiguously indicate this.

I have a couple more comments, but we should focus on the above first. But just as as heads up:

- 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?
- 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.
- 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.
- 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.
- One additional concern is that Garnier et al. (2022) uses the same neural network approach (citing this discussion paper). If I am correct in my analysis above, I think the authors should be prepared to re-analyse the data and show that their Envisat TFMRA50 RFb values have a practical impact on their calibrated RFb values. I think this situation may be better, as the calibration from Envisat → CS2 is more direct than for ERS2 → EnviSat → CS2.