EGUsphere, referee comment RC3
https://doi.org/10.5194/egusphere-2022-214-RC3, 2022

Comment on egusphere-2022-214

Anonymous Referee #3

Referee comment on "Arctic sea ice radar freeboard retrieval from ERS-2 using altimetry: Toward sea ice thickness observation from 1995 to 2021" by Marion Bocquet et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-214-RC3, 2022

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.

General comments:

- 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.

- 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)

- 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 comunity.

- 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.

- 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!

- 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.

- While not a native English speaker myself, I further suggest some general English language editing before publication.

Specific Comments:

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

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

L284: This should be ‘the trained NN’ not the ‘the NN trained’.

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

Figure6: This definitely needs a much larger figure caption!