Synopsis

This work seeks to address several open research questions surrounding the amount of snow on sea ice, the ability of the CryoSat-2's radar waves to penetrate snow, and radar waveform interpretation. Reducing one of these uncertainties may move the subsequently retrieved value further away from "the truth" due to the presence of compensating biases, so it makes sense to consider several sources of uncertainty at once and then optimise. This work is therefore well motivated.

In this submission the authors have interchanged several aspects of the radar-freeboard to sea-ice-thickness processing chain. In particular, they have used two different "off the shelf" radar freeboard products (AWI and Bristol/UIT), several snow products (AMSR/W99, SnowModel-LG, NESOSIM, TOPAZ4, FY3B/MWRI), and have compared the assumption of full radar penetration of the assumed snowpack to an ice-type dependent radar penetration factor. These radar penetration factors were derived from comparison to several separate (but often not independent) data sources such as airborne estimates of sea ice freeboard/thickness and drifting ice-mass-balance buoys (Fig. 3). Having made these various interchanges, the authors appear to compare their new SIT data to similar evaluation data as before, which seems like a conflict between training and testing (Table 2; Fig. 5).

To be honest, I found it difficult to identify any clear conceptual or practical advances in this work. I was also left somewhat doubting whether the claims in the abstract and conclusions were supported by the results. It seems to me that rather than a "comprehensive optimisation", this work represents several experimental recombinations of different existing components in the processing chain. Furthermore, the authors' recombinations are potentially improving one metric of skill (e.g. RMSE) at the expense of another metric (e.g. correlation R value). I also note that not all combination possibilities were explored.

I was also concerned about the treatment of OIB QuickLook data as "in situ validation" (Table 2; Figure 5), since that particular data set has several known issues and is of course an airborne remote sensing product (not in-situ). Similarly, AWI IceBird and CryoVEX airborne data are also referred to and treated as "in situ validation", but are of course also the results of airborne remote sensing with their own biases and uncertainties. Particular to CryoVEX, it is **not** appropriate to use Ku-band ASIRAS data processed with the assumption of full radar penetration to then work out CS2 penetration factors.

Relatedly, the issue of how representative an ice-mass-balance buoy's measurement is of the ice sampled by (a) a CryoSat-2 SAR footprint, or (b) the ice in a 25x25km grid cell such as those analysed here, was mentioned (Sect. 2.4) but not meaningfully tackled. These issues strongly affect the suitability of IMBs for radar-altimetry evaluation. More broadly, the extent to which the authors may be optimising towards an uncertain or even biased truth is not discussed.

Finally, it was not possible for me to reproduce this analysis since it relies substantially on an unpublished snow dataset from the FY3B radar altimeter (which one coauthor has previously led a publication on). Since the manuscript relies so heavily on publicly available data, I thought it was a shame not to act in this spirit. The reason for the data not being made available was not given (despite an explanation being required by the Copernicus data policy). On a related note, no code was made available with this manuscript for review. These factors will limit the impact of any publication, and also limit my confidence as a reviewer.

It is therefore my recommendation that this manuscript should be rejected, and a similar analysis reperformed and resubmitted in service of a more carefully constructed research question. I appreciate this might be received as an overly negative recommendation, but I believe our community has a long way to go until we can meaningfully conduct the comprehensive optimisation marketed in this manuscript. Such an exercise would require reprocessing of several key datasets by many different research groups for a truly "apples to apples" comparison, not to mention a deep new look at the sampling biases, systematic uncertainties and independence of airborne and in-situ evaluation data.

Major Concerns

Operation IceBridge snow depth, freeboard and thickness data are not "in-situ validation"

Firstly (and this is a somewhat trivial point), they are not "in-situ" because they are taken from an aircraft. They are much better described as data products from airborne radar (and sometimes lidar) remote sensing.

Less trivially, because they're radar-based products, they're subject to significant uncertainties involving radar penetration of snow, sea ice roughness, and particularly sidelobes. Much has been written about OIB snow depth retrievals, and the quicklook product that the authors have relied upon in several sections is notoriously poor. In particular, it has relatively poor performance over multiyear ice (King et al., 2015), and the underlying algorithm suffers from a persistent issue of misidentifying range sidelobes from the snow-ice interface as the snow-air interface (Kwok and Maksym, 2014; Kwok and Haas, 2015). Figure 4 of Kwok et al. (2017) illustrates that the GSFC-NK product (which I think corresponds to the OIB QL product?) shows persistently low snow depths. In any case, the figure definitely shows that several very different interpretations of the OIB raw data exist, and therefore any individual snow or ice product cannot represent some kind of "in-situ" truth.

Even in the case where there are no snow-penetration, roughness or sidelobe-related biases, OIB sea ice thickness estimates still rely on exactly the same uncertain conversion of freeboard to thickness as CryoSat-2, so cannot form an independent benchmark for optimisation. For instance, I was hoping for some analysis of how the snow and ice densities used the in OIB hydrostatic conversions compared to those being used in the satellite conversions at hand, but that was glossed over. This type of material must be foundational to any activity that aims to "optimise" satellite retrievals to these OIB products. Otherwise the optimised product be biased in all the ways in which the OIB products are biased. Along these lines, describing OIB thickness products as "observations" is also dubious. The confidence with which you can "observe" SIT even with accurate measurements of snow depth and freeboard is nicely documented by Alexandrov et al. (2010; see numbers in second half of abstract).

CryoVEX Data

The authors need to be much more specific and thoughtful about how CryoVEX data on total thickness are generated – it's not enough to say "airborne electromagnetic sensors". If the CryoVEX method is similar to CryoSat-2, it will suffer from similar biases as CryoSat-2 and therefore is unsuitable for independent evaluation.

Furthermore, it appears that the authors have used a very limited subset of CryoVEX data from 2014. Aside from limiting the applicability of their results, my understanding is that SIT for the CryoVEX '14 campaign is from combination of the Ku-band ASIRAS system and the ALS system (Hvidegaard et al., 2015). Given the study by Willatt et al. (2011) on ASIRAS radar penetration through snow on sea ice, I think a lot more consideration needs to be given to the positioning of this data as some kind of objective truth that can be used to optimise satellite data. It's also worth pointing out that if you look at Fig. 9 from Garnier et al. (2021) which analyses CryoVEX 2017 data, it's clear that the difference between the ALS and ASIRAS data is often negative towards the end of the track, further indicating that the "off-the-shelf" approach to CryoVEX ALS/ASIRAS data in this manuscript may need considerable further scrutiny. At an absolute minimum, the authors should understand they cannot get at the CS2 radar penetration factor using Ku-band CryoVEX data that assumes a 100% penetration factor from ASIRAS. When comparing CS2 freeboards to ASIRAS freeboards derived from the assumption of a 100% RP, is it any wonder that we find high CS2 radar penetration factors?

Ice mass balance buoys

There are major issues with using single IMBs to validate measurements by radar altimeters such as CryoSat-2. IMBs are generally deployed on level ice, and are of course point measurements. This is in contrast to as CryoSat-2 footprint which is several hundred metres long and several kilometers wide depending on the ice roughness. And of course this is very different to the 25x25km scale on which this manuscript's analysis is conducted. It's also worth considering that IMBs essentially break as soon as the ice on which they’re situated begins to grow dynamically, and dynamic growth can be quite a significant contributor to the thickness in a CS2 footprint and also in a 25 km grid cell. So much more serious consideration needs to be given to how and whether the IMB data correspond to that collected by CryoSat-2, and whether they are suitable for a tuning/optimisation exercise such as this.

Radar Penetration Factors

I assume 0.77, 0.96, and 0.91 for the penetration factors were generated from the means of the histograms in Fig. 3A (not the modes, and not the mean of the Gaussian fits) – this should be clarified. More information is also required on how the “all data” histograms are constructed. This is because there are presumably many more data points from some sources than from others: were their contributions weighted for measurement reliability or quantity? It seems like the only reason that FYI penetration is significantly lower than MYI is because of the contribution of IMB data (panel f), but this is probably related to the fact that IMBs have a distinct sampling bias relative to both the other data and their surrounding environment (discussed above). So is it reasonable to allow the derived penetration factor to be so strongly influenced by the IMB contribution? This again goes back to how different sources of evaluation data should be weighed against each other in terms of reliability and quantity.

With the exception of the IMB panel, the histograms generally exhibit a large number of data points with penetration factors >1. This is often >1/3 of the data in a histogram. Since this is unlikely to be physical, it points to other biases existing in the radar altimetry processing chain that need to be accounted for before this manuscript’s method will actually work to retrieve the real penetration factor. At the moment, the penetration factor is simply being used as a tuning parameter to make the satellite products match the evaluation data. When deployed in this way the penetration factors are just making up for insufficiencies in the retracking approach/ snow data /hydrostatic conversion, but are also being tuned towards evaluation data that may themselves be biased. This is a totally different concept to what the authors claim to be doing elsewhere in the manuscript, i.e. recovering the penetration characteristics of Ku-band radar waves into snow.

How do the authors justify the width of the uncertainty bars in Figure 15? The shaded bars around the line seem surprisingly thin for a quantity that’s eluded the CryoSat community so effectively over the last decade. Could it really be the case that the penetration-factor consistently goes up in January and then back down in Feb by a magnitude considerably larger than the uncertainty bars? What is the spatial range of applicability of these statistics? How are the printed numbers of distinct in-situ observations generated? After all, there are hundreds of thousands of SnowRadar/ASIRAS waveforms and ALS spot heights going into this analysis, So I guess the figures in the thousands correspond to the number of contributing 25x25km grid cells? If more observations go into an aggregated 25x25 km data point, then we should probably weight that data point’s reliability as being higher than a grid cell that contains only a few data points. Has this been factored into the uncertainty analysis?

I have several further questions about the radar penetration factor analysis, both concerning how it was done and how it is reported, but will leave it here.

Radar freeboard data and retracking

I don’t think it’s legitimate to claim to have applied a “comprehensive optimisation of an improved retracking algorithm” (from the abstract) when no retracking takes place in the manuscript. This is misleading to the casual reader. A comparison has not been made between retracking algorithms (TFMRA & LARM), but instead between two radar freeboard products (AWI & Bristol/UIT). To imply that differences in the RF products only represents differences in the retracking approach is a risky business. The authors state “although the different classifying waveforms, geophysical corrections, and sea level tie-point interpolation also contribute to a relatively small extent (Landy et al., 2020).” That’s not exactly my understanding from reading that paper; Figure S3 shows there definitely are some differences that derive from the “in-house” treatments of AWI vs Bristol/UIT. That’s the reason that Landy et al. 2020 emulated the other retrackers with the same processing: to learn about the retrackers in isolation. I think that’s the approach that should be taken here if the retrackers are to truly be “optimised” rather than the radar freeboard products.

Optimisation Metrics

Firstly the metric referred to as *R* needs some clarification. Is this the Pearson product-moment correlation coefficient? *R* is also often used for the coefficient of determination, which essentially captures the data's deviation from the line y=x (i.e. it also captures the bias, where Pearson does not).

Using the product-moment correlation coefficient has its pitfalls. i.e. with heavy optimisation for Pearson you end up capturing the variance of your evaluation data, but at the cost of increasing your bias. This issue should be handled explicitly. These concepts are at the core of the "bias-variance tradeoff", and allied concepts in optimisation such as overfitting. While RMSE is sensitive to the bias, it also can be large when spread about the y=x line is large and the bias is small. So I think the metrics against which the processing chain is optimised need much more careful consideration when defining the optimisation exercise.

Some More Minor Things

The effect of reducing the radar penetration factor from unity to some fractional value is basically to reduce the subsequently calculated ice freeboard and thus the derived thickness. This does the same thing as assuming a deeper snowpack. Since the authors are varying the snow and the RP at the same time, they should grapple with this conflict explicitly: when we observe the sea ice to be "too thick" relative to our evaluation data, we often don't know whether it's because our assumed RP is too high, or whether our assumed snow is too thin. How can we optimise in light of this? One option is to just assume that airborne surveys of snow depth are correct, and use that snow depth to then optimise the RP. The authors haven't done this, but it might be interesting? But it does take a bit of a leap of faith regarding airborne snow depths.

Title: I think we should steer clear of subjective adjectives like "comprehensive" in titles. One could easily argue that this analysis is not comprehensive as it uses a fairly limited set of in-situ sea ice thickness measurements relative to those available, and uses limited subsets of CryoVEX and IceBird data. I recommend removal of this word.

L42: "Recent advances in satellite altimetry began in 2003." – This is a very subjective sentence and I'm not sure what it adds. Envisat began operating in 2002 for instance. What defines "Recent" here?

L68: Lower errors, not minimal.

L140: I think it would be better to cite the official NSIDC source for these data: https://nsidc.org/data/nsidc-0758/versions/1

L509: Perhaps I've missed this, but I don't think the authors have specified whether they're using the ERA5 or MERRA2 run of SnowModel? When comparing SnowModel and NESOSIM it's of course important to make sure they're forced by similar data, or else you're basically just comparing reanalysis precipitation data rather than the models themselves. Same with TOPAZ – I assume this is driven by some ERA-type product?

Fig. 1: There's no reference to the bathymetry in this paper, so I think shading it into this plot confuses things.

Fig. 2: If you label the "Reference ellipsoid" in this figure you should reference/explain it in the text.

L379: Think this sentence needs rewording for clarity.

L447: The authors should discuss their optimised radar penetration factors with reference to Nab et al. (2023), who derived RP factors based on the relationship between radar freeboard and SnowModel-LG depths.

L594: I was disappointed to see that the snow depth data from Li et al. (2021) are not made publicly available, so many of the results presented in this paper are not replicable by either me as a reviewer, or the community at large. As well as limiting my confidence as a reviewer, this will limit the impact of the paper since the supposedly optimised products are not available or reproducible in future. The Copernicus

Publications data policy states “if data are not publicly accessible, a detailed explanation of why this is the case is required”, and I can’t see why this was ignored. Furthermore, no code was made available with the manuscript, which also limits my confidence in recommending the paper for publication.