General comments

The paper aims to address several difficulties encountered in CS-2 (CryoSat-2) data analysis, including waveform retracking, snow depth estimation, and radar penetration. It is noteworthy that the paper summarizes a substantial amount of observational data for validation, compares snow reconstructions, and employs two CS-2 retracking algorithms to generate an ice thickness dataset. However, I share Reviewer 1's opinion that the paper lacks comprehensive optimization. Instead, it leans towards combining algorithms and products in a somewhat selective manner, which is inappropriate for a scientific paper, particularly one focused on datasets or products. Additionally, there is a consistent mixing of sensitivity studies and validation studies, making the paper challenging to read and diminishing its overall credibility. Furthermore, there is a critical need for clarification and correction regarding the definition of radar penetration and potential misunderstandings arising from previous papers.

I acknowledge the usefulness and instructive nature of referring to Armitage et al. (2015) to gain a quick understanding of the priorities concerning the CS-2 radar penetration problem. However, the algorithms employed in the current paper are essentially the same as those presented in Armitage et al. (2015), which calls for innovation and reorganization given the passage of several years. It is important to exercise caution, particularly in two areas: (1) the potential misuse and conflict arising from using training data for radar penetration estimation and validation data for sea ice thickness validation, an issue that Reviewer 1 has already highlighted; and (2) a genuine misunderstanding of how the paper defines the radar penetration factor. The authors seem to suggest that the radar penetration factor depends on the waveform retracking algorithms and snow depth product. However, this may lead to significant misunderstandings since the radar penetration effects are a result of snow and ice scattering, which, in turn, depend on the properties of the upper snow and ice layers, such as wetness, density, and grain size. Therefore, the radar penetration of CS-2 should be based on the properties of the snow and ice itself, rather than solely relying on the waveform algorithm or snow product.

If the authors aim to address the radar freeboard penetration or snow scattering problem, a more appropriate and physically robust approach would be to follow the methodology outlined by Slater et al. (2019), where the penetration depth over Greenland was derived, or refer to the study by Kurtz et al. (2014), which used a model to address the uncertainty of radar penetration. If the authors genuinely seek to reduce bias in a specific radar freeboard estimation, I would suggest the following steps: (1) select an appropriate snow product; (2) choose an appropriate radar freeboard estimation (or retracking algorithm); and (3) based on current in-situ observations, correct the selected radar freeboard bias. In this approach, the fundamental tasks involve selecting the proper snow product and radar freeboard estimation, which have not been adequately addressed in the paper. Additionally, the datasets used for correction and validation purposes lack clarity. Therefore, I strongly recommend that the authors consider changing the term "radar penetration factor" to "radar correction coefficient" to better align with their protocols and enhance the overall structure of the paper.

Based on the aforementioned concerns and the following comments, I suggest rejecting this manuscript. Here are my main comments:

(1) Line 15, '*applying a comprehensive optimization of an improved retracking algorithm, corrected radar penetration rate*'. When reading this line, it gives the impression that the paper will introduce a genuinely improved retracking algorithm, such as better waveform fitting or more reasonable treatment of different ice types. But when I looked through the

paper, the whole method is the combination choices from different productions, which is kind of disappointed. This is NOT 'improved retracking algorithm' you wrote in the abstract.

(2) Line 68, *'found that the radar freeboard derived from the LARM has minimal errors compared'*, the authors must be very clear here that the validation from Landy et al., (2020) is based on the OIB 2011, 2012, and 2013 L4 NSIDC product.

(3) Line 80, *'Second, the calculation of radar…'*. As discussed in the previous section, it would be more appropriate to refer to it as the "radar correction coefficient" rather than the actual radar penetration. The real radar penetration is dependent on the properties of the snow and ice, not the retracking algorithm.

(4) Line 81, *'Because…, the radar freeboard errors were transferred to the radar penetration rates estimation'*, this sentence appears somewhat unaware. The empirical method is not the cause of radar freeboard errors or the existence of radar penetration rates.

(5) Line 99, '…we used LARM to replace TFMRA…', I don't see why the authors use 'replace' here since there is no consensuses on the algorithm choices until now.

(6) Line 102, 'For the snow depth, we…'. Up until this point, the authors have not highlighted any strengths of the FY3B/MWRI snow depth product. I would suggest that the authors incorporate the benefits of this product in the paragraph discussing snow depth.

(7) Line 103, *'Using the three improvements above, we ran four test cases—three individual and one combined…'*. Exercise caution when using the term "improvements" when there has been little discussion of their strengths.

(8) Line 188, *'The difference between AWI CS2 and LARM-derived radar freeboard is mainly due to the different retracking algorithms…'*, I am pretty sure this is NOT from the Landy et al., (2020), and in fact, what they did is aligning these filtering, corrections and schemes to focus on the effects from retracking algorithm itself. And they continued finding there still exist significant discrepancies from retracking itself. They NEVER said these filtering, corrections and schemes contributed to a relatively small extent. It is definitely sure that classification, waveform filtering, geophysical correction and se level tie-point interpolation exert nonnegligible effects on the final gridded radar freeboard product from each developer.

(9) Line 132, *'In this study, the MW99/AMSR2 was used in some optimization cases….'*, nstead of providing a vague explanation, the authors need to clarify where the MW99/AMSR2 dataset was used and the reasons for its inclusion. As of now, it appears that the optimization is limited to the four case studies. Therefore, calling them optimization schemes is questionable, especially considering the authors have not addressed the uncertainty associated with each product. Case studies CANNOT be equated to an optimization scheme.

(10) Line 135, I still do not understand why the authors also chose NESOSIM, SnowModel-LG, and TOPAZ4, since in Line 102, the authors mentioned the use of FY3B/MWRI. If the authors aim to compare different products to determine the best combination, they should refrain from stating that FY3B/MWRI is used for improvement in the Introduction part.

(11) Section 2.2, the whole section should have specific description of the spatial and temporal resolution used in this paper, e.g. monthly? Daily? Time span? From which month to month?

(12) Line 178, In the Data gridding section, the authors need to explain the data protocol for daily/subdaily datasets (NESOSIM, SnowModel-LG, TOPAZ4, and all observational data) and the monthly dataset (W99/AMSR2, CS-2). They should describe how these datasets are coordinated in this study, such as whether all datasets are averaged into a monthly

setting. Additionally, it is important to provide a clear explanation of the method used for spatial interpolation.

(13) Section 3.1, I have several questions about this section. As I understand it, this section calculates the radar penetration based on all observation radar/snow freeboard and CS-2 LARM radar freeboard, right? In that case, the total freeboard should be calculated from AWI IceBird and IMB ice thickness and snow depth datasets. It is necessary to specify which density is used for these calculations. Furthermore, OIB products have their own protocols for calculating total freeboard. How are these protocols coordinated fairly or placed within the same context? Additionally, since you have already used the results of MYI and FYI penetration factors based on all observations, it is unclear why these datasets are used for further validation. It does not seem fair to use them again for validation, considering they were already used for radar penetration correction.

(14) Line 258, *'The differences in radar penetration rates…'*. Once again, it should be noted that the differences in radar penetration can be explained by factors such as frequency, sensor, and period, but not solely by the spatial resolution.

(15) Line 259, *'For example, for the OIB, the radar penetration rates may be applicable only in the spring.'*, so, you did not use the OIB from October to November, right? (That's why the clear information in datasets using in the Data and Method part is very important)

(16) Line 267, 'The relationship between FYI and MYI penetration rates supports the previous studies…', It is not clear why you consider all of these relationships to be consistent. Nandan et al. (2017) deduced a depth-dependent saline snow correction factor from observations, and Landy et al. (2022) used 0.9 as a first approximation due to the difficulty of quantifying snow cover changes between May and September. It would be helpful to provide further clarification on how these studies align with your findings.

(17) Fig. 4(a). It is intriguing why the snow depth from FY3B/MWRI is higher in October compared to November. Additionally, it would be beneficial to clarify whether Figure 4 represents Arctic basin-scale mean values. If so, it is puzzling why radar freeboard and thickness are larger in October than in November. Providing possible explanations for these observations would be valuable.

(18) From the Table 2 and Section 3.3, the improvements observed among different cases are only reflected in the RMSE, which is expected since you corrected or generally reduced the values based on the observations. However, it is frustrating that these four cases differ in at least two products, making it challenging for readers to make direct comparisons.

(19) Line 313-314, I assume you consider AWI CS2 as your baseline and aim to determine whether the results are better than AWI CS2. If that is the case, you should provide this context from the beginning. However, I have some concerns since the work now uses a completely different algorithm and observed-corrected coefficient for comparison, which may be unfair to AWI CS2.

(20) Line 310-333. Among the in-situ observations, only CryoVex provides actual independent validation. Upon closer examination of the third column in Figure 6, all cases show high correlation coefficients, and the combination cases reduce the RMSE by over 23% compared to AWI CS2. Therefore, there does not seem to be a significant improvement in the LARM+FY3B/MWRI+RP choice compared to the other cases. The differences lie in the slopes, but it is unclear whether you placed the retrieved data on the x-axis and the in-situ/real data on the y-axis. Mathematically, the x-axis in linear fitting should represent the true/validation data, or else there might be considerable uncertainty in data validation. Therefore, if you were to switch the axes, the slope would likely be different. Additionally, there is a concern that the LARM+FY3B/MWRI+RP combination might result in significant underestimation of sea ice thickness.

(21) Section 4. It is unclear what the main takeaway is from the entire Section 4, where numerous pictures and discussions focus on the differences between each combination, ranging from spatial patterns to spatial-temporal trends. Since the previous parts have already discussed the improvement in the optimization case, it seems unnecessary to include all combinations here and analyze their differences. This approach might cause readers to lose focus and miss the main points. Additionally, in the abstract, Section 4 is summarized in just one sentence stating that MYI ice thickness is decreasing, which is already quite obvious since lowering the radar correction would naturally reduce the ice thickness. To simplify the paper, it might be better to move Figures 9 and 10 to the Supplementary section.

(22) Line 455-459, by combining Equations (8) and (9), it is evident that there is a linear relationship between density and radar penetration, with ice density having a larger coefficient than snow density. It would be beneficial to see more uncertainty quantification, such as considering the combined effects of IMB and LARM, and the uncertainties associated with radar penetration derived from observed ice thickness, snow depth, ice density, snow density, and radar freeboard.

(23) Line 474-477. The temporal sampling is a significant concern in the paper. As mentioned, only IMB data was used from October to February, which raises questions about representativeness and could compromise the results. It would be helpful to provide further explanation on this issue.

(24) Figure 15, I am very curious how the radar penetration rates vary from year to year. Including such information in the figure would be valuable.

(25) Section 5.2. Like I suggested before, it is important to combine the sensitivities from all parameters. However, it is unclear whether this section focuses on the sensitivity of radar penetration or sea ice thickness. If it is about radar sensitivity, then it is unnecessary to bring up other snow products and their effects, as you have already compared them earlier. It would be better to concentrate on the uncertainties of the FY3B/MWRI snow product in relation to radar penetration. If you also want to discuss the sensitivity study of ice thickness, you should systematically address the uncertainties associated with LARM radar freeboard, FY3B/MWRI snow product, density choice, and derived radar penetration. Additionally, in Figure 16, you introduce another validation on sea ice thickness, which is confusing. It is unclear whether this figure is part of the sensitivity analysis or a validation study.

(26) Section 5.3, once again, it is crucial to clearly distinguish between sensitivity studies and validation studies. When discussing the uncertainty of density on sea ice thickness results, it is important to recognize that this pertains to the density choice and its impact on ice thickness. You have already validated the results above and concluded that LARM+FY3B/MWRI+RP is the optimization case. Therefore, please utilize the validated results from earlier and avoid reintroducing these combinations here. Otherwise, it will confuse readers and undermine the confidence and trustworthiness of the previous results. Moreover, it is not appropriate to select densities or refer to them as an “updated density scheme” solely based on having lower RMSE than others after several rounds of validation. We want the paper to avoid cherry-picking results.