

Response to Referee #2

In this document, we outline our responses to comments from the referee, including modifications that we intend to make to the manuscript where necessary.

Referee comments are shown in black and our responses in blue.

This manuscript uses AMSR2 passive-microwave brightness temperatures, satellite-altimetry-derived snow depth, SnowModel-LG snow density, and OIB and MOSAiC observations to retrieve pan-Arctic sea-ice snow depth. It compares four machine-learning approaches (multiple linear regression (MLR), random forest (RF), LightGBM, and long short-term memory (LSTM)) and further evaluates the effect of the retrieved snow depth on sea-ice-thickness (SIT) estimation. The topic is within the scope of The Cryosphere. The use of multiple data sources, the comparison of different model architectures, and the feature-importance analysis are potentially valuable. I recommend accepting this article after moderate revisions.

We thank the reviewer for the positive assessment of the manuscript's scope, multi-source data integration, model comparison, and feature-importance analysis. We also appreciate the recommendation for moderate revision and the constructive comments, which have helped us improve the clarity, consistency, and robustness of the manuscript. In the revised manuscript, we have (i) corrected the inconsistencies in Table 3 and its caption; (ii) replaced the overly specific attribution of the internal-independent validation discrepancy to overfitting with a more cautious discussion of dataset shift, sampling differences, scale mismatch, reference-data uncertainty, and transferability; (iii) revised the interpretation of permutation importance to avoid causal or unique physical claims and added an explicit discussion of feature collinearity; (iv) softened the MOSAiC-based conclusions by acknowledging the low correlations and small MAE differences and by avoiding any definitive ranking of product accuracy; (v) revised the SIT discussion so that the results are interpreted as sensitivity to the snow-depth input rather than confirmation

that MW99 is the dominant error source; and (vi) standardized terminology, model naming, abbreviations, language, and figure/table captions throughout the manuscript. Point-by-point responses are provided below.

Major comments:

1. There appear to be inconsistencies between the table headings, the numerical values, and the descriptions in the text. For example, Table 3 contains an additional “ALL” column, whereas the numerical rows appear to contain seven values corresponding to LGBM, MLR, RF, LSTM, SM, Rost, and MW99.

Thank you for identifying this inconsistency. We carefully checked Table 3, the associated table caption, and all references to the table in the main text. The “ALL” column was removed, so that the number and order of column headings are now fully consistent with the numerical entries. The revised table now presents results in the following order: LGBM, MLR, RF, LSTM, SM, Rost, and MW99 in the revised manuscript.

2. The manuscript interprets the difference between the internal testing results and the independent OIB/MOSAiC results as a “classic indicator of overfitting.” This is possible, but the difference may also result from several other factors, including differences between the ASD and OIB/MOSAiC error structures, spatial and temporal sampling, scale mismatch, and differences in snow-depth distributions. The current experiments do not fully isolate overfitting from dataset shift or reference-product uncertainty.

We agree. Our original wording attributed the performance difference too specifically to overfitting. Although the discrepancy between the internal testing results and the independent OIB/MOSAiC evaluations may be consistent with limited out-of-sample generalization, potentially including overfitting, the

available experiments cannot distinguish this explanation from dataset shift and uncertainty in the reference datasets.

We have therefore revised the text to state that the reduced performance in the independent evaluations may reflect a combination of factors, including: (i) differences in the measurement principles and error characteristics of the ASD, OIB, and MOSAiC snow-depth datasets; (ii) differences in spatial and temporal sampling; (iii) point-to-grid and footprint-scale mismatches; (iv) differences in the snow-depth distributions represented in the datasets; and (v) potential overfitting or limited transferability beyond the training domain. We now explicitly state that the present analysis does not isolate the relative contributions of these factors.

We revised the subsection “Synthesis of validation results and implications for model selection” in the Results section to avoid attributing the discrepancy between the internal testing and independent OIB/MOSAiC evaluations solely to overfitting. We also revised the corresponding interpretation in the “Quantification and analysis of retrieval uncertainties” subsection and the Conclusions. The revised text now explicitly discusses dataset shift, differences in reference-data error structures, spatial and temporal sampling, snow-depth distributions, and scale mismatch, and states that the present experiments do not isolate the contribution of overfitting from these factors.

3. The Results repeatedly interpret high importance of PR(89), PR(06), and other features as direct evidence of specific snow microstructural processes, such as depth-hoar development and peak metamorphism. Permutation importance indicates the predictive contribution of a feature within a particular model and dataset. It does not, by itself, establish a unique physical mechanism, especially in the presence of highly correlated GR and PR variables and possible atmospheric effects at 89 GHz.

We agree with this important distinction. We have revised the interpretation of permutation importance throughout the manuscript. The revised text now clarifies that permutation importance measures the predictive contribution of a feature within the specific model, training dataset, and evaluation framework. It does not independently establish a unique physical mechanism.

We have removed statements that interpreted the high importance of PR(89), PR(06), or other predictors as direct evidence of depth-hoar development, peak metamorphism, or other specific snow microstructural processes. Instead, these variables are described as statistically informative predictors of snow depth in the present retrieval framework. We also added a statement noting that the importance rankings may reflect overlapping information among correlated GR and PR variables, as well as combined effects of snow properties, surface and ice conditions, and, particularly for the 89 GHz channel, atmospheric influences.

We revised the subsection “Optimizing input parameters for machine learning algorithms” in the Results section, the relevant discussion of feature importance, and the Conclusions. The revised manuscript now interprets permutation importance as model- and dataset-dependent predictive relevance rather than direct evidence of a unique physical mechanism.

4. Please replace causal or overly definitive statements such as “directly tracking,” “demonstrates,” “signifies,” and “providing direct evidence” with more cautious language. The authors should also mention the effect of feature collinearity on permutation-importance rankings.

We appreciate this suggestion and have systematically revised the wording throughout the manuscript. Expressions implying direct causality or definitive physical attribution, including “directly tracking,” “demonstrates,” “signifies,” and “providing direct evidence,” have been replaced with more cautious language,

such as “is associated with,” “is consistent with,” “suggests,” “may reflect,” “contributes predictive information,” and “is statistically informative in the present model.”

In addition, we added an explicit statement on feature collinearity and its implications for permutation importance. Several GR and PR predictors are derived from overlapping brightness temperature channels and are therefore correlated. Their permutation importance values may reflect shared or redundant predictive information rather than the independent contribution of an individual variable. Consequently, the importance ranking of one predictor may be reduced when correlated predictors remain available to the model, whereas its ranking may increase when those correlated predictors are absent or less informative. We therefore interpret the rankings as model- and dataset-dependent measures of predictive relevance, rather than unique measures of physical relevance or causal influence.

We added a statement on predictor collinearity and the interpretation of permutation importance in the Methods section and revised the relevant feature-importance interpretations in the Results and Conclusions.

5. The correlation coefficients against MOSAiC are very low for all products, and the LSTM correlation is reported as 0.13 with a p-value of 0.08. The manuscript correctly notes the point-to-grid scale mismatch, but it should not state that LSTM “most accurately captures the mean snow depth state” solely on the basis of a low bias and MAE, particularly when the differences in MAE among the products are small.

We thank the reviewer for this important comment. We agree that the low correlations with the MOSAiC observations do not support a definitive ranking of product accuracy or a strong statement that LSTM captures the mean snow-depth

state most accurately. In particular, the LSTM correlation is low and not statistically significant ($R = 0.13$, $p = 0.08$). Although LSTM has the smallest bias (0.30 cm) and a low MAE (5.57 cm), its MAE differs only slightly from those of RF (5.60 cm), SnowModel-LG (5.51 cm), and MLR (5.69 cm). Therefore, these metrics alone are insufficient to establish that LSTM is more accurate than the other products.

We have revised the text in the MOSAiC validation, the synthesis of validation results, and the Conclusions. The revised manuscript now states that LSTM showed the smallest mean bias and an MAE comparable to those of several other products. We further clarify that the weak correlations, together with the point-to-grid scale mismatch and small differences in MAE, prevent a robust ranking of product accuracy based on the MOSAiC comparison alone. The MOSAiC evaluation is therefore interpreted as an assessment of broad agreement in mean error metrics, rather than evidence that any product accurately reproduces point-scale snow-depth variability or local mean conditions.

6. The SIT results show a much smaller bias for the machine-learning-based estimates than for FT4 using MW99. This is an important result, but the conclusion that this “confirms” MW99 as the major source of operational SIT error may be too strong. Differences may also result from sea-ice density, snow density, freeboard processing, ice-type classification, matchup selection, and correlations between the snow depth and SIT reference products.

We thank the reviewer for this important comment. We agree that the lower bias of the machine-learning-based SIT estimates relative to FT4 should not be interpreted as confirmation that MW99 is the major source of error in the operational FT4 SIT product. Although the SIT sensitivity experiment was designed to isolate the effect of replacing the MW99 snow-depth input while holding the density assumptions and CryoSat-2 freeboard input constant, the

comparison does not account for all sources of uncertainty in operational SIT retrievals.

In particular, uncertainties associated with spatial and temporal matchup procedures, and possible dependencies between the snow-depth and SIT reference products may also affect the observed differences. We have therefore revised the manuscript to state that the results demonstrate a strong sensitivity of SIT bias to the choice of snow-depth input under the assumptions adopted in this study, rather than attributing the FT4 bias uniquely or predominantly to MW99. We also clarify that, although the machine-learning-based SIT estimates substantially reduce mean bias, their RMSE values (0.47–0.51 m) are comparable to that of FT4 (0.48 m). Thus, the experiment primarily indicates an improvement in systematic bias rather than a uniform improvement in all SIT error metrics.

Minor comments:

1. The title should consistently use “pan-Arctic.” such as CryoSat-2, AMSR2, and sea ice thickness should be used consistently throughout the manuscript.

We agree. The title and manuscript have been revised to use “pan-Arctic” consistently. We also standardized the terminology and spelling of “CryoSat-2,” “AMSR2,” “sea ice thickness,” and related expressions throughout the manuscript.

2. The manuscript requires comprehensive language and copy editing. Several language and formatting errors are present, including: “root-means-square-error,” which should be “root-mean-square error”; “horizontal,” which should be “horizontal”; “addictive linear interactions,” which should be “additive linear interactions”; “oceanatmosphere,” which should be “ocean–atmosphere”; “year-around,” which should be “year-round”; and “Cryosat2,” which should be consistently written as “CryoSat-2.”

We have conducted a comprehensive language and copy edit of the manuscript. The specific terms identified by the reviewer have been corrected.

3. Suggested revisions to key statements in the abstract: Original statement: The 89 GHz polarization ratio has a significantly greater influence on snow-depth retrieval over multi-year ice. Suggested revision: The 89 GHz polarization ratio showed greater predictive importance over multi-year ice than over first-year ice. Original statement: Our long-term snow products offer users a reliable, high-accuracy dataset. Suggested revision: The resulting long-term snow-depth products provide an observationally evaluated dataset for Arctic sea-ice and snow studies, with uncertainties that vary across ice types, seasons, and the extrapolation period.

We agree and have revised the abstract to avoid overstating the physical interpretation and reliability of the products.

The original statement:

The 89 GHz polarization ratio has a significantly greater influence on snow-depth retrieval over multi-year ice compared to that over first-year ice.

has been replaced by:

The 89 GHz polarization ratio showed greater predictive importance over multi-year ice than over first-year ice.

The original statement:

Our long-term snow products offer users a reliable, high-accuracy dataset for advancing Arctic energy budget modeling and sea ice studies.

has been replaced by:

The resulting long-term snow-depth products provide an observationally evaluated dataset for Arctic sea-ice and snow studies, with uncertainties that vary across ice types, seasons, and the extrapolation period.

4. Use consistent model naming and capitalization. Please use one form consistently:

LightGBM rather than alternating between LGBM and LightGBM, unless LGBM is explicitly defined as the abbreviation. Similarly, use consistent forms for MLR, RF, LSTM, OIB, MOSAiC, FYI, MYI, and SIT.

We have standardized model names throughout the manuscript. At the first occurrence, the following definitions are now used:

Light Gradient Boosting Machine (LGBM), Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM).

We use LGBM consistently thereafter, including in the tables and figure captions. The abbreviation “LightGBM” has been replaced by “LGBM” throughout the manuscript.

The following abbreviations have also been standardized: LGBM, MLR, RF, LSTM, OIB, MOSAiC, FYI, MYI, SIT.

We use “first-year ice (FYI)” and “multi-year ice (MYI)” at their first occurrence and use FYI and MYI consistently thereafter.

5. The manuscript contains several typographical and language problems, including: “generalisability” and “generalization” should be standardized according to the journal’s preferred English style;

We have standardized the manuscript to American English and use generalization consistently. Related terms have also been harmonized, including: generalization capability, generalization performance.

6. In Figure 5 caption, “Artic” should be “Arctic”.

The caption has been corrected.

7. Line 676, “retrievals..” contains an extra period;

The extra period has been removed.

8. Line 490, “The Rost product” should use lower-case “the” when it appears in the middle of a sentence.

The sentence has been corrected.

9. Line 514, “The value in parentheses represent” should be “The value in parentheses represents”.

The sentence has been corrected.

10. The captions of Figures 2–7 should provide sufficient methodological information. The figure captions should indicate whether the plotted values are daily or monthly, whether the distributions are based on all collocated samples, and whether the comparisons use the same spatial and temporal mask for all products. Figure 4 should also clarify the meaning of the “point of flawless agreement.”

We agree. Because two new figures (Figs. 2 and 3) were added in Section 4.1, the original Figures 2–7 have been renumbered as Figures 4–9. The captions of Figures 4–9 have been revised to specify the temporal resolution, the number of collocated samples, the use of common spatial and temporal masks, and the meaning of the reference line in Figure 4.

The revised captions are provided below.

Revised Figure 4 caption

Figure 4: Probability distributions of monthly snow depth from the OIB observations and seven snow depth products: (a) LGBM, (b) MLR, (c) RF, (d) LSTM, (e) SM, (f)

Rost, and (g) MW99. The comparison uses 3,167 collocated 25-km grid cell samples from the common spatial and temporal mask for which all products and the OIB observations are available.

Revised Figure 5 caption

Figure 5: Comparison of monthly gridded snow depth estimates from LGBM, MLR, RF, LSTM, SM, and MW99 with MOSAiC snow depth observations from October 2019 to April 2021. The comparison includes 192 collocated samples obtained using the same spatial and temporal mask for all products. The x-axis represents the number of valid grid points. The y-axis represents the corresponding gridded snow depth estimates. The Rost product is excluded because its temporal coverage does not overlap with the MOSAiC period.

Revised Figure 6 caption

Figure 6: Comparison of daily snow depth estimates from LGBM, MLR, RF, and LSTM with daily OIB snow depth observations. Panels (a–d) show the corresponding probability distributions, and panels (e–h) show the differences between the retrieved and OIB snow depths for 5,133 collocated samples. The same spatial and temporal mask was applied to all four products. In panels (e–h), the black line denotes zero retrieval error, corresponding to exact agreement between the estimated and OIB snow depths.

Revised Figure 7 caption

Figure 7: Spatial distributions of monthly mean Arctic snow depth (m) during the 2012–2013 growth season (October–April) derived from the four model retrievals (LGBM, MLR, RF, and LSTM). The maps were generated using the valid sea ice mask ($SIC \geq 80\%$) and the corresponding ice type classifications for each month.

Revised Figure 8 caption

Figure 8: Daily snow depth retrievals on the 15th of each month during the 2012–2013 growth season from the four models (LGBM, MLR, RF, and LSTM). The

maps use the same sea ice mask and ice type classifications as Figure 7.

Revised Figure 9 caption

Figure 9: Distributions of the differences between directly retrieved monthly snow depth and monthly mean snow depth calculated from the corresponding daily retrievals during the 2012–2013 growth season. The comparison is performed for the same valid 25-km grid cells and months for all model. The dashed and solid lines in each violin plot represent the median and mean differences, respectively.

