

Response to Referee #3

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 study uses four algorithms, including MLR, RF, and LSTM, to retrieve snow depth on Arctic sea ice and further estimate sea ice thickness. However, the authors do not identify an optimal method, and the study appears to produce four separate datasets. The authors should clarify the applicability of each algorithm in the abstract, conclusions, and other relevant sections. Such recommendations should not be based solely on the overall comparison with OIB, but should also consider differences in ice type, season, and temporal resolution. In addition, several figures, tables, and descriptions require further improvement.

We thank the reviewer for the constructive and insightful comments. We agree that the four retrieval algorithms should not be treated as having a single universal ranking. Their relative performance depends on the evaluation dataset, spatial and temporal scale, ice type, season, and selected performance metric. We have therefore revised the Abstract, Results, Discussion, and Conclusions to describe the four products as complementary rather than to identify one universally optimal algorithm.

Specifically, we added a model-applicability summary that distinguishes: (i) performance on the ASD-derived internal test set; (ii) agreement with monthly and daily OIB observations, including separate FYI and MYI evaluations and seasonal performance differences; (iii) mean-bias behavior in the MOSAiC comparison; (iv) daily-to-monthly internal consistency; and (v) differences in temporal resolution. We also clarified that the four products are not intended as four independent datasets, but as complementary algorithmic retrievals that characterize structural uncertainty and support application-specific model

selection. In addition, we revised statements that previously implied that MLR, LSTM, or any other model was universally superior, clarified the roles and characteristics of the benchmark products, added a feature-importance and final-feature-selection figure, revised the figure and table captions, corrected and rechecked the MOSAiC validation statistics, and conducted comprehensive English editing.

Specific comments:

1. Please provide the full name of OIB when it first appears in the abstract.

Thank you. We have revised the Abstract to define OIB at its first occurrence as Operation IceBridge (OIB).

2. The abstract should be reorganized to describe the datasets, retrieval algorithms, feature selection, temporal coverage of the products, and main conclusions more clearly. At present, the abstract places too much emphasis on the comparison between MLR and OIB. This result only represents model performance for a specific validation dataset and set of evaluation metrics and does not fully reflect the main contribution of the study. Because the model rankings differ between the internal test and independent validation, the general expression "highest accuracy" should be avoided. The performance of the four models and the main characteristics of the four products should be summarized more objectively.

We agree. The Abstract has been reorganized to present the input datasets, retrieval framework, feature-importance analysis, temporal coverage, independent evaluations, and the complementary applicability of the four products. We removed the statement that MLR has the “highest overall accuracy,” because this result depends on the OIB validation dataset and selected metrics. The revised Abstract emphasizes that no single model is optimal in every evaluation setting.

Revised Abstract:

Snow depth is a critical climate indicator and a key parameter for Arctic sea ice retrieval. In this study, we retrieve pan-Arctic snow depth from 2012 to 2021 by integrating satellite altimetry, passive microwave brightness temperatures, and multi-source ground/airborne data. Four machine learning models, Light Gradient Boosting Machine (LGBM), Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM), are trained using altimeter-derived snow depth (ASD) for 2018–2021 and applied to generate daily and monthly 25-km products. Permutation importance indicates that snow depth and microwave gradient and polarization ratios provide complementary predictive information, with the 89 GHz polarization ratio generally more importance over multi-year ice than over first-year ice. In independent monthly OIB comparisons, MLR shows the lowest RMSE (7.19 cm) and highest correlation ($R = 0.67$), whereas daily OIB and MOSAiC evaluations show complementary model behavior; MOSAiC correlations are weak for all products and do not support a definitive ranking. The four products therefore provide complementary options rather than a universally optimal retrieval. In a sea ice thickness sensitivity experiment, replacing the Modified Warren 1999 snow depth input with the retrieved snow depth reduces mean thickness bias by nearly an order of magnitude, while RMSE remains comparable among products. The resulting observationally evaluated daily and monthly 25-km snow depth datasets support Arctic sea ice and snow studies, with uncertainties that vary across ice types, seasons, and the extrapolation period.

3. Section 4.1 discusses the importance of different frequencies and polarization parameters for different months and ice types, but no corresponding figure or table is provided. Please add a figure or table to support these descriptions and list the final parameters selected as inputs for each model.

Thank you for this constructive comment. We agree that the importance of different frequencies and polarization parameters should be presented more clearly. In the revised manuscript, we have added two new figures in Section 4.1:

The factors listed in Figure 2 are the final input parameters selected for the FYI model, and the factors listed in Figure 3 are the final input parameters selected for the MYI model. We have also clarified in Section 4.1 that these parameters were selected according to their importance rankings and are used in the corresponding FYI and MYI models.

4. Please clearly state which datasets are represented by SM, Rost, and MW99 in Lines 471–472.

Thank you. We have defined these abbreviations explicitly at their first occurrence and in the relevant table captions and Results text.

Revised text in Section 4.2:

For comparison, we evaluated three benchmark products: the SnowModel-LG snow-depth output (hereafter SM), the passive-microwave snow-depth product of Rostosky et al. (2018) (hereafter Rost), and the Modified Warren 1999 snow-depth climatology (hereafter MW99).

Revised Table 3 caption:

Table 3. Evaluation of monthly snow depth from LGBM, MLR, RF, LSTM, SnowModel-LG (SM), the Rostosky et al. (2018) passive microwave product (Rost), and the Modified Warren 1999 climatology (MW99) against OIB airborne snow depth observations. (RMSE/bias/MAE units: cm).

5. Please further discuss the main differences and advantages of the products developed in this study compared with Rost and other existing products, not only in terms of model validation results but also with respect to spatial and temporal resolution, temporal coverage, seasonal coverage, and other relevant aspects.

We agree. We added a dedicated comparison subsection to distinguish the products by input data, retrieval approach, spatial and temporal resolution, period of

coverage, seasonal availability, ice-type treatment, and intended use. We emphasize that the principal contribution of this study is not only a validation comparison, but also the extension of an altimeter-informed retrieval framework to a continuous daily and monthly pan-Arctic record for the 2012–2021 growth seasons.

6. Lines 531–532 suggest that some models accurately capture the snow depth at the MOSAiC locations based on their relatively small bias and MAE. However, the correlation coefficients of all products are very low. If the low correlations are mainly caused by the scale difference between point measurements and gridded products, the authors should clarify which aspects of product performance can be evaluated using MOSAiC data and weaken the conclusion about "accurately capturing" snow depth variations. In addition, please check the MW99 statistics in Table 4. The reported RMSE of 7.64 cm is smaller than both the absolute bias of 12.29 cm and the MAE of 12.70 cm. This is mathematically impossible, suggesting that some values or labels in the table may be incorrect.

We agree. We revised the MOSAiC evaluation to state explicitly that, because MOSAiC observations are point measurements while the evaluated products are 25-km grid-cell averages, the comparison is mainly informative for systematic differences in the sampled mean state and broad error magnitude. It cannot robustly assess point-scale spatial or temporal variability, and the low correlations prevent a reliable ranking of product accuracy at the MOSAiC locations.

We also thank the reviewer for identifying the inconsistency in the MW99 statistics. The reported MW99 RMSE of 7.64 cm is mathematically inconsistent with the reported absolute bias of 12.29 cm and MAE of 12.70 cm. We rechecked the collocated data, metric calculations, units, and table-transcription process. The corrected values are now reported in Table 4. (The corrected RMSE is 14.47 cm).

7. Table 2 shows that MLR performs relatively poorly in the internal test but performs well in the OIB validation. In Line 549, the authors explain this difference as possible overfitting of the more complex models. Since overfitting should normally be considered and controlled during model design and training, the authors should explain what measures were used to reduce overfitting and provide more evidence to support this explanation. If RF, , and LSTM are clearly affected by overfitting, the generalization and applicability of their products should also be reassessed.

We agree that the available comparisons do not isolate overfitting from dataset shift, reference-product uncertainty, sampling differences, and scale mismatch. We have therefore removed statements that described the internal-versus-independent performance difference as a definitive indicator of overfitting. Instead, we now characterize it as evidence that model rankings are sensitive to the validation setting and that the transferability of the retrievals must be interpreted cautiously.

To reduce overfitting during model development, RF and LGBM hyperparameters were selected using five-fold cross-validation and grid search. RF complexity was constrained through maximum tree depth and the number of candidate predictors used at each split. For LGBM, hyperparameters were selected using cross-validation, with a low learning rate and controlled tree-building settings. The LSTM architecture was restricted to a single layer with 64 neurons, and model selection was based on validation performance. Early stopping, dropout, and validation-loss monitoring were also applied during LSTM training. We have added these details to the Methods section.

However, cross-validation within the ASD-derived training dataset does not fully test transferability to OIB or MOSAiC because these datasets differ in measurement principle, spatial footprint, sampling distribution, and error structure. We therefore revised the Results and Conclusions to state that the four products should be interpreted as complementary, and that the independent evaluations do not provide definitive evidence that RF, LGBM, or LSTM are severely overfitted.

These measures, together with the independent evaluations, indicate that the ranking differences are more consistent with validation-setting dependence than with severe overfitting. Therefore, we did not reclassify these products as overfitted; instead, we emphasize their complementary applicability.

8. The English should be improved throughout the manuscript.

We appreciate this comment. The manuscript has undergone comprehensive language editing to improve grammar, sentence structure, terminology, consistency, and readability. We standardized terminology and capitalization, including “pan-Arctic,” “CryoSat-2,” “,” “Operation IceBridge (OIB),” “sea ice thickness (SIT),” “first-year ice (FYI),” and “multi-year ice (MYI).” We also corrected typographical errors, punctuation, hyphenation, and inconsistent American/British English usage.