Response to Referee #1

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 presents a study on snow depth retrieval over Arctic sea ice using multi-source satellite data and four machine learning models. The authors compare MLR, RF, LGBM, and LSTM, evaluate their performance against independent OIB and MOSAiC observations, and demonstrate the practical utility of their retrievals by improving sea ice thickness estimates. The work is methodologically sound, and the validation framework is rigorous. The long-term (2012–2021) snow depth product is an important dataset for Arctic climate studies.

We appreciate the reviewer's positive assessment and constructive comments on our manuscript entitled "Snow depth retrieval over Pan-Arctic sea ice (2012–2021) using multi-source data and machine learning models". These comments have been invaluable in improving the clarity and quality of our work. We have carefully addressed each point and provide detailed responses below.

Major Comments

1.The study employs four models (MLR, RF, LGBM, and LSTM) and reveals complex complementary performance—MLR performs best against OIB, while LSTM achieves the lowest bias against MOSAiC. However, the manuscript does not provide explicit guidance on how readers should select among these models for different application scenarios.

We agree with this constructive suggestion. To translate our comparative results into actionable guidance for end-users, we have added a model selection recommendation at the Section 4.2.3 of the revised manuscript. Based on the comprehensive validation and uncertainty analysis, we recommend the following model selection strategy: Given these inherent limitations, the choice among the four algorithms should be guided by the specific application priorities rather than by absolute performance rankings. MLR is preferred for large-scale mean state estimation and long-term consistency; LSTM is recommended when minimizing local bias is critical; RF offers a balanced trade-off between accuracy and robustness; while MLR remains the most interpretable option for studies requiring physical transparency."

2. The MW99 product exhibits very large bias (7.04 cm, Table 3), far exceeding that of all other products. The manuscript notes that MW99 is outdated, but does not elaborate on the physical reasons for its poor performance.

We agree that the physical rationale for MW99's poor performance should be made explicit. We have added a brief discussion in Section 2.3.4 explaining that MW99 is based on aircraft routing data from the 1970s-1980s and does not account for the rapid changes in Arctic snow distribution and ice type composition over the past four decades. This explains why it introduces systematic overestimation, particularly on FYI, where snow accumulation patterns have changed substantially.

We have revised the following paragraph in the Section 2.3.4:

"The Warren snow depth climatology (W99) has been widely used to calculate sea ice thickness (SIT) (Laxon et al., 2013). However, validation against in situ measurements has revealed that W99 significantly overestimates snow depth over FYI. This bias is partly attributable to the fact that W99 is based on historical in situ observations from Soviet drifting stations between 1954 and 1991 (Warren et al., 1999), and therefore does not capture the substantial changes in Arctic snow distribution and the shifting FYI/MYI composition observed over the past four decades (e.g., Stroeve et al., 2014; Lindsay and Schweiger, 2015). To address this issue, Kurtz and Farrell (2011) proposed a modified version of W99, which reduces snow depth over FYI by 50% to better represent contemporary Arctic Ocean conditions. This modified climatology, referred to as MW99 (Modified Warren 1999), has been adopted in several recent studies and has demonstrated improved accuracy. In this study, we use the MW99 snow depth data (2012-2021, gridded at 25 km resolution) as a benchmark for evaluating our snow depth retrievals."

3. The Introduction mentions both Rostosky et al. (2018) and Li et al. (2024) as existing snow depth products, but their roles in this study are different and may confuse readers.

We agree that this distinction should be made clearer. We have added a clarifying paragraph in the Introduction explaining that Li et al. (2024) is used as the training target (the ASD product, derived from CryoSat-2 and ICESat-2), while Rostosky et al. (2018) is used as a benchmark for comparison (purely passive microwave-based). We have also added a brief sentence noting the key methodological difference between the two products.

We have added the following discussion in the Introduction section, near the end of the Second graph of the revised manuscript:
"In this study, two existing snow depth products serve distinct roles. The altimeter-derived snow depth product from Li et al. (2024) is used as the training target (the ASD product), as it represents the accurate satellite-based estimate of pan-Arctic snow depth combining CryoSat-2 and ICESat-2 observations. By contrast, the Rostosky et al. (2018) product derived solely from passive microwave data is used as a benchmark for comparison with our retrievals, providing a reference point against which the improvement of our machine-learning-based approach can be assessed."

4. The models are trained on data from 2018-2021 (the ASD product) and applied to retrieve snow depth back to 2012. While the temporal extrapolation test (Section 4.3.1) demonstrates generalizability, it remains unclear whether the training period (2018-2021) is representative of the longer period. Did the snow-microwave

relationships remain stable across this entire period, or have there been systematic shifts due to changing Arctic conditions?

This is an important point. We have added a discussion in Section 5 addressing the temporal representativeness of the training data. We note that while there have been significant changes in Arctic sea ice and snow conditions over the past decade, the fundamental relationship between brightness temperature and snow depth particularly the strong sensitivity to PR(89) and snow density remains stable. We also cite literature confirming the stability of AMSR2 measurements over the 2012-2021 period. The successful temporal extrapolation test (Section 4.3.1) further supports the robustness of the models across this time span.

We have added the Section 5.3 in the Discussion:

" 5.3 Temporal representativeness of the training data

A question arises as to whether the training period (2018-2021) is representative of the full 2012-2021 record. Arctic sea ice and snow conditions have indeed undergone significant changes over this period, including declining MYI extent, earlier melt onset, and altered snow distribution patterns (e.g., Stroeve et al., 2014; Lindsay and Schweiger, 2015). However, the fundamental physical relationship between passive microwave brightness temperatures and snow depth remains governed by volume scattering processes, which are inherently stable over decadal timescales. In particular, the strong sensitivity of PR(89) to snow depth supports the temporal transferability of the learned relationships. This, combined with the relatively stable behavior of snow density in the microwave domain and the demonstrated stability of AMSR2 measurements over the 2012-2021 period, provides a strong physical basis for temporal extrapolation. Beyond these physical and instrumental justifications, the successful temporal extrapolation test (Section 4.3.1) provides direct empirical evidence. In that test, models trained on 2018-2021 data were applied to 2012-2017 and validated against OIB observations. The evaluation yielded consistent performance metrics, bias, mean absolute bias (MAB), and correlation coefficient (R),at both daily and monthly timescales, confirming that the retrieval quality is stable across temporal resolutions and throughout the 2012-2021 period. This confirms that the learned relationships are not period-specific and are robust across the full temporal domain of our product. Nevertheless, we acknowledge that the training period does not fully capture the extreme conditions of earlier decades. Future work will explore the incorporation of additional observational constraints, such as snow freezing depth or reanalysis-derived snow properties, to further enhance the long-term generalizability of the retrieval framework."

5. The manuscript's methodological contribution lies in the combination of systematic model intercomparison, rather than simply "applying four ML models to snow depth retrieval." However, the current framing may give readers the impression that this is just another ML-based retrieval study.

We thank the reviewer for raising this important point. We acknowledge that our initial framing may not have adequately conveyed the novelty of our work. In the revised manuscript, we have substantially strengthened the articulation of our methodological contribution.

To clarify, while previous studies have applied machine learning to snow depth retrieval (e.g., He et al., 2024; Zhou et al., 2025), they have primarily focused on product generation without systematically interrogating the behavior of different model classes under consistent conditions. Our contribution lies not in the application of ML itself, but in the systematic intercomparison framework that:

(1) evaluates four fundamentally different models (MLR, RF, LightGBM, LSTM) across three dimensions-accuracy, interpretability, and temporal generalizability;

(2) explicitly tests temporal extrapolation (trained on 2018-2021, validated on 2012-2017), which has not been reported in previous retrieval studies;

(3) provides a transparent benchmark that allows future studies to select models based on their specific priorities rather than relying on untested assumptions.

We believe this framework-level contribution, rather than a one-off product, offers transferable value to the broader cryospheric remote sensing community. To reflect this, we have revised the Introduction (Section 1, Lines 80-87) to more clearly state our novelty to emphasize the implications of our benchmark framework.

6. The authors should consider whether it is necessary to include a discussion of its role in ice thickness retrieval, as this study mainly focuses on snow depth retrieval.

We thank the reviewer for this comment. We agree that the core focus of this study is snow depth retrieval. However, we believe the sea ice thickness (SIT) application section is necessary for the following reasons: (1) The introduction explicitly identifies snow depth uncertainty as the dominant error source in satellite-derived SIT. Applying our retrieved snow depths to SIT estimation and evaluating the improvement directly tests the practical utility of our product and closes the loop on the problem raised in the introduction. (2) This exercise serves as a physical consistency check: the spatial correlation between snow depth and SIT (Figure 9) reveals that snow depth uncertainty most critically affects thin ice in marginal zones-finding that has independent scientific value. (3) The SIT application also aligns with The Cryosphere's focus on advancing the understanding of cryospheric processes, as it serves as a geophysical consistency check that demonstrates the physical plausibility and practical value of our snow depth product for improving sea ice thickness estimates- central parameter in polar climate studies.

Minor Comments:

1. The abstract states that the validation "reveals complementary strengths of snow retrieval among the models," but does not specify what these complementary strengths are.

We have specified the strengths briefly in the abstract. For example: "MLR achieves the highest overall accuracy (RMSE = 7.19 cm), while LSTM minimizes mean bias (1.98 cm against OIB; 0.30 cm against MOSAiC)."

2. Line 60, in the sentence introducing data fusion approaches, the term "bridges" should be changed.

We have replaced "bridges" with a more precise term "combines" to better reflect the methodological approach.

3. Line 112, the authors note that 89 GHz is "known" to be affected by atmospheric water vapor but do not cite a reference to support this claim.

We have added a citation (Lu et al., 2022) to substantiate this statement.

4. Table 3 shows seven products compared against OIB, but the layout is dense and could be made more readable.

We have grouped the product columns into two sections: "Our products (Machine learning retrievals)" and "Benchmarks (Existing products)" with a blank column between them for visual separation in the Table 3 of the revised manuscript.

5. Line 548, "This scale issue affects all products equally". This statement is slightly too strong. While it is true that all products suffer from scale mismatch, the degree of impact may differ. Some models (e.g., LSTM) might be more or less sensitive to the scale mismatch than others.

We agree that "equally" was too strong. We have revised this to "to a comparable degree" to acknowledge that while all products are affected by scale mismatch, the exact magnitude of impact may vary among models. The change has been made in the revised manuscript.

6. Line 574, "A black line represents the point of flawless agreement with OIB standards". "flawless agreement" is informal.

We agree that "flawless agreement" was too informal. We have revised the sentence to "A black line represents the 1:1 reference line" to use standard and objective language in the figure caption of the revised manuscript.

7. In section 5.1, the uncertainty analysis uses 1000 Monte Carlo iterations for non-linear models. The manuscript does not justify the choice of 1000 iterations.

We have added a brief sentence stating that “A sensitivity test showed that increasing the number of iterations beyond 1000 yielded negligible changes in the estimated uncertainties, confirming that this choice was adequate for stable convergence.”

8. Lines 674-676, the manuscript contains several long, multi-clause sentences that could be broken up for improved readability.

We have reviewed the following locations and splitted long sentences:

Lines 519-520: The comprehensive validation against the independent OIB and MOSAiC datasets reveals critical insights. These findings directly address the concerns regarding model performance and selection.

Lines 674-676: To ensure a direct comparison with FT4 and to isolate the effect of snow depth input, all other parameters were kept constant. Only the snow depth was substituted with our retrievals.

9. Figures 5 and 6 have relatively brief captions that do not fully explain what is being shown.

We have revised the brief captions of Figures 5 and 6 in the revised manuscript:

Figure 5: Spatial distributions of monthly Arctic snow depth (m) during the 2012-2013 growth season (October-April), retrieved by LGBM, MLR, RF, and LSTM. Each panel represents the multi-model ensemble mean for the corresponding month.

Figure 6: Same as Figure 5, but for daily snow depth retrievals on the 15th of each month during the 2012-013 growth season. Each panel shows the snow depth on the 15th day of the corresponding month from the four models (LGBM, MLR, RF, and LSTM).

10. The language still needs further polishing, and consistency should be improved, such as ensuring that symbols are used consistently throughout the manuscript.

We thank the reviewer for this careful reading and for pointing out these issues. We have thoroughly revised the language throughout the manuscript to improve readability and scientific precision.

References

Lindsay, R. and Schweiger, A.: Arctic sea ice thickness loss determined using subsurface, aircraft, and satellite observations, The Cryosphere, 9, 269-287, https://doi.org/10.5194/tc-9-269-2015, 2015.

Lu, J., Scarlat, R., Heygster, G., and Spreen, G.: Reducing weather influences on an 89 GHz sea ice concentration algorithm in the Arctic using retrievals from an optimal estimation method, J. Geophys. Res.-Oceans, 127, e2019JC015912, https://doi.org/10.1029/2019JC015912, 2022.

Stroeve, J., Barrett, A., Serreze, M., and Schweiger, A.: Using records from submarine, aircraft and satellites to evaluate climate model simulations of Arctic sea ice thickness, The Cryosphere, 8, 1839-1854, https://doi.org/10.5194/tc-8-1839-2014, 2014.