https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

Snow depth retrieval over Pan-Arctic sea ice (2012-2021) using multi-source data and machine learning models

Mengmeng Li[1] , Jianwei Ma[2,*], Yang Li[3], Juha Karvonen[4], Bin Cheng[4], Yingfei Wang[1], Haili Li[5], Yafei Nie[6] and Zheng Duan[7]

[1]School of Mathematics and Statistics, Zhengzhou University, Zhengzhou, China;
[2]Institute of Water Resources and Hydropower Research, Beijing, China;
[3]University of Helsinki, Institute for Atmospheric and Earth System Research / Physics, Helsinki, Finland
[4]Finnish Meteorological Institute, Helsinki, Finland
[5]School of Geography and Ocean Science, Nanjing University, Nanjing, China;
[6]School of Atmospheric Sciences, Sun Yat-Sen University, and Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China
[7]Department of Physical Geography and Ecosystem Science, Lund University, Sölvegatan 12, SE-223 62, Lund, Sweden

Correspondence to: Jianwei Ma (majw@iwhr.com)

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. We employ four machine learning models—Light Gradient Boosting Machine (LightGBM), Multiple Linear Regression (MLR), Random Forest (RF), and Long Short-Term Memory (LSTM)—to leverage the complementary strengths of altimetry and microwave datasets while evaluating the performance of different machine learning (ML) architectures. Through permutation feature importance analysis, we identified that 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. Validation against Operation IceBridge and MOSAiC measurements reveals complementary strengths of snow retrieval among the models. The MLR model achieves the highest overall snow depth accuracy (root-means-square-error = 7.19 cm, correlation = 0.67 against OIB), while the LSTM demonstrates minimal mean bias of snow depth between satellite-based and in situ observations (1.98 cm against OIB; 0.30 cm against MOSAiC). All ML models exhibit robust generalization capabilities. Our retrieved snow depth products improved sea ice thickness estimation significantly, reducing bias between satellite-based and a standard climatology-based ice thickness product by nearly an order of magnitude. Our long-term snow products offer users a reliable, high-accuracy dataset for advancing Arctic energy budget modeling and sea ice studies.

1 Introduction

Snow on Arctic sea ice is a critical component of the climate system. Its high albedo and low thermal

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

conductivity regulate the surface energy balance, modulating both ice growth in winter and melt in summer (Perovich et al., 2002; Liston et al., 2020). The depth and distribution of snow further exert dynamic and thermodynamic controls on sea ice thickness (Koo et al., 2021, SIT), making it an essential variable for mass balance studies. Crucially, accurate snow depth is a prerequisite for retrieving SIT from satellite altimetry, with uncertainties in snow depth contributing dominantly to the total thickness error (Zygmuntowska et al., 2014; King et al., 2018). However, obtaining reliable, pan-Arctic snow depth estimates remains a significant challenge.

Satellite remote sensing offers the only viable means for basin-scale monitoring. Traditional approaches primarily rely on empirical relationships between in-situ snow depth and passive microwave brightness temperatures, using linear regression (Markus and Cavalieri, 1998; Comiso et al., 2003; Markus et al., 2006; Kilic et al., 2019). Although widely used in practice, these methods are limited by their dependence on sparse ground data and perform poorly for deep snow, wet snow, and over complex multi-year ice (MYI) (Markus et al., 2006). The incorporation of lower-frequency brightness temperatures (e.g., 6.9, 10.7, 1.4 GHz) has extended retrieval capabilities (Maaß et al., 2013; Rostosky et al., 2018; He et al., 2022). However, fundamental challenges persist due to the non-linear and ice-type-dependent interactions between microwave radiation and snowpack properties, which recent studies have aimed to address through advance radiative transfer modeling and machine learning techniques (Li et al., 2022; Rückert et al., 2023). A distinct, physics-based pathway involves the synergistic use of multi-sensor observations within a coupled framework. By combining radar altimeter freeboards with thermal infrared and passive microwave measurements, these methods solve a system of physical equations including hydrostatic balance and wave speed correction to simultaneously retrieve snow depth, SIT, and density (Lee et al., 2021; Shi et al., 2023). This approach provides physically consistent multi-parameter estimates but relies on the accuracy of its underlying assumptions and the precise synergy of inputs from different satellite platforms. Active remote sensing, notably radar and laser altimetry, provides another physical pathway by directly measuring snow and ice freeboards (Armitage and Ridout, 2015; Guerreiro et al., 2016; Lawrence et al., 2018; Kwok et al., 2020). The advent of ICESat-2 has enabled the derivation of pan-Arctic snow depth with reliable quality (Kacimi and Kwok, 2022; Li et al., 2024). However, this altimeter-based record is inherently short, extending only from 2018 onward, and lacks the long-term continuity needed for climate studies.

To bridge these complementary gaps—the long-term coverage of passive microwave data and the physical accuracy of altimetry—recent studies have turned to machine learning-based data fusion. In this framework, altimeter-derived snow depth (ASD) serves as a reference to train models that predict snow depth from passive microwave brightness temperatures, yielding promising results (He et al., 2024; Zhou et al., 2025). Yet, as this approach gains traction, key methodological questions remain unaddressed. First, no systematic intercomparison has been conducted to evaluate how different machine learning architectures—from simple linear models to complex deep networks—perform in this specific retrieval task, particularly regarding their accuracy, robustness, and generalization capability across ice types and seasons. Second, the black-box nature of many machine learning models often obscures the underlying physical drivers of the retrieval, limiting interpretability and hindering the

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

extraction of geophysical insight. Achieving low validation error alone is insufficient; understanding why a model succeeds or fails is essential for building scientific trust and advancing retrieval methodologies.

In response to these gaps, this study presents a systematic evaluation of snow depth retrieval over pan-Arctic sea ice using multi-source data and machine learning models for the 2012–2021 period. Our investigation is structured around three core objectives. (1) We conduct a benchmark intercomparison of four fundamentally different machine learning models—Multiple Linear Regression (MLR), Random Forest (RF), LightGBM (LGBM), and Long Short-Term Memory (LSTM)—trained on a fused dataset combining altimeter-derived snow depth with passive microwave brightness temperatures. Our evaluation emphasizes model generalization, overfitting susceptibility, and performance stratification by ice type (first-year ice, FYI, and multi-year ice, MYI). (2) To enhance physical interpretability, we apply permutation feature importance analysis to identify and quantify the passive microwave channels most critical for snow depth retrieval, thereby translating machine learning outputs into geophysically meaningful insights across seasons and ice regimes. (3) We further demonstrate the practical utility of our machine learning-based snow depth products by assessing their impact on SIT retrieval relative to a conventional climatology. Through this comparative and interpretable framework, our work not only delivers a validated, long-term snow depth product for 2012–2021 but also provides a clear methodological reference for selecting, evaluating, and interpreting machine learning models in cryospheric remote sensing applications.

2 Data

All datasets used in this study are described in three categories. Training datasets include passive microwave brightness temperatures from AMSR2 and the active remote sensing snow depth product from Li et al. (2024) used as the target variable. Validation datasets consist of airborne Operation IceBridge (OIB) measurements and in-situ MOSAiC observations. Additional datasets employed for feature inputs, masking, and intercomparison cover SnowModel-LG snow density, OSI SAF sea ice concentration and ice type, the Rostosky et al. (2018) snow depth product, the Modified Warren 1999 (MW99) snow depth climatology, and the CryoSat-2 sea ice thickness product.

2.1 Training data sets

2.1.1 Passive microwave data

Daily, Level-3 gridded brightness temperature data from the Advanced Microwave Scanning Radiometer 2 (AMSR2) instrument were utilized for the period from 1 July 2012 to 30 June 2021. These data were obtained from the AMSR2 Unified L3 Daily 25-km Brightness Temperature product (Version 1), accessible via the National Snow and Ice Data Center (NSIDC) at https://nsidc.org/data/au_si25. The product provides brightness temperature at six frequencies (6.9, 10.7, 18.7, 23.8, 36.5, and 89.0 GHz), each with both vertical (V) and horizontal (H) polarization (Table 1). All data were regridded to a consistent 25 km EASE-Grid 2.0 projection. Lower frequencies

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

provide deeper penetration and sensitivity to thick snow (Rostosky et al., 2018), while higher frequencies are sensitive to surface scattering. The inclusion of the 23.8 and 89.0 GHz channels, despite known atmospheric influence (mainly water vapor), is justified as they contain critical information on near-surface snow properties necessary for discriminating thin-to-moderate snow covers. Potential atmospheric effects were implicitly accounted for by the machine learning models trained on collocated data.

2.1.2 Active remote sensing snow depth

The target variable for model training was the satellite altimeter-derived snow depth product from Li et al., (2024). This product was generated by combining sea ice freeboard from CryoSat-2 and snow freeboard from ICESat-2, following the methods described in the original publication. It provides pan-Arctic coverage over both FYI and MYI domains during the sea ice growth seasons (October to next April) from 2018 to 2021. The data were processed to a 25 km EASE-Grid 2.0 resolution to match the passive microwave data. Independent validation against Operation IceBridge (OIB) measurements showed a root mean square error (RMSE) of 0.06 m (Li et al., 2024). This ASD product served as the primary reference dataset for training our machine learning models.

2.2 Validation datasets

2.2.1 Operation IceBridge

The NASA (National Aeronautics and Space Administration) initiated OIB in 2009, equipping aircraft with laser altimeters, snow radars, and digital cameras to measure snow depth and SIT in the Arctic. Snow depth was estimated by radar echoes of the snow-ice and air-snow interfaces, while snow freeboard measurements were obtained using the laser altimeter. For this study, we used the IceBridge Sea Ice Freeboard, Snow Depth, and Thickness Quick Look data (Version 1) from 2012 to 2019, provided by NSIDC (https://nsidc.org/data/idcsi4). The uncertainty of OIB snow depth is 0.06 m (Kurtz et al., 2012). These data were utilized for comparison with the SIT and snow depth retrievals in this study. The OIB measurements are geographically limited to regions north of the Canadian Arctic Archipelago during March and April (Figure 1a).

2.2.2 MOSAiC

The MOSAiC is a year-around comprehensive international Arctic observation program. The drifting ice camp (RV Polarstern) started from the marginal ice zone in the eastern Amundsen Basin and ended in the Fram Strait (Figure 1b). Snow and ice mass balance buoys were deployed during MOSAiC (Nicolaus et al., 2022). Various snow measurements alongside sea ice properties were measured extensively. The MOSAiC dataset, covering 2019-2021, provides spatially distributed measurements across various sea ice types (https://data.mosaic-expedition.org). The MOSAiC snow depth data was used to validate the retrieval snow depth product.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

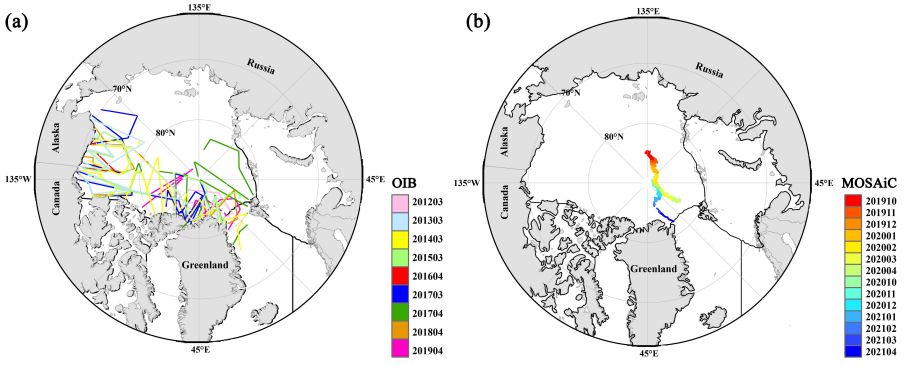


Figure 1: Spatial distributions of (a) OIB measurements from 2012 to 2019 and (b) MOSAiC measurements from 2019 to 2021.

2.3 Additional datasets

2.3.1 SnowModel-LG

Snow density is a critical governing factor for microwave volume scattering and dielectric properties. Therefore, snow density serves as a critical input feature constraining the physical relationship between brightness temperature signals and snow depth during machine learning algorithms. We used daily snow density outputs from SnowModel-LG (Version 1.0) (Petty et al., 2018; Liston et al., 2020), available at https://github.com/snowmodel/snowmodel-lg. This Lagrangian snow-evolution model, forced by ERA5 reanalysis and other datasets, provides physically consistent snow property estimates (including depth and density) across the Arctic at a 25 km resolution. For comparison, the snow depth output from SnowModel-LG (hereinafter SM) was also used in our evaluation.

2.3.2 Sea ice concentration and sea ice type

To define the sea ice domain and distinguish ice types for model training and retrieval, we used two complementary products from the Ocean and Sea Ice Satellite Application Facility (OSI SAF). Daily sea ice concentration (SIC) data were obtained from the OSI SAF Global Sea Ice Concentration Climate Data Record (Version 3.0) (Lavergne et al., 2019), accessible at https://osi-saf.eumetsat.int/products/osi-450. Only grid cells with SIC $\geq$ 80% were included in the analysis to define valid sea ice pixels and minimize contamination from open water effects. Daily ice type classification data were obtained from the OSI SAF Global Sea Ice Type product (OSI-403-b, Version 3.0), which categorizes pixels as FYI or MYI. This product provides the fundamental ice type mask used throughout our study for partitioning the training dataset and models by ice type, and guiding the final retrieval. Both the SIC and sea ice type products have a native spatial resolution of 10 km. To ensure consistency with all other input datasets, they were regridded to a common 25 km EASE-Grid 2.0 projection.

2.3.3 Rostosky snow depth

Rostosky et al. (2018) estimated Arctic snow depths for 2012-2018 using passive microwave satellite

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

observations from AMSR-2 and OIB data and regression analysis algorithms. The datasets (hereinafter referred to as Rost) separate snow depth over MYI and FYI. Over MYI, snow depth data is available only in March and April, as OIB observations are limited to this period. Over FYI, snow depth data covers the snow accumulation season, from November to April. Rost snow depth was used to compare with snow depth retrieved by this study. We obtained the data from 2012 to 2018 which has a spatial resolution of 25 km. The algorithms for deriving snow depth are detailed in Rostosky et al. (2018), see product manual for data processing, quality control, and limitations.

2.3.4 MW99 snow depth

The Warren snow depth climatology (W99) has been widely used to calculate SIT (Laxon et al., 2013). However, validation using in situ measurements revealed that the W99 snow depth significantly overestimates snow accumulation over FYI. To address this issue, Kurtz and Farrell (2011) proposed an improved version of the W99 climatology, reducing snow depth over FYI by 50% to better reflect the changing conditions in the Arctic Ocean. This modified version, referred to as MW99 (Modified Warren 1999), has been adopted in recent studies for its improved accuracy. In this study, the MW99 snow depth data, spanning from 2012 to 2021, was utilized for comparison with snow depth retrievals derived from our methods. MW99 is monthly data that has been processed to a 25 km spatial resolution.

2.3.5 Sea ice thickness products

For SIT estimation, we used Cryosat-2 sea ice freeboard, covering the period from 2012 to 2021, obtained from NSIDC (https://nsidc.org/data/rdeft4/versions/1). This dataset includes monthly SIT and sea ice freeboard measurements derived from CryoSat-2 satellite altimetry, which have been rigorously validated using field observations and airborne measurements (Xia and Xie, 2018; Li et al., 2020). Sea ice freeboard combined with the retrieval snow depths was utilized to calculate SIT while the NSIDC SIT was used for comparison with the SIT estimates derived from our snow depth retrievals. The Cryosat-2 data has a spatial resolution of 25 km.

3 Methods

3.1 Theoretical basis of snow depth retrieval

The theoretical basis for retrieving snow depth from passive microwave data resides in the interaction of microwave radiation with snow properties. Volume scattering becomes significant when the snow grain size is approximately one-tenth of the wavelength, and its intensity increases with frequency. As scattering intensifies, the brightness temperature of snow over the sea ice surface decreases. Consequently, higher frequencies result in lower snow brightness temperatures. Deeper snow enhances the scattering effect, leading to a further reduction in brightness temperature as snow depth increases. Additionally, snow volume scattering modifies the emission from the underlying ice layer, further altering the brightness temperature. In summary, snow brightness temperature decreases with both

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

increasing snow depth and frequency. Snow depth exhibits the strongest correlation with GR derived from brightness temperatures at 37 GHz and 19 GHz (Markus and Cavalieri, 1998).

$$\mathrm{GR} = \frac{T_{bice}(f_1,p) - T_{bice}(f_2,p)}{T_{bice}(f_1,p) + T_{bice}(f_2,p)}, \quad (1)$$

$$\mathrm{PR} = \frac{T_{bice}(f,V) - T_{bice}(f,H)}{T_{bice}(f,V) + T_{bice}(f,H)}, \quad (2)$$

Here, T_{bice} is the sea ice brightness temperature; f represents frequencies; H and V represent horizonal and vertical polarization, respectively.

$$T_{bice}(f,p) = \frac{T_b(f,p) - (1-SIC)\times T_{bow}(f,p)}{SIC}, \quad (3)$$

Here, $T_b(f,p)$ represents the total brightness temperature. T_{bow} is the brightness temperature of open water. SIC is the sea ice concentration.

We applied all available vertical polarization and horizontal polarization frequency combinations (Table 1) in this study. To explore the relationship between passive microwave brightness temperature and ASD, we employed four machine learning methods to establish relationships between GRs, PRs, snow density, and snow depth over FYI and MYI.

Table 1: Spectral and polarization gradient ratios (GR, PR) for various frequency combinations of vertical and horizontal polarization modes

Frequency (GHz)	6	10	18	23	36	89
6	PR(6)	GR(10/6V)	GR(18/6V)	GR(23/6V)	GR(36/6V)	GR_V(89/6V)
10	GR(10/6H)	PR(10)	GR(18/10V)	GR(23/10V)	GR(36/10V)	GR (89/10V)
18	GR(18/6H)	GR(18/10H)	PR(18)	GR(23/18V)	GR(36/18V)	GR(89/18V)
23	GR(23/6H)	GR(23/10H)	GR(23/18H)	PR(23)	GR(36/23V)	aGR(89/23V)
36	GR(36/6H)	GR(36/10H)	GR(36/18H)	GR(36/23H)	PR (36)	GR (89/36V)
89	GR(89/6H)	GR(89/10H)	GR(89/18H)	GR(89/23H)	GR(89/36H)	PR(89)

Note: The frequencies of passive microwave data were rounded to integer values. PR(6) represents PR(6V, 6H). GR(10/6V) and GR (10/6H) represent vertical and horizonal polarization for gradients ratios GR(10V/6V) and GR(10H/6H), respectively.

3.2 Machine learning methods

Four models were selected to cover a spectrum of complexity and interpretability. MLR serves as a baseline statistical model. Although linear regression is often considered a classical statistical method rather than a machine learning algorithm, it is included here to provide a lower-bound reference for performance. The remaining three models represent distinct nonlinear machine learning approaches. RF and LGBM are ensemble tree-based methods known for handling feature interactions and outliers robustly, while LSTM is a recurrent neural network designed to exploit temporal dependencies in the input time series. These models were chosen because they are widely used in geophysical remote sensing, have different inductive biases, and allow a meaningful comparison between simplicity (MLR), ensemble averaging (RF, LGBM), and sequential memory (LSTM). The goal is not to claim that any

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

single model is universally superior, but to understand their relative strengths and weaknesses for snow depth retrieval under different ice regimes and validation scenarios.

3.2.1 Data preprocessing

To construct the machine learning training dataset, the passive microwave brightness temperatures and the ASD product required temporal alignment due to their different temporal and spatial resolutions. The spatial collocation was straightforward as both final datasets share the same projection and grid. Both the AMSR2 brightness temperature data (Section 2.1.1) and the ASD product (Section 2.1.2) were ultimately provided on a common 25 km EASE-Grid 2.0 projection. No additional spatial regridding or resampling between these two core datasets was necessary, ensuring spatial co-location at the grid cell level. The ASD product is a monthly composite (e.g., representing conditions for October 2018), whereas the AMSR2 provides daily observations. To align them temporally, we processed the daily brightness temperature data into monthly composites that correspond exactly to the temporal period of each ASD value. For each calendar month and each 25 km grid cell, all available daily brightness temperature observations (from all frequencies and polarizations) within that month were averaged to create a single monthly mean brightness temperature value. This monthly compositing of brightness temperature effectively filters out sub-monthly variability and noise, aligning the temporal scale of the input feature (brightness temperature) with that of the target variable (monthly ASD snow depth).

Each valid sample in the final training dataset thus represents a single 25 km grid cell for a specific month. It contains input features, including the monthly averaged brightness temperature values (and their derived gradient ratios, PR/GR), along with the collocated monthly snow density from SnowModel-LG and ice type from OSI SAF. Target variable is the monthly ASD snow depth value for the same grid cell and month. Only grid cells where all required data (monthly brightness temperature, ASD, snow density, ice concentration $\geq$80%, and ice type) were available for a given month were retained. This rigorous collocation and compositing process ensures a spatially and temporally consistent dataset for model training and retrieval.

3.2.2 Multiple linear regression

Multiple linear regression was implemented as an interpretable baseline model to establish linear relationships between Arctic snow depth and predictor variables (snow density, selected GRs and PRs). The input features consist of selected GRs and PRs. Rather than fixing a universal feature set, a recursive feature elimination procedure was applied to identify the optimal subset of GRs and PRs for each month and ice type (FYI vs. MYI), based on the permutation importance rankings. We prioritized this classical algorithm for its computational efficiency and parametric transparency, which enables explicit coefficient analysis-a critical advantage when establishing preliminary physical relationships. Unlike the subsequent nonlinear ensemble methods (LGBM, RF) and temporal models (LSTM), MLR assumes strictly addictive linear interactions, providing a performance benchmark for complex model comparisons. As MLR requires no hyperparameter tuning, we employed ordinary least squares

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

estimation to minimize residual errors. The regression model can be represented as follows.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon \ , \quad (4)$$

Here, $X_1, X_2, \ldots, X_n$ denote the independent variable, Y represents the dependent variable, β_1, β_2, …, β_n are the coefficients of each variable, β_0 represents the intercept, and ε is the error term. In the linear regression, the loss function is defined as:

$$L = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} (y_i - X_i\beta_i)^2 \ , \quad (5)$$

Where, y_i is the actual value, and $\hat{y}_i$ is the predicted value. This loss function estimates the discrepancy between the actual values and the model's predictions. The objective is to minimize this discrepancy.

3.2.3 Random forest

The Random Forest was selected for its proven capability in capturing complex nonlinear relationships within our snow depth dataset, while maintaining robustness against overfitting and noisy measurements - a distinct advantage over simpler linear models. RF builds upon decision trees and incorporates bootstrap sampling (with replacement) and ensemble learning concepts from the Bagging algorithm. During model construction, the algorithm creates k decision trees by training each on a random subset of the original training data (both samples and features).

After training the regression decision trees, each leaf node may have several value predictions. The predicted value for each leaf node is determined by calculating the average of all these values. Following the classification of feature attributes in the training dataset for each decision tree, the regression output of the entire tree is obtained. For the overall RF regression prediction model, the end result is the average of the outputs from all decision trees. The specific formula is as follows:

$$f(x) = \frac{1}{T}\sum_{i=1}^{T} f_i(x) \ , \quad (6)$$

Where, x represents the input variable, $f(x)$ represents the model output, $f_i(x)$ denotes the base learner, and T is the number of base learners.

The RF model includes several tunable hyperparameters, such as the depth of each tree (TD), the number of trees (NT), and the number of features considered for splitting a node (NF). Here, NF was configured to one-third of the total predictive variables in the regression task. To determine the best values for NT and TD, a grid search method with 5-fold cross-validation was employed. NT was varied from 50 to 250 in increments of 50, while TD was adjusted on a logarithmic scale between 2 and 16. Ultimately, NT was selected to be 200, and TD was set to 5.

3.2.4 LightGBM

The LightGBM algorithm was implemented to leverage its superior computational efficiency and predictive accuracy with large feature sets - particularly advantageous given our high-dimensional

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

input data (snow density, GRs, PRs). Unlike RF, LGBM employs a sequential tree-building approach and utilizes histogram-based optimization for faster training while maintaining precision. Similar to XGBoost, LGBM leverages the Taylor expansion of the loss function to approximate residuals and incorporates regularization to control model complexity and prevent overfitting.

The basic process of the LGBM algorithm consists of several steps. First, the continuous values in the dataset are counted, and a histogram with a specified width (k) is constructed. Second, based on the statistical intervals of the histogram, the continuous values are discretized into k discrete values. These discrete values are then used as indexes to traverse the histogram and accumulate statistics. Finally, the optimal split point is determined from the accumulated statistics to create splits in the decision tree.

A systematic hyperparameter optimization framework was employed, integrating 5-fold cross-validation with grid search algorithms, to augment the predictive robustness of the LGBM architecture within our computational paradigm. The primary hyperparameters we focused on were boosting type, NT, and the learning rate. Specifically, we varied NT from 500 to 2500 in increments of 100. For the boosting type, we evaluated three different methods: the traditional Gradient Boosting Decision Tree, the Dropout-integrated method known as Dart, and the Random Forest-style boosting. Furthermore, the learning rate was adjusted systematically, starting from 0.0001 and increasing by orders of magnitude up to 0.1, to assess its effect on model performance. Ultimately, we identified the optimal combination of hyperparameters: NT was set to 2000, the Gradient Boosting Decision Tree was chosen as the boosting type, and the learning rate was established at 0.01. This configuration exhibited the best balance between model complexity and predictive accuracy.

3.2.5 Long Short-Term Memory

Long Short-Term Memory networks were specifically deployed to capture temporal dependencies within our chronologically ordered snow depth dataset - a critical capability absent in non-sequential models like RF and LGBM. This architecture uniquely preserves long-range temporal patterns through its memory cell design, effectively mitigating the vanishing gradient problem inherent in standard recurrent neural networks. This makes LSTM particularly well-suited for modeling the complex evolution of Arctic snow across seasons.

The LSTM's core innovation lies in its gated mechanism, comprising forget, input, and output gates. These gates dynamically regulate information flow using sigmoid activations ($\sigma \in [0,1]$) for feature-wise gating and hyperbolic tangent for nonlinear state transitions, enabling selective preservation and integration of sequential information over extended time horizons.

To optimize the LSTM model for our study, we focused on four specific variables: the numbers of neurons per layer, the depth of LSTM layers, the maximum training epochs, and the choice of optimizer. We began with 16 neurons and increased this number exponentially by a factor of 2, up to a maximum of 128 neurons. The number of LSTM layers was increased incrementally from one to three. The optimizers evaluated included Stochastic Gradient Descent, Adam, and RMSprop. Ultimately, the optimal configuration for the LSTM model was determined to be 64 neurons, a single LSTM layer, and

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

the Adam optimizer. For LSTM, the loss function is same as that of MLR (Equation 5). The input data have been arranged in chronological order, with no missing data, to ensure the continuity of the time series.

3.3 Sea ice thickness estimation

SIT is essential for understanding Arctic mass balance, oceanatmosphere interactions, and climate change. Satellite altimetry (e.g., CryoSat-2) provides sea ice freeboard, but converting freeboard to thickness requires knowledge of snow depth and densities. The hydrostatic equilibrium equation is the standard physical framework for this conversion. Snow depth directly affects the calculated SIT because snow loads the ice floe and its freeboard is partly compensated by snow. An overestimation of snow depth leads to an overestimation of SIT (and vice versa). Therefore, improving snow depth estimates is expected to reduce systematic biases in SIT products, which currently often rely on outdated climatologies such as MW99.

In this study, we apply the snow depths retrieved by our four machine learning models (LGBM, MLR, RF, LSTM) to derive SIT from CryoSat2 freeboard data. SIT is calculated by sea ice freeboard, snow depth, snow density, sea ice and seawater density based on hydrostatic equilibrium (Equation 7).

$$T = \frac{\rho_w \times h_f + \rho_s h_s}{\rho_w - \rho_i}\ , \tag{7}$$

T is SIT. ρ_w is the seawater density which is a constant (1025 kg/m3). h_f and h_s are sea ice freeboard and snow depth, respectively. ρ_s and ρ_i are snow density and sea ice density, respectively. h_f is from NSIDC Cryosat-2 dataset and the snow density is from the Warren climatology (Warren et al., 1999). The sea ice density is different for different sea ice types. For MYI, we set to 882 kg/m3 for MYI and 917 kg/m3 for FYI (Alexandrov et al., 2010). The same density values are used for all SIT estimates to ensure that any performance difference is solely attributable to the snow depth input. For comparison, we also include the operational NSIDC SIT product (FT4), which uses the MW99 snow climatology.

4 Results

Results of snow depth retrieval over Pan-Arctic sea ice using the machine learning models described in Section 3 are presented. Optimisation of input parameters is first examined through permutation importance analysis, revealing the seasonal evolution of key features for FYI and MYI separately. Optimal feature combinations are then selected using recursive feature elimination, and the performance of four algorithms (LGBM, MLR, RF, LSTM) is compared on the testing subset. Independent validation against airborne OIB data and in-situ MOSAiC measurements follows, benchmarking the retrievals against three existing snow depth products (Rostosky et al., 2018; SnowModel-LG; MW99). This validation reveals critical insights into model generalisability and the trade-off between internal testing metrics and true independent performance. Finally, a robustness assessment is performed, evaluating temporal extrapolation to the 2012-2017 period and scale transfer from monthly-trained models to daily retrievals.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

4.1 Optimizing input parameters for machine learning algorithms

The permutation importance method was employed to identify the most influential input parameters for snow depth retrieval across different months and ice types. This analysis not only ranks feature importance but also provides a foundation for interpreting the underlying physical processes governing the relationship between microwave observations and snow depth.

Over FYI, the feature importance analysis reveals a clear seasonal narrative governed by snowpack evolution. Snow density is the unvarying cornerstone, ranking first every month. Its supreme importance is physically fundamental, as it directly controls the snowpack's dielectric constant and volume scattering efficiency, thereby modulating the overall microwave emission and penetration depth. The low-frequency gradient ratios GR(10V/06V) and GR(10H/06H) are persistently among the top five features, underscoring the critical year-round role of deeper-penetrating 6 and 10 GHz signals for sensing total snow accumulation down to the snow-ice interface. A defining seasonal signal is the dramatic rise in importance of the 89 GHz polarization ratio (PR(89)). In the early season (October-November), it ranks moderately (8th and 6th, respectively). Its importance undergoes a marked transition, surging to become the 2nd most important feature from December through April. This shift pinpoints the timeframe when the FYI snowpack develops sufficient thickness and metamorphic complexity (e.g., formation of larger depth hoar crystals) to become an effective volume scatterer at high frequencies. Consequently, PR(89) transforms from a moderate contributor to a primary indicator of snow presence and microstructure during the core winter period. The rankings also highlight other dynamically important features. The gradient ratio GR(36/23V) consistently appears in the top ten, especially prominent in late winter (March-April), likely related to scattering from intermediate snow layers. The importance of GR(18/10V) is notable in the early to mid-season (ranking 5th in October) but diminishes by spring, possibly reflecting changes in the relative sensitivity to different snow layers as the pack develops.

In summary, the retrieval on FYI is supported by the constant physical constraint of snow density and the deep-sensing capability of low-frequency gradients. The clearest seasonal dynamic is the mid-winter emergence of PR(89) as a leading feature, directly tracking the development of a scattering snow microstructure. This detailed understanding allows for a seasonally-aware retrieval approach, optimizing model performance by emphasizing the most physically relevant parameters at each stage of snowpack evolution.

Over MYI, the permutation feature importance reveals a distinct and stable hierarchy, strongly influenced by the unique properties of aged snow. Snow density is the predominant feature, ranking first in every month except March. Its consistent supremacy underscores its role as the fundamental physical constraint governing the snowpack's dielectric and scattering properties.

A hallmark of MYI is the exceptional and early prominence of the 89 GHz polarization ratio (PR(89)). It ranks as the second most important feature from October through February and in April, and notably ascends to first place in March. This demonstrates that the high-frequency, polarization-sensitive

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

scattering signal is a dominant and nearly constant source of information for MYI snow retrieval throughout the entire growth season. This is a direct consequence of the mature, metamorphosed snow on MYI, which contains well-developed depth hoar and larger ice crystals. This microstructure creates an efficient volume-scattering medium, making PR(89) a robust indicator of snow presence and its internal structural state. The importance of lower-frequency polarization ratios (e.g., PR(06), PR(36)) exhibits a clear increasing trend from early to late winter. While present in the top features from October, their rankings systematically improve. For example, PR(06) moves from the edge of the top ten (11th in October) to a stable position within the top five from January onward. This progression signifies the growing dominance of the snowpack itself as the primary scattering layer for low-frequency emissions. In early winter, a shallower snow cover may allow the underlying rough MYI surface to influence the signal. As the snowpack deepens and its microstructure becomes more complex by mid-to-late winter, it becomes the main controller of volume scattering even at lower frequencies like 6 GHz. A striking pattern emerges in March, where the top ranks are dominated by polarization ratios across multiple frequencies (PR(89), PR(06), PR(10), PR(36), PR(23), PR(18)), with GR(10/06V) as a key exception. This polarization ratio dominance likely signifies a period of peak snowpack metamorphism and optimal scattering conditions across a wide spectrum of microwave frequencies. The persistent presence of GR(10/06V) among the top features across all months confirms the complementary, vital role of deep-penetrating signals for sensing total snow depth.

In summary, MYI snow retrieval is characterized by the unparalleled, year-round importance of high-frequency volume scattering (PR(89)), grounded by the physical baseline of snow density. The systematic rise in low-frequency PR importance and the late-winter shift to a PR-dominated regime provide a clear, data-driven signature of the deepening and maturing of the MYI snowpack, offering profound insight into its seasonal electromagnetic evolution.

After determining the monthly importance of input parameters, optimal feature combinations for snow depth retrieval were identified using the recursive feature elimination method. Using this method, the best input parameters for snow depth over FYI and MYI were derived. Once the best input parameters were identified for machine learning algorithms, we assessed them with commonly used metrics presented in Table 2. The matching data sample points between ASD snow depth and machine learning input parameters were 99,347 over FYI and 61,540 over MYI, respectively. The dataset samples were randomly split into testing and training subsets at a ratio of 1:4.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

Table 2: Statistical comparison of machine learning-retrieved snow depth based on the testing subset data. Metrics include: R, RMSE, Bias, and MAE.

	R	RMSE (cm)	Bias (cm)	MAE (cm)
FYI(n=99347)				
LGBM	0.83	3.96	-0.07	2.87
MLR	0.71	4.95	-0.07	3.74
RF	0.84	3.87	-0.04	2.73
LSTM	0.81	4.12	0.01	2.99
MYI(n= 61540)				
LGBM	0.85	4.51	-0.02	3.23
MLR	0.70	6.01	-0.05	4.54
RF	0.84	4.52	0.01	3.16
LSTM	0.86	4.45	0.04	3.13

Note: R is the Pearson correlation coefficient; RMSE is the root mean square error; MAE is the mean absolute error; Bias is the mean bias.

Over FYI, RF slightly outperforms LSTM and LGBM among the three models (RF, LGBM and LSTM). This is primarily because RF has advantages in managing nonlinear relationships and feature interactions. Although LGBM performs slightly worse than RF in FYI, its results are still acceptable, indicating that it can still provide reasonable estimations of snow depth. MLR has the lowest correlation coefficient (R = 0.71) and the highest RMSE (4.95 cm). This can be attributed to its simple model structure, which is unable to capture the complex nonlinear relationships between input parameters and snow depth. Over MYI, LSTM showed a strong correlation with the actual snow depth followed by LGBM, RF and MLR; it has a low RMSE, but a slightly larger bias, which may be attributed to its sensitivity to extreme values or outliers during training. Among all algorithms, RF exhibits the best overall statistical performance, reflecting its advantages in prediction accuracy, particularly in deviation control.

4.2 Validation against in situ observations

To rigorously assess the accuracy of our retrievals, we performed independent validation using airborne observations from OIB and in-situ measurements from the MOSAiC expedition. Furthermore, we contextualized our results by benchmarking them against three established snow depth products from the passive microwave-based retrieval by Rostosky et al. (2018), the SnowModel-LG output, and the Modified Warren climatology.

4.2.1 Validation with OIB airborne measurements

OIB provides direct, high-quality snow depth measurements derived from snow radar, serving as a crucial independent reference. For comparison, OIB data were regridded to a 25 km resolution (Li et al., 2024). To ensure a fair and consistent evaluation across all seven snow depth datasets (four from our algorithms and three benchmarks), we used only collocated grid cells where all data were

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

simultaneously available, resulting in 3,167 validation points. The statistical results are summarized in Table 3.

The MLR algorithm demonstrated the most balanced performance, achieving the highest correlation (R = 0.67), the lowest RMSE (7.19 cm), and the lowest MAE (6.02 cm), with a bias of 2.53 cm. The Rost product, whose regression model was trained on OIB data, showed comparable accuracy (RMSE = 7.16 cm, R = 0.66). This is expected because the Rost snow depth was trained directly using OIB data as the target variable. The tree-based models, LGBM and RF, also performed robustly (R = 0.65 and 0.62, respectively). A notable trade-off was observed for LSTM network. It achieved the smallest mean bias (1.98 cm) but the lowest correlation (R = 0.47). The model-based products, SM and especially MW99, showed larger errors, with MW99 exhibiting the highest bias (7.04 cm) and MAE (8.55 cm).

Performance varied significantly between ice types, underscoring the importance of analysis for different sea ice types. Over FYI, MLR again showed strong performance with the lowest RMSE (5.61 cm) and highest R (0.55). LSTM had the weakest correlation (R = 0.32). MW99 was the least accurate product for FYI, showing the highest bias and MAE. Over MYI, The Rost product performed best on MYI (lowest RMSE=7.73 cm, highest R=0.46), closely followed by MLR. LSTM maintained a very low bias (1.59 cm) but the lowest correlation (R = 0.23). SM had the highest RMSE (12.21 cm) and MAE (10.34 cm) on MYI.

The probability density distributions (Figure 2) provide further insight into error structures. LSTM's distribution aligns most closely with OIB, consistent with its minimal bias. LGBM, MLR, and RF show good agreement but with a slight overestimation in the 0-0.2 m range. The Rost product underestimates for depths less than 0.15 m and overestimates for deeper snow. The SM distribution is broad, while MW99 shows the poorest agreement, significantly overestimating snow depth, particularly for values greater than 0.3 m.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

Table 3: Accuracy evaluations of monthly snow depth retrieved by LGBM, MLR, RF, LSTM, SM, Rost and MW99 using OIB airborne snow depth data (RMSE/bias/MAE units: cm).

	ALL	LGBM	MLR	RF	LSTM	SM	Rost	MW99
ALL	N	3167	3167	3167	3167	3167	3167	3167
	RMSE	7.80	7.19	7.50	9.59	11.36	7.16	7.59
	bias	2.84	2.53	2.91	1.98	4.85	2.66	7.04
	MAE	6.49	6.02	6.32	7.57	9.34	6.23	8.55
	R	0.62	0.67	0.65	0.47	0.53	0.66	0.63
FYI	N	1093	1093	1093	1093	1093	1093	1093
	RMSE	5.98	5.61	6.02	7.22	8.81	5.80	6.70
	bias	2.74	2.69	2.95	2.92	4.07	3.79	5.86
	MAE	5.07	4.79	5.20	5.76	7.03	5.77	7.37
	R	0.51	0.55	0.48	0.32	0.48	0.56	0.41
MYI	N	2074	2074	2074	2074	2074	2074	2074
	RMSE	8.56	7.85	8.14	10.49	12.21	7.73	7.79
	bias	2.90	2.48	2.90	1.59	5.08	2.15	7.67
	MAE	7.21	6.62	6.89	8.43	10.34	6.48	9.15
	R	0.35	0.45	0.42	0.23	0.35	0.46	0.37

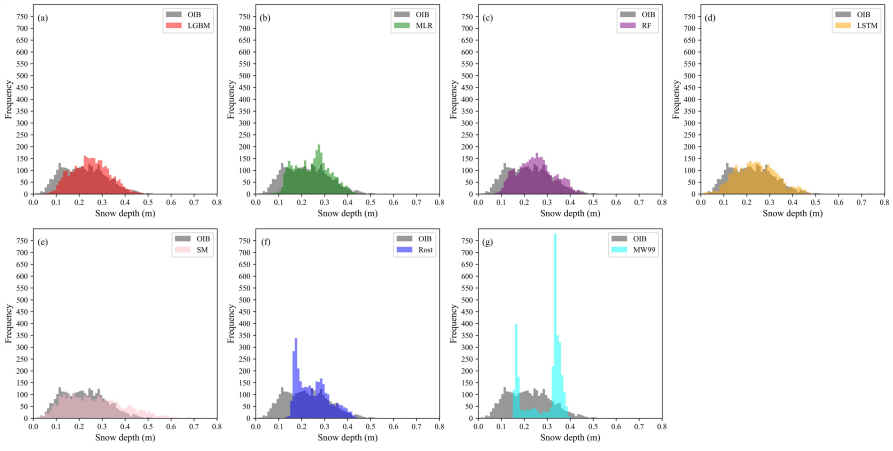


Figure 2: OIB snow depth distributions are shown in grey against seven snow depths estimated by various algorithms in distinct colors. (a) LGBM, (b) MLR, (c) RF, (d) LSTM, (e) SM, (f) Rost, (g) MW99.

4.2.2 Validation with MOSAiC in-Situ measurements

We further validated the snow depth products using independent, high-frequency point measurements from MOSAiC buoys (2019-2021), which were not used in any model training. Due to limited temporal overlap, the Rost product (2012-2018) was excluded, and 192 collocated grid points were available for comparison with the remaining six products (Table 4, Figure 3).

The results reveal a different performance landscape compared to the OIB validation. The RMSE values across all algorithms and SM were similar (5.97-7.00 cm). Notably, LSTM achieved the lowest

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

bias (0.30 cm) and MAE (5.57 cm), followed closely by LGBM and RF, which showed slight negative biases. In contrast, MLR, SM, and MW99 exhibited positive biases, with MW99's being exceptionally large (12.29 cm). Correlation coefficients with MOSAiC data were low for all products (R = 0.13-0.20), reflecting the fundamental scale mismatch between point measurements (buoys) and areal averages (satellite products). Despite the low correlations, the biases and MAEs indicate that LSTM, LGBM, and RF most accurately capture the mean snow depth state at the MOSAiC locations.

Table 4: Accuracy evaluations of snow depth retrieved by LGBM, MLR, RF, LSTM, SM and MW99 using MOSAiC snow depth (RMSE/bias/MAE units: cm).

	LGBM	MLR	RF	LSTM	SM	MW99
N	192	192	192	192	192	192
RMSE	7.00	6.71	6.94	6.83	5.97	7.64
bias	-0.33	2.55	-0.31	0.30	3.52	12.29
MAE	5.74	5.69	5.60	5.57	5.51	12.70
R	0.19	0.17	0.19	0.13(0.08)	0.20	0.18

Note: The value in parentheses represent its p-value; except for LSTM, other snow depths show significant correlations with MOSAiC snow depth.

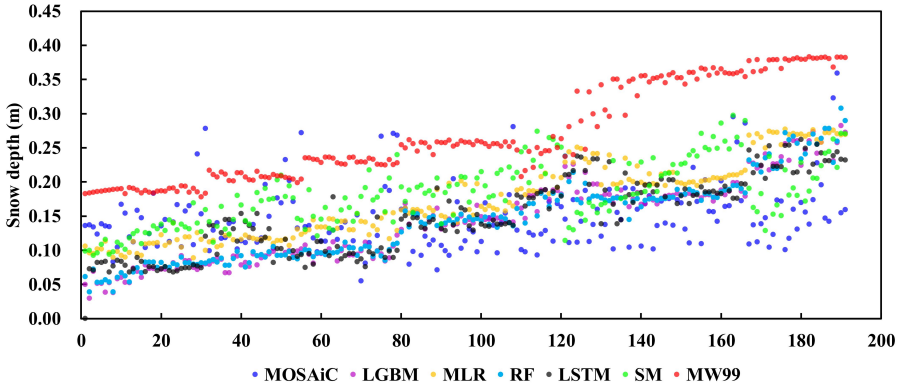


Figure 3: Comparison of snow depth between MOSAiC observations and estimates from LGBM, MLR, RF, LSTM, SM and MW99 from October 2019 to April 2021. There are 192 valid grids, with the x-axis representing the number of valid grid points.

4.2.3 Synthesis of validation results and implications for model selection

The comprehensive validation against the independent OIB and MOSAiC datasets reveals critical insights that directly address the concerns regarding model performance and selection. A central finding is the apparent discrepancy between model performance on the internal testing subset (Section 3.1, Table 2) and their performance on truly independent validation data (Tables 3 and 4). Within the testing set (derived from the same ASD data used for training), the complex models (RF, LGBM, LSTM) consistently outperformed the simple MLR, achieving higher correlations (R) and lower errors. However, this ranking inverted or shifted significantly when validated against OIB and MOSAiC.

This divergence is a classic indicator of overfitting. The complex models, with their high capacity for non-linear fitting, likely learned not only the genuine geophysical relationship between microwave

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

signals and snow depth but also the specific noise, biases, and spatial sampling characteristics inherent to the ASD training product. Consequently, they excelled on data from the same distribution (the testing subset) but showed reduced generalizability to data from different sources (OIB, MOSAiC) with distinct error structures and observational scales. In contrast, the linear constraints of the MLR model acted as an implicit regularizer, preventing it from fitting spurious patterns. This resulted in its poorer performance on the testing set but conferred superior robustness and generalizability, leading to its high correlation and lowest RMSE against the OIB areal averages.

The validation also highlights a fundamental trade-off between accuracy metrics and the influence of validation data type. The LSTM model demonstrated a unique strength in minimizing mean bias against both OIB and, most notably, the MOSAiC point measurements, where it achieved near-zero bias. This suggests LSTM is particularly effective at learning and correcting for systematic offsets to capture the accurate local mean state. However, this came at the cost of lower correlation coefficients, especially with OIB, indicating a weaker ability to replicate precise spatial variability at the satellite grid scale. This pattern (low bias but low R) further supports the hypothesis of overfitting to specific patterns in the training data that do not translate to spatial variability in independent sets.

Furthermore, the scale mismatch is a primary reason for the generally lower correlations with MOSAiC compared to OIB. Validating a 25-km grid-cell average against a point measurement inherently introduces representativeness error, which suppresses correlation coefficients regardless of the algorithm's intrinsic skill. This scale issue affects all products equally, as seen in their uniformly low R values against MOSAiC.

4.3 Robustness assessment: temporal extrapolation and scale transfer

To rigorously evaluate the generalizability of our models beyond their training constraints, we conducted a comprehensive assessment focusing on two dimensions: temporal extrapolation (applying models to a period not included in training) and scale transfer (generating daily retrievals from monthly-trained models). This analysis addresses key questions about model robustness and the feasibility of producing a continuous, high-resolution snow depth record.

4.3.1 Validation on temporally independent daily data

We applied the models, trained on 2018-2021 monthly data, to retrieve daily snow depth for the independent period of 2012-2017. Validation against coincident daily OIB measurements (N=5,133) reveals performance trends (Table 5, Figure 4) that strongly corroborate findings from the monthly validation (Section 4.2.1). MLR again demonstrates superior balance, achieving the highest correlation coefficients overall (R=0.65) and on FYI (R=0.53), along with the lowest MAE (6.27 cm) and a low bias (1.83 cm). LSTM maintains its distinct strength in minimizing mean bias, achieving the lowest overall bias (1.03 cm) and an exceptionally low bias on MYI (0.42 cm), albeit with the lowest overall correlation (R=0.44). The ensemble models (RF, LGBM) show intermediate performance. The distribution analysis (Figure 4a-d) confirms these patterns. LSTM's distribution aligns most closely with OIB, while others show a slight positive shift. This successful application to a temporally

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

independent era confirms that the learned relationships are not period-specific.

Table 5: Accuracy evaluations of daily snow depth retrieved by LGBM, MLR, RF and LSTM using OIB airborne snow depth. (RMSE/bias/MAE units: cm)

		LGBM	MLR	RF	LSTM
ALL	N	5133	5133	5133	5133
	RMSE	7.57	7.79	7.42	10.51
	Bias	3.20	1.83	3.44	1.03
	MAE	6.61	6.27	6.59	8.06
	R	0.64	0.65	0.66	0.44
FYI	N	1627	1627	1627	1627
	RMSE	6.40	6.29	6.41	8.01
	Bias	4.28	1.96	4.32	2.45
	MAE	6.36	4.99	6.40	6.39
	R	0.45	0.53	0.45	0.43
MYI	N	3506	3506	3506	3506
	RMSE	7.98	8.38	7.80	11.41
	Bias	2.71	1.78	3.04	0.42
	MAE	6.72	6.84	6.68	7.24
	R	0.41	0.43	0.43	0.20

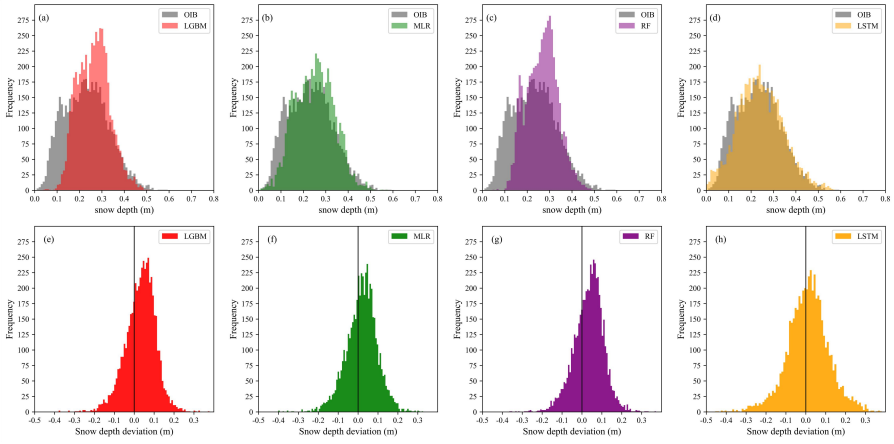


Figure 4: OIB snow depth distributions are shown in grey against daily snow depth estimated by various algorithms in distinct colors (a-d). The deviations between estimated snow depth from various algorithms and OIB snow depth are displayed in panels (e-h). A black line represents the point of flawless agreement with OIB standards (e-h).

4.3.2 Spatial and seasonal plausibility of the extended product

The application of the models enables the generation of a continuous pan-Arctic snow depth record from 2012 onward. The spatial distributions for a sample season (2012-2013) are shown in Figure 5 (monthly means) and Figure 6 (daily). The products exhibit physically plausible and coherent patterns. Snow depth increases progressively from autumn to spring, and MYI regions consistently exhibit

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

greater snow accumulation than FYI regions. For instance, using the MLR product, the pan-Arctic average snow depth increased from approximately 7.9 cm (FYI) and 17.7 cm (MYI) in October to about 16.7 cm (FYI) and 26.4 cm (MYI) by April. The spatial patterns are largely consistent across algorithms, with LGBM and RF showing the highest similarity, while LSTM exhibits minor seasonal differences.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

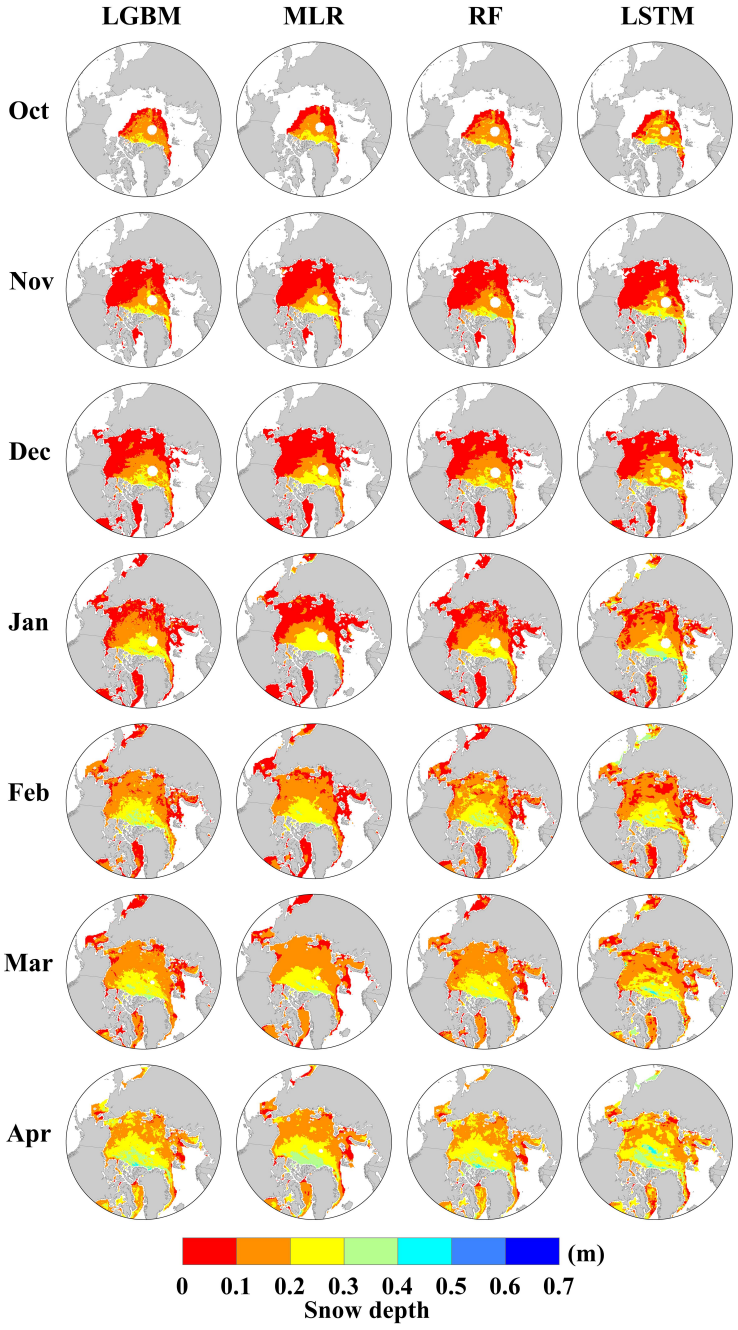


Figure 5: Spatial distributions of monthly Artic snow depth in the growth period of 2012-2013 based on LGBM, MLR, RF and LSTM.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

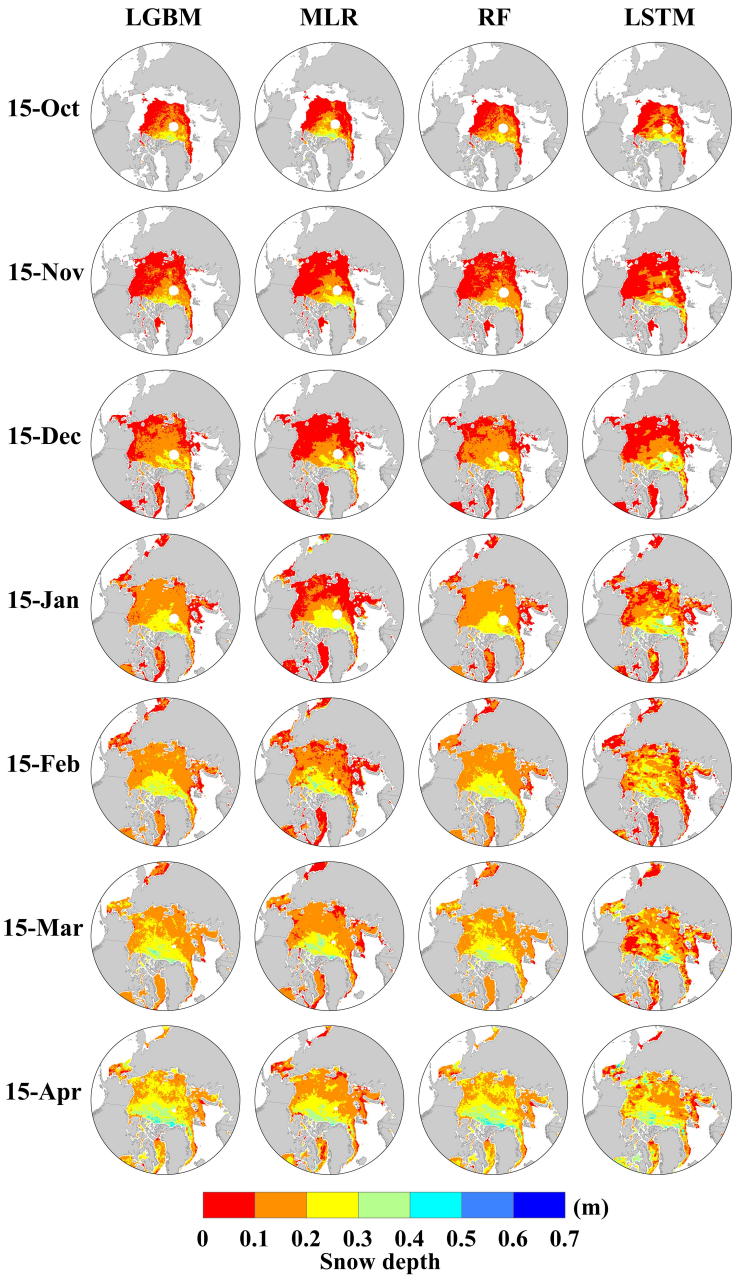


Figure 6: Spatial distributions of daily snow depth over Arctic sea ice in the growth period of 2012-2013 based on LGBM, MLR, RF and LSTM.

4.3.3 Internal consistency: daily-to-monthly scale transfer

A critical test of scale transfer is the internal consistency between the directly retrieved monthly snow depth (the model's native training target) and the monthly average computed from the derived daily

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

retrievals. The differences between these two estimates are quantified in Figure 7. MLR exhibits the smallest and most tightly clustered differences (mean and median nearest to zero), indicating remarkable stability and that its daily predictions aggregate robustly to the monthly scale. LSTM also shows high consistency with low variability. In contrast, LGBM and RF show larger, positively skewed differences, suggesting a systematic effect when downscaling to daily resolution. This analysis confirms that MLR (and to a similar extent, LSTM) effectively captures the underlying physical signal invariant to temporal aggregation, validating the feasibility of generating daily-scale products from monthly-trained models.

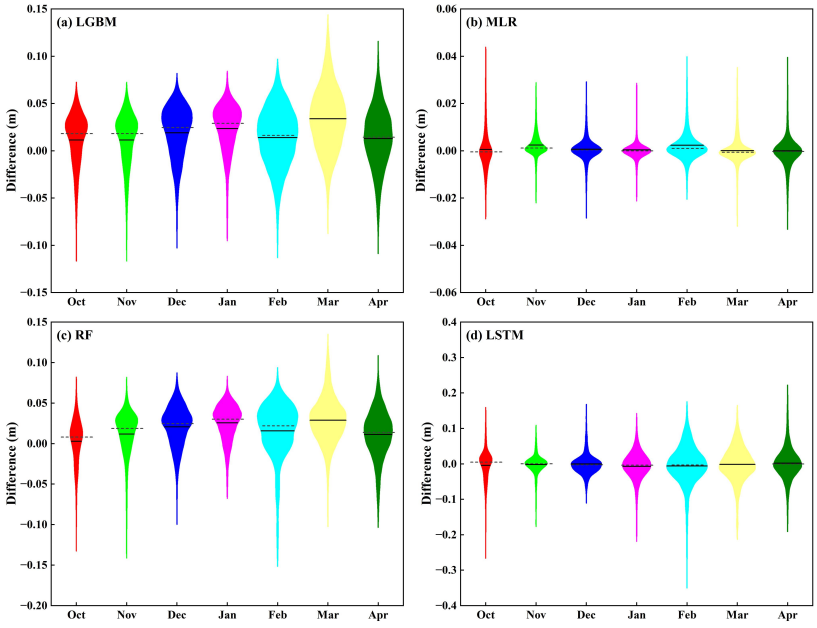


Figure 7: Violin plots showing the statistical distributions of the differences between estimated monthly snow depth by machine learning and calculated monthly average snow depth based on daily snow depth for the growth season of 2012-2013. The dashed line and solid line in each violin plot represent the median value and the mean value, respectively.

This multi-faceted robustness assessment delivers two key conclusions. First, the consistent model performance rankings between the temporally independent validation (this section) and the contemporary validation (Section 4.2) provide strong evidence that the models have learned generalizable geophysical relationships rather than fitting dataset-specific artifacts. Second, the high internal consistency in daily-to-monthly aggregation, particularly for MLR and LSTM, validates the practical feasibility of scale transfer. It demonstrates that our framework can reliably produce a continuous, daily-resolution snow depth dataset extending back to 2012, thereby bridging a critical data gap prior to the modern altimeter era. For applications prioritizing long-term consistency and robust error metrics, MLR is affirmed as the most reliable choice, while LSTM remains optimal for minimizing mean bias in the derived products.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

5 Discussion

Quantification of retrieval uncertainties through error propagation is first presented, comparing the uncertainty characteristics of the four algorithms and interpreting why ensemble methods exhibit lower internal uncertainty. The retrieved snow depths are then applied to sea ice thickness estimation, with validation against airborne measurements and an examination of the spatial correlation between snow depth and ice thickness as a diagnostic of physical coherence.

5.1 Quantification and analysis of retrieval uncertainties

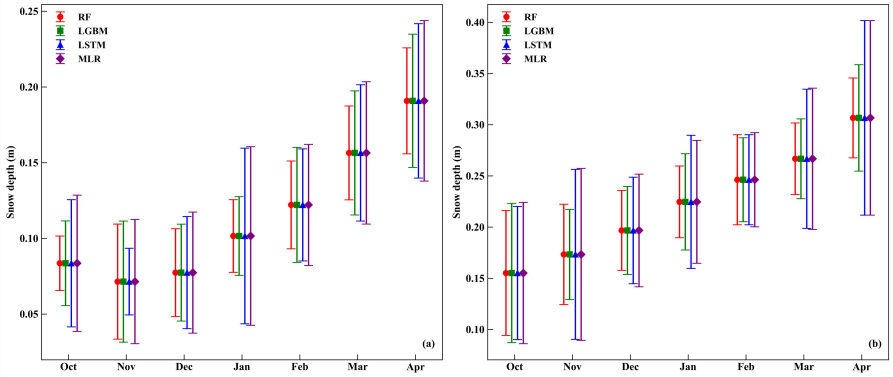
A systematic uncertainty assessment is crucial for evaluating the reliability of the snow depth products and interpreting differences among algorithms. We quantified the total retrieval uncertainty by propagating key input uncertainties through each model. Input variables (brightness temperatures, ASD) were modeled as normal distributions using their observed means and standard deviations. The primary uncertainty sources included 1) AMSR2 brightness temperature measurement uncertainty (standard deviation of 0.5 K); 2) the uncertainty of ASD training target, represented by its standard deviation against OIB measurements (0.06 m; Li et al., 2024); and 3) SIC retrieval uncertainty (±5%), analyzed across the relevant SIC range (80-100%). For the non-linear models (RF, LGBM, LSTM), uncertainties were estimated using a Monte Carlo approach with 1000 iterations. For MLR, uncertainty was derived analytically via Gaussian error propagation.

The estimated uncertainties for monthly snow depth, illustrated in Figure 8, reveal distinct magnitudes and patterns across algorithms and ice types. Over FYI, uncertainty ranges were: RF (1.8-3.8 cm), LGBM (2.6-4.4 cm), LSTM (2.9-5.8 cm), and MLR (4.5-6.9 cm). Over MYI, all uncertainties increased, with ranges expanding to: RF (3.5-6.1 cm), LGBM (3.9-6.8 cm), LSTM (4.4-9.5 cm), and MLR (5.9-9.8 cm). This amplification on MYI is consistent with the greater physical complexity and heterogeneity of the snowpack on older ice.

The systematic differences in uncertainty among the algorithms stem from their inherent methodological characteristics. The ensemble tree-based models (RF and LGBM) exhibited the lowest uncertainties. This robustness can be attributed to their built-in variance reduction mechanisms. RF utilizes bagging and random feature selection, while LGBM employs gradient boosting. Both techniques effectively average out noise and stabilize predictions against perturbations in input data. The LSTM network, while powerful for capturing temporal dependencies, showed higher uncertainty. Its sequential architecture and larger parameter space can make it more sensitive to input noise and variability, allowing small errors to propagate through its memory cells, which is reflected in the broader uncertainty bounds. The MLR model exhibited the highest quantitative uncertainty. This is primarily due to two factors: 1) Multicollinearity among the input microwave features, which inflates the variance of the estimated regression coefficients under perturbation; and 2) its inability to model nonlinear relationships, meaning any unaccounted-for, non-linear signal variability is treated as irreducible uncertainty.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

675 Importantly, this uncertainty ranking (RF/LGBM < LSTM < MLR) provides a complementary perspective to the validation-based performance ranking discussed in Sections 4.2 and 4.3. It underscores that lower predictive uncertainty does not universally equate to higher independent accuracy. For instance, while MLR has the highest quantitative uncertainty, it demonstrated the most robust correlation against OIB. Conversely, LSTM's higher uncertainty aligns with its greater 680 sensitivity to data specifics, which enables low-bias estimations but also leads to higher variance in independent validation (e.g., lower R). This analysis confirms that ensemble methods (RF, LGBM) offer the most stable and precise retrievals from a repeatability standpoint, whereas the choice for optimal accuracy must balance this intrinsic stability with the model's generalizability, as validated externally.



685 **Figure 8: Monthly average snow depth on FYI (a) and MYI (b) from LGBM, MLR, RF and LSTM. The error bars display the uncertainty of the snow depths.**

5.2 Application to sea ice thickness retrieval and validation

SIT was calculated using the hydrostatic equilibrium principle. The primary input was the Level-4 Sea 690 Ice Freeboard product from CryoSat-2 (FT4), which provides the elevation of the ice surface above sea level (sea ice freeboard). It is important to note that this dataset is already corrected from the original radar freeboard. Four distinct SIT datasets were generated by applying the snow depths retrieved by our four machine learning algorithms (LGBM, MLR, RF, LSTM) into the hydrostatic equation. For a definitive baseline comparison, the operational NSIDC SIT product (FT4), which utilizes MW99 snow 695 climatology, was also included. To enable a direct, like- for- like comparison with FT4 and to attribute performance differences solely to the snow depth input, all other parameters were kept identical, with only the snow depth replaced by our retrievals..

The derived SIT estimates were validated against independent, airborne-derived thickness measurements from OIB. After applying a quality filter to exclude points with high OIB uncertainty, 700 427 collocated matchups were retained for analysis. The validation statistics are presented in Table 6. SIT derived using the machine learning-based snow depths shows a substantial improvement over the climatology-driven FT4 product. Biases for the LGBM, MLR, RF, and LSTM-based SIT range from 0.04 to 0.07 m, in stark contrast to the 0.36 m bias exhibited by FT4. This result quantitatively

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

demonstrates that the systematic error introduced by the outdated MW99 snow climatology is a major contributor to SIT uncertainty in the Arctic, and that our dynamically retrieved snow depths effectively mitigate this error.

Table 6 Comparison between OIB SIT and CryoSat-2-derived SIT using snow depths from four algorithms and MW99 snow depth (RMSE/bias/MAE units: m).

	LGBM	MLR	RF	LSTM	FT4
N	427	427	427	427	427
RMSE	0.48	0.47	0.48	0.51	0.48
bias	0.05	0.04	0.04	0.07	0.36
MAE	0.38	0.38	0.39	0.42	0.48
R	0.61	0.61	0.61	0.57	0.62

The hydrostatic equilibrium equation inherently links snow depth and SIT. Therefore, the observed correlation between these variables (Figure 9) is expected and is not presented as a novel physical discovery. Instead, this analysis serves as a diagnostic tool to assess the physical coherence of the retrieved snow depth within the SIT retrieval system. The spatial pattern of the correlation coefficient provides valuable insight. The relationship is significantly stronger in the marginal ice zones, which are predominantly covered by thinner FYI, compared to the central Arctic dominated by thicker MYI. This pattern indicates that uncertainty in snow depth is a more dominant controlling factor for the accuracy of retrieved thickness over thin FYI. This finding validates the geophysical plausibility of the snow depth product and highlights the regions where accurate snow information is most critical for SIT monitoring.

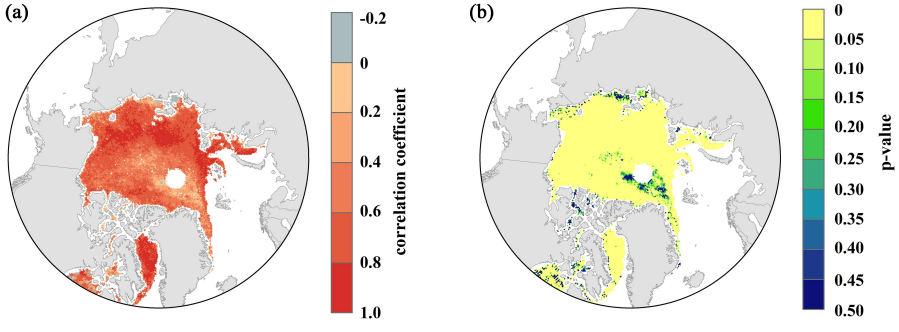


Figure 9: Correlation between SIT and snow depth retrieved by the MLR algorithm on the left (a). Reliability of the correlation relationship (p-value) is shown in the panel on the right (b).

6 Conclusions

This study developed a multi- model machine learning framework to retrieve snow depth over Arctic sea ice (2012-2021) by synergising satellite altimeter and passive microwave observations. The core value lies in providing a data-driven solution to algorithm selection, addressing multi-source data fusion and temporal coverage.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

Four models were rigorously intercompared: MLR, RF, LGBM, and LSTM. While complex models (RF, LGBM, LSTM) performed better on internal testing data, the simpler MLR showed superior generalisation against independent OIB data (R = 0.67, RMSE = 7.19 cm), highlighting overfitting risks when training data are limited. MLR is therefore recommended for long-term, pan-Arctic snow depth records prioritising correlation and error minimisation. LSTM is recommended when accurate local mean estimates are critical, given its near-zero bias against MOSAiC.

Permutation feature importance identified snow density and the 89 GHz polarisation ratio (PR(89)) as the most influential parameters. Importantly, PR(89) was markedly more important for MYI than for FYI, providing direct evidence for ice-type-dependent volume scattering and enhancing physical interpretability.

The framework demonstrated temporal extensibility: daily snow depth retrieved for 2012-2017 (outside the 2018-2021 training period) with performance consistent with contemporary validation. High internal consistency between monthly aggregates from daily retrievals and direct monthly estimates validated temporal downscaling, enabling a continuous daily-resolution snow depth record.

Applying our snow depths to sea ice thickness (SIT) estimation reduced SIT bias against OIB from 0.36 m (using MW99 climatology) to 0.04 m. This confirms that inaccurate snow depth is a major error source in operational SIT products and that our method offers a significant correction. Spatial correlation further indicated that snow depth uncertainty most critically affects thin FYI in marginal ice zones.

Future work should incorporate spatiotemporal deep learning (e.g., ConvLSTM, graph neural networks) and higher-resolution active microwave data. The recently launched AMSR3 sensor, with similar channels to AMSR2 plus additional high-frequency bands, could extend our framework to improve retrievals over thin snow and melt onset. This study underscores that systematic model comparison and physical interpretation are as valuable as the geophysical product itself, providing practical guidance for future algorithm selection in multi-source satellite data fusion.

Acknowledgements

Data support is gratefully acknowledged: AMSR-E/AMSR-2 brightness temperatures, snow density, sea ice thickness, MW99, and Operation IceBridge snow depths from the National Snow and Ice Data Center (NSIDC); MOSAiC data from the Alfred Wegener Institute (AWI); sea ice type and concentration from the Ocean and Sea Ice Satellite Application Facility (OSI SAF); and Rostosky snow depth from PANGAEA. We thank colleagues who provided helpful discussions and technical assistance. This work was supported by the National Natural Science Foundation of China (Grant No. 42301151) and the China Postdoctoral Science Foundation (Grant No. 2022M712853). Bin Cheng was supported by and the Research Council of Finland project IceScales 714 (grant 364939). Zheng Duan also acknowledges support from the Crafoord Foundation, Sweden (Grant No. 20240857).

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

Conflict of Interest

The authors declare no conflict of interest. At least one of the (co-)authors is a member of the editorial board of The Cryosphere.

Authors contribution

Mengmeng Li conceived and designed the study, developed the methodology and wrote the original draft. Yingfei Wang, Yang Li and Yafei Nie developed the methodology and revised the manuscript. Jianwei Ma, Juha Karvonen, Bin Cheng, Haili Li, and Zheng Duan contributed to the discussion and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Data availability statement

The daily and monthly gridded snow depth estimates and the machine learning codes generated from this study are available at figshare https://doi.org/10.6084/m9.figshare.31123483. AMSR-E/AMSR-2 bright temperature, snow density, sea ice thickness, MW99, and OIB snow depths can be obtained freely from the National Snow and Ice Data Centre (NSIDC, http://nsidc.org). MOSAiC data were provided by the Alfred-Wegener-Institute (AWI, https://www.meereisportal.de/en/). Sea ice type and sea ice concentration data were obtained from the Ocean and Sea Ice Satellite Application Facility (OSI SAF). Rostosky snow depth were from the website(https://doi.pangaea.de/10.1594/PANGAEA.902747?format=html#download).

References

Alexandrov, V., Sandven, S., Wahlin, J., and Johannessen, O. M.: The relation between sea ice thickness and freeboard in the Arctic. The Cryosphere, 4, 373-380. https://doi.org/10.5194/tc-4-373-2010, 2010.

Armitage, T., and Ridout, A.L.: Arctic sea ice freeboard from AltiKa and comparison with CryoSat-2 and Operation IceBridge. Geophys. Res. Lett., 42(16), 6724-6731. https://doi.org/10.1002/2015GL065439, 2015.

Comiso, J.C., Cavalieri, D.J., and Markus, T.: Sea ice concentration, ice temperature, and snow depth using AMSR-E data. IEEE T. Geosci. Remote, 41(2), 243-252. https://doi.org/10.1109/TGRS.2002.808317, 2003.

Guerreiro, K., Fleury, S., Zakharova, E., Rémy, F., and Kouraev, A.: Potential for estimation of snow depth on Arctic sea ice from CryoSat-2 and SARAL/AltiKa missions. Remote Sens. Environ., 186, 339-349. https://doi.org/10.1016/j.rse.2016.09.007, 2016.

He L., Xue B., Huang S., Hui F., Chen Z., and Cheng X.: On the Synergy of SMAP and AMSR2 for Estimating Snow Depth on Arctic Sea Ice. IEEE Geosci. Remote S., 19, 1-5.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

https://doi.org/10.1109/LGRS.2022.3188001, 2022.

He, L., Xue, B., Hui, F., Xu, S., Chen, Z., and Cheng, X.: Towards daily snow depth estimation on Arctic sea ice during the whole winter season from passive microwave radiometer data. IEEE T. Geosci. Remote, 62, 1-15. https://doi.org/10.1109/TGRS.2024.3358340, 2024.

Kacimi, S., and Kwok, R.: Arctic snow depth, ice thickness, and volume from ICESat-2 and CryoSat-2: 2018-2021. Geophys. Res. Lett., 49(5), e2021GL097448. https://doi.org/10.1029/2021GL097448, 2022.

Kilic, L., Tonboe, R.T., Prigent, C., and Heygster, G.: Estimating the snow depth, the snow-ice interface temperature, and the effective temperature of Arctic sea ice using Advanced Microwave Scanning Radiometer 2 and ice mass balance buoy data. The Cryosphere, 13, 1283-1296. https://doi.org/10.5194/tc-13-1283-2019, 2019.

King, J., Skourup, H., Hvidegaard, S.M., Rösel, A., Gerland, S., Spreen, G., Polashenski, C., Helm, V., and Liston, G.E.: Comparison of freeboard retrieval and ice thickness calculation from ALS, ASIRAS, and CryoSat-2 in the Norwegian Arctic, to field measurements made during the NICE2015 expedition. J. Geophys. Res.-Oceans, 123(2), 1123-1141, https://doi.org/10.1002/2017JC013161, 2018.

Koo, Y., R. Lei, Y. Cheng, B. Cheng, H. Xie, M. Hoppmann, N. T. Kurtz, S. F. Ackley, A. M. and Mestas-Nuñez: Estimation of thermodynamic and dynamic contributions to sea ice growth in the central Arctic using ICESat-2 and MOSAiC SIMBA buoy data. Remote Sens. Environ., 267, 112730, https://doi.org/10.1016/j.rse.2021.112730, 2021.

Kurtz, N.T., and Farrell, S.L.: Large-scale surveys of snow depth on Arctic sea ice from Operation IceBridge. Geophys. Res. Lett., 38(20), L20502. https://doi.org/10.1029/2011GL049216, 2011.

Kurtz, N., Studinger, M., Harbeck, J., Onana, V., and Farrell, S.: IceBridge sea ice freeboard, snow depth, and thickness. NASA National Snow and Ice Data Center. http://nsidc.org/data/idcsi2.html, 2012.

Kwok, R., Kacimi, S., Webster, M.A., Kurtz, N.T., and Petty, A.A.: Arctic snow depth and sea ice thickness from ICESat-2 and CryoSat-2 freeboards: A first examination. J. Geophys. Res.-Oceans, 125, e2019JC016008. https://doi.org/10.1029/2019JC016008, 2020.

Lawrence, I., Tsamados, M., Stroeve, J., Armitage, T., and Ridout, A.: Estimating snow depth over Arctic sea ice from calibrated dual-frequency radar freeboards. The Cryosphere, 12, 3551-3564. https://doi.org/10.5194/tc-12-3551-2018, 2018.

Laxon, S.W., Giles, K.A., Ridout, A.L., Wingham, D.J., Willatt, R., Cullen, and R., et al.: CryoSat-2 estimates of Arctic sea ice thickness and volume. Geophys. Res. Lett., 40(4), 732-737. https://doi.org/10.1002/grl.50193, 2013.

Lee, S. M., Shi, H., Sohn, B. J., Gasiewski, A. J., Meier, W. N., and Dybkjaer, G.: Winter snow depth on Arctic sea ice from satellite radiometer measurements (2003–2020): Regional patterns and

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

trends. Geophys. Res. Lett., 48(15), e2021GL094541. https://doi.org/10.1029/2021GL094541, 2021.

Li, H., Ke, C.-Q., Zhu, Q., Li, M., and Shen, X.: A deep learning approach to retrieve cold-season snow depth over Arctic sea ice from AMSR2 measurements. Remote Sens. Environ., 269, 112840. https://doi.org/10.1016/j.rse.2021.112840, 2022.

Li, M., Ke, C., Cheng, B., Ma, J., Jiang, H., and Shen, X.: Inter-comparisons of Arctic snow depth products. Int. J. Digit. Earth, 17(1), 2376286. https://doi.org/10.1080/17538947.2023.2376286, 2024.

Li, M., Ke, C., Xie, H., Miao, X., Shen, X., and Xia, W.: Arctic sea ice thickness retrievals from CryoSat-2: seasonal and interannual comparisons of three different products. Int. J. Remote Sens., 41(1), 152-170. https://doi.org/10.1080/01431161.2019.1637961, 2020.

Liston, G.E., Itkin, P., Stroeve, J., Tschudi, M., Stewart, J.S., Pedersen, S.H., Reinking, A.K., and Elder, K.: A Lagrangian snow-evolution system for sea-ice applications (SnowModel-LG): Part I—Model description. J. Geophys. Res.-Oceans, 125(10), e2019JC015913. https://doi.org/10.1029/2019JC015913, 2020.

Maaß, N., Kaleschke, L., Tian-Kunze, X., and Drusch, M.: Snow thickness retrieval over thick Arctic sea ice using SMOS satellite data. The Cryosphere, 7(6), 1971-1989. https://doi.org/10.5194/tc-7-1971-2013, 2013.

Markus, T., and Cavalieri, D.J.: Snow depth distribution over sea ice in the Southern Ocean from satellite passive microwave data. Antarct. Res. Ser., 74, 19-39. https://doi.org/10.1029/AR074p0019, 1998.

Markus, T., Cavalieri, D. J., Gasiewski, A. J., Klein, M., Maslanik, J. A., Powell, D. C., Stankov, B. B., Stroeve, J. C., and Sturm, M.: Microwave signatures of snow on sea ice: Observations. IEEE T. Geosci. Remote, 44(11), 3081-3090. https://doi.org/10.1109/TGRS.2006.883134, 2006.

Nicolaus, M., Perovich, D.K., Spreen, G., Granskog, M.A., von Albedyll, L., Angelopoulos, M., et al.: Overview of the MOSAiC expedition: Snow and sea ice. Elementa, 10(1), 000046. https://doi.org/10.1525/elementa.2021.000046, 2022.

Perovich, D. K., Tucker, W. B., and Ligett, K. A.: Aerial observations of the evolution of ice surface conditions during summer. J. Geophys. Res.-Oceans, 107(C10), 8048. https://doi.org/10.1029/2000JC000449, 2002.

Petty, A. A., Webster, M., Boisvert, L., and Markus, T.: The NASA Eulerian Snow on Sea Ice Model (NESOSIM) v1.0: Initial model development and analysis. Geosci. Model Dev., 11(11), 4577-4602. https://doi.org/10.5194/gmd-11-4577-2018, 2018.

Rostosky, P., Spreen, G., Farrell, S.L., Frost, T., Heygster, G., and Melsheimer, C.: Snow depth retrieval on Arctic sea ice from passive microwave radiometers-Improvements and extensions to multiyear ice using lower frequencies. J. Geophys. Res.-Oceans, 123(10), 7120-7138. https://doi.org/10.1029/2018JC014028, 2018.

https://doi.org/10.5194/egusphere-2026-2504
Preprint. Discussion started: 1 June 2026

865 Rückert, J. E., Huntemann, M., Tonboe, R. T., and Spreen, G.: Modeling snow and ice microwave emissions in the arctic for a multi-parameter retrieval of surface and atmospheric variables from microwave radiometer satellite data. Earth Space Sci., 10(10), e2023EA003177. https://doi.org/10.1029/2023EA003177, 2023.

Shi, H., Lee, S. M., Sohn, B. J., Gasiewski, A. J., Meier, W. N., Dybkjær, G., and Kim, S. W.: 870 Estimation of Arctic winter snow depth, sea ice thickness and bulk density, and ice freeboard by combining CryoSat-2, AVHRR, and AMSR measurements. IEEE T. Geosci. Remote, 61, 1-18. https://doi.org/10.1109/TGRS.2023.3265274, 2023.

Warren, S.G., Rigor, I.G., Untersteiner, N., Radionov, V.F., Bryazgin, N.N., Aleksandrov, Y.I., and Colony, R.: Snow depth on Arctic sea ice. J. Climate, 12(6), 1814-1829. 875 https://doi.org/10.1175/15200442(1999)012<1814:SDOASI >2.0.CO;2, 1999.

Xia, W., and Xie, H.: Assessing three waveform retrackers on sea ice freeboard retrieval from Cryosat-2 using Operation IceBridge Airborne altimetry datasets. Remote Sens. Environ., 214, 456-471. https://doi.org/10.1016/j.rse.2018.06.016, 2018.

Zhou, Y., Wang, X., Lei, R., Zhang, C., and Zhang, Y.: AdaSA-SD (v1. 0): An Adaptive Seasonal 880 Algorithm for Snow Depth Retrieval Over Arctic Sea Ice. IEEE T. Geosci. Remote, https://doi.org/10.1109/TGRS.2025.3593433, 2025.

Zygmuntowska, M., Rampal, P., Ivanova, N., and Smedsrud, L.H.: Uncertainties in Arctic sea ice thickness and volume: New estimates and implications for trends. The Cryosphere, 8(2), 705-720. https://doi.org/10.5194/tc-8-705-2014, 2014.

885