Dear Reviewer #1,

We highly appreciate you taking your time to review our manuscript and provide valuable feedback to aid in addressing weak points and areas of concern, while also seeking to strengthen the novelty of the presented work. In the following, we address each of the points raised. Black text indicates the reviewer's comments. Blue text indicates our response and changes to the manuscript.

General Comments:

Brodylo et al.'s manuscript is well-written, structured clearly, and supported by strong graphical presentation, providing a straightforward exploration into snow depth and snow water equivalent (SWE) estimation using an ensemble machine learning approach. The integration of LiDAR, remote sensing imagery, and in-situ observations is logical and aligns well with the type of studies frequently published in this journal. However, I have several significant concerns regarding the novelty of the approach, methodological clarity, and the limited sample size—particularly for SWE estimation—that need to be thoroughly addressed before the paper can be considered for publication. I have outlined these major concerns, along with specific suggestions for improvement, in detail below.

Major Comments:

1. Currently, the paper's primary novel contributions are unclear to me. While the presented approach effectively integrates established practices (ensemble machine learning methods, LiDAR-based snow depth estimation), the methodological novelty seems incremental and primarily focused on application in the specific context of Sodankylä, Finland. Intuitively, an ensemble approach should outperform individual techniques; however, given the limited sample size—especially with SWE data (only around a dozen observations)—it becomes challenging to conclusively demonstrate superiority over simpler, more traditional methods such as multiple linear regression. Indeed, as highlighted in Table 3, some machine learning models significantly underperform in certain months, likely due to this limited dataset. Thus, at present, the main takeaways and broader scientific significance are somewhat ambiguous. I encourage the authors to clearly articulate the core contributions of their approach, considering the constraints posed by dataset size. If a stronger case for novelty can be made, particularly in comparison to simpler or previously established methods, this would greatly strengthen the manuscript, as I am currently unsure of the main takeaways.

As noted in comment 3, we will add a convolutional neural network (CNN) model into the modeling approach. This deep learning model will serve as a contrast to the more commonly utilized and well-known machine learning methods and will serve to demonstrate how well each of these (RF, SVM, ANN, and MLR) perform against CNN. We will also be able to compare the CNN results to the machine learning-based EA results and identify which may be superior for snow depth or SWE estimation from our study site, with possible implications for other snow prone areas. The CNN model will also be included into the EA modeling in a hybrid-like approach of machine learning and deep learning, with it being possible to note how much of a positive influence, if any, this may have on the output metrics compared to a machine learning only EA model. A weighted, hybrid model that combines deep learning with traditional machine learning can attempt to fuse the intricacy of multiple layers of neural networks with the relative simplicity of more traditional models for snow depth estimation over a winter timeframe, before then estimating SWE with the predicted snow depth values. To aid in providing more valuable input data, we will generate polygons with a ~3 m radius at each observed field point that will contain average raster band values. Previously, the band values for each point were linked to image objects that could be long and thin, or have the point located near the corner or edge and thus not truly represent the band values of nearby features such as trees or water. This will ensure that the input data for this new modeling approach better incorporates the spatial context of surrounding features and will improve modeling performance.

2. Further clarity is needed regarding the training and validation processes for the machine learning models. The authors briefly mention using a "k-fold" validation but do not clearly specify how the data was partitioned into training, validation, and test sets at each step. Important details are missing, such as whether splits were random or sequential—random splits could inadvertently introduce spatial autocorrelation issues. Additionally, specifics on the machine learning implementations are essential. For instance, how deep were the random forest trees allowed to grow? What structure was adopted for training the multi-layer perceptron—including the number of hidden layers, neurons per layer, activation functions, epochs, and optimization methods? Providing visualizations of training and validation curves for MLP models would also help clarify the model training and generalization processes. These details are crucial for reproducibility and fully understanding the robustness of the results.

Input data was randomly assigned into a 10-fold CV approach that separated the input data into a similar number of observations for the training or testing partitions. In each iteration, 90% of the data was utilized for training the model, while a different 10% was used to test that model. In the next iteration, a different 10% of the data would be used to test the model, with 90% of the data being for training. Thus, each observation is in the testing group 1 time and used to train the model 9 different times. Given that each iteration is run independently, each successive iteration does not result in the model learning from previous training/testing CV results which may result in biased outcomes. The outcomes are compared once all the iterations have been run with a mean of the model scores used for the final outcome. This will be included in the text. Chosen hyperparameter settings for each of the model inputs will be added into section 3.2 and explained how they contributed to model performance. In addition, the model parameters will also be placed into a table in the appendix to help with clarity about what was set for each model, as suggested by Reviewer #2.

3. Given the inherently spatial nature of snow depth and SWE, I'm curious if the authors considered employing machine learning methods specifically designed to leverage spatial dependencies in data. The current choice of models—MLR, RF, and MLP—generally treats each data point independently, potentially losing valuable spatial context unless explicitly provided as an input feature. Models that explicitly capture spatial information (e.g., convolutional neural networks like U-Nets, or vision transformer approaches) could better represent the spatial variability across diverse land types. Exploring spatially aware methods, despite your current dataset limitations, could significantly increase the novelty and impact of your study.

Currently our input data for snow depth and SWE contain raster values associated with image objects that may not properly account for the spatial variability of nearby features. For instance, the field data may have been in the corner of an image object, or in one that is long and thin, with either case not providing a true indication of the surrounding terrain in reference to the location of the actual field data. To better integrate spatial context to the modeling procedure, we will instead generate a ~3 m radius polygon around each field measurement that will include the average and standard deviation of the raster data such as spaceborne imagery band values, elevation, and canopy height. This will allow nearby features which may have affected the real-world snow depth and SWE values to be better connected to the observed field snow depth and SWE data before then repeating the modeling. We will also include a CNN model into our approach to compare how this method compares with the previous models, how it can be integrated into the ensemble approach, and what benefit that may result for EA with estimating snow depth and SWE.

4. Finally, I also feel that this paper would really benefit from a more comprehensive comparison to existing approaches in the literature. Although your method is LiDAR-derived, related studies by Bair et al. (2018), King et al. (2020), Liljestrand et al. (2024), Shao et al. (2022), and Vafakhah et al. (2022) (amongst others) have utilized similar ML methodologies (RF and neural-network-based architectures) to predict regional variations of SWE. A clearer positioning of your work in relation to these papers would not only help justify the novelty of your method but also allow readers to better appreciate your contributions relative to the current state-of-the-art approaches. Such contextualization could also probably help address some of the concerns I raise in Comment 1 regarding methodological novelty.

The mentioned studies will be included in the updated manuscript, and it will be discussed how the methodology here compares with the methodologies presented. These studies and many like them (ours included) have noted the use of ML methodologies like RF, SVM, and other regression-based models in predicting snow related features, often with the support of various remote sensing data. Here we focused on comparing commonly utilized regression-based ML models and a weighted ensemble model to first estimate snow depth in six instances over a winter period, before then utilizing the more numerous snow depth data to aid in estimating more limited SWE data over the same period. To further distinguish our work, we will incorporate a deep learning CNN model for comparison to the ML models and integrate it into the weighted ensemble approach. Thus, the final model would be a weighted, hybrid ensemble approach of machine learning and deep learning.

Minor Comments:

- **Lines 89:** With all the different datasets being used here, I wonder if a summary table listing their names, variables, resolution, and source would help better situate readers?

  A summary table listing different data types, sources, names, resolutions, etc. will be provided in section 2.2 to help more clearly visualize the datasets used in this approach.

- **Lines 162-163:** It wasn't totally clear to me what this RF classification scheme was referring to here? Why is this step necessary?

  The acquired Land Use Land Cover (LULC) data, while very helpful, was limited at a 20 m resolution that was coarse for the chosen study site and limited the ability to make clear connections with obtained field data and vegetation types. In addition, the LULC data was from 2018, and may have become more outdated since that time. Thus, we needed to downscale the data to the 2 m resolution of the WorldView-2 imagery and the LiDAR data to provide a clear connection between these values and the landcover types. Just changing the resolution would result in many misclassifications, especially with artificial features and in heterogenous areas. As a result, we needed to utilize a classification-based scheme to better connect findings with properly downscaled and classified land cover types. While many different classification models would have served well, the best performance was obtained from Random Forest, which was why it was chosen for this purpose.

- **Section 3.1:** I also don't fully understand this image segmentation step and how it is "*utilized as the spatial unit for image assessment*". Why does this need to be done for this project, and how are the resulting segments used in the models afterwards?

  All pixels found in a specified image are separated into groupings of similar pixel values. These grouped pixels are then converted into polygons across the entire study area, with each polygon representing grouped pixels that match real-world features such as a cluster of bushes or a small body of water. Each of these polygons, now referred to as image objects, will then contain the average and standard deviation of all raster-based data inputs which are separated into columns. The field data are then placed into image objects based on spatial location and modeling is performed. This was accomplished for this project as a pixel-based method can result in heightened variance in nearby pixel values that can result in extreme predictions, such as with the presence of shadows which were in the obtained imagery, or with rapid differences in predictions in very heterogenous areas. In addition, there may also be a potential mismatch between the imagery bands and the LiDAR data that may result in an individual pixel being incorrectly assigned a value. However, by using the image objects for the image assessment, it provides averaged band and LiDAR values that help to minimize extreme values, and thus provide more realistic values into the modeling approach. Once the modeling is completed, the predicted values are then assigned to all the image objects in the entire study area, as except for the field snow depth and SWE data, every image object contains the same set of raster data.

- **Lines 189-192:** I think this section is important, and I would add a little more detail describing each of these models and how they've been used in other studies, as they really underpin your main results. For instance, I'd mention bootstrapping and aggregation in the RF, and I would rework your description of the ANN (as the linkage to the human nervous system is somewhat spurious) and not a clear description of how it actually works (i.e., a feedforward directed acyclic graph connected with artificial neurons with nonlinear activation functions)

  We will expand upon the descriptions of each of the models listed in section 3.2 to provide greater clarity of what these models are and how they function. We will revise what was written for ANN and update it with text that better explains how it functions. Hyperparameters that were used in the models will also be included in this section.

- **Lines 203-204:** Do you know why the SVM performance so poor? I'm wondering if the sample was simply too small for this approach? This goes back to my earlier major point that the same issue with the limited SWE data is also likely impacting the other models. However, it does feel a bit odd to me to just choose to not include a model in some cases due to poor performance when using an ensemble approach

  SVM was dropped for SWE likely due to the lack of available field data, which as was seen with RF could result in poor performance, especially if outliers were present. As noted earlier, many of the field input data were joined to image objects but may have been spatially located in a corner or edge or be a part of a long and thin image object. All of these may have not represented the true surroundings of the obtained field data. This will be addressed by changing the field inputs to collect raster values within a ~3 m radius of each field point, thereby providing proper spatial context into the modeling. It is also possible that hyperparameters were overlooked or not functioning as expected due to human error. Based on our findings, we will update our reasoning for the performance of SVM in the revised manuscript.

- **Eqs. 1/2/3:** This is personal preference but these are all very common metrics that don't need to be explicitly defined in this work

  These three equations will be removed from the manuscript.

- **Lines 258-260:** From a physical perspective, what do you think is causing this large swing in performance for the ANN over these months? Is there something about the onset snow in December that makes this an especially challenging task for the NN?

  A possibility is that in December, which is in the early middle of the winter period, there is relatively little snow. Snow depth is thus less variable and is somewhat more uniform across the landscape regardless of canopy cover or vegetation type when compared to further in the winter period such as in March or April. In addition, in late autumn and early winter temperatures may rise above freezing and rain events may also occur, both of which may reduce and flatten the snowpack. Over the course of the winter period, the effects of frequent snowfall and wind patterns may have led to more noticeable differences in snow depth based on the landcover.

- **Table 1:** For this table and the others after, I am wondering if this would be more interpretable as a bar graph? Comparing so many numbers in a table like this can bit a bit challenging

  For the revised manuscript we will attempt to update these tables into graphs that convey the same information, but ideally in a manner that is more clearly legible. For Tables 1 and 3 these graphs will show the MAE and RMSE, alongside the coefficient of determination (R^2) instead of the Pearson's correlation (r) as requested by Reviewer #2. Tables 2 and 4 will likewise be updated into bar graphs that include the mean and standard deviation values, and be color coded to match the LULC values in the respective maps for ease of comparison.

- **Table 2:** Similar to my previous table comment

  See comment above for Table 1.

- **Figure 5:** The red->green color scheme for snow depth can be challenging to view for color blind individuals, and I would recommend moving to something more accessible

  Thank you for pointing this out. The color scheme for snow depth will be updated from red->green to a different variation of blue->orange, as is seen in Figure 5 (h) to make it more accessible for color blind individuals. This same color scheme change will be applied to Figures 7 and 8 for consistency and ease of comparison between figures.

- **Lines 318-319:** Was the SVM left out because it had bad performance everywhere for SWE? As you state, the RF was also inconsistent for SWE prediction, but was still included in this part of the analysis

  Correct, SVM was left out as it was largely producing poor metrics for SWE. It is valid that RF was also inconsistent with modeling results. It was chosen to remove SVM as neither individually nor when it was added into the ensemble analysis did it provide meaningful outcomes, and in all cases reduced ensemble performance. With RF, while it sometimes did result in poorer outcomes, it also had instances where it provided meaningful outcomes and benefited the ensemble analysis, and was thus included in the SWE modeling.

- **Lines 344-362:** I appreciate the detail the authors put into comparing SWE over various land cover types, however this section (and other similar paragraphs) are a bit challenging to parse in their current form. Currently, you list many statistics in a row, and it isn't fully clear to me what I am to take from all of these stats? I wonder if you could restructure these paragraphs to highlight the most important findings and relate those to what the predictive accuracy means for each land cover type?

  We will revise how this paragraph and other paragraphs like it are structured to particularly highlight the most important findings, while minimizing or eliminating findings that are minor or add little value. This will result in text that is more substantial with less repeated listings of statistics, and with a greater emphasis on connecting more notable findings to what the respective predictive accuracy indicates for the vegetative land cover types.

- **Lines 428-429:** When referring to EA here, it sounds as if it is it's own technique, but really it is just a combination of the MLR/RF/MLP. And this enhanced performance in the EA is because of high variability in individual models with biases which mostly cancel out resulting in a more stable prediction. So is this section speaking primarily to the high variability of individual models?

  Yes, the EA outputs are the result of a mixture of model outputs with high variability that can often cancel out, and thus generally lead to more stable predictions. The text in this section highlights that RF had the highest variation out of the listed models, specifically in terms of having a large range between positive and more discouraging modeling metrics, such as having a *r* value of .05 and 0.71. None of the other models listed experienced such a dramatic variation, despite utilizing the same input data but in different instances. This contrasted with ANN, MLR, and EA which tended to be more stable in each instance. We will clarify this in the revised text.

- **Line 430:** I would reword this sentence "*EA consistently produced the best or second best metrics, and generally produced the best metrics*"

  This sentence will be reworded in the revised manuscript as "*EA consistently produced the best or second best metrics compared to the base models.*"

- **Lines 471-475:** Could you have included reanalysis estimates from say ERA5 to provide temperature, humidity and pressure data to your models? While coarse, this would perhaps give you some additional information about the surrounding environmental context at the time of observation?

  We will investigate providing data such as temperature, humidity, and pressure into our models at the times of observations whether from remote sensing or ground-based systems. Similar data was also available with instruments in and around the chosen study area, however there were very few, with some spaced far away from the field transects.

Ideally, had such data been readily available at a finer resolution, we would have inputted it into our modeling as these factors can be related to the development and change of snow depth, SWE, and snow density over time. At minimum, including such data would allow for enhanced understanding of the surrounding environmental context at the time of field observations and would be worth including in the discussion.

- **Lines 501-502:** I would strongly recommend including some code for reproducing at least a subset of these results, perhaps in an interactive notebook uploaded to Google Colab with some test data? Then others could more easily test and build on what you have provided here

  In the revised manuscript, we will provide a link for code to be made available to the public. It will also include data so that others who are interested can test and build upon what was done in this manuscript, and to verify that the outcomes are reproducible.

References

Bair, E. H., Abreu Calfa, A., Rittger, K., & Dozier, J. (2018). Using machine learning for real-time estimates of snow water equivalent in the watersheds of Afghanistan. The Cryosphere, 12(5), 1579–1594. https://doi.org/10.5194/tc-12-1579-2018

King, F., Erler, A. R., Frey, S. K., & Fletcher, C. G. (2020). Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada. Hydrology and Earth System Sciences, 24(10), 4887–4902. https://doi.org/10.5194/hess-24-4887-2020

Liljestrand, D., Johnson, R., Skiles, S. M., Burian, S., & Christensen, J. (2024). Quantifying regional variability of machine-learning-based snow water equivalent estimates across the Western United States. Environmental Modelling & Software, 177, 106053. https://doi.org/10.1016/j.envsoft.2024.106053

Shao, D., Li, H., Wang, J., Hao, X., Che, T., & Ji, W. (2022). Reconstruction of a daily gridded snow water equivalent product for the land region above 45° N based on a ridge regression

machine learning approach. Earth System Science Data, 14(2), 795–809. https://doi.org/10.5194/essd-14-795-2022

Vafakhah, M., Nasiri Khiavi, A., Janizadeh, S., & Ganjkhanlo, H. (2022). Evaluating different machine learning algorithms for snow water equivalent prediction. Earth Science Informatics, 15(4), 2431–2445. https://doi.org/10.1007/s12145-022-00846-z

Thank you for the references, these will be included and cited in the manuscript.