Dear Reviewer #2,

We greatly appreciate your time reviewing our manuscript and providing valuable feedback to aid in addressing weak points and strengthening the clarity and methodology of the manuscript. 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.

The paper "Object-based ensemble estimation of snow depth and snow water equivalent over multiple months in Sodankylä, Finland," authored by Brodylo et al., investigates the use of four machine learning techniques and their ensemble for snow depth estimation. The estimated snow depths were then used to estimate SWE. Finally, the ratio of the modeled SWE to snow depths was taken to estimate snow density. In my estimation, the paper is well written. However, I have major comments regarding the methodological clarity.

1. In section 3.2, the authors mentioned using Artificial Neural Networks (ANNs), among other models. However, they did not mention the exact architecture of the ANN (e.g., feed-forward, convolutional, transformers, etc.) used. Without this information, it is difficult to evaluate the appropriateness of the ANN architecture used in the study.

   The architecture of the ANN model was a feed-forward model with a single hidden layer. Specifically, we used "nnet" from the "caret" package in the R programming language. This will be expanded upon and added to the text in section 3.2.

2. In section 3.2, the details of the hyperparameters of the ML models (SVM, RF, and ANN) used were not mentioned. For example, for ANN, in addition to the architecture type, it would be beneficial to add the number of layers and neurons per layer, the activation function used, regularization (if any), the number of epochs, and other important hyperparameters used. For SVM, the kernel used, gamma, tolerance, and other important hyperparameters should be specified. For RF, the number of trees, the maximum depth, the minimum number of samples required to be at a leaf node, the minimum number of

samples required to split an internal node, and other important hyperparameters should be specified. These details are essential for reproducibility.

The hyperparameters for the RF, SVM, and ANN models will be added into section 3.2 to provide greater clarity for how the models were tuned and brief explanations for their purpose in improving model performance. This section will include the hyperparameters for the convolutional neural network (CNN) model as suggested by Reviewer #1. These will also be included in a table of chosen hyperparameters in the appendix for added legibility and clarity.

3. Also, in section 3.2, the authors mentioned using 10-fold cross-validation. However, important details are missing.

    1. Was the 10-fold CV done on the entire dataset or just the training set?

        The 10-fold CV was performed on the entire dataset to generate model performance when making predictions with the RF, SVM, ANN, and MLR models. For EA, it was indirectly applied, in that the outcomes were based off the weighted combination of base model outputs which were obtained with CV.

    2. No details about the train/test split ratio and strategy (random, stratified, etc) were mentioned.

        The utilized 10-fold CV approach was utilized to randomly separate the input data into either training or testing. 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.

3. During the CV, how were hyperparameter configurations selected? Was it a grid search or Bayesian? A table of the hyperparameters tuned and their optimal values can be placed in the appendix.

   Hyperparameters were selected and configured based on a manual trial-and-error approach. We will include a table of chosen hyperparameters in the appendix of this work alongside their optimal values. This will be for all models including the addition of a convolutional neural network (CNN) model as suggested by Reviewer #1.

4. In section 3.3, the authors used Pearson's correlation as a measure of prediction accuracy. However, a perfect correlation does not necessarily mean that the model is good or that the predicted values are close to the true values. For example, cor(y, y) = cor(y, 20y) = cor(y, 300y) = cor(y, 10000y) = 1. That is to say, a model could be doing significantly worse and still have a perfect correlation. I encourage the authors to use the coefficient of determination instead. Please do not square the correlation coefficient; you can use r2_score in sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html) or see this link for the formula (https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score).

   For the measure of prediction accuracy, we will use the coefficient of determination (r^2) instead of the Pearson's correlation (r) for all models. As the code was produced in the R programming language, the coefficient of determination will be determined from "r.squared". This will ensure that the coefficient of determination is directly utilized and is not based upon squaring previously obtained correlation coefficient values.

5. This study uses 13 points of SWE and 88 points of depths to train the ML models. This is an extremely limited sample size for training any machine learning model, especially when trying to predict across 37,917 image objects with varying characteristics. This raises a serious concern about overfitting. With such a small training set, for example, for the SWE estimation problem, there's a high risk that the model would simply memorize the patterns in those 13 objects rather than learning generalizable relationships. Therefore, the authors should comment on how to validate the SWE across the upscaled

10 km^2. How did the authors ensure that the model wasn't overfitting for the SWE estimates? These points should be added to the discussion.

The issue of limited sample sizes, especially for SWE was considered during the development of this study. As the study site was in a 10 km^2 study area, this limited the influence of major weather patterns that would influence a portion of the study area, while not impacting other areas in the study. Here, if major weather events such as snowfall or change in temperature did occur, it would have been relatively uniform and over a similar start and stop period. In addition, land cover types were largely well covered with field snow depth data that may be related to SWE measurements, even if the SWE data was more limited. To address the overfitting issue, one method was using the cross-validation approach with 10 folds. With this, we could tune hyperparameters on the original training set, while also not letting the testing set be seen for selecting the final model. The cross-validation approach was also modified to minimized data leakage by incorporating preprocessing techniques such as standardization. In a simpler approach such as one time split between training and testing, the model could get lucky with a certain set of inputs for the testing/training phase, and then provide an unusually optimistic outcome that would give caution for overfitting or bias. If this occurs during the cross-validation approach that one iteration has extremely positive metrics, it would not be the final choice given that the final output would be an average of all 10 iterations, including those that provide poor metrics. In addition, multiple trial runs were done with a pseudorandom number generator when assigning the training / testing for the cross-validation modeling to verify if the outputted metrics were consistent or if there were any unusually high spikes. As multiple models were also tested in each instance, it was possible to compare the mapped results between all models. While in theory all models could be overfitted, they should ideally at minimum reveal similar trends in the data. If one or each model has notably differing results, it is a warning that something may not be right with the modeling and warrants further investigation. Lastly, given the limited data availability, model complexity was halted from become too elaborate and thus considering noise in the data that would invite the possibility of overfitting.

6. Line 204: The model weights should use another metric since correlation is not reliable based on comment 4. Also, I think adding the weighting formula would be helpful to readers.

   We will change the model weights to the coefficient of determination (r^2) based on comment 4. The weighting formula for how the ensemble analysis is performed will be placed in section 3.3 and explained.

7. Line 203: SVM was dropped due to poor performance. Could you please quantify "poor" in this scenario?

   Poor performance indicated instances where SVM provided unsatisfactory metric values that would have negatively impacted the ensemble analysis metrics and outcomes. SVM metric outputs for SWE estimation will be included and expanded upon in the revised manuscript.

8. Figure 3: One might think field snow depth and field swe are inputs. The authors should clarify in the caption that they are the outcome variables, not the input. Or they could represent output data with a different color.

   We will update Figure 3 to better reflect the distinction for snow depth and SWE being outcome variables and not as input variables. The figure caption mentions “Blue indicates input/output data”, and we see how this can confuse readers or misrepresent the methodology framework given that this includes input and output data but does not clarify which is which in the figure. We will apply a unique color to field snow depth and SWE, and then clarify in the caption that these are output variables.

9. Tables 1-4: Were these metrics obtained from the entire dataset or just the testing set?

   The values in Tables 1 and 3 were obtained from the data inputs in the cross-validation approach. The field data values in Tables 2 and 4 were obtained from the entire data where field input values were available and matched to vegetative landcover types. Local scale outputs here were obtained from the entire dataset and matched to vegetative landcover types.

10. The authors should comment on the transferability of the ML models in this study. Can we grab this model and apply it elsewhere? The authors could dedicate a paragraph to model transferability in the discussion.

    We will add a paragraph in the discussion that discusses the potential of the applied model to be utilized in other snow-prone regions of the world, and how this may be transferred to such regions. We will note the similarities in regions that have similar terrain, while also noting potential challenges in other snow-prone regions with differing conditions such as in mountainous terrain. In addition, it will be mentioned what data inputs would be necessary or beneficial to include to make the model function well.

11. Line 167: A period is missing between "scale" and "In OBIA".

    Thank you, this will be corrected.