The authors have greatly improved the quality of the manuscript. Here are some comments to enhance the scientific rigor:

Data Preprocessing (L215-216):

- **Standardization**: The authors mentioned that all inputs were standardized, but did not specify the standardization method used. Was it Z-score, min-max scaling, or robust scaling?
- **PCA**: The authors mentioned that PCA was used, but provided no implementation details. How many principal components (PCs) were used for the machine learning model fitting? Were the PCs selected based on explained variance? If yes, what threshold was used? The PCA step should also be represented in the flowchart in Figure 3, as the authors' statement about using PCA (L215-216) suggests that the raw data weren't used as features in the ML models but PCs. This information is important because it helps the reader assess the severity of overfitting due to the limited SWE samples.

Neural Network Architectures (Section 3.2):

- **ANN vs FFN**: The authors clearly specified that the ANN model used in the manuscript was a feed-forward model (L205). However, because all deep learning models are considered ANNs, I suggest that the authors change ANN to feed-forward network (FFN).
- **Lack of details about CNN**: The only description of CNN given by the authors was on L207-208. While the description is technically accurate, it leaves a crucial detail about the dimension of the CNN used. Was it 1D, 2D, or 3D CNN. The parameters in Table 1 of Appendix A suggest that a 1D CNN was used. However, these details could be mentioned in section 3.2, as not all readers read appendices. Most readers would assume a 2D CNN for image analysis, making the 1D architecture choice counterintuitive without explanation. Additionally, a brief justification would strengthen the rationale for a 1D CNN.

Metrics:

- **R^2 vs. r^2:** It appears that the coefficient of determination reported in this study was obtained from regressing the observed SWE/depth on the predicted depth. This was confirmed from Line 114 of Snow_Depth_Code.R of the source code. Mathematically, this is equivalent to the square of a Pearson correlation coefficient (r^2) between the observed and the predicted values. With this approach, there is a potential for overestimating the model performance. The reason is that as long as the observed and the predicted values trend together, the r^2 will be high. For example, cor(y, y)^2 = cor(y, 20y)^2 = cor(y, 300y)^2 = cor(y, 10000y)^2 = 1. That is to say, a model could be doing significantly worse and still have a perfect r^2. What should be reported is R^2: 1 – SSR/SST. The question R^2 answers is: "If I compare these predictions directly to the true values, how close are they to lying on the perfect 1:1 line?" That would give us a more accurate assessment of predictive accuracy.
- **K-fold metric vs. overall metric**: According to L213 – L214, metrics are determined from the average of the test folds. However, Figures 5 and 9 suggest that the coefficient of determination was obtained after the final predictions. Can the authors clarify this?

L449: The authors mentioned that "In comparison to the snow depth, there was a much smaller sample size, which led to greater model uncertainty and disagreement." I am not clear what disagreement is referring to in this situation. Is it disagreement between the different models or disagreement between the observed and the predicted SWE values? If the latter, the disagreement aspect doesn't seem supported by the results in the study. According to Figure 5, the coefficient of determination ranges between 0.86 and 0.92 for snow depth, but 0.93 and 0.97 (Figure 9).

L508: Could the authors clarify what kind of models they refer to as "uncomplicated?" Examples would be helpful.

Table vs Figures: Figures 4 and 8 could benefit from a tabular presentation.

Table 1 in Appendix A: the parameter column displays inconsistent information. For CNN, the column shows the actual layer configuration used, while for RF, SVM, and ANN, it displays the hyperparameter search grid rather than the selected values. The authors should standardize this

column to either show: (1) the final selected hyperparameters for all models, or (2) create separate columns for "hyperparameter search space" and "selected values."