Reviewer 3

Summary

This brief communication describes a new method to use inexpensive temperature sensors and machine learning to estimate snow depth in the Arctic, with cross-validation in temperate regions. The manuscript presents the results clearly and succinctly, and my only major comments relate to the presentation of information, rather than the analyses conducted. I recommend that this manuscript be published following minor revisions.

Thank you for your review of our manuscript. We appreciate the time and thought that you have put into your comments.

Major comments

The non-Arctic sites should be introduced somewhere in the methods – as it is, they come as a bit of a surprise in the results, making it difficult to track what data are used and how.

Thank you for your suggestion. We have introduced these sites earlier in the manuscript now.

Our abstract is very limited on space (100 words maximum), but we have added the following sentence:

P1-L[17]: "***It performs poorly at temperate sites with deeper snowpacks, partially due to training data limitations.***"

We have also added text to the introduction:

P2-L[8]: ***"The model was trained at two small sites on the Seward Peninsula, Alaska, USA, and evaluated at ten sites distributed across Alaska, Colorado, and New Mexico (USA), Svalbard (Norway), and Siberia (Russia)."***

These sites are also mentioned in the methods section:

P4-L[3]: ***"Further, we applied RF-Seward and RF-Below to ten evaluation datasets where T_{SG} and snow depth measurements were collocated (within approximately 5 m of each other). Sites were located in the United States (Alaska, Colorado, and New Mexico), Norway (Svalbard) and Russia (Siberia), with temperature sensors placed at the snow-ground interface or within the top 5 cm of soil (see Table C1)."***

We hope that these sites come as less of a surprise now that they are mentioned in the abstract and introduction.

I'm sure space is short, but I worry that the description in the abstract noting that the model performed "well" is a little bit misleading, as the RMSE = 0.15 m is among the lowest you report, and whether or not that should be considered good performance is a matter of judgement. I'd

like to see a little more nuance in the abstract – maybe a brief description of the conditions under which the model performs best and worst, with the relevant RMSE values provided.

Our abstract is limited to 100 words, but we have added that the model performs poorly at temperate sites (e.g., Colorado). We hope that our changes to the abstract clarify that the statistic of RMSE = 0.15 m only holds true for Arctic sites:

P1-L[16]: ***"The model performed well on Alaska's Seward Peninsula where it was trained, and at Arctic evaluation sites (RMSE 0.15 m). It performed poorly at temperate sites with deeper snowpacks, partially due to training data limitations."***

A full description of model limitations is provided in the conclusions.

Percent bias could also be helpful here, given that snow depth is so important to model performance.

We chose to present an RMSE value rather than a percent bias value in the abstract because we compare model performance across sites, and RMSE is directly comparable between sites whereas percent bias is not. For example, small errors at a site with low snow depths would likely result in high percent errors, even though the magnitude of errors is small.

However, we have added mean bias values to our results section (see P4-L[27-29], for example).

Minor comments

Line 19 – citation needed here, as this probably refers mainly to potential for increasing snow depth?

Thank you for noting this. We agree, and we have added the following citations:

Bigalke, S. and Walsh, J. E.: Future Changes of Snow in Alaska and the Arctic under Stabilized Global Warming Scenarios, Atmosphere, 13, 541, https://doi.org/10.3390/atmos13040541, 2022.

Pedron, S. A., Jespersen, R. G., Xu, X., Khazindar, Y., Welker, J. M., and Czimczik, C. I.: More Snow Accelerates Legacy Carbon Emissions From Arctic Permafrost, AGU Advances, 4, e2023AV000942, https://doi.org/10.1029/2023AV000942, 2023.

Line 25-26 – Sonic sensors are deployed at SNOTEL stations, along with snow pillows, but this currently reads as though sonic sensors and SNOTEL stations are two distinct types of monitoring equipment. Suggest rewording.

You are right that the wording was misleading. We have reworded this sentence:

P2-L[5]: ***"The temporal evolution of snow can be monitored using automated instruments (e.g., snow sonic sensors deployed at Snow Telemetry (SNOTEL) stations; Fleming et al., 2023), but spatially distributed deployment is time consuming and expensive. "***

Line 28 – Can you say why these remain a challenge in Arctic regions? In fact, I would expect IceSat-2 to provide better observations in polar than temperate regions, due to the higher sampling density.

Arctic snow depths vary across very fine spatial and temporal scales. This is partially because winds in many Arctic regions are very high, so snow blows across the landscape and redistributes quickly, which creates a patterned landscape of drifted and scoured areas. Satellites cannot capture those fine scale patterns because they operate at relatively coarse temporal and spatial scales. It is important to actually measure the fine-scale variations in Arctic snowpack because drifts may impact (warm) permafrost. For example, as shrubs expand, it is likely that where drifts form on the landscape will change.

We tied this reasoning into our introduction by adding a few sentences:

P1-L[26]: ***"As shrubs expand in the Arctic (Mekonnen et al., 2021), the spatial distribution of snow drifts and subsequent impacts on permafrost may change (Lathrop et al., 2024). Thus, monitoring and modeling fine-scale drifting processes are crucial to understanding permafrost evolution"***

P2-L[1]: "***Satellite data can be used to estimate snow depth (Besso et al., 2024), but spatial and temporal resolutions are too coarse to capture the complexity of Arctic snowpacks."***

Line 59 – I don't think this permutation importance is unique to RF; should remove as a reason for selecting RF. Your other reasons for selecting RF are perfectly good, though.

Thank you for catching this, we have rephrased those sentences. We still say that random forests are easier to interpret than other models because more feature importance metrics exist for random forests (e.g. gini importance) and individual trees can be examined to understand how the model is making its decisions.

The updated text is shown below:

P3-L[12]: ***"We chose a random forest as it outperformed or performed similarly to other models. A random forest is simple to design, computationally inexpensive, and easy to interpret. We identified key model features using permutation importance, which reflects how model performance changes when an input feature is randomly shuffled (Breiman, 2001). Larger decreases in performances indicate greater feature importance."***

Line 90 – I think this is the first time the other training sites are being introduced. They should be briefly described somewhere.

We have now introduced these sites in the abstract and introduction (see previous response).

Line 134 – I think you should define the zero-curtain period the first time you use the term.

Thank you for this suggestion. We first use the term "zero-curtain" when discussing Figure 1. We have updated the manuscript text to provide a more detailed description of a zero-curtain period:

P5-L[4]: ***"Further, warmer and/or wetter sites (e.g., Teller27) undergo more freezing and thawing than colder and/or dryer sites (e.g., Kougarok64), producing zero-curtain periods where the key snow depth predictor (temperature variability) flattens at 0°C as water changes phase (Staub and Delaloye, 2017)."***

Line 180-181 – I question whether future work should try to improve the technique for deeper snow – it seems that for physical reasons, this may be unlikely. Perhaps it would be more productive to discuss how the technique could be combined with other types of observations.

We agree that it is possible that this technique will never work for deep snow. However, the dataset used in this study had no deep snow estimates in it at all, so the model could not possibly predict deep snow even if some relationship with depth and temperature still existed. Because of this, we think it is worth exploring whether this technique works given a more representative training dataset. We did try to test this using "RF-Deep", but the data used to train that dataset was not as high quality as what we used to train RF-Seward and RF-Below, and we used far fewer training data points. We have provided a more nuanced discussion of this in the conclusions.

We also agree that this technique could be combined with observations/models to improve estimates even in regions where deep snow limits model performance. We have added some brief discussion around this in the conclusions as well. See below:

P9-L[26]: "***Future research should focus on developing this technique for locations where peak snow depths exceed 1.5 m (e.g., Colorado, USA), as these regions are crucial for water security across the world. While deep snow may completely dampen T_{SG}, it is possible that the ML model will perform better given a larger and more representative training dataset and/or additional input features. Alternatively, this technique could be combined with other monitoring and/or modeling efforts. For example, snow depth estimates made early in the snow season (e.g. when snow is shallow) could be used to estimate snow variability across the landscape and to downscale coarse model or remote sensing snow depth estimates."***

I also wonder about discussing a more thorough investigation of the relative merits of different ML models; an LSTM would make more sense conceptually but is probably harder to implement, and we're not given much information about the implementation you tried that didn't outperform the RF.

This is a great suggestion. LSTMs have the potential for modeling snow dynamics given their ability to capture temporal dependencies. However, they require sufficient data to learn these relationships. The lack of a complete snow cycle likely hindered the LSTM's ability to effectively learn the seasonal patterns. Additionally, there is a trade off between having more training samples with a shorter look-back window and having less samples with a longer look-back window. With a longer dataset encompassing multiple years, we anticipate that an LSTM could potentially improve performance. We summarized this briefly in the conclusions:

P9-L[31]: ***"Further, the application of a ML model tailored towards time series estimates (e.g., a Long Short Term Memory Model; LSTM) could improve predictions. In this study, we only had one year of data, which likely limited the LSTM's performance. With a longer-term dataset, we could provide the LSTM with more training points and a longer look-back window (e.g., an entire snow season), which would likely enhance its performance. "***

Citation: https://doi.org/10.5194/egusphere-2024-2249-RC3