Reviewer 2

This brief communication presents an interesting approach to derive snow depth through temperature data recorded with easy-to-deploy sensors. Authors exploit machine learning models (random forest) to predict snow depth from snow-soil interface temperature. While the brief communication reads well and it is suitable to be published in The Cryosphere, some points must be addressed before publication.

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

First of all, the approach tested is trained in two sites and then evaluated in these two sites, but also in 10 other sites, what I might highlight in both the abstract and the introduction.

Per your suggestion, we have highlighted this point in the introduction of the revised manuscript:

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).***"

The abstract is limited to 100 words so we did not have space to highlight this point there.

Through this test on model transferability it is clear that this approach works well in cold and high latitude areas, but in temperate areas where ROS events can occur or temperatures are milder, it fails. This has to be highlighted in the abstract and the conclusions.

We agree with your suggestion that the shortcomings of this approach should be highlighted earlier on. One thing to note is that some issues related to zero-curtains and warm, ephemeral snowpacks can largely be avoided when using temperature collected at the snow-ground interface. This is because below ground sensors are impacted by the soil freeze-thaw cycle, whereas the above ground sensors are not. In the abstract and introduction, we highlight results from RF-Seward, and therefore do not discuss the additional shortcomings of using RF-Below.

We added text to the abstract to highlight the shortcomings of our model at temperate sites:

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

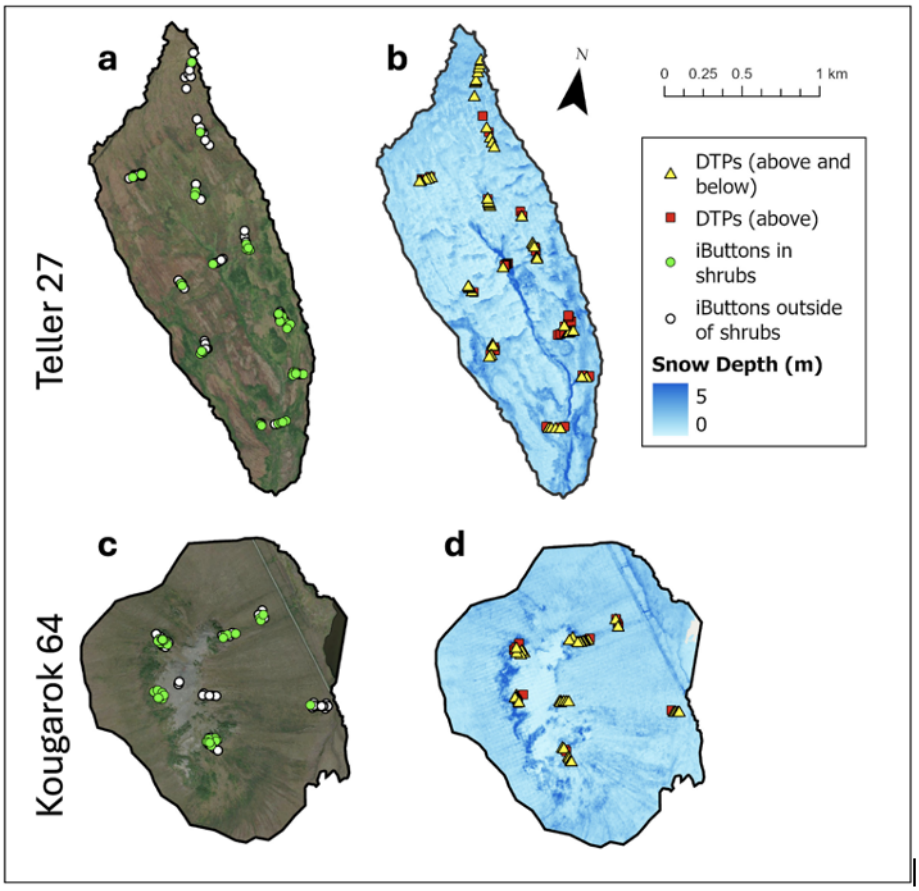
We were unable to mention the ROS limitation in the abstract because we are limited to 100 words, but that shortcoming is highlighted in the conclusions. The full description of shortcomings in the conclusion now reads:

P9-L[19]: "***While the model generally performed well, rain-on-snow events and zero-curtain periods cause the model to erroneously predict snow accumulation events. Further, the model failed to replicate deep snow depths (greater than 1.5 m) observed in Colorado, USA. For optimal performance, the model should be applied to temperatures***

recorded at the snow-ground interface. Predictions made using temperatures recorded below the ground surface were impacted by varying soil types, vegetation properties, and latent heat processes."

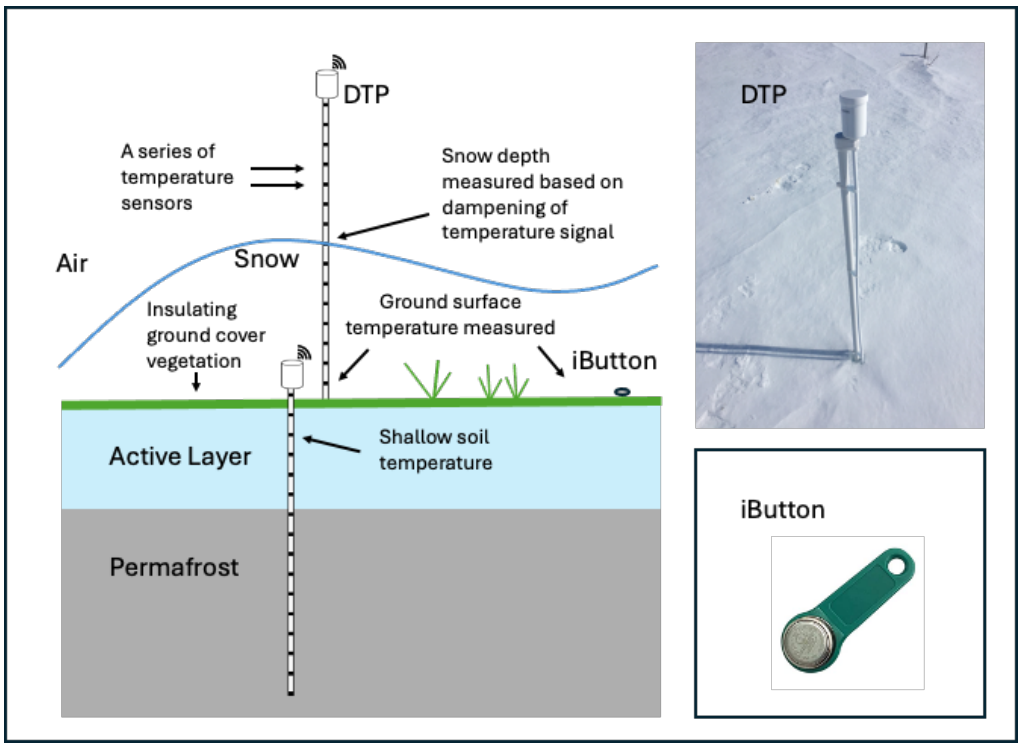
More details about the training study sites (spatial distribution of DTP's within the domain), image of the DTPs, and photograph of them would be desirable. I guess number of figures are limited, but some of these can be included in Figure 1.

A map of DTP locations was provided in the supplemental material and is copied below for your convenience:



"Figure A1. Locations of iButton Link Thermochron (DS1921G-F5#) temperature sensors deployed in (green circles) and outside (white circles) of shrubs over the 2022 – 2023 snow season at a) Teller27 and c) Kougarok64. Background imagery from Esri, Garmin, USGS, Maxar, 2024, ArcGIS RGB Basemap. Locations of DTP temperature sensors that recorded both above and below ground temperature (yellow triangles) or only above ground temperature (red circles) over the 2021 – 2022 snow season at b) Teller27 and d) Kougarok64. Blue background imagery shows snow depth in April 2022 estimated using Light Detection and Ranging (LiDAR) data (Singhania et al., 2023b, a)."

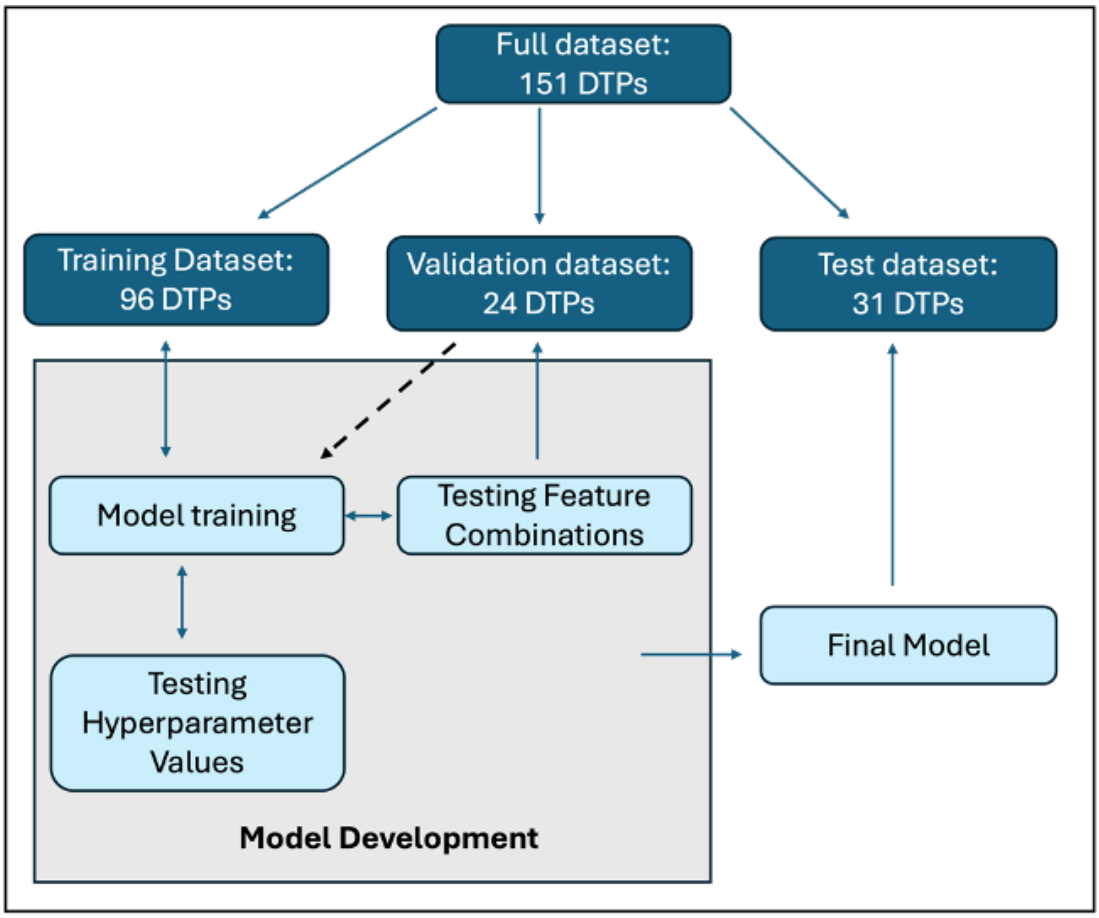
We also introduced an additional figure to the supplemental material (Figure A2) to provide more information on the DTP sensors:



"Figure A2. Set up of DTP and iButton sensors."

The application of the training, validation and evaluation datasets it is not clear. This point has to be clarified in methods section.

We have added a figure to the supplemental material (Figure B1) to visually show how the training, validation, and test datasets are applied during model development. The validation dataset is used to test how different feature combinations impact model performance. This way, none of the test dataset is used to inform the development of the final model. The new figure is shown below and referred to in the updated methods section.



"Figure B1. Use of training, validation, and test datasets in model development. We split the training data into groups of DTPs rather than groups of daily data points to maintain the independence of entire snow depth/ temperature time series during model testing. Different combinations of input features were tested using the validation dataset. After the best-performing set of input features was determined, the final model was trained using both the training dataset and validation dataset. The test dataset was excluded completely from the model development process."

Similarly, it is not clear if, for sites where the models are transferred, these are evaluated with a similar dataset of observation (DTPs spatially distributed) or just data compared with automatic weather station data from a single location.

Model predictions at these sites are compared with snow depth data collected at a single location within 5 m of the temperature data. Snow depth predictions were recorded using sonic sensors, except for at a site in New Mexico where we manually recorded snow depth. To clarify that we are using individual snow depth measurements and not measurements averaged over a network of sensors, we add an additional sentence below:

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). Snow depth was also recorded at the sites (e.g., snow sonic sensors at automated weather stations), and was used to evaluate model performance. "***

Minor comments

Line 30: I assume you already know somehow the spatial distribution of the snowpack in the study area (lidar/uav data?) or you are just modeling and testing in the exact location of your DTP sensors? I think it is the second but it is not clear.

The depth and density values presented here were from end-of-winter snow surveys conducted at the study sites. We clarified this in the text:

P2-L[16]: ***"According to end-of-winter snow surveys, the average peak snow depth from 2017-2019 at Teller27 was 0.96 m, with an average density of 310 kg/m^3 (Bennett et al., 2022). In 2018, snow depth was shallower at Kougarok64 than at Teller27, with an average end-of-winter depth of 0.75 m and density of 290 kg/m^3 (Bennett et al., 2022)."***

Line 38 and 39: Please include snow density units in the international system (Kg/m3).

We have corrected this throughout the manuscript and supplemental material.

Line 9. There are some works which have already exploited random forest to analyze, and simulate snow distribution, showing suitable performances. You might cite here: Meloche et al., 2022 (https://doi.org/10.1002/hyp.14546), Revuelto et al., 2020 (https://doi.org/10.1002/hyp.13951) and Hsu et al., 2024 (https://doi.org/10.31223/X57391)

We agree that these studies are relevant to our research. Bennett et al. (2022) developed a random forest machine learning model to predict peak SWE at our study site. Because we are limited on the number of citations we can include, we chose to cite the Bennett et al. (2022) study in our introduction as it is most relevant to our paper:

P2-L[3]: ***"Machine learning (ML) models can be used to extrapolate snow survey data, but these estimates still only represent a single point in time (Bennett et al., 2022)."***

Bennett, K. E., Miller, G., Busey, R., Chen, M., Lathrop, E. R., Dann, J. B., Nutt, M., Crumley, R., Dillard, S. L., Dafflon, B., Kumar, J., Bolton, W. R., Wilson, C. J., Iversen, C. M., and Wullschleger, S. D.: Spatial patterns of snow distribution in the sub-Arctic, The Cryosphere, 16, 3269–3293, https://doi.org/10.5194/tc-16-3269-2022, 2022.

Line 69-70: Did you apply an “out of the bag” approach to validate evaluate? I do not understand why you use a 24 DTP validation data and a 31 DTP evaluation dataset, which is the difference here? If not, why don’t you use an out of the bag test?

We did not use an “out of the bag” approach because the model is trained using daily data, and we wanted to hold out entire sensors for validation/testing rather than individual daily data points. We suspected that if we held out individual (daily) data points (as done in an “out of the bag” approach), our error estimates would underestimate model error, as the model likely would have seen similar data from neighboring days recorded using the same DTP sensor during model training. By holding out entire sensors, we hoped that our error estimates would be more realistic.

To clarify why we chose to split our training/validation/test datasets into groups of sensors, we add the following sentence to the caption for Figure B1:

“We split the training data into groups of DTPs rather than groups of daily data points to maintain the independence of entire snow depth/ temperature time series during model testing.”

We hope that our addition of Figure B1 helps clarify how we use the training/validation/test tests. Mainly, we use the validation set to evaluate how different combinations of input features impact model performance.

Lines 72-77: Impact of sensor burial. I would present this section on section 2.2.

Thank you for this suggestion. We have made this change in the updated manuscript.

Line 90: How many sensors are used to train RF-Deep in senator Beck Basin? is this a similar test area (i.e. same number of DTPs or equivalent sensors)?

Far fewer training data points were used to train RF-Deep than the other machine learning models. At Senator Beck Basin, there were two automated weather stations which recorded both snow depth and snow-ground interface temperature. We combined this data with the data collected at the Seward Peninsula. We then balanced the combined Seward Peninsula + Senator Beck training dataset such that deeper snow represented a reasonable proportion of the training data (10 %) to reflect the distribution of snow depths at the sites in Colorado. If we included the entire DTP training dataset, we worried that the model would remain biased low, as any snow depths above 1.77 m would reflect a very small percentage of data points. Our updated description of these methods is given below:

P4-L[13]: ***“The training data at our study sites was limited to a maximum of 1.77 m due to the length of DTP probes, and thus RF-Seward and RF-Below cannot predict depths greater than 1.77 m. To test if ML could accurately predict deeper snow depths, we trained a third ML model, which we refer to as “RF-Deep”. To train this model, we***

supplemented our original Seward Peninsula training dataset with additional data from two model evaluation sites in Senator Beck Basin, CO, USA with deeper snowpacks (Table C1). The model was applied to one site and trained with data from the other (in addition to the Seward Peninsula DTP data). To mimic the distribution of snow depths at these sites, we ensured that 10 % of the training data consisted of snow depths above 2 m. This reduced the training dataset size compared to other models (Table B2).”

We also added Table B2 in the supplemental material to show how many training data points were used to train each machine learning model:

Model	**Number of training data points**	**Related figure**
RF-Seward (applied to the test dataset)	**20,963**	**1a**
RF-Seward (trained at Teller27 and tested at Kougarok64)	**17,171**	**1b**
RF-Seward (trained at Kougarok64 and tested at Teller27)	**9,272**	**1c**
RF-Seward (retrained on all DTP data; applied to evaluation sites)	**25,418**	**2a-g, h**
RF-Below (applied to the test dataset)	**15,197**	**1d**
RF-Below (trained at Teller27 and tested at Kougarok64)	**11,396**	**1e**
RF-Below (trained at Kougarok64 and tested at Teller27)	**7,980**	**1f**
RF-Below (retrained on all DTP data; applied to evaluation sites)	**18,968**	**2c-h**
RF-Deep (applied to first Senator Beck Basin Site)	**1,305**	**2i**
RF-Deep (applied to second Senator Beck Basin Site)	**3,294**	**2j**

Table B2. Number of training data points (days) used to train the random forest models.

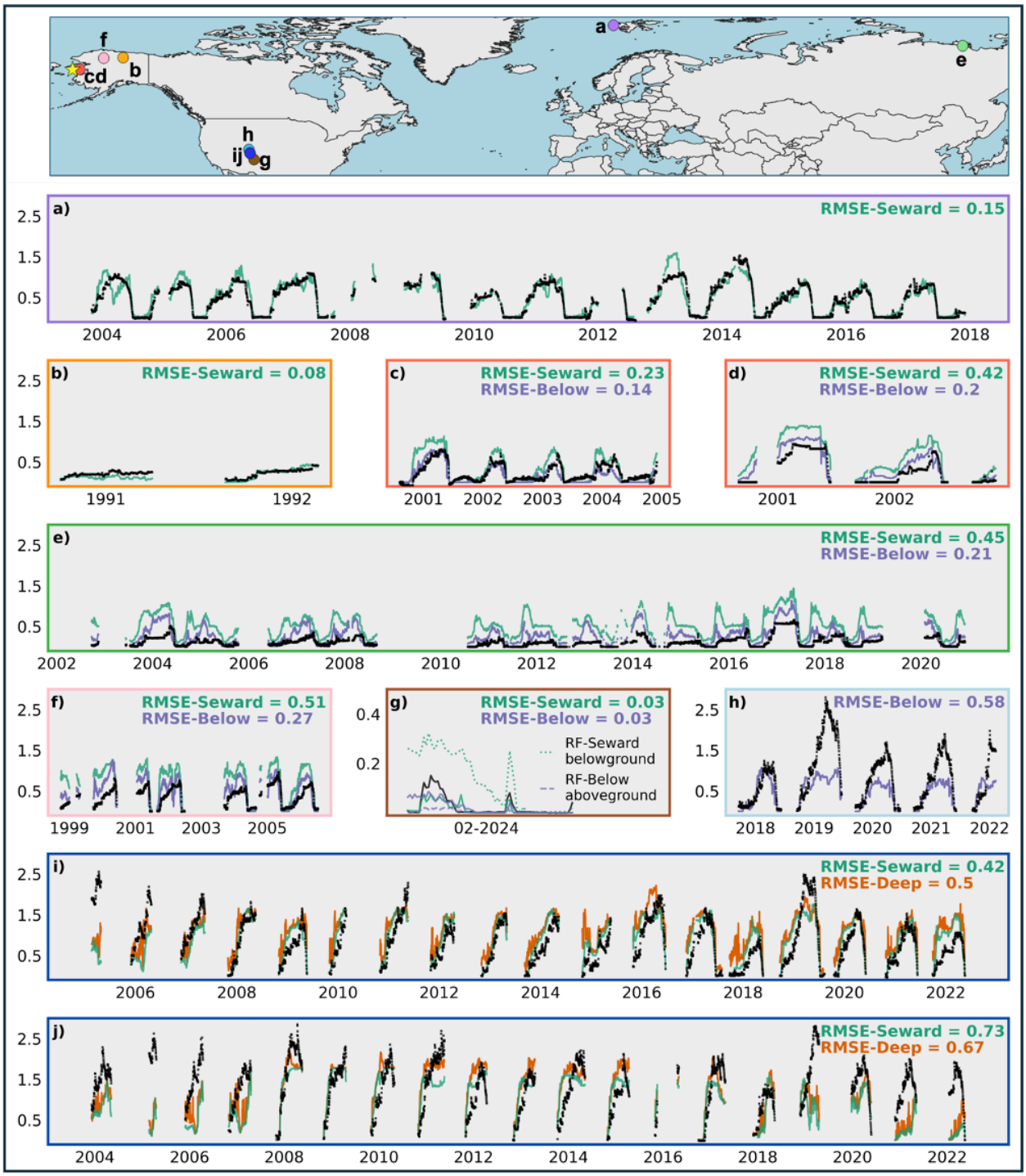
Line 114. I would briefly state here how do you test these models. You are directly comparing the observed snow depth at the sensor location in different stations with that modeled, right?

Thank you for this suggestion. We are testing these models by comparing them to snow depth measured at the site. We have clarified this in the methods:

P4-L[6]: “***Snow depth was also recorded at the sites (e.g., snow sonic sensors at automated weather stations), and was used to evaluate model performance.”***

Figure 2. Some symbols of the study area are quite difficult to identify (eg. Bayleva station or Siberian), please increase their size. Also captions and graphs sizes are too small, can this figure be extended and increase captions size. For instance, you can remove the names above the graphs and just include the letter inside each one (a), b), c),…).

We have made the caption and graph sizes larger as you suggested. We also added letters (a,b,c, etc.) to the site map to make the symbols easier to identify and pair with their corresponding time series plot. The updated figure is shown below:



"Figure 2. ML performance at a) Bayelva Station, Svalbard, Norway; b) Imnaviat Creek, Alaska, USA; c,d) Council, Alaska, USA; e) Samoylov Island, Siberia, Russia; f) Ivotuk, Alaska, USA; g) Los Alamos, New Mexico, USA; h) Grand Mesa, Colorado, USA; and i,j) Senator Beck Basin, Colorado, USA. Locations are shown on a map, with the yellow star indicating the Seward Peninsula of Alaska, where RF-Seward was trained. Black lines show measured snow depth at each site. Y-axis and RMSE values indicate snow depth in

meters. f) Note adjusted y-axis for Los Alamos, New Mexico. For this site, we also show RF-Seward and RF-Below predictions when RF-Below was applied above ground and RF-Seward was applied belowground (dotted lines).”

Conclusions: It must be highlighted that this method is suitable to predict snow depth in cold regions and that its applicability in temperate areas must be further investigated.

Thank you for your review of our manuscript. We hope that our additions to the abstract, introduction, and conclusions help highlight this point.

Citation: https://doi.org/10.5194/egusphere-2024-2249-RC2