Reviewer 1

Summary & General Comments:

This brief communication presents a novel approach to derive snow depth from low-cost temperature sensors deployed at the snow-ground interface using a random forest model. This method would hypothetically allow snow depth monitoring at far greater number of sites than currently available, a significant finding well within the remit of The Cryosphere journal. The manuscript is concise and well-written. I recommend this manuscript be published subject to minor revisions, as detailed below.

Thank you for your review of our manuscript. We appreciate your thoughtful feedback.

As an aside, I would be interested to see what you think the impact of snow stratigraphy may be on the depth estimates from your model (particularly when it comes to evaluating snow depth at sites beyond the Arctic, such as the New Mexico site), but I understand that you are unlikely to have this data for comparison.

This is an interesting question which we have not yet explored. I would think that the accuracy of the machine learning algorithm may decrease at locations where the stratigraphy (and more generally, snowpack physical properties) is different from where the model has been trained, but we do not have the data to test this. We have added a mention of snow stratigraphy to the conclusions to address this area of uncertainty:

P10-L[3]: ***“Additionally, how snow stratigraphy and density affect model results remains unclear .”***

Minor/Technical Comments:

This is a broad and minor stylistic comment, but I would remove the italics for above and below ground throughout.

We have removed italics from above and below ground and agree that it improves the manuscript.

Section 2.1: Could you add a photo of one of your DTPs to Fig S1? Please also give an indication of how deep into the soil these profilers go, and when they were deployed relative to the start of the snow season.

Thank you for your suggestion. We have added to the Supplemental section Figure A2 showing the DTP and iButton setup, and we provide pictures of both instruments in this figure (see below). The DTP sensors measure temperature down to 100 cm to 160 cm depending on the DTP design. We have incorporated this information into the figure caption below. DTPs were deployed late September 2021 and snowfall started on October 20, 2021, information which we have also incorporated into the Methods section of the paper.

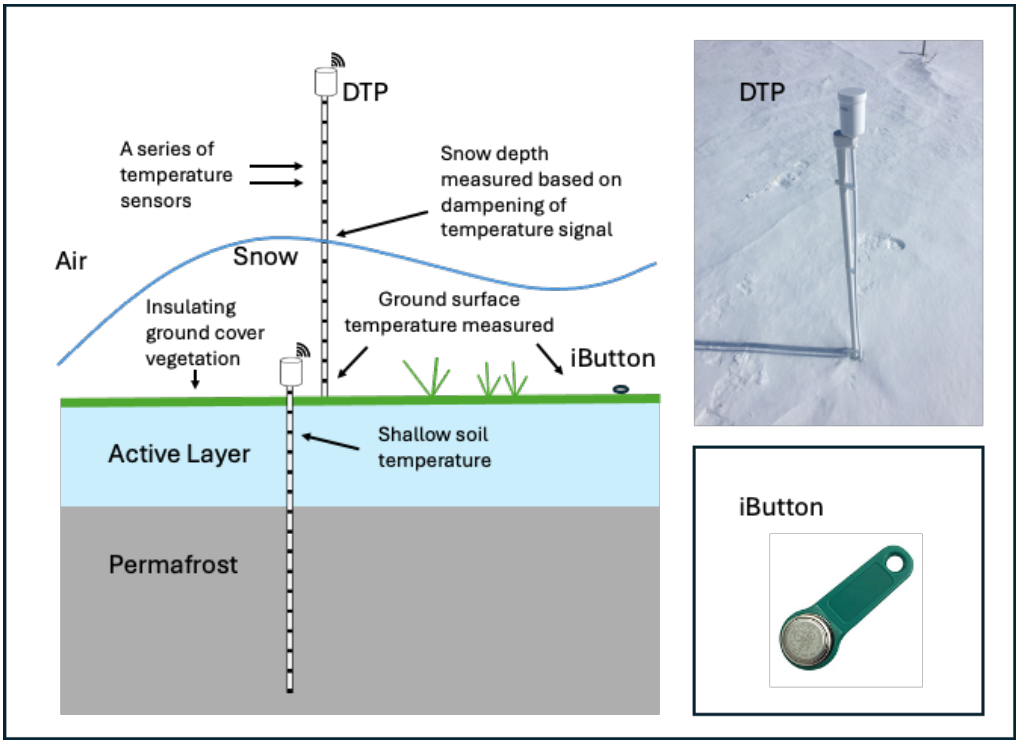


Figure A2. Set up of DTP and iButton sensors. Only shallow soil temperature data was used in this study, but the soil DTPs can measure temperature down to 100 to 160 cm of depth.

Line 46: Please give the precision of the snow depth estimates.

The estimated snow depth has an uncertainty of ±2.5 cm or ±5 cm, depending on the sensor spacing (5 and 10 cm, respectively). We have incorporated this information into the Methods section of the paper.

Line 47: Is the value of the closest temperature sensor used as the value for T_{SG}, or is T_{SG} estimated from the sensor temperature using another method (such as a linear extrapolation)?

The value of the closest temperature sensor is used. We do not apply linear interpolation because temperature data collected below soil/moss is affected by insulation from that layer, and therefore incorporating those measurements into our T_{SG} estimate could confound snow depth predictions.

To clarify this, we provided more detail in the revised manuscript:

P2-L[29]: ***"We estimated T_{SG} from the temperature sensor closest to the snow-ground interface, which ranged from 1 to 5 cm above the ground surface and thus avoided impacts of soil or moss on the T_{SG} estimate."***

Line 76-77: Does "shallow subsurface" refer to the 1 - 5 cm temperature measurements from the previous sentence? Consider rephrasing these two sentences for clarity.

Yes, it does. However, we agree that how we phrased it was unclear. Thank you for catching this. We have rephrased these sentences in the revised manuscript.

P2-L[31]: “***Additionally, we extracted shallow subsurface temperature measurements recorded 1 to 5 cm below the ground surface from soil DTPs deployed into the ground.”***

P3-L[29]: “***We used the same hyperparameters and features as RF-Seward, but calculated features from DTP subsurface temperature measurements recorded 1 to 5 cm below the ground surface.”***

Line 84: Consider adding vegetation type for all sites to table S3 and refer to this after the statement “Vegetation also varied across sites”. Vegetation for 2 sites is given in the following text but this info isn’t currently in the table, whereas vegetation for other sites is included in the sensor details column.

We have added vegetation type to the table (Table C1 in revision) for all sites where sensors were buried beneath the ground surface. We did this because vegetation type only affects snow ground interface temperature measurements when the sensor is placed beneath that vegetation. We then referred to this table after the line “Vegetation also varied” as you suggested. We also added a description of vegetation at the Teller27 and Kougarok64 sites where the machine learning models were trained:

P2-L[19]: ***“Vegetation at Teller27 consisted of mixed sedge-willow-Dryas tundra and mixed shrub-sedge tussock tundra-bog, with some areas of tall willow shrubs (Bennett et al., 2022). Vegetation at Kougarok64 consisted of tussock-lichen tundra, alder savanna, tall willow shrubs in willow-birch tundra, tall alder shrubs in alder shrublands, and rocky areas with birch-ericaceous-lichen and sparse Dryas-lichen dwarf shrub tundra (Bennett et al., 2022; Breen et al., 2020).”***

Lines 92-94: I am confused as to how you trained RF-Deep when you are unable to derive depth estimates for snowpacks deeper than the 1.77m length of the temperature probes. Please clarify what data was used to train the deep model.

We have rewritten our description of RF-Deep. RF-Deep was trained using some of the original DTP training data collected on the Seward Peninsula combined with data available at two sites in Senator Beck Basin, Colorado. This data was not collected using DTPs, but rather collocated snow sonic sensors and temperature sensors at an automated weather station. We only used “some” of the original DTP training data because we wanted to balance the 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***

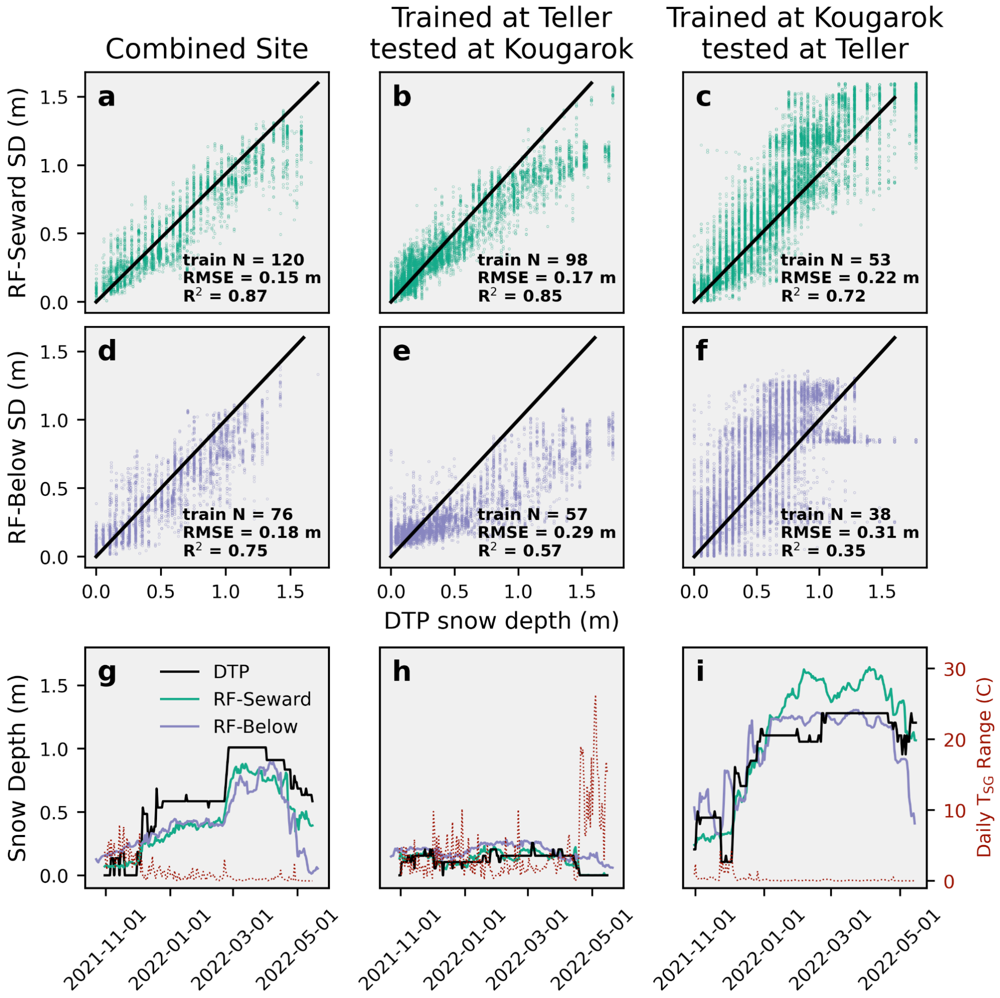
greater than 1.77 m. To test if ML could accurately predict deeper snow depths, we trained a third ML model, which we refer to as "RF-Deep". To train this model, we supplemented our original Seward Peninsula training dataset with additional data from two model evaluation sites in Senator Beck Basin, 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)."

Line 95: I would role this section into the previous one.

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

Line 115/Figure 1: My initial thought was that "temperature range" referred to the range of temperature measured along the whole depth of the DTP, not just as the snow:ground interface. Consider changing this to "daily T_{SG} range" in both line 115 and the red y-axis for plots g - i. Additionally, the use of blue and green to distinguish between the two different models is not accessible to those with colour vision deficiencies. Please change one of these colours – something like blue and orange or green and purple would work.

Thank you for these edits for Figure 1. We have made the suggested changes and updated the color scheme. See below:



"Figure 1. Performance of RF-Seward a) evaluated using test data, b) when trained at Teller27 and tested at Kougarok64, and c) visa versa. d-f) Same as a-c but for RF-Below. Time series plots of DTP snow depth data vs. ML estimates when g) trained at both sites, h) trained at Teller2, and i) trained at Kougarok64. The dotted red line shows daily T_{SG} range, with narrower temperature ranges occurring under deeper snow cover. "Train N" refers to the number of DTP sensors used to train each model."

Line 138: Could this poor performance for ephemeral snowpacks be improved by including more ephemeral snowpacks in the training dataset?

This is a good question. Because zero-curtains completely decrease temperature variability, they mask the impact of snow depth on snow-ground interface temperature and remove the predictive value of temperature data during that period. Therefore, we expect that even with a larger/ more representative dataset, the model would perform poorly during these periods. However, incorporating new features related to zero-curtain periods could potentially reduce these errors.

To address this question in the manuscript, we add the following lines:

P6-L[21]: ***“ZCPs completely dampen T_{SG} variability and therefore uncouple T_{SG} from snow depth. Even given training data more representative of ZCPs, snow depth estimates may remain unreliable during these periods. Incorporating features into the model which indicate the presence of ZCPs may reduce these errors. Further, deploying iButtons at the snow-ground interface (rather than below ground) decreases the number of ZCPs in the temperature data.”***

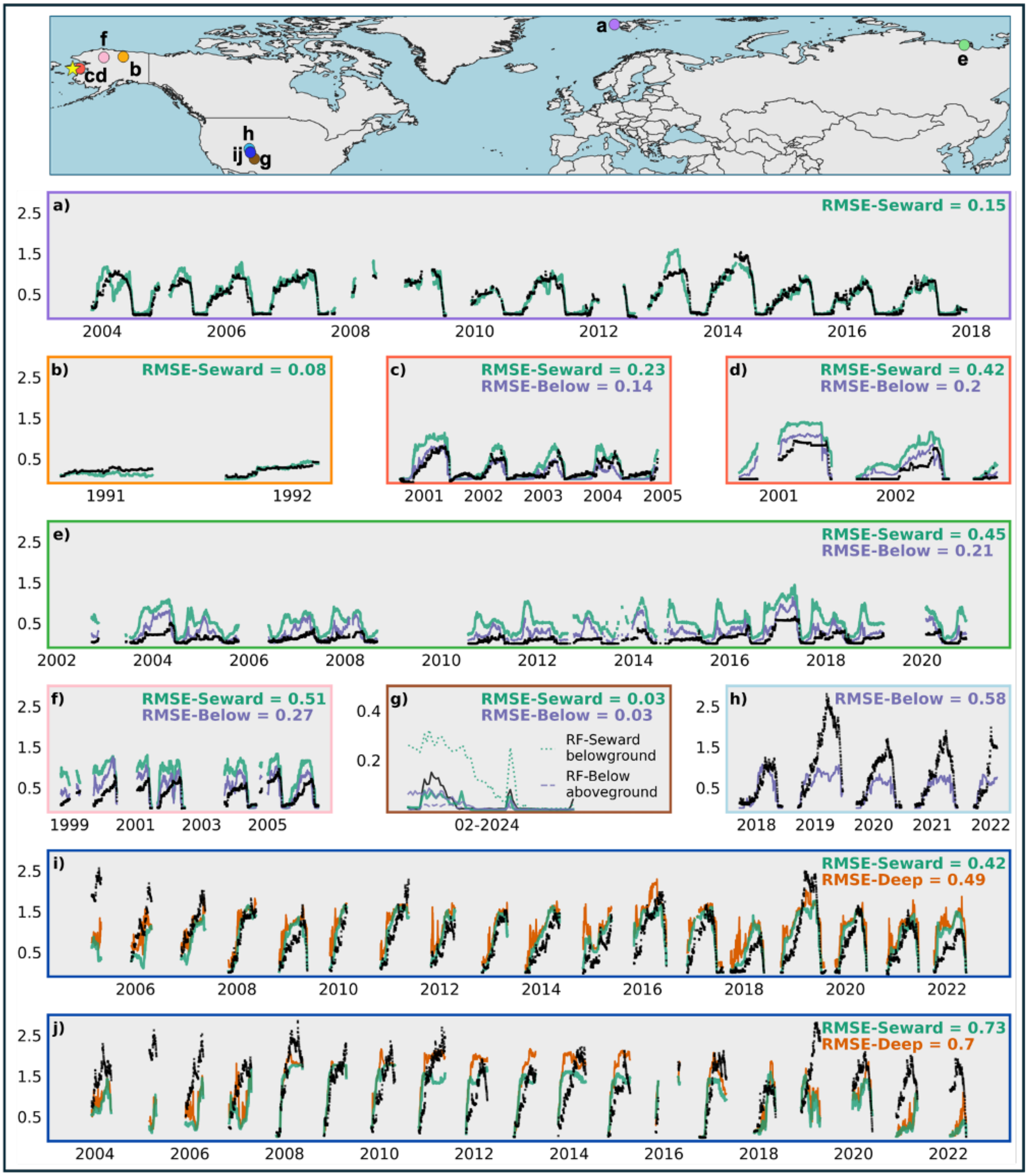
We also reorganized the paragraph to better highlight this message.

Line 149: The insulative capacity of some snowpacks has been shown to be reached at much shallower depths than 1 m (e.g., Slater et al, 2017), particularly in Arctic environments like where the original model was trained. Potentially reconsider the use of this 1m value.

Thank you for sharing this reference with us, after reviewing it, we realized that the 1 m value was too high, as you suggest. Therefore, we have changed this value to 50 cm in the text.

Line 156/Figure 2: The figure caption refers to a colour bar for subplot f), when I think you mean the y-axis for subplot g). Please double check. Some units on the y axes are also needed. Please also clarify what the black lines refer to – measured snow depth? Also, as for the previous figure, the use of blue and green to distinguish between the two different models is not accessible to those with colour vision deficiencies. Please change one of these colours.

Thank you for these suggestions, we have updated the figure and figure caption. See below:



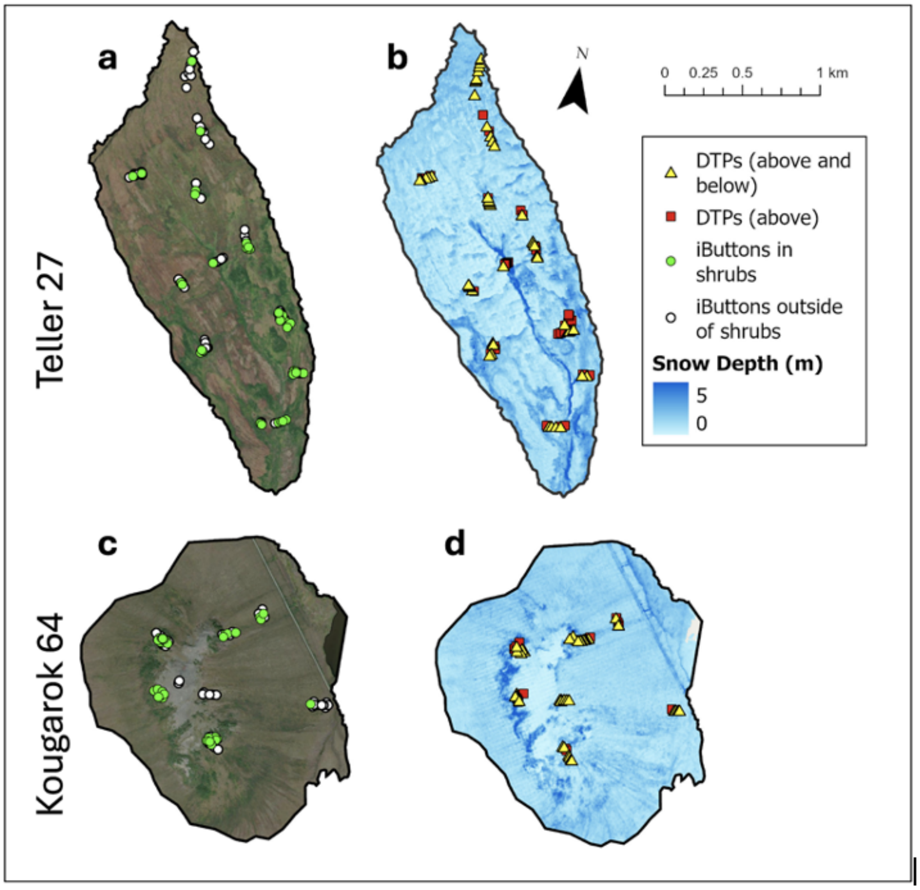
“Figure 2. ML performance at a) Bayelva Station, Svalbard, Norway; b) Imnaviat Creek, Alaska, USA; c,d) Council, Alaska, USA; e) Samoylov Island, Siberia, Russia; f) Ivotuk, Alaska, USA; g) Los Alamos, New Mexico, USA; h) Grand Mesa, Colorado, USA; and i,j) Senator Beck Basin, Colorado, USA. Locations 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 below ground (dotted lines).”

Figure S1: Please clarify what is meant by WY2023 and WY2022 in the figure caption. Can you also confirm that snow depth data shown in b) and d) is for a different year to the temperature data on which the snow depth model is based. Also see comments for Section 2.1 above.

By WY2023 we meant the 2023 water year (October 1 2022 - September 30 2023). We have clarified this in the figure caption by saying the “2022 - 2023 snow season” instead of WY2023 and the “2021 - 2022 snow season” instead of WY2022.

The background snow depth imagery shown in b) and d) was collected in the same year (April 2022) as the DTP data on which the model was trained. We hope that by clarifying the date ranges of data collection we have resolved your uncertainty around this. The figure and updated figure caption are shown 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).”

References:

Slater, A.G., Lawrence, D.M. and Koven, C.D. (2017) ‘Process-level model evaluation: a snow and heat transfer metric’, *The Cryosphere*, 11 (2), 89–996. https://doi.org/10.5194/tc-11-989-2017.

Citation: https://doi.org/10.5194/egusphere-2024-2249-RC1