Point-by-point replies to the question and comments by Reviewer 1

Dear Reviewer 1,

we are pleased to submit the replies to your questions and are thankful for the insightful comments and many good suggestions, as well as we are grateful for your time and effort in providing valuable feedback. We believe that addressing the issues raised by you, have now substantially improved our manuscript.

We hope our answers meet your approval. Your comments and our point-by-point responses are presented below. Please note, that we added a detailed description of a new RF modelling approach in the appendix A.

Reviewer #1 comments	**Action**	**Response**
1. The paper "Comparing High-Resolution Snow Mapping Approaches in Palsa Mires: UAS Lidar vs Machine Learning" by A. Störmer et al. aims to quantify the accuracy and efficiency of mapping snow depth over three palsas in northern Finland, in a spatially continuous raster-based map. Specifically, they choose two methods to compare: 1) using a Lidar sensor on a drone with two acquisition dates of data (no snow and snow), and 2) modelling snow depth based solely on a digital elevation model and using the machine learning algorithm "Random Forest". In situ data of snow depth are collected and used for training and validation. It is an interesting idea, and the need of mapping snow-depths over permafrost features is of great interest. It is also hard work, as noted by the authors in the Discussion, and the contribution of this paper will be of use for those wishing to map snow cover over terrain that has large variations over short distance, such as palsas. The conclusion was that the Random Forest model gave superior results as compared to the UAV Lidar. However, I have some major questions about the process and conclusions that must be addressed, as I question the overly optimistic result presented from the Random Forest model. The two larger issues to be addressed are below, followed by general and specific comments.	Answered	We appreciate your recognition of the relevance of our study and acknowledge the concerns raised regarding the optimism of the Random Forest model results. In response to your suggestions, we have conducted an additional model run incorporating the recommended adjustments. Specifically, we have removed vegetation from the initial DSM and applied hyperparameter tuning as well as cross-validation to the Random Forest model. The updated results, which provide a more robust evaluation, are presented in Appendix A.
Larger issues that need to be addressed: 2. Why was a Digital Surface Model and not a Digital Terrain Model used for calculating the ground in the no-snow data, and how does	Changed/ Answered	We have provided detailed answers to the two larger issues in Appendix A.

this affect the snow-depth measurements, and even the topographic derivatives used in the RF Model? 3. If the authors used cross-validation and present it as the accuracy of the model, then this result is over-optimistic and the comparison of UAV-Lidar to the Random Forest model result is biased and not a fair comparison to make.		
4. 1 - Use of a DSM to represent ground level - It appears that the authors have made a Digital Surface Model (DSM) from the Lidar point data to represent the ground, rather than create a Digital Terrain Model (DTM) from the Lidar data. The DSM represents the height of all objects on the surface, and if there are shrubs on the palsas (which is typically the case in degraded palsas), they may be 35-50 cm tall. Therefore if a DSM was used to represent the ground in August, while insitu snow-depth measurements were taken from the ground up, the reported snow-depth will be highly affected by the height of the vegetation, and this will then vary over the whole surface of the palsa. If the authors have a reason for using a DSM rather than DTM, it is not clear in the article, and it needs to be motivated. Using a DSM will result in error in the snow depth measurements as presented. To create a DTM from your existing data is not difficult. If you look at the paper by Jacobs et al., 2021, you will see reference to papers that discuss the potential errors of snow depth measurements when DSMs are used. In addition if the DSM was used to calculate the Topographic derivatives used as input parameters to the RF model, are these derivatives valid?	Changed/ Answered	See Appendix A.
5. 2- Cross validation - As I understand what has been done, the results of snow-depth for UAV Lidar and RF Modelling have been evaluated differently. In the case of UAV Lidar, the in situ data act as a fully independent data set used for calculating RMSE and the accuracy of the snow-depth measurements. In the case of the RF Modelling, the in situ data are used for training of the model, and the validation of the model as presented (see Fig 8) seems to have been made using a 10-fold cross-validation. In any case, the latter means that the data used to create the model are also used to evaluate the model. Cross-validation is never an assessment of the resulting map accuracy but is an assessment of the fit of the model. So it is no surprise that the authors get seemingly much better results for the RF	Changed/ Answered	See Appendix A.

Model – the comparison is biased in the favor of the RF Model. Figure 8 shows this clearly, and to me is misleading. So the conclusion, as in the Results on Line 367/368, that the RF Model is showing its strength without high bias, I think is not valid. The only way to fairly compare the assessments of these two would be to develop a model using in situ data from one palsa and apply the RF model developed to the other two palsas and assess the accuracy using the in situ data from those two palsas. Or, you could take insitu data from half of each palsa and developing training and accuracy datasets. (Note that if you consider taking a random selection of the insitu data for training/accuracy it is not optimal, since you will have spatial autocorrelation issues due to the proximity of the points, which is why the previous suggestions are better.)		
Other general		
6. The title: Rather than using the term "Machine Learning", I think it would be better to refer to this as "Modelling", because it doesn't make sense to me to compare it to the specific algorithm that is used, but rather that you have created a model to predict snow depth.	Changed	We agree with your suggestion to change the specific term "Machine Learning" to "Modelling".
7. There have been scientific articles that have mapped snow with UAV Lidar, eg, Jacobs, J.M. et al., 2021 "Snow depth mapping with unpiloted aerial system lidar observations: a case study in Durham, New Hampshire, United States" in The Cryosphere. (https://doi.org/10.5194/tc-15-1485-2021). While this may be the first paper to be published using UAV Lidar for snow on a palsa, I think that the Introduction should review and refer to articles that have generally applied UAV Lidar mapping of snow over other landscape types.	Changed	We acknowledge the importance of previous studies that have applied UAV LiDAR for snow depth mapping and appreciate your suggestion. Our initial focus was primarily on demonstrating the feasibility of snow depth modelling using Random Forest and assessing its implications for palsas. However, as the study evolved, the focus shifted more towards a comparative analysis between Random Forest modelling and LiDAR-based snow depth estimation. In response to your suggestion, we have incorporated additional references on UAV LiDAR-based snow depth mapping. Specifically, we adopted the vegetation removal approach inspired by Jacobs et al. (2021) and further reviewed relevant studies, including those by Avanzi et al. (2020) and Harder et al. (2020).
8. Section 2.1 is lacking a description of vegetation heights on the palsas.	Changed	We have added a description of the typical heights of common palsa vegetation. *Dwarf birches* are 15 and 60 cm (Betula nana) and *dwarf shrubs* 5 – 20 cm, while *sphagnum*

		moss (up to 3 cm) and *lichens* (<3 cm) are considerably smaller. These height estimates are based on our field observations.
The following points all refer to Section 3.1 – Data collection		
9. Did you Post-Process the UAV Lidar data with RINEX data from a base station? If so, what was the base station (ie, source of the RINEX data)?	Changed/ Answered	We used RINEX data for post-processing the LiDAR flight trajectory in POSPac. For both datasets, we obtained the RINEX data from the National Land Survey of Finland (NLS) CORS station in Kilpisjärvi (*KILP 2147250.4266 820562.0462 5930136.8831*). Further details can be found on the NLS website. This information has been incorporated into the newly added paragraph 3.1.1 (see comment #17).
10. Parameters for the UAV flights are needed, eg, flying altitude, were cross-wise flights used? Knowing the directions of the flight lines is important because there are some Lidar measurements of 0 cm snow depth, and 50-60 cm snow depth in the insitu data, and it might be explained (possibly?) by not acquiring Lidar data in multiple angles – but I am not sure what has been done.	Changed/ Answered	We have included additional details about the LiDAR flights in the newly added paragraph 3.1.1 (see comment #17). The flight altitude during the summer data collection was 30 m for each palsa, while in winter, the flight altitude was 60 m. All flights had a 50% side overlap. Cross-flights were not conducted, as we determined that they would not provide additional valuable data due to the specific environmental conditions (flat terrain with low vegetation). The flight direction was primarily along the longitudinal axis of the palsas, except for the summer flight over Peera palsa, which followed an east-west orientation. To improve clarity, we have included the flight trajectories in Figure 4 (see Appendix C). Additionally, we have provided point density values for each flight in m^2. The RGB flights were conducted using an Autel EVO II Pro V2 UAV at a flight altitude of 80 m, with a 75% side overlap for each flight. Initially, we had stated that the orthophotos were acquired with the integrated RGB sensor of the LiDAR mapper. However, this was a misunderstanding, and we have now corrected this statement. The orthophotos do not contribute specific data to the analysis but were solely used for figure creation. Regarding the occurrences of 0 cm snow depth in the UAS-LiDAR data: These values

		result from our initial processing step, in which we set all negative values in the computed DSM to zero. This aspect is now explicitly mentioned in Appendix A, and we have adjusted this approach in the updated modelling.
11. Line 151/152 says that GCPs were set out. Was this for both the Lidar and the RGB images? How many GCPs? And then, what was the horizontal and vertical accuracy of your data – both the Lidar and the RGB images?	Changed/ Answered	For all winter flights (LiDAR and RGB) and palsa sites, four GCPs were used, positioned around the palsa. The accuracy for each GCP is between 1–2 cm. We have established several permanent GCPs (measured with RTK-GPS) located on known points of large stones in the study sites. Permanent GCPs have been established because we are monitoring changes in the palsas by collecting drone data annually since past 8 years. The accuracy of these RTK-GPS-measured GCPs is between 1–2 cm. For all UAS-LiDAR summer flights we utilized these GCP's: three for Peera, 20 for Pousu, and 30 for Puolikkoniva. The RGB flights were conducted using the drone's internal RTK system. However, we consider the accuracy of these flights to be not relevant in the context of this study, as the orthomosaics are used solely for overview purposes. We have added this information to section 3.1.1.
12. Line 153 – Change orthopictures to images, since the raw images are not orthorectified yet. That's a later step.	Changed	We agree and changed the term.
13. Line 157/158 "Structure from Motion techniques were not applied…" I do not understand why this sentence is here. If you created an orthophoto, which you say you do in the next sentence, then you have applied photogrammetric image matching (how you define SfM and if you define it differently than photogrammetric image matching determines what term you like to use). But why even say what you haven't done? State what you have done to produce the orthophoto.	Changed	We acknowledge the potential for misunderstanding and have clarified our statement in lines 157/158. Specifically, we used SfM techniques solely for the creation of orthophotos. As noted in comment #10, there was an initial misunderstanding regarding the generation of the orthophotos. We have now revised this section to ensure clarity and accuracy.
14. Line 164 – I think you mean snow depth rather than snow cover.	Changed	We agree and changed the term.
15. Line 166 – RTK-GPS.	Changed	We agree and added "GPS".
16. It says on line 173 that there are randomized points on the edges of Puolikkoniva, but I do not see very many of these (maybe 5 at	Changed/	In Puolikkoniva, 20 randomized points are located at the edges, which corresponds to one fifth of the total point dataset for this

most?). In hindsight, I would guess that you would want to have made cross-wise transects on this palsa. Take this up in the Discussion if so.	Answered	palsa. We consider this distribution to be adequate. To provide a clearer visualization of the point distribution, we have added an additional figure to the appendix, categorizing the points by classes. In this figure, we have also included an orthophoto without snow cover to enhance the visibility of point distribution in the most extreme areas. Furthermore, we have adjusted the point style to improve differentiation and recognition (Appendix E).
Reference (in situ) data		
17. I think you need a separate section to describe Reference data collection – either two sub-sections under 3.1 or else 3.1 for UAS data collection and 3.2 for Reference data collection. Under the reference data collection, there should be a better description regarding how the insitu snow depth measurements were made, specifically, was the GPS Z-measurement made from the ground level? Was it a yardstick, and was a level used to make sure it was normal to the surface?	Changed/ Answered	We have revised the manuscript to create a clearer structure for data collection. In particular, we have introduced two separate subsections: 3.1.1 for data collection at universities of applied sciences and 3.1.2 for the collection of reference data. In response to your suggestion, we have expanded the description of the reference data collection. The snow depth measurements were carried out with a heavy wooden yardstick. GPS-Z measurements were not taken from the ground, as this approach is subject to large uncertainties. In addition, no level was used to ensure that the measurements were perpendicular to the surface. These methodological details were explicitly mentioned in the revised section to improve transparency and reproducibility.
18. For the insitu data you need at some point to say that these also may have errors and what these errors may be caused by, and how they may affect your result. Since the RF model is completely based on the insitu data, the errors of the insitu data are simply propagated, but do not affect the evaluation. For validating the Lidar data derived snow depths, the potential measurement errors of the insitu data are only accounted for in the evaluation.	Changed/ Answered	We have expanded our discussion on the uncertainties associated with snow depth measurements in lines 409–411. In this revision, we have explicitly acknowledged that measurement errors not only influence the modelling results but also affect their statistical evaluation. We have critically assessed the extent of this impact. However, we still consider that manual snow depth measurement is the most accurate method available, with only minor deviations. Given the overall reliability of this approach, we do not expect these uncertainties to significantly affect the conclusions of the study.

19. Also, think about whether the section on UAS data collection is only about data collection or if you want to describe the processing of the data here – in which case you might just name it "UAS data" or "UAS data collection and processing".	Changed	We agree with your suggestion and have renamed section 3.1.1 to "UAS Data Collection and Processing" to better reflect its content. As noted in previous comments, we have also expanded this section to include additional details on both the data collection and processing procedures.
20. The in situ data particularly in the case of the largest palsa Puolikkoniva were run in two transects lengthwise along the palsa, but not crosswise, over the edges where the deepest accumulation of snow may have been. Therefore the values where some of the largest differences are between the Lidar and the RF Model cannot really be assessed, making the assessment incomplete – the shortcoming must be acknowledged.	Answered	We respectfully disagree with this point. As discussed in comment #28, we believe that our dataset is well-suited for the objectives of this study. Our measurement strategy included data points in key areas of the palsa, such as the summit, steep slopes, and internal trenches, ensuring that the primary variations in snow depth were captured. However, we acknowledge that extreme values may still have been missed due to the inherent limitations of in situ snow depth measurements. We have addressed this explicitly in the discussion part about limitations.
21. Also the Lidar may measure extremes in snow-depths, while the model will not if it does not have representative data for the extremes. Therefore there will be more variability in the Lidar data, but we cannot tell which is "wrong".	Answered	As mentioned in response to the previous comment, we believe that our dataset adequately captures most of the extreme snow depth variations. However, we acknowledge that some extreme values may not have been fully represented in the model due to limitations in the in-situ data collection. Therefore, we address this issue in the limitations section and emphasize the differences in variability between LiDAR measurements and the RF model.
Section 3.2 – RF algorithm		
22. The authors state on Line 189 that no explicit hyperparameters were specified. So this means that they were not analyzed, although the outcome of the model is what is being assessed as the main objective of the article. It is not difficult to assess the hyperparameters using Grid-Search or another comparable function.	Changed/ Answered	See Appendix A.
23. Permutation mode was used for variable importance – do you know how this works? Is it a single run of the RF model? When you run PI repeatedly, do the same variables have the same importance? The random nature of RF often requires running variable importance (or in this case PI) many times (eg, 100) and taking an average. Even then, one needs to be	Changed/ Answered	See Appendix A.

careful with their interpretation of variable importance.		
24. For Line 187-188 - I'm not really sure what you have done with the model and the in situ data. You state that you have split 70% training and 30% test. Is this used by RF for internal cross-validation of the model (if you split the data 70/30 in the RF model, then it is likely this is how it is being used). Is this done with replacement? If you have removed 30% of the data for independent evaluation, then you need to clearly state this, but I don't think this is what you have done.	Changed/ Answered	We separated the dataset into 70% training data, which was used for the Random Forest model, and 30% test data. The test dataset was not utilized for internal cross-validation but was exclusively used to compute performance metrics, including RMSE, R^2, MAE, and standard deviation. Further details on this methodology are provided in Appendix A.
25. Line 184 – The dependent variable for your model is snow-depth.	Changed	We agree, and to clarify this, we have removed the term "dataset" in line 184.
26. Line 185 –"Input parameters" are mentioned here but we don't know what they are until later. Couldn't you refer to Table 2 here? Otherwise we are left wondering what the parameters are.	Changed	We agree and added Table 2 as reference.
27. Line 189 – delete "precise" – This is a judgmental word – leave it to your results to be the judge of that.	Changed	We agree and removed the word *precise*.
28. In addition, RF models are sensitive to imbalance in the training data, and also do not extrapolate beyond the minimum and maximum snow-depth values (or whatever the target variable may be). How are your results affected by this, and how might others in the future be affected by this and what would your recommendations be to future applications of this method?	Answered	We acknowledge the sensitivity of Random Forest models to imbalances in the training data, as well as their inability to extrapolate beyond the observed minimum and maximum values of the target variable. To ensure the robustness of the model, it is crucial to capture a dataset that adequately represents the full range of snow depths within the study area. A thorough understanding of the investigation area is essential, and snow depth measurements should specifically target extreme locations, such as exposed areas on the palsa summit and accumulation zones along the edges. However, conducting such measurements in these environmental conditions is both time-consuming and labour-intensive. For future applications of this method, we recommend careful planning to ensure representative data collection. This includes identifying extreme locations in advance by analysing orthophotos and previously acquired digital elevation models (DEMs) to optimize the placement of measurement points. We have added this to our discussion.

Section 3.3 –		
29. The first sentence needs rewriting. First of all, which "collected airborne data" is referred to here? I assume it was the August DSM from Lidar that was used? It is not stated. Were these data processed differently than what was described in Section 3.1? Declare which DEM you are working with and say specifically that you are creating parameters from this. What happens if you use a DSM and create all of these topographic derivatives as parameters? Are those new derivatives valid, such as Topographic Wetness Index, if they are based on the surface elevation which includes vegetation? This must be well-motivated if the authors believe that there is a valid reason for this.	Changed/ Answered	We acknowledge the need for greater clarity in the first sentence and have specified that we are referring to the summer dataset. To improve transparency, we have moved this information to the newly added section on UAS-LiDAR processing (Section 3.1.1). Regarding the influence of vegetation on the derived topographic parameters, we have addressed this aspect in previous comments and provided further details in Appendix A.
30. Line 210 – If a 0.3 m buffer was used were the values for any parameters averaged within this area?	Answered	The input parameters were not averaged within the 0.3 m buffer areas. Instead, each input parameter value was directly linked to the corresponding in-situ snow depth ($SD_{in\text{-}situ}$) measurement. Consequently, each snow depth value is associated with an average of 28 input parameter values. For Puolikkoniva, 100 snow depth measurements were linked to a total of 2,819 input parameter observations, while for Peera and Pousu, the corresponding values were 39 to 1,097 and 46 to 1,306, respectively. In total, 5,222 points were generated for the Random Forest model, with 70% used for training and 30% for validation. This approach was chosen to reduce noise and smooth the resulting dataset. You can find further information in Appendix A.
31. Table 2 – 12 parameters were used, but 21 are in the table. Could you indicate in a way what parameters were used?	Changed	We have described the input parameters used for each model run in lines 205–209. However, based on your suggestions and the revised model design, we have decided to focus on a single model run. As a result, we used only 12 parameters and removed the information related to previously unused parameters and former model runs. The section on input parameters and model runs has been updated accordingly.
32. For the Discussion: When you made the insitu measurements, it was August, and the palsa	Changed/	We have already incorporated the findings of Renette et al. (2024) into our discussion (lines

had likely subsided. Renette et al., 2024 show that the difference between elevation in September (likely maximum thaw depth of the Active Layer) and April (minimum thaw) was on average 15 cm, and up to 30 cm in some areas, albeit on a taller palsa than in the study presented here. In any case, this may mean that trying to measure snow depth using a DTM from September may introduce errors if the terrain is actually elevated some cm more than this. This is hard issue to solve with UAV Lidar, since you would need to be in place to create a DTM right after snow-melt, and all snow would need to have melted. So, you need to discuss what implications this has to your results. Also, since you have RTK-GPS data, and you have measured to the ground I assume, you actually have a dataset where you could compare the Z-measurement from March to the DTM from August, and get an estimate of the difference in height between the max-thaw and min-thaw state of the palsa.	Answered	384–387). However, based on your comments, we have expanded this discussion to explicitly address the potential impact of seasonal elevation changes on the accuracy of SD_{LiDAR} measurements. In contrast, the SD_{RF} results should be less affected, as the modelled snow depth is independent of seasonal elevation fluctuations. Furthermore, we have considered this aspect when refining our overall evaluation of the comparison between RF and LiDAR results. As you pointed out, accurately capturing snow depth using LiDAR is only possible if data collection occurs immediately after snowmelt, once all snow has disappeared. Regarding the RTK-GPS measurements, we did not measure directly to the ground. During fieldwork, we observed that the thick and frozen vegetation layer made it challenging to reach the true ground surface using the RTK stick. Instead, we found that the fine yardstick provided a more accurate way to measure snow depth. Consequently, we are unable to compare Z-measurements within our datasets. However, we have incorporated this consideration into our discussion of future research implications (lines 386–387).
Language		
33. It's my feeling that some value judgement words don't belong in a scientific article. Such as "exemplarily" on line 53.	Changed	We reviewed the manuscript for judgemental words and changed these accordingly.
34. Line 38 – deepening instead of growth. Line 58 – deeper instead of higher.	Changed	We changed these words.
35. Otherwise some minor grammatical fixes once the paper is revised can be looked over.	Changed	We will have a final grammar check after implementing all changes to the manuscript.
Specific		
36. Line 35 – it is not only bound by peatland presence but also climatic parameters	Changed	We agree and added "and driven by climatic parameters".
37. Line 69 – "Satellite data" only names the platform. What kind of satellite data are you referring to? Optical? Radar? That is the more important aspect. Similar issue is on line 74 where the sensor type should be mentioned and not just the platform which is UAS/UAV. Look through your paper for these kind of omissions.	Changed	We acknowledge the need for greater specificity regarding the types of satellite data referenced. In the respective sections, we now explicitly state that we are referring to both optical and radar satellite data.

		Additionally, we have specified the type of UAS sensor used in each mentioned study to ensure clarity. These adjustments have been implemented in lines 9, 14, 49, 71, 74, 75, and 95.
38. Line 70 – change technical limitations to properties	Changed	We agree and changed the term.
39. Line 86 – the authors mention 3 methods, but the title takes up two. The third method seems to be the insitu data, but that has been used to train the RF Model, and I don't think you are really assessing the accuracy of the method, so I would stick to the two methods.	Changed	Thank you for highlighting this inconsistency. We have revised our focus to explicitly center on the two primary methods - LiDAR-based snow depth estimation and RF modelling. Accordingly, we have adjusted the structure of our objectives and intentions.
40. Line 89 – delete simulation. You are just modelling.	Changed	We agree and deleted "simulation".
41. Table 1 – the photos are rather small. Can they be made bigger. Put the date (day-month-year) of the photos in the Table text.	Changed	We agree and changed the caption and increased the size of the images.
42. Line 129 – For what year or years is that the annual mean temperature?	Changed	We inserted "For the time period 1991 – 2020, …".
43. Line 137 – For what location is that the duration of permanent snow cover?	Changed	This value is specific to Kilpisjärvi, and we have incorporated it into the text.
44. Figure 2 – What is shown in Fig 2? It needs to be said clearly in the Fig text. Is this an average value for 1990-2020? It would be very helpful to know what the climate conditions were for the years in which you acquired the snow data. Was it a very snowy year? Windy in the days before you visited? Warm temperatures so that the snow melted some? Knowing these conditions can help us to explain any differences between the various results, particularly if the model is solely based on the DEM. I see you mention this on Line 401/402.	Changed/ Answered	Figure 2 presents the average monthly snow depth (cm), temperature (°C), and precipitation (mm) recorded at the Kilpisjärvi weather station for the period 1990–2020. This timeframe was selected to align with the 30-year reference period established by the World Meteorological Organization (WMO). However, we recognized that the appropriate reference period should be 1991–2020 and have updated the figure accordingly (Appendix B). The purpose of this figure is to provide a general overview of the climatic conditions in the study area. Since snow depth data were collected on only two days (March 23–24, 2023) under stable weather conditions, we do not believe that presenting weather data from the preceding days or the entire winter season of 2022/23 would provide additional meaningful insights. To clarify this, we have explicitly stated in line 141 that all snow depth measurements were conducted on March 23–24, 2023.

45. Line 141 – Write which day the data were acquired. If you cannot fit it reasonably in the text, because it was different dates for different palsas, I suggest you put it in Table 1 – dates for image and Lidar acquisition.	Changed	We have added the specific dates of data acquisition in line 141. The UAS-LiDAR data were collected on August 27, 2022 (summer) and March 23, 2023 (winter). Snow depth measurements were conducted on March 23, 2023, for Puolikkoniva and Pousu, and on March 24, 2023, for Peera. Additionally, we have included the exact dates of LiDAR data acquisition in Table 1 to ensure clarity and consistency.
46. Several of the Figures have such small text that they are difficult to read. Eg Fig 3.	Changed	We have increased the font size for Figures 3, 6, 7 and 8.
47. Section 3 – Is August the season for maximum thaw? It's not September? Does Verdonen et al. 2023 state that August is the max ALT? If it is August, I think you should more specifically say the end of August. If you aren't sure or don't have a reference to back it up, then maybe it is more reasonable to say that the end of August is near max ALT.	Changed	We agree that this statement requires greater accuracy. We have revised it to indicate that the maximum ALT is typically reached between the end of August and mid-September, depending on annual weather conditions and the onset of the freezing season.
48. Line 231 – 240 feel like they belong in the section describing the RF model.	Changed	We agree and moved this part to the description of the RF algorithm and modelling data preparation.
49. Line 231/232 – Was the 10-fold cross-validation done when creating the initial RF model, or was this something that was done afterwards and used as the "validation" data presented in Figure 8? If it is the latter, you cannot say that it was used to reduce over-fitting in the model? There is an option in Random Forest to use cross-validation to create the model, and that is one tool of several to reduce over-fitting. Other ways to reduce over-fitting is to limit tree depth, -- by the way, in Section 3.2 you mention target node depth, but I don't see in the caret package what that refers to. Is it "maxdepth"? In that case I suggest you name the parameter in parentheses.	Changed/ Answered	See Appendix A.
50. Line 236/237 – What are "the initially calculated values"? You are using the insitu data to train a RF model and then evaluating the model based on a cross-validation that using that same insitu data. See my point #2 under "Larger issues".	Changed/ Answered	See Appendix A.
51. Line 273/274 – "Only a few narrow structures with significantly higher snow can be recognized based on the UAS LiDAR data" – I do not know what this sentence is about.	Changed	Our intention was to highlight that only small areas within the study region exhibit significantly higher snow depths in the UAS-LiDAR dataset. To clarify this, we have revised the sentence as follows:

		"On the other hand, only small areas with significantly higher snow depth in the UAS-LiDAR dataset compared to the RF dataset are detectable in certain regions surrounding the palsas."
52. Line 281 and Fig 7 and Table 3 – I don't think we need to see all 3 model runs, just the best one.	Changed	We agree and changed the text, figure and table accordingly. See comment #31 and Appendix A.
53. Line 285 – rather confusing that it is stated that Elevation was removed, and now it is important. Also Fig 7 text is impossible to read because it is so small.	Changed/ Answered	Please refer to our previous responses regarding the removal of the initial model runs. As a result, it is no longer necessary to elaborate on the exclusion and reintroduction of the *Elevation* parameter. For clarity regarding our initial approach: *Elevation* was excluded in the second model run because all other input parameters were derived from it. This step was taken to assess whether *Elevation* might introduce bias into the modelling results. After analysing the outcomes, we found no indication of such bias and subsequently decided to retain *Elevation* as an input parameter in the final model.
54. Line 295 and Table 4 – these areas of "Top", etc, could you have a figure somewhere – maybe supplemental where these areas are shown? Do we know the number of samples (n) in each group?	Changed/ Answered	See Appendix A and E.
55. Line 323 also Line 346 – Fig 9?	Changed	Thank you for the note, we have changed that.
56. Figure 9 – Is B (Slope in degrees) based on the DSM? Is this valid then to calculated slope based on vegetation?	Changed	See Appendix A.
57. Line 404/405 – I guess you are referring to reflectance of the lidar from the snow/ice surface? If so I think you should have a reference here.	Changed	We agree and have added a reference to Deems et al. (2013), which investigates the influence of reflectance and scattering by snow and ice surfaces on the accuracy of LiDAR sensors.

Appendix

In this section, we provide additional information addressing comments #4, 5, 10, 16, 22, 23, 24, 44, 49, 50, 54, 56 from Reviewer 1 and #5, 6, 8, 10, 38, 40, 41, 52, 53, 63, 66 from Reviewer 2.

We sincerely appreciate your insightful comments and suggestions, which have significantly contributed to improving both the modelling approach and the overall quality of the manuscript.

Appendix A

To ensure high-quality modelling results and accurate snow depth distribution maps derived from UAS-LiDAR, we implemented your recommendations, including the removal of vegetation from the LiDAR-derived products and a re-evaluation of the modelling approach.

Additionally, we incorporated hyperparameter tuning and cross-validation to determine the most suitable parameter settings for the Random Forest model. To further improve model robustness and prevent overfitting, we also adjusted the data splitting strategy by testing the RF model on an independent external dataset.

1. Removal of vegetation from UAS-LiDAR DSM

Our initial decision to retain vegetation in the modelling process assumed that small and dense vegetation, as present in our study sites, is difficult to remove - even from point clouds. Testing several vegetation filter algorithms, such as the *Cloth Simulation Filter* (CSF) and *Statistical Outlier Removal* (SOR) in CloudCompare, confirmed this assumption, as the vegetation was not properly removed in the resulting products.

Additionally, we considered that vegetation significantly influences snow depth distribution by enhancing snow retention capacity. Therefore, we initially decided to include vegetation in the modelling process, expecting it to be beneficial for RF modelling.

However, based on your suggestions, we tested the *Progressive Morphological Filter* (PMF) Algorithm as described by Zhang et al. (2003) and Jacobs et al. (2021) and obtained satisfactory results with an effective removal of vegetation. We applied PMF filtering using the following parameters:

- Window sizes: 0.5, 1, 2, and 3
- Thresholds: 0.05, 0.1, 0.3, and 0.5

The extracted ground and vegetation points were saved in point cloud format. Using CloudCompare, we generated a DTM for each palsa using the Rasterize function. Empty cells within the point clouds were interpolated with a triangle max edge length value of 5.0.

The newly created DTMs were then used to recalculate the snow depth distribution for all three test sites in GIS, following the methodology described in the manuscript. In our initial calculations, all negative values were set to zero. However, in this revised approach, we retained negative values to highlight areas where either the LiDAR sensor produced inaccuracies or surface degradation occurred between the summer and winter flights.

Based on these refined DTMs, we recalculated all input parameters used in the final RF model run in SAGA GIS. The following 12 parameters were included: *Aspect, Elevation, Channel Network Base Level, Channel Network Distance, Negative Openness, Positive Openness, Relative Slope Position, Slope, Topographic Position Index, Valley Depth, Wind Effect, Wind Exposition*.
A detailed description of these parameters is provided in Table 2. We have now focused on a single model run, and accordingly, we have removed descriptions of other parameters from the manuscript to ensure clarity and consistency.

2. Splitting data into training and test datasets

In the initial study design, we used the entire buffered $SD_{in\text{-}situ}$ dataset to extract the input parameters from the raster stack, resulting in a data frame with 5222 points. We then split this dataset into 70% training and 30% test data. However, this approach introduced a risk of overfitting, as each $SD_{in\text{-}situ}$ point was represented an average of 28 times in the dataset. Consequently, many points appeared in both the training and test datasets, reducing the independence of the validation process.

To address this issue, we revised our study design by first separating 70% of the point features from each $SD_{in\text{-}situ}$ dataset for training and 30% for testing. Only after this separation did we extract the input parameter values for the training dataset, ensuring a clear distinction between training and validation data. The test dataset was reserved exclusively for model validation. The following extract from the R script illustrates these steps:

```r
#===== Function to split training and test dataset =====
split_shapefile <- function(shp) {
  set.seed(42)  # Ensure reproducibility
  num_samples <- nrow(shp)  # Get the number of samples
  train_indices <- sample(num_samples, size = round(0.7 * num_samples))  # Select 70% of the samples for training
  test_indices <- setdiff(1:num_samples, train_indices)  # The remaining 30% for testing

  shp_train <- shp[train_indices, ]  # Create training dataset
  shp_test <- shp[test_indices, ]  # Create test dataset

  return(list(train = shp_train, test = shp_test))  # Return the split datasets as a list
}

# Splitting the dataset for all three locations
split_pousu <- split_shapefile(shp_pousu)
split_peera <- split_shapefile(shp_peera)
split_puolikkoniva <- split_shapefile(shp_puolikkoniva)

# Combine training and test datasets for all Palsas
shp_train_all <- rbind(split_pousu$train, split_peera$train, split_puolikkoniva$train)  # Merge training datasets
shp_test_all <- rbind(split_pousu$test, split_peera$test, split_puolikkoniva$test)  # Merge test datasets
```

After extracting the input parameters from the raster stack, the final dataset consisted of:

- Training dataset: 3,645 points (Puolikkoniva: 1,983; Pousu: 905; Peera: 757)
- Test dataset: 1,577 points (Puolikkoniva: 836; Pousu: 401; Peera: 340)

To prevent errors and miscalculations, all NoData values were removed from the datasets, resulting in a final training dataset of 3,504 points and a final test dataset of 1,548 points for further modelling and validation.

3. Hyperparameter tuning and cross validation

To determine the optimal values for *mtry*, *min.node.size*, and *sample fraction*, we performed hyperparameter tuning using the *mlr* package in R (Bischl et al., 2016).

To prevent overfitting, we restricted the search range for *min.node.size* to 10–15 and for *sample fraction* to 0.7–0.85, following the recommendations of Probst et al. (2019) and Breiman (2001). Allowing an unlimited search range initially resulted in better model performance, but at the cost of reduced generalization, indicating signs of overfitting. We selected the final search range based on multiple test runs with different settings.

For cross-validation, we tested different fold sizes to identify the most effective configuration. The best results were achieved using a 4-fold cross-validation. The following R script extract provides details on the tuning process:

```r
#===== Hyperparameter Tuning with tuneRanger (Regression) =====

# Define the regression task
task <- makeRegrTask(data = all_train, target = "Class")

# Define the cross-validation strategy
cv_desc <- makeResampleDesc("CV", iters = 4)  # 4-fold cross-validation

# Define the Random Forest learner with hyperparameters as tuning options
learner <- makeLearner("regr.ranger", num.trees = 1000)

# Define the hyperparameter search space
param_set <- makeParamSet(
  makeIntegerParam("mtry", lower = 2, upper = ncol(all_train) - 1),  # Number of variables to consider at each split
  makeIntegerParam("min.node.size", lower = 10, upper = 15),  # Minimum number of observations per node
  makeNumericParam("sample.fraction", lower = 0.7, upper = 0.85)  # Proportion of samples used in each tree
)

# Define the tuning control (e.g., Bayesian optimization or random search)
control <- makeTuneControlRandom(maxit = 70)  # 70 iterations for tuning

# Hyperparameter tuning with cross-validation
tuned_params <- tuneParams(
  learner = learner,
  task = task,
  resampling = cv_desc,  # 4-fold CV
  par.set = param_set,
  control = control,
  measures = rmse  # Root Mean Squared Error as the performance metric
)

# Display results
print(tuned_params)

# Best Random Forest model with tuned parameters
best_learner <- setHyperPars(learner, par.vals = tuned_params$x)
```

The final tuned hyperparameters were as follows:

- mtry: 9
- min.node.size: 10
- sample fraction: 0.79

4. Permutation Importance (PI)

In our initial study design, we conducted the RF modelling once and directly used the permutation importance (PI) values provided by the model.

In our revised approach, we refined this process by repeating the calculation 100 times to obtain a mean PI value for each input parameter, ensuring more robust and reliable importance rankings.

The following R script extract details the implementation of this approach:

```r
#========== Permutation Importance
num_repeats <- 100
importance_values <- matrix(NA, nrow = num_repeats, ncol = ncol(all_train) - 1)

for (i in 1:num_repeats) {
  cat("Iteration:", i, "\n")

  # Train the model using the identical hyperparameters from tuning
  temp_model <- ranger(
    x = all_train[, -ncol(all_train)],
    y = all_train$Class,
    mtry = tuned_params$x$mtry,  # Optimized mtry value
    min.node.size = tuned_params$x$min.node.size,  # Optimized min.node.size
    sample.fraction = tuned_params$x$sample.fraction,  # Optimized sample.fraction
    num.trees = 1000,
    importance = "permutation",
    seed = i  # Different seed per run for robustness
  )

  # Store the feature importances in the matrix
  importance_values[i, ] <- importance(temp_model)
}

# Compute the mean Permutation Importance over the 100 runs
mean_importance <- colMeans(importance_values)
```

We modified Figure 7 to display only the 12 selected parameters along with their respective mean PI values over 100 iterations. Additionally, we normalized the values, setting the most important parameter (Topographic Position Index) to 1.

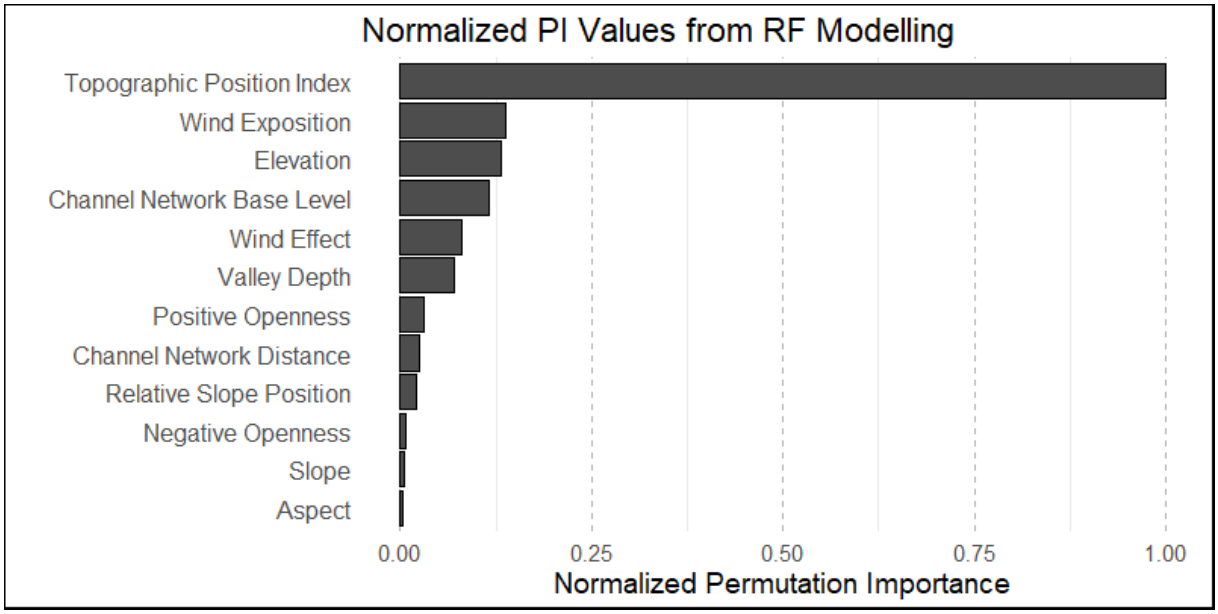

Figure 7. Overview of normalized mean Permutation Importance values from RF modelling over 100 iterations.

5. Final results and validation

Both the RF-based and UAS-LiDAR-based results were validated using the initially separated test dataset. Additionally, we conducted three further RF model runs, where in each iteration, two palsa sites were used as the training dataset, and one was used as the test dataset. This approach further validated the generalization capability of the model.

The validation results indicate that the RF-based approach now exhibits lower peak accuracies compared to the initial study design. However, by reducing overfitting, the results are more plausible and robust, while still achieving high accuracy and outperforming the UAS-LiDAR-based approach:

Table 3. Overview of the calculated Root Mean Square Error (RMSE), Coefficient of Determination (R2), Mean Absolute Error (MAE) and Standard Deviation (SD) for RF- and UAS-LiDAR-based snow depth estimations. Additionally, external validation results (RMSE and R^2) for RF-based snow depth at each palsa site (Peera RF, Pousu RF, Puolikkoniva RF) are provided.

Parameter	*RF*	*LiDAR UAS*	*Peera RF*	*Pousu RF*	*Puolikkoniva RF*
RMSE	18.33	23.49	16.67	21.31	27.13
R^2	0.77	0.691	0.628	0.767	0.578
MAE	13.26	17.49	-	-	-
SD	18.11	20.84	-	-	-

We recalculated all metrics for different point groups and included the number of points per group. These groups were classified visually, based on orthophotos, slope data, and elevation characteristics of the respective locations.

The results show that the accuracy differences between RF and UAS-LiDAR-based approaches are now less pronounced. However, in certain categories, such as *Thermokarst* and *Open Area*, the UAS-LiDAR-based results show lower accuracy, likely due to measurement inaccuracies caused by water surfaces and irregularities in areas with higher vegetation.

Table 4. Overview of RMSE, R^2, MAE and SD divided by validation point locations within the investigation areas.

	RMSE		**R^2**		**MAE**		**SD**	
	RF	*LiDAR*	*RF*	*LiDAR*	*RF*	*LiDAR*	*RF*	*LiDAR*
On Top (n = 69)	8.33	8.33	0.841	0.730	3.84	3.84	8.32	10.83
Edge (n = 66)	13.12	13.12	0.894	0.768	5.85	5.85	12.82	19.09
Thermokarst (n = 16)	10.99	33.73	0.893	0.592	5.42	30.35	10.69	25.08
Open Area (n = 26)	4.54	14.23	0.926	0.519	1.56	9.84	4.40	12.59

Figures 5, 6, 8, and 9 have been updated based on the new results.

Figure 5 now includes the recalculated snow depth maps. We have incorporated all areas where SD_{LiDAR} values are below 0, visualizing these parts in red to highlight regions where the LiDAR sensor may have measured incorrectly or where degradation has occurred between flights.

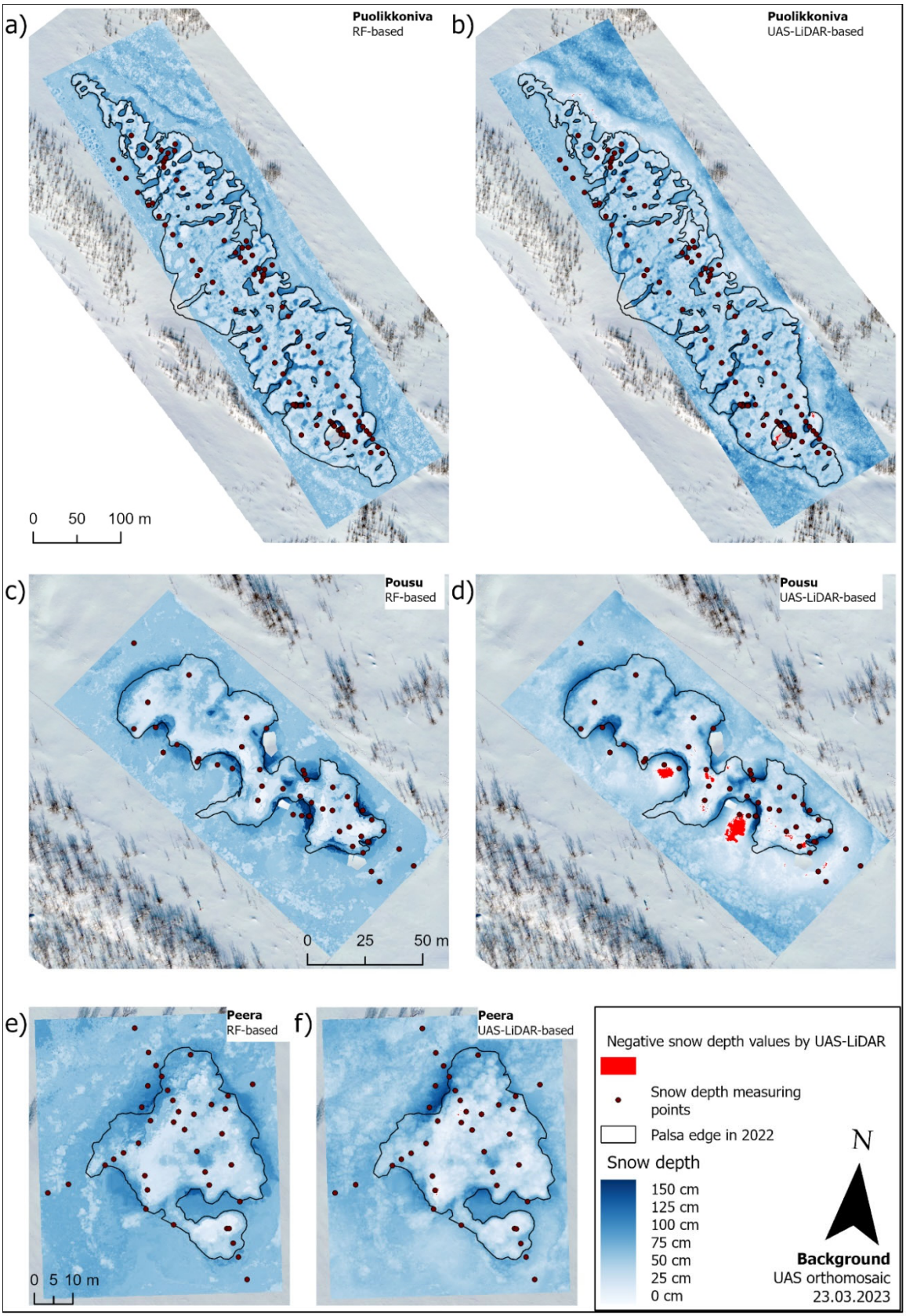

Figure 5. Snow depth predictions based on the RF model (left) and the UAS-LiDAR (right) at site Puolikkoniva (a, b), Pousu (c, d) and Peera (e, f) palsas. Red points are showing the in-situ snow depth measurement locations.

In Figure 6 we inserted the new calculated difference maps and we also included the parts with negative values in red:

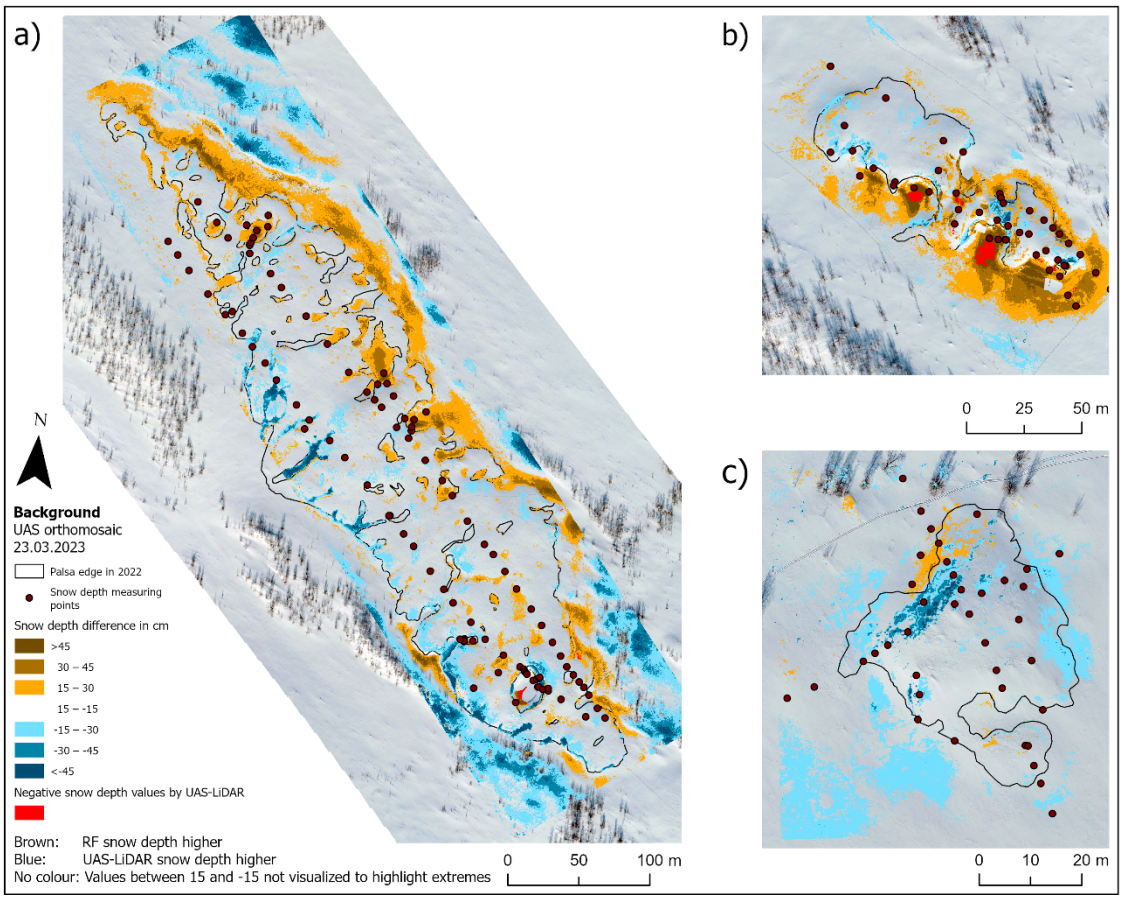

Figure 6. Snow depth differences between modelled and UAS LiDAR results at a) Puolikkoniva, b) Pousu and c) Peera palsas.

Figure 8 shows the scatter plots based on the 30% test dataset. Here we used only the single values of the $SD_{in\text{-}situ}$, not considering the values within the buffer areas of the test data. We decided to do it like that, to obtain a very fine validation of both methods:

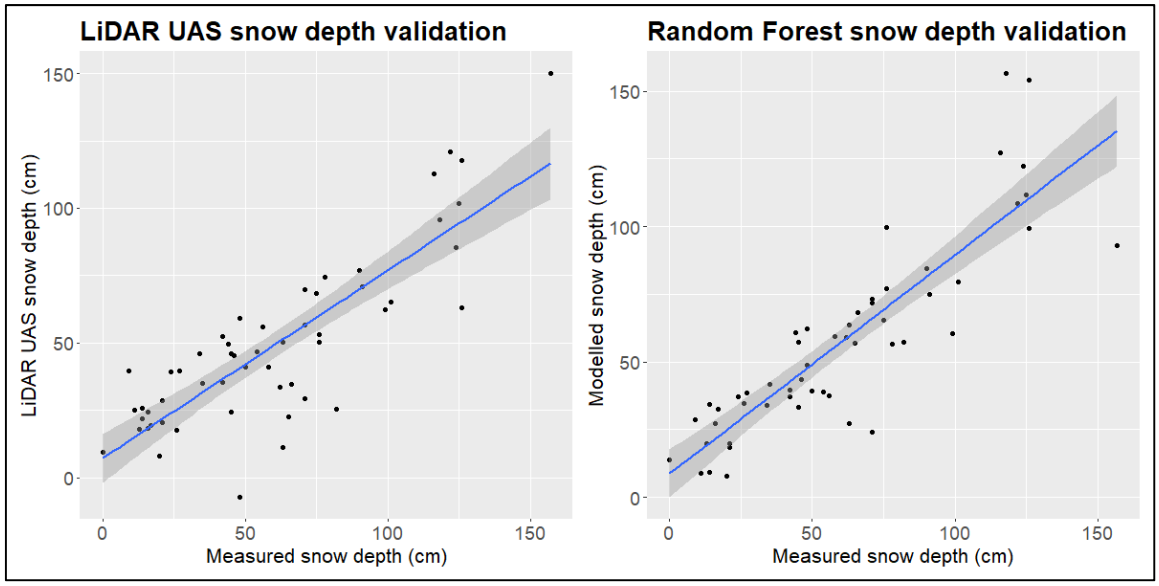

Figure 8. Scatter plots with regression lines for UAS-LiDAR-derived and RF-modelled snow depths, based on the external test dataset.

Figure 9 has been updated to reflect the new results. Additionally, we have incorporated the calculated slope derived from the DTM of Pousu palsa.

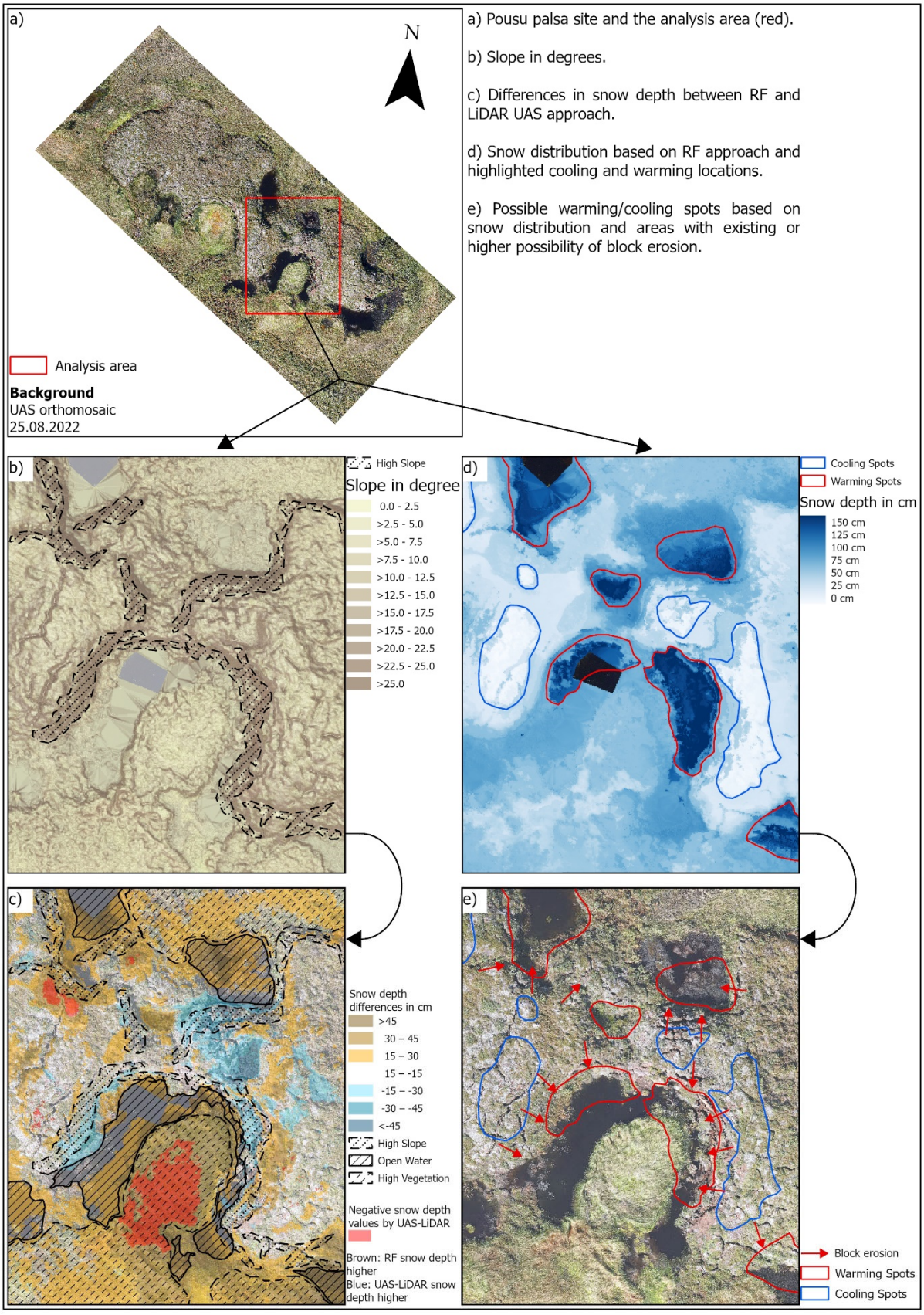

Figure 9. Explanation of differences between UAS LiDAR-derived and RF-modelled snow depths.

Appendix B

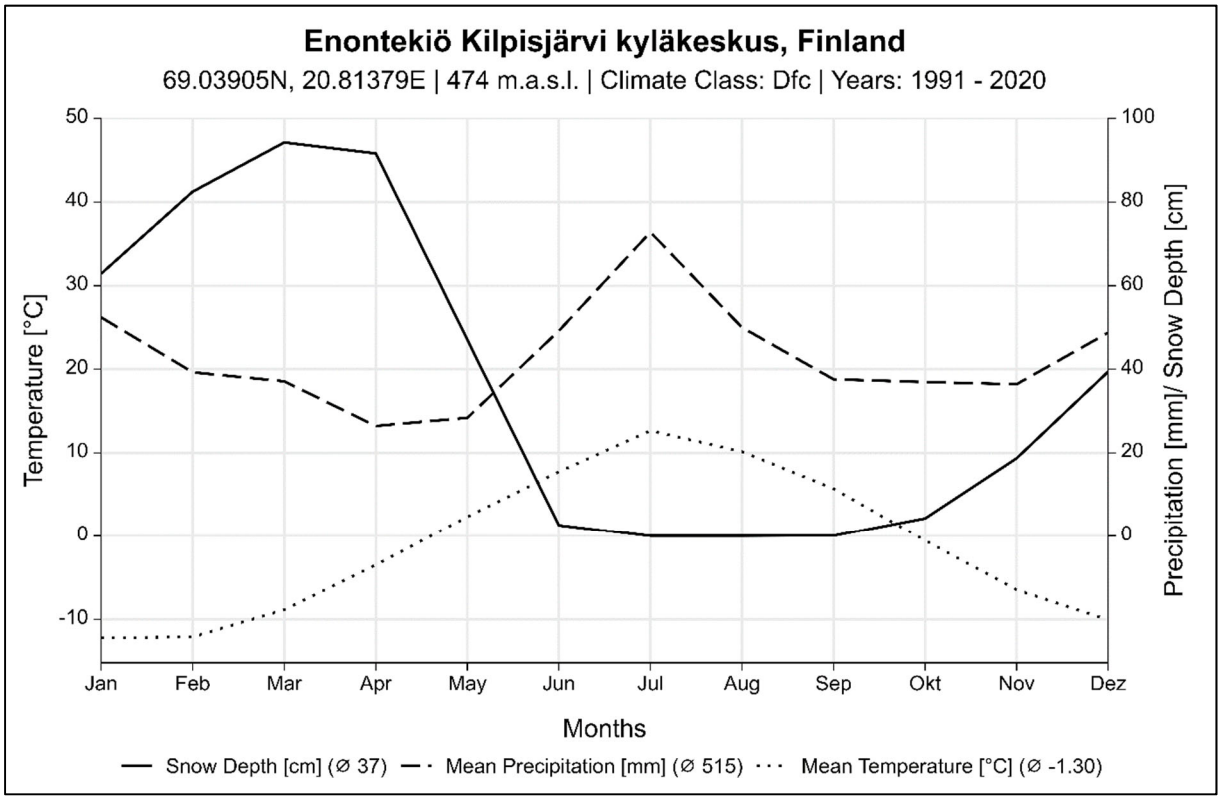

Figure 2. Climate chart of Kilpisjärvi (FMI, 2022). Dotted line shows 2 m above ground temperature in °C, dashed line shows precipitation in mm and solid line shows snow depth in cm.

Appendix C

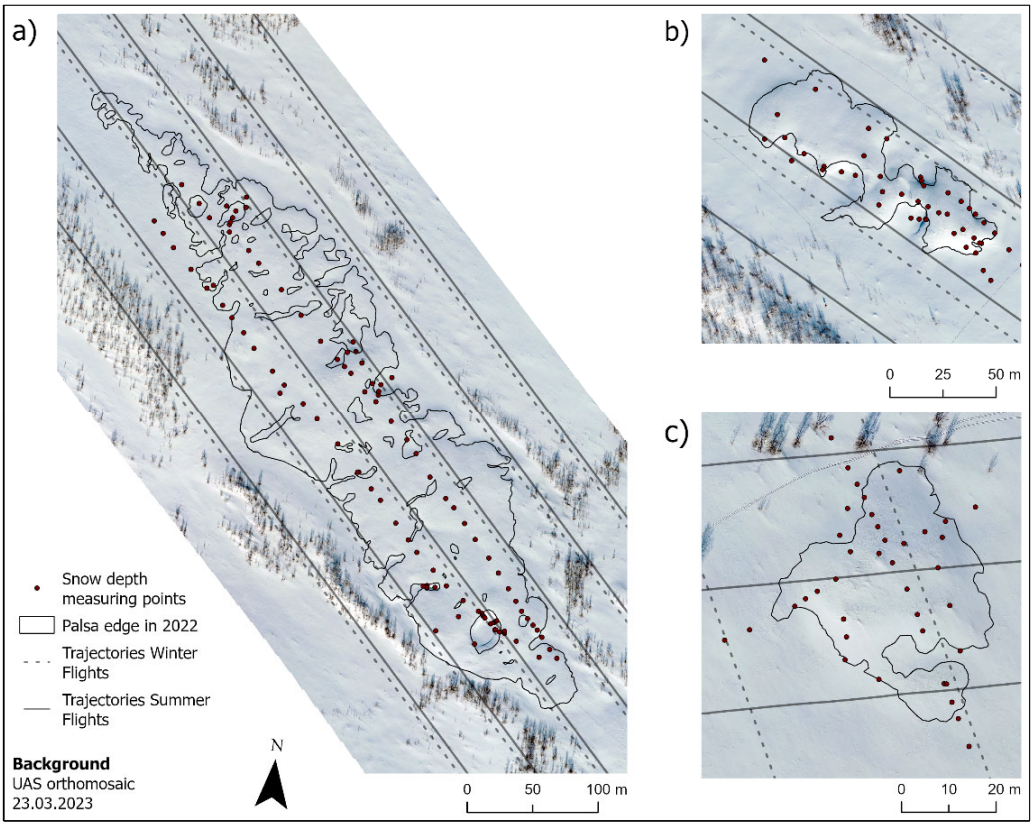

Figure 4. Snow depth measuring points within the investigation sites at Puolikkoniva (a), Pousu (b) and Peera (c) palsa illustrating different methods for recording snow depth (transects, randomized, crossed).

Appendix D

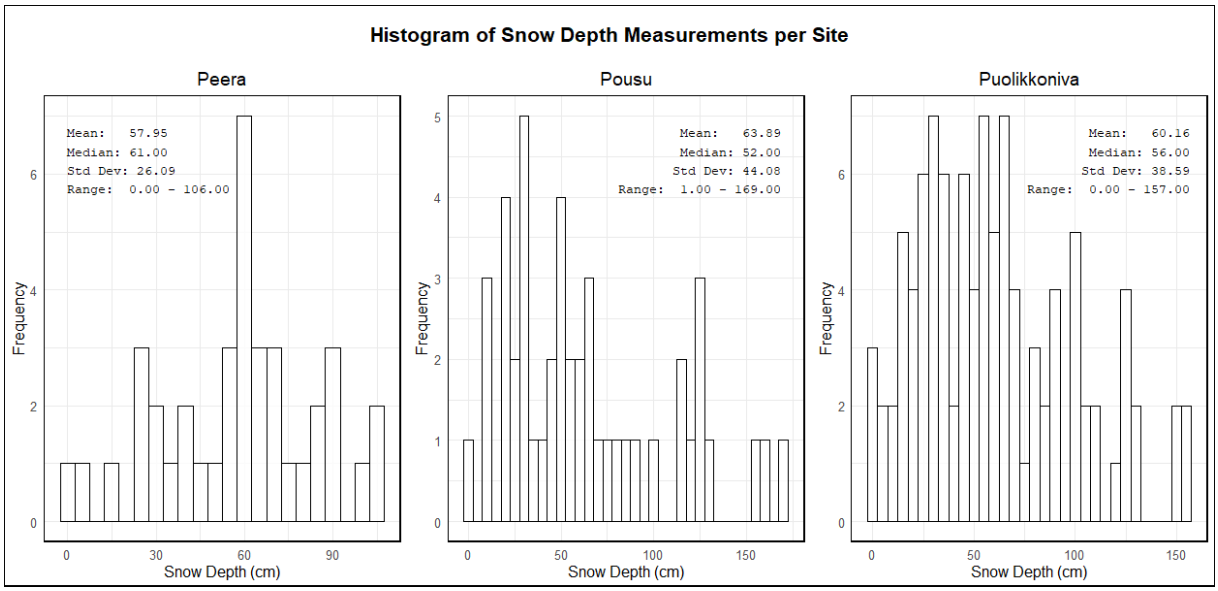

Figure A1. Histogram of $SD_{in\text{-}situ}$ points and respective statistics per palsa site.

Appendix E

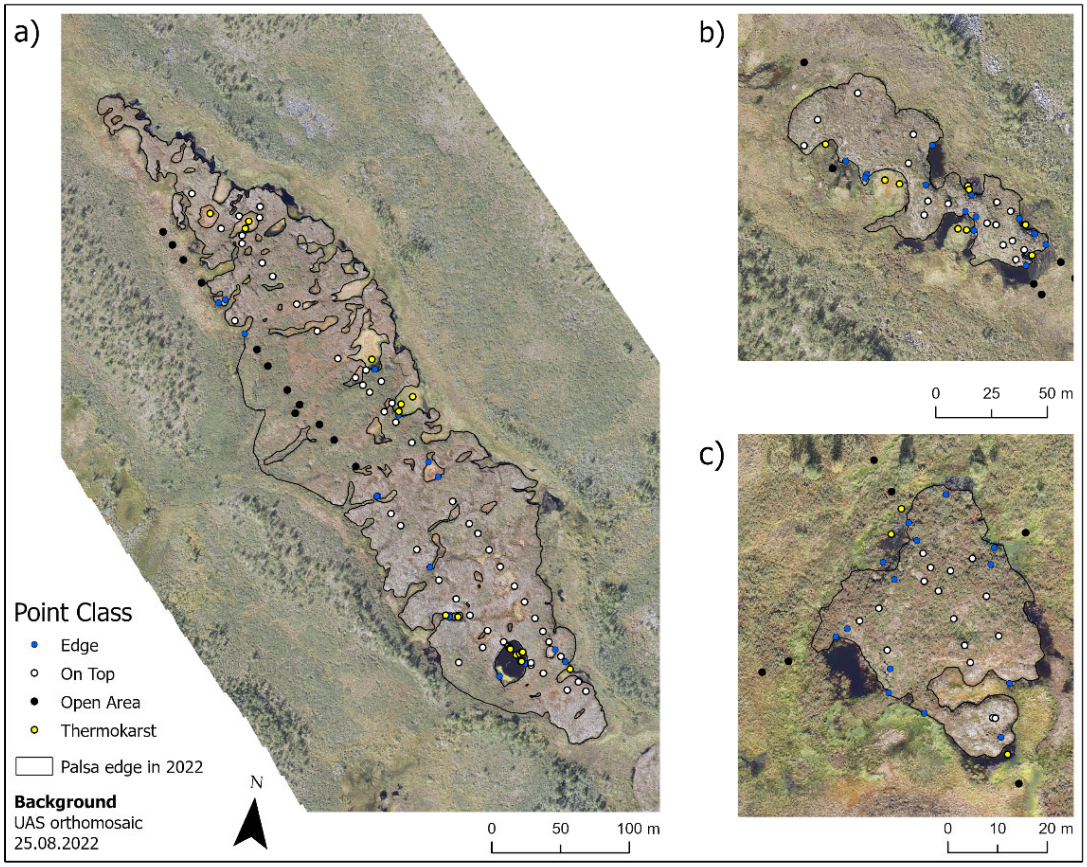

Figure A2. Overview of classification of all $SD_{in\text{-}situ}$ points into classes Edge, On Top, Open Area and Thermokarst.

Appendix F

Table A3. Correlation between each input parameter and RF-modelled snow depth.

Parameter	Correlation to SD_{RF}	Parameter	Correlation to SD_{RF}
Aspect	0.09	Relative Slope Position	-0.49
Elevation	-0.12	Slope	0.08
Channel Network Base Level	-0.09	Topographic Position Index	-0.87
Channel Network Distance	-0.45	Valley Depth	0.50
Negative Openness	0.22	Wind Effect	-0.55
Positive Openness	-0.50	Wind Exposition	-0.80

References

Avanzi, F., Zheng, Z., Coogan, A., Rice, R., Akella, R., and Conklin, M. H.: Gap-filling snow-depth time-series with Kalman Filtering-Smoothing and Expectation Maximization: Proof of concept using spatially dense wireless-sensor-network data, Cold Reg Sci Technol, 175, https://doi.org/10.1016/j.coldregions.2020.103066, 2020.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M.: mlr: Machine Learning in R, Journal of Machine Learning Research, 1–5 pp., 2016.

Breiman, L.: Random Forests, Mach Learn, 45, 5–32, 2001.

Deems, J. S., Painter, T. H., and Finnegan, D. C.: Lidar measurement of snow depth: A review, https://doi.org/10.3189/2013JoG12J154, July 2013.

Harder, P., Pomeroy, J. W., Helgason, W. D., and Helgason, W. D.: Improving sub-canopy snow depth mapping with unmanned aerial vehicles: Lidar versus structure-from-motion techniques, Cryosphere, 14, 1919–1935, https://doi.org/10.5194/tc-14-1919-2020, 2020.

Jacobs, J. M., Hunsaker, A. G., Sullivan, F. B., Palace, M., Burakowski, E. A., Herrick, C., and Cho, E.: Snow depth mapping with unpiloted aerial system lidar observations: A case study in Durham, New Hampshire, United States, Cryosphere, 15, 1485–1500, https://doi.org/10.5194/tc-15-1485-2021, 2021.

Probst, P., Wright, M. N., and Boulesteix, A. L.: Hyperparameters and tuning strategies for random forest, https://doi.org/10.1002/widm.1301, 1 May 2019.

Renette, C., Olvmo, M., Thorsson, S., Holmer, B., and Reese, H.: Multitemporal UAV lidar detects seasonal heave and subsidence on palsas, Cryosphere, 18, 5465–5480, https://doi.org/10.5194/tc-18-5465-2024, 2024.

Zhang, K., Chen, S. C., Whitman, D., Shyu, M. L., Yan, J., and Zhang, C.: A progressive morphological filter for removing nonground measurements from airborne LIDAR data, IEEE Transactions on Geoscience and Remote Sensing, 41, 872–882, https://doi.org/10.1109/TGRS.2003.810682, 2003.