Author response to Editor and Reviewer comments

Manuscript "UAV LiDAR surveys and machine learning improves snow depth and water equivalent estimates in the boreal landscapes" by Ylönen, M. et al.

We are grateful for the reviewers' comments and their positive assessment on our manuscript. We have now carefully addressed all the comments, and outline corrections, clarifications and deeper discussion as requested. Here we provide a point-by-point response addressing the reviewer's comments. At the end, there is also a comprehensive list of all relevant changes made to the manuscript.

Yours sincerely,

Maiju Ylönen

Referee #1 comment of egusphere-2025-1297 manuscript entitled "UAV LiDAR surveys and machine learning improves snow depth and water equivalent estimates in the boreal landscapes" by Ylönen et al. 2025.

This manuscript presents a study combining UAV LiDAR surveys with machine learning approaches to estimate snow depth and snow water equivalent (SWE) in boreal landscapes. The work addresses a gap in snow monitoring at local scales and demonstrates the potential of integrating high-resolution remote sensing with clustering techniques. However, several methodological aspects require clarification and improvement. In addition, the English writing also needs to be generally improved. Some sentences are very colloquial and redundant.

Answer: Thank you for the positive and encouraging comments, which helped us to improve the manuscript. You can find our detailed answers below to all suggested points, especially concerning methodologies. The language in the new version of the manuscript will be checked by a native professional proofreader.

Here are major comments:

RC1. The authors tested several DTM processing methods, but there is a lack of quantitative comparison of different methods. Also, suppose the MinZ produced "notably poorer accuracy" in the Sodankyla May campaign. In that case, it seems better to use another DTM model because the May campaign represents critical snowmelt conditions where accuracy is more important than the consistency of the method.

Answer: Thank you for your insightful comment. You are correct in noting that the May terrain model was clearly of poorer quality than the other models in terms of flooded areas. The reason we ended up using a terrain model calculated using minimum values rather than averages is that we wanted to use the same model processing protocol throughout the entire study and also ensure that it was the same in Pallas. It might be justified to vary the processing method used in the terrain model depending on the accuracy of the outputs, but we believe it is more consistent to use just one and be more comparable within each other. In other methods, we would have had to compromise on

accuracy for other months, because the terrain model based on minimum values was the most accurate for the winter period. We will add a note to the manuscript to clarify our goals.

RC2. Regarding the ground classification parameters, such as steepness, minimum object height, and so on, can you provide relevant rationale for parameter selection and sensitivity analysis?

Answer: We used an iterative, trial-and-error approach to optimize ground classification in Yellowscan CloudStation. Multiple combinations of parameters—such as minimum object height and slope tolerance — were tested and visually evaluated against field observations and GCPs. The final configuration effectively minimized misclassification and produced the most accurate and realistic DTMs for our boreal study area. This same parameter set was applied consistently across all campaigns, including both bare-ground and snow-covered conditions. Although snow accumulation can smooth terrain features and influence classification (e.g., reducing local slope), the selected settings yielded stable and reliable results across all conditions. This information will be added to the manuscript to clarify the parameter selection.

RC3. The authors combined K-means clustering and random forest, but they are not sufficiently clear about their machine learning input strategy. For each site, how were multiple campaigns combined? I assume 4 campaigns were stacked to create a 4-dimensional feature vector for each 1 m pixel, but you need to clarify this more clearly. Are there any quantitative indicators (or cross-validation in space) to assess the clustering results? Did you do some sensitivity testing for different k values? When you used the random forest, I am confused about what drivers you considered. Vegetation? Terrain? How about their importance? Terrain is a very important predictor at a small scale, LAI and canopy cover significantly affect the intercepted snowfall by branches. Air temperature, solar radiation, and wind speed alter snow accumulation and ablation. It's necessary to include more predictors in the random forest and show readers the key parameters and prediction accuracy or validation results.

Answer: Thank you for your suggestions and you pointed out very well, where the text does not answer clearly the questions mentioned. In the random forest model, we do not use predictive factors. The random forest model uses the following variables: the number of clusters, number of samples for k-means, maximum iterations for k-means, number of starts for k-means, number of variables randomly sampled as candidates at each split, and number of trees in the random forest. These are the only parameters that are given, and the model then classifies the data based on the given number of clusters. The explanatory factors you mention are certainly all relevant to the spatial distribution of snow, but they cannot be included separately in k-means clustering. We have discussed their significance in the Discussion section in paragraph 4.2 and will clarify the K-means clustering method in paragraph 2.3.2.

RC4. Line 35: delete "on".

Answer: Will be corrected in the manuscript.

RC5. Line 39: delete "water".

Answer: Will be corrected in the manuscript.

RC6. Line 156: How about the spatial distribution of 5 GCP?

Answer: Will be clarified in the manuscript.

RC7. Line 157: Can you explain a bit more regarding the MinZ and Meanz?

Answer: That's an insightful question, thank you for pointing it out. In Yellowscan CloudStation, MinZ and MeanZ refer to different gridding approaches for generating a digital terrain model (DTM) from the classified ground points in the LiDAR point cloud. The MinZ method assigns the minimum elevation value (Z) within each grid cell, favoring the lowest detected point, which is commonly used to reduce vegetation or object influence and approximate bare-earth surface. The MeanZ method calculates the average elevation within each cell, which can smooth out noise but may include non-ground points if classification is imperfect. We included both to examine how sensitive DTM accuracy is to these surface interpolation strategies. We will add this further clarification to the manuscript.

RC8. Lines 160-162: Rephrase these sentences.

Answer: Will be rephased and clarified in the manuscript.

RC9. Line 170: In section 2.2.3, can you indicate the specific number of manual snow observations? It's better to put a table in this section, including the mean snow depth, mean snow density, and the number of samples for each snow course in both regions.

Answer: This information will be added to Table 2 to describe the averages of SWE and snow depth of each campaign and their standard deviations.

RC10. Line 235: Can you introduce something about the Δsnow model? What is the input and output? What is the method of snow density calculation that is very related to SWE?

Answer: Good point. The Δsnow model relies solely on daily snow depth observations as input. Snow density is estimated using a set of seven calibrated parameters. The model comprises four distinct modules, which are selectively activated based on the magnitude and direction of snow depth changes between successive time steps. The model outputs daily estimates of snow water equivalent (SWE). We will remove the Δsnow model from this chapter and move it entirely to section 2.3.4, where the model and it's equations for density calculations are there fully explained.

RC11. Line 227: "a smaller, random sample of the data", what is the specific percent?

Answer: By this sentence we are referring to a random subset of the data. The number used is 1600 and can be found in the code published. We used 1600 sample size as it was recommended by Geissler et al. (2023). As we are doing the sensitivity analysis to be added to this manuscript, we will add this information as supplementary material.

RC12. Table 2: Are they the mean values? Please describe the title clearly and show the standard deviations.

Answer: Will be corrected in the manuscript.

RC13. Lines 305-306: You attributed the low trueness (13.2) to flooding. Is that also related to the DTM processing method you mentioned before?

Answer: This question is related to RC7. Yes, indeed the usage of the minimum z value in flooded areas might be part of the reason for the poorer quality of DTMs. Also, the reflection of laser beams from water surfaces caused issues for surface detection. We will clarify this in the discussion part of the manuscript, because this issue has also been noted by the other reviewer.

RC14. Lines 700-705: Can you discuss the weather conditions for 2021-2024 and the corresponding snow evolution, combining Fig. 8?

Answer: Very good point and thanks for a valid suggestion. We will add a connection to the weather figure (Fig. 2) and the model outputs for previous winters (Fig. 8).

RC15. It's better to add some content regarding the uncertainty propagation, which is helpful to understand the error sources. The workflow of this study is "UAV processing – clustering – random forest – snow model- final SWE and snow depth estimates". Each step can introduce errors that may compound, so some comprehensive uncertainty analysis and confidence level discussion are necessary. Except for some systematic errors, some uncertainties are from the method limitations. For example, there are not enough snow density observations. And the daily spatial distribution of snow depth and SWE from the snow model is identical in each cluster. The cluster probability of each pixel from the random forest is static. These errors or limitations are worth discussing.

Answer: Good point, and this has also been requested by the other reviewer. We have decided to add a sensitivity analysis part to our model and output evaluation. Uncertainty will be discussed when we conduct a sensitivity analysis for the model and respond to these uncertainties in the discussion section.

RC16. Importantly, please polish the English writing throughout the whole MS. Sometimes I can't follow.

Answer: Next manuscript version will be checked by native professional proofreader for English writing and grammatics.

Referee #2 comment of egusphere-2025-1297 manuscript entitled "UAV LiDAR surveys and machine learning improves snow depth and water equivalent estimates in the boreal landscapes" by Ylönen et al. 2025.

General comments

This contribution by Ylönen et al. provides a comprehensive assessment of UAV-based LiDAR combined with machine learning techniques for improving snow depth and snow water equivalent (SWE) estimations in boreal landscapes. Data collection was carried out in two study sites in Northern Finland. Overall, the study is scientifically robust, methodologically sound, and contributes some novel insights into snow hydrology and closing the observation gap between local, in-situ and regional-level snow depth and SWE

mapping. It successfully integrates high-resolution UAV LiDAR data with ground-based measurements and advanced clustering methodologies to generate spatially detailed snow characteristics. Overall, the manuscript is well-structured. However, to enhance comprehensibility, some sections could be revised (see comments below). I also found several typos and formulations that should be double checked by a native-speaker / typesetter. Some sections are somewhat difficult to understand due to a convoluted and partially unclear sentence structure. Figures in the manuscript are generally well presented and organised, captions are clearly written and comprehensible.

Answer: Thank you for the positive and encouraging comments, which help us to improve the manuscript. You can find our detailed answers below on all suggested points. The language in the new version of the manuscript will be checked by a native professional proofreader.

Specific comments

RC17. I realise there is up to now no operational implementation of satellite-based snow depth monitoring, but I feel this point should be further elaborated on in the introduction, rather than just mentioning it in passing.

Answer: Thank you for your kind comment. We had originally written more extensively about the use of satellite data in snow monitoring but ultimately left it out to condense the text. However, satellite observations are not directly related to our research, except that new methods are needed to verify them and that their limitations have motivated our research on this scale. We will add a few more sentences to address this issue more in the introduction part of the manuscript.

RC18. I missed a mention of the publications on ALS-based snow depth monitoring in the introduction, e.g. the ASO (Deems et al.) – feel this should be included here.

Answer: Good point. This information will be to the introduction part of the manuscript.

RC19. The terms ‘snow course network’ or ‘snow course measurements’ is maybe slightly misleading – in-situ snow depth measurements? Also, I did not find any reference to AWS in the introduction. At least in the Alps, AWS are traditionally the main drivers of spatially explicit snow depth maps – maybe something that could be included in the introduction too.

Answer: The definition of snow course network will be clarified in the manuscript and we will also include information about automatic weather stations (AWS). The term “snow course network” refers to the manual snow depth and density measurement protocol that is carried out as part of the hydrological monitoring in Finland. AWS are also an important part of hydrological monitoring, so including this information would improve the introduction part of the manuscript.

RC20. Overall, I would advise to outline more clearly, how the integration of UAV LiDAR with machine learning significantly improves upon traditional remote sensing methods. This improvement could be emphasized more explicitly to distinguish the presented approach from previous methods.

Answer: This will be clarified in the manuscript introduction.

RC21. Regarding the comparison with the GCPs: In which study area was this comparison carried out? I take it the GCPs where present for all flights during a given campaign, as the May campaign is stated as being the one with the poorest results? For anyone not familiar with the mentioned processing methods, a brief explanation of the routines and the employed methodology would be helpful. Were multiple echos available for processing? How come the R-methods were outperformed? Also: Part of the paragraph, where this is first described (2.2.1) is redundant to part of chapter 2.3.1. Structure-wise it would seem clearer to me, that the method-section of 2.2.1 was moved to the 'data analysis' section below. Regarding the issues with the overlapping points clouds mentioned in the former section, I feel this could also better be moved to and discussed in the analysis chapter.

Answer: GCPs were used on all sites in all campaigns. This will be explained better in the manuscript. R-methods were tested to see their accuracy in comparison to the Cloudstation's outputs – it was tested in order to make the best possible decision between different processing methods, but the result analysis and further discussion is out of the scope of this manuscript. The laser scanner can produce up to three laser echoes, and this information is in the supplementary material and the reader will be guided to check the information there.

About the structure of the manuscript, we respect and agree with the reviewer that it is clearer for the reader if the discussion about the DTMs and their inaccuracies would be moved to the data analysis section, so it will be moved to chapter 2.3.1.

RC22. Adding to the comments on chapter 2.2.1 above, what was the point cloud density of the recorded ULS-scans? What kind of accuracy is stated by the manufacturer? Generally speaking, the term DTM is used somewhat confusing to me in the manuscript, since it usually refers to the terrain, i.e. surface of the bare ground or vegetation. Thus, I suggest either using the term DSM for the snow-on datasets and DTM for snow-off, or the generic term DEM for both.

Answer: Regarding the laser scanner and point cloud density, we will clarify for the reader that this information can be found in the supplementary material.

About terminology, we understand that the term DTM may be the term used in certain research areas to refer specifically to the ground surface and that it may be confusing to the reader when we are talking about the snow-covered ground surface. However, we have used the term DTM instead of DSM because it is a terrain model (DTM) and not a surface model (DSM), as snow forms the "terrain" in winter. The DSM would include trees and other environmental features, which are not present in the DSMs we are using. On the other hand, the use of the term DEM in this context would seem too broad, even if it covered both terms. The terms we use are explained in Chapter 2.2.1 and their use is consistent throughout the manuscript.

RC23. In 2.3.1: How was the choice of the parameters for ground classification made? Was this same parameter set used for all campaigns, i.e. were both bare ground and snow cover identified with this

setting? I guess at least the parameter 'steepness' would change with the snowpack build up and terrain features being smoothed out?

Answer: This question has also been addressed by the other reviewer (RC2), so it clearly indicates that this part needs to be addressed better in the manuscript. We used an iterative, trial-and-error approach to optimize ground classification in Yellowscan CloudStation. Multiple combinations of parameters—such as minimum object height and slope tolerance—were tested and visually evaluated against field observations and GCPs. The final configuration effectively minimized misclassification and produced the most accurate and realistic DTMs for our boreal study area. This same parameter set was applied consistently across all campaigns, including both bare-ground and snow-covered conditions. Although snow accumulation can smooth terrain features and influence classification (e.g., reducing local slope), the selected settings yielded stable and reliable results across all conditions. This information will be added and clarified in chapter 2.3.1.

RC24. In chapter 2.3.3 I'm having some trouble grasping the method – I advise restructuring and rephrasing this section to improve clarity. Overall, it remains unclear to me how the k-means and random forest methods were connected and how the random forest model was parameterised.

Answer: Thank you for pointing this out. This question is similar to the other reviewers' comments (RC3). To clarify the methodology, we will explain better both the random forest and the models better in the methodology part of the manuscript, as they seem to be written unclear. They are also part of the sensitivity analysis we will add to the manuscript final version. K-means clustering is used to assign initial class labels based on LiDAR based snow depth maps. These labels train a random forest, which generalizes the clustering over the entire dataset. The random forest model uses the following variables: the number of clusters, number of samples for k-means, maximum iterations for k-means, number of starts for k-means, number of variables randomly sampled as candidates at each split, and number of trees in the random forest. The random forest model outputs probability maps for each cluster, which are later used in the synthetic snow depth and SWE calculations.

RC25. In chapter 3.1 (linking to the comments above on 2.2.1), I take it the accuracy of the RTK GNSS measurements was recorded? This could give some indication, whether the varying accuracy is in fact the cause for poorer May results as indicated in the manuscript.

Answer: The GCP plates have been measured with RTK GNSS device (Trimble or Emlid), which reports 7-8mm horizontal and 14-15mm vertical RTK accuracies. Therefore, this magnitude of measurement errors cannot alone explain the poorer results. We will add the accuracy information to the manuscript.

RC26. Generally, I suggest providing justification for selecting three clusters despite methodological indices suggesting varied optimal numbers. Clarify the criteria beyond "simplicity and comparability" for site comparisons.

Answer: Good point and in agreement with the other reviewer's comments (reviewer #1). This point will be better addressed by including the results of the sensitivity analysis in the manuscript and further elaborated in the chapter on that analysis. This will be done for the final version of the manuscript.

RC27. In the results chapter, the discussion of errors, particularly for the May campaign in Sodankylä, could be expanded. How exactly do flooding conditions influence LiDAR returns? Provide more explicit details or references that explain why these conditions cause significantly larger errors.

Answer: The comment is similar to the comment RC13, so this part clearly needs to be improved. This will be better explained and discussed in the manuscript.

RC28. The discussion chapter could more explicitly show how the presented clustering method generalizes to other boreal or similar ecosystems, particularly considering interannual variability. Are there landscape conditions or climatic contexts where this method might face significant limitations?

Answer: This topic has been discussed in chapter 4.2, and after the sensitivity analysis we will add arguments to the discussion to support further investigation and discussion. Climatic conditions do not directly influence the final clustering result, provided that several snow depth maps and daily reference data are available for the area. For example, areas where snowfall and thaw timing vary drastically and snow falls repeatedly directly on bare ground may make clustering difficult. It is not possible to make comparisons between different years based on the data we have collected, so it would be important to collect similar data in another winter and compare the results of clustering these. However, the original developer of the ClustSnow model, Geissler et al. (2023), obtained a similar result in their study area in the Alps with four clusters so we could believe that the number of clusters has the most influence on the accuracy of the model.

RC29. Consider expanding on how the presented findings specifically support operational hydrological forecasting and climate adaptation strategies. Providing more context on potential applications would strengthen the impact of the manuscript.

Answer: Good suggestion. More detailed discussion about the applications will be added to the discussion part of the manuscript.

Technical corrections (non-exhaustive)

RC30. Title: Maybe the title could underscore the connection between ULS and ML clearer, e.g. 'Combining UAV LiDAR surveys and machine learning…

Answer: Thank you for a good suggestion.

RC31. Line 13: regions are experiencing

Answer: Will be corrected in the manuscript.

RC32. Line 18: and a machine

Answer: Will be corrected in the manuscript.

RC33. Line 26: patterns in at the

Answer: We are not sure about this comment – does the reviewer suggest changing the "in" to "at"? Corrected so in the manuscript.

RC34. Lines 29f: management, and offering new

Answer: Will be corrected in the manuscript.

RC35. Line 30: forecasting, and climate

Answer: Will be corrected in the manuscript.

RC36. Line 34: in at high latitudes and in mountainous

Answer: Will be corrected in the manuscript.

RC37. Line 35: cover, the timing

Answer: We are not sure about this comment – it seems the text is already in requested format.

RC38. Line 35: distribution directly influences on climate

Answer: Will be corrected in the manuscript.

RC39. Line 42: this passage could be more specific re the mentioned impact on the snowpack (i.e. volume, extent, snow depth, season?)

Answer: Will be corrected in the manuscript.

RC40. Line 44: flood monitoring/early warning/prognosis (?)

Answer: Will be corrected in the manuscript.

RC41. Line 52: However, for the accurate

Answer: Will be corrected in the manuscript.

RC42. Line 67: re lighting conditions, maybe add that this applied to snow in particular

Answer: Will be corrected in the manuscript.

RC43. Line 83ff: Not sure I understand the point that is being made here

Answer: Will be clarified in the manuscript.

RC44. I'm confused, why the manuscript sometimes mentions two (e.g. line 107) and other times three study sites/areas (e.g. line 117).

Answer: Will be corrected in the manuscript – the real number is two.

RC45. Table 1: Instead of 'LiDAR extent', maybe 'Area mapping with ULS' → potentially a good move to introduce ULS (UAV Laser Scanning) as an acronym in the manuscript for brevity (see related publications)

Answer: Thank you for your suggestion to better specify the use of lidar in connection with UAV flights. However, according to the EU Agency for the Space Programme, ULS is already reserved for a different use ("uplink station"), so we will continue to use the established term UAV LiDAR.

RC46. Figure 2 caption: UAV drone ULS

Answer: Similar response as to RC45 above – we have considered the change of this term and decided to keep with the original, which better corresponds to the current vocabulary in the research field.

RC47. Line 175: GPS GNSS

Answer: Will be corrected in the manuscript.

RC48. Throughout: Ensure consistent use of abbreviations (e.g., SWE) once introduced to avoid unnecessary repetitions

Answer: Will be corrected in the manuscript.

List of all relevant changes made in the manuscript

- Grammar review
- Standardization of terminology and abbreviations throughout the manuscript
- Clarification of the classification of the ground surface in the point cloud
- Reorganization of chapters in the methodology section
- Better explanation and argumentation of the reasons for the poorer quality of the DTMs in May
- Clarifications to the methodology used in accordance with the reviewers' wishes, as well as descriptions of model parameterization, input, and output
- Addition of sensitivity analysis and review of its results, also covering the selection of the number of clusters