Review comments for Liljestrand et al.

Liljestrand et al. employ a two-step machine learning framework to derive basin-wide, high-resolution snow depth from a limited number of in-situ snow depth samples. The framework identifies optimal sampling locations based on areas with representative physiographic features, then applies a Gaussian Process Regression model to estimate snow depth at other locations based on the physiography at a pixel. This work presents a compelling method to estimate basin wide snow depth based on a limited number of user collected samples. In this document I outline some minor revisions to improve the manuscript prior to publication.

General comments:

- The authors emphasize that the model relies on in-situ samples and static terrain features (lines 85-86). This overlooks the role of snow-on lidar in the snow depth estimation process. The GPR is a supervised ML model which relies on the snow-on lidar data for model training. Model results are then validated on the same day as the lidar flight, meaning that the model had access to proximal snow depth data from the same day in its estimations. A key question that is not addressed is how transferable the GPR model is to other times. If citizen scientists collect data in different years or periods within the snow season, how would this impact model results? I understand it would be difficult to address this question given lidar data availability in the study area, but this seems worthy of discussion since it impacts the applicability of the methods for their intended use.
- The introduction discusses the importance of snowpack as a water source, but the manuscript focuses on snow depth with little mention of SWE. I recommend adding a paragraph regarding the decision to focus on snow depth and the potential future applicability to SWE.
- The authors primarily use 10 as the number of optimally placed sampling locations. At one point using five samples is mentioned, but I do not see results for this. To me, this is a key question of the study: how few locations can be sampled while still getting quality results? Additionally, Snotel represent a ‘one sample’ framework. How much advantage does multiple locations pose over a single sample? This is important since Snotel provide the advantage of temporally continuous data. I see this is mentioned in the discussion as a subject of future work. Based on the available data it would be feasible to reduce the number of sampling sites and produce results. If this is out of the scope of this paper, some justification could be provided.
- Table 2 is the only place where results are presented for more than 10 samples. If results were calculated iteratively up to 100, it could be helpful to visualize errors with the number of samples on a line plot, even if just added to the supplement.
- The figures only visualize model error. It would be helpful to include figures which visualize both lidar snow depth and model snow side-by-side (addressed in line-by-line comments).
- The Discussion section could benefit from subheadings to improve readability.

Substantive line-by-line comments:

Title: The title broadcasts the “leveraging of citizen science data” but no such data appear to be utilized in this study. I understand that the approach/findings of the paper have implications for

guiding citizen science data collection, but I feel the title has potential to be misleading. Consider reframing.

Line 2. ‘Address gaps in basin-scale snowpack modeling’ is a bit vague. Can you briefly describe the gap you are addressing? Increasing spatial information of the snowpack?

Line 9. Maybe state which dataset represents the “true snow depth distribution”.

Line 12. Add “in the training data” after “excluded”.

Line 34. Meromy et al., 2013 and Herbert et al., 2024 (references at end of document) are more recent papers which explore Snotel representativeness.

Line 37-39, 144. Is the assumption of normally distributed snow depth key to the methodology? As in, does the GPR make assumptions about the distribution of snow depth when making predictions? If yes, this assumption could be explored further in the discussion. Does the methodology deteriorate if snow depth is not normally distributed? If this is not key to the methodology, this information feels unnecessary. Additionally, it would be useful to show the lidar snow depth histogram to convince readers it follows a normal distribution.

Line 69: Well, not “anytime” (weather/clouds can still be a factor in the snow-free season. Consider rephrasing.

Line 70. I was a bit confused here when you mention ‘snow-free lidar data’. If my understanding is correct, saying something like: ‘physiographic data from snow-free lidar scans’ would make this sentence easier to follow.

Line 74 (paragraph). Citing examples of papers which use citizen science could be beneficial here. See Crumley et al. 2021.

Figure 7. I recommend adding panels which show the modeled snow depth and lidar snow depth (in addition to the delta snow depth). This allows for the reader to make an easier visual comparison of the two maps. I also recommend adding sampling locations to the map.

Figure 7. Is there an explanation for the horizontal blocks of similar error? There appear to be blocks of ~5 horizontal pixels that tend to register the same error (and maybe the same snow depth?). Is this an artifact of the GPR? Topography?

Line 132: How was it decided that northerly is used as the upwind barrier direction? Is this the dominant wind direction in the area?

Figures 2, 10, etc.: The use of red (FB) and green (HK) lines/markers may render some figure unreadable those with red-green color vision issues. Please consider revising to make the figures more accessible.

Line 142: Please provide the date of the sampling effort. I don’t see it anywhere in the document.

Lines 142-148: Was ground-truthing conducted for the lidar-derived snow depths? If so, what differences/errors were found?

Line 147-148. The explanation for why one study area wasn't upscaled could be clarified.

Figure 3 and Lines 134-138: One could argue that the avalanche runout zones (flatter slopes below the avalanche slopes) should be included as "avalanche-prone terrain", since the goal is to avoid measurements in dangerous zones.

Line 175-177. I don't see any results for the model which used five sampling locations.

Line 201. Any justification for the use of the GPR model? Pros/cons versus other models? Or just following the Oroza methodology?

Figures 7 & 8: It would be useful and interesting to show the maps of lidar snow depth and modeled snow depth in addition to the map of estimate errors (what is currently shown).

Figure 9. The left two plots are difficult to interpret based on the current colors. On the left I could barely find the curve with the lowest peak and in the middle plot I can only see four curves. Consider changing colors, using different line styles, or removing the fill on the curves.

Line 309: "streamflow forecasting" – it is odd that this is only mentioned in the conclusions but not earlier in the paper. Consider removing or providing more context earlier.

Line 310 (paragraph). In a similar vein to the avalanche terrain exclusions, I wonder how many appropriate sampling locations could be found. You currently choose the 'best' sampling location, but what if you selected the top 5 for each cluster? Then the samplers could select the location which is easiest to access. Maybe out of scope here, but just a thought!

Line 401. 'significant' has statistical implications. Maybe something like 'minimal losses' instead.

Formatting and wording comments

Lines 12, 29, and elsewhere: What does 'seamless' mean in these sentences?

Line 29: 'Products to produce' is a bit awkward.

Line 35: should be "snowpack" (no hyphen).

Line 37: Should be "snow depth" (no capital S)

Line 62: 'Aerial flown' is awkward. Maybe just 'aerial'.

Line 76. ‘collected by such users via a mobile app platform’. Maybe ‘reported’ would be more appropriate.

Line 105: Careful with capitalization of cardinal directions here and in the rest of the document. No need to capitalize East, Easterly, etc.

Line 115: ‘Snow-free’ shouldn’t be capitalized.

References:

Crumley, R. L., Hill, D. F., Wikstrom Jones, K., Wolken, G. J., Arendt, A. A., Aragon, C. M., et al. (2021). Assimilation of citizen science data in snowpack modeling using a new snow data set: Community Snow Observations. *Hydrology and Earth System Sciences*, *25*(9), 4651–4680. https://doi.org/10.5194/hess-25-4651-2021

Herbert, J. N., Raleigh, M. S., & Small, E. E. (2024). Reanalyzing the spatial representativeness of snow depth at automated monitoring stations using airborne lidar data. *The Cryosphere*, *18*(8), 3495-3512. https://doi.org/10.5194/tc-18-3495-2024

Meromy, L., Molotch, N. P., Link, T. E., Fassnacht, S. R., & Rice, R. (2013). Subgrid variability of snow water equivalent at operational snow stations in the western USA. *Hydrological Processes*, *27*(17), 2383-2400. https://doi.org/10.1002/hyp.9355