Response to Reviewer 2 regarding: “Accuracy of snow depth estimation in mountain and prairie environments by an unmanned aerial vehicle”

By: Phillip Harder, Michael Schirmer, John Pomeroy, and Warren Helgason

We thank reviewer 2 for the constructive comments and the specific edits that will clarify our findings. Our responses are provided in red text.

Regarding General Comment 1:
“There appear to be a low number (or potentially a low number) of snow depth data used to evaluate depths retrieved from SfM. In some areas of the manuscript this is clear (e.g. observations range between 3 to 19 in the Alpine), but in the prairie, measurements ‘between and at 34 snow stakes’ is ambiguous. In addition, the reader is left unaware of the spatial coverage of these measurements (within each airborne measurement area) nor how representative they are. At the very least I would expect the n-value to be included in tables 1 and 2. Currently in the literature the amount of in-situ evaluation data for airborne SfM studies are highly variable, e.g. De Michele et al. (2015) tens of depths, Bühler et al. (2015) hundreds of depth, Nolan et al. (2015) thousands of depths. So while this comment should not be seen as an impediment to publication, where very low numbers of in-situ data exist, this needs strong justification or perhaps judicious exclusion from analyses.”

We agree that the number of verification points in this analysis is quite variable. Manual snow depth observation protocols were different at the alpine and prairie sites due to the dynamics of the melt processes, and logistics. The locations of the manual snow observations were fixed throughout time at the prairie site. Each stubble treatment zone had 17 observation points identified by a physical stake for a total of 34 points at the prairie site. In contrast, the alpine site did not have a fixed snow course and snow depth measurements were limited by logistics and thus ranged between 3 and 19 sites. While the number of snow measurements is limited and variable at the alpine site, there were 100 surface measurements that were continually snow free which that had very similar errors over the course of the campaign to those of the snow surfaces. Considering the snow covered and non-snow covered surface errors together one can see that despite the limited n of error measurements specific to snow, these were not different from the large sample over bare ground. In contrast to other studies which are limited to assessing accuracy over a single or small number of flights we assessed accuracy over a large number of flights over a season. Therefore, the total number of surface observations available to assess accuracy was high. At the alpine site, absolute snow surface accuracy was assessed at 101 points and snow depth accuracy was assessed at 83 (five probe average at each point corresponds to 415 individually probed depths) points. At the prairie site, absolute snow surface and snow depth accuracy was assessed at the same 646 points. This information is now included in the tables. The locations of the points used to assess snow depth and the alpine bare surfaces are plotted in the site figure (Fig 1ab). The prairie site is very homogenous so evaluation points are quite representative of the study area. The alpine evaluation points are not as representative of the areal variation in snowpacks due to steep and inaccessible slopes but do reflect the variabilities in snow depth observed. These points are clarified in the manuscript.

Regarding General Comment 2:
“Quantification of SCA is demonstrated in Fig 8, and only briefly mentioned in section 3.4. The authors mention this is not discussed in this paper. This leads the reader to ask why not? If data are available to do this in a more thorough manner than currently presented, then this analysis would make an exceptionally valuable contribution to the literature, increase the scientific value of this paper and should definitely be included.”

This is a good comment. The quantification of SCA has been added as an objective of the paper and the manuscript section on quantification of SCA has been expanded. The discussion of orthomosaic accuracy is complementary to that for the DSM so not much text is needed to include this. The additional step needed to assess SCA from orthomosaics is to implement a classification scheme and some options such as traditional supervised/unsupervised classification as well as object-oriented classification are now discussed with a clearer example. Compared to estimating snow depth from DSMs, calculating SCA from an orthomosaic is relatively simple and so is discussed concisely.

Specific Edits:

While NIR imagery was attempted, as it is not used in any of the results or discussion I suggest excluding it from this paper.

- For the sake of brevity and lack of results all references to NIR will be removed.

While written in a very readable style, the manuscript in its current form could be shortened in many areas, losing extraneous text that is not relevant to the main thrust of the argument. This will provide room for select expansion of sections in greater detail that are currently vague. Some suggestions for sections to delete or shorten considerably are: Ln 11-14; Ln 29-32; Ln 93-97; Ln 98-104; Ln 115-118; Ln 146-149; Ln 152- 155; Ln 266-269; Ln 342-345; Ln 408-412. Could much of the information in Ln 168- 181 be put in a table, making this section much more concise?

- Many of the identified sections have been edited to reduce redundancy and/or make more concise.

Ln 137: Could the size of the areas measured be explicitly mentioned?

- The prairie site was 65 hectares but the UAV consistently mapped ~100 hectares (to ensure the area of interest was captured). The alpine site was 24 hectares in size. These areas are listed in the revised manuscript.

Ln 205: Why was vegetation negligible? I'd like more information about the nature of the vegetation here to justify this claim for the creation of DSMs.

- Alpine site vegetation was sparse and where it did exist was limited to short grasses on the ridgetop (<10cm) and shrubs and coniferous trees in deep gullies on the shoulders of the ridge. To avoid potential errors in detecting change associated with vegetation obscuring the snow, springing up as snowpack ablated or growing, accuracy assessment points (the 100 points surveyed) with no vegetation (bare ground or exposed rock) were selected. Other errors, such as offsets or tilts, which are minimized through inclusion of GCPs, had a greater impact on DSM accuracy than vegetation. This is clarified in the revised manuscript.

Ln 205 – 'most of the flights' – this is vague. How many flights? Did this affect the analyses?

- Not all flights throughout the measurement campaign had concurrent snow measurements. Only 8 flights did and this is clarified in the revised manuscript

Ln 219 – (linked to previous vegetation comment) While vegetation is said to be negligible I need more convincing that grasses, particularly on 24 July at the Alpine site after 'spring up' once the snow has cleared, would not have any impact on the on the ability to pick the ground surface from photos. I expect this concern can be allayed through local knowledge, but it needs to be made explicitly and clearly here as it has been a big issue in the past at other sites.

- See answer to previous comment regarding vegetation. These grasses were very sparse.

Ln 240: Please give more details describing what 'dynamic conditions' and 'surface characteristics' are.

- Dynamic conditions reflect changes in lighting due to variability in cloud cover and wind over the course of the flight and surface characteristics reflect changes in vegetation exposure and their shadows. This is clarified in the revised manuscript.

Ln 242: Please define either here or very clearly in 3.3.1 how 'problematic flights' are defined. Currently this is, at best, vague.

- Agreed and fixed. Problematic flights were identified upon on examination of the DSMs - we could easily see that the generated surfaces clearly did not represent the snow surface (rough, with gaps in point clouds). For four of these flights this was due to high wind conditions (> 10 ms-1) and challenging light conditions that were also reflected in quite high RMSE values. One flight at the alpine site had a bias much larger than the other flights. To date we have not been able to come up with a reasonable explanation for this situation beyond the fact that it increases with the inclusion of GCPs. Diagnosis of this error is hampered by the "black box" nature of the software, we cannot examine intermediate steps to determine where the error originates. The identification of these 'problematic flights' is more rigorously defined in section 3.3.1 of the revised manuscript.

Ln 255: Give more explanation on what is meant by 'limited observations' and why this doesn't affect the detection of differences.

- That sentence was poorly constructed and did not convey what was intended. It is changed in the revised manuscript

Ln 283: No correlation is presented. Do you mean 'related'? If so please change the terminology? If not, please add the statistical correlations.

- For the sake of brevity, the brief discussion of bias correction and the associated figures is now removed.

Ln 325-340: Uncertain that this section on SGM is that useful. Proprietary software (last sentences of this paragraph) is always problematic for scientific understanding, but somewhat unavoidable for much SfM processing. Also, please explain what '2.5D' means.

- The section of SGM is very specific to the processing software that we did use and while important to replicate/understand how we dealt with the erroneous points it is now shortened to be more concise. 2.5D refers to the type of point cloud that is used in the DSM generation. 2.5D point clouds are point clouds that do not have overlapping elements. The best way of conceptualizing this is to consider the figure at the following link: https://support.pix4d.com/hc/en-us/articles/202556289-Difference-between-a-3D-and-a-2-5D-Model#gsc.tab=0. This is clarified in the revised manuscript.

Ln 376-381: I consider this just speculation. Suggest removal.

- Removed in the revised manuscript.

Ln 335: 'were' rather than 'where'.

- Corrected in the revised manuscript.

Ln 373-375: Repetitive use of 'This'. Hard to understand what 'this' is referring to. Please re-write this section with increased clarity.

- Agreed. Section is rewritten.

Ln 472: De Michele et al. 2015 is now in TC rather than TCD.

- Reference is now updated

Ln 597 & 601: Is the mean of the absolute values not the same as RMSE? If so, then stick with RMSE as terminology.

- This is the mean of the bias values from the various flights. Since bias can be negative the absolute of bias values is used to ensure that the magnitudes of the biases are preserved. This should read (is updated in revised manuscript) “mean of absolute bias values”. This is different from RMSE, which is the root of the mean squared error.

Fig 1 c) – Is this short or tall stubble – please specify.

- Tall stubble and is now specified in the caption.

Fig 5 – Opening sentence of caption - introduce ‘Alpine’ as well as the prairie sites.

- Corrected in the revised manuscript.

Fig 7 – Add ‘100’ on the y-axis of both plots.

- Corrected in the revised manuscript.