Answer to P. Harder

The authors would like to thank Phillip Harder for his review. We provide below a point-by-point response to his comments and we explain how we intend to modify the manuscript in order to take them into account.

P.H. This work provides a deeper evaluation of the snow depth mapping from stereo satellite imagery approach proposed by Marti et al (2016). The advancement of this work is to consider some of the various/emerging DSM processing options and a more thorough error analysis with basin scale airborne lidar data from ASO versus manual discrete manual probe observations in locations not fully representing the variability present in the landscape. The prospect of obtaining high resolution (3m scale here) snow depth with errors reported here to be less than 0.8m from a space borne platform is tremendously exciting and the benefits of such a capability are clearly articulated herein better than I can summarise in this space.

Much of this article is clear and well written but there are a couple aspects which would benefit from some clarification and/or clearer justifications. I will begin with some main comments and then provide a list of more technical comments/edits/suggestions. Overall I think this work is well suited to The Cryosphere. The previous review of Buhler makes many important observations which I fully agree with. I would highly recommend the authors make those edits in addition to some more articulated here.

Main Comments:

1) Error model: there is a lot of discussion of methodology and results of the scaling of the random error with length scale. I have a couple concerns on this. First we are looking at the random error metric, articulated later on as the standard deviation of the snow depth residual error. This is only one part of the error, as in overall error is comprised of random components (captured here) as well as biases (not a part of this model). Correct me if I am wrong but my read is that increasing length scales will lead to decreases in random error and will therefore not comment on the bias error? Or, this model shows that increasing the length scale will increase precision but does not say anything about the accuracy? We could have really large biases in the dDEM but these will not be reflected in the model? A quick read from the abstract doesn't articulate this nuance that I am perceiving. Without lidar or ground observations to correct for the bias this suggests there will be operational challenges to implement this method in data sparse regions. Would co-registration on common snow-free stable areas be a reasonable approach to provide these relative differences? Would you have any other suggestions to improve applicability in data sparse areas?

We modified the error model to use the one proposed in Rolstad et al. (2009) which is a more rigorous and better defined than what was shown in the initial manuscript. We deeply modified the presentation of it (Method 4.2, Results 5.5, Discussions 6.5) and hope that it is more understandable now.

Anyway this model indeed concerns random spatially correlated error. The version we use assumes that the data are not biased. We first compare it to uncorrected HS residual to match what one would obtain on another study site without validation data and without the possibility to unbias or detrend the data. We then compare it to corrected HS to measure the impact of the satellite jitter. We now comment on what level of bias should always be expected with satellite photogrammetry (L487):

"Finally, although the bias or systematic error is corrected on stable terrain, there remains a bias on HS of the order of ~0.20 m (Table 4) that should be taken into account in the error calculation. According to the literature, this bias can be estimated by comparing the mean and median of elevation differences over stable terrain (Gardelle et al., 2013) or by calculating the residual of co-registration vector when more than two elevation datasets are available (Nuth and Kääb, 2011)."

About co-registration in data-sparse areas, we added a paragraph in discussion L527:
"A lack of well distributed stable terrain in snow-on and snow-off DEMs can complicate the co-registration process in some regions. The horizontal component of the co-registration vector can be measured without differencing stable terrain and snow covered terrain (Marti et al., 2016) but the vertical component requires some stable terrain or an elevation reference. GCPs could be used but would limit the applicability of the method in remote mountains. Besides, it remains to be tested how many GCPs would be required and how precisely their position should be measured."

Second, there is an assumption that stable terrain residual random errors apply to snow surfaces. Is this a valid assumption? In most topography the snow surface will reflect the underlying ground surface. In complex alpine topography the range in surface elevations can be orders of magnitude greater than the hs corresponding to the dDEMs and therefore this is a reasonable assumption. But in many areas prone to wind redistribution parts of the landscape can be smoothed out, for example, wind blown snow fills up gullies.
Therefore, I would expect the random components of the error to vary between snow and snow-free area. While the ability to predict error of a snow depth product is of interest I worry that this is perhaps a little too simplistic and detracts from the main point of the paper which is a detailed evaluation of satellite stereo imagery snow depth measurement and its comparison to airborne lidar data. The comments from Buhler about partitioning the error the errors of the snow-free and snow covered surfaces would be very valuable and provide a clearer interpretation of what the HS error is derived from.
We agree that many reasons point to the fact that stable terrain residual and HS residual might not be equal. However there is often no other choice than using statistics over stable terrain to estimate uncertainty of HS. We try to evaluate this approach in 6.5 comparing either statistics of stable terrain or HS in the error model with the measured residual at different resolutions. We comment on that in discussion L481:
"This analysis shows that the model proposed by Rolstad et al. (2009) provides a good first order estimation of the random error after spatial aggregation under the assumption that there is no spatial drift in the error at scales beyond the correlation length. In most cases, the statistics of the HS residuals are not available and might be only measured on stable terrain. Interestingly in this study, the length of correlation of the error is similar over stable terrain and snow terrain. However, the dispersion (NMAD, standard deviation) is two folds larger over snow covered terrain than stable terrain, which leads to a proportional underestimation of the error."

We now comment on individual DEMs evaluation. We include figure S2 and table S1 in the supplement and this discussion in section 6.4 of the manuscript L448:
"We further compared Pléiades snow-off DEM with the ASO snow-off DEM and Pléiades snow-on DEM with the ASO snow-on DEM. The latter was calculated by adding the ASO snow-off DEM and the ASO HS. Both Pléiades DEMs are co-registered as described in 4.1.4. We find a mean bias over snow-covered terrain of +0.13 m for snow-off conditions and +0.21 m for snow-on conditions (Table S3). These biases are of the same order of magnitude and suggest that a bias in the Pléiades snow-on DEM is partially compensated by the difference of the surface observed in the snow-off DEM (see above). In addition, the ASO snow-off DEM was acquired in October 2015 and the Pléiades snow-off DEM in August 2017. Growth or decay of the vegetation can occur over almost two years, leading to elevation differences between the snow-off DEMs. The NMAD is larger for snow-off DEMs (0.80 m) and snow-on DEMs (0.93 m) compared to HS residual (0.69 m). This shows that some errors are consistently present in the snow-off and snow-on DEMs of each type (airplane lidar or satellite photogrammetry). Pléiades DEMs are indeed over-estimating the surface elevation as the terrain slope increases (Figure S3). This suggests that combining satellite photogrammetry and airplane lidar DEMs may lead to larger errors than keeping homogeneous sourced DEMs."

Table S3. Comparison of the snow depth residual (HS Pléiades minus HS ASO) and stable terrain elevation difference (Pléiades). All metrics are in meters. The bold line is common to this table and Table 2.

	Area (km²)	Mean	Median	NMAD	RMSE	standard deviation
difference of elevation difference ($HS_{Pléiades}$ minus HS_{ASO})	**138.02**	**+0.08**	**+0.10**	**0.69**	**0.80**	**0.79**
difference of elevation SNOW OFF ($DEM_{Pléiades}$ minus DEM_{ASO})	138.02	+0.13	+0.01	0.80	0.96	0.95
difference of elevation SNOW ON ($DEM_{Pléiades}$ minus DEM_{ASO})	138.02	+0.21	+0.13	0.93	1.09	1.07

2) Structure: There is a lot of detailed technical discussions but many sections of this paper would benefit from stepping back for a moment and explaining the justification for what is occurring and how it fits in to the overall story. As it is there are some disjointed sections. (2 examples: line 125-126 – this sentence requires context, line 240 onwards – why do we care about developing an error model?)

We modified many parts of the article to make the scientific goals clearer for instance in the introduction (L77). We also modified the error model calculation and presentation. Its use is now justified in introduction (L84). We think that presentation of the error model makes sense in this article as it makes use of this rare datasets combination and is of interest for future similar studies. We present the error model L243 and associated results L331 and L463.

Specific comments:

Line 21 "A recent new method" -> "Recently, a method"

We modified according to the reviewer's suggestion (L21).

Line 32: "up to" -> "down to"? I don't see anywhere else in the paper where it is mentioned that there is a factor of two decrease in random error going from 3m to 20m (short of a reader interpretation of Figure 10). Also as a sample size increases the standard deviation (aka random error here) will always decrease – can you articulate the nuance that the modelling modifies this relationship by accounting for spatial correlation?

It is true that we did mention this interpretation of Figure 10 out of the abstract. We added it in the model error results L332:

"The measured error of the HS map decreases with increasing resampling resolution (Fig. 10). The NMAD of the HS residuals is reduced by a factor of almost two by resampling from the original resolution of 3 m (NMAD=0.69 m) to 36 m (NMAD=0.38 m)."

We agree that the random error decreases as the sample size increases. The point of the error model is to define at which rate it decreases with the sample size. This especially depends on the length of correlation of the error. We tried to explain this and why we use Rolstad et al. (2009) error model in the rewritten method (L243):

"The accuracy of HS maps is often discussed at (or close to) the highest resolution that is allowed by the sensor (e.g. Nolan et al. 2015, Marti et al., 2016). In practice however, HS maps may be subject to spatial averaging to assimilate in a snowpack model, to estimate catchment-scale HS or to compare with coarser satellite products and model output (Painter et al., 2016; Margulis et al., 2019; Shaw et

al., 2019). The accuracy of the mean HS of a set of contiguous pixels is expected to be higher than a single pixel accuracy but depends on the spatial correlation of the errors (Rolstad et al., 2009).”

Line 35-37: A great conclusion!

Line 44: Nolan et al. 2015 is not a UAV reference, they use SfM from a manned airplane. Other early reference options would be Buhler et al. 2016, Harder et al. 2016 or De Michele et al. 2016 for starters.

We agree and replaced the Nolan et al. 2015 reference with the ones suggested by the reviewer (L41).

Line53-55: One criticism from abstract was that the Marti paper only considered limited numbers of observations and terrain representation with manual insitu probe depths while here there is acknowledgment that there was also validation versus UAV data? Clarify this contradiction?

We agree that this contradiction is misleading. We modified the abstract (L23) in:

“*However, the validation was limited to probe measurements and UAV photogrammetry, which sampled a limited fraction of the topographic and snow depth variability.*”

Line 72: Harder et al. 2016 also showed that SGM improved performance over low texture snow in the UAV-SfM snow depth mapping context.

We missed this information. We now added the reference to Harder et al. 2016 (L75).

Line 136-137: this is not a method to do forest snow depth mapping so is it important to include forest lidar data processing steps?

We agree with the reviewer and removed this sentence and the following one explaining how SWE is calculated in ASO program.

Line 162: if including coordinate system information should also put in the projection.

We now state clearly that we used datum WGS84 L153:

“*The output DEM resolution and coordinate system was defined to match those of the ASO product (UTM 11 north, WGS 84).*”

Line 175: What was the extent of the Pleiades HS values that was outside these thresholds?

We mention this result in results section L306:

“*No HS were higher than 30 m but 0.25 km² of HS were excluded because HS was less than -1 m. This occured in areas covered with low density deciduous vegetation which was classified as snow.*”

Line 188: Is “eroded” the proper term to describe this? I feel this may be confusing for those with geomorphology mindsets.

Erosion is a term used in image morphology which can be indeed confusing. We replaced this sentence with L221:

“The stable terrain and snow masks were shrunk (morphological erosion) with a radius of two pixels (4 m) and patches smaller than 30 pixels (270 m²) were removed.”

Figure 2: Line passes through the “co-registration” step. Does this need a box or to be offset like other processing step descriptions?

We modified the figure so that no line crosses the processing step description.

Line 240 -255: empty super and sub script boxes appear in many of the equation/symbol text. Add space between equations and the equation numbers as well?

The equations were replaced and should appear normally now.

Section 5.1. Can you clarify the results and discussion around the comparison of pairs versus triplets? In parts (like the last sentence of this section you say the triplet is the best) while in others (line 360) it is justified that a pair of images is just fine. Line 382-384 says tri-stereo is best to provide the best coverage and reduce distortion. Can this be clarified to allow for more consistency throughout?

We agree that we did not handle conclusion about the benefit of bi- or tri- stereo clearly. We do not find a clear benefit of using tri-stereo images in this study case since front-back images resulted in snow depth with comparable accuracy and no data gaps. However, we do not think that this will hold for any terrain, image acquisition angle (B/H of the front-back pair) and processing workflow.

We tried to make this clearer through the manuscript.

For instance, L281:

"In the following sections, the HS map from the front-nadir-back geometry is used as it yielded the lowest bias, RMSE and NMAD of all the geometries although similar to the front-back geometry."

and L382:

"We do not find a large added-value of the tri-stereo images for the map accuracy compared to an optimal bi-stereo configuration. Tri-stereo might provide greater benefits in case of image occlusion in steep slopes, which is more prone to occur with higher B/H."

Line 320-325: the low frequency undulation in HS residuals. Where is this coming from? How pervasive is this error with satellite stereoscopy or is it specific to this site/processing options? For this approach to be of value what techniques can be employed to address this error (while low amplitude could be important) where there is no ASO like lidar data for validation?

These error are likely due to satellite movement (jitter) not measured by the onboard device and thus unmodeled in the RPCs. We describe this L326:

"Such undulation pattern was observed in other Pléiades products, ASTER images (Girod et al., 2017) and World-View DEMs (Fig. 10 in Shean et al., 2016, Fig. 6 in Bessette-Kirton et al., 2018). It is attributed to unmodeled satellite attitude oscillations along-track (jitter)."

This kind of error can occur in any site and at least with optical satellites which acquire images with a pushbroom sensor. This effect can be mitigated by identifying the jitter in the elevation difference map either on stable terrain (Girod et al., 2017) or over the complete area in rare cases (this study). ASP has a module to handle the jitter effect at the triangulation step which calculates the point cloud from the disparity map. However this module is "highly experimental" according to ASP documentation (https://stereopipeline.readthedocs.io/en/latest/index.html on april 29th). Therefore we did not make any attempt to use it. We added a reference to Girod et al. 2017 (L467):

"To verify this explanation, we applied an empirical correction to remove the undulation pattern from the residuals map. We averaged the HS residuals by pixel rows in the across-track direction and used a Fourier transform to identify the undulation frequencies (adapted from Girod et al., 2017). Then, we modelled this error by selecting the frequencies lower than 4 10^{-4} m^{-1} (i.e. wavelength longer than 2.5 km) and removed it from the HS map."

Line 330-334: are you applying the error model to the undulation removed HS residual or the raw HS residual? Can you clarify that? If you consider a semivariogram that extends out to the amplitude of the undulation does the undulation length scale appear in the semivariogram (extending figure 9)

Figure 9.b. now shows the semi-variogram for a larger range of distance. The impact of the undulation and its correction is visible and commented L322:

"The semi-variogram of the residual increases from 0.2 to 0.8 linearly for lag distances between 3 and 20 m (Fig. 9.a). Low amplitude undulation for lag distances between 2000 m to 8000 m (Fig. 9.b) are related to a low frequency undulation in the HS residual map, which has an amplitude of

approximately 30 cm and a wavelength of about 4 km (Fig. 8). The crests of the undulation are oriented in the east-west direction (Fig. 8). Such undulation pattern was observed in other Pléiades products, ASTER images (Girod et al., 2017) and World-View DEMs (Fig. 10 in Shean et al., 2016, Fig. 6 in Bessette-Kirton et al., 2018). It is attributed to unmodeled satellite attitude oscillations along-track (jitter). A similar semi-variogram shape is obtained over stable terrain. From this semi-variogram analysis we estimate that the correlation length of the residuals (see 4.2) is about 20 m for both snow and stable areas."

In part 5.5 we apply the error model to the raw HS residual (not corrected). We tried to make it clearer L332:
"The measured error of the HS map decreases with increasing resampling resolution (Fig. 10). The NMAD of the HS residuals is reduced by a factor of almost two by resampling from the original resolution of 3 m (NMAD=0.69 m) to 36 m (NMAD=0.38 m). As explained in Sect. 4.2, we computed two error models using either the HS residuals (= 20 m, = 0.69 m) or the stable terrain residuals (= 20 m, = 0.40 m) to parameterize Eq. (1) and (2). We find that the NMAD of the HS residuals matches well the error modelled in for averaging areas smaller than 103 m² when (,) are calculated with the HS residuals (Fig. 10). However it does not match with the modeled error for averaging areas larger than 103 m² (Fig. 10). This is due to the lower decrease of the residuals dispersion with spatial resolution. The measured NMAD decreased by 0.07 m between 36 m resolution and 180 m resolution while the modeled error decreased by 0.22 m between the same resolutions. We attribute this mismatch to the undulation pattern identified in Sect. 5.3 (see Sect. 6.5 in Discussion)."

We apply the error model to the corrected HS residual in Discussion, part 6.5, L464.
"The error predicted with Eq. (1) and (2) does not agree with the NMAD of measured HS error for averaging areas larger than 103 m² (Fig. 10). This is likely because Eq. (1) assumes a randomly distributed error beyond the short distance correlation length (here 20 m), while the undulation pattern identified in Fig. 8 introduces an additional spatial correlation at larger scales in the HS residuals map. To verify this explanation, we applied an empirical correction to remove the undulation pattern from the residuals map. We averaged the HS residuals by pixel rows in the across-track direction and used a Fourier transform to identify the undulation frequencies (adapted from Girod et al., 2017). Then, we modelled this error by selecting the frequencies lower than 4 10^{-4} m^{-1} (i.e. wavelength longer than 2.5 km) and removed it from the HS map. As expected, this correction makes the semi-variogram of the HS residual flatter for lag distances between 2000 m and 8000 m (Fig. 9.b). As a result, there is a better agreement between the HS residuals NMAD and the modeled error with and l_{cor} estimated from the HS residuals (l_{cor} = 20 m, = 0.69 m) (Fig. 10). The improvement is more marked at lower resampling resolution. For instance, the HS NMAD is reduced after correction by 50 % at a resolution of 180 m. The improvement is under 10 % at 20 m resolution as expected since the correction only dampers a low frequency signal. When the stable terrain residuals are used to compute Eq. (1) and (2) (l_{cor} = 20 m, = 0.40 m), the modeled error is lower than the measured error. This is expected since the NMAD of the stable terrain residuals is lower than the NMAD of HS residual. However, the discrepancy between both models decreased at coarser resolution."

Line 367-387: there is a justification being made here to bi-stereo imagery versus tri-stereo. Can you also articulate/quantify what the differences there may be in terms of cost differences (financial and computing). Will help to justify from a resource perspective why we should consider doing this all with pairs of images if it can articulated that there are significant savings terms of money and computing time/power requirements.
A tri-stereo increases the cost by 50% since an additional image must be purchased (cost is proportional to imaged area). In addition larger areas can be imaged in bi-stereo configuration.

However, we have been told by Pléiades operator (Airbus D&S) that the tri-stereo mode does not significantly reduce the probability of a successful acquisition. Last, with our workflow, using tri-stereo images instead of bi-stereo images roughly double the computational time as the longest operation (disparity map calculation) occurs twice as many times.

Line 410: can you clarify "decimetric accuracy"?

We rephrased in L53:

"The results showed that snow depth could be retrieved from Pléiades images with an accuracy of roughly ~0.5 m (standard deviation of residuals 0.58 m for a pixel size of 2 m), suggesting that the method had the potential to become a viable alternative to airborne campaigns in mountain catchments with the benefits of a space based platform: access to any point on the globe and lower cost for the end-user."

Line 421-424: merge with following paragraph?

The structure of the paragraph in this part was largely modified.

Line 430+: is this supposed to be a new paragraph?

The structure of the paragraph in this part was largely modified.

Line 436: "squares of length 210m". squares themselves don't have a length – can this be expressed differently?

This sentence has been removed in the modification of the manuscript.

Line 438: "unvalidated" -> "invalidated"?

We modified according to the suggestion.

Line 455: "probably the vast majority of mountain regions with seasonal snow cover" -> "the vast majority of mountain regions with snow cover"

We modified according to the suggestion.

Line 458-459: "high competitions especially" -> "high tasking competition"

We modified according to the suggestion.

Line 459-260: can you identify specific satellite platforms that we should keep a look out for?

Several very-high resolution stereo satellites are supposed to be launched in the years to come. Airbus, which operates Pléiades, advertises the launch of Pléiades-Neo in 2020 and 2022 (four satellites with image resolution of 30 cm). Maxar (former Digital Globe), which operates the WorldView fleet, should start the launch of the WorldView legion fleet in 2021 (at least six satellites with image resolution of 30 cm). The CNES, french space agency, is also working on the CO3D project in which four stereo satellites with 50 cm resolution should be launched in 2022.

There is no hint at which conditions images from these satellites would be available for research purpose.

We added a reference to the commercial project L502:

"More frequent acquisitions should, however, become easier as new stereo satellite fleets are to be launched in the coming years (Pléiades Neo, WorldView legion)."

Line 471: this is an important point to make that needs more than 1 sentence at the end of the discussion. Can you emphasise the implications of this on where snow-depth mapping with this technique is valid and possible steps that may be available to address this limitation?

We added a new paragraph in discussion to address specifically the topic of the applicability of snow-depth mapping with this technique in other regions (6.7. Generalization to other regions).

Line 474-475: “satellite very high resolution stereo images” -> “high resolution stereo satellite images”
We modified according to the suggestion.

Figures: Information in the figures is good but the layout of the figure themselves need a fair bit of formatting work.

Figure 3, 5, 6, 7, 8, 9 , 10: Can you pull out the text/lines from inside the plotting areas in to legends (based on color) outside .
We made the requested legend for figures 7, 9 and 10 but rather like to keep the other figures as it is to keep the figure more readable by color-blind readers.

Figure 4: all of the lettering (referencing inset maps etc.) is a little confusing in the legend. Perhaps add a upper-lower case distinction?
We modified the labeling and the presentation of the lower panel to ease the reading of the figure.

Figure 3, 7: Put x-axis descriptions outside of the plotting areas
We modified according to the suggestion.

Figure 7: slope residual plot quantiles exceed the plot area.
As pointed, this leads somes boxes not to be fully visible. These boxes represent very small portions of terrain (< 1 m²). We think that it is more beneficial to keep the axis adapted for the interpretation of the largest areas and to keep the same axis in all three subplots to ease comparison.
We now precise it in the caption of Fig. 7 (L610):
“*Boxes where data were covering less than 1 km² are slightly transparent.*”

Figure 8: could the a? panel have coordinates (UTM?) to provide scale and could then remove the scale bar. Also remove “Map” title and on b remove the titles and add x-axis label and units.
We updated the figure according to the suggestion.

Figure 9: “distance” -> “Distance”
We updated the figure according to the suggestion.

Figure 10: can you provide an explanation for “raw and “corrected” in the figure caption?
We now avoid to use the term “raw” which is confusing but rather “before” and “after” correction. For instance in this caption L632:
“Empty blue circles are the NMAD of the residual HS maps averaged at different resolutions before the undulation correction. Filled red circles are the NMAD of the residual HS maps averaged at different resolutions after the undulation correction”

References:
Bühler, Y., Adams, M. S., Bösch, R., Sto, A., Buhler, Y., Adams, M. S., Bosch, R. and Stoffel, A.: Mapping snow depth in alpine terrain with unmanned aerial systems (UASs): Potential and limitations, Cryosphere, 10, 1075–1088, doi:10.5194/tc-10-1075-2016, 2016.

Harder, P., Schirmer, M., Pomeroy, J. W. and Helgason, W. D.: Accuracy of snow depth estimation in mountain and prairie environments by an unmanned aerial vehicle, Cryosph., 10, 2559–2571, doi:10.5194/tc-10-2559-2016, 2016.
De Michele, C., Avanzi, F., Passoni, D., Barzaghi, R., Pinto, L., Dosso, P., Ghezzi, A., Gianatti, R. and Vedova, G. Della: Using a fixed-wing UAS to map snow depth distribution: An evaluation at peak accumulation, Cryosphere, 10(2), 511–522, doi:10.5194/tc-10-511-2016, 2016. Interactive comment on The Cryosphere Discuss., https://doi.org/10.5194/tc-2020-15, 2020.