Review of “Intercomparison of photogrammetric platforms for spatially continuous snow depth mapping” by Eberhard et al.
https://tc.copernicus.org/preprints/tc-2020-93/#discussion

David Shean
July 29, 2020

This paper considers several photogrammetric approaches that can be used to produce gridded surface elevation products (satellite stereo, aerial, UAS and terrestrial), which can in turn be differenced to provide snow depth estimates. These methods and products are evaluated over one test site in the Swiss Alps, using photogrammetric data collected within a short time period, contemporaneous ground-truth observations and several external datasets. The authors provide some analysis of their snow depth map quality, and qualitative comparisons of the strengths and weaknesses of the different methods.

In general, this is a nice methods paper. The amount of work presented is substantial. The campaign planning was excellent, the photogrammetry processing and methods are sound, the resulting datasets are impressive, and the writing is generally good. There are many DSM processing challenges involved with precise snow depth mapping, especially when surfaces in the “snow-on” DSM are almost entirely snow-covered. The authors use more traditional, labor-intensive approaches to integrate control data early in the process, rather than alternative approaches involving point-cloud or DSM co-registration later in the process. Some decisions about data processing and presentation should be revisited, as they will impact the quantitative results, though they may not have a substantial impact on the qualitative conclusions. The writing in the results and discussion requires some improvement and I recommend that the more senior authors on the paper dedicate time to help refine these sections.

As a fellow photogrammetry enthusiast/evangelist, I enjoyed reading this paper, and had a lot to say, offering many general and specific comments. I realize that not all of these changes will be possible, but I hope that the authors will consider these suggestions to improve this paper and their future work. I look forward to reading the response and seeing this nice study published!

General Comments

- There is ambiguity with the term “ground-based”, as many readers may interpret this to be “ground-truth” (I made this mistake a few times). I would suggest using “terrestrial photogrammetry” or “close-range photogrammetry” throughout the text to avoid issues.
- I would offer that snow depth on glaciers and ice sheets is critical for properly measuring glacier mass balance, beyond the applications you list
- The summer reference data sets were acquired on June 27, 2018 and August 5-6, 2018. Were all surfaces completely snow-free during both periods? For some mountain sites with similar elevation here in Pacific Northwest, we see considerable snow in late June. Do you expect differences in the vegetation during these periods? How do you account for errors introduced by these issues?

- I wasn't clear on the strategy used for blending of the independent Pleiades DSMs to create a final triplet DSM product. If the bundle block adjustment was successful, and both P12 and P13 are usable, shouldn't P23 also be usable, even with more limited convergence angle? The text cites Sirguey and Lewis (2019), which is listed as "Sirguey, P., and Lewis, C.: Topographic mapping of Franz Josef glacier region with Pléiades satellite imagery, University of Otago, 2019." - I can't find a published version of this reference, which I think is a technical paper? Need to describe the method in this text if it is not yet published elsewhere. The last sentence of this section mentions a filter to remove triangulation error >0.5 m on the blended DSM. Why not apply this filter to the individual DSMs before blending?
- The +/-1 m (total of 2 m magnitude) over 15 km "tilt" of blended Pleiades DSM is a concern, and seems to indicate that the bundle adjustment was unsuccessful, or the GCPs used are inaccurate or poorly distributed. It's not clear why a more rigorous co-registration approach was not used here rather than a hyperplane fit (presumably least-squares? Need to provide more details on this correction) through residual values at sparse points along roads. I would think a 3D rotation would be more suitable. How does the reader know that this grid correction did not introduce additional error over snow-covered surfaces at greater distances from the points on the roads in valley bottoms?
- Figures 5 and 12 show a clipped version of the Pleaides orthomosaic and snow depth products, with data only shown for the polygon defining the Dischma valley study area. It seems like it might be valuable to show the full unclipped extent of these products, maybe in a supplementary figure. The unclipped products should cover a larger area, adjacent valleys/peaks, and hopefully several of the AWS stations identified in Figure 1.
- The 0.1 m accuracy (is this both horizontal and vertical? Why don't we see expected ~2x higher error in vertical?) for GCP and CP is relatively high for modern GNSS systems, and this will propagate to all of the photogrammetric solutions. Does the 0.1 m value represent RMSE? How was the differential GNSS processing performed? I don't remember seeing any mention of this, so perhaps a real-time correction was used. In the future, I recommend using a survey-grade GNSS for your positions, or longer occupation times with static occupations at each if RTK accuracy is poor.
- For validation, several different reference datasets were used. In the snow depth error analysis, it's unclear whether you can isolate the error from the HS_platform from the error in the "spatial ground truth". Do you assume that HS_ALS(Summer) has 0 error? The lidar vendor documentation/metadata should include some assessment of product error (likely ~1-10 cm). This needs to be accounted for. I think what we really care about here is the accuracy for each HS_platform and whether we can, for example, combine two Pleiades DSMs to independently measure snow depth. If you can isolate platform-specific error (not integrating an external reference DSM), that would be really valuable, though I realize this is challenging.
- Using a 2*STD filter is not necessarily more rigorous. More aggressive, yes. With this filter, you're inevitably classifying many "valid" points as outliers and removing. This filter will always reduce your apparent error, and some might perceive this "cleaning" followed

by lower reported error as a bit dishonest. Since you're also defining robust metrics of error - median for bias, NMAD for spread, and using quantile/percentile values (25, 50, 75%) for the boxplots, you should be able to rely on those, rather than these "cleaned" values. I would avoid calling them "cleaned" - you can use "filtered". There is value in presenting and working with "filtered" maps in figures, but try to use the robust descriptive stats in the text/tables.

- Using Agisoft's "aggressive" filtering will remove many of the largest outliers. ASP also uses an outlier-removal filter during gridding of triangulated points to produce a DSM (point2dem utility removes points with triangulation error > 3*75th-percentile). If you had used different filter settings during DSM generation, your 2*STD filter would produce different results.
- Do your errors actually have a normal distribution? Looking at box plots in figure 9, I would say no. So some of these assumptions about 2*STD and percentiles may not hold. Showing histograms (with some transparency) of delta_HS from the different sources on a single plot might be more informative than the box plots.
- I don't think the variable "HS" was ever defined in the text or captions, though it is used in several figures and tables.
- The results section was a bit hard to follow at times - dense presentation of numbers and statistics. The writing could use some additional proofreading as well - hopefully one of the more senior authors can help with this. There was a lot of interpretation and discussion presented in the results section. Some of the more speculative interpretations should probably be moved to the discussion section.
- Rather than computing a snow depth for 4 DSM inputs against the same summer reference DEM (with it's own spatially variable error), and then doing an accuracy analysis of the snow depth rasters, you might perform your analysis directly on the DSMs. If you trust the DSM_eBee(winter), that can be your reference. Differencing each input DSM against this reference will allow you to characterize the error of the input DSM, without any additional error introduced by the summer reference or the resampling process. Alternatively, you could compute a per-pixel median grid from the input DSM values, and then difference each input against that median, though this works better with n >> 4
- I'm not used to seeing the "mean bias error" (MBE) and "median of the bias errors" (MABE). You are calculating these as the mean and median of the signed (not absolute) error values, right? Both will capture any bias in the data. Also, should be definition of MABE be "median absolute bias error"? The equation in Table 3 does not indicate that absolute values were used. You're missing a "model" superscript on m_BE in the NMAD definition in Table 3.
- I'm not sure that the "comparison 1" is representative of true accuracy of the snow depth products, as the sample size of the reference manual/pole measurements is pretty limited with n=4-37, and the spatial distribution of the "ground-truth" is not ideal, with manual samples clustered along the valley bottom, where gridded snow depth values show stronger spatial autocorrelation. The ~11 usable pole samples on valley walls have good spatial distribution. I expect that if you isolated the two "ground-truth" sources

(manual vs pole), the resulting accuracy statistics would differ considerably. My guess is the apparent error for manual points will be small, but you'll see much larger errors for the ~11 pole samples.

- If you have maps of vegetation (perhaps from differencing a lidar DSM and DTM, though I think your lidar was unclassified?), you could produce a vegetation mask. This could be used to analyze your HS product accuracy only over pixels that are snow, which will allow you to test whether most of your negative snow depth values are from vegetation (primary hypothesis presented), or due to some other issue (bias, noise in DSM, etc). This is important b/c your HS maps don't cover the same extent, so some are preferentially sampling vegetated areas, and may appear to have more error as a result.
- Figure 7 shows some of the detailed snow depth products, which is really valuable. It shows the noise in the Pleaides HS product compared to the ultracam and eBee HS products. But the focus is on artifacts in all HS maps due to vegetation. It is clear that the source for these artifacts is the summer ALS DSM product that was subtracted from each of the 4 photogrammetry DSMs provided. It might be valuable to show a panel with a color shaded relief map of your summer ALS DSM and a snow-free orthoimage showing this vegetation. Could also show similar panels for a snow-on DSM and orthoimage. This is tied to the suggestion of providing an evaluation of the DSM products, rather than the derivative HS products.
- Can you add some discussion of the maximum snow depths that you're seeing in the gullies? The scales are cut off at 3 m, and I expect that you could have real snow depth of >5 m locally. Your products capture this detail really well, which is a major selling point!
- It's important to remember (and state) that snow depth maps are just a specific application of a DEM difference map. There are many studies out there offering accuracy analysis of the latter, and it's important to consider these. Photogrammetric DEMs can also suffer from elevation- and/or slope-dependent errors (e.g., Shean et al., 2016). It may be worth considering how this affects your results.
- There are several figures/tables that could be moved to a supplement, which would help reduce length and improve flow
- In your Table 7, you present some qualitative advantages and disadvantages. You might consider breaking out into columns for coverage, acquisition time, cost, accuracy, resolution, repeat, and other. One thing that is missing from this is the "processing time," in addition to acquisition time. It might be useful to mention somewhere the time required to process each dataset (manual interaction and compute), software costs, equipment costs, etc. At the end of the day, these are often the deciding factors...

Specific Comments

Page 1

Rather than "RMSEs" and "NMADs", I recommend "RMSE and NMAD values are…"

25: "too few"

25: Based on abstract, unclear why ground-based obs were used with eBee but not with snow pits; I think the issue is that the ground-based obs don't intersect with "manual and snow pole measurements"
29: specify "more than two photogrammetry platforms"
30: replace "the specific advantages and disadvantages of them" with "their specific advantages and disadvantages"

Page 2
18: Specify the sensors or methods used for "manual" or "AWS" - e.g., probing, GPR, sonic ranging, etc.; Also, point measurements themselves are not the problem - the issue is their sparse spatial density, especially over large regions. One can probe every 10 cm on a relatively small grid and get lots of valuable information about the local spatial distribution of snow depth.
27: Should include Painter et al reference for ASO: https://doi.org/10.1016/j.rse.2016.06.018
31: What kind of accuracy? Is this position accuracy, and 0.1 in horizontal and vertical? Or snow depth? Several important factors for TLS, so careful with a blanket statement like this. See also study by Currier et al (2019): https://dx.doi.org/10.1029/2018WR024533

Page 3
8: Would be valuable to mention other VHR satellite image options for snow depth mapping. See our preliminary publication on WorldView stereo snow depth: "Spatially extensive ground-penetrating radar snow depth observations during NASA's 2017 SnowEx campaign: Comparison with In situ, airborne, and satellite observations
D McGrath, R Webb, D Shean, R Bonnell, HP Marshall… - Water Resources Research, 2019"
9: Should include medium-format (Phase One) camera SfM mapping: "Assessing the Ability of Structure From Motion to Map High-Resolution Snow Surface Elevations in Complex Terrain: A Case Study From Senator Beck Basin, CO
J Meyer, SMK Skiles - Water Resources Research, 2019"
Chris Larson's work in Alaska?
23-24: Consider rewording to emphasize that few studies have evaluated these platforms for the specific application of snow depth mapping. Many past studies have done this for other surface types and scientific/mapping applications.
26: Mention high-albedo surfaces (sensor saturation), limited surface texture for fresh snow (poor stereo correlation results)
28: UAS can offer cm-scale products; all depends on altitude
31: Other issues beyond ownership can prevent flying whenever required - regulations, certified pilot availability, weather
35: "scales"

Page 4
19: "surface slope"
20: "contrasted snow depth distribution" is a bit unclear - I think you're saying that downslope winds from gullies lead to snow deposition at the base of the gullies?
24: "Data… provide" (singular, not plural)

24: not just “snowfall” though, as you have wind redistribution and melt, I think “snow depth evolution” might be better term
25: “it” - I think you mean the snow stations?
28-29: I agree that things didn’t change much, but if you consider the observed change as a percentage, then 20 cm of a ~65-70 cm snowpack is pretty substantial! You might consider framing these observed changes as a percentage of your expected measurement error. For example, 20 cm is well within the Pleiades snow depth accuracy, but much larger than the sUAS snow depth accuracy.

Page 6
10: eBee+ RTK is a “fixed-wing” - should mention this, as many readers may think “drone = quadcopter”
13: “triangulation” is not the word I would use here, as it could be confused with stereo triangulation to produce a point cloud. I think you mean consistent geolocation.
17: OK, so these are visible in the Ultracam and eBee imagery, but probably not visible in Pleiades?
19: “positioned” to ”The positions of all GCPs were determined using…”

Page 7
2: “Pleaides-1B stereo image triplet”
7-9: I would also recommend providing the combined off-nadir angles for each image, rather than the signed across and along-track components. I think your B/H numbers are just accounting for along-track off-nadir angles? If possible, good to provide convergence angle for each pair in addition to B/H.
10: Given your scene relief, this is probably acceptable
15: Provide time zone for times (and all other instances in paper)
16: Could delete sentence about not acquiring on same day, as we don’t know what the technical issues were
20 and Table 1: would be nice to provide sensor dimensions in addition to total pixel count
21: “Large-format CCD”
22: focal length in Table 1 is 122.7.
24: “I” is near-infrared? Recommend consistency with terminology used for Pleiades section
28-29: OK, so three flights total? Better to state that, rather than changing the battery twice.

Page 8:
7: time format 10.37 vs. 10:37, and time zone
7: What is imaging interval? 1 second, 1 minute? Ah, it wasn’t clear that this was a terrestrial photogrammetry survey! I thought you had installed time-lapse stereo cameras. Maybe state this earlier in the section, before you get into details on variable GSD. Could move lines 16-22 or at least 16-19 here.
11: That sounds really large for a convergence angle, and earlier you said B/H of 0.25-0.42 is acceptable? Why different here?

19: Really cool to see this setup! :) I used a similar rig for terrestrial and oblique aerial surveys in Greenland and Pacific NW: https://dshean.github.io/technology/sfm/ (Fig 4), back before consumer UAS and integrated cameras were up to the task

Page 10
12: Hmm. I think you can do better than 5 cm on the pole shown in fig 3b. Probably down to 1-2 cm precision. Is the inaccurate reading issue also related to the depression in the snow surface immediately around the pole? If so, state this. For future work, you might consider a small quadcopter that you can fly to "inspect" the poles from closer range to make more precise readings.

Page 11
4-6: provide date range for ALS survey. I see August 5-6, 2015?
9: Provide citation for LAStools. Martin is pretty clear about this, especially if you're using an unlicensed version :)
16: "NASA Ames Stereo Pipeline (ASP, Shean et al., 2016, Beyer et al., 2018) version 2.6.2 (corresponding Zenodo DOI)" - please follow latest citation instructions https://github.com/NeoGeographyToolkit/StereoPipeline#citation. Please include Shean et al. 2016, as you are using core ASP functionality for processing Earth observation imagery that was implemented during that effort.
16: not clear what you mean by "GDAL (for satellite data)" - for orthorectification? Resampling?
17: For reference, you can also use the ASP geodiff command or generic gdal_calc.py for simple raster operations outside of ArcGIS.
17: No mention of co-registration before subtraction here? I assume this will be discussed elsewhere…

Page 12:
2: Suggest "analyzing" rather than "using"
9-16: I didn't completely follow all of this, but it sounds like something you clearly thought about carefully :). Any coordinate system transformation of raster data will require some resampling/interpolation, it's a question of error introduced by the transformation. If you can provide some sense of this error, that would be valuable. It's likely several cm.
I initially thought the LV03/LN02 were from 2003/2002, but just checked and it was 1903/1902! Good call to convert to LHN95. I would think the lidar vendor should update their products…
20: Already mentioned version number and citation earlier, don't need to repeat here.
21: Should state you're using the RPC camera model first, then talk about updating during bundle adjustment
22/30: Should avoid starting sentence with "14" and acronyms "DSMs"
27: Is it expected that the CE90 and LE90 are identical? I would expect a factor of ~2 degradation in LE90.

Page 13:

2: See general comment about the citation here and question about blending. I wouldn't say "with GDAL". This is not a standard GDAL command line utility or API function. I think you created a custom script that would weight values from your 3 input DSMs based on the ASP triangulation error map? I recommend you consider the ASP dem_mosaic weighted average blending approach.
9-15: This "tilt" is a bit troubling, and potentially indicates issues with the GCPs, either their spatial distribution and/or recorded position errors. See general comment.
13-15: The last sentence should be moved to the end of the above paragraph (line 7).
19: Note that Agisoft can process grayscale images, and RPC models these days
29: OK, so refining interior camera parameters is not desirable, so you explicitly disabled from the initial alignment solution? Did you add calibrated lens distortion coefficients, or did you allow Agisoft to solve for these? If so, which coefficients? Provide more detail here.
29: Careful about interchanging GCP and CP here. I don't think you have 29 GCPs, and if you're using points to control geolocation, they are GCPs, not check points.

Page 14
2: I'm really glad that you used a factor of 2x here, and did not export an oversampled DSM at the native image GSD! This is a common issue.
5: I think this was done using the SenseFly eMotion software? This first paragraph is a bit confusing, as you're already talking about the accuracy of the DSM but you haven't told us how the DSM was produced. I recommend you move this paragraph later, and integrate with the last sentence of paragraph 2 (lines 21-22), where you actually report the observed RMSE.
If your base station position is accurate (I think you used NTRIP caster for real-time corrections, not a local base), then you should be able to achieve the accuracy you report without GCPs.
15: "EXIF metadata"
32: dGNSS - check for consistency in terminology used elsehwere

Page 15
1-2: I think you mean it was not possible to determine the precise offset of the GNSS antenna phase center relative to the center of the camera detector. Could probably estimate vertical offset and constrain a bit better than 0.2 m horizontal and vertical, but not a huge issue.

Page 17
3-5: What you did here should work. Technically, you want to use a consistent grid for all output products, with same grid cell size, projection and origin. If the spatial extent of the two rasters varies significantly, then you can use something like GDAL's -tap option to force output grid extent to use whole integer multiples of the output grid cell size. Can also do this manually in Agisoft when specifying output extent - round left, top, right and bottom bounds to nearest multiple of the cell size. A shared raster origin (upper left coordinates) or this -tap approach avoids the need for a second resampling step, which is going to further degrade resolving power of your products and can introduce additional error.
7-9: I like this approach. Approximately how many pixels were sampled within this circle?

10: So this sounds like nearest neighbor sampling for Pleiades snow depth grid. I recommend bilinear or cubic for point sampling like this, esp if you see relatively large pixel-to-pixel variability in snow depth grid values near the sites.

Page 18:
12: See general comment about the 2*STD filter
12-13: Do your errors actually have a normal distribution? Looking at box plots in figure 9, I would say no. So some of these assumptions may not hold. Overlapping histograms of delta_HS from the 3 sources might be more informative than the box plot.

Page 19:
3-5: recommend using “area” for all instances instead of “surface”
6: already said ultracam only covers northern part of valley due to clouds elsewhere (which is really too bad, data quality over cloud-free areas looks great). Does “Dischma valley” = “Dischmatal”?
6-7: what does “good quality of eBee+ flight” mean, and how does an orthoimage illustrate this?

Page 21:
3: “graphically” sounds strange to me, I think you mean “qualitatively”
7-8: “errors introduced during photogrammetric processing” - the processing itself is not responsible for the negative values
9: OK, yes, using a DSM is a problem, but the acquisition date of your summer reference DSM will also be very important in terms of the vegetation growth cycle, leaf-out, etc. Also, you are assuming that your photogrametrically derived DSM is capturing the true top of canopy, which is not necessarily the case (esp with “aggressive” filtering that will remove isolated shrubs/trees)
9 and 14: recommend “depressed” or “compressed” rather than “pressed down”
17-22: What is the total sample size for manual and snow pole measurements again? I would argue that some of your larger apparent error is due to limited sample size, and likely the nearest neighbor sampling approach for the coarser Pleaides snow depth grid.
21-23: Careful with this kind of statement, as it sounds like subjective error reporting…
23: “decreased” instead of “deteriorated”
23: The NMAD and median should not be strongly influenced to one outlier, unless your sample size is an issue.
25: Again, triangulation here is a bit ambiguous, as it could be confused with the triangulation to produce the point cloud. I think you’re talking about the integration of GCPs during the bundle adjustment routine.
27-29: This information (about the roads used to correct Pleaides DSM) should be moved to the methods section.
29: I think you’re trying to say that one sample Pleaides triplet is not necessarily representative of the capabilities of the sensor.
30: “...approach for snow-covered images” - this problem will be mitigated with snow-free conditions
30-31: Not sure this last sentence is necessary

Page 26:
10: recommend changing “data analysis” to “accuracy analysis of comparison 1 is potentially not representative of the true accuracy of the snow depth products” or something along those lines. See general comment on this issue.

Page 27:
1: Note that this is correlation of per-pixel values, and should not be confused with spatial correlation
6-7: Might be useful to report the percentage of data removed with your 0-5 m filter here. BAsed on Figure 8, this should be minimal for ultracam, but you’re removing a nontrivial sample of the Pleiades HS values.
8: Don’t start sentence with “0 m…”
11: “R^2 values”

Page 29:
3: “imagery” should be “DSMs”
4: You can only use the ultracam HS as ground truth to evaluate the Pleiades HS, right? The way it is stated, it sounds like you’re using to evaluate both.

Page 32:
8: There are also studies by Shaw et al. (2019?). Please review Simon Gascoin and Etienne Berthier’s recent publications, as they are listed on other recent papers using Pleiades for accuracy of DEM difference maps, including snow depth
12: I don’t think it’s fair to present your “cleaned” values here
12: Again, rather than comparing DSM accuracy, you’re comparing snow depth accuracy, which includes error from the reference dataset.
13: “This is higher” here is a bit confusing, as your numbers are lower. Consider rephrasing. Also “this” is ambiguous.
14: Why is one pixel considered the maximum??? Sub-pixel correlation should be capable of 0.1-0.2 px accuracy. Is the issue with your manual GCP identification? Still should be able to achieve sub-pixel with this, esp with modern GCP markers.
18-19: I don’t follow the last sentence, but I may just be getting tired :) Consider rewording. What is “they”
21: Need to qualify this with “from Pleiades, without further correction”. We have demonstrated better accuracy with WorldView-3 DEM difference (same as snow depth) products (see McGrath et al, 2019; Shean et al., 2016). And there are several systematic errors in the individual DSM products (e.g., CCD offsets, unmodeled attitude error [“jitter”], parallax issues in L1B camera models) that can be corrected to further improve accuracy. We are actively working on this, so hopefully Pleaides DSM accuracy can be improved!
25: I would avoid using the world “profit” to avoid confusion with commercial applications, “benefit” would be better

28-30: I disagree. ICP co-registration approaches can work very well, even when only sparse exposed surfaces are available. Co-registration using methods like Nuth and Kaab (2011) can also be used to take advantage of the slope and aspect-dependent dh - can then use limited snow-free surfaces for final vertical correction. I think you may be referring to cases when the entire scene is 100% snow-covered (rare for mountains).

Page 33:
14: Issues of contrast can also be problematic in satellite imagery. In my experience, it is more about fresh snowfall and whether image GSD is fine enough to capture relevant length scales of surface roughness.
18: An aircraft outfitted with high-end GNSS and IMU should be able to provide very accurate camera position and orientation data, eliminating the need for GCPs (as with your eBee RTK results)
19: Again “higher GSD” - careful here, as lower numeric value means better resolution
24-25: This is consistent with my experience using eBee RTK platform. But careful about generalizing to all “UAS photogrammetry” - a DJI Phantom with no GCPs is not nearly as capable

Page 34:
8: This is where using a 3-4 m pole can be beneficial.
14: “non-negligible”
20: Careful with this - using a DTM may help in some areas, but not others; “bushes” is pretty generic, and a bush with leaves appears very different than a leaf-free bush to a laser or camera

Figure 2:
Check TC date formatting standards - I initially read this as MM.DD
If you’re going to remake this figure, it might be better to alter the aspect ratio to reduce the width of the x axis, so we can better see the magnitude of the change over the study period (right now all lines look really flat). Either that or add thin horizontal gridlines and a complementary right axis label.

Figure 4:
I would move this to supplemental figure - it is awfully large for the information you’re trying to convey, or could be presented as a table
Need to define HS in caption
I’m still confused by comparison 3. Why is comparison 3 HS_platform_ALS?

Figure 5:
d) I don’t see the violet stars? If they are present, recommend making larger and using a color that won’t blend in with the orthoimage

Figure 6:

The color ramp used here makes it very difficult to distinguish snow depth values between 0-1 m. It would be better to use a perceptually uniform, linear color ramp, ideally with labels for increments of 1.0 or 0.5 m intervals. I realize this may not be straightforward in ArcGIS.

Figure 7:
Maybe use “apparent snow depth” or “snow depth estimate” here and elsewhere in the paper, as it’s not physically possible to have a negative snow depth
Why did color ramp change here to what looks like matplotlib plasma? Also, since you are showing values from -3 to 3 m, could be better to use a diverging color ramp. Lots of good resources on the theory behind these visualization approaches:
https://matplotlib.org/3.1.1/tutorials/colors/colormaps.html

Figure 8:
This is a nice figure, but why such a large bin size??? The quantization here makes comparison between different sources really challenging. I would recommend bin size of <5-10 cm, so we can properly assess each distribution.
Also, it would be valuable to create a mask for the common intersection of valid HS pixels (ie pixels where all 4 DSMs have a valid elevation), clip each input HS to the same mask, then produce a similar histogram, maybe as a second panel in this figure. This provides an “apples to apples” comparison, as your current histograms are sampling different portions of the domain, and there is no reason to expect your reported mean and std statistics to be the same.
In the caption, you can just say “normalized histogram” without details about dividing by total number and multiplying by 100.
What does “all not shown” mean? You have the Manual and snow poles measurement on histogram on the plot.

Figure 9:
Is “the median” the same as your MABE metric? I think MBE is just the mean error, right?
5th and 95th are not quartiles, they are percentiles. Fix in all other captions for box plots.

Figure 11:
I think this is a 2-D histogram showing density? So, not a scatterplot. I don’t think your current caption says what color represents.
Probably want to say “y = x”
Should mention in caption why the bottom row is limited to range of 0-5 m on y axis. I think this is labeled “cleaned” but you’re adding another filter here.

Table 1:
Time is local or UTC?
Add a row for sensor dimensions in pixels and/or mm. 450 MP is not really a “resolution” and we don’t know dimensions.

What do you mean by Pleaides mean GSD of 0.7 (resampled to 0.5). Did you intentionally oversample, or are the L1B products delivered at higher res after some super-resolution processing beyond normal TDI?

Table 3:
Threshold for classifying outliers

References:
Several of these are “gray literature” and some contain errors (e.g., Deems and Painter, 2006 has no journal information, year listed twice)
Authors should review all references carefully, update according to TC policy, remove gray literature, and update lingering errors from citation manager software