Reviewer: Hannah Besso

General Comments:

The paper constitutes an important contribution to the field. Snow station data are used for many applications in hydrology. This study adds to our understanding of these stations' representativeness of basin snow quantities and is an important addition to snow hydrology. The scale of the analysis and use of lidar sets it apart from previous studies. However, the authors should explain better and/or reevaluate the temporal analysis. They should also remove the landcover component from Research Question 3, since the author states in the discussion that the dataset used for this component of the analysis was inadequate. Additionally, the manuscript (especially the Analyses section) should be reorganized or condensed to make the story clearer.

Organization: The Introduction includes a deep dive into several relevant papers, whose details are repeated later in the paper. These details should be removed from the Intro. The final paragraphs of the Intro, starting at line 91, would fit better in Methods. The Analysis section seemed to jump from one thing to the next and was confusing to keep track of what you were doing. Either introduce the whole section with a list (could even be bullet points) or make separate section headers with titles that describe what you do for each of these paragraphs/separate analyses. Then in the Results it wasn't always clear which part of the Analysis you were reporting on. It would probably be best to maintain the order in both sections, consistent with the order of the Research Questions.

Vegetation Impacts: The lack of a strong vegetation component is a big missing piece of this paper, and it should be highlighted as future work that should be done. I think the paper still stands without a veg component, because the impact of the paper is the bias-correction of the stations, not the reasoning behind that bias. However, understanding the impacts of vegetation on snow quantities at snow stations relative to the surrounding basins is important. Somewhere near the beginning of the manuscript you should acknowledge that vegetation has been proven to impact snow depth, but that your dataset was too coarse to adequately investigate the impacts. Or, if you want to include a vegetation component, you could come up with a simple metric such as distance from station to canopy edge (even just using imagery like on Google Earth). As the manuscript stands currently, the discussion provides good citations of others' work on snow-vegetation interactions, but your Results section reads as if you think there is no impact of vegetation on snow.

Temporal component of the analysis: How did you decide on the different groups (that were "typically" low, unbiased, or high)? I'm not fully convinced that these groups are distinct since there's so much overlap in Fig 9D-F.

Pflug and Lundquist (2020) (see Figure 6 of that paper) show that basin snow variability can change throughout a season based on snow covered area and whether it was a 'big' snow year or not. This seems relevant to your Section 3.1, where you argue that a larger range of snow depths increases the maximum magnitude of the RSD. So it would follow that there might be an inter- and intra-season temporal component to changes in RSD at a basin. And Figure 9A-C does show that stations can have a range of RSDs of up to about 50 cm, which seems large

relative to your 10cm threshold. I don't find myself convinced that stations are so temporally consistent in their bias that it would be easy to bias-correct them based on just a few lidar flights. I think this would be a huge finding that would have implications for everyone who uses snow station data - I just want to see this proven/investigated a bit more thoroughly. A discussion also might be warranted of whether 3 years of data is enough to develop this relationship.

Snow depth instead of SWE: I've been told by people more familiar with CA data than I that snow depth measurements (especially at sites in California, managed by CDEC) are less accurate because they're not maintained or quality controlled as well as the SWE measurements are. I think your reason for using snow depth is valid (given that it's directly measured by the lidar data) but it's worth thinking about and maybe mentioning more than the brief description of your quality control method.

Technical Corrections:

Line 65: Define 'area-mean snow depth' since you use it throughout the paper.

Line 134: "provides no advantage" - higher accuracy of SWE vs snow depth measurements from CA snow stations

Lines 129 - 132: repetitive with the intro. I think you should remove these details from the Intro.

Lines 150 - 152: Might deserve further discussion.

Line157: "requires much less storage and computational expense to manage" in comparison, I assume, to the 3 m data set instead of the "range of larger scales" you reference in the previous sentence. Be explicit here.

Lines 154 - 159: How do they produce the 50 m product? Is it derived from the 3 m product?

Lines 161 - 164: DTMs can vary quite a bit in their RMSE, and errors can be spatially variable. For example, areas of a certain DTM with steep slopes or dense vegetation can have larger errors or even systematic bias than in areas that are flat with less vegetation. I think here you could just get away with reporting the error published in Painter et al., 2016. Especially because this is the only lidar dataset you use, so I don't see any added benefit to making generalizations about other lidar products.

182 - 183: Confusing sentence, I don't understand what you did.

Line 184 - 185: Reword this sentence to be more clear. "We compared the snow depth from different data sources at each point location" or something. "Use different data sources to represent the point snow depth value" is confusing.

Line 185 - 187: Can you explicitly define SD? I understand the sentence but had to read it twice to make sure SD was defined.

Line 187 - 189: Better suited when you introduce the datasets in 2.1.3.

Paragraph starting with 192: Equation 1 and surrounding text would fit better here than in the introduction.

Line 195: Why did you choose 10 cm? Also, “acceptable” is undefined. Why not say something like “this threshold will change based on the application, but here’s why we chose 10cm”.

Lines 212-213: Confusing sentence and probably unnecessary. This section would likely benefit from a summary sentence at the beginning or end that lists your analyses (see my above comments on organization).

Line 215: “proportion of representative sites” was confusing, had to read twice. Make it more clear what you’re referring to (I assume it refers to 2 paragraphs previous, where you talk about using each cell as the “station” location.)

Figure 2: I like that you show the different scales using the boxes. But I want the boxes to be different colors than the SD and Elev scales. Also make the SD and Elev gradients different color scales.
For D can you plot the 50m SD and Station SD as vertical lines that intersect your different CDFs instead of points? Otherwise there are 3 points on this graph representing the same data.

LIne 229: Typo: extra d in “Dd Cumulative”

Figure 3: Same critiques as Fig 2.

Line 233: You use Cumulative Density Function in Figure 2 caption but CDF in Figure 3 caption. Be consistent.

Line 234: What do you mean by “truncated”? Be explicit.

Line 245: Why include lidar flights that occurred when the study sites were mostly snow-free if this will skew your statistics?

Note: you already defined a threshold of 10 cm magnitude RSD. Why so much emphasis on percentiles? This seems like a useful tool in characterizing site variability, but you say it yourself that it’s a problematic indicator of representativeness, so emphasize it less. Also, how does the timing of the lidar flights play into this quantile analysis see above comments about the temporal analysis? Do periods of ablation change this relationship?

Line 264: stick to cm units for consistency

Figure 4: Are the vertical lines on your CDFs supposed to represent the 5th and 95th percentile? They don't look like they do (they're all the same width just located in different places - maybe check your code for generating these). If they don't represent those quantiles, I think they need to be labeled/explained.

Line 289: what are "low sites"? Do you mean "low-biased sites"?

Lines 291 - 294: I like this summary at the end of the section.

Figure 5: Explain what the gray vertical lines are.

Lines 305 - 308: See my above comments on the vegetation component.

Lines 313-314: perhaps due to vegetation effects.

Figure 6: The different colors overlap such that they block each other. Is there a way to make both visible via a different type of plot or by using transparency? This is especially a problem at the 0.5 km scale where I think the pink is plotted on top of the blue. Also why is the .5 vertical line lighter than the others? I missed it at first.

Line 349: "sensing scale"? Does this refer to remote sensing or something else?

Figure 8 caption: use consistent labels. "50 m Lidar pixel" vs. "50 m SD". Also, the 10 cm lines look gray to me instead of black.

Figure 9: See my above comment about how you grouped the stations. There's a lot of overlap between groups (D-F).

Line 384: I don't think "overrepresent" is the right word here.

Lines 386 - 388: this fits better here than in the Intro.

Lines 397 - 398: Rephrase. I think you're saying that any bias correction would need to be site specific. And do you mean "positively biased" not "oversampling"?

Line 400: Instead of a list you should present the infrastructure, flat terrain, etc as components of the location bias. Otherwise this conflicts with the other 2-component list you give in Results.

Lines 444 - 445: See above comments on vegetation component of the analysis.

Line 465: "pixel", not "point" for the lidar data