Reply to Adam Winstral, Referee

Referee comments are left-justified, in black. Author replies are indented, in blue.

The aim of this study is to develop a simple means of estimating snow densities to convert observed snow depths to snow-water-equivalent. The authors seek to use long-term climatological variables rather than station or modeled data so that snow depths garnered in remote locations without direct meteorological observations (e.g. crowd-sourced or Lidar data) can be easily and accurately converted to SWE. There is a growing need for improved means of characterizing snow density as greater amounts of snow depth data are becoming available (e.g. Lidar). Therefore this type of research is certainly warranted. Given that snow depths have always been more readily available than SWE or density data, other researchers have similarly produced methods of estimating densities. While not all of the previously developed approaches tackle the specific case presented here (i.e. meteorological data immediately preceding snow depth observation not required), the Sturm et al. (2010) and Jonas and Magnusson (2009) approaches do. The authors clearly acknowledge this and make a positive comparison of their method measured against the Sturm method. However, I find the presented Sturm comparison to be biased against the Sturm method (see further comments). They also hint at (lines 413, 493), but never provide evidence nor specifically claim to be, better than the Jonas approach. I would like to see the authors present in a more convincing manner how and why their method represents a substantial advancement over the previously published methods before I am ready to consider this manuscript worthy of publication.

This is my major concern:
The authors randomly split the aggregated CONUS, AK, and BC data into training and validation datasets (Section 2.2). They then use the "held-out" validation dataset to make the Sturm comparison (Section 3.1). So, essentially they have trained their model on data from the same locations with the same statistical metrics present in the comparison dataset. On the other hand, the comparison dataset is 100% independent of the Sturm training data. In order to present a fair comparison this needs to be done with a dataset that is totally independent from the derivation of both.

This is an important point. Our current approach aggregated all western North America snow pillow data (some ~2M points) and then randomly split it in two. So, for each station, some data at each station was used for model building, the other data at each station ended up being used for validation. We can see why it would be important to test a validation approach that separated the training and validation data either by location, or by time.

To address your concerns, we took all of the snow pillow data and we split up the stations randomly into two groups. We took all of the data from the first group and we used that to train the regression model. We then validated the regression model against the second group. We did several realizations of this process and found that the results were extremely close to those presented in the original manuscript. Anonymous

> Referee #3 also raised a similar concern, and suggested an 80/20 cross validation (80% of the data used to train, 20% of the data used to validate) approach. This method also generated similar results. We believe this to be due to the very large N of our dataset.
>
> Given how similar all of these approaches were, and given the lack of any clear 'preferred method' in the literature, we decided to retain our original approach.
>
> We strongly agree with the referee that it would be ideal to have a perfect test between the two methods (our model and that of Sturm et al.). However, that would require that the two models be developed with the same training datasets and then validated using the same validation datasets. Unfortunately, we don't see a way to create this perfect 'laboratory test' for two models developed with different data.

The northeast dataset would be one ideal dataset for conducting this test and I'm not sure why this wasn't done. That said, it would certainly be more convincing if the inter-model comparisons were conducted over a wider range of conditions.

> You are correct in that it would be ideal to have inter-model comparisons over a wide range of conditions. We believe that applying both models to the NE data set would not accomplish that. We prefer to keep our inter-model comparisons to the larger dataset from western North America snow pillow data, and we will retain the NE dataset for our model only.

I would also like to see direct comparisons to the Jonas method. As I stated in the above paragraph, the authors must present a convincing case that the new methodology represents an improvement over existing procedures. I just don't find that in the current manuscript.

> With regards to Jonas et al. (J09). We specifically chose not to apply that model for the following reason. The J09 model has coefficients that depend upon month of year and elevation. In addition to this, there is a geographic 'offset' term that depends on boundaries drawn in the Swiss Alps. Therefore, the model cannot be applied in other regions (since we would have no idea what to use for an offset). We do not wish to ignore the offset and apply a 'partial model' since that is not what those authors constructed.
>
> One thing that we have done is to apply the very simple Pistocchi[1] model which depends only on day of year (DOY). In Pistocchi's paper, he claims comparable performance to both Sturm and J09. We now include summary results (RMSE and bias only, no figures) for the Pistocchi model applied to the western North America snow pillow data.
>
> We believe that the results for our model demonstrate an improvement (lower bias and RMSE than existing methods) and also a strength of our approach is that it allows for a

[1] https://www.sciencedirect.com/science/article/pii/S2214581816300131

> continuously varying snow density in space rather than discontinuities due to discrete snow classes. Our plots below, provided in response to another comment, help illustrate this point.

Moderate concerns that need addressing:
I don't understand why rmse was normalized with respect to mean annual precipitation (Section 3 and Figure 8). This obviously biases the normalizations low where summer precipitation is more common. Artifacts of this can be seen in Figure 8 (e.g. low ratios in Arizona, New Mexico, Alaska where summer precipitation can be considerable compared to winter; high ratios in eastern Sierras where synoptic summer storms are rare). This type of normalization might be appropriate for annual or longer hydrologic studies, but for this snow-based, winter-focused research the normalization should be based on either mean wintertime precipitation or better yet, mean annual snowfall. Both mean wintertime precipitation and mean annual snowfall should be easily derivable from the PRISM data already used in this study.

> This is a reasonable suggestion. Our intent was simply to provide some sort of 'relative' measure of the magnitude of the RMSE. We have actually redone this using the mean annual peak SWE to normalize the RMSE, which makes good sense.

Graphs. There are way too many data points in the scatter plots to understand what is really going on in Figures 6 and 9, and some of the plots in Figure 11. These should be presented as either heat plots or randomly select and plot a subset of these data. Additionally and partly due the aforementioned reason, the overlapping plots in Figure 9 are impossible to fully discern.

> With regards to Figure 11 (Fig 12 in the revision), the symbols are colored by DOY in the left column, so we are unable to show that column as a heatmap. We have changed the center column to show the data as a heatmap.
>
> With regards to Figure 6, we have changed the plot to a heatmap (which is just a 2d histogram). The 'footprint' or 'envelope' of the data cloud is unchanged of course.
>
> With regards to Figure 9 (Fig 10 in the revision). The important point is how the 'width' of the data cloud is different between the two methods. The envelope that is closer to the 1:1 line indicates better performance. Our original approach was chosen since, in each case, our envelope was narrower (so we plotted ours on top). We cannot show two overlapping heatmaps. What we have done in the revision is to show Sturm's results as scatter symbols (as before) and to then plot our results as a transparent heat map on top.

I had difficulty accepting the reasoning for the residuals and mean biases apparent in the Figures 6b and d. I think these residuals, which are present in the validation dataset are also related to the choice of fitting a power law relationship rather than a linear least

squares one. Given that the training and validation data should maintain the same statistical metrics then these residuals should be present in the training data as well. If, in fact, this is the case then the combination of a power law fit and the predominance of accumulation season samples would be the reason. My suspicion is that if a linear least squares fit was chosen then there should be near zero mean biases in both the training and validation sets given that the two sets maintain the same characteristics. I would expect that in the linear scenario, there should be a wider spread in residuals (i.e. higher rmse) but very little change in mean bias. Of course, this would be entirely different if the validation set was truly independent.

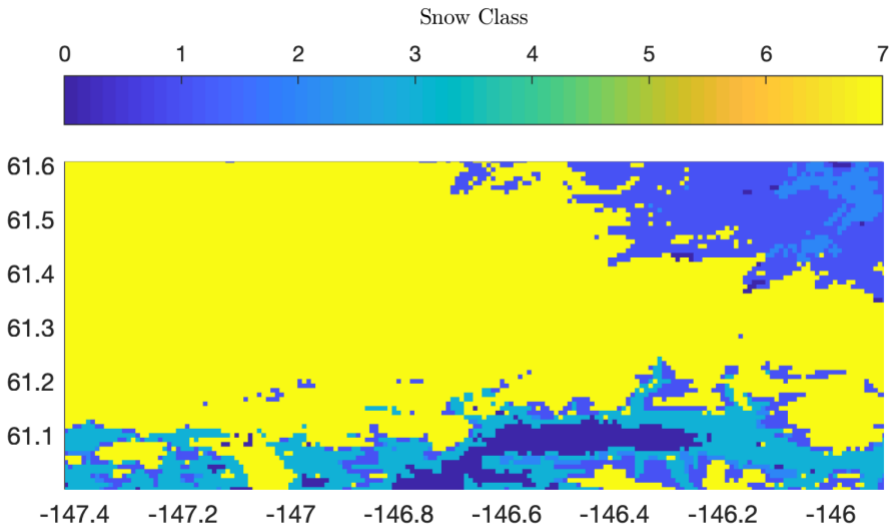
> This is a fair comment, and our initial remarks may have been too speculative. We adopted a power law relationship based on the hysteresis loop (Figs 1 and 4) suggesting something other than a linear relationship between h and SWE. We feel that the best course of action is to remove our overly speculative comment.

How the different datasets were used needs better clarification. I didn't understand the purpose of the manually sampled Chugach data. As far as I can tell, these data were not included in the calibration nor the validation analyses. What do these data show? Why were they included? How do these data add anything new to the analysis? This should be clearly stated and incorporated into the story or leave the Chugach data out.

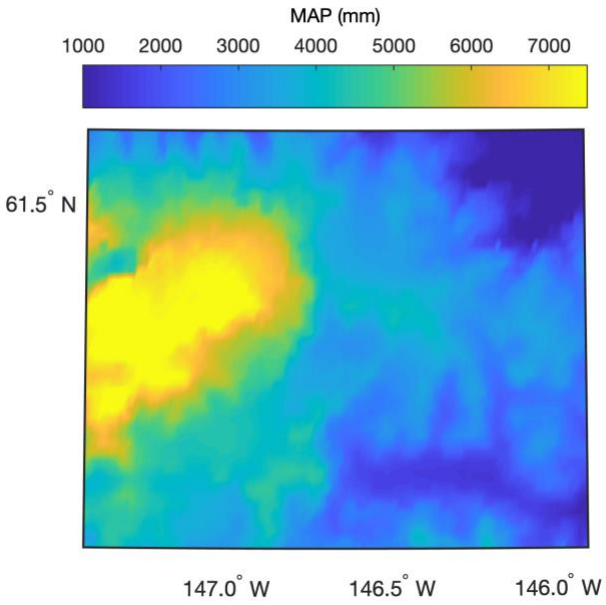
> Two reviewers of this paper noted this. We have essentially dropped that dataset from the paper, with one exception. The large ensemble (80 or so) of collections (8 at each site) of probe measurements is valuable since it helps to quantify the variability in snow depth over small distances (in discussion section).

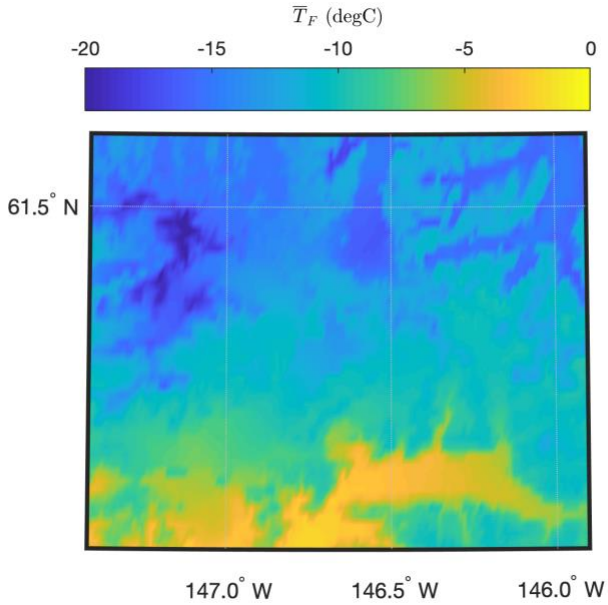
Section 2.1.2. Do these PRISM climatological variables, based on sparse station data and resolved at 800m, really pick up the heterogeneity you're aiming to capture as expressed on lines 132-37. It would be nice if you could show a spatially explicit example showing these capabilities.

> We feel that the continuously variable PRISM data does a better job of capturing climate than 5 snow classes. Let us illustrate this with some sample figures. First, consider the map of snow class in the region just northeast of Valdez, Alaska.



Note that there are only a few snow classes and that the landscape is dominated by class 7 in this case. Now, for the exact same lat / lon bounding box, let us look at the MAP and Feb_T_Mean:





In both of these climatological rasters, we see very considerable variation over a region that is monolithic in snow class. These, we do feel that the use of 800m PRISM data will allow for smoother variability in snow density.

Tidbits:
The residuals (e.g. Figure 6) should be presented as modeled minus observed. In this manner the underestimations of SWE appear as negative residuals rather than the positive residuals currently presented. I find this much easier to understand.

Actually, the residuals are done correctly. Please see the new version of Fig 6, which has been much improved by showing it as a heatmap. Look at the top row. The residuals are indeed computed as model-observed. The vertical black line in the right column (panel (b)) is the mean residual. It is negative. And that makes sense since the cloud of data points appears to be, on average, below the 1:1 line. So, thank you for your suggestion, it was good for us to double check, but we do have the residuals defined correctly, we believe.

Lines 44-47 and 72-74. Each of these sentences contain two distinct thoughts that would perhaps be better if split into two sentences.

Thank you for the suggestion. We will improve the clarity of these lines.

Lines 120-22. I didn't think this sentence was necessary . . . unless you turn it into reasoning that this just adds a layer of computational costs / complexity that aren't necessary for your desired application.

We slightly adjusted the sentences there to improve the clarity.

Lines 141-2. Might want to add something about why you would also prefer to not use NWP data that could possibly substitute for the lack of observations (i.e. computational costs, errors in NWP data).

> The purpose of this work is to provide a rapid, easy to use tool. Relying on external daily or sub-daily datasets and/or model output moves the work away from that goal and towards more sophisticated snow models. So, yes, it could be done, but at significant expense and effort.

Line 169. You also used snow pillow data from the northeast US. You might want to make that clear here . . . as in "Snow data for this project, aside from the aforementioned SNOTEL data, . . ."

> Yes, thank you. We fixed this.

Section 2.1.1.5. Might want to mention that these issues are most common in summer when vegetation grows beneath the sensor.

> We are not sure we fully understand this remark. Which particular issues are you referring to? The data that we considered was only winter time data, where snow was present.

Line 440. Roughness of underlying terrain is certainly one factor, but couldn't there be others as well (e.g. wind redistribution).

> We have now noted this explicitly.