Reply to Anonymous Referee #3

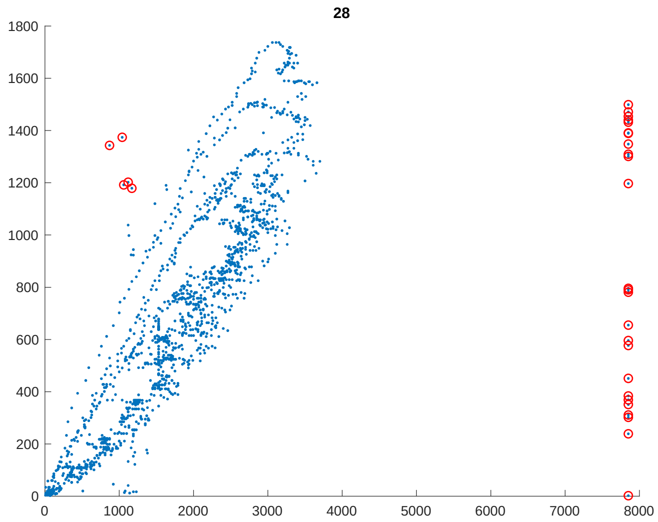
Received and published: 13 March 2019

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

The authors address the issue of converting spatiotemporal snow depth measurements to estimates of snow water equivalent (SWE). This topic is relevant to many areas of research because of the relative ease of taking snow depth measurements over SWE. Framed in the context of citizen science or field work, snow depths collected by nonexperts and experts alike can be leveraged as a low-cost input to hydrological or climate analysis. In an era of high-availability altimetry (lidar or radar) and photogrammetry (structure from motion), an ensemble of methods to convert surface heights into SWE will be critical for both targeted basin studies (ASO) as well as future satellite missions. The authors develop three regression models to evaluate a snow depth to SWE conversion. Regression skill is evaluated using depth alone, depth separated by accumulation and ablation phases, and depth in combination with climate normal for precipitation, temperature as well as elevation. Their work differs from previous studies such as Sturm et al. 2010 in that the climate inputs are regressed as continuous variables. As such, any measurement of snow depth with coordinates could potentially be converted, independent of measurement scale. In general the paper to well written and clear in its advancements. The focus on estimates during the ablation phase is a clear contribution, where methods fail. Addressing that ‘not all snow is equal’ is a strength of the approach.

Prior to publication, I would like to see the outlier detection and validation portions of the paper revisited to reinforce the statistical analysis. While I agree that outlier detection is necessary, an enhanced description of where and when the outliers originate would help to identify potential seasonality or spatial clustering. For example, if many of the outliers are from the early snow season, does this preclude ability of the models to convert measurements that include fresh snow? There are artifacts in Figure 4 where SWE varies drastically but depth does not, are these melt events? A histogram of the outlier DOY or a table of the outlier properties may be all that is needed to address this. These additions could be used to reinforce the statement that the reduced dataset is physically plausible (Lines 229-230).

> These are very good points. One other referee had similar remarks. Manual examination of many of the SNOTEL time series revealed the presence of clearly wrong data (Figure 3 of the paper). We wanted to develop a wholly objective method for removing those data points. The approach that we used seems like a good one, based upon the characteristics of the bivariate distribution. We recognize that some valid data points (mostly at low SWE-Hs values) are undoubtedly removed as well. Given the very low number (less than 1%; so the valid points removed are some small fraction of this 1%) of points that were removed in our process, we feel that this is acceptable. Here is a figure of the process at one particular station. SWE on vertical axis, h on horizontal. Red circles are removed points.



We particularly like your suggestion of looking at the characteristics of the removed points, and now include specific information on the DOY values of these points. It turned out that removed points were occurring throughout the snow season, and not just at the beginning and the end.

Your comment about events in Figure 4 where SWE changes a lot while Hs remains fixed is an interesting one. It is hard to understand how SWE could drop from 1 m to near zero while Hs remains fixed at 5 m. The lack of an accepted and easy to implement protocol for addressing snow pillow data quality control is an obstacle to analysis.

For the validation, it may be of benefit to use a cross-validation (CV) to determine if the model skill is overly optimistic. Using an N-folds CV with a 80/20 train/test split would be a simple approach to achieve this. In this regard, I'd also be interested to know if the non-SNOTEL datasets actually influence to the regression coefficients (What happens when the training datasets are SNOTEL only). The remainder of my comments addressed to specific lines or figure.

A few comments. First of all, the regression coefficients were constructed with snow pillow data only from the western United States and Canada. We have now tried to make this more obvious in the section discussing datasets. For example, in Table 1, we now use bold font to highlight which datasets were used to build the model.

Second, with regards to validation. We looked into this at some length before beginning this work, since we wanted to determine if there was some preferred way of doing validation in the snow density literature (or streamflow prediction, or any other discipline for that matter). We found no 'best' or 'preferred' method. We ended up doing a 50/50 split (aggregating all snow pillow data points and randomly dividing them up) in the first draft of the paper. Upon receiving the manuscript reviews, we also tested your suggested 80/20 split, and a 50/50 'station split' (divide up the stations, not the aggregated data points). We found that all methods provided essentially the same results We feel that this is likely due to the large N (number of observations) of our dataset. Given the lack of consensus in the literature, we feel that our approach is acceptable, and we are clear and upfront about our methods.

Lines 64-65: Are there additional references available to support this statement regarding L-band? The only cited application in the field is a conference proceeding.

Another referee though we should simply remove the remote sensing discussion and that seemed like a sensible change to us, so we did so.

Line 172: Each style of corer has its own associated bias. Could this be considered to bound or constrain errors for each region/dataset?

Corer data were not used to build the regression model. So, those biases would not affect the regression model coefficients. Any depth measurement that has a bias or random error and that is used to estimate a SWE value using the methods in this paper would propagate through into a bias or error in the SWE. We do try to present some discussion on this in the manuscript.

Line 185: I would expect readers to be unfamiliar with some of coring devices. For example, the Mt. Rose snow tube could be supported with Church, J. Improvement in Snow Survey Apparatus, TAGU, 1936.

Thank you for this suggestion, and we can add this citation.

Line 228: See concerns about outlier detection in the main comments. It would be important to describe the temporal aspect of the outlier detection.

Yes, as noted above, we provide information on this now in that section.

Line 228: uncleaned data -> source data

Good catch, we made this change.

Line 229: State how many outliers were removed from the other datasets via this process. Figure 4: An axis label is needed for the DOY color bar.

This has been handled with added parenthetical notes to column 4 of Table 1.

Line 231: How does this work for ‘stations’ where there are a very low number of observations, ie AK?

The process was objectively applied to all stations. Stations with low numbers of observations could still be processed, in terms of computing the characteristics of the bivariate distributions and then removing points that did not satisfy the criteria.

Table 1: Can this table be augmented with a % of retained points or an omission %? Is the BC survey missing the # of ultrasonic sites?

As per the remark just above re:line 229, yes we have done this. Regarding the BC comment. The first row of that table has two sets of numbers. One for the Western USA SNOTEL. One for the eastern USA SCAN. The BC row only has one set of numbers since we grouped all BC snow pillows together. In revised Table 1, we have split up the USA NRCS data into two rows to eliminate this confusion.

Line 250: Is this 50% of all measurements or 50% of each subset. If it is all of them, it could be such that the only ones removed are CONUS because of the low numbers elsewhere.

All of the aggregated snow pillow data were grouped (data points were grouped in one large bin) and then divided in two. Given the random nature of the division, each station should have ~50% of its data represented.

Line 256: Figure 3 is used as support for the outlier detection due to poor correlation (ie increasing h with no SWE) and but is referenced here as strongly correlated. It might be confusing to do both.

This is a good catch. We meant to refer to just the winter (snow present) portions of Figure 3. The noisy bits in that Figure are at times when there is no SWE. We will clarify our language.

Line 283: If this is an important consideration, why is the SCAN dataset not used to train the models?

There are several reasons. Foremost, we wanted to leave the northeastern USA data alone so that we could use those data as an independent test of the ability of the model to work in completely different regions / snow regimes. Second, the N (5 sites) of the northeastern USA dataset is a tiny fraction of the rest of the available data. Locations with multi-peak SWE curves may do better with a more complex model that is able to capture this behavior.

Line 290: Interesting that a static 180 works best as the DOY separator. Could a sentence on why this might occur be added to the discussion?

To be frank, we do not have a great explanation for this. When we discovered a fairly strong correlation between day of peak SWE and April temperature, we were confident that the variable DOY approach would produce the best results. In this case, it appears that simpler is better.

Line 332: I see how it would not be possible to use an absolute value here but are snow-covered regions where the February normal is below -30C.

We chose this offset value based upon the lowest February temperature values observed at the snow pillow stations. This may limit our methods to not apply in some extremely cold regions.

Figure 6: Titles for each plot might make this easier to read if someone skips the caption.

We appreciate this stylistic suggestion. Our approach favors using the figure caption to provide details on the content in each figure panel, which is consistent with the approach of other papers in The Cryosphere. We are open to modifying this if the editors request it.

Table 5: Include the normalized errors for completeness of the table.

We are not able to normalize the errors for these datasets in the way that we do for the snow pillow sites (Figs 8-9 of new version of paper). For the snow pillow stations, we normalized the RMSE at each station based on (this is a change for the 2nd draft of this paper) the mean annual maximum SWE at that station. The information in table 5 is different. The RMSE values there are essentially being averaged ‘spatially’ over a distributed dataset, rather than being averaged temporally at a snow pillow station. Thus, we do not have a mean annual maximum SWE available for normalization in a consistent fashion. Note that in the 2nd paragraph of the discussion section, we do talk a bit about the east coast results and how they differ from western North America (smaller snowpack, etc.).

Line 423-430: Might be helpful to discuss measurement errors as a contributor.

We do discuss this (measurement errors) in lines 447 → 472 (numbering of original draft). In the specific context of the northeastern USA data, those data are generally high-quality coring data. Having not taken those data ourselves, it is hard to quantify the measurement errors. In some cases, the supporting documentation for those datasets is brief to non-existent. Also, note that, in response to another reviewer as well, we have added more general discussion of both coring error and snow pillow errors to the manuscript.