

Reply to Matthew Sturm, Referee

Review of Converting Snow Depth to Snow Water 1 Equivalent Using Climatological Variables

February 18, 2019

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

In this paper, the authors address the problem of converting more readily obtained snow depth measurements to snow water equivalent values. The problem is highly topical as airborne lidar and airborne and satellite-based photogrammetric snow depths become more readily available for widespread use. The authors primarily build on the method described by Sturm, Taras, Liston, Derksen, Jonas and Lea (2010), with the main difference in their method being the replacement of *climate classes of snow* by continuous climate variables (*mean annual precipitation and February mean temperature*) obtained from the PRISM data set. Though not explicitly stated, the authors also establish their regressions using a larger data set than the Sturm et al. study and most other studies of which I am aware. They reach the conclusion that their regressions show an improvement over the 2010 work.

As the lead author of that prior depth-to-SWE study, I find this a fine piece of work, clearly and honestly written, and useful to many practitioners. It should be published. That said, I am not sure that I fully agree with the conclusion of the authors as to the extent of the improvement, whether their improved method is more easily applied than the old, and I find the omission of any discussion of the well-known errors in the data set used to develop the regression equations troubling. I would like to see the authors grapple with this last issue explicitly in the paper before a version of the paper is published.

Thank you for this general overall assessment. Below, we provide point-by-point responses to your comments and we indicate where and how we plan to revise the manuscript prior to publication.

Examining the input data for this study (Table 1), 98.5% is essential SNOTEL snow pillow data; 1.5% comes from coring. Both types of data are known to contain biases. My personal experience for the latter (coring) is that it tends to undersample SWE (or produce low-biased density values), and across prior studies, there is agreement the method is no more accurate than about $\pm 10\%$.

We would like to point out that all of the data used to construct the regression model are snow pillow data. Table 1 summarizes the data used to build the model, and also the other independent data sets used to validate the model. We have clarified in the manuscript (beginning of section 2.2) and table (using bold font in the table) which data are used for what purposes.

It has been some time since I worked through the literature on snow pillow data, but I recall significant biases from these instruments as well. One source of error is due to snow bridging with, particularly, low biases during the melt when percolating meltwater can run off the pillow to the surrounding snowpack due to the shape of the pillows. Sonic sounders also can exhibit some measurement errors (in this case the ones near the SNOTEL site paired to the pillow SWE values, chiefly in not being representative of the snow depth on the pillow.

This is an important point. There are some studies¹ that show that SNOTEL sites can report SWE > accumulated precipitation, attributed to drifting snow. However, this would not bias snow density assuming that the SWE and Hs measurements are co-located. There are other studies² that have looked at the measurement bias in SWE depending on whether or not the snow pillow is steel vs. hypalon. One comprehensive study³ of biases notes a complex situation, where SWE is sometimes under-reported due to 'snow bridging', but over-reported at other times (see Table 1 of that paper). While that paper proposes methods for correcting SWE measurements, it is complex in practice, requiring continuous SWE, Hs, and near-ground temperature measurements. Please continue to our next remark below.

Given these potential sources of error, and the fact that the authors are attempting to develop general depth-SWE regressions, they should examine how these errors might cause their results to deviate from the "true" local conversion functions. For example, hypothetically, in a maritime regime, perhaps the natural snow packs retains frequent rain-on-snow water, but at the actual measurement sites it runs off from the pillows. Then there would be a consistent tendency in this February-warm location with high MAP (mean annual precipitation) to have light (or low) SWE vs. depth values. At least describing in what ways the modeled SWE values might diverge from the on-the-ground values would alert readers to limitations in the methodology.

This is a sensible suggestion. In the first draft of the paper, we did investigate the effect of measurement precision. In our revision, we now provide more discussion about potential errors in snow pillow measurements (to help alert readers, as you suggest). One complicating issue is that many studies that report on 'errors' in SWE from snow pillows define this error as the difference between the snow pillow and a coring measurement. The implicit assumption is that the coring measurement is the 'ground truth' but as you note, coring is good to +/- 10%. Given the lack of any consensus information on the distributions of errors in snow pillow measurements (we provide some citations to show the divergence of studies out there), we are unable to provide any good quantitative information on the effects of the pillow errors on the SWE estimates.

¹ <https://journals.ametsoc.org/doi/pdf/10.1175/JHM-D-12-066.1>

² https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs141p2_032059.pdf

³ <https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.5795>

As far as whether this study is an improvement over our 2010 study, it really comes down to which ancillary data set one wants to work with: a gridded data set of snow classes or of PRISM climate data. Each has advantages and disadvantages in terms of computational cost and hassle. Looking at Table 4 which compares our prior work to the new work, most of the statistical improvement comes from the taiga snow class, which, as the authors note (Line 415), is because in 2010 we assigned a fixed value to this class (e.g., a fixed value performed better than regressed values). This snow class was only 6% of our training set, and I suspect the sample we chose tended to be quite “stiff” because of the high percentage of depth hoar found in taiga snow, thus it did not tend to densify due to overburden stress (probably something of an Arctic bias we showed). The authors taiga sample set is deeper with greater SWE.

With regards to relative model performance. In our first draft, we tried to be as objective and factual as possible, in the sense of simply providing the comparative results (both figures and RMSE values). We feel that this is fair and appropriate. We also felt it appropriate to break out results by snow class so that readers could see how the comparison varied based on that. We have added information about how many data points in the aggregated CONUS, BC, and AK dataset are in each snow class (in section 3.1).

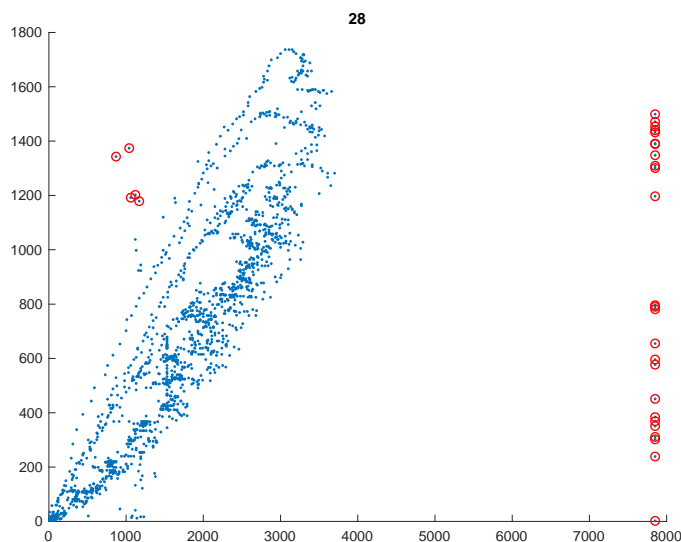
Regarding computational cost and hassle. The Sturm approach uses a straightforward equation and only requires access to the 1km snow-cover raster. Our approach uses more ‘data’ in the sense that numerous PRISM grids are required. However, we have packaged all necessary files into a freely available (will be released on GitHub upon acceptance of this paper) function that is very easy to use. By doing so, we alleviate any cost and hassle concerns.

One last substantive comment: The authors have an entire section on outlier detection and removal, but I would argue they have potentially removed real data. I applaud them for recognizing the hysteresis loop that is produced by depth-SWE seasonal evolution (Fig. 1) and their clever way of handling it in their regressions (Equation 5). We had actually during our work looked at using a rotated lemniscate to model this behavior, but dropped it because we could not make it work right. But if one recognizes that physically the bulk density increases during the melt during the Spring, then one also has to recognize that very early in the winter, deep fluffy snow will be found on some snow pillows....snow with bulk density values of that are less than 150 kg/m^3 . Figure 4 (clean version) has a lower depth-SWE line that at 2000 mm is about 350 mm SWE, a density of 175 kg/m^3 , and a density of 180 kg/m^3 at 3000 mm depth. I believe actual depth-SWE data on the low end has been removed, not erroneous data. Now one might argue we may in general seem to introduce a low bias when we do these sort of regressions, but that is not reason to label what may be accurate physical data as outliers. As further confirmation, the color of the removed data in Figure 4 is mostly blue (early season) and this removal would impact thin climate classes (e.g. taiga) more than thick classes.

This is a great point, and one which the authors have discussed at some length, following your review. 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. There is a lack of clarity and/or consensus in the literature about how to do this. 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.

This figure of the output of the data removal process illustrates things. SWE on vertical axis, h on horizontal. Removed points are in red



We acknowledge that referring to this process as ‘outlier’ detection is perhaps too strong and we have modified the language accordingly, notably re-titling Subsection 2.1.1.5. We also note that Anonymous Referee #3 had a similar comment and wanted to see a histogram of the DOY of the removed data points. We have gone back and looked at the distribution of DOY for all removed points. It turns out that the mean value of DOY was 160 and the standard deviation was 65. So, the bulk of the removed points comes from the middle of the snow season, not at the beginning or the end. This seems to alleviate a bit of the concern that you raise above.

One final comment, and this would be not only to the authors of this paper, but virtually every author out there. Please try to cite the seminal or original papers on a topic if possible...not the newest or easiest to cite. The authors here do well in citing Alford and Church, but when it comes to recognizing how snow depth and SWE are related in time and space, the seminal work of G. A. McKay should not be overlooked.

McKay, G. A., and B. F. Findlay, 1971: Variation of snow resources with climate and vegetation in Canada. Proc. 39th Western Snow Conf., Billings, MT, Western Snow Conference, 17–26.

McKay, G. A and D. M. Gray, 1981: The distribution of snowcover. Handbook of Snow, 1st ed. D. M. Gray and E. D. H. Male, Eds., Pergamon Press, 153–190.

We have added the former citation.

Detailed Comments:

Line 36: Surely someone before 2018 recognized that snow was important to hydrology...like Gerdel (see U.S. Army Corps of Engineers [1956] monograph on snow hydrology.

This citation has been added.

Lines 41-50: This paragraph is a little jumbled and doesn't address some of the well-known errors present in snow pillow measurements (see major comments), yet in the next paragraph, errors in SWE core values, which may actually be smaller in some cases, are identified. Perhaps here is where errors in the input data could be discussed in greater detail.

Yes, we have reworked the introduction a fair bit to bring in some information upfront about errors in coring and in snow pillows.

Lines 60-66: This little sections seems uneven, and given the huge literature on trying to extract SWE from remote sensing, particularly radar, very one-dimensional. Why even talk about snow remote sensing in the paper? I would simply say if falls outside of the scope of the work....and if there really is a reliable operational way to get SWE now from space, I don't know it.

We were trying to be comprehensive in laying out all of the options (in-situ vs. remote) for acquiring snow information. Your suggestion (huge literature that we don't do justice to) is on point and we have removed this section from the paper.

Line 74: Again, Goodison did the seminal work on the sonic sounders. Perhaps you could cite him.

This is a sensible suggestion and we will do so.

Line 96: I think this citation should be: Jonas, Tobias, Christoph Marty, and Jan Magnusson. "Estimating the snow water equivalent from snow depth measurements in the Swiss Alps." Journal of Hydrology 378, no. 1-2 (2009): 161-167.

Correct. We had it right in the references, but incorrect in the in-line citation.

Lines 117-118: I do not agree that *a priori* complexity produces more accuracy. What is really going on here is that proximity to high quality input data tends to produce better accuracy. But that may be true whether the model used with the data is complex or simple. Basically, in a very heterogeneous snow world, when we have local driving data, the results regardless of the model, get better. One might even be able to argue, given the difficulties of measuring radiation in snowy locations, that energy balance models can introduce errors. I don't think you need to work so hard on making a case for the type of statistical approach developed in this paper. Ease of use, and generally the lack of driving data most places, make the case for you.

This is a fair point and we revised our wording.

Line 282: I like this section on DOY, even though in the end you fix the value to 180. Just the fact that the regressions are insensitive to the DOY of peak SWE is interesting.

Agreed, thank you for noting this. It was an unexpected result.

Lines 336-337: I wish the authors would expand this section. It is the heart of why the regressions work, and it is how this study and our 2010 study are related. Climate classes tell us which snow is warm and deep and tends to densify rapidly; high MAP and high February temperatures tell us the same thing, perhaps as the authors claim, even better (or maybe it is just that the training set being larger is better?) This said, the authors I think are aware that there are several snow packs in which due to development of depth hoar and wind slab, there is very limited increase in density over time. Icy snow too can resist densification.

We have expanded this section somewhat. We fully recognize that all bulk-density methods that rely on simple inputs like DOY or climatological weather characteristics are unable to capture numerous features of snowpacks. That is a limitation of the emphasis on simplicity.

Line 342: Figure 6 is nice and clear.

Thank you. We have slightly modified this to show the data clouds as heat maps (2d histograms, essentially) at the suggestion of another reviewer.

Line 357: The model errors will have NO impact on the local snow regime...I think you mean the impact will be on the predicted results.

Correct, this was poorly stated, and we have reworded this.

Lines 405 – 410: I realize that the authors are fond of their Chugach results, undoubtedly obtained with much effort, but these data constitute 0.004% of the entire ensemble and could readily be omitted, with the space saved a deeper examination of why the systems is working, and where I might fail to work well.

Another referee was also lukewarm on the inclusion of this dataset. We have removed most of it, except for the useful information that it provides on the variability in Hs that is observed over short distances. That is a valuable point to retain.

Lines 431-432: Consider why this is: early in the winter, the addition of new snow to a thin pack makes a dramatic change in the bulk density (e.g., called here noise, but which is real) while later in the winter that noise dissipates because the addition is an increasingly small percentage of what is already on the ground. While a model using historical data cannot adjust for this effect, one could talk about how the uncertainty in the modeled result decreases with time. Does it then increase again after the DOY of peak SWE?

These are good points. Yes, our model is using only climatological weather data, which know nothing about individual snowfall events. We have added remarks on this issue and we have added a new figure that shows the errors as a function of the DOY.

Lines 447-460: This is much too cursory a discussion of precision and accuracy, and it sets up a false strawman: more stations or better precisions? The real question is how do we achieve better accuracy, and by this I suspect we mean better more accurate assessments of snow water resources. Given that 95% of the data being used is SNOTEL measurements, then this question has to start with whether the SNOTEL sites were actually designed to be “representative” or “index” sites....and I believe they were always meant to be the latter. Next it has to proceed to the issue of representativeness, as increasingly as we get depths from lidar or photogrammetry, we will be converting depths to SWE in locations not sampled by the SOTEL network. Are we moving into locations where the bulk density is likely to be higher or lower than at an index station? Why? I would rather see the authors just bypass the issue than trivialize the problem in a statistical experiment that doesn't tell us much about the core issue.

We have removed the commentary on whether or not future investments should be in more stations vs. better stations. We agree that SNOTEL stations are largely index stations in that their measurements are often directly regressed against downstream streamflow. However, for our purposes here (providing an equation to estimate SWE from h ‘anywhere anytime’) we do feel that it is valuable to discuss the effects of source data accuracy and precision on the estimated SWE. This will help the reader to understand how much uncertainty there will be in SWE derived from the current paper. So, we retain some of the statistical testing.