Wagner et. al., 'Snowfall and snow accumulation processes during the MOSAiC winter and spring season'

The authors introduce novel work completed as part of MOASiC to constrain snow mass balance and relate measured SWE (on the ice) to precipitation estimates from sensors and reanalysis on sea ice. A detailed introduction is provided defining contributions to snow mass balance along with examples of previous efforts to constrain these terms in terrestrial and sea ice domains. Strong covariance between snow depth and water equivalent is leveraged to characterize spatiotemporal change in SWE along two transects on second year ice. The depth-to-SWE relationship is built from a combination of traditional corer and SnowMicroPen estimates of bulk SWE. Blowing snow events, as indicated from particle counters, and derived horizontal mass flux, are used to delineate accumulating from drifting snow in the transect SWE data and identify periods for process attribution. Cumulative precipitation estimates from a range of sensors are compared against the transect estimates of SWE to estimate retention on the ice against erosion terms. It is demonstrated that between 54 and 68% of precipitation is lost to erosion in terms of cumulative SWE.

The introduced dataset is one of a kind and provides a critical reference for a broad range of Arctic science. I would like first to congratulate the authors on this ambitious work which will be of benefit to many research communities. I have identified areas where I felt strongly that clarification was needed to validate the measured SWE dataset as a reference. Overall, the work described in this paper adds novel information to inform our understanding of the snow mass balance on sea ice. I hope my comments are not misinterpreted as I view the underlying work as a good contribution, that with refinement, would see uptake. Thanks to the authors for their efforts to bring novel snow on sea ice and precipitation information to Arctic science.

Josh

Reference Dataset Methods: I struggled at times to keep track of the reference dataset methods, where for example, modifications were made within the results section to the accumulated SWE data (ie. P25 L524; P25 L532). There were also several datasets elements that appear to be strong support for the reference validation which are introduced but not used (ie. SMP measurements in known drift locations P9 L193, P10 L233 to delineate areas where the HS-SWE regression might fail; Co-located snow pits with SMP profiles P7 L190 that could be used for validation). My general feeling after several reads was that consolidation of the methods along with tabular descriptions of final reference data (number of observations + descriptive statistics + dates) would provide support for the precipitation analysis and improve flow.

HS-SWE Conversion and Evaluation: The conversion of snow height to SWE follows an assumption that covariance with bulk density is generally weak as compared to height. To exploit this the regression in Eqn. 6 show mixes coefficients from two sperate fits (SMP and corer derived). Given that this is an atypical statistical approach I would like to see a clear justification for the use of an arbitrary function (ie not fit to any specific dataset) instead of a validated fit from a combined SMP + ETH height/SWE dataset. Additionally, I was a bit surprised that previous work construct HS-SWE relationships on Arctic sea ice were not contrast against the methods introduced here (ie. SWE = 0.348 * hs; in CM units from SHEBA doi:10.1029/2000JC000400). Using the heights from the public SMP derived dataset from this paper, general agreement can be demonstrated between the SHEBA experiment derived conversion and the one introduced here. Evaluating against the ETH + SMP derived dataset would be of interest. This would particularly interesting given that the ETH + SMP dataset covers an entire accumulation season, suggesting seasonal and spatial bulk density are well constrained.

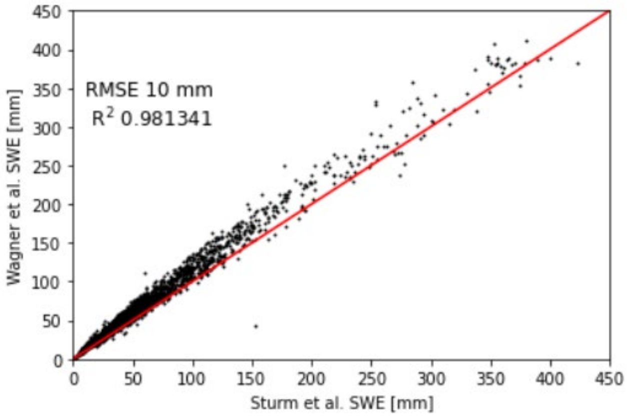

Figure 1: SWE as derived from Eqn 6 in this paper and Sturm et al 2002 based on the SMP height data. Red line indicates 1:1.

Additionally, an improved explanation of how RMSE was used to determine applicability of the SMP coefficients would be beneficial for the discussion of retrieval skill. RMSE as shown in figure 4 and described in text (P11 L288) appears to describe residuals related to Eqn. 6 rather than retrieval skill associated with a specific set of coefficients. A straightforward opportunity for commenting on skill would be comparison against the collocated snow pits where 5 SMP profiles are noted as available at each (P7 L190). Without comparison against a known reference (ie snow pits) it should be made clear the RMSE characterizes the fit to the regression and not explicitly skill of the SMP derived density.

Comparison with precipitation information: Temporal reduction of the observed SWE dataset to only dates where both loops were available is justified as improving how representative the dataset is. I would like to see an improved example to support this. At times the comprehensive datasets introduced as distilled to very few aggregate comparisons (ie P25 L529 n = 9) but no discussion is given on the impact of limiting the temporal steps.

<u>Specific Comments</u>

P2 L29: Can you define what small and large-scale are in terms of area or length scale? Doing so will would help to frame the scales for the analysis and anticipated processes.

P2 L53: Similar to my last point, a definition for what a ‘larger area’ would be helpful to frame this.

P3 L127: ‘snow pack’ and ‘snowpack’ are both used in the manuscript.

P6 L167: Using the word direct to describe the ETH measurement might be confusing. To me, direct would imply melting down the snow and measuring it in graduated cylinder. I certainly leave this up to you, but I tend describe these in general terms as snow cores.

P6 L167: Additional information on the ETH methods would be helpful to understand the precision/accuracy of the bulk SWE estimates. For example, what type of scale was used (& with what uncertainty), what were the uncertainties associated sample capture (was it hard to get full samples in shallow snow with snow ice?), were measurement replicates were made at each location (how were outliers detected). This is not something that needs extended discussion but if the cores are to be considered baseline (instead of snow pits) I would like to understand the expected accuracy.

P7 L190-194: Can a table be provided to understand the number of SMP measurements by date and site for quick reference? Is not clear in text where and when the data points come from.

P9 L215: Are these measurements simply paced out or was there was distance reference used? Average distance between measurements here is less than the accuracy of the GPS (~5 m) so it might be helpful to provide this information.

P9 L216: Can you elaborate on what 'tip sinking into the ice' means? Mangaprobe users often describe the issue of 'overprobe' but in terrestrial domains this is unwanted penetration into the soil or vegetation. What is the interface that is being penetrated? How was it decided that a 1 cm correction was appropriate for all measurements?

P6 217: More information is needed on this error detection step for the magnaprobe. The magnaprobe is generally known for good repeatability (mm-scale). What is the justification for the 3-sigma approach? Given that snow depth on sea ice distributions are generally log-normal if feel like this could lead to omission bias for lower-frequency deep snow.

P10 L231: When I read this I felt like it was saying that the SMP was easier to use than a traditional corer. The SMP is a great tool but I'd strongly argue against it being easier to use then a corer and hanging scale in harsh conditions. They seem to be complementary datasets here where both are considered as a bulk value, by increasing N and thus providing confidence in the regressed coefficients.

P10 L233: This line indicates availability of both measurements in drifts and at pits. Why is this information not leveraged to demonstrate layering as a non-issue for drifts (not saying it isn't!) or provide a well-known reference to validate against?

P11 Eqn (3). Proksch et al 2015 describes median force (F bar) within a defined window as input but it is not indicated here or in the following text.

P11 L253: See major comments. Its not clear how the mixing of coefficients from two different regressions from two different datasets is justified over fitting the full dataset.

P11 L259: See major comments. Additionally, there is context as to why the original SMP-density coefficients may not be appropriate (using SMPv3 vs SMPv4 hardware) that is not noted here. I'm concerned that RMSE as used here is not an indicator of SMP skill while the text reads as such.

P13 L267: I don't agree that its beyond the scope of the measurements to distinguish relative depth hoar and wind slab contributions to SWE. It's a demonstrated capability of the SMP. Please rephrase this to indicate that it is beyond the scope of the study. Examples of SMP based layer classification with Random Forest (10.1109/TGRS.2012.2220549) , SVM (10.5194/tc-14-4323-2020), and ML methods (10.5194/egusphere-egu21-15637)

P13 L271: Its not clear why GPS is a critical consideration here if SMP profiles are noted as located in direct proximity to the snow pits (L273). Please revise this sentence to indicate that this comparison was not completed or provide a justification for needing cm-scale precision.

P14 P287: There appears to be sufficient data to demonstrate this quantitatively. I feel strongly that analysis would be strengthened with a brief assessment of skill rather than a qualitative assessment of Figure 5.

P20 L435: Could you provide more detail on why the loops are not considered separately and are immediately averaged? The temporal record is also greatly impacted by the averaging, does this not potentially minimize the potential comparison periods? This needs to go beyond stating that averaging the two loops is 'more representative' and demonstrate why.

P20 L441: Could the periods selected based on this criterial be highlighted on one of the time series plots. I'm finding it hard to visualize the temporal frequency and range of compared periods.

P21 L451: See major comments. Its atypical to add processing steps in the results, removed from the original dataset description. Consolidation of methods to develop the reference dataset may improve flow and reduce overall length.

P21 L459: The statement regarding 'slight' changes is quite vague. Consider removing qualitative statements and refining the description to provide clear quantitative results where the data is available to support.

P24 L497: Is it possible to frame the statistical significance for mass change in terms of % lost. What was the mean SWE in this small transect relative to the 12 mm of change? From the diagram itself it would seem new drift structures formed but its challenging to determine total loss. Histograms may also be useful here to show the change.

P34 L660: This is an important statement and one that I am a bit surprised is not in the site description. If the evaluated sites are catch for larger domains of the site this is information that can be included in the original site description to indicate the presence of significant ridges outside the observations that potentially act as catch.

P36 L767: The terminology 'snow surface roughness' has not been used before in this study. Do you mean variability in surface height of the snowpack?