The authors would like to thank the referee for the prompt and precise comments to our reply. We want to re-emphasize that the purpose of the study is to investigate the current status-quo of the community in snow cover reconstruction and retrieval for Arctic sea ice. The variety of the methods used in the products, as well as the potential lack of independent data pose unique challenges, but we consider this a timely update of the community's efforts, and the collateral trade-offs during the analysis necessary at this stage of research progress. Moreover, the general consistency among many of the products reflect the exciting progress that has been made, and provide confidence for further improvements of snow depth retrieval. Also, messages are conveyed relating to both observation and representation issues. In specific, with respect to the comments from the referee, we have made replies and accompanying revisions as follows.

The referee's comments 1:

Review of Inter-comparison of snow depth over sea ice from multiple methods by Zhou, Stroeve and others.

In this paper the authors explore how well various methods of determining the snow cover (mostly depth) across the sea ice of the Arctic Basin match each other, and where possible, match in-situ data. The methods consist of a) in-situ depth values from two ice buoy systems, b) IceBridge airborne measurements from radar, c) the Warren et al. (1999) and newer climatologies based on Russian measurements, d) satellite retrievals and e) physical models driven from reanalysis products. The authors find similarities between products in some geographic areas and at some periods of the winter, but also large discrepancies in both absolute values and patterns of depth. This comes as no surprise given the diversity of the ways in which the snow fields are derived, and the limitations within the various methods.

This is a useful paper, and quite an impressive job in merely handling the massive data sets involved, but it does not go far enough in examining the "whys" of the discrepancies, and for that reason I think needs another round of revision. To simply state that radical differences in footprint sizes and spatial scales of the various snow depth products are at the root of the observed differences, while perhaps correct, seems of little practical use.

Part of the problem with the paper seems to start with this statement, given as a motivation for the study: (line 48) [to] provide an inter-comparison of these products so that recommendations can be made to the science community as to which data product best suits their needs. I think I understand what the authors mean here, but it is not what they actually wrote. One would hope that the science community is not only looking for products that suit their needs, but in fact, are as accurate as possible. So while we all can accept that with footprint size and scaling issues, snow depth "truth" may be elusive, conceptionally it exists and it really is what users, and snow product developers, need to strive for. This murkiness in purpose reappears throughout the paper in that none of the models or methods are ever labeled as "wrong," even when the results seem to be utterly improbable. I understand no one wants to "slag" a model in print, but clearly one conclusion from the paper is that some of the models, in some situations, should be avoided for now (until improved) and the authors could be more explicit in saying so when that conclusion is clear (e.g. The PMW-DMI model has 13 cm as the end-of winter average across the entire Basin: Fig. 3 left).

Reply:

The authors understand the urge for the 'accurate' snow product for the community. The motivation of the paper is to investigate the status and (to try) to attribute the discrepancies among current snow products constructed from multiple methods, and further validate against available observations. However, as discussed in Section 4.3, due to the resolution differences of the various snow depth measurements, there exist distinctive representation issues when local snow depth measurements are used. Therefore, it is hard to judge which is the “true” snow depth based on limited validation datasets.

Although it is impossible to pick one “perfect” snow product, we do find outliers. For example, the positive trend in DuST snow depths during ICESat and CryoSat-2 periods is caused by the limitation in processing the calibration with OIB in the product, which is inconsistent with other products. In spite of lacking thorough analysis of independency and representative issues, we pointed out “unreasonable” performances in some products, including low correlation with OIB in UW product and no correlation and small snow variability in the PMW Bremen and PMW DMI product and buoy comparison although existing representation issues. The above four products are therefore not recommended.

According to the comments, he revised paper further emphasizes the choices of the candidate products with consistent performance according to our analysis. Regarding the manuscript, relevant revisions are highlighted in yellow (Line 311; Line 337; Line 361; Line 388; Line 542).

The referee’s comment 2:

Before getting to the heart of my review, I wanted to raise a point that I am not an expert in but I believe is important. In any inter-comparison of models, the comparison will be skewed when the models are being forced by different reanalysis products, with the precipitation forcing, notoriously difficult to get right in the Arctic, varying a lot in both time and space. How did the authors sort out input reanalysis precipitation differences from model biases. It seems to me that the models need to be run on the same input.

Reply:

The authors are working with data products as provided by various contributors. For the four reanalysis-based products, contributors use different reanalyses as input. Furthermore, the methodologies differ a lot as well. Regarding reanalyses, a comprehensive assessment of available products was made recently by Barrett et al. (2020). They found inter annual variability in reanalysis precipitation was consistent among the different products, but that MERRA-2 was wetter than ERA-I or ERA-5. Spatial patterns are however consistent. Further, Stroeve et al. (2020) also found that running SnowModel-LG with ERA-5 and MERRA-2 gave snow depths within a few cm of each other. We anticipate that the differences between the various reanalysis-based estimates has more to do with how snow processes related to sea ice are considered, such as the snow accumulation or loss and physical process, rather than the reanalysis product used. The distinctive snow initial conditions between SnowModel-LG and NESOSIM is another example of hurdles in aligning these products. More aligned comparison of the reanalysis-based products, which is suggested by the referee, is beyond the scope of this study, including attributing to uncertainties (or differences) to driving datasets and methods. This would potentially require another round of intercomparison with much higher coordinated activities across the contributing institutes.

The referee's comment 3:

I found the results section of the paper quite good, and the graphics, of which perhaps there are a few too many, both useful and explanatory. If some shortening were to occur here, I would suggest the spatial trends in Autumn (Fig. 1), and the temporal trends over the period of record (Fig. 5) could be deleted. Autumn is a tough time to model or map snow, since open water (no platform) determines snow depth to a large extent, depending how it is handled in the model or with the satellite products. Figure 5, while interesting, is a dangerous figure to put out there, as it is likely to be cited as depicting real changes, when in fact the point of the paper is that the models and methods need significant improvement, as attested to in Figures 6, 7, 8, 9, 10 and 11. Personally, I would remove Figure 5. But if it is to be left in (see the next paragraph for this same point), then it would be more useful to explore why all of the models/methods converge in getting increasing snow depth in the Greenland-Canadian sector, while the Russian-Siberia sector is losing snow then to just put it out there as "trends." Can this convergence be traced back to some basic climate variables that enters all of the various models? What drives the convergence?

Reply:

Thank you for referee's suggestion. Figure 1 is removed from the manuscript, and moved into the supplementary material as Figure S1. Further explanation is added about newly Figure 7 (denoted as Figure 5 in old version). As mentioned in the paper, snow depth decreases over Eurasia as a result of delayed freeze-up while the increasing over Greenland-Canadian sector from the three reanalysis-based products (SnowModel-LG, NESOSIM and CPOM) use precipitation (or snowfall) from reanalysis. Existing studies of reanalysis-based precipitation in the Arctic indicate more snowfall and ensuing accumulation, hence thicker snow. Serreze et al. (2012) found that in summer and early autumn, the precipitable water from MERRA, CFSR and ERA-Interim both show positive trend in the Canadian Arctic Archipelago, and Stroeve et al. (2020) suggests that widely open water may increase winter precipitation and further affect snow accumulation. Further, the snow accumulation process largely depends on ice motion fields and the above all three products (SnowModel-LG, NESOSIM and UW) use NSIDC ice drift, which may potentially cause similar pattern in snow depth trend.

The referee's comment 4:

And that type of analysis is what I think the paper currently lacks. For example, starting with Figure 3, depth histograms offer a wealth of interpretive information if examined closely. SnowModel and NESOSIM show a distinct drift shoulder (deeper snow) in Autumn that indicates they allow the snow to drift immediately, while CPOM and DuST show little or none. So this feature is about drifting: is it real? Should it be there? Strangely, the drift shoulder disappears by Spring for SnowModel and NESOSIM, yet appears in UW and DuST. Therein lies some behavioral differences that could provide insights for model/method improvement.

Reply:

In general, the authors want to explain that the observable modes in the PDF is caused by snow accumulation on different ice types. The shoulder shape in Figure 2 mentioned by the referee is the second mode in PDF. This bimodal PDF is more obvious in Figure S3 both for SnowModel-LG, NESOSIM and CPOM as a result of different snow accumulation over FYI and MYI. MYI may accumulates initial snow cover early in autumn, while FYI won't be able to catch any snow until it forms. Another minor issue in the comparison is that, the shoulder shape in Figure 2.a is due to the smaller common coverage area compared with Figure S3. The common coverage without DuST (shown in Figure S2) covers more over Canadian coastal regions, where MYI manifests. All spring snow PDFs in Figure S3.b show the bimodal or long-tail features, which is also found in other observations (Kwok et al., 2011).

The referee's comment 5:

Figure 4 offers similar analysis possibilities. A lot is known at least at local scales about the patterns of snow build up on sea ice in various locations. The slope of these seasonal trajectories, and whether they curve off late in the season or not, is a diagnostic that could readily provide insight into what is or isn't working. The W99 curve is suggestive; only the CPOM curves seem to carry that shape, yet these curves fail to reach reasonable depth values by Spring. That should tell us something. Finally, considering Figure 10, which is fascinating, the paragraph (lines 426-431) discussing it barely scratches what could be gleaned from the data. Not only is W99 significantly deeper than the model/method products, in some cases by more than 2X, but also the histogram shapes are so different as to look like they are from totally different fields. SnowModel, DuST, and DESS all have a zero snow fraction, while the others do not. PMW-DMI is as peaked as the W99 data, but is about 1/3rd as deep. Surely, contained in this plot, which took considerable effort to develop, are many useful suggestions for model and method improvement (most of the same comments apply to Fig. 11 and SS18), but to get to them requires thinking about why the histograms have the shape they do. In the old days, a lot of emphasis was place on interpreting skewness and kurtosis, and that literature might be of use here.

Reply:

The authors have included more analysis of the seasonal cycle in the manuscript (now a dedicated section, Sec. 3.2). Regarding to the comment on the specific details of intercomparison, the authors agree that the seasonal curve shape in CPOM is similar with that in W99, however, SnowModel-LG, NESOSIM also has the similar curve shapes in some years, especially from 2011 to 2012 and from 2015 to 2016. However, the interannual variability of this seasonal shape varies widely in all reanalysis-based products. As sea ice freeze-up has happened later and later over the last 40 years, it is expected that the snow accumulation during the early stage of winter is no longer similar to W99.

As in Figure 3 and S5 (which are Figure 8 and 9 in the previous version of the manuscript), the authors apologize that the previous result of the comparison between SnowModel-LG and W99 was not confined into Arctic basin, which is not the same region as the other products. These two plots are updated into the paper and the discussion part is revised. Figure 3 basically shows the results of Figure 2 but for different regions. In Figure 2, common regions only include Baffin Bay, north of Barents and Kara Sea but in Figure 3 and S5, only the Arctic basin is included (as shown by W99 in Figure 1). Within the Arctic basin, thin snow (less than 10cm) is observed in SnowModel-LG, DuST and DESS, while NESOSIM, CPOM, UW, PMW Bremen and PMW DMI show deeper snow packs. Other studies have indicated that snow depth is decreasing over the last 30 years (Webster et al., 2014), and that the snow depth since 2000s is thinner than in W99 and SS18. The majority of the snow products are consistent with this decreasing snow depth, but their mean values and slopes differ.

As measurements in W99 are mainly over the western of Arctic, where more of the ice was MYI, SS18 is adopted for providing more details over eastern Arctic, where FYI dominates. The shapes of W99 and SS18 PDF are different as a result of spatial coverage and the inclusion of FYI. Since the ice is getting younger, the histogram shapes of the various snow products are expected to differ from W99 and SS18 especially after 2000. Figure 3-4 also provide snow depth estimations during 1980s or 1990s from reanalysis-based products, suggesting how snow depth has changed in each product over the longer time series. The mode in SnowModel-LG is decreasing while that in UW is still unchanged over the 40 years. The authors agree that the comparison with climatology is helpful for model improvements and we have revised relevant parts, highlighted in red (Line 320; Line 325). As for the skewness and kurtosis of snow depth distribution, Kwok et al, (2011) found that the snow depth is left skewed from OIB and snow depth from ICESat-2 and CryoSat-2 (Kwok et al., 2020) also shows the left skewness especially during early winter. PDF shapes from the snow products and OIB observations is quite different from W99 and SS18.

Summary comments:

In summary, I recommend this manuscript be returned to the authors for revisions without acceptance, that they be commended for undertaking a very useful and difficult set of analyses, and that they be urged now to reap the rewards of that analyses by gleaning more pertinent and useful information from their work. That that could prove very useful in improving existing and future models and methods of extrapolating snow over the Arctic Basin.

Reply:

The authors sincerely thank the referee for the comments. Revisions to the manuscript have been made to: (1) restructure Section 3 to 5 for better clarity, with each subsection covering an aspect of intercomparison, and (2) include more analyses of the intercomparison results, especially for the consistency of the products and PDF of snow depth and comparison with climatology, and (3) explicitly picking out the consistent products and outliners. We sincerely hope that through these revisions, we can convey more clear and informative results.

Reference:

Barrett, A. P., Stroeve, J., and Serreze, M. C.: Arctic Ocean Precipitation from Atmospheric Reanalyses and Comparisons with North Pole Drifting Station Records, Journal of Geophysical Research: Oceans, 2020.

Kwok, R., et al. Airborne surveys of snow depth over Arctic sea ice. Journal of Geophysical Research: Oceans 116.C11, 2011.

Kwok, R., Kacimi, S., Webster, M. A., Kurtz, N. T., & Petty, A. A. Arctic Snow Depth and Sea Ice Thickness From ICESat-2 and CryoSat-2 Freeboards: A First Examination. Journal of Geophysical Research: Oceans, 125(3), e2019JC016008, 2020.

Serreze, M. C., Barrett, A. P., and Stroeve, J.: Recent changes in tropospheric water vapor over the Arctic as assessed from radiosondes and atmospheric reanalyses, Journal of Geophysical Research: Atmospheres, 117, 2012.

Stroeve, J., Liston, G. E., Buzzard, S., Zhou, L., Mallett, R., Barrett, A., Tschudi, M., M. Tsamados, P. I., and Stewart, J. S.: A Lagrangian snow-evolution system for sea-ice applications (SnowModel-LG): Part II - Analyses., Journal of Geophysical Research: Oceans, in revision, 2019.

Webster, M. A., Rigor, I. G., Nghiem, S. V., Kurtz, N. T., Farrell, S. L., Perovich, D. K., and Sturm, M.: Interdecadal changes in snow depth on Arctic sea ice, Journal of Geophysical Research: Oceans, 119, 5395–5406, 2014.