

Author response to Anonymous Reviewer #1 on: “The NASA Eulerian Snow on Sea Ice Model (NESOSIM): Initial model development and analysis” by Alek A. Petty et al.

Reviewer comments are in black, our responses are in blue.

We will also submit the revised manuscript and a word document highlighting the tracked changes we have made based on these comments.

The paper describes the development, calibration and validation, and sensitivity of a snow on sea ice model. Given that the model is likely to be used to retrieve ice thickness from CryoSat and ICESat altimetry, the model should be documented in the literature. However, the paper needs to be improved and some points clarified before it can be published. I would suggest that the authors consider reorganizing the paper to make it more readable and potentially shorter.

We sincerely thank the reviewer for providing this review. It included some very helpful suggestions on how we can improve the manuscript, and areas where we need to provide more justification etc.

General Comments

While I recognize that a model cannot include everything, I am surprised that the authors do not include snow melt. I see this as a major shortcoming in the model. The top-left panel of Figure 3 shows, what might be interpreted, as a melt signal between March and April. I suspect that melt is also a factor in the North Atlantic sector. A warming Arctic is almost certainly likely to have melt earlier. This years warming ‘spike’ likely caused melting in some sections of the Arctic. While this might not have resulted in a loss of snow mass, it would have increased snow density and caused a reduction in snow depth - refreezing/metamorphosis is another process. The authors should address in more detail leaving these key processes out.

We have added more details in the summary of planned future model improvements, including the lack of snow thermodynamics (highest priority). We feel this discussion combined with the comments in the introduction make clear we see the lack of melt processes as the main shortcoming of the model.

We have also added the following to the end of the abstract: ' Potential improvements to this initial NESOSIM formulation are discussed in the hopes of improving the accuracy and reliability of these simulated snow depth and density.'

Regarding the decrease in the observed snow depths in the Soviet Station data: we actually doubt this is due to a melt event. Air temperatures remained below freezing for all of April for each station in 1980-1991. None of the stations showed a decrease in snow depth except for 1 (Station 31) in 1990 – the air temperatures rose to -4C, which could potentially have been a melt event given that it’s a daily average. Looking at the sea level pressure for that time, there was a big decrease between April 10 and April 30 (the time when the decrease in snow depth occurs) – which suggests there may have been a storm event, which likely redistributed the snow and may have contributed to a decrease in mean depth along the survey line. A storm event would also explain the decrease in density (e.g., fresh snowfall).

Another point to consider is that we only took data from four stations (1980-1991). Despite some of them lasting for multiple years, this is still a relatively small sample size, and the mean in snow depth will be strongly affected by variations in a single data point. We refer the reviewer to

Warren et al. (1999) which shows data from all stations (1954-1991) and thus is more representative of the seasonal cycle.

Another concern is the use of the Polar Stereographic grid. The model tracks snow volume but only sea ice concentration, and not area, appears in Equation 1. Are the authors assuming that Polar Stereographic grids are equal area? This is not the case, they are conformal but not equal area. Cells at 70 N are about 10% smaller than cells at the pole. Maybe I am missing something here, and maybe this erroneous assumption might not have a big impact, but the authors should satisfy both themselves and readers that the choice of grid does not have an impact.

This is an interesting point, and working with projections/grids always provide some complexity that should not be quickly overlooked. In this case we are using equal areas in our polar stereographic projection, so quantities are being conserved in the model. However the reviewer is correct that the polar stereographic projection does provide a distortion to the real world such that our grid cells will be larger at lower latitudes. We do not see this as a huge concern as this is similar to the issue one always has when using polar stereographic data. We did consider using equal area projection (the Lambert azimuthal projection) but this can introduce potential issues with the vector projections in this non-conformal projection that could offset the potential benefits of less distorted area grids.

An obvious question, given the prevalence of Warren 1999 snow depths in sea ice thickness retrievals, is how different are the results presented here from W99? I would argue that W99 is the current benchmark for evaluation of snow depth products. The authors should include some discussion on this topic. I think it would also be useful to put the uncertainty reported in this paper (~ 10 cm) in the context of thickness retrievals. Ten centimeters is at least 30% of snow depths in the Arctic and similar to inter-annual variability in snow depth. While there is clearly room for model improvements, quality of precipitation fields and other forcing fields also come into play. It would be good to discuss these issues.

Another good point. Instead of simply comparing the results against W99, we believe the more interesting comparison is between W99 (actually the modified W99) and OIB (ground truth) and NESOSIM and OIB. This was similar to the approach taken by Blanchard-Wigglesworth et al., (2018) in their recent model evaluation efforts (published during this discussion period).

We have carried out this analysis showing that the difference between the model and mW99 compared to OIB is similar, but also depends on the chosen OIB product. See the figure attached:

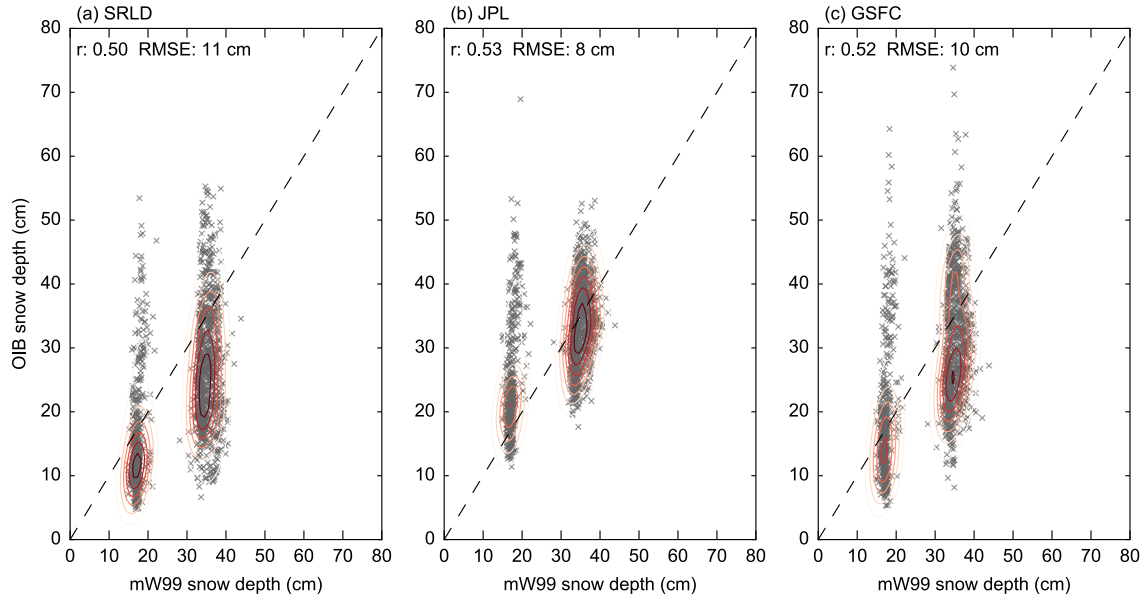


Figure 1: Comparison of modified Warren snow depths against OIB snow depths for the years 2009-2015. The data are binned to the 100 km model grid, only for grid-cells with significant (>1000 OIB data points).

While the model doesn't really offer an improvement over modified Warren in terms of these comparisons, our hope is that with continued model developments and calibrations we can improve the performance and reliability. Uncertainty in the OIB data and the possibility they have been tuned to fit the modified Warren climatology make this interpretation more challenging.

We have added more discussion of these comparisons, and the comparisons to mW99 to the revised manuscript.

I find the model formulation as described in section and equations 1 through 8 confusing. This may be because I think of the process in a different way to that described here. Hopefully, my interpretation in the following paragraph, whether right or wrong, will help improve the model description.

Yes we appreciate these efforts to make the model formulation more readable. As you say, we are really tracking an effective snow depth, and we have tried to make this clearer in the revised manuscript.

I see the model as analogous to the evolution equations for ice thickness (or any other tracer), where the change in snow depth is the sum of a dynamic component ($-\text{del} \cdot \text{dot} (hu)$ [I think] not $\text{del}(hu)$), and a static “snow depth evolution” component that represents snow accumulation, ablation by wind, and compaction ($h^{\text{acc}} - h^{\text{wp}} + h^{\text{bs}}$ for the “new” snow layer and h^{wp} for the “old” snow layer). What I find confusing is that h^{acc} is the product of snow accumulation (snowfall/density) and sea ice concentration. Shouldn't what I call the sum representing the “snow depth evolution” component be multiplied by concentration (A) not just (Sf/density); e.g. $A \cdot (Sf/density - h^{\text{wp}} + h^{\text{bs}})$ [Note my comment in the paragraph above – this only applies to a uniform grid]. Similarly, should h in $-\text{del} \cdot \text{dot} (hu)$ be Ah .

I think our use of effective snow depth was confusing things. As you say we are really tracking Ah , which is why our use of A only really comes in when we introduce the accumulation term (if

we keep using it we double count). The other terms then act on this effective snow depth, so we don't need to apply the ice concentration again in those terms.

Maybe this is what the authors mean by “we track snow volume” – e.g. (Ah) – and “An effective snow depth...”. However, I would suggest that it is Ah that is the effective snow depth because this term represents a mean grid cell snow depth (including open water areas). Whereas, h can be thought of as a physical snow depth because it represents the process of accumulation, wind ablation and compaction at point on an ice floe. I think what I describe is a conceptual difference rather than an error in the model because h^{wp} and h^{bs} are tuned, so the wind packing and blowing snow coefficients can be thought of as including sea ice concentration, i.e. you can change the model description in the paper without changing the model.

Agreed. We have changed this in the revised manuscript to read as: 'Note that in the model we track the evolution of an effective snow depth within each grid-cell (the volume of snow per unit grid cell area) for simplicity. The actual snow depth over the ice fraction is calculated by dividing the effective grid-cell snow depth by the grid-cell ice concentration.'

Two further issues are with model calibration and the use of the ensemble mean snowfall. If I understand correctly, the parameters are tuned using ERA-Interim snowfall and these “best” parameters are applied to model runs with MERRA, JRA55 and the Ensemble Snowfall. I would suggest this is the wrong approach. The calibration process will compensate for biases in ERA-Interim snowfall. However, biases in the other reanalyses are different. A different “best” parameter set should be expected for each reanalysis snowfall product. Conceptually, model equations, parameters and forcing data are all part of the Model. Using parameters obtained for ERA-Interim, might have detrimental effects on snow depths when other reanalyses are used. I would recommend that the MERRA, JRA55 and (maybe) ensemble runs should be calibrated separately.

We did consider this, and the idea makes sense. However, in reality the tuning was not highly optimized and instead we attempted to find parameters that achieved a good balance between capturing the seasonal cycles in snow depth and density in the Soviet Station data. Again, this was mostly due to concerns regarding how representative the Soviet Station data were for calibration purposes. We have added the following to the revised manuscript to explain this:

P17: "We also decided against specific model configuration parameter tuning due to the limitations in the calibration data, however this should be considered when analyzing the model performance, especially with regard to our validation efforts (i.e. more sophisticated and/or configuration specific tuning could improve the comparisons shown)."

This was also discussed in Section 4: " Specific model configurations may be required based on user demands, and our expectations is for these calibrations to evolve as new calibration data are made available and physical parameterizations introduced/updated."

We also added the following to this paragraph: " As discussed in Section 3, it is likely that specific model configuration tuning could improve these comparisons and the later validation efforts, but we decided against a more optimized calibration approach due to the limitations in the Soviet station data."

With regards to the Ensemble snowfall, the assumption behind an ensemble average being a better estimate is that individual ensemble members bracket “reality”. Is this the case with reanalysis snowfall? If all ensemble members are biased in one direction, the ensemble average

will also be biased. My understanding is that both ERA-Interim and MERRA precipitation are both biased high compared to land stations. Based on Fig 11, JRA55 is also high. So is the ensemble snowfall an improvement over the individual ensemble members?

This is a good point and one we perhaps didn't make clear enough in our discussion of the ensemble 'median' snowfall data. We have dropped the idea that the median snowfall dataset might be somehow 'better' from Section 4 ("as we expect these results to be less prone to errors in the individual reanalyses etc.), so this now reads:
"we choose to mainly focus on the MEDIAN-SF forced results using the default configuration (Table 1) for simplicity"

We don't feel that land stations are wholly representative of conditions over the sea ice (confirmed from conversations with colleagues in NASA's GMAO for example) hence our recent study currently in press looking at precipitation estimates over the Arctic Ocean from reanalyses (Boisvert et al., 2018). While again we are limited in direct observations over sea ice, comparisons of the precipitation converted to snow depth (using constant snow density approximations and lagrangian feature tracking) against drifting snow buoys in the central Arctic Ocean showed no obvious bias in the differences between the reanalysis derived snow depth estimates used in this study and the buoy snow depths (in Boisvert et al., 2018). It did appear that MERRA-2 has a clear positive bias so was dropped from our analysis. Clearly more work needs to be done to better understand the precip within the Central Arctic, however.

There needs to be more detail about the ensemble snowfall was generated. For example, I can envisage ERA-Interim, MERRA and JRA55 all having snow but the location of this event being shifted by one or two grid cells. On one side of the event, while ERA-Interim might have no snow, MERRA and JRA55 do have snow. By taking the median, snowfall from MERRA or JRA55, would go into the ensemble product. On the other side of the event, MERRA might not have snow but ERA-Interim and JRA55 do, so snowfall would go into the ensemble product. This would result in a larger region receiving snow. How do you deal with that situation.

The Boisvert et al., (2018) analysis showed that the reanalyses tended to agree well in terms of the presence of a precip event, but differed strongly in the magnitude of the precip during an event. The fact the reanalyses assimilate the same sea level pressure observations means they tend to simulate storms in the same locations, , but they differ in the magnitude and phase of the precip. There was also no obvious suggestion from this analysis that there were significant biases in the location of precipitation events, although the different model grids complicate this slightly. The fact we re-grid everything onto our coarser 100 km grid should help somewhat with this issue. In general, we see this as a crude effort to produce a 'consensus' snowfall estimate on our model grid. Ideally we would have more reanalyses (that provide a direct snowfall estimate) to increase our confidence that the reanalyses are bracketing reality.

With respect to the flow and structure of the paper, sections 4 and 5 seem repetitive, especially where model sensitivity to reanalyses is discussed. Essentially, sensitivity analyses for both periods give the same results. It makes sense (to me) that if you are going to compare the two time periods then the discussion and plots are merged. This would reduce repetition, make it easier for readers to compare the two time periods, and maybe shorten the paper.

We have followed your suggestion and merged these sections, mainly through dropping most of the figures and discussion of the 1980s results. As you say, the differences are not large in terms of the model performance evaluation, and our interpretation of the different figures was pretty

similar. We have kept the budget figures in the SI. We refer the reviewer to our revised manuscript which highlights this change and our justification.

Many of the figures could be improved. In many figures, the colors are not sufficiently distinct, dark purple and dark blue. This is the case with figures where dots are used. Maybe get rid of the black borders to the symbols. Also use different symbols. For many of the line graphs, increasing the weight of lines in the legend and in plots would help. Also consider whether or not you need to show the spread. The overlapping shading obscures the lines showing the means. Using shading works for two, or possibly three, series, especially if they are separated, but it starts to detract from a plot and not convey the information you want it to with more series. For example, the North Atlantic plot in Figure 11: I can't see the lower limit of the JRA55 spread or the upper limit of ASR. In many cases the spread is not discussed in the text. If you still need or want to show spread, you could just show May 1 snow depth spread as vertical bars off to the right hand side of each panel. The issue of including plots in figures but not describing the plots in the text occurs in several figures (e.g. Fig 9). I would suggest that if it not discussed, then don't include it.

We have made several changes to the figures, especially based on the comments below, including clearer legends, thicker lines, better axes. Thanks for that.

The shading appears to be an issue in Figure 11 and 13. I am keen to include the shading as it helps show where the model results may significantly differ from each other over interannual variability. While it isn't always mentioned, the presence of clear differences beyond the spread (e.g. some of the JRA and ASR regional results) is clearer than showing just the means. Showing just some shading for some of the lines would be odd in my view and questionable to the reader. We have changed the axes to prevent the cropping, lightened the shading, and changed the colours, to make this clearer. These figures should also be more colorblind friendly (dropped the green).

The spread as of May 1st is summarized in the Table, which also allows for an easier comparison across the sensitivity studies if the reader struggles to see this in the figure.

Specific comments

L3, P5. "...to avoid complexity of snow melt processes". As I note in General Comments, Fig 3 shows what could be interpreted as melt. I would like to see more justification. Furthermore, a simple temperature index approach could have been used to account for melt.

We refer to our response to the general comment above and our inclusion of more discussion about this in the revised manuscript.

Equation 1. See General Comments.

See response to that comment

Equation 3. Should this be $-\nabla \cdot (\mathbf{h}u)$ Equation 4. Should the divergence be $-\nabla \cdot \mathbf{u}$

Yes, thanks. The code was correct at least! I have changed these equations.

Equation 5. You have a wind speed threshold for wind packing but not for blowing snow. Why is this? Studies for prairie environments indicate blowing snow initiates above ~ 4 m/s, which is similar to your wind compaction threshold.

We agree it does appear odd to not apply this threshold to the blowing snow loss term, so have now added this in. This only had a small impact on the results, but has meant that we had to re-run the model and reproduce all the figures to highlight this small impact. We didn't change any of the coefficients as the changes were so negligible.

Section 2.5. Suggest this section is moved to 2.1 as an introduction to the modelling framework. This sets up the discussion of the parameterizations of the accumulation and sink terms.

We have made this change to the model description, as recommended. We hope this has improved the readability of the model formulation.

L12, P11. An advantage of reanalyses is that they produce consistent outputs. Mixing and matching fields from different reanalyses breaks this consistency. How similar are the ERA-Interim winds to MERRA and JRA55.

Yes true, but as we accumulate snow, the seasonal growth is more important than the daily variability. The study of Lindsay et al., (2014) compared several reanalyses in the Arctic (not including JRA) and showed that ERA-I winds were slightly higher (~0.5 m/s) than winds measured on drifting stations, and MERRA was slightly lower (~0.5 m/s). We decided not to add this to the investigation as it should be second order compared to the other variables and would likely just involve a recalibration of the wind packing and blowing snow loss coefficients.

L17, P11. "We linearly interpolate..." Do you mean bilinear interpolation? See General Comments. This needs more detail.

We use the Python SciPy interpolation package. The linear interpolation uses a triangular (barycentric) not bilinear interpolation approach. We have added a line: 'Gridding scripts written in Python are included in the GitHub code repository' to indicate that the reader can refer to our scripts to learn more about this and reproduce our gridding.

L13, P14. Are significant amounts of snow in summer likely to be present in recent years? The data in Warren 1999 is 30 years old at a minimum. Are there observations from N-ICE or other field campaigns to justify non-zero initial snow depths. Furthermore, how do you initialize new sea ice? This needs to be explained.

The inclusion of more recent summer initial snow depths was also justified to correct a low bias in snow depths when we compared to OIB. We agree there is not much direct evidence to justify this, and the location of the N-ICE campaign will limit its utility here (highest initial snow depths north of Canada/Greenland coastlines).

The initial snow depth is only applied to regions where we have grid-cells with a concentration above 15%. As we track the snow in a given a grid-cell, new ice forms with no new snow, but can accumulate snow instantly. We have added this to the manuscript: 'New ice that forms in a grid-cell is assumed to be snow free, but these grid-cells can accumulate snow instantly.'

L24, P14. "The snow depth is distributed evenly over the old and new snow layers...". Is there a reason why initial snow depth was not just assumed to be dense old snow.

As discussed after this line, albeit briefly, we assumed based on the sparse old observational studies that this was a combination of snow that didn't melt/persisted through the melt season, and

some early summer/fresh snowfall. We admit this is somewhat unconstrained, but it based on the observations in Radionov et al. (1997)

L19, P16. “We carried out initial model calibration...” For this study, you only calibrated the model once, right? Suggest drop “initial” throughout this discussion unless multiple calibrations were made.

True, dropped initial from the text in a few places where it seemed inappropriate.

L22, P16. “...calibration involved manually tuning NESOSIM to improve the general fit...” Was this fit judged “by eye” or was some metric used? Also were all years 1980 to 1991 used, or did you leave a year out for validation during this period. While I recognize that you validated for 2000 to 2015 using OIB data, measurement accuracy and conditions might be different between the two periods.

Yeah this was by eye. We would have carried out a more optimized calibration effort if the in-situ data were more consistent, but as it was, we didn't want to over fit the model to this sparse dataset. We agree with your statement about old and new accuracies but feel this was the best compromise between producing a calibrated model we want to run primarily for a new arctic time period (because of the available altimetry missions).

L14, P17. It looks as if there is a larger difference in snow depth between January and March. Modelled snow depths gradually increase, while observed depths appear to increase in accumulation rate. Is this a shortcoming of the snowfall products. Also, you should mention the decrease in depths in April that could relate to melting and or compaction.

As discussed earlier, the limited coverage of the in-situ data prevents us from saying more about these comparisons (i.e. to be confident of the presence of a particular high or low bias in the model in our view.

L2, P18. It is difficult to believe a correlation of 0.6 for density given the spread of points in the plot.

We checked and stand by these values. The change in axes limits perhaps makes this clearer/more believable! Note that the density of points increases to the upper right which is the main reason for the moderate correlation strength.

L7, P19. “Including the blowing snow loss... but no significant change in snow density.” My first thought here, is why expect any change in density? The only mechanism by which density can be influenced by the blowing snow parameterization is a reduction in the “new snow” depth. So how deep is this “new snow” layer and how quickly does it get redistributed to the “old snow” layer?

It could change the snow density by essentially removing fresh snow (only this layer can be blown into leads) before it gets transferred into the old snow layer. We have added: ' This parameterization can impact the bulk density implicitly by reducing the amount of fresh snow contributing to the total snow depth/density.'

L7 to 15, P19. Maybe add that blowing snow loss in the central Arctic are small because sea ice concentration is close to 100%.

Adapted that line to read: 'As the drifting station data are collected primarily within the Central Arctic where ice concentrations are near to 100%,'

Section 4.2 and Figure 8. I am struggling to make sense of this section. I think part of the problem is that evolution terms are shown as cumulative, which makes comparison difficult: a big snow storm could deposit several 10's of centimeters of snow, dominating the snow depth for the rest of the season. I think you can compare the magnitudes of the terms at the end of the season (May 1) (as you do in the text) but not during the season. To compare terms during the season, I think you need to compare the timestep change in each component. The comparison is not helped by the fact that it is very difficult to distinguish lines in Figure 8. The lines in the legend need to be thicker. I would suggest leaving snow volume out of the figure.

We have tried to make the lines and shading similar, through similar efforts to Figure 11 and 13. I generally like the idea of showing the daily timestep change, although they were pretty noisy and harder to distinguish than just the cumulative plots (a lot of lines crossing over).

Our aim was really to show the cumulative impact as this fits in with the fact we validate the model in spring, at the end of the accumulation season. We agree understanding the model budget terms in all seasons is also crucial but hope to explore this more in follow up work.

L12, P24. Prefer “advected” to “drifting”. For snow, drifting implies blowing snow.

Agreed, changed.

Figure 9. I would suggest showing only the evolution terms that you discuss in section 4.2. Other plots can be put in supplementary figures. While I suggest you don't show the snow volume, note that the units are a depth. Moreover, I think Ah_s (sea ice concentration * snow depth) is better thought of as a gridcell mean thickness.

In response to another reviewer we have added better labels to this figure. We are keen to keep all these panels in as we think readers could be interested in seeing the spatial importance of the various fields. We have changed volume to snow depth, and changed the final panel to 'snow depth over ice'.

L13, P26. Prefer “Soviet Station” to “old” period.

We have actually changed this to the '1980s' time period here and throughout the revised manuscript. The Soviet Station era goes back to the 1950s so didn't want to use that label to avoid confusion.

L13, P26. Given the spread in snow depths in the “New Arctic” and “Soviet Station” periods, are they really that different?

The spread in snow depths based on the input data make such conclusions hard to state and we have tried to avoid this in the paper. We have added more discussion around this issue in response to the other reviewers, but have also tried to remove the focus on comparing time differences as we think this beyond the scope of this initial model evaluation paper in line with your comments of too much regional analyses. As such we have shifted the focus to the 2000s analysis and put the 1980s results in the SI.

L8, P29. Maybe use “difference” instead of “bias” as you have no “truth”.

Agreed, changed.

L2, P30. See General Comments. If Median-SF is biased it might not be that useful.

We refer to our response to the general comment.

L21, P30. “regional variability” – suggest “regional scale”. Regional scale is contrasted with Pan-Arctic Scale.

Agreed. Changed this to ' capturing the variability in snow depth at this regional scale'

Figure 17. Why does NESOSIM have zero snow depths but OIB has non-zero snow depths. It is difficult to interpret the panels with the OIB datasets overlaid. Maybe just show All-years but with separate panels for SRLD, JPL and GSFC OIB products. The individual years can be included as supplementary figures and/or discussed in the text. Also maybe use dots rather than x's to avoid symbols overlapping.

Good spot, we realized there was a small error in the gridding and zero values in the model along the coastline were being included in the binning/regression. We have now removed these values (adding in a zero snow depth mask) which has removed this issue. This is another thing to consider in this comparison as the various OIB products have different treatments for including low snow depths, and some may use cut-offs for higher snow depths. We decided to keep all the OIB snow depths in this comparison (especially as we bin the data up to 100 km), but is something to consider in future validation efforts.

We have also taken your advice and now just show the all year regressions in separate panels. The figures of the individual years are shown in the SI and the r/rmse values are still shown in the table. We also included Kernel Density Estimate contours to highlight more clearly the differences in the distributions.

Technical Comments

Abstract, L9, P1. “Several simple parameterizations to represent key sources and sinks”. The number of processes is not large, so you might as well list them explicitly, rather than keeping the reader guessing :).

Agreed, we have added: '(accumulation, wind packing, advection/divergence, blowing snow lost to leads)' in parentheses.

L22, P3. Suggest “availability” rather than “presence”.

Agreed, changed.

L12, P4. “(Show later)” Give a figure number.

Agreed, added.

L14, P4. “Ice drift”. Suggest “Ice Motion” to avoid confusion with drifting snow.

Agreed we have changed drift to motion here and in several other places in the manuscript.

L8, P5. "...our reanalysis data..." Suggest "...reanalysis fields...".

Agreed, changed.

Table 1. Add symbols for snow densities.

Added, along with the model variables (suggested by other reviewers).

Figure 3. As snow depths and densities are binned, could the data in the upper panels be shown as "box and whiskers" or just boxes. That way readers can see the amount of overlap between depth and density estimates. I suggest you spell out Soviet Stations in the figures. Add 1:1 lines on the lower panels. It would be nice to have a single symbol in the legend.

We have added Soviet Station to the legend and axis labels, reduced number of markers to 1, and added the 1:1 line. We decided box and whisker plots were too much as we already show the mean and spread (+/- 1 SD) and have the raw values below.

Figure 4. Does No Initial Conditions (NO IC) mean the model was initialized with 0 cm snow depth?

Yes! We have added more information to the caption to help clarify the model runs.

L1, P26. Shouldn't this be Section 5?

Yes! changed.

L13-14, P29. Reference needed.

Added reference to Boisvert et al., (2018).

Figure 16. Why two symbols in the legend. Also the colors are difficult to distinguish. Maybe no black border on symbols. Also use different symbols.

We have changed this to just show one marker, made the markers bigger and made the OIB markers squares to more clearly differentiate them from the model markers.