Author response to Anonymous Reviewer #3 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 authors present a new open source model, the NASA Eulerian Snow on Sea Ice Model, for estimating daily depth and density of snow on sea ice. The authors note at a few points in the paper that the model is being developed primarily with application to altimetry-based ice thickness determination in mind, though other applications are likely. The model is a simple representation of the snow that is largely an accounting of snowfall produced by reanalysis data, similar to prior efforts (e.g. Maksym and Markus 2008; Kwok and Cunningham, 2008), with terms for snow compaction, loss to leads, and transport on sea ice. It is Eulerian, but features pseudo transport by exchange between grid cells, features only 2 layers, and is forced with available spatially and temporally complete datasets that are known to be of limited accuracy (e.g. Reanalysis, passive microwave concentration). The model is calibrated/validated against limited available snow on sea ice data from Operation Ice Bridge and from 1980s era Soviet drifting stations. The description of the model is complete and in this regard the model is publishable with minor revisions – but reviewer doesn't feel the model is very good or useful in its current form for its intended purpose. Reviewer focuses most of this review on highlighting its shortcomings. In fact, a possible conclusion of this data presented would be that simple treatment of snow on sea ice will not meet the accuracy levels required for altimetry applications. The reviewer encourages the early career team to put the paper aside for awhile and take the time write a model that would actually be highly used.

We thank the reviewer for putting the time into providing this review. As also highlighted by the other reviewers, the model we developed has provided a useful means of thoroughly assessing reanalysis-derived snow depths and their sensitivity to the input forcing data used. It is hard to develop a highly sophisticated model and explore true sensitivity to input forcing data, and the latter has been extremely lacking in the literature to-date. We believe this study has provided the needed baseline from which we hope to increase model sophistication in future as needed.

The reviewer feels that the key issues are that the model is excessively simplistic, not representative of known physical process (even at the level of simplicity targeted), and that its results show it is inadequate for the intended purpose. There are errors in the equations presented, many compromises appear to have been made that make accuracy and/or realism lower in favor of rapid release, and as a result the work is unlikely to have much impact as presented. The presentation in the paper is quite long, and focuses on trying to convince the reader that the model is good, rather than taking a hard look and comparing against a reasonable standard.

We know of no other pan-Arctic snow depth & density product that provides significantly higher accuracies (e.g. the rmse when compared against OIB say).

Based on the comments of Reviewer 1 we have added similar comparisons of the OIB snow depths against Warren and modified Warren climatologies, highlighting similar/better agreement when using NESOSIM (in terms of correlation coefficient and/or root mean squared errors). We are somewhat limitied by the uncertaintiy in the OIB products (several centimeters) that prevent us from carrying out a more complete validation.

We have also added a lot more justification at the start and end of the revised manuscript regarding our approach and expectations for future model development/calibration efforts to improve these error.s

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 in these simulated snow depth and density'

The development of a snow product for improving retrieval of sea ice thickness from altimetry is critical for ICESAT 2 to be useful and this team should have NASA's support to do just that. Such a snow model's accuracy goal must be based on a desired accuracy in thickness retrievals. (e.g. retrieval of ice thickness accurate to +-0.5m over a given domain demands snow depth accurate to O5cm over the same domain). The model presented here is not up to meeting these kinds of needs, and does not leverage the existing (more sophisticated) models of snow on sea ice (e.g. LIM, SnowModel, CICE).

We are not aware of more sophisticated approaches providing the requirements you list, and our belief is that uncertainties in the input forcing data needed to be explored further, along with efforts to improve model sophistication. We see this model as a contribution towards this end goal. Based on the comments of the other reviewers we have added more discussion of the other snow models available and the physics they include.

We agree that working towards a 5 cm error makes sense, and we believe this provides a useful framework to build towards that goal. This is also discussed in a new study by a co-author of this study: Webster et al. (accepted) which speaks to several issues of our current abilities in observing and treating snow in models. Existing basin-scale observations unfortunately do not have 5-cm accuracy; snow depths derived from merged satellite data do not have 5-cm accuracy. So it is difficult to produce model output with 5-cm accuracy given that no observations of 5-cm accuracy (or better) at the basin-scale exist for model validation and assessment, and especially on a seasonal time-scale.

Some major issues include: Model design relative to state of knowledge: 1. The two layers used in the two layer model (new snow and windslab) do not represent the two layers of the snowpack discussed in literature (wind slab and depth hoar). Authors cite and discuss the literature indicating that windslab and depth hoar dominate the mass of the pack and have quite different density – then ignore these decades of observation to invent a new scheme unsupported by observations. Respecting the effort to create a simple, 2 layer model, new snow should not be one of the two layers. The references cited clearly state that new snow rarely comprises much of the Arctic snowpack, because it is very rapidly converted to windslab. The preservation of a new snow layer appeared to be designed for modeling loss of snow into leads – but little is known about the magnitude of this flux, and it was minor in this model. 2. The model is operated on a 100x100km grid, which is very coarse relative to the variability in ice – which is shown to be important in impacting the accumulation of snow. The data sets used provide much higher ice concentration, and movement information – this data should be used at full resolution and atmospheric data can be downsampled. 3. Melt is neglected despite it being important during part of the timeframe and having significant impact on results.

1. We expanded on the description in the manuscript about the second layer representing depth hoar and wind slab and their respective densities. We treat wind slab and depth hoar by taking the average ratio of these two layers, based on Radionov et al. 1997 and Sturm et al. 2002, over the

accumulation season. This ratio is then used to take a weighted average of the wind slab and depth hoar densities. We agree that new snow provides a smaller fraction of the total snow, especially towards the end of the accumulation season, which occurs in both the model and in observations (Warren et al., 1999).
2. We respectfully disagree, especially as we are taking gradients in ice drift fields on daily time-scales that can be very noisy, so other recent studies, e.g. Holland et al., (2014) smooth the data before using them. We hope to explore specific high res configurations of this model in the future, but wanted to carry out more extensive sensitivity studies across various forcing data in this initial model development stage.
3. Yes we neglect it and have added more discussion on this point based on the input from reviewers 1 and 2 in the model development and future priority sections.

Quality of the Model Results and Characterization thereof 1. Validation shown indicates the model produces results that do not capture the variability in observed snow depth or density reliably. Authors focus on averages of model output over decadal timeframes, which can be made to match observations by tuning of the arbitrary, non- physical constants in the model. This focus fails to acknowledge the inability of the model to capture interannual or spatial variability. 2. Prediction intervals are not pro- vided, but scatter plots show little relationship between individual observations of snow depth and modeled snow depth. No discussion is provided of how these errors would propagate in the intended use (altimetry retrievals of ice thickness) but it appears errors are sufficient to radically alter retrievals of depth and appear to indicate the data would not be useful for altimetry retrievals of ice thickness from ICESAT2. Authors fail to acknowledge any of these shortcomings and go to great pains to make the results appear good. 3. Modeled variability in density appears to have very little relationship to observations. 4. Comparison with the southern ocean, are pushed to a future effort, but validation statements in the paper suggest the model applies to ‘polar oceans’. 5. Results from the median of the three reanalysis products are declared ‘better’ repeatedly with no reasonable support. Taking the median of atmospheric reanalysis models would result in nonphysical jumps between atmospheric states and the removal of extreme events from the record, and is challenging to support physically.

1. We are confused by this comment as the assessment of how well the model captures spatial and interannual variability is shown in the OIB validation section.
2. Not sure what you mean here either. We show both the correlation coefficient (how well it agrees on the relationship between the two distributions) and the root mean squared error which provides a model error compared to OIB. We did not hide these numbers and have been very open with the performance of the model. The RMSE of ~9-10 cm is clearly not ideal but a value like this was broadly expected considering the challenges inherent in modelling snow accumulation and snow depth/density. The observational uncertainty on the order of several centimeters should also be considered here and we have made note of this in the discussion. It's thus hard to translate this into a snow depth uncertainty.
3. Correct, as we stated in the manuscript.
4. We have made clearer this is focused on the Arctic Ocean.
5. We have made clearer the justification for showing more of the median results in the New Arctic time period discussion was mainly for simplicity!

DETAILED COMMENTS

Page 1, line 16. “very strong agreement” Delete “very strong”

It does show very strong agreement with the seasonal cycles (very high correlations with the seasonal correlation plots) which is what we are referring to here.

Page 1 line 22 descriptions of agreement too subjective. The use here is altimetry. Tell the reader about the error in estimates implied.

We have changed this to: showing moderate/strong correlations and root mean squared errors of ~10 cm depending on the OIB snow depth product analyzed. These are similar to the comparisons of OIB-derived snow depths and the commonly used modified Warren snow depth climatology. Potential improvements to this initial NESOSIM formulation are discussed in the hopes of improving the accuracy and reliability of these simulated snow depths and densities.' We are wary of translating this into an error, due to the uncertainty and differences in the OIB snow depth products.

Page 2 line 5-8. Poorly worded sentence. Consider modifying. One suggestion is: The altimetry technique involves measurements of freeboard, the extension of sea ice or snow surface above a local sea level. Estimates of snow depth are required to derive sea ice thickness from either snow surface freeboard or ice freeboard, because snow depresses ice freeboard and adds to snow surface freeboard. Snow depth is one of the primary sources of uncertainty for both laser and radar altimetry (e.g. Giles et al., 2007).

We appreciate the suggestion but prefer the sentence as is.

Page 2 line 10. Replace 'lacking' with something more descriptive/accurate (they aren't lacking they are just not complete/good enough).

Changed this to 'very limited'. Also added 'direct' to observations at the start of the line.

Page 2 line 22-24. The sea ice community often relies on simple models of snow depth forced by reanalyses – please clarify how this is different. To the reader, it still looks like a simple model forced by reanalyses!

Here we are just making a comment here about what the community are using currently. Ours is also a simple model forced by reanalysis as you say and we don't think this suggests otherwise.

P 3 Line 16 "and two snow layers to broadly represent the evolution of both old/compacted snow and new/fresh snow." The assignment of the two layers in this two layer model is not consistent with the widespread understanding of the primary two layers on sea ice as depth hoar and windslab. New snow is occasionally present but usually rapidly transformed to windslab. It may be an acceptable third layer. See many of the snow on sea ice references cited here, such as Sturm et al., 2002 – generally the snow is treated in these two layers. The author's choice here to take the two layers to represent layers that the extensive literature reviewed does not discuss is perplexing.

We completely agree with the reviewer that the primary snow density layers are wind slab, depth hoar, and to a much smaller degree, new snow which, in reality, is more like a "transient" snow layer that gets redistributed by the wind. If reduced to two layers, the snowpack would be wind slab and depth hoar.

Given that the model does not have a temperature dependency, we were not able to parameterize a depth hoar layer since this is dependent on the temperature gradient within the snowpack.

Instead, we chose to include a “new snow” second layer, representing recent snowfall and blowing snow, which reduces the bulk density of the snowpack. Likewise, for the “old” snow layer, we explicitly chose density values that result from the mean ratio between wind slab and depth hoar from works by Matthew Sturm and historical data from Radionov. Although the model doesn’t explicitly take into account the seasonal cycle of this ratio, we feel that it’s a better treatment than applying the higher-end density value of wind slab and ignoring depth hoar altogether. Related, we chose not to apply a bulk climatological density because of its questionability in representing regions where observations are lacking.

P2 line 18 replace “detailed” with “iterative”. The simplified scheme does not permit a ‘detailed’ assessment of connection between input data and snow depth given its lack of physical complexity – it permits an easier iteration of possibilities.

Removed detailed from this line.

P4 line 13 Input data from passive microwave higher resolution than 100x100km, even if atmospheric data is not. Since ice concentration is so important, reviewer questions if 100km resolution is adequate. Further - does observed snow depth vary over 100km resolution? Since this is the motivation, what resolution is needed for useful for altimetry based determination of sea ice thickness?

As both drift and snowfall products come on grids with resolutions above 60 km, we did not want to use the higher res 25 km ice concentration grid. We think in these budget models the grids should be at least as coarse as the coarsest input data or it could look misleading. This is the approach taken in the concentration budget studies referenced in the manuscript. In reference to the later part of the question, higher resolutions are clearly better, but accuracy is really more important. One can always downscale data to higher resolutions but doing so doesn’t add additional information at that higher resolution.

Page 4 line 14 add “from reanalysis data” after the word ‘drift’.

The ice drift is not from reanalysis data. We have modified this to 'The model is forced with daily data of snowfall and near-surface winds from reanalysis data, satellite passive microwave ice concentration, and satellite-derived ice drifts.'

Page 4 line 16 – (volume of snow per unit grid cell in units of meters) – doesn’t make sense volume is meters cubed. Throughout the treatment of snow varies between depth and volume freely, but this free transition between volume and depth is challenged for some considerations of snow – particularly convergence/divergences. Since the goal here is to understand depth for altimetry retrieval, a convergence, which moves volume into a cell, is not the same as a change in depth.

This kind of terminology is common in models (e.g. CICE) to describe a quantity that is expressed over the entire grid-cell. Based on the recommendation of reviewer 1 we have tweaked the terminology used so that we make it clear we track the effective snow depth (over the entire grid cell) but can derive the snow depth over the ice fraction by dividing by ice area.

Page 5 table one – put formal references to data sources, e.g. “bootstrap” is not sufficient.

We provide references in the text and in the data section of the paper.

Page 5 delete "snow pit and density data. . . helped guide. . . parameterization . . . seasonal evolution." There is no prescribed seasonal evolution of density, use of snow pit data etc. in this model. Two constant snow densities are selected and declared. This sentence obfuscates the very simple, non-experimentally supported nature of the scheme.

The model produces a prognostic density from the ratio of old and new snow and the calibration of the model was guided by this data. The average ratio of wind slab and depth hoar within the second (old snow) layer is parameterized based on snow pit observations from Radionov et al. 1997 and Sturm et al. 2002.

Page 6 line 8 replace bulk density with mass.

Agreed, changed.

Page 6 – here authors note that the community of snow science experts and prior literature they have created generally group the snow into two layers (wind slab and depth hoar). They further note substantial differences observed in density of these two layers, and that these two layers comprise the majority of the snowpack. Not noted, but available in the literature is data showing that the contribution of the two layers to the overall snowpack varies from the approximately 50-50% contribution seen at SHEBA. So it seems windslab and depth hoar are the two layers to model. But. . . these two layers are different than the layers the authors have chosen (new snow/old snow). It seems a major departure from decades of snow research is being made here and it is not being well defended. Why?

We discuss this in more detail in response to some of your earlier comments, but briefly we were aiming to keep the model simple in this fist iteration. As we have no snow thermodynamics, explicitly capturing both snow layers and their different densities is less important for our given purpose.

Page 6 line 12 "for this reason we use the average of higher end values of ws and dh". Reviewer sees no reason provided supporting the use of the higher end of the range of values for each of the two common layers. The mean density of each layer, multiplied by the mean fraction of each layer should provide a more representative density for the combined wind slab and depth hoar. Further, the value selected is not the average of the higher end of the range of values for each of the two common layers, leaving it unclear how it was determined.

The density values chosen for the wind slab and depth hoar in the "old snow" layer are based on the average ratio of these properties within the snowpack from Radionov et al., 1997 and Sturm et al., 2002. The compacted layer uses a value of 350 kg/m3, which is a value calculated using the average ratios of depth hoar (40%, 250 kg/m3) and wind slab (60%, we found values of 410 kg/m3 in the referenced works above) within a snowpack. Admittedly, the ratio of depth hoar and wind slab seasonally changes and the model does not account for this seasonal change. However, we feel this is a better treatment than neglecting the influence of depth hoar altogether in the "old" snow layer. We have revised this description in the manuscript to make this clearer. In the conclusion, we expand the discussion to include future work in incorporating a temperature-dependency of the model, which will enable the separation of wind slab and depth hoar layers rather than applying a crude treatment for the old snow layer.

Page 6 Line 16 "Our simple parameterization is thus expected to be generally representative" No reasonable evidence provided supports this. Statements like this are found throughout this paper.

Delete or support with concrete evidence that quantifies what the range of uncertainty they will work within.

We have deleted this statement from the revised manuscript.

Page 6, Line 23 (default of 5m/s). Default or for the purposes of this work is it simply always set to this?

It's fixed in the model but can easily be changed in the open source code.

Page 6, Line 24 "determines the fraction... transferred..." Over what time? (seems that the coefficient is model timestep dependent. . . and perhaps shouldn't be)

Yes, we have changed this to not be timestep dependent (units of s $^{-1}$ now)

Page 6 line 26 'Wind threshold of 5m/s was determined based on. . .' studies. Please add a description of the range of wind thresholds indicated by these studies, and why 5m/s was selected from within that range.

We implemented a 5 m/s threshold based on Liston and Sturm (2002) and the dry snow transport in Li and Pomeroy (1997). In reality, this threshold depends on the snowpack's physical properties (grain size, water content, etc.) and atmospheric conditions (humidity, air temperature, etc.).

We conducted sensitivity tests on the wind speed threshold and found this to have a negligible effect on the modeled snow depth distributions. However, there are some regions where the wind threshold and wind loss term will play more important roles, such as the Antarctic environment where more leads are present, more snowfall events occur, and windier conditions occur more frequently relative to the Arctic (Massom et al., 2001; Toyota et al., 2016; Massom and Sturm, 2017; Massom, pers. comm., 2018; Webster et al., accepted).

Page 6 Line 8 Daily gridded ice drift is still required in this Eulerian scheme, eliminating it as a reason for choosing Eulerian over lagrangian, discussed above.

Our statement referred to the fact you do not need consistent ice drift - i.e. the model can be run for periods/regions with no ice drift.

Page 7, line 19. Reviewer is not aware of any evidence indicating that the loss of snow to leads in the North Atlantic sector of the Arctic is significant relative to the thick snowpack in that region. No evidence seems to be coming out of the N-ICE experiment to that effect. Some quantification of loss to leads in the Antarctic has been made by Leonard and Maksym as noted, but this was in the southern ocean. Please cite appropriate literature or delete speculation.

Correct, the snow lost to leads in the North Atlantic may not be significant relative to the thick snowpack in that region. However, we speculate that a greater proportion of snow is lost to leads there than in other Arctic regions given the lower ice concentrations, more open leads, more frequent snowfall events making more fresh snow available to redistribute, and windier conditions in the North Atlantic.

Page 8 line 4 – This parameterization doesn't make sense and is under supported for several reasons. 1. It appears that a constant coefficient beta is multiplied by 10m windspeed NOT by the amount which the wind speed exceeds the threshold velocity! So snow is lost to leads even when winds are too slow to move snow. 2. The amount of the snow lost to leads increases linearly with windspeed, when the drifting snow volume is well known to vary more rapidly than linearly 3. The loss to leads varies linearly with open water area, again this is likely more rapid than linear, and a thought experiment with random lead spacing/size could arrive at a better approximation. 4. The parameterization removes a fraction (2.5%) of the new snow layer to leads on each windy timestep – timestep is then important due to compounding what timestep is this defined for? 5. Is this parameterization/ value supported by any field quantification of loss to leads or is it simply made up due to lack of available observation. Either is fine, but state which it is. Page 6 line 9 – missing parenthesis on equation

1. We have changed this such that the wind action threshold is applied to the blowing snow loss term.
2. Our approach is simple, agreed. we hope to explore each term in more detail in future developments.
3. Agreed.
4. This term is now time step independent.
5. Yes, this was not based on any studies. We have added clarification to the text: ' We have no observational constraints for this parameter and is a free-parameter in this model, chosen through our model calibration efforts.'
6. Added the parenthesis, thanks.

Page 6 Equation 7 – appears incorrect. Change due to blowing snow is added (last term), but this should be a loss term (loss into leads). It appears that the term calculated in Eq 5 is always positive, so adding here will result in addition of snow, not loss. Similarly, how signs are handled on dynamics, convergence and divergence as well as advection depends on how (+-) ui is defined in equation 3 and 4, and this is not (but should be) specified above. . . so the reviewer is unsure if the sign here is handled correctly.

Correct, this was a mistake. We have added a negative sign to Eq 5 and 6. This was correct in the model code. Dynamics are all correctly signed.

Page 8 line 21 August is mid- late summer. Change “early” to ‘late’ or delete.

Changed this to 'middle'.

Page 9 line 2-3 Do these melt events invalidate the results here? Is this model useful before these ‘hoped for’ additions occur? It sounds like this is being hurried along.

We have no data to test if this invalidates the results here but the extra snowfall did improve our initial model testing efforts so we included it despite the obvious concerns of missing melt events.

Page 9 line 8-13 This paragraph appears to handle a specific test case, not discussed here. Seems out of place possibly a draft fragment. Unclear what tests this new density applies to, or how this test relates to the model released for community use. (update after later reading, now understand what this refers to, but still feel it was out of place and not well enough contextualized here)

Changed calibration to testing to make this clearer. The 1-layer approach is not included in the model released to the community.

Page 9 line 16 Soviet - capitalize.

Changed.

P 9 line 26 one OF the

Changed.

P9 – would be appropriate to acknowledge the lack of validation sites or validation data over Arctic sea ice, and uncertain accuracy of the products in that region.

We have added some comments regarding the uncertainty in precip earlier in the revised manuscript.

P11- Taking the median of the reanalysis products is an interesting idea if one has no idea which of the different products is best, but don't authors have better information about which is doing best from the comparison studies in literature?

Not really over Arctic sea ice. A new study of precipitation comparisons over the Arctic Ocean has recently been published (in early form) in the Journal of Climate but co-authors of this study, but the lack of validation data makes it challenging to recommend a particular product that one should focus on for such studies. This exercise was useful to guide us to exclude some reanalyses from our study - e.g. MERRA-2 - as stated in the MERRA data description.

P14 L10 – Initial conditions the Warren climatology is quite outdated. It is good you are trying to update them somehow. Is there evidence, e.g. from current autonomous ice mass balance buoys, that snow still regularly survives summer? Can you 'calibrate' this adjustment scheme based on those observations? Would a degree-day model be better than number of melting days? Also, what category is this snow placed in? Does it have a density reflective of melting snow (i.e. 400-500 kg/m3)?

Good question, and one that we can't answer with certainty about large-scale snow depth distributions in summer given the lack of observations. However, we can piece together information to hint at an answer. Based on these "pieces" (observations), some snow persists and the amount of snow that persists in summer has decreased relative to the Warren climatology.

In the IMB data, for example in 2012 and 2013, four IMB buoys show snow is present while six buoys show snow-free conditions in mid-to-late August. Based on survey line data from SHEBA and HOTRAX, snow can persist in summer where the largest drifts exist (next to ridges, etc.). However, when in the field, it's extremely difficult to tell the difference between a melting, slushy snow layer and a melting slushy ice surface (see the SHEBA field notes for an interesting reference for this – some scientists called this slush snow while others called it ice). What we can say with more confidence is that there is less snow in summer than there used to be based on the decreasing trend in surface albedo from AVHRR and melt pond information from MODIS data.

If the reviewer means a model based on the number of freezing-degree days, then our opinion is that the persistence of above-freezing days would have a larger effect on the removal of snow than the amount that's present in May.

We don't explicitly use a density in the initial conditions since we scale the climatological snow depth based on the number of consecutive above-freezing days.

P14 L22 – explain how this is 'linearly scaled' a bit better. Provide an equation. Is the fraction by which duration of melt is different from mean simply multiplied by snow depth? Does it mean that at 2x duration no snow is left and at 0x duration 2x snow is left?

We have provided more detail in the revised manuscript.

scaling factor = (mean duration for climatology - duration for individual summer)/mean duration for climatology

initial snow depth = august climatology*(1+scaling_factor)

We could add this to the manuscript if the reviewer thinks necessary.

*We only have ERA-Interim data for 1979-1991 rather than the full 1954-1991 climatological period to calculate the mean duration of above-freezing days in summer.

P14 L 28 – were necessary. . . Could this be because the model doesn't handle melt processes?

We don't believe so since deeper snow depths were needed to improve the comparison between the model and Soviet station observations (hence the initial conditions). Melt would reduce the snow depths so we don't think that is why.

P15Fig2 – These substantial August snow depths in 2012 and 2013 should be compared against available buoy data to determine if they are reasonable. The reviewer believes they are not and that this is ultimately a nonphysical tuning mechanism that helps account for lack of melt processes and poor representation of precipitation phase at this time of year in reanalyses.

The IMB data for 2012 and 2013 show four IMB buoys with snow present in summer (and six buoys with snow-free conditions). For the cases where snow is present, three of the four buoys show depths of ~10 cm and larger. The buoy IDs for persistent snow are: 2012C, 2012D, 2012G, and 2012I. Note, some of these buoys show data for summer 2013 and are not just limited to the year 2012, when they were deployed.

P 16 L 17 – this section is missing a clear statement of how accurate OIB data is expected to be.

The recent STOSIWIG study (Kwok et al., 2017) stated that snow depth can be estimated from a Snow Radar echogram with an uncertainty of 'several centimeters' although this depends strongly on the ice conditions, the particular Snow Radar system being used, and various other factors (e.g. geolocation errors associated with the plane pitch and roll. We have added a sentence regarding this to the discussion at the end of Section 3, where the OIB data are introduced.

P16 L 19-27 – pretty hand wavey – not rigorous.
Show plots of how snow evolves in model – what fraction of the snowpack is new snow layer vs time (it would have to be small to be realistic.)

We feel the plots already included in the manuscript provide the seasonal evolution of the budget terms highlighting the small contribution from the new snow layer.

P 17 Fig 3 – modeled data appears systemically low by about 5 cm depth. Snow density has essentially no relationship between modeled and observed. Individual year data – which is how this data would be used to derive altimetry based estimates of ice thickness – appear poor.

We didn't want to over fit the model to the data. Obviously it would be easy to calibrate this to increase snow depth in this comparison but we wanted to be flexible to the fact these depend strongly on the chosen forcings and the time period of analysis.

An r = 0.58 does not translate to no relationship. Clearly the snow depth shows a better relationship, but the density relationship isn't awful.

The validation part of the model is more in reference to the modern OIB data record, so we refer the reviewer to this.

P17 L 13 delete "extremely" – an error of ~5cm on a snowpack of ~20cm is still a 25% error.

We are referring to the seasonal cycle here and have attempted to make this clearer that we are referring to the seasonal cycle. We dropped extremely.

P18 L 1 – r of 0.74 would not generally be characterized as 'strong' P18 L2 - delete 'more' . . . its just moderate. Also, what is actually suggested here is that the model is good at predicting the MEAN – because you can tune your constants to make the mean look very nice, but not very good at capturing the interannual variability that is key to getting snow depth right for altimetry.

We are unsure how you have decided it's not good at getting the interannual variability? That's not really shown in this figure due to the lack of coincident data across years.

P18,L4-5 "In General, the moderate/high correlations. . . provide confidence. . ." This statement is hand waving and cheerleader- y without content. Delete this statement and replace it with a statement that articulates the degree of certainty with which output of the model should be treated. Suggest authors calculate the +-95% prediction interval for a modeled density or depth relative to this dataset. Suggest authors do this for individual locations/months on individual years, as well as mean.

At this stage we are just calibrating the model. The validation and expression of model errors comes in the OIB section.

P18, Figure 4 – This comparison is really just showing how well tuned the models are on average. Since the model is not presented as a mean climatology, but rather is presented as a deterministic snow product for specific locations on specific years, this comparison is inadequate.

Again, the OIB comparisons provide this later.

P19 L 3-6 Here authors make an odd argument. The model does not reproduce the climatology observations as well as a single mean over the entire timeframe. They argue this is OK because the model will handle interannual variability better because of its 'more advanced' density parameterization. The density parameterization is not particularly physically realistic, however, and fails to meaningfully capture interannual variability of the climatology density data (figure 3d). Reviewer therefore finds this statement lacking.

We state that the model is able to respond to expected interannual variability in the forcing, which is different to how the reviewer expresses this. Again we test this later with OIB.

We agree our approach is very simple, so have changed the more advanced line to read 'include a simple bulk density parameterization;

P19 l10 – not all marginal ice zones have low concentration – clarify that low concentration areas are where greater impact is expected

We have changed this to 'low ice concentration regimes'

P 19 L 19 – Reviewer disagrees that wind threshold velocity for blowing snow is un- constrained. Resources reviewed can establish that under all but extreme conditions (e.g. recent rain on snow) a threshold of 10 m/s is pretty high, maybe unreasonable. It would be better to range this within the values observed in the references cited earlier.

Changed this to poorly constrained. As you say this is on the high side, but we wanted o test the sensitivity so wanted to provide a big change, and doubling was consistent with the other changes. It does appear from this that increasing this value only slightly might lower the low bias in snow depths compared to the Soviet station data, but would also lower the density to create a low bias. This is obviously a tough balancing act and further calibration efforts will be explored in future.

P 20 L 15-18 – Speculative. Reviewer finds no reason to believe the median should be superior in regions of heavy snowfall. Defend or delete.

We have changed this to 'for simplicity'. As you say we provide no evidence that the results are better, but it is a useful synthesis forcing data to use - preventing the creation of figures across all the different forcing sets.

P21 L 5 – again this represents the mean over the decade being presented/compared. The model performance over this timeframe is highly tunable and not the performance metric of interest to an end user taking this data as an input to altimetry – that user would want to know the prediction interval for individual or moderate size groups of snow data points, and probably also whether there is any change in mean bias over time.

We have provided more information about the comparisons in terms of the root mean squared errors, which we feel are the best and most fair way of comparing these snow depth distributions.

P 22 Fig 6 – are standard deviations in depths this low comparable to any observations? If so they suggest a single climatology would be adequate for most end uses.

It's challenging to compare the SD from a pan-Arctic model with the SD from an in-situ dataset. We also dont have enough confidence in the snow density to make such a statement.

P22 L 15 plurals

Not sure what this refers to.

P23 L4 The fact that there is scatter among the reanalyses is not necessarily an argument for taking the median of them. Delete.

The point is that the snowfall forcings show strong differences so also having some kind of 'consensus' (e.g. median) snowfall dataset is useful. We have changed this line in the text.

P23 L19 providing significant VOLUME REDUCTIONS and sinks of snow (wind packing is not a sink)

Agreed, we have changed this to reductions of snow volume.

P24 L1 – Convergence really causes an increase in snow volume, not an increase in depth. The use of depth vs volume for a cell needs to be sorted out and treated consistently throughout this paper. Reviewer recalls a section way up at the top saying snow would be handled in volume throughout the paper, but has seen treatment vary.

We have made it clearer that we are tracking an effective snow depth (snow volume distributed across the entire grid-cell) and replaced volume with depth throughout the revised manuscript.

P25 Fig 9b – unclear what "Ocean" refers to. Snowfall directly into water? Caption needs to be more descriptive and the figure subcomponents should be linked back to which equation # they represent.

We have completed re-labeled this figure to improve clarity.

Fig 9f – this underscores the issue with treating convergence as a change in depth- in reality convergence/divergence of the ice at scale does not change depth in the sense that such would be used to interpret remote sensing. It changes snow volume, further, it appears the impact of convergence/divergence is noisy at best.

We treat this as a change in effective snow depth over the grid-cell. This is the most consistent way of treating this and avoids issues with the changing ice concentration.

Fig 9 I – density map warrants discussion. For example, density appears highest in central arctic (far from melt) This is likely untrue and an issue with not including melt/rain on snow processes. Very low density is indicated in marginal areas around N Greenland and in Baffin bay/Canadian islands. Can these be supported at all?

We agree that melt processes, especially in the marginal seas, should increase the density in these locations. This has been added as a future priority for the model.

Fig 9 h-k colors appear to fade toward lower value indication near land in general. Is this valid or a plotting artifact? Again volume and depth are used interchangeably and plotted on a depth scale. This needs to be resolved consistently.

Not sure I really see what the reviewer is referring to. Maybe some lower values in the older snow but likely due to wind packing as not seen in the new snow results..

Page 24 L 5-7– Authors state these measurements are explicitly for altimetry retrievals, so they must have characteristics useful for such, (more than matching seasonal evolution on aver- age) including: 1. Capture interannual variability 2. Capture spatial variability 3. No long term bias or trends in error (that could be mistaken as trends in ice thickness). If these cannot be shown, perhaps a discussion about whether this approach is viable is needed.

This all comes in the OIB validation section.

Page 26 Figure 10 – "As in figure 6" This is a nice addition to convey consistency, but pls also provide full description in caption, don't make reader hunt back several pages.

OK added.

Page 26 L15 Without melt processes included, what explains the loss in depth?

I believe in this case it is no significant accumulation and wind packing reducing the density. Can be shown in Figure S4.

Page 26 L 5 Why the depths are different during the later time period IN THIS MODEL is within the scope of demonstrating a model, even if understanding why they are changing in reality is not. Please answer: Are depths less due to less ice for snow to fall on? Or due to less precip? The reader must know if the model is representing the changing Arctic – since it is calibrated on old data. This cannot reasonably be scoped out of the study.

The snow depths are different between periods primarily due to the difference in the timing of sea ice freeze-up. Maximum snowfall rates occur in autumn (Warren et al., 1999; Webster et al., 2014; Lique et al., 2016; Boisvert et al., 2018), so if sea ice forms later, more snow falls into the open water relative to when sea ice formed earlier. This has been modeled in Blanchard et al. (2018) and Webster et al. (accepted), studies implementing a similar modeling approach to NESOSIM, and another study using a more sophisticated model (CCSM) (Hezel et al. 2012) as well as in ongoing work using CESM 1. This relationship has also been shown in Webster et al. (2014) based on ice mass balance buoy, *in situ*, airborne, and satellite data. The *in situ* and buoy data suggest no significant change in snowfall rates, while the combination of *in situ*, buoy, and airborne data show a decrease in snow depth that corresponds to later sea ice freeze-up (derived from passive microwave data).

We deliberately do not include a more in-depth discussion on this topic because 1) the manuscript is already at 13,000 words, and 2) we're preparing a manuscript on this topic and feel that including this discussion would detract from the purpose of the other manuscript. A comparison of the reanalysis precipitation between the 1980s and 2000s show little difference, which suggests that less precipitation (magnitude, phase) is not the primary driver of these differences. This is still open to debate, however. As the reanalysis description was very similar across the time periods we have now dropped this from the manuscript (following the recommendation of Reviewer 1). The 1980s budget figures have been moved to the SI too. Our aim is to focus on the model performance in the 2000s period, along with the OIB validation analysis. If the reviewer is interested in seeing more results on this topic, we would encourage him or her to contact Melinda Webster at melinda.a.webster@nasa.gov

P 27 Fig 11 – see comment on Figure 10

OK, changed.

P 29 Fig 12 – please clarify what positive and negative deviation mean (is the product higher or lower snow than the median product)?

Red (blue) colours indicate the individual reanalysis-forced simulations have higher (lower) snow depth.

P 30 line 3 – it is not clear that the median provides a result any more useful than the others. One should note which product compared best to coastal stations data and any other indications from literature which might be best.

The comparisons with coastal stations (e.g. Lindsay et al., 2014) are not wholly representative of precipitation biases over the Arctic Ocean. This is discussed more in Boisvert et al., (2018) and we reference this in the revised manuscript.

P 30 line 20 – this is not surprising and should be noted as such. Advection/convergence/divergence was much less important than snowfall in the plots above.

This is showing the spread based on the product spread, rather than its importance to the total mass balance. We agree this isn't surprising, but it is worth documenting and including.

P30 L 23-24 – here is where the idea of snow depth vs snow volume is really important. Dynamics are perhaps not important in depth over a 100km cell average, but they are important to the DEPTH on the actual subgrid ice, since divergence creates new ice with no snow, rather than rearranging all the snow into a gridcell average. The averaging over the 100km cell at each timestep may be particularly important in ice generating areas, where snow is continually averaged back into source regions, rather than being advected out entirely. Tracking ice classes within the cells, as is done in CICE may be critically important.

We agree this is worth thinking more about in future.

P 31 figure 13 – the drift scheme matters little over huge areas because convergence and divergence cancel. This plot is just not the right way to consider this, particularly in the context of use fore spatially distributed altimetry observations. Figure 14 suggests that the drift products don't differ that much between them in the central basin, but that having drift represented at all is very important, altering snowpack by O50% in large areas of the Arctic.

The point here is that if one cares about regional mean snow depths, the choice of product isn't that important really. Agree it matters more on the grid-scale hence the reason to show maps in the following figure.

P32 L6 – There are actually substantial biases in the peripheral seas – which may not be important overall, but cannot be ignored in the statement about biases.

Added ' and issues around the ice edge'

P33 – given the importance of concentration product, better understanding the role of changing concentration in the changing modeled snow depth above is important.

Agreed.

P34 – Tough to compare to observational data this noisy. Reviewer agrees they can be considered 'in agreement' within the bounds of the error of either. . . both of which are large. Are any of the OIB algorithms emerging as superior? Must all three be treated as equally likely?

Not sure noisy is the right phrase to use here, but there are clear differences in the products (an observational uncertainty). They seem to all have different pros and cons depending on the region/year/scale being analyzed. The STOSIWIG paper (Kwok et al., 2017) made no clear recommendations in this regard unfortunately.

P35 – comparison of 100km grid cells still includes substantial averaging, but already shows poor agreement. Agreement should be presented in terms of a 95 % prediction interval so user knows the capability of the method in useful terms – if the model says snow was xx OIB will say snow was xx +- yy 95% of the time.

Figure 14 shows the spread in snow depth across all the OIB campaigns, with the spread/uncertainty based on the product spread. If we wanted to add the individual product uncertainty we would have to guess at this as, they don't necessarily provide this. It is also thought to be highly variable and a function of the ice type and snow depth profiled. The spread is expected to be very large for a regional mean. The RMSE comparisons in Figure 15 etc. are a useful way of showing how the products and NESOSIM compare, but we acknowledge throughout the high uncertainty in both NESOSIM and OIB estimates making such comparisons challenging to interpret.

P36 – this discussion of the comparison of the scatter plots goes to great lengths to avoid describing the obvious. The model isn't very good at reproducing variability on OIB data, and if you believe OIB snow data is in any way representative of the variability in snow depth on ice, the modeled snow depth isn't very good at capturing spatial or interannual variability. The conclusion should then be that more sophisticated model representations are needed or that OIB data is trash. Since the model didn't agree with the Soviet drift station data scatter plot very well either, I don't think you can conclude that the model is adequate but OIB is trash.

We are confused by this statement. We very clearly present the correlations and rmse values for the model and OIB comparisons, then also show the mean interannual variability comparisons. The reader can clearly make their own opinion regarding the comparison but I don't think that would be that either the model or OIB are trash. The correlations are moderate/strong in general, there are no obvious skews/biases and the rmse is ~10 cm, with some values lower than this. Both the model and OIB have uncertainties associated so saying stronger statements than we have should be done with caution. Hopefully our inclusion of the mW99 comparisons help put thse in context.

P37 – "in general, however, the moderate to strong correlations. . . gives us confidence" Reviewer cussed in exasperation when reading this. This is a science paper not an opinion piece. These are not moderate to strong correlations! They clearly show NE- SOSIM cannot capture the variability observed well. Get this subjective language out of the paper and replace it with quantifications of how well the model does at both rep- resenting means (where performance is good because of tuning) and variability (where the model is not working so well). Talk about whether the model is good enough to be used in altimetry honestly and present some paths forward to getting there if it isn't.

See discussion above. We did indeed include an interpretation of the comparison metrics we presented in this study based on the chosen metrics, which is pretty standard practice. The choice of moderate/good/strong was based on standard definitions for the interpretation of correlation coefficient values.

P 37 L 19 – data is yet to be released in parenthesis? Thin its out now. . .

We changed this to 'the data was not available for this study but was made available during the review phase of this paper'

P 37 L 20 – There must be some field data available that you could at least spot check it against!

Unsure what the benefit of a spot-check would be in this instance as it wouldn't be a particularly robust comparison.

P 37, L 26 delete very strong, delete good

We would like to keep these in. The agreement with the mean seasonal cycle was very good (very high correlation).

P 37 L 28 contributing to the MODELED seasonal evolution in snow depth

Added.

P38 L5 uncapitalize New, consider replacing with 'more recent.'

New Arctic was cited earlier so we wish to keep this statement.

P38 L7 There is no evidence presented that this median product is better, and good reason to believe it just averages in erroneous values and non-physically jumps be- tween atmospheric states toward limited representation of extreme events. Defend the use on scientific merit or consider deleting the median product.

We dropped the last part of this sentence.

P38 L10 use consistent language. . . it is 2nd order on mean, but first order in some regions.

Added 'in our regional mean analysis'

P 38 L 14 "moderate/strong correlations" This statement is flatly unsupported by the results shown, and authors 'confidence' in line 16 is unfounded. The product does not represent the OIB data well in terms of the intended use – in retrieving thickness from freeboard.

We are discussing correlation coefficients above 0.5, with some above 0.6/0.7. Moderate/strong are the appropriate way of describing these values based on the statistical guidelines we referred to.

Please provide a variable list

Thank you. We have added a variable list to the paper.