Response to RC1. The authors provide responses to the reviewer comments in italic serif typeface throughout whereas the reviewers comments remain in normal sans serif typeface.

This study investigates a modelling approach to estimate ice sheet wide time series of Surface Mass Balance (SMB) and Firn Air Content (FAC) evolution on both the Greenland and Antarctic ice sheets (GrIS and AIS). Using a set of firn cores, the authors recalibrate a firn densification model and establish a new formulation for surface snow density. The MERRA-2 atmospheric reanalysis product is used to compute climatic conditions and to force the firn densification model. The importance of different SMB components as well as their evolution is analysed for both ice sheets. The authors partition ice sheet wide volume changes associated with surface processes between mass and FAC changes, showing that the latter dominates seasonal variability, while the former dominates multi-annual trends.

I believe that this study demonstrates a comprehensive approach to estimate ice sheet SMB and FAC evolution. The modelling framework is robust, well-explained and is undoubtedly a great contribution to the firn modelling community. The use of MERRA-2 is also noteworthy, because this is the first assessment of ice sheet SMB using this product. The authors propose innovative ways to deal with challenges associated with decades-long, large scale simulations, and their results demonstrate an extensive work to perform these simulations. The objective of the study has direct implications for ice sheet mass balance assessments performed via satellite-based altimetry techniques or via the input-output method. As for any model-based study, assumptions and simplifications had to be made. I list my reservations concerning some aspects of the approach in this review. Given the quality of the study, I am confident that a slightly revised version of the manuscript will be accepted for future publication. I realise that modifying the modelling approach to account for any suggestion from the reviewers would subsequently require to re-run the simulations. I do not consider this necessary to address the reservations I raise. However, I expect the authors to provide strong justifications or to better acknowledge limitations with respect to my reservations in the revised manuscript.

I have separated my review in Specific comments requiring more clarity in the manuscript and/or strong justifications from the authors in their response, and Technical comments related to the structure of the text. Despite my numerous comments, I strongly encourage the authors to re-submit the manuscript after the revisions have been made.

We thank Dr. Verjans for their support of the work presented and the time dedicated to providing a very thorough review that has substantially improved our research and publication quality. We apologize for the lengthy wait for the response; both reviews provided substantive feedback for improvement of our work, which required additional work. We acknowledge Dr. Verjans' review revolved more around adding clarifications and limitations to the modelled work rather than actually re-running simulations. As they will see, we have re-run the simulations accommodating many of the comments within both their review as well as that of reviewer 2, which we believe have led to an improvement in the quality of work.

Specific Comments

1) The surface density formulation ($\rho 0$)

Constraining surface snow density is both important and difficult. The authors took the approach of using a very large range of possible predictor variables and a single formulation for both ice sheets. My first concern relates to the neglect of melt sites. The formulation is calibrated only to the set of firn cores previously selected for the dry firn compaction calibration. However, surface snow density values

are also used in the percolation and ablation areas of the GrIS, which are rejected from the calibration selection but account for a predominant part of the GrIS total area (see Figure 1). The estimated $\rho 0$ values are particularly low in the GrIS southwest, which is the highest melt area. This disagrees with observational studies (e.g. Machguth et al., 2016; Fausto et al., 2018). Why did the authors decide to use the same selection criteria for the surface snow density calibration as for the dry firn compaction calibration? Surface snow density should be accurately represented, including for higher melt areas. Underestimating surface snow density there can lead to overestimation of the firn retention capacity, and thus lower runoff values. I believe that this largely explains why the GrIS-wide runoff estimate of this study is on the lower end of the GrIS SMB intercomparison (Fettweis et al., 2020). A similar effect can be perceived on the Antarctic Peninsula, which has $\rho 0$ values in the lowest range of the AIS.

The reviewer brings up an excellent concern regarding our choice of eliminating sites that experience significant melt from our constraint on surface density. We have opted to include ***all*** *observations of density, irrespective of their melt rates, in our calculation of surface density. The initial reason that the melt-rich sites were excluded was under an interpretation of "new snow" as "dry snow." The thought process being that snow is deposited at an initial dry density and then subsequently modified as melt occurs. However, we realize that this surface density is more of a long-term integrated mean density, sometimes being dry and other times wet; thus, melt should be considered when using such a simplified surface density scheme. The new parameterization provides more realistic results when compared to observations, and we use the references provided to acknowledge their contribution.*

My second concern relates to the use of the northward wind speed in the formulation. This climatic variable has very different physical meanings in various areas of the AIS (i.e. wind from the inner continent or from the shore), and even more between AIS and GrIS. I am thus skeptical about the physical sense to include it in the parameterisation. Furthermore, given the few GrIS cores used in the calibration, I believe that including the northward wind speed may undermine the validity of the parameterisation on the GrIS.

Both reviewers expressed concern regarding the lack of physical basis for the use of northward winds. We agree with the reviewers and have reformulated the relationship using total windspeed rather than just northward.

Finally, the model performs worse for the lowest $\rho 0$ values. It should be mentioned that these correspond to surface densities most critical for FAC calculations because firn layers of low density have high FAC values.

Yes. The simple regression is not capable of reproducing the lowest initial densities, which we express on Line 273. We have added a line in the section on FAC (Section 3.1) to express more clearly the impact of overestimation of surface density on our model results.

2) The use of the effective temperature (TE)

It is well-explained in Section 2.1.3 that stage-1 firn compaction rates cannot be assumed to depend simply on the mean annual temperature. Computing an effective temperature is a novel approach that accounts for the impact of high temperatures on compaction rates, but that leads to other potential problems. Firstly, I think that more details should be given about Equations (15) and (16). In Eq. (16), is TE meant to be $\overline{TE}$? It is stated that Eqs. (15) and (16) are only used for stage-1 densification, thus should Ec be $Ec0$? Is $\bar{k}$ the average of all hourly k values of the climatic forcing?

We thank the reviewer for the close inspection of the equations presented and indeed, they were lacking some key details to fully understand the calculations performed. Because of the concerns regarding the potential "downstream" impacts of the use of an effective temperature, we have elected to re-run the simulations using mean temperatures rather than mean effective temperatures. We have added a section that compare the results between the two approaches with simulations at daily resolution. The use of a traditional mean temperature better matched the 1-day simulations better than using the effective mean. Thus, we have eliminated use of the effective temperature. We do, however, add in the text that there are likely other ways in which an effective mean could be utilized in future work.

Secondly, several shortcomings related to the use of TE should be mentioned. As far as I understand, k represents the compaction rate of a firn layer at the skin temperature. The temperature signal is dampened in depth and k thus overestimates the compaction rate of the whole stage 1 firn. In turn, this leads to an overestimation of the effective temperature. Also, the use of TE as temperature forcing in the temporal coarsening of the climate input can have a significant impact on the firn temperature profile in the simulations. In the CFM, a newly deposited layer has its temperature set at the temperature of the time step (thus TE in this approach). Subsequently, the layer is buried and carries this temperature signal, causing advective heating. I guess that TE is significantly higher than the mean skin temperature for a large majority of the grid cells and coarsened time steps. As such, advective heating is significantly greater, which in turn leads to overestimated firn compaction rates. I do not know how much the 5-, 10- and 20-days TE values differ from their mean temperature and thus how strong this bias in compaction rates is.

The reviewer is entirely correct that effective temperatures will be larger than traditional mean temperatures; thus, the modeled temperatures at depth are larger than if they were modeled with mean temperatures. Thus, use of the temperature profile from the initial simulations would provide temperatures that are not realistic. Thus, we abandoned the effective temperature, but do note that this could be implemented in a future CFM release by simulating both effective and physical temperatures with time.

3) Densification Model Calibration (Section 2.1.3)

3a) The assumption that the "logarithm of the firn density profile with depth is approximately linear" was stated by Herron and Langway (1980) but never mathematically verified. I would appreciate if the authors could validate their use of this assumption for their calibration process. I suggest that summary statistics are provided to evaluate this validity. It should be straightforward to compare point density measurements of the firn core dataset to corresponding density values taken from the estimated log-profile (and re-converted to kg m-3 units). I ask the authors to provide RMSE and R2 values of the fit for both stage-1 measurements and stage-2 measurements. These statistics should be computed before removal of the measurements via the iterative 3-sigma edit. I would welcome any valid alternative way to validate this assumption put forward by the authors.

The reviewer raises another reasonable concern regarding the approach to using linear fits to the logarithm of the firn depth-density profile. In order to address the concern, we have executed a comparative analysis of the fit statistics between deriving a linear fit to the log-profile as well as the original density profiles. We believe that we have designed it as the reviewer suggested, including keeping all sites (rather than just using those that remained after the 3-sigma edit. The results are presented in Section 2.1.3 and we have added a plot comparing the RMSE/R2 values between the linear fits to the actual density profile against the log(density) profile.

3b) I am not sure to understand how "the firn density measurements and model output are binned into half-meter depth increments to obtain similar sampling intervals before slopes are estimated". Are all measurements (resp. model outputs) averaged in intervals of 0.5 m and the slopes computed on these 0.5 m averaged density points?

We bin the observations to 0.5 meter increments as well as the respective model output. Slopes are then calculated on these binned profiles (both observations and model output). We have reworded the sentence on Line 167 to better express the actual technique.

3c) There are mathematical inconsistencies when substituting Eq. (11) in Eq. (18). The final formulation of the firn model assumes: $\dot{b}\beta 0 \times \dot{b} = \dot{b}1+\beta 0\ exp\ (\ -60000\ RT\ ) \times exp\ (\ -Ec0\ RT\bar{E}\ ) = exp\ (\ -60000-Ec0\ RT\ )$ Both these assumptions are mathematically wrong. I understand that these are made for practical purposes, but they should at least be mentioned in the manuscript. Similar concerns hold for the stage-2 formulation (substituting Eq. (12) in Eq. (19)), even though they are less critical because $\bar{b} \approx \dot{b}$ and $\bar{T} \approx T$ in deeper firn.

The reviewer is absolutely correct; we had to make assumptions regarding how the calibration translates into modification to the densification rates, which we did not fully expand. We have clarified the assumptions made (as well as the equations presented) to ensure mathematical correctness.

3d) Why do the authors reject sites falling in a same grid cell? And how do they choose which depth-density profile to exclude? They could very well compute two different pairs ($R0$,$R1$) within a single grid cell. This would illustrate the natural small-scale heterogeneity of firn structure.

We actually do not reject sites that fall within the same grid cell and have added more context to the sentence beginning on Line 159. In reality, what we meant to convey is that we do actually compare ***all*** *observed profiles with model output, but that the number of model simulations is fewer than the total amount of observations because several observations fall within the same grid cell. We realize that the language used was vague and have tried to be more explicit in our description.*

3e) Why is the intercept forced to 0 in the regression? Is it to make the estimation of the parameters well-determined?

The choice to set the intercept to zero was to ensure modification to the functional form presented by Arthern et al. (2010) and limit the possibility of overdetermination. Basically, we wanted to ensure that the observed variability was linked to an atmospheric control rather than an unknown. We have added justification to this choice in the text on Line 193.

3f) Is $E0$ exactly 0? Or was it sufficiently close to 0 to set it equal to 0?

Excellent question. The uncertainty in the linear fit was larger than the prediction, so we assume that the E0 does not differ significantly from 60kj per mol, or rather, the observations are not sufficient to resolve a significant deviation from 60kj per mol. We have clarified how we came to the results in Equation 17.

4) The degree-day model (Section 2.2.1)

4a) In their study, van den Broeke et al. (2010) used $T0$< 273.15 K with the justification that: "on days with a negative average T2m, the method predicts zero melt if T0 = 273.15 K is used, while melt may have occurred during a short period. This problem may be avoided by applying the method to hourly T2m values or by applying a lower value for T0". Because hourly $T2m$ values are used in this study, citing

van den Broeke et al. (2010) to justify the choice of $T0$< 273.15 K is inappropriate. Also, this raises the question of the physical sense of using $T0$< 273.15 K. Should the calibration not rather fix $T0$=273.15 K and tune DDF only?

We appreciate the concern from the reviewer as it is one of the largest assumptions made in this work and can be improved in future studies. The reviewer is correct that we should not cite van den Broeke et al. (2020) to justify use of a T0 < 273.15 as they were using daily means whereas we use hourly data. We did, however, learn from van den Broeke et al. (2010) that DDFs become more realistic at lower T0 threshold, which is something that we found in our analysis as well. We have refined Section 2.2.1 to acknowledge the reference in a more appropriate manner. When using T0 at 273.15K, we ended up with unrealistically large DDFs over regions that experience little melt. As we moved to lower T0, the values normalized more, and we use an evaluation of how well the model predicts the training data to select the most realistic model. The use of a degree-day model is not the most physically robust model; we have attempted to expand on the limitations. Additionally, we do not explore the potential for temperature bias within MERRA-2, which would also bias the temperature threshold needed to reflect melt.

4b) The selection of the best $T0$ threshold depends on maximizing r 2 and minimizing $RMSE$. However, it is not explained which particular r 2 and $RMSE$ are considered since many grid cells are used. It is only in the caption of Figure 7 that the authors mention that "the median r 2 and $RMSE$ of every grid cell" are used. This should be stated in the main text. Also, I wonder about the relevance of the choice of the median values. Most of the grid cells have very low melt rates. It is not important to capture the low melt rates with great accuracy. It is much more important to capture melt rates of the grid cells in high melt areas. Thus, why choosing the median?

We agree that we need to provide more context in regards to how the curves in Figure 7, so we have added additional explanations in Section 2.2.1 to make it more clear how we performed our analysis, but also explain many of the limitations. We have used the median because we were more interested in representing most values, regardless of magnitude. When interpreting ice elevation change, while the largest melt rates impart the largest height changes, we must consider the relative importance of the changes. So even small melt rates can be relatively important outside of regions of fast-flowing ice and strong ice melt. We discuss this choice and present the T0 solution if we had used the mean to show they do not differ by much.

4c) I ask the authors to provide the final ranges of DDF values used for the GrIS and for the AIS.

We have provided the distributions in a new figure and present the ranges in the text in Section 2.2.1.

5) Wet firn compaction

Simulating wet firn compaction and liquid water processes is a major weakness of firn models. I certainly do not blame the authors for this and addressing this shortcoming is not the subject of this study. I appreciate that the results of the compaction model are also evaluated at high melt sites (Figure 8). I think it is important to also provide the bias of the compaction model at the zero-, moderate- and high-melt sites to know if the model tends to overestimate/underestimate densities in such melt conditions. Also, I believe that the text should remind the reader in the Discussion section about the wet compaction shortcoming and that it can have a significant impact on FAC results for the GrIS and ice shelves. As stated by the authors themselves, only a limited area of the GrIS satisfies the criteria used for the dry firn compaction calibration. This implies that firn compaction can only be expected to be well represented in that limited area.

The reviewer brings up a very fair point, and we have carved out additional discussion within the FAC discussion on how poorly constrained wet firn processes are over the ice sheets. We appreciate that this lack of knowledge translates really challenges our ability to model FAC changes, especially over the Greenland Ice Sheet, which experiences a significant amount of melt. We still use the results but have made it clear that the results are limited significantly by our lack of understanding of wet snow/firn processes and that the field needs improved process studies to build better models.

6) The Reference Climate Interval (RCI)

In Section 3.2, the authors are perfectly right: "The RCI is ideally representative of long-term steady-state conditions". However, when they evaluate the assumption of their RCI choice, they only assess the "steady-state" aspect and neglect the "long-term" aspect. Indeed, the RCI should show no trend in any climatic field and this is thoroughly investigated for both the GrIS and the AIS RCIs. But the RCI should also be representative of the climate under which the firn column was established (i.e. of the past centuries in AIS). Some studies contradict the assumption that 1980- 2019 is representative of the long-term AIS climate and that there are very likely some pronounced regional trends (e.g. Medley and Thomas, 2019). Similarly, in Greenland, regional long-term trends may exist (e.g. Hanna et al., 2011). This impacts the spin-up process because the initial firn column should be in equilibrium with the past climate. Again, such difficulties are inherent to firn simulations because reliable climate forcing covers only the recent decades. Thus, one cannot expect the authors to have a solution to this particular problem. But I would like this limitation to be mentioned in the manuscript, as well as its potential impacts on the findings.

The reviewer highlights a very important assumption that all firn evolution simulations must consider, and they rightfully point out that we do not fully convince the readers of the work as to our selection of RCI. We have added a "Limitations" section within the discussion to highlight this challenge as well as several other challenges we face when modeling firn column processes (surface density, degree-day modeling, etc.). While we cannot solve this problem, we do provide an evaluation of the potential impact of this selection. We have done several additional simulations for select sites to show how this choice impacts our modeled FAC changes.

7) Comparing SMB and FAC components

Seasonal variability in height is shown to be driven more by FAC than by snow mass. However, FAC gain/loss is essentially governed by snowfall amounts. For example, if we assume 1 m i.e. accumulation over a given month and a surface snow density of 300 kg m-3 , the corresponding FAC gain is ~2 m. In other words, without considering compaction, one should expect FAC variability to be around 2 times larger than SMB variability. The values found in this study are around 3 and show the additional effect of seasonally varying compaction rates. But the reader should be explicitly informed about the direct dependence of FAC variability on the SMB variability and, as a consequence, about its expected larger magnitude. Most of the change in FAC is not simulated by firn densification models but is determined by the climatic forcing. In regards to this aspect, I find the statement in the Conclusion line 505 misleading ("Thus, determination of seasonal mass change using satellite altimetry requires a substantial FAC correction, highlighting the importance of firn densification models, especially when investigating shorter intervals of change as not being mindful of the seasonal cycles of SMB and FAC can generate large biases.").

We entirely agree that FAC variability is driven predominantly by snow accumulations (or lack thereof), which include both air and ice mass. Our intent was not to mislead, but rather highlight the difference between changes in "air" versus "ice" because of the importance for satellite altimetry studies of ice-sheet mass balance. Specifically, we must remove the changes in air to measure the mass balance. How does one convert SMB to height change? You must have a density, which is a large unknown as the reviewer has already pointed out. We have reworded many of the sentences discussing this to clarify that SMB controls ice and air, whereas firn processes control air content. In regards to the sentence of interest, we have clarified to combine surface mass balance AND firn models importance.

Technical comments

All changes suggested by the reviewer were included except those with specific responses.

p.2 l.39: "few hundred meters", I am not sure that the firn column can be that deep (e.g. Ligtenberg et al., 2011), please provide a reference.

p.2 l.53: Make sure to consistently use either "solid earth" or "solid-earth" throughout the manuscript.

p.3 l.71: I suggest adding a statement underlining the sensitivity of Eq. (3), such as "Mass balance estimates are highly sensitive to small errors in the height change measurements and in the modelled firn signal."

p.3 l.72: I think that "Variable rates of the height change due to compaction" should be replaced by "Height changes due to variable rates of compaction".

p.3 l.73-76: I suggest not introducing the variables dhc/dt, dhm/dt and dha/dt because these are not used in the remainder of the manuscript.

p.3 l.80: SMB and FAC appear in the wrong order: "air thickness and the thickness of ice: surface mass balance (SMB) and firn air content (FAC), respectively"

p.3 l.86: Add a comma: "(…) mass fluxes at the surface, including (…)"

p.4 Eq. (6): There is a typo in the equation, which should have ρi in the denominator: $FAC = \int \rho i - \rho(z)\ \rho i\ dz\ z\rho i\ 0$

p.4 l.110: Why do the authors simulate grain-size evolution?

It was simulated as a test. We have clarified that the grain sizes were simulated, but are not likely useable because we were testing the capabilities.

p.5 l.141: "subset of 256 published firn depth-density profiles" The authors should provide a little more detail about the dataset of firn cores used in this study. I suppose that the SUMup dataset is used. If this is the case, the authors should cite the work of Koenig and Montgomery (2019) (https://doi.org/10.18739/A26D5PB2S). If other datasets are used, they should also be cited. All the references can be provided in the section Code and data availability or in the section Acknowledgements.

We did not use the SUMup dataset, but rather compiled data from the literature. It was a complete oversight to not reference the data. We have now included the references.

p.6 l.156-157: I do not understand the point of this sentence. The authors introduce a model in which grain growth is only a function of mean annual temperature, which is also the case for the model presented above. Do they mean that Arthern et al. (2010) actually developed two different models? However, the model in which grain growth does not depend on the mean annual temperature is not the one calibrated in this study. Please clarify the purpose of this sentence.

p.6 l.160: Typo, change "form" to "from".

p.6 l.171: "did not contain more than 7 data points for that stage" before or after the 3-sigma edit?

Before the 3-sigma edit. We have clarified the language to express that.

p.6 l.174: Note that stage 1 and stage 2 were not previously defined, which might be confusing for readers less familiar with firn densification models.

Thank you for the comment. We have introduce the stages earlier.

p.7 Eq. (14): Typo, there should be no space in ln.

p.8 l.215: Change "equations" to "Eq.".

p.8 l.216: I think it is worth mentioning the good agreement of the calibrated coefficients with the calibration of Verjans et al. (2020), despite using very different statistical techniques. This reinforces the reliability of the calibrated dry densification model.

We agree and have included it in our revision.

p.8 l.217: Change "equations" to "Eq.".

p.8 l.221: Remove the italic from "any".

p.8 l.223: Define "peripheral ice".

p.8 l.233: What is meant by "interpolate between these neighbors"? I believe that the same SMB and FAC time series are taken for all grid cells classified as neighbours. If so, I suggest changing the statement to "we use model output of a single cell as representative for all neighbors".

p.8 Section 2.1.5: In my opinion, the reader should be informed about how the climatic output is processed to the coarsened resolution. This could be summarised in a single sentence by specifying that precipitation, evaporation and melt fluxes are cumulated and by reminding about Eq. 15-16 for the calculation of TE.

p.9 l.243: Use "GrIS" instead of "Greenland Ice Sheet".

p.9 l.244: Please be more precise about "several calibration sites".

p.9 l.245: Change "when simulated at five, ten, and twenty days" to "when simulated at resolutions of five, ten, and twenty days".

p.9 l.246: I am not sure to understand how the residuals in $dFAC/dt$ are computed. Are $dFAC/dt$ values computed at each time step (five, ten, twenty days) or is only the total change in FAC considered?

We have run all of Antarctica at 5-day resolution, so this analysis is no longer relevant.

p.9 l. 247: Are the mean snow accumulation and skin temperature good predictors of residuals in $dFAC/dt$ in the regression model? Could the authors provide summary statistics of the fit?

See prior comment.

p.9 l.262: "the 151 depth-density profiles (stage 1)" but in Section 2.1.3, the authors mention that they reach 141 depth-density profiles for stage 1.

p.9 l.264-265: Is the regression performed with the mean annual climate of the RCI or the mean annual climate of the entire MERRA-2 climatic forcing?

The mean annual climate of the RCI. We have clarified.

p.9 l.266-267: I do not understand the iterative removal process. If points with the largest residuals are iteratively excluded and the model is subsequently re-evaluated, there will always be points having residuals outside of the 99th percentile. I am probably missing something.

We have clarified the process. We do iteratively remove (and recalculate statistics) because one or two extreme outliers skew the statistics. We present the results when throwing out outliers in a non-iterative fashion as well to show it doesn't impact the results.

p.9 l.271: Change "surface mean temperature" to "mean surface temperature".

p.9 l.272: Specify "we capture more than 50% of the variability for measurements used in the calibration".

p.10 l.282-284: Use the abbreviations "GrIS" and "AIS".

p.10 l.294: I believe that "Sect. 2.4" should be changed to "Sect. 2.1.4".

p.10 Eq. (21): "Eq. (21)" refers to two different equations. The references to Eq. (21) in the text should subsequently be adjusted.

p.11 l.320: Please clarify what is meant by "the threshold if determined by one evaluator alone".

p.11 l.321: I think it is important to insist on the DDF being different for each grid cell. Thus, I suggest changing "and the DDF calibrated to that threshold" to "and the grid cell specific DDFs calibrated to that threshold".

p.11 l.328: I suggest changing "realistic magnitudes" to "realistic annual magnitudes".

p.11 l.329: Again, I suggest changing "and the DDF calibrated to that threshold" to "and the grid cell specific DDFs calibrated to that threshold".

p.12 l.353: Change "against observations" to "against the calibration data set".

p;12 l.354: Clarify what is meant by "shared variance". Should it be "explained variance"?

p.12 l.356-357: I have some doubts about the values given for % decrease in model error. I believe that the authors calculate them as $MAE\ mean(values)$, which is not the same as $mean$ ($|error|\ value$).

We calculate the mean absolute error as the mean(abs(error)). The MAE numbers presented in Figure 4 are compared to the values presented on Lines 355-357 to generate the percentages. Specifically, Figure

1a shows MAE = 0.030 while the mean observed rate is 0.066, which we call a mean absolute error of 45%. We have attempted to clarify how these numbers were generated.

p.12 l.357: Remove "Interestingly". Every reader might not consider it as interesting, although I certainly do.

p.12 l.362: Typo, "a" should be "an".

p.13 l.382: I suggest changing "locally" to "local".

p.13 Eq. (22): This equation is already given as Eq. (5).

p.14 Section 3.2.1: In my opinion, an interesting and valuable extra-contribution of this study would be to quantify the extension of the GrIS ablation area. That is, how does the extent of the area with $SMB <$ 0 has increased post-2003 with respect to the RCI? I leave it to the authors to decide whether to include it in the manuscript or not.

We have included this analysis as it is an interesting contribution to provide.

p.14 l.405: Change "Figure 11a" to "Figure 11b".

p.14 l.407: Change "statistical difference" to "significant difference".

p.14 l.415-416: Is "followed" the appropriate word? It seems to me that the decrease in runoff and the increased precipitation are simultaneous.

p.14 l.419-420: Clarify to which period the "gains" and "increases" refer to. Since 2003 or post-RCI?

p.14 l.425: Typo, there is no verb in this sentence.

p.15 l.428: I think that another word than "yet" should be used here.

p.15 l.435-438: If the authors compare grounded- and floating-ice numbers, they should clarify that they consider absolute values here because their extents are very different.

p.15 l.449: I think there might be an error about the value "142 km3 ". Here, the authors use it to quantify the post RFI annual net volume loss, but the same value is given below for the post-2003 period.

p.15 l.458: Consider replacing "Like the GrIS, the change in FAC is 3 times larger than SMB" with "The change in FAC is more than 3 times larger than SMB".

p.16 l.463-464: Change "the height and volume changes begin and end with zero" to "the height and volume changes in our model experiments begin and end with zero". It is important that the reader understands that this feature is due to a modelling assumption and is not necessarily representative of reality.

p.16 l.473: I think that there should not be a dash between "best" and "fit".

p.16 l.483-488: Please note that the Arthern et al. (2010) model was not developed to capture compaction of the very low density firn layers. The shallowest depth range for which it was calibrated for is 0-5 m. The densification process of very low density fresh snow is governed by different mechanisms, which are likely not well captured by firn densification models.

p.17 l.493: "GrSMBMIP" is not defined. I think that the sentence would be clear even without the abbreviation.

p.17 l.494: I think that "results" should be singular.

p.17 l.497: Note that the study of Wang et al. (2016) shows that MERRA has a similar bias than other models concerning SMB in Antarctica.

p.17 l.517: "" can maybe be updated.

Figures: In general, for all figures using different colour scales for the GrIS and AIS, please make sure to add a statement such as "note the different colour scales" in the caption.

p.22 Figure 1: I believe that the open circles were not used in any of the calibration steps of this study. If so, they should be removed from the maps or the statement "The open circles are calibration site locations" should be modified.

p. 26-27 Figures 5-6: Please provide the period over which the mean annual climatic values are considered (because MERRA-2 and M2R12K do not have the same time span).

p.32-33 Figures 11-12: Increase the size of the labels of the subfigures a.

p.34-35 Figures 13-14: If possible, increase the size of the axes-labels and of the legends.

References used in this review:

Arthern, R. J., Vaughan, D. G., Rankin, A. M., Mulvaney, R., and Thomas, E. R.: In situ measurements of Antarctic snow compaction compared with predictions of models, J. Geophys. Res.-Earth, 115, 1–12, https://doi.org/10.1029/2009JF001306, 2010.

Fausto, R. S., Box, J. E., Vandecrux, B., van As, D., Steffen, K., MacFerrin, M., Machguth H., Colgan W., Koenig L. S., Mc-Grath D., Charalampidis, C., and Braithwaite, R. J.: A Snow Density Dataset for Improving Surface Boundary Conditions in Greenland Ice Sheet Firn Modeling, Front. Earth Sci., 6, 51, https://doi.org/10.3389/feart.2018.00051, 2018.

Fettweis, X., Hofer, S., Krebs-Kanzow, U., Amory, C., Aoki, T., Berends, C. J., Born, A., Box, J. E., Delhasse, A., Fujita, K., Gierz, P., Goelzer, H., Hanna, E., Hashimoto, A., Huybrechts, P., Kapsch, M.-L., King, M. D., Kittel, C., Lang, C., Langen, P. L., Lenaerts, J. T. M., Liston, G. E., Lohmann, G., Mernild, S. H., Mikolajewicz, U., Modali, K., Mottram, R. H., Niwano, M., Noël, B., Ryan, J. C., Smith, A., Streffing, J., Tedesco, M., van de Berg, W. J., van den Broeke, M., van de Wal, R. S. W., van Kampenhout, L., Wilton, D., Wouters, B., Ziemen, F., and Zolles, T.: GrSMBMIP: Intercomparison of the modelled 1980–2012 surface mass balance over the Greenland Ice sheet, The Cryosphere Discuss., https://doi.org/10.5194/tc-2019-321, in review, 2020.

Hanna, E., Huybrechts, P., Cappelen, J., Steffen, K., Bales, R. C., Burgess, E., McConnell, J. R., Peder Steffensen, J., Van den Broeke, M., Wake, L., Bigg, G., Griffiths, M., and Savas, D.: Greenland Ice Sheet surface mass balance 1870 to 2010 based on Twentieth Century Reanalysis, and links with global climate forcing, J. Geophys. Res.- Atmos., 116, D24121, doi:10.1029/2011JD016387, 2011.

Herron, M. and Langway, C.: Firn densification: an empirical model, J. Glaciol., 25, 373–385, https://doi.org/10.3189/S0022143000015239, 1980.

Koenig, L. and Montgomery, L.: Surface mass balance and snow depth on sea ice working group (SUMup) snow density subdataset, Greenland and Antarctica, 1950–2018, Arctic Data Center, https://doi.org/10.18739/A26D5PB2S, 2019.

Ligtenberg, S. R. M., M. M. Helsen, and M. R. van den Broeke : An improved semi-empirical model for the densification of Antarctic firn, The Cryosphere, 5(4), 809–819, doi:10.5194/tc-5-809-2011, 2011.

Machguth, H., Macferrin, M., van As, D., Box, J. E., Charalampidis, C., Colgan, W., Fausto, R. S., Meijer, H. A. J., Mosley-Thompson, E., and van de Wal, R. S. W.: Greenland meltwater storage in firn limited by near-surface ice formation, Nat. Clim. Chang., 6, 390–393, https://doi.org/10.1038/nclimate2899, 2016.

Medley, B., and Thomas, E. R.: Increased snowfall over the Antarctic Ice Sheet mitigated twentieth-century sea-level rise. Nature Climate Change, 9, 34–39. https://doi.org/10.1038/s41558-018-0356-x, 2019.

van den Broeke, M., Bus, C., Ettema, J. and Smeets, P.: Temperature thresholds for degree-day modelling of Greenland ice sheet melt rates, Geophysical Research Letters, 37(18), 2010.

Verjans, V., Leeson, A. A., Nemeth, C., Stevens, C. M., Kuipers Munneke, P., Noël, B., and van Wessem, J. M.: Bayesian calibration of firn densification models, The Cryosphere, 14, 3017–3032, https://doi.org/10.5194/tc-14- 3017-2020, 2020