Dear Editor,

We would like first to thank both reviewers who will help to improve our manuscripts.

According to the reviewers recommendations, we plan in the revised version of our manuscript to
- improve our introduction to put better in the context the interest of such a models intercomparison as well as to better describe the advantages and drawbacks of each kind of models.
- add a table before the description of the models to summarise the information given afterwards (model name, type, resolution, forcing, …)
- start the discussion with the model intercommunion for ending with the comparison with observations.
- discuss more in-depth the comparison with GRACE. We plan notably to add a scatter plot comparing the bias with GRACE against the mean snowfall/runoff simulated by the models.
- generally improve the text of our manuscript with respect to the suggestions of the reviewers.

Thanks for considering these remarks in your decision.
Best regards,

Xavier Fettweis, on the behalf of all the co-authors.

Anonymous Referee #1 **(R#1)**

The following is a review of "GrSMBMIP: Intercomparison of the modelled 1980-2012 surface mass balance over the Greenland Ice sheet" By X. Fettweis, et al. The manuscript presented describes the Greenland Ice Sheet surface mass balance (SMB) model inter-comparison results for the historical period 1980-2012. The authors assess the ability of different types models (including regional climate models, radiation balance models, positive degree day models, and general circulation models), 13 in all, to estimate the surface mass balance over the Greenland continent. Skill criteria for the models are derived from observational datasets, including MODIS bare ice extent, ice cores/snow pits/ in-situ observations, and the calculation of regional SMB as the C1 difference between GRACE estimates of mass and previously-published ice discharge into the ocean. A large amount of effort is taken to design the experiment and compile model submissions that have the same source of forcing, the same spatial resolution, and cover the same overlapping time periods. Through this comparison, the authors derive an ensemble mean and standard deviation of Greenland SMB over the 1980- 2012 period, as well as trends. The authors find the largest model discrepancies are along the ice sheet margins, and it is the increase in meltwater runoff along the margins that drive the prevailing negative trend in Greenland mass balance over the study period. Results suggest that regional climate models have strong skill in matching observed patterns of SMB, though computationally expensive compared to the positive degree day or radiation balance models. Overall, the authors find that it is the ensemble mean that best matches observations, meaning that errors from the various models balance each other out and do not convey any obvious systematic biases.

The work presented here is critical for cryosphere scientists, especially to the scientists interested in quantifying and simulating the evolution of ice sheets (including atmosphere/surface/ice sheet modelers). This is clearly a massive effort, and as observation of SMB in many areas are quite sparse, the authors have done a very nice job of compiling meaningful comparison criteria as a first attempt at this type of exercise. Such an effort is quite necessary to build a SMBMIP community

and launch similar efforts in the future. The work presented here is especially a nice basis on which to build future comparison efforts that may focus on sea level projections. This is especially true considering the conclusion that the current compilation of models does not show systematic bias. For these reasons, publication of this work is timely and critical.

Thanks for these comments.

That being said, the manuscript in its current form needs a lot of work, especially the text which requires major revision. The tables and figures, in general, are adequate for conveying the discussion and conclusions of the manuscript. However, the model descriptions take up most of the text, and the rest is very concise. I think expanding upon the scientific results would make this manuscript much less of a technical paper C2 and much more appropriate for publication in The Cryosphere. Such improvements would also help broaden the audience for this paper. As is, the manuscript is difficult to digest by other cryosphere scientists, and the authors do not make it immediate clear to the reader why these impactful results may be of interest to their research.

Below, I outline my general comments to the authors:

R#1 1. Introduction

R#1 1.1 In general, the introduction should be expanded to discuss more clearly the topics of observed variability in SMB over the historical time period assessed and why it is important and/or difficult to capture them with models. In addition, an introduction to the types of models that are assessed should be given, since those reading this manuscript might not be familiar with how and why these types of models differ. This could be a good way to let the reader know about the general advantages and disadvantages of each model type also. Another helpful topic to cover would be a short discussion on why this effort is so important and what the authors are aiming to learn about model bias (i.e. why a historical assessment is helpful to complete before assessing projections from this group of models). This pertains to statements that are made in the conclusion section of the paper, especially those about implications on model coupling and about quantification of uncertainties in sea level rise projections. Introducing these concepts before mentioning them in the concluding remarks would help highlight their importance and future inter-comparison goals.

We fully agree with these suggestions and we plan to improve our introduction following them.

R#1 1.2 The first paragraph of the introduction mentions glacial water storage, and notes that this is the first time it has been evaluated. However, the GS term is not included in Equation 1, and GS is not discussed explicitly anywhere else in the manuscript. Please be more specific here about how GS is evaluated, and how it is being included in this analysis.

Sorry that our sentence aiming to tell that GS is not considered here was ambiguous. We will therefore rewrite our sentence to
Moreover, as it is simulated by no model considered here, the water glacial storage (GS; lakes, melt pond, channels,…) is neglected in this intercommunion although the mass changes coming from GS, when SMB is integrated over the whole ice sheet, could be relevant (but has never been evaluated until now).

R#1 1.3 Line 70-71, This line might be better coming after Eq. 1. C3

OK. GS is also missing in Eq. 1

R#1 1.4 Line 78, Please specify the type of variability you refer to here

OK. It is the surface water runoff increase.

R#1 2. Model Section

R#1 2.1 Maybe it would be helpful if there was an overarching Methods section, since the Model section would really benefit from a short introduction describing what your methods in general are, and what you plan to do as an inter-comparison exercise. If the Model, Observations, and Evaluation all came under a larger section, it might be a good way to add some explanation prior to the reader going through all the details right away without understanding what type of inter-comparison is being presented.

Such a section will be added at the end of the introduction.

R#1 2.2 As it is, it is very difficult and quite boring for the reader to be introduced to a list of models and their descriptions up front with no introduction to them. Maybe a table of model names, types, native resolution, downscaling type, etc., could help serve as a reference/summary to this section. Such a table/figure might help the reader to have something to refer to while looking through the tables and the figures. Easier access to model type (by a table or color coding in the figures?) would help make the results easier to read.

This is an excellent suggestion. A table summarising all of this information will be added at the beginning of Section 2.

R#1 2.3 If at all possible, it also might be helpful to push this list deeper into the section
R#1 2.4 maybe with the observations or data described first? (Though this might be fixed by section summary I mention earlier). - It is also important to note, that many model descriptions refer to RACMO within their write-ups, but no reference for RACMO, what it is, or what it stands for has been included prior to these sections.

As reviewer #2 recommend to put the model inter-comparison (Section 5) before the evaluation of models (Section 4), we prefer to leave the order of Section 2 and 3 as it as the comparison with data will be discussed after the model.
The new table at the beginning of Section 2 will mention RACMO allowing the reader to better know what is RACMO before describing it afterwards more in depth.

R#1 3. GRACE estimation Section

R#1 3.1 It could be helpful to include the equation of glacier mass balance here, so that it is clear to the reader how SMB is calculated from GRACE and ice discharge.

OK, the equivalent of Eq 1 for SMB will be added.

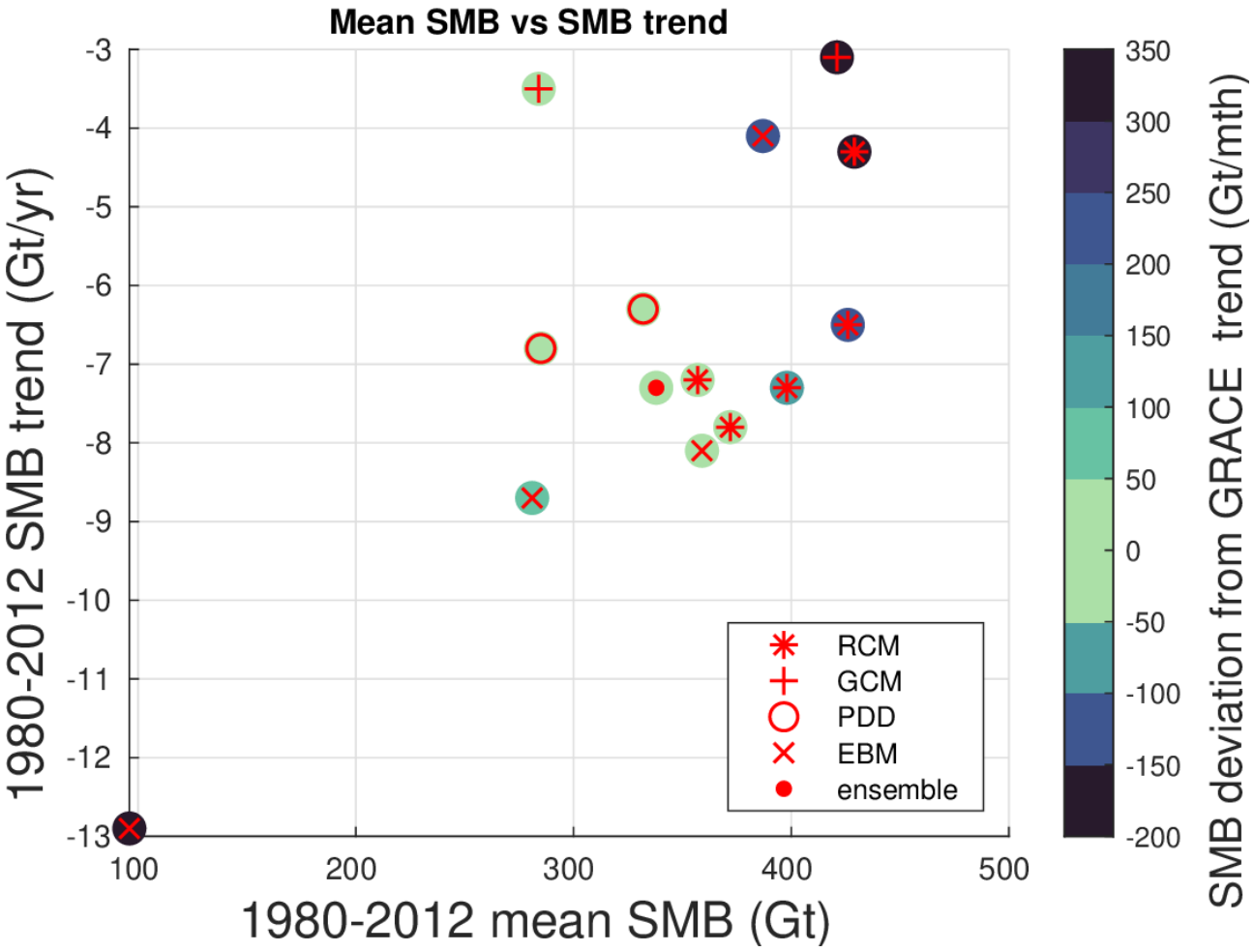
R#1 3.2 These last two sentences can probably be simplified to just say that you are using the methods of King et al. (2018), but instead of RACMO, you use each of your different C4 SMB products. These sentences are awkward that way that they read currently. Comparison with GRACE measurements Section

Excellent suggestion. We agree that our last two sentences are not very clear.

R#1 3.3 Results here are very interesting and there is plenty to point out to the reader. In general, I

don’t think there is any advantage to being extremely concise. It would be nice for you to lead the reader from figure to conclusion for some of the statements made in this section.

We plan to add figures similar to the one shown below, discussing the mean SMB rate/trend simulated by the models vs the mean bias with GRACE. This figure notably shows that the models with the largest SMB rate (due to higher snowfall or lower runoff) systematically underestimates the recent GRACE derived surface mass loss.



R#1 3.4 The discussion about the seasonal cycle here is a nicely suggested by your presented results. Could you please add some more explanation in order to lead the reader a bit more on why the RMSE from your Supp. Fig. 2 would imply how the seasonal cycle is modeled?

OK. We will better explain in the revised version that removing the trend allows us to evaluate the seasonality of the signal (which is a combination of both the seasonal cycle and the Global Warming induced mass loss).

R#1 3.5 For discussing the GCM’s, could you please be explicit about the difference between the forcing of variability on these models by ERA-Interim, and how it pertains to the RMSE?

OK. Idem. We agree that the problem of GCMs not using ERA-Interim as forcing is not sufficiently explained in-depth.

R#1 4. Conclusion Section

R#1 4.1 Here, some of new concepts that were not brought up earlier in the manuscript are mentioned. This includes the mention of coupling with an ice sheet model (i.e. a topographic feedback scheme was not used, maybe add a reference to a paper that shows the feedback may be important) and quantification of uncertainty in climate projections. I think the manuscript would be improved if some space was taken in the discussion section to mention more of how results presented here do have implications for these other applications. Implications of interest could range from estimates of historic sea level contribution to forcing of ice sheet models for historic and future simulations, and the ability to now give those applications error bars. I would even say that bringing these applications up in the Introduction as justification for conducting this MIP could help improve the manuscript's impact.

OK.

R#1 5. Below, I offer some more specific comments/suggestions:

Thanks for all of these suggestions that will be taken into account in the revised version of our manuscript.

Page 2, line 62: Please rephrase, "of the same order as RCMs compared with observations and therefore remain useful tools. . ." or something similar C5
Page 3, line 98: maybe, "although each model prescribes the reanalysis forcing in different manner". Please refrain from referring to the forcing as data.
Page 3, line 99: "(EMBs)"
Page 11, lines 1-2: A reference to Fig. 1 would be helpful here
Page 11, line 336-337: "This allows. . ." Please rephrase this sentence. It is very awkward and difficult to understand.
Page 12, line 359: maybe, ". . .compared to the resulting mass balance estimates from the GRACE product".
Page 13, line 411: "with the GRACE-derived. . ."
Page 15, line 452: Instead of mainly, maybe "largely"?
Page 15, line 457: Not sure what you mean by "oscillates" in this context. Maybe "deviates from the mean"?

Anonymous Referee #2 **(R#2)**

The manuscript presents an experiment in which the surface mass balance (SMB) output of five regional climate models, four surface energy balance models, and two positive degree day (PDD) schemes for the Greenland Ice Sheet (GrIS) are each forced with ECMWF-Interim atmospheric reanalyses over the period 1980-2012. They are compared with each other, with available in situ observations, with MODIS-derived bare ice extent, and with a derived gravimetric data set in which an observed terminal glacier discharge has been incorporated. The output from two global general circulation models is also considered. The main results presented are that the models simulate a statistically significant decrease in SMB over the period, that the largest differences between models occur on the ice sheet margins, and that regional climate C1 models generally perform well in comparison to the validation data.

The manuscript is around 7500 words with 6 figures and 4 tables, which is a reasonable length for the topic. It represents a considerable community effort in organizing and executing the experiment.

The author list comprises most major modeling efforts for contemporary GrIS SMB with the outstanding exception of atmospheric reanalyses (e.g., the Arctic System Reanalysis; MERRA-2). The initial reaction is that this is a significant update on earlier efforts of Vernon et al. (2013) and perhaps Rae et al. (2012) in model assessment. While those studies were mostly focused on regional climate models, this manuscript aspires beyond that with the inclusion of a large number of surface energy balance and PDD models. I have a few points for the authors to consider below, and so would suggest some revision of the manuscript.

Thanks for these comments.

R#2 1. The experiment necessarily relies on the common use of one forcing product, ECMInterim. By itself, this study is then not a complete characterization of SMB and its uncertainty from model sources, as the uncertainty of the forcing would also need to be considered. The use of different forcing products is beyond the scope of this study, but it would seem that the forcing selection plays a significant role in determining trends. Consider that if one wished to comprehensively evaluate a forcing product for the GrIS, a possible approach would be this experiment: a comparison of many forced models with observations may be seen as an elaborate validation of the forcing product. Is that not so? The purpose of this study is an appraisal of the different models, and for this purpose the key results are in how models compare with each other, and the systematic differences between them. These would seem to be the results that should be emphasized. Comparisons with observations are of interest (e.g., Fig. 1 is very interesting) but would not seem to be the principal outcome to be emphasized. The manuscript presents a considerable amount of information on the intercomparison in the form of figures and tables. Beginning in section 4, the text focuses primarily on the direct comparison with observations. The intercomparison is largely covered in the second paragraph of section 5. It is suggested that the results be re-ordered with the C2 intercomparison presented first. Some additional aspects of the intercomparison may be highlighted, as suggested below.

As suggested to reviewer #1, the models inter-comparison (Section 5) will be put before the evaluation of models (Section 4) in the revised version. We agree with the comments of the forcing (ERA-Interim) vs observations. However, according to Fettweis et al. (2017) who forced MAR with 6 reanalyses, the impact of the forcing on the modelled results over the recent decades (in particular over 1980-2012 considered here) remains negligible with respect to the model discrepancies we found here over this same period. The modelled results' dependence on the forcing as well as the associated impacts on the comparison with observations will be taken into account in the revised version of our manuscript.

R#2 2. For tables and Figs. 2, 4, 5, and 6, the plots should be sorted by model type rather than alphabetically, and perhaps labelled accordingly. It may also be useful to plot the spread for each model type.

Initially, we thought to do this but we prefer to show the legend alphabetically sorted and not put models of the same kind together, because the spread inside a model type is of the same order as the spread over all models.

For example, the spread around the mean SMB, snowfall and runoff listed in Table 4 is generally of the same order for the total of 13 models than for a class of model, except for the 2 PDD models, which are very similar in their design and underlying assumptions (except the resolution).

	Std dev around the mean SMB	Std dev around the mean Snowfall	Std dev around the mean runoff
EBM (4)	131	43	144
RCM (5)	32	73	117
PDD (2)	33	11	34
GCM (2)	96	78	43
Total (13)	91	81	109

R#2 3. It is useful to continually compare this experiment with the previous efforts cited in the introduction. Vernon et al. found SMB estimates were within 34% of the multimodel mean of 4 models. Table 4 suggests this value is now something like 22% for the 5 RCMs but close to 100% when all of the models are considered. Does this suggest an increasing proficiency within the RCMs. Also, it is noticeable that the manuscript does not indicate surface temperature sensitivity. It is difficult to include and assess the PDD and EBM models without that consideration. For example, this was a focus of Bougamont et al., who found that PDD models were more sensitive than EBM models. Given the same forcing and the trends shown in Table 2, it does not appear that a similar conclusion holds here, is that correct?

We think that the largest difference between the 34% found by Vernon et al. and the 22% shown here is the use of a common grid and ice sheet mask. As the models in Vernon et al. did not use the same ice sheet mask, a great part of the spread around the mean was only due to the fact the ice sheet mask was larger in some models. Therefore, the difference between the models presented here can not be compared like-for-like with the differences shown in Vernon et al. (2013).

Indeed, the trend in runoff shown in Table 4, driven by the temperature increase from the end of the 1990's, is generally lower for PDDs than for EBMs except BESSI and RCMs. It is nevertheless important to note that the increase of solar radiation (not taken into account in the PDD) has also played an important role in this meltwater increase (Hofer et al., 2017).
As explained in Fettweis et al. (2013), the recent and future changes are more sensitive to the ability of the models to simulate the current mean runoff, independently of the formulation used as the melt does not increase linearly with temperatures. About GCM, this trend is also lower, mainly because they do not simulate the general circulation changes in summer as recently observed (Hanna et al., 2018) and driving in part the recent surface meltwater runoff.

Hanna, E., Fettweis, X., and Hall, R. J.: Brief communication: Recent changes in summer Greenland blocking captured by none of the CMIP5 models, The Cryosphere, 12, 3287–3292, https://doi.org/10.5194/tc-12-3287-2018, 2018.

This issue will be discussed more in-depth in the revised version of our manuscript.

R#2 4. At 445 words, the abstract is too long by half. A large part of the abstract is devoted to motivation, which should instead be mostly left to the introduction. As suggested in the previous point, may consider adding more text regarding the resulting differences between the models.

Ok, thanks for these suggestions. The abstract will be shortened.

R#2 5. A concern is the very lengthy description of the models contained in section 2. The descriptions include sub-model components, the forcing time scale, vertical resolution, ancillary

forcing data, etc. It is of course useful for close examination of individual model results, but this is generally available elsewhere from the cited literature, and it is not clear that all of it is directly pertinent to the aggregate results presented for understanding cryospheric modeling. It is suggested that this material may be incorporated into supplementary text and/or condensed with a table that includes model type, references, and major points. Otherwise it could be argued that this type of material is C3 more appropriate for a publication such as Geoscientific Model Development.

As requested by Reviewer #1, a table listing model name, type, resolution, forcing, … will be added at the beginning of Section2. As Reviewer #1 tends to request more details about the models, we think that the current description is a good compromise.

R#2 6. Lines 455 and following. As indicated, it is apparent from Fig. 6 that the snowfall from the EBM models, which is directly imported from the forcing, is low (mostly blue) compared to the mean over the interior regions of the ice sheet. Would it be correct in saying that models that compute snowfall generally show higher amounts than the forcing? This appears to be true for most of the HIRHAM, NHM-SMAP, and RACMO RCMs and to some extent for the BOX13. Is that an expected systematic response?

Yes and no.
Ettema at al. (2009) found that higher the resolution, higher the simulated precipitation with RACMO is but it is no more the case with the more recent versions of RACMO (Noël et al., 2019). Moreover, Franco et al. (2012) found that lower the resolution is, higher the simulated precipitation with MAR in the interior of the Greenland ice sheet is, mainly because the topographic barrier effect is less efficient in the MAR model. We think that these differences are more driven by the physics of the models and by the different downscaling methodologies + corrections (eg. for PDD) applied to the ERA-Interim based forcing data.

Ettema J and 6 others (2009) Higher surface mass balance of the Greenland ice sheet revealed by high-resolution climate modelling. Geophys. Res. Lett., 36(12), L12501 (doi: 10.1029/2009GL038110)

Franco, B., Fettweis, X., Lang, C., and Erpicum, M.: Impact of spatial resolution on the modelling of the Greenland ice sheet surface mass balance between 1990–2010, using the regional climate model MAR, The Cryosphere, 6, 695–711, https://doi.org/10.5194/tc-6-695-2012, 2012.

Noël, B., van Kampenhout, L., van de Berg, W. J., Lenaerts, J. T. M., Wouters, B., and van den Broeke, M. R.: Brief communication: CESM2 climate forcing (1950–2014) yields realistic Greenland ice sheet surface mass balance, The Cryosphere Discuss., https://doi.org/10.5194/tc-2019-209, in review, 2019.

R#2 7. A list of acronyms in the appendix would be useful.

OK, we will add this.