Dear Dr. Morin,

Thank you for your thorough review of our manuscript and for your constructive suggestions.
Please see below our response to each of your remarks.

Sincerely,
Baptiste Vandecrux on behalf of the co-authors

The manuscript by Vandecrux et al., entitled "The firn meltwater Retention Model Intercomparison Project (RetMIP): Evaluation of nine firn models at four weather station sites on the Greenland ice sheet", provides results from a recent intercomparison of firn models dedicated to their handling of meltwater retention and runoff. This topic has attracted a lot of interest over the past decade, following the discovery of subsurface aquifers on the Greenland ice sheet and the fact, more generally, that liquid water transfer is a critical, yet very complex process governing the energy and mass balance of firn below the surface. The manuscript undoubtedly fits the scope of The Cryosphere. It is generally well designed and easy to understand, although, like many other manuscripts reporting on intercomparisons, the reading can be a bit cumbersome and the results will mostly appeal to experts in this field. I recommend publication of this article, and I provide here below some feedback and suggestions, which may be considered by the authors in case they lead to improvements of the manuscript.

Page 1, line 1-2: "Perennial snow, or firn, covers 80% of the Greenland ice sheet and has the capacity to retain part of the surface meltwater, buffering the ice sheet's contribution to sea level". This sentence could in fact be quite misleading and I suggest reformulating, either at least acknowledging that the "buffering" acts upon sea level *rise*, or maybe more generally, leave out the term "buffering" (which could be inappropriate in some cases where runoff generation can increase the contribution of the ice sheet to sea level rise), and use a more neutral term referring to the fact that firn processes influence the behaviour of the ice sheet and affect its contribution to sea level rise.

Revised as suggested: Perennial snow, or firn, covers 80% of the Greenland ice sheet and has the capacity to retain surface meltwater, influencing the mass balance of the ice sheet and its contribution to sea level rise.

Page 1, line 3 : "weather-station-derived" : maybe consider unpacking the wording, this is quite tedious to read.

Revised as suggested: forced by mass and energy fluxes derived from weather stations at four sites.

Page 2, line 60 : I think it could be useful, and appeal to a wider community of readers, if this article could provide more information on what a "firn" model is, or what a "snow and firn" model is if the two terms here are meant to be combined. Indeed, GCMs and RCMs all feature a "snow" component in their land surface models, and many such models do not handle ice sheets differently from a long-lasting seasonal snowpack. It could thus be quite appropriate to provide more background on how snow and firn processes are considered in GCMs and RCMs (or other tools used for reanalyses etc.), in order to make this effort even more useful in a CMIP6/IPCC context. This could make it possible to establish how the firn models used for this intercomparison fit in this wider context, thereby providing information on

whether the results of this intercomparison are relevant when discussing the results of existing GCMs and RCMs projection. Also(see below point too), it would be good to introduce what are the typical input/output to firn models, because they seem to deviate from typical input/output of land surface models, hence the link to GCMs/RCMs/etc. is not direct.

Regarding the use of firn model in GCM/RCM we added: “The performance of these models, when coupled to regional and global climate models, has a direct impact on the quality of ice-sheet mass-balance calculations (Fettweis et al., 2020) and sea-level change estimations (Nowicki et al., 2016).”. A general description of the firn model structures and traditional use is presented in the Method section. Since it is rather technical, we did not want to have it in the Introduction.

Page 2, line 64 : Although I know this can be debated, I have strong personal reservations against the term “validation”, which I believe is beyond reach in any geosciences field because the “truth” is never known hence “validation” (in the strong sense) can only be elusive. I much prefer the term “evaluation”, which better encapsulates the fact that the evaluation results from a comparison with observations, which also carry some uncertainties. I was happy to see the term “evaluation” in the title and in the abstract, so it was disappointing to see the term “validation” popping up in the introduction. Maybe this can be harmonized throughout. Along the same line, in several instances the comparison between model output and observations is referred as a “bias”. Here again, “bias” is a term which includes some judgement of value, implicitly assuming that observations are the “truth”. I think that observations are never the “truth” and, especially regarding in-situ snow and firn observations, we know that observations are intrinsically prone to significant errors, in addition to large spatial variability for many variables at all scales, which induces representativeness issues. In this context, I much prefer referring to “deviations” between model results and observations, without using the term bias. I note that the term “deviation” is used in several places in the manuscript, including in figure captions (e.g. caption of Figure 6, which I think is perfectly worded), so maybe this could be harmonized in the text.

We agree with these points and have revised the text to harmonize the use of the terms “evaluate” and “deviations”, recognizing that true validation is a difficult exercise, especially in field settings around variables that vary widely in space and from year to year. That said, some model results are arguably invalid or biased (e.g., deep infiltration of meltwater and anomalously warm firn temperatures at Summit, counter to available observations and physical expectations), so we still use of the term “bias” in the manuscript.

Page 2, line 67 : Here is a good example where more information could be provided on whether (some of) the firn models included in this intercomparison are representative to how firn processes are handled in GCM/RCM/NWP models, so that the results can be used to analyse some of the output of such models. At present, and even though some of the information is provided in section 2, the models included in the intercomparison are not categorized explicitly according to their use, and I think it could be helpful to the scientific community to provide the rationale and the results of the intercomparison in a way that can (somehow) be transferred to the interpretation of other model results.

Considering that the history and description of each model are already reported in the Methods section, we would like to keep the introduction concise and not list the models there (which would require

references and explanation of the acronyms). We also hope that the introduction now connects better the RetMIP with broader applications of firn models.

Page 3, line 91 : it seems that there is a typo in this reference, I believe this should be 1998. I haven't checked all reference, but if they were typed by hand and not using literature management software, there may be other errors in the references.

Revised to 1998, thank you.

Page 4, line 1 : This table is very useful, I suggest adding a column for providing the extended name (developing the acronym) and, more to the point, adding a column on how the model is typically used (included/coupled to a land surface model in RCM/GCM/NPW context, or purely offline for process investigations etc.). I'm convinced that the authors can easily define several categories within which models can be classified (these categories could even be used in the results and discussion, if common features, or not, emerge from these various categories, in addition to discussing results referring to process representation in models).

We thank you for this suggestion but would like to keep the model use within each model section within the method and not as a grouping criterion. Many of the models have been used for various purpose but with very different settings, making it unclear whether a label apply to the model set used here. Lastly, I would like to point out that, like for other earth surface models, the choice of a firn model for a specific application is rather determined by "legacy rather than adequacy" (https://doi.org/10.1029/2018WR022958). With our study, we hope we highlight the tasks for which each model is adequate independently of how it has been used in the past.

Page 4, Line 98 : I think it would be good to spell out the acronyms in the titles of subsections 2.1, 2.2 etc., this is otherwise quite obscure for non-expert readers not accustomed to the acronyms of these firn models.

To keep the titles of subsections concise we wish to keep the acronyms there. They are then defined in the paragraphs just below. Another motivation for this choice is that most of the acronyms used as model names do not carry information about the model type or design and therefor are not deemed worth to be highlighted in the subsections' titles. We note that Fettweis et al., (2020, https://doi.org/10.5194/tc-2019-321) used a similar strategy.

Page 5, line 116 : extra "-" between ASIRAS and instrument

This sentence was revised.

Page 5, line 131 : it seems to me that layers defined by a w.e. (e.g. mass per unit surface area) should not be qualified by a "thickness", which refers to a distance (in m). If the model is formulated in terms of layers with a given mass, then I believe the text should refer to this, and the substitution of "thickness" with "mass" will accurately represent how the model is formulated.

We agree that the phrasing was awkward here. We changed for: " DMIHH employs 32 layers within which snow, ice and liquid water fractions can vary and where each layer has a constant mass."

Page 7, Table 2 : I think that the column on "Hydraulic conductivity" needs some attention. The van Genuchten (1980) article provides a way to link between the saturated hydraulic conductivity and the hydraulic conductivity, which for snow has been addressed by several studies such as Shimizu (1970), Calonne et al. (2014) or can be used using geometrical estimates such as Carman-Kozeny (see Calonne et al., 2014 for a review of existing formulations, and Wever et al., 2014, for context). Hence I suggest to double check, for each model, what is the parameterization used for estimating the saturated hydraulic conductivity (corresponding to the permeability) from the microstructure (density, specific surface area/grain size) and the formulation used to derive the hydraulic conductivity (van Genuchten (1980) is probably widely used). This column seems to be lumping and mixing the two.

Thank you for spotting this. We now specify for each study using Darcy flow the saturated and unsaturated hydraulic conductivity, with the source of the coefficients that have been used if necessary.

Page 8, line 177 : I'm not fully convinced by the formulation of how the forcing data are introduced. "Any bias in forcing data propagates into the model output" : I'd rather suggest that any *difference* in forcing data propagates into differences in model outputs. The term "bias" is here inappropriate I think (see comments above).

Here we respectfully disagree and would like to continue with the word "bias". Regional climate models are known to have systematic deviation from observations both locally (AWS) and on a larger scale (against remote sensing products). For example, Noël et al. (2018, https://doi.org/10.5194/tc-12-811-2018) use 93 times the word bias when describing RACMO2.3p2.

Further, "To make sure we compare and evaluate the models independently of biases that may exist in forcing datasets that come from RCMs, we use meteorological fields derived from five weather stations at four sites." I think this sentence needs rephrasing, because it gives the impression that only RCM atmospheric fields can be biases, and in-situ atmospheric observations are not biased. I don't see the point in referred to RCM here at all, but simply state that "To make sure we compare and evaluate the models independently of differences due to forcing data, we use for all models the same meteorological fields derived from five weather stations at four sites." The references to RCM data is absolutely not needed here, and in the current formulations I consider it misleading and improper.

We agree that AWS data can be biased and we mention weaknesses in our dataset (f.e. lack of tilt correction for the radiation data) to be corrected in future intercomparisons. But we believe that AWS are the best estimation of local meteorological conditions. We would like to continue with the phrasing and raise awareness about the deviations that exist between RCM and AWS measurements: for instance RACMO2.3p2 (Noël et al., 2018) give air temperatures that are on average 2.7°C colder than observed at Summit station.

This is now also introduced in the introduction when mentioning previous model intercomparisons: “Steger et al. (2017) and more recently Verjans et al. (2019) investigated the impact of meltwater infiltration schemes on the simulated properties of the firn in Greenland. These studies highlighted the potential of deep-percolation schemes, for instance for the simulation of firn aquifer, but also the sensitivity of simulated infiltration to the firn structure and hydraulic properties. In these previous studies, the surface conditions were prescribed by a regional climate model. Inaccuracies in this forcing could therefore explain some of the deviation between model outputs and firn observations and prevented a full assessment of different firn model designs.”

Page 9, line 202 : If I understand well, the firn models are driven by 3-hourly skin temperature, meltwater generation (what it this ?) and net snow accumulation.

Indeed, we clarified the phrasing: “This surface energy and mass balance provides, at three-hourly resolution, the three surface forcing fields that were used by all models: the surface “skin" temperature, the amount of meltwater generated at the surface, and net snow accumulation (precipitation – sublimation + deposition). “

I think this warrants an explicit statement on the forcing data for firn models (see my comment regarding the introduction), because this appears to be quite different from forcing data of land surface models (including the snow component), usually driven by air temperature, relative humidity, incoming shortwave radiation, incoming longwave radiation, wind speed (and direction) and snowfall and rainfall rate (in offline or online applications).

I do not have much experience with reanalysis datasets and land surface models but RCMs like HIRHAM and RACMO use the same energy budget closure approach to calculate surface temperature and surface melt that are then passed to the firn module:

At the surface, snow mass is updated with snowfall, rainfall, melt and deposition/sublimation at each subsurface scheme time step (1 h). Likewise, the surface temperature is updated via energy budget closure with radiative and turbulent surface energy exchange above and diffusive and advective heat exchange with subsurface layers. If the surface temperature exceeds 0°C, it is reset to 0°C and the excess energy supplies heat for melting (Langen et al., 2015).
Langen et al. (2017, https://doi.org/10.3389/feart.2016.00110)

In RACMO2, the skin temperature (Tskin) of snow and ice is derived by closing the surface energy budget (SEB), using the linearised dependencies of all fluxes to Tskin and further assuming, as a first approximate, that no melt occurs at the surface (M D 0). If the obtained Tskin exceeds the melting point, Tskin is set to 0 C; all fluxes are then recalculated and the melt energy flux (M > 0) is estimated by closing the SEB.”
Noël et al. (2018, https://doi.org/10.5194/tc-12-811-2018)

The forcing fields that were provided to the RetMIP participants were therefore close to what each of these firn modules traditionally take as input. Surface models that could not take these prescribed forcings, and required to be given meteorological fields instead, were not considered in our study as they

would have had different temperature and meltwater input at the top of their simulated firn column. We do not see the need for listing the firn models that are not compatible with our forcing fields. The MeyerHewitt model was an exception to this since Colin Meyer found a meaningful way to re-calculate the surface energy fluxes that would give, for his energy balance scheme, similar surface temperature and melt as prescribed (details in the supplementary material).

In this context, it would be good to quickly introduce how the firn model data input are typically computed within GCM/RCM/NWP models where they are implemented.

In the introduction we now mention "Firn models traditionally take as input energy and mass fluxes at the surface and calculate the evolution of firn characteristics and meltwater retention at scales ranging from tens of metres to tens of kilometres ". Making a review of currently-used surface energy and mass balance models that can be used to force firn models (from full energy budget to PDD and statistical approaches), is out of the scope of this manuscript.

I also think that it would be beneficial for the manuscript, if the material provided in the Supplement regarding how the models were adapted in order to contribute to the intercomparison, could be placed within the body of the article. [In fact, I'm in favour of moving the whole supplementary material into the article, which is quite technical anyway, and which I think would benefit from having all the information in the same document].

We brought in the main text Table S1 (now Table 4) and removed some of the unnecessary material from the supplementary. Considering the length and the complexity of the study we would like to keep the remaining information (complementary information regarding our methods) in the supplementary and only have the necessary information in the main text. If there is a specific result or discussion point that cannot be understood without the supplementary, we will be happy to move the required material to the main text.

Page 10, Table 3 : Mean annual air temperature data should probably be homogenized in terms of formatting (why is the number provided by 0.1C resolution of KAN_U and rounded to the unit C for other sites ?)

Revised to common precision (nearest degree).

Page 20, line 371 : Eulerian

Revised.

Page 20, line 384 : it seems that the sentence does not end as planned. Was it the intent to refer to Calonne et al. (2019 ; https://doi.org/10.1029/2019GL085228) which may provide some hints into how to parameterize firn thermal conductivity ?
The sentence fragment has been corrected and this reference has been added - thanks for flagging this very relevant paper.

Page 21, line 404 : The last sentence “Especially, the heterogeneous nature of the firn, the presence of vertical ice features in the firn, the variability in surface snow density/thermal conductivity as well as firn ventilation are processes not currently included in the models and should be subject of future research.” is certainly true but is quite vague. Based on existing literature (e.g. Albert et al.) is it possible to elaborate more on the expected direction of change and whether accounting for such processes would be beneficial (and at what numerical cost) ?

Following the recommendation from Reviewer 1, we now only elaborate on topics for which we have data and/or model outputs. Topics such as spatial heterogeneity, firn ventilation or surface snow density are only mentioned for future research and are therefore outside of the scope of this study.

Page 21, line 413 : add “with” after “match”

Revised.

Page 22, line 433 : on the “fresh snow” density issues and beyond, there is little discussion in the manuscript on the connection between snow cover models (such as those embedded in land surface schemes) and firn models. Wouldn’t it be adequate that surface processes are handled by snow cover model rather than firn models, taking advantage of the features of both types of models ? This connects with the question on how firn models are used within coupled GCM/RCM or offline (see comments above).

As recommended by reviewer 1, we reduced the discussion of surface snow density as we do not present data or model output bringing new insight on this topic. We agree that surface snow density is tightly linked to the surface meteorology and consequently opens the question of how firn models are coupled to surface and atmospheric conditions. But we wish to focus here on the inter-comparison of firn models and leave this issue to future work.

Page 22, line 443 : It should be mentioned here that the uncertainty envelope corresponds to +/- 2 standard deviation (+/- 2 sigma). This information is missing here (although it is provided later in the conclusion). Regarding the temperature, I see a limitation to the estimate provided here, in the sense that the temperature of the firn cannot exceed 0C. How does this impact the uncertainty range provided for temperature?

Now clarified where we first introduce the uncertainty estimation. Indeed, as a consequence of this upper bound of firn temperature, model spread indeed decrease at sites where firn reach 0°C. It simply implies that it is easier to simulate firn temperature at a temperate site than at a percolation or dry snow site.

Page 23, line 451 : What is the argument for stating that this uncertainty range applies in situations where observations are not available ?

We have reworded this for clarity; they are modelled uncertainties, in the absence of observational constraints. Where observations are available, it is possible to calibrate models or reject models that do not perform adequately, reducing the model uncertainty envelope.

Page 24, line 482 : “much faster drainage of the aquifer” : faster than what ?

Sorry for the ambiguity: faster than the Zuo-Oerlemans parameterization. Now clarified in the text.

Page 24, line 484 : A qualifier is probably missing before “models” : “existing firn models” ? “firn models considered in this intercomparison” ?
Revised as suggested.

Page 24, line 490 : I suggest not using the term “reality” in scientific publications. Furthermore, a reference is missing to support the statement in this sentence.

This has been reworded here and at two other instances in the manuscript.

Figures : I didn’t notice any major flaw in the design or content of the figures, which are appropriate to convey the results of this intercomparison.