Review: *The firn meltwater Retention Model Intercomparison Project (RetMIP): Evaluation of nine firn models at four weather station sites on the Greenland ice sheet*, Vandecrux, et al.

Summary

In *The firn meltwater Retention Model Intercomparison Project (RetMIP): Evaluation of nine firn models at four weather station sites on the Greenland ice sheet*, Vandecrux et al. present an assessment of nine state of the art firn meltwater retention models, forced by data from four sites spanning a range of conditions in Greenland's percolation and dry snow zones. The authors find that densification schemes perform similarly, as evidenced by limited model spread at Summit station, where melt events are rare. In contrast, at sites in Greenland's percolation zone, where seasonal melting is commonplace, models simulate a wide range of density and temperature, which is supposedly primarily influenced by different infiltration schemes between models. Based upon the results, a host of discussion points are raised indicating shortcomings of particular models that need to be addressed.

The fate of pore space in Greenland's firn package has important mass balance implications for the ice sheet. The runoff/retention of surface melt depends on the detailed physics of meltwater infiltration, which remains a central challenge for the field (van As, 2016). While this manuscript does not present any fundamental scientific advancements towards overcoming this challenge, the intercomparison is an opportunity to synthesize the present state of the modeling field and range of output from various published models. This, I believe, has merit and is worthy of publication.

The topic is certainly of high relevance, but I believe the manuscript requires major revisions before it can be considered for acceptance. In its present state, the manuscript fails to present model output which is most relevant to the community. This could be alleviated by refocusing the results section with greater attention to the details of simulated infiltration between the different models. Additionally, I found the Discussion to be quite unorganized, reading largely as a scattered list of model details, and introducing new results. Reorganizing the Discussion, detailing how different infiltration schemes lead to wide differences in modeled density and temperature, and providing better synthesis of the model results with the objective of outlining directions for improvement, would substantially improve the manuscript.

These primary concerns are detailed below. In addition, I have included specific comments where language, content, and figure edits would improve the manuscript.

Major concerns

Presentation of results

As the title of the paper states, the manuscript presents a comparison of firn models with explicit schemes for accommodating the infiltration, refreezing, and runoff of surface melt. This distinguishes the study from past firn model intercomparisons (e.g. Lundin et al (2017)), and compels the authors to present intercomparison results that leverage the unique capabilities of the models. Specifically, I believe the manuscript would be of much greater interest to the community if the authors present meltwater infiltration and refreezing results explicitly. In its current form, the intercomparison results are focused predominantly on density and temperature. Minor focus is given to infiltration, refreezing, and retention, and the reader is forced to interpret the differences in infiltration/refreezing among the models through the density and temperature results.

The depth of infiltration at the different sites/years, and the annual depth distribution of refreezing are quantities that would be very useful in presenting the range of simulated meltwater infiltration among the models. I recognize there are few observational validation metrics, but the authors do have the upward-looking radar results (I found this figure to be the most informative in the manuscript). But even so, one of the objectives of the manuscript I believe is to communicate the range of meltwater infiltration in the retention models that are out in the community. Since infiltration/refreezing is also of first order importance in controlling the simulated density and temperature range, a greater focus on these quantities provides a much richer context for discussing the density and temperature results. This is especially true considering that observations of these variables are more common.

Organization of discussion
Much of the discussion reads as a scattered list of interesting things the authors noticed in their results, rather than a synthesis of the model results with direction for the community. This is especially true of section 5.1. I think the authors have an opportunity to use detailed infiltration results (see above) as a foundation to discuss why density and temperature are so variable. The Summit density results show that the underlying densification schemes in the models are similar, even when temperatures span a wide range. So the large range in density results at percolation zone sites is due primarily to differences in infiltration. This point would be strengthened if the authors present the refreezing results. By framing the density/temperature output within the context of infiltration, and comparing these (density/temperature) fields to observations, the authors can expand on how the general infiltration approaches perform and comment on directions for improvement. This approach seems more logical than the scattered listing of 10 details and improvements. It may also feed more naturally in to the discussion of density and temperature uncertainty in models.

Section 5.2 focuses on uncertainty in density, temperature, and mass balance from the suite of model output. I found the mass balance uncertainty particularly interesting, but I also consider these to be results, and should be presented as such in section 4 (Results). The Discussion section here is a unique opportunity to discuss the implications of the infiltration schemes for quantities that are of perhaps the most importance to many -- runoff. The authors offer discussion of specific sites but any sort of upscaling is lacking. Moreover, the results from Dye-2 are used to predict future conditions at the site, which I believe is both a non-sequitor and not necessarily supported by the results. Based on the presented results from KAN_U it is apparent that the different infiltration schemes result in a huge range in runoff/refreezing from model to model. If the Machguth (2016) results are to be accepted, then it suggests that perhaps the retention models underestimate runoff. However, Machguth et al. (2016) also define Dye-2 as a deep percolation site outside the runoff regime. Yet, the DTU model, which shows the highest runoff at KAN_U, also indicates that 30-40% of surface melt runs off at Dye-2. Based on: a) the wide spread in modeled runoff, and b) the fact that the model which supposedly most closely honors the interpreted runoff at KAN_U also appears to overpredict runoff at Dye-2, my interpretation is that the current generation of infiltration/refreezing models do not have the capacity to quantify runoff with fidelity. Is that an accurate assessment? If so, this would be a very important point to make to the community.

I struggled with section 5.3. It outlines many shortcomings of the models' ability to treat firn aquifers but does not draw explicitly from anything in the Results. The Results section simply states that the case of the firn aquifer is discussed in 5.3 (line 312). If the authors are to discuss model shortcomings, then they need to at least support these criticisms with presented results. Restructuring the results section by site (presenting infiltration, density and temperature site by site) would be one way to help achieve this. This way, for instance, the reader will be primed on whether models simulate firn reaching the melting point down to 20m as is shown in observations. This appears to occur in a number of models in Figure 5 but the text in the Discussion implies that this may not be the case.

Specific comments

ln 28: change '...capacity to retain part of the surface meltwater...' to '...capacity to retain part or all of the surface meltwater...'

Abstract: The abstract contains only methods and results. No synthesis of interpretations or summary statement.

ln 44: Consider replacing 'seen' here and elsewhere in the manuscript. 'occurred'?

ln 45: See comment re: ln 44

ln 60-65: In the first para, the importance of firn for meltwater retention and observed changes are established. In this second paragraph it would be stronger to communicate the challenges associated with meltwater infiltration in firn. Reijmer presented early work modeling infiltration/refreezing. But new models are on line with differing and, in some instances, more physically based schemes for meltwater infiltration/refreezing. Others have also tested different melt infiltration schemes. What about Steger (2017)? The authors need to do a better job of placing this work in the context of existing comparisons of meltwater infiltration/retention.

ln 66: Why just 'some' of the models currently used on the GrIS? What other models are the authors explicitly *not* using, and why is this the case?

ln 70-73: I would consider reframing the questions to more explicitly treat the different infiltration schemes. Isn't this the point of the study, to compare the infiltration schemes of different model schemes and the impact on the firn framework (density and temperature), and resulting runoff/refreezing partitioning?

ln 117-119: Reeh (2005) limits melt to the annual layer while the bucket approach redistributes based on cold and liquid water content. Seems to be contradictory?

ln 124-125: Is this statement relevant to the output? It appears to be a technical detail of the model mechanics that is a non-sequitor.

ln 141: Vandecrux (in review) does not appear in the references?

Table 2: GEUS model, Runoff Calculation -- change 'identical' to 'adjacent'? Also, DMIHH Meltwater routing is empty. Should this also be Darcy's law?

ln 210: The authors state that 'as many boundary conditions as possible' were given to all modes. Is there any guarantee that the models are all being forced by the same surface melt? This is imperative to understanding the results. Hopefully the answer is 'yes'. I believe, based on reading of the supplemental, that this the case for all but the Meyer and Hewitt model? I think this should be stated explicitly in the manuscript. Moreover, the authors should state in the main manuscript an estimate of the high/low melt bias in the Meyer and Hewitt model.

Figure 4: With such narrow panels this plot is impossible to decipher. Are models grouped and cores repeated 3x for clarity? What about 2 plots for each core? 1) the measured core density, and 2) the difference between measured and modeled for each model?

Figure 6: Again, the plot is indecipherable. All the text overlaying the plots implies that the visuals themselves are unimportant. If so, then just present the ME, RMSE, and R2 in a table and be done with it.

ln 303: 'more' is very qualitative. The results also appear to be very model-dependent at KAN-U. So what is 'more'? Consider just eliminating it.
ln 319-321: I found this lead-in confusing b/c the next figure call (Figure 7) references is all sites. Further, the next Figure 8, is focused on Dye. As stated above, I would consider restructuring the results by site.

Figure 8: State in the caption that a, b, c are all the same. Just grouping models for clarity.

ln 339-340: awkward sentence

ln 369: what is 'dissipate contrast'?

ln 382: What is inappropriate? Nearly all of the models use a thermal conductivity referenced in the literature.

ln 384: Incomplete sentence.

ln 407-409: Interestingly, Charlampidis et al. (2015) show what they believe to be refreezing below this very same ice slab. Just delayed in time.

ln 414-415: This could then indicate the challenge of comparing a firn model, which operates over such large spatial scales, with measurements of density that are local and prone to large spatial variability.

ln 526-527: These are new results in the conclusion!

ln 532: The paragraph leading to this statement summarizes what I believe to be the primary purpose of the manuscript. Synthesis of the model results beyond site-by-site agreement would vastly improve the paper. The final sentence is also incomplete.

Supplemental: Considering the revisions required in the manuscript, I have not gone through the supplemental material in detail. However, figure call-outs do not follow a consistent order (e.g. Figure ST1 is followed by Figure S1).