Review F. Saito

This paper discusses uncertainties in the simulated century-scale projection of the Antarctic ice sheet. A multi-model ensemble approach by the SeaRISE project, which has a large dispersion of the projection, is qualitatively evaluated. MISMIP experiment is proposed as a benchmark in order to filter the models whether or not have an ability to grounding line dynamics, which is expected to reduce an uncertainty in simulated projection. The approach of this paper is very unique and interesting. I think this paper is fairly well written and can be accepted with minor changes as follows.

We thank the reviewer for his general comment on the interest of our approach and for his remarks that help to improve the manuscript. Responses and related changes are detailed below.

One point which need more discussion is Figure 4. The authors show the uncertainty range of the models including capability of the grounding line dynamics (dark gray) is smaller than that not including (light gray). It is possible, however, that the dispersion of the SeaRISE result is (partly) due to other differences among the participants. It is possible that an opposite impact is shown for the SeaRISE five models (not including grounding line dynamics) if they implement a method to represent grounding line dynamics. Actually I doubt it personally, but I suggest to include more clear explanation for the core message of this paper.

The reviewer is correct; part of the spread comes certainly from other differences within the models than specifically their ability to cope with grounding line dynamics. However models selected against MISMIP3d criterion show also such differences in their numerical schemes, boundary conditions… but the multi-model spread is much lower. It is one of our arguments to point that grounding line dynamics is a key issue. It is now explicitly written that part of the spread may come from many other sources.

Another point is the SISM experiment configuration. An accumulation field is provided by the SeaRISE. Why not use this dataset for the SISM boundary conditions? Actual computation methods of the surface mass balance vary among the SeaRISE models, but at least all of them seem to based on this field. It is not necessary to reperform the experiment, but is to describe and discuss the influences of different boundary conditions. (Although I believe the difference of the surface mass balance has less impact on the conclusion of this paper).

To our understanding this point is pretty similar to the previous one and is related to the fact that other differences in the model are parts of the spread in the projection. As mentioned, we clarified this point.

Regarding specifically the accumulation field, we now force the model with the observed mass balance; climate forcings were left out, because this is also to make a point: the mass changes due to the perturbations are all much larger than those due to the climate forcing. This way, we filter out the dynamic response of the ice sheet, and not its climate response.

Some minor points
P2626 L11. ‘biais’ → ‘bias’.

Done

Eq. (2) Computation of A depends also on other factors (e.g., age) than the temperature in some models. Also AIF spatially tunes this factor in a sense.

To our knowledge, A is solely a temperature dependant facto potentially adjusted through an enhancement factor E. It is now more clearly stated in the manuscript.

P2632 L10. ‘a least’ → ‘at least’?

Done

Sec 2.3. Not enough description for SISM. What value is used for the rate factor A? Is it spatially uniform? What combination of the coefficient and the exponent is used for Weertman type basal sliding? How do you determine the basal sliding grids, or basal sliding is imposed on all the domain?

We now give a complete description of the physics and numeric of SISM. The details are now put in Appendix A. At this point we would also like to remark that during this review process we discovered an error in the calculation of the VAF (volume above floatation) in the SISM model. This is now corrected. The result is that we have lower values for the mass loss due to grounding-line melt and our results are better embedded within the SeaRISE spread. This does not alter our results, discussion, nor conclusion. Since it is an isothermal model, sliding was allowed on the whole domain, but given the very low driving stresses in the interior, its impact on overall ice dynamics remains limited.

Section 3.1. It may be better to split into the two: SeaRISE and the others (PIG).

We partly followed reviewer recommendation. Section 3.1 is now split into two subsections, on for each modelling initiative.

Fig 1 or Tab.2. In this manuscript, Fig. 1 appears earlier than Tab.2 in the main text. So it is better to move the description of sea-level computation from the ice-sheet volume, to the caption of Fig.1.

Done

Fig. 1. Is it possible to plot SISM results also in this figure?

Of course, it is in principle possible. However we refrain to do so. The argument going with figure 1 is to show that coastal changes are unavoidable when regarding large contribution to SLR (“having models able to cope with grounding line dynamics is a prerequisite before establishing projections of upper bound dynamic contribution of the Antarctic ice sheet to SLR »). We believe that it weakens the argumentation if we add the result of a model which is explicitly described as unable to model coastal changes (« Grounding line dynamics are not explicitly included in SISM… Hence, a marine ice sheet instability (retreat of the grounding line on a retrograde slope in absence of melt perturbation and significant buttressing) is not simulated with this simplified)

Tab. 2. I prefer to see the SeaRISE mean/SD and the SISM results, not the mean/SD of SeaRISE and SISM.

Discussion related to Table 2 shows how adding one ice sheet model (SISM) may impact mean and standard deviation of the model ensemble. This is the reason why Table 2 shows mean and SD for SeaRISE alone and SeaRISE + SISM. It seems thorny to follow exactly reviewer recommendation without substantially rewriting the discussion and probably loose the point. However, Table 2 has been adjusted and contribution of SISM alone have been added.

Fig. 2. Unit of the time (year) is missing in the caption or the color bar.

Corresponding label has been added.

P2639 L23. ‘SSA models reacts...’ better to add more explanation.

Corresponding sentence has been changed to “Grounding line of SSA models moves faster in reaction of a perturbation”.

Fig 4. Dark/light grey are not explained in the caption.

Done

Reviewer 2

By combining and comparing output from ice dynamics models with a large range of complexity, the authors argue that future inter comparison projects, which use model ensemble means to improve predictions of ice sheet mass loss, should only contain MISMIP (and its successors) approved models. They show that results from a simple model with a biased (i.e. wrong) description of the grounding line dynamics are not significantly different from results previously obtained in the SeaRISE inter comparison, which mostly employed models that did not comply with MISMIP requirements. They further show that the spread in predictions of future mass loss of Pine Island Glacier is largely reduced when only MISMIP approved models are included in the ensemble.

I agree with the authors that recent advances in model verification (incl. MISMIP and MISMIP3d) have unambiguously shown the importance of a correct description of the grounding line dynamics. Only when this dynamics is described correctly can we as a community attempt to proceed towards model validation and reliable predictions. In this light, this is an important paper, which provides newly acquired evidence that only models which contain the correct physics, should be used in ensemble predictions. Other models might produce equally compelling results, though for the wrong reasons. Although such conclusions have been reached in less clear terms before, this work sheds a new (qualitative and quantitative) light on the matter, and publication of this work is recommended.

We thank the reviewer for his very positive comments on our work. Adjustments of the manuscript as proposed by the reviewer are detailed below.

The authors provide two important test cases, a pan-Antarctic and a basin scale case. Only for the basin scale case, a comprehensive comparison is provided between MISMIP and non-MISMIP models. It would also be interesting to see how MISMIP approved models, such as [Cornford et al., 2015, doi:10.5194/tcd-9-1887-2015] perform in the pan-Antarctic case. As such results have been obtained in the aforementioned reference, it is presumable not a major effort to include those. Do you expect similar conclusions to emerge as in the basin scale case, i.e., are the mean mass loss and the spread reduced?

Indeed a similar study at the pan-Antarctic scale would be of interest. However, even with the most recent publications there are today not enough models to proceed to such a study. Cornford et al., 2015 publication is only focussing on west Antarctica. Today, to our knowledge, only the PennState3D model has been MISMIP-verified and run at the continental scale.

We would expect similar results at this scale. This can be presumed based on figure 2, as PennState3D heuristic model (panel c), known to be faster than other MISMIP-verified approaches (e.g. Fig 4), exhibits much smaller retreat on East-Antarctica when compared to the other models.

In the abstract, and later in the main body of the manuscript, there seems to be a contradiction. On the one hand, it is stated that the "representation of the grounding line dynamics is essential to infer future Antarctic mass change" and yet, "the biased model (with wrong GL dynamics) can hardly be discriminated from the ensemble based on its estimation of volume change". Does this mean that GL dynamics is important, but models without GL dynamics can produce similar predictions in mass loss (but for the wrong reasons)?

Absolutely, a model with an inappropriate description of grounding zones can fortuity mimic results exhibiting ungrounding of large regions. Or not! This is the reason why we argue that such models should preferably be excluded a-priori from SLR projections. So, this is not a contradiction in our argumentation. This has been clarified in the abstract.

Although differences in described physics seem to dominate the spread in Figure 4, even for models with the correct first-order physics (dark shaded area), inter-model differences are large compared to differences within a single model for e.g. different melt rate parameterization. This was a surprise to me. Should this be a reason for concern, as ultimately we want to obtain a range of predictions based on uncertainties in physical parameters such as mass balance, sliding etc., rather than model design? Does this mean that only full-Stokes models should be used, as they describe the full physics?

This is a good point. We however believe that it remains difficult to be firmly conclusive with current knowledge for at least the two following reasons:

(i) the number of models belonging to the same category is very limited (one full-Stokes, one heuristic, two SSA and two L1Lx). Classification versus the physics appears consistent with MISMIP3d intercomparison but remains to be confirmed. As an example (mentioned in the text), the two SSA models covers almost completely the spread of verified models (Fig 4, black and red squares in the dark grey envelope). Would it be similar with other full-Stokes models?

(ii) Most of the models we compare use different set-up, grids and forcing. So not alone the physics certainly impact the spread in a manner that can hardly be untangled. Typically numeric may have a very significant impact (e.g., different resolution to sample the grounding line).

We agree that in the future, the community should try to reduce the inter-model spread through specific well-designed intercomparison. This is now more clearly mentioned in the manuscript.

A minor comment about the Figures: It is difficult to see the shading of blue in Figure 1. Similarly the dashed black and red lines are difficult to see in Figure 4.

Colorscale and symbols have been adjusted on both Figures 1 and 4 to improve clarity.

2 small typos: abstract l11: biais -> bias p2632, l10: a -> at

Done