We thank the reviewer for their constructive criticism of our manuscript, and would hereby like to respond to their concerns. Their comments are shown in italics, our response in regular type.

General assessment

> *While the authors address most reviewer concerns in the response and the revised manuscript, the additional speculations and discussions are still based on the original set of experiments. I was hoping that some additional test simulations were possible to quantify the effects of the size of the problem (see below) or the mesh refinement frequency on scalability (Sect. 3.1) or of a wider mesh around the grounding line (Sect 4.3).*

We have performed an additional set of simulations of the scaling experiment (i.e. the spin-up phase of the modified MISMIP experiment), this time with a grounding-line resolution of 2 km instead of 8 km. The poor scaling persists, suggesting that there is indeed an underlying problem with the parallelisation. As we currently do not have funding for IT support anymore, investigating this issue further will be reserved for future work. We will mention this in the manuscript.

We have re-done the MISMIP+ ice1r simulations with a wider high-resolution band around the grounding line for the high-resolution (< 2km) versions, resulting in less-frequent mesh updates. This does not seem to have much effect, suggesting that the numerical diffusion resulting from the remapping is not that important after all. Another suspect is the fact that the flow factor is tuned for each simulation individually to achieve a steady-state grounding-line position at x = 450 km (following the MISMIP+ protocol), so that the different simulations have different flow factors. We have performed an additional set of experiments using the same flow factor for each resolution (see Figure below). This does indeed result in the retreat curves in Fig. 6b being more parallel, but at the cost of a larger spread in initial grounding-line positions (ranging within ~435 – 445 km), so that the final spread is not reduced. We will mention this in the manuscript.

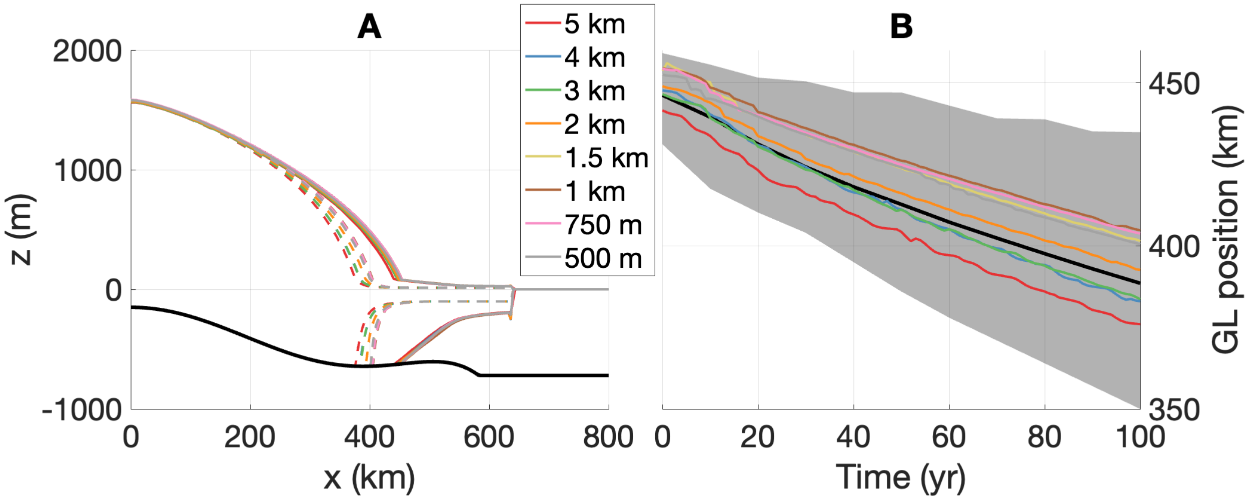

Fig. 1: results of the MISMIP+ ice1r experiment with the same flow factor for every resolution.

> *The authors now emphasize that DIVA is the default stress balance approximation used in the experiments (and not BPA). As outlined earlier in the review, there is some fundamental critique in parts of the ice model community on the applicabilty of DIVA in general, and it is therefore good to gain some more experience. However, the authors mention in their response, that "validating the equations themselves" would not be the focus of their manuscript.*

We maintain that a fundamental study of the relative merits of different approximations to the Stokes equations, is beyond the scope of this study.

> *l. 166: "UFEMISM currently does not include a stress boundary condition at the ice front for any of the momentum balance approximations. Instead, it uses the "infinite slab" approach…"*
> *Doesn't this imply numerical diffusion of the front? And if yes, how would you then define a calving front (line segment) in the model, e.g. for calving experiments?*

While we have not performed such experiments with UFEMISM, earlier experiments with our older model IMAU-ICE showed that using the infinite-slab approach resulted in very small differences in the modelled velocities, compared to using a stress boundary condition at the calving front. We will mention this in the manuscript.

Although the conservation-of-momentum solver uses the infinite-slab approach (by 'imagining' a very thin – 10 cm – ice shelf in areas that in reality are open ocean), the conservation-of-mass solver does see the difference between ice-covered and ice-free vertices. The calving front can then be identified per edge, and calving can be applied there. We have contributed simulations with UFEMISM to the ongoing CalvingMIP project, and our results seem to agree well with the other models' submissions so far.

l. 424: "Another contributing factor could be that the model set-up used for the scaling test was too 'small' (i.e. had too few vertices), so that the communication latencies between cores begin to dominate the total computation time."
This would be worth a few test simulations, not over the full length.

See above.

l 141: Antarctic ice sheet → Antarctic Ice Sheet (also l 427)

Fixed.

l. 198. we used to DIVA

Fixed.

l. 468: extrapolation→ extrapolated

Fixed.

l. 485: "perfect restart" → this would be a prerequisite of reproducibility from (intermediate) restart states

Yes.

Fig C1: panel C may not be needed

We will remove panel C.