Reviewers' comments are *in italics*, while our replies are in roman text. Serif font indicates our action.

Reply to Reviewer 3

We thank reviewer 3 for providing a detailed review.

This paper advances the concept of hindcasting, or predicting prior observations, as it may be used to assess the skill of ice-sheet models. This is the first paper to do so, probably because others have avoided the issue due to a paucity of data and concerns about the differences between glaciological and observational time scales. Even as more data have become available, the questions remain, and prevent this paper from having quite the impact it is trying to achieve. Additionally, I have real concerns about what is being attributed to the ice-sheet model vs. what is attributed to the climate forcing. Nevertheless, in this paper I found a reasonably well-defined plan for hindcasting with ice-sheet models, which I am sure will be dissected and reassembled by future investigators. I also found several useful guidelines relating to model initialization, and real insight into the tradeoffs associated with different initialization methods. Finally, the work has been carried out with an attention to detail regarding the input data sets, and the model runs are openly available to the public through the PISM repository. These facts do much to assure the work has a legacy, and is followed up on by other investigators. Hence, I recommend publication of the manuscript, but only after some of the significant issues having to do with time scales, and surface mass balance vs. dynamical thinning have been addressed in a substantive way.

The reviewer is right that the issue of validating ice sheet models by simulating the recent past has not been thoroughly addressed before. This remains a challenge and our paper is a first attempt to tackle this. As outlined in our general reply, we rewrote the introduction to better define the scope, and discuss the current limitations of our approach in the discussion.

First, the issue of time scales. There isn't much question about if hindcasting is useful. Plainly, it would be; any means of improving the quality and reliability of ice-sheet models highly desirable. The question is how hindcasting could be done with ice-sheets which are characterized by changes taking place on very long (decades to centuries) timescales, and a very short, incomplete observational record. So, in the language of the paper: what are the "known inputs" for "past events". Presumably this is the 8 year GRACE record, the present day surface and its rates of change, and the present day surface speed. All based on a short, 20 some years of climate forcing. In some cases the model output matches observations well, and in others it doesn't. Given the complexity of the ice sheet models, and the vagaries of the model initialization processes, I'm not at all confident the time periods being compared are reasonable, and am not convinced that a good or poor match

(however those are defined) to those observations reveals much. The authors do attempt to address the short time periods; considering model "drift", and discussing the potential for transients to be introduced into the system as the climate forcing is changed. In the end I still think more is needed. The ergodic hypothesis is that the time average is the same as the ensemble average. In this experiment we don't have a large ensemble, or a long time period. Why not continue the runs another century under or more (under the same climate forcing) and see how well the trends hold up? If they don't hold up, the authors should address the issue of why. The authors should dispel the notion that we are observing a transient, or a short term fluctuation in ice dynamics that is not representative of the systems true behavior.

As outlined in our general reply we reworded the introduction to better define the scope of our paper.

Regarding the comment "I'm not at all confident the time periods being compared are reasonable, and am not convinced that a good or poor match (however those are defined) to those observations reveals much." We acknowledge that there is an open question of the time-scale of hindcasting, however, whether the time periods compared are reasonable is beyond the scope of our paper. More importantly, the duration of hindcasts is currently bounded by the length of observational records. Regardless of the ideal hindcast duration, validation using hindcasting allows one to use time-dependent observations gauge relative performance of models or model choices. As outlined in our general reply, we added a discussion about the currently limitations of hindcasting to the Discussion section.

We agree with the reviewer that more work is needed. In fact we see our manuscript as a step towards better methods for ice sheet model testing. Because using hindcasting as a method of assessing performance of an ice sheet model system is a relatively new concept, our paper may indeed raise more questions than it answers. There is not doubt that future work will explore hindcasting further and develop other frameworks for model evaluation. We made this more prominent in the rewritten conclusions. Also, we now discuss the limitations of our approach in an additional paragraph in the conclusion section (see our general reply).

We wonder what the reviewer means with *"Why not continue the runs another century under or more (under the same climate forcing)"*. In our understanding, the "same climate" forcing refers to the 1989–2011 climate forcing from HIRHAM5. In this case we don't see a way how this climate forcing could be meaningfully applied to continue the simulation. On the other hand, if the reviewer meant that it would be useful to perform prognostic simulations with GCM-derived climate forcing for the 21st century then we agree that it is an interesting experiment, but not the scope of this study. 21st century projection runs are, however, made in a companion paper by Aðalgeirsdóttir and others (manuscript in prep.)

Second, issues related to surface mass balance vs. dynamic thinning need to be addressed more clearly. PISM is an ice-dynamics model, giving estimates of dynamic thinning.

GRACE is measuring surface mass balance and dynamic thinning. Before the comparisons can be made, GRACE should have the SMB signal removed. I get the impression, from figure 5, that the ice dynamics is very sluggish, and on the 20 year time scales, accounts for an insignificant part of the total dynamics. As such, I'm not even convinced that this is a paper about hindcasting with ice-sheet models. Rather, it may be more a paper about differences in the SMB emerging from different dynamic land models and climate parameterizations. The authors need to confront this more directly. The initializations are different, and the SMB model forcing the ice-sheet in each of the three cases is not quite the same. Lapse rates are applied, and in some cases (flux corrected) anomalies are used. These differences in how the SMB is applied may account for most or all of the differences reported. This might be treated discursively, and parts of the paper related to this should be re-written. It is possible that I've somehow missed the point, but the authors should concede that the presentation is confusing.

In our general response we write that we are not evaluating the ice dynamical part of PISM, but rather consider the ice sheet system as a whole. As mentioned there, other combinations may give different results.

GRACE indirectly measures variations in total ice sheet mass (after accounting for changes in water storage and glacial isostatic rebound) without differentiating between processes that lead to the mass changes. Therefore mass changes from GRACE should be compared to simulated changes in total ice mass. Current whole-ice sheet models are not capable of simulating the (observed) increase in ice discharge due to perturbations at the lateral boundary. Admittedly this complicates the evaluation of the performance of an ice sheet system model.

The surface kinematical equation comprises two terms, flux divergence (divergence of the product of vertically-integrated horizontal velocities and ice thickness) and climatic mass balance (for simplicity we ignore basal mass balance). We demonstrate that ice sheet models are capable of transforming monthly climate forcing in a realistic way into surface elevation changes. This study shows that the ice dynamic sub-system does not transform the applied climate forcing field into an unrealistic response.

We are not quite sure what the reviewer means with *"I get the impression, from figure 5, that the ice dynamics is very sluggish, and on the 20 year time scales, accounts for an insignificant part of the total dynamics."* If this is a question about whether inter-annual variations in ice dynamics or climate mass balance are larger, then we agree that inter-annual variations in climatic mass balance dominate the signal.

We only apply a lapse rate correction to temperature, which is used as a boundary condition for the enthalpy equation. No lapse rates are used for climatic mass balance. Therefore, the same climatic mass balance forcing is applied to both "constant-climate" and "paleo-climate". We have explored the consequence of applying the climatic mass balance as anomalies for "paleo-climate" (p. 5078, l. 4–7). For "flux-corrected", anomalies are needed. Otherwise, switching from the initialization climate (using flux correction)

to the ERA-interim forcing would result in a shock. Anomalies are expected to reduce this shock. However, as can be seen from Table 2, the shock remains relatively large.

Lateral boundary conditions, and calving criteria in particular need to be addressed in the paper. The reader needs some assurance that the observed behavior is not arising from whatever is happening there.

Lateral boundary conditions and calving criteria are explained in the Supplement. Figure 5 of the main paper shows that ice discharge of all three hindcasts remains nearly constant over the hindcasting period. Therefore the simulated mass loss is not strongly influenced by lateral boundary conditions and calving criteria.

A final point is that I think it's misleading to report the thickness and surface height for flux corrected models. Either it's going to be very close, or the flux correcting scheme is wrong. Reporting it with graphical weight equal to the other runs gives a casual reader the impression that you've got things working well, when in reality all that works is the flux conservation.

We report ice thickness and surface height for "flux-corrected" on purpose; first to show that the flux-correction method works and, second, to point out that one *might* get the wrong impression that things are working well by comparing against too few data sets. Both in Table 1 and Figure 2 we explicitly state that "flux-corrected" is only shown for comparison but is not available for validation.

p. 5072:, **l. 15:** *equilibrium with modeled present-day climate*

Changed as suggested.

p. 5072:, **l. 20:** *mass balance from interpolated surface temperature... and model constrained precipitation.*

We changed "model-constrained precipitation" Temperature, however, is not interpolated, but a parametrization based on latitude, longitude, and elevation (c.f. Fausto et al., 2009)

p. 5073:, **l. 5:** *specify that the flux correction is the same between forward and spinup. Also, the flux correction changes, right? But it doesn't change in the forward run? Justify. How large is the flux correction? How does it compare to accumulation? Is it unreasonably large or small?*

Changed to "This initial state is not in equilibrium with the applied forcing fields (Price et al., 2011). To prevent a subsequent shock (i.e. model drift), climate forcing is applied as anomalies at the hindcasting stage (Supplement)." We updated the Supplement to

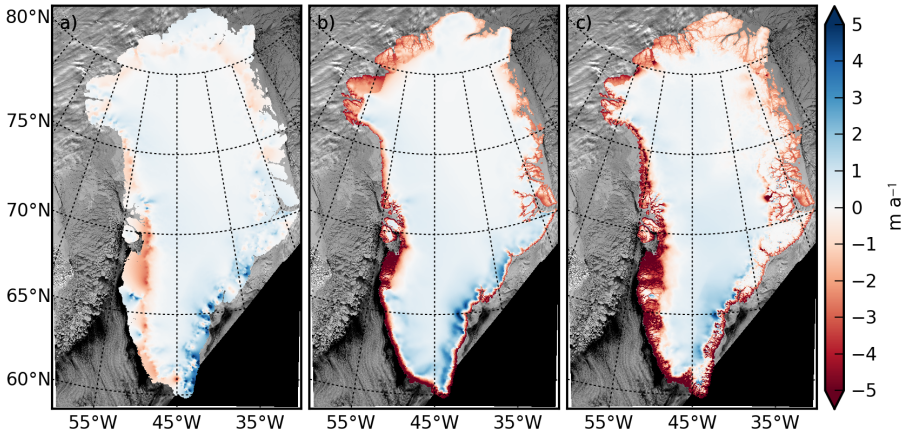

Figure 1: Mean climatic mass balance of the last 10 years of the initialization. **(a)** "constant-climate". **(b)**. "paleo-climate". **(c)** "flux-corrected".

clarify how the flux correction is applied and how we calculate anomalies. We also added a figure for comparison, see Fig. 1. Because PISM does not provide the correction term as a diagnostic model output, we show the climatic mass balance instead (for "flux-corrected", this includes the correction term). As can be seen from Fig. 1, the climatic mass balance at the end of the initialization period of "paleo-climate" and "flux-corrected" are very similar. Thus the flux correction term itself is small.

p. 5073:, **l. 26:** *"Good match" please be more quantitative! I have no idea what good is.*

We think that, in the methods section, a qualitative rather than a quantitative statement is appropriate, making the paper to be read more fluently. Later, in the results section, we quantify the agreement between observed and simulated surface speeds using root mean square error.

p. 5074:, **l. 4:** *remove 'dynamical'*

Changed as suggested.

p. 5075:, **l. 18:** *No fair making a big deal of how well the flux corrected run works.*

It is not our intent to make a big deal, we just report the results for completeness. See our reply further above. In fact we conclude that "flux-corrected" is not a good initial state as it shows unphysical transients.

p. 5075:, **l. 24:** *What does "normalized" mean here?*

We changed

"The mass change time-series (Fig. 4) are normalized to the beginning of the GRACE period (January 2004)."

to

"Figure 4 shows the time-series of mass change since the beginning of the GRACE period (January 2004)."

p. 5076:, **l. :** *This might be the place to outline how to differentiate between SMB and dynamical thinning, and how the analysis will be done.*

As outlined in our general reply we compare simulated to observed surface elevation changes. Partitioning the spatial elevation change signal into contributions from ice dynamics and surface mass balance is not the scope of our paper.

p. 5076:, **l. 26:** *a too weak—reword this.*

Changed to "Despite having comparable ice volumes, the three initial states respond differently to the applied climate forcing, and the difference is significant. Therefore reproducing observed ice volume is not a sufficient metric. "

p. 5077:, **l. first two paragraph:** *rework this entirely. It's a critical component of the paper, but in its present form, very challenging to understand. Please consider the points made above, and attempt to clarify the partition between mass balance and ice dynamics.*

We replaced

"Trend under-estimation is expected because of the absence of ocean forcing that could lead to an increase in ice discharge."

with

"As mentioned earlier, observations show a rapid increase in ice discharge since the late 1990s, which was attributed to changes at the ocean boundary. Our model does not include ocean forcing that could lead to an increase in simulated ice discharge. Therefore, under-estimation of the simulated mass loss trend is expected."

p. 5077:, **l. 17:** *Not sure what is being referred to here 'before ice discharge started to increase rapidly'...*

We believe our suggestion above clarifies this question. We removed "'before ice discharge started to increase rapidly"

p. 5078:, **l. 18:** *-50 Gt "constant climate" OK, but then how do you reconcile the previous page's statement on lines 25-28? Is it ALL due to changes in climate forcing? Some appears to be due to drift, right?*

On p. 5077, l. 18, we changed

"The constant-climate initialization is in equilibrium with its climate, and, consequently, surface elevation changes must result from the climate forcing, either directly through the applied climatic mass balance or, indirectly, through dynamical adjustment due to changes in climatic mass balance. "

to

"The "constant-climate" initialization is expected to be in equilibrium with its climate, and, consequently, surface elevation changes should result from the climate forcing, either directly through the applied climatic mass balance or, indirectly, through dynamical adjustment due to changes in climatic mass balance."

We also changed

""Constant-climate" is in equilibrium with the present-day climate."

(p. 5072, l. 15–16) to

""Constant-climate" is expected to be in equilibrium with the present-day climate."

p. 5078:, **l. 25–28:** *move to conclusions, this is an interesting result.*

We agree that this is an interesting results regarding the performance of the "flux-corrected" hindcast. However we think it is mostly a matter of taste whether the Discussion section or the Conclusions section is the more appropriate place.

p. 5080:, **l. 7–10:** *I'm just not convinced the conclusions are as strong as you make them out to be. Spend some time in the conclusion discussing the shortcomings of the approach.*

As formulated in our general reply, in the new introduction we now clearly outline the scope of the paper. Also, we added a paragraph to the conclusion section discussing the limitations of our approach.

References

Fausto, R. S., Ahlstrøm, A. P., Van As, D., Bøggild, C. E., and Johnsen, S. J.: A new present-day temperature parameterization for Greenland, J. Glaciol., 55, 95–105, doi:10.3189/002214309788608985, 2009.

Price, S. F., Payne, A. J., Howat, I. M., and Smith, B. E.: Committed sea-level rise for the next century from Greenland ice sheet dynamics during the past decade, P. Natl. Acad. Sci. USA, 108, doi:10.1073/pnas.1017313108, 2011.