Review of a manuscript "Hindcasting to measure ice sheet model sensitivity to initial states" by A. Aschwanden, G. Aðalgeirsdóttir, and C. Khroulev

This study aims to quantify sensitivities of a large-scale ice-sheet model (PISM) to the ice-sheet initial configurations for short-term simulations (22 years). The authors propose to use hindcasting as a method of quantification. Although the investigated problem (sensitivities to initial state) is extremely important and needs to be addressed, and the manuscript have interesting thoughts and considerations, it cannot be published in its present form. The manuscript requires substantial revisions at least in presentation, if not in substance (methods and simulations).

General comments

The manuscript compares results of three simulations performed for a very short (∼20 years) period that are initiated from different states obtained in the following ways: the first one running the model for 125 kyrs with the present day climate forcing, the second one with the "paleo-climate" and the third one using the "flux-correction" method. All three simulations are run till the year 1989 using climate parameters specific to a particular simulation. The next 22 years of simulations are done using the same climatic forcing in all simulations, and the obtained states are compared to observations at the year 2011.

There are several major concerns with the study itself. They are listed here in no particular order of importance. First, the hindcasting period of ∼20 years is extremely short by glaciological standards. Apart from the fast flowing outlet glaciers, whose dynamics is barely resolved in this model due to a number of reasons (e.g. resolution, the absence of appropriate boundary conditions/forcing at fjord terminated glaciers etc.), that can move some appreciable distance (∼10-60 km), the rest of the ice sheet hardly moves at all, perhaps, remaining within the original model grid cell. Hence, all possible changes that simulated during this hindcast period are due to the surface mass balance, which is identical in all simulations. Therefore, the authors could have easily omitted this hindcasting stage, and compared the ice-sheet configurations at 1989. If this assessment is incorrect, the authors need to demonstrate that significant *dynamic* changes happened in these 20 years of simulations. In other words, the authors need to demonstrate that hindicasting for such a short period of time is indeed a reasonable approach to test sensitivities of ice-sheet models to their initial states.

Second, the climate forcing used during the hindcasting stage come from an atmosperic model (HIRHAM5). Undoubtedly, as any model, it has errors and biases. Since the changes during the hindcsting period are dominated by the evolution of the surface mass-balance, one could argue that the effects of errors in climate forcing on the final states is substantial. There is no analysis or discussion of such errors or performance of this model. How different would the results be if outputs from other atmospheric models would be used as climate forcings?

Third, the choice of ways to obtain initial states is not obvious. Though, it **is** the authors' choice, and they are free to use any approaches, the manuscript does not provide any discussions or justifications for these approaches. It is clear, that the climate was not identical to the present-day climate for the past 125 kyrs, therefore, it is reasonable to expect that an ice-sheet state obtained with the "present-day" approach is quite unrealistic. In the "paleo-climate" approach, there are many poorly known and unconstrained parameters (not mentioning the validity of the degree-day method) that also result in an ice-sheet state that is, most likely, very far from the present-day state. As the authors mention, the “flux-correction” method is dynamically inconsistent, therefore, it is natural to expect that in the hindcasting simulations the ice-sheet system will be rapidly changing in order to adjust to the internally-consistent state resulting in unrealistic ice-sheet configurations.

Forth, the comparison of the final states to observations, although not surprising, in my view, is not very informative. There are many other issues (e.g. unknown and/or unresolved physics, unknown/unresolved boundary conditions at the ice-sheet margins and bed) that can result in the simulated present-day states that are very far from the observed one. As a side comment, comparison to the GRACE observations, most likely, cannot be treated as an appropriate metric, or at least cannot have the same weight as comparison to other observations (e.g. surface elevation and its changes, surface velocities, ice sheet extent, etc.). This is due to the GRACE coarse resolution and a number of issues in processing (e.g. postglacial rebound, etc).

To summarize general comments: there is no doubt that short-time (decadal to several thousand years) simulations of ice-sheet behaviour are strongly affected by its initial states. How to obtain “good” initial states (whatever that means) is a large and unresolved problem, which is not addressed in this study. The proposed hindcasting method **may** be a valuable, or at least, interesting approach, however, this study does not demonstrate that. By comparing the final states to the present-day observations the authors mix two issues. One is the performance of the hindcasting method, and the other is the choice of an appropriate metric that can be used to evaluate a simulated ice-sheet state. Neither of them are addressed in the manuscript with sufficient clarity. There are many questions with the former (e.g. is 20 years a sufficient period?), and the latter is barely touched. The choice of an appropriate metric or a set of metrics is an important question that has not been studied before. It definitely deserves in-depth investigations, and perhaps, should be left for further focused studies.

Of course, it is the authors’ choice how to conduct their investigation, however, if I may offer a suggestion, it would be to use a synthetic approach to illustrate the usefulness of hindcasting. One could create an artificial state of an ice sheet (either using a realistic bed topography or idealized) by forcing an ice sheet with prescribed climate conditions. This state would be the true “present-day” ice-sheet configuration. Then, one could repeat a procedure somewhat similar to what is done in this study, i.e. create different initial states for the hindcasting period and compare the different ice-sheet configurations obtained at the end of hindcasting to the true state. It would be interesting to see what the optimal hindcasting period might be. My hunch is that 20 years is too short. By doing so, the authors would be able to truly isolate the effects of initial states and investigate/demonstrate advantages of the hindcasting approach.

Specific comments

Page 5070

Lines 4-5: what is “the quality of projections” and how can one measure it?

Line 6: “. . . initial states” of what?

Line 11: what is “dynamic state”?

Line 21: “Ice sheet models integrate such physical process understanding.” an awkward sentence

Line 25: what are “spatially-reach” observations?

Page 5071

Lines 9-23: this paragraph is out of place and unnecessary, either remove it or re-write it to be stylistically similar to the previous and the next ones.
Page 5072
Lines 1-4: this paragraph is unnecessary, the paper is fairly short.
Line 6: why is 1989 used as a datum?
Page 5073
Lines 6-7: flux-correction is not used in climate models anymore.
Line 9: what does "overall dynamic state" means?
Lines 16-17: what does that mean "reduce model complexity"? What are boundary conditions there?
Line 25: join with the previous paragraph.
Page 5074
Lines 6-7: what are resolutions of spinup simulations?
Line 23-: maybe say something why these data sets are used. This might be a good place to say something about metrics.
Page 5075
Line 16 and throughout the text: suggest to use "constant climate", "paleo-climate", "flux correction" in quotation marks
Lines 24-25: what does it mean "normalized to the beginning of the GRACE period"?
Page 5076
Lines 24-26: "...the three initial states" do not "respond differently" they adjust to a shock (or a jump) in the surface forcing.
Page 5077
Lines 3-6: the last two sentences of this paragraph are unclear.
Line 11: why should the split between ice dynamic and surface processes be the same?
Page 5078
Lines 1-7: for how long has the "interim forcing" been applied? What is "ERA-forcing"? What are its errors?
Lines 23-29: This paragraph is too cryptic, either clarify or remove it.
Page 5079
Line 5: what does that mean "Surface elevation changes corrected for model drift"?
Line 9 an Line 14: once again, what is "spatially reach"?
Section 5 Conclusions needs to be rewritten. The presented conclusions have a very loose connection to the material described in the manuscript.