RESPONSE TO REFEREE #1

General comments:

This paper presents a procedure of refining large ensembles (LE's) of geophysical model simulations, iteratively narrowing parameter ranges based on fits to observations. The core of the paper is the statistically based method of refining each parameter range, Eqs. 10-12. The procedure is tested using modern simulations of the Antarctic Ice Sheet with the PISM model, run to equilibrium with modern climate, and testing on continental average fits to observed ice extents and thicknesses.

The procedure is put into perspective and is well motivated in a particularly broad-based introduction. It is well considered and should be a useful tool in many large-ensemble studies. It is explained succinctly and clearly, and the results presented for the PISM application illustrate the potential and limitations of the method.

We thank the reviewer for their positive feedback and for their constructive comments on the manuscript.

Specific comments:

As shown in Table 2, after 5 iterations the reductions in parameter ranges from those at the start do not seem very impressive - considerably less than 50% in all cases. The same is true for the reductions in error metrics Ecrit_A and Ecrit_V. I think the procedure could still be useful for the community, but perhaps with some caveats along these lines.

We acknowledge this point, but we also note that our approach succeeds in eliminating 85.4% of the possible parameter combinations. To expand upon this, and to add the caveats requested by the reviewer, we have added the following paragraph at lines 334-341 of the revised manuscript:

"During the optimisation process, the ranges for all four of the parameters used to determine the till friction angle remain unchanged. However, for the other six parameters, the ranges are reduced in width by between 14.5% (the shallow ice enhancement factor) and 44.0% (the exponent of the basal resistance model). Overall, the volume of the parameter space has been reduced to just 14.6% of the original size, meaning that 85.4% of the possible parameter combinations have been eliminated. We note that the application of the technique described in this manuscript involves a trade-off between computational expense (as determined by the ensemble size) and precision (as measured by the reduction in parameter uncertainty). Increasing the ensemble size might allow a greater reduction in the volume of parameter space, but at the expense of increased computational cost."

As an alternative to Eqs. (10) and (11), these could be combined into a single step by considering: p = ((Xmax - Xmin)/(XB - XA))^N. Could there be any advantage to this, compared to the separate consideration of the max's and min's in the paper? There is none that I can see, but it might be interesting to mention briefly (or to rule out).

This is theoretically possible. However, in the case where the null hypothesis can be rejected, this combined approach would not allow us to determine whether it is XA, XB or both that should be changed: it would merely allow us to determine that the parameter range should be reduced. As such, we consider that Equations 10 and 11 should remain separate.

The basic concept of refining parameter ranges in an iterative series of LE's is analogous to Lee et al. (2020), who use a new method of "re-sampling" parameter values at given steps within MCMC (Markov Chain Monte Carlo) sequences. However, the procedure here is considerably less complex and will be more accessible to the cryospheric modeling community.

Thank you for bringing this reference to our attention. We have added the following text at lines 411-414 of the revised manuscript:

"While it is analogous to other techniques that use large ensemble modelling to refine parameter ranges (e.g. Solonen et al., 2012; Lee et al., 2020), the approach developed here is considerably simpler and should therefore be more accessible to the geoscientific modelling community."

The large number of parameters (10) and the relatively small number of simulations (100) in the Latin HyperCube (LHC) ensembles here is a concern. Chang et al. (2014) found that coarsely spaced LHC sampling inadequately resolves parameter ranges and interactions in LE's for high-dimensional parameter spaces (their Fig. 4). One alternative in future work could be to greatly reduce the number of parameters, still with LHC's or possibly performing runs with all possible combinations of parameter values.

Thank you for bringing this reference to our attention. As acknowledged above, the application of the technique that we describe involves a trade-off between expense (ensemble size) and precision (reduction in parameter uncertainty). We gratefully acknowledge the reviewer's suggestion and have added the following text at lines 389-394 of the revised manuscript:

"The size of the ensemble presented in this study (100) is relatively small, particularly given the large number of parameters being optimised (10). Chang et al. (2014) show that a 100-member Latin Hypercube ensemble cannot adequately resolve the interactions between parameters in an ice sheet model, even when being used to study a five-dimensional parameter space. Ideally, our technique would therefore use a larger ensemble size or would be used to target a smaller number of

model parameters. While the former would increase the computational cost, either of these modifications should allow for greater refinement of parameter ranges."

The sentence of line 368-370 may be questioned: "We have also shown that regions of parameter space exist that cannot be meaningfully reduced in volume any further, given constraints arising from the availability of data and limitations on our understanding of the underlying physical system." "Meaningfully" here relies on the assumption in line 299 that the top tercile of ensemble members identifies "good" vs. "bad" simulations. That is reasonable, but the chosen fraction (1/3, vs. 1/4 or 1/5 say) is fairly arbitrary and possibly model dependent. Also considerably more observational modern data than ice extent and thickness are available for validation, e.g., surface ice velocities and ice temperature profiles in deep cores (and also paleo data as in some other studies referenced here).

We agree. In response to comments by both reviewers, we have replaced this sentence with the following text at lines 382-386 of the revised manuscript:

"We have shown that the optimal ranges for each parameter can be dependent on other variables. While we have been able to substantially reduce the volume of plausible parameter space, limitations on our understanding of the underlying physical system ensure that the plausible ranges remain large for some parameters. Using additional observational and palaeoclimate datasets to evaluate the model, such as the surface ice velocity and vertical profiles of temperature and age from ice cores, might allow us to constrain the parameter ranges further."

Some of the sentences in the concluding sections sound negative concerning the ability of LE's to address probabilistic ranges of parameter values and future-projection results (lines 379-380), or whether it has been addressed before (lines 399-400). That ability may indeed be limited due to other sources of uncertainty (lines 382-383), but is prominent in some LE studies, including MCMC applications (Chang et al. 2014; Edwards et al., 2017; Gilford et al., 2020; Lee et al., 2020). Perhaps the extent of these points could be clarified.

We agree. These statements are intended to refer to the application described in the manuscript, rather than to refer to all possible techniques for parameter optimisation. We have therefore revised the text at lines 379-380 (lines 403-405 of the revised manuscript) as follows:

"Finally, we note that, whereas the approach developed in this study allows for the rigorous quantification of uncertainty in model parameters, and therefore the quantification of uncertainty in model projections, the technique presented here does not allow these uncertainty ranges to be interpreted in probabilistic terms."

We have also expanded the text at lines 399-400 (lines 427-430 of the revised manuscript) as follows:

“The parameter uncertainties identified in this study, and in other studies that have used analogous large ensemble modelling approaches (e.g. Chang et al., 2014; Edwards et al., 2019; Gilford et al., 2020; Lee et al., 2020), represent a source of uncertainty in future climate projections. Further exploration of these uncertainties should form the basis of further work.”

Detailed points:

Lines 7-8: In the abstract, the sentence "We find that co-dependencies between parameters preclude the identification of a single optimal set of parameter values" might be misunderstood. Does it mean "the identification...is impossible with any technique", or "it cannot be done simply, for instance by varying just one parameter at a time, and so needs LEs" ? This is addressed but not quite clearly answered for me on lines 389-393.

As per our response to the previous comment, this statement is intended to be specific to the technique described in the manuscript. We have revised the text at lines 7-8 by replacing “the identification” with “any simple identification”.

Lines 284-288: Does Eq.(8) distinguish between floating ice and grounded ice? i.e., if a model grid point has grounded ice and observed has floating ice, or vice versa, do the masks M agree or disagree? I would think they should disagree, as suggested by line 287, but perhaps not from line 288(?).

Equation 8 does not distinguish between grounded and floating ice, although we acknowledge that there would be potential benefits to this. We have revised the paragraph at lines 286-291 (lines 292-297 of the revised manuscript) accordingly.

Line 325: "Convergence is achieved..." may be questioned. Only one parameter minimum or maximum changes in the last iteration (Table 2), but only one has changed in most of the previous iterations as well. If one more iteration was performed, would no parameter ranges change?

At the final iteration, the statistical tests described by Equations 10 and 11 do not result in a rejection of the null hypothesis for any of the ten parameters. As such, if one more iteration was to be performed, it would repeat the previous iteration exactly.

To clarify this, we have taken the text at lines 325-326 and expanded it to the following paragraph (lines 331-333 of the revised manuscript):

"Table 2 shows the progression of the iterative parameter optimisation process. Convergence is achieved after five iterations, at which point the statistical tests described by Equations 10 and 11 do not result in a rejection of the null hypothesis for any of the ten parameters. No further changes are therefore made to either the minimum or maximum values for each parameter."

References:

Chang, W., P. J. Applegate, M. Haran, and K. Keller. Probabilistic calibration of a Greenland Ice Sheet model using spatially resolved synthetic observations: toward projections of ice mass loss with uncertainties. Geosci. Model. Devel., 7, 1933-1943 (2014).

Edwards, T. L., Brandon, M. A., Durand, G., Edwards, N. R., Golledge, N. R., Holden, P. B., Nias, I. J., Payne, A. J., Ritz, C., and Wernecke, A.: Revisiting Antarctic ice loss due to marine ice-cliff instability, Nature, 566, 58-64, https://doi.org/10.1038/s41586-019-0901-4, (2019).

Gilford,, D.M., E. L. Ashe, R. M. DeConto , R. E. Kopp, D. Pollard, and A. Rovere. Could the Last Interglacial Constrain Projections of Future Antarctic Ice Mass Loss and Sea-Level Rise? JGR-Earth Surface, 125, e2019JF005418. https://doi.org/10.1029/2019JF005418. (2020).

Lee, B.S., M. Haran, R. Fuller, D. Pollard and K. Keller. A fast particle-based approach for calibrating a 3-D model of the Antarctic Ice Sheet. Annal. Appl. Stat., 14, 605-634 (2020).

RESPONSE REFEREE #2

The following is a review of a Geoscientific Model Development manuscript on "An iterative process for efficient optimisation of parameters in geoscientific models: a demonstration using the Parallel Ice Sheet Model (PISM) version 0.7.3" By S. J. Phipps et al.

This manuscript describes a systematic brute-force statistical approach to determining the most realistic combination of values for multiple parameters of a geoscientific model. The authors adopt an iterative sampling procedure in which they continuously down-select from a collection of ensemble members until they reach their criterion for convergence. This procedure allows them to identify the areas of the multi-dimensional parameter space that will result in the most realistic model outcome. For this paper, the authors use the Parallel Ice Sheet Model (PISM) to illustrate the utility of their approach, and they find 14 different configurations that best match with observational data. In conclusion, however, they are not able to identify a truly optimal set of parameter values due to computational limitations and the fact that a number of the variables being explored co-vary. They suggest that ice sheet models may not be able to be tuned to a truly optimal state, and that model complexity and non-linearity demand the use of ensemble modeling in order to adequately quantify model uncertainty due to variable tuning.

Overall, this manuscript is well-written and comprehensive. The theme is appropriate for GMD, especially since the outlined approach can be used for exploration of other types of geoscientific models. The methods are sound, and the authors do a good job describing the ice sheet model experiments. The figures illustrate the findings adequately, and the conclusions are well-founded, especially given the described caveats and challenges associated with using a state-of-the-art ice sheet model. As a result, I recommend acceptance for publication to GMD, after minor edits.

We thank the reviewer for their positive feedback and for their constructive comments on the manuscript.

Below, I outline a few comments/suggestions for the manuscript:

Line 77: Please rephrase. Evaluate is used twice and the wording is confusing as I am not sure what "the ability of a model to evaluate multiple different states" means.

Thank you for spotting this typographical error. The second instance of the word "evaluate" should be "simulate", and this phrase should therefore read "the ability of a model to simulate multiple different states". As well as correcting this error, we have expanded upon this point by adding the following text at lines 78-79 of the revised manuscript:

“A model of the Antarctic Ice Sheet, for example, might be evaluated for its ability to simulate past warm or cold intervals (e.g. DeConto and Pollard, 2016).”

Line 197: Please define F here. In the PISM User’s Manual I think they define it to mean a function, but please specify here.

F is the function that describes the flow law. We have added this information at line 198 of the revised manuscript.

Line 255: Can the variable distribution referred to at this point be described as uniform (as opposed to normal for example)? Please specify in the text.

We consider that the parameters studied in the manuscript are insufficiently understood to enable any robust assumptions to be made in regard to the distribution of prior probabilities. As such, we consider that the simplest possible assumption (i.e. uniform) is the most appropriate. To clarify this, we have added the following text at lines 262-263 of the revised manuscript:

“In the absence of any information on the distribution of prior probabilities, a uniform probability distribution is used for each parameter.”

Line 256: I think this means that for each of the 100 ensemble members, all 10 variables are perturbed independently, and then one ensemble member is run. Please rephrase this part of the procedure so that it is clear to the reader.

The reviewer is correct. The text at lines 259-260 of the revised manuscript now reads:

“A 100-member perturbed-physics ensemble is constructed. Each of the ten parameters is perturbed independently, using a Latin hypercube approach (e.g. Helton and Davis, 2003) to sample the ranges of possible values.”

Line 256: Based on comments later in the manuscript, you are aware that 100 ensemble members for variation of 10 variables statistically not enough to fully characterize the parameter space. Of course, a lot of work went into just running the 100 members presented here, so I am not suggesting that you run more. However, while this caveat is not completely ignored by the authors in the manuscript, I would like to see it specifically addressed either here or in the discussion (i.e. where the need for more systematic, larger ensembles is discussed).

As in the response to Referee #1, we acknowledge that, when applying the technique described in the manuscript, there is a trade-off between expense (ensemble size) and precision (reduction in parameter uncertainty). We also acknowledge that an ensemble size of 100 is not sufficient for us to

fully characterise the properties of a ten-dimensional parameter space. We have therefore added the following text at lines 389-394 of the revised manuscript:

"The size of the ensemble presented in this study (100) is relatively small, particularly given the large number of parameters being optimised (10). Chang et al. (2014) show that a 100-member Latin Hypercube ensemble cannot adequately resolve the interactions between parameters in an ice sheet model, even when being used to study a five-dimensional parameter space. Ideally, our technique would therefore use a larger ensemble size or would be used to target a smaller number of model parameters. While the former would increase the computational cost, either of these modifications should allow for greater refinement of parameter ranges."

Line 343: Awkward use of "from" twice. Please rephrase.

Thank you for spotting this typographical error. We have removed the first instance of the word "from" in the revised manuscript.

Line 359: Please specify what smaller means in this context. I think you refer to the area extend of the ice sheet, but it is not clear.

We are referring to the volume of the ice sheet, and have replaced "a smaller ice sheet" with "an ice sheet with a smaller volume" at line 373 of the revised manuscript.

Line 370: I am curious as to why you did not also choose to use surface velocities as an observational constraint. I realize that since the ice sheet is in balance, the thickness profile encompasses velocities in a way, however, I would be interested to see results showing that the use of thickness and mask is adequate on its own. The choice of which of these constraints is chosen to be uses actually could, in itself, contribute to uncertainty. I think is worth adding some sentences addressing this in this discussion section.

We agree. In response to comments by both reviewers, we have replaced this sentence with the following text at lines 382-386 of the revised manuscript:

"We have shown that the optimal ranges for each parameter can be dependent on other variables. While we have been able to substantially reduce the volume of plausible parameter space, limitations on our understanding of the underlying physical system ensure that the plausible ranges remain large for some parameters. Using additional observational and palaeoclimate datasets to evaluate the model, such as the surface ice velocity and vertical profiles of temperature and age from ice cores, might allow us to constrain the parameter ranges further."

Lines 375-377: Agreed. This is especially the case because model response will become even more convolved when model forcing is not held constant (i.e., transient forcing through time). This is a potent point. Could you expand upon it with 1-2 sentences that give a concrete example of why this is the case (so that a reader can clearly discern how you make the jump from your results to this claim)?

We agree and we will expand upon the important points raised in this paragraph. We do not provide future simulations in the current manuscript. Therefore, to provide a concrete example, we have added the following text at lines 398-402 of the revised manuscript:

"The importance of these points is demonstrated by DeConto and Pollard (2016) and Edwards et al. (2019), who find that parameter uncertainty and ensemble design influence the probability distributions for projections of future sea level rise. In particular, Edwards et al. (2019) emulate an ice sheet model and find that the probability distributions are skewed towards lower values; failure to take this into account might lead to over-estimates of the most likely rate of sea level rise during the coming centuries."

Lines 378-380: This is a very strong statement. While this claim may be accurate, I do not see how it is shown in this manuscript. That is, the results shown in this manuscript do not directly illustrate how the approach shown here explicitly quantifies parameter uncertainty and that this derived uncertainty is propagated into projections. The experiments only derive uncertainty in model steady-state spin-up due to chosen, appropriate spreads chosen for key model variables. I think this is what you are trying to say in this paragraph, but I think it can be said in a more direct, clear way to the audience that your approach can be expanded to projections using ice sheet models, but will require expanded ensembles and further investigation into the model system. Expanding on why this is true with a couple of sentences would strengthen your final discussion point.

We agree. We have therefore replaced this paragraph with the following text at lines 403-408 of the revised manuscript:

"Finally, we note that, whereas the approach developed in this study allows for the rigorous quantification of uncertainty in model parameters, and therefore the quantification of uncertainty in model projections, the technique presented here does not allow these uncertainty ranges to be interpreted in probabilistic terms. Extending our approach to generate future projections with associated probability distributions would require larger ensembles and further understanding of the uncertainties inherent in the physical system. This would include uncertainties in our physical understanding of that system, in the numerical representation of that physical understanding within the model, and in the boundary conditions applied to the model."

Lines 399-400: With respect to my last comment, this statement does a good job of summarizing what I think you are trying to say at the end of the discussion. Leading the reader to how and why you make this conclusion is exactly what I am hoping you can accomplish in the last paragraph of your discussion.

Thank you. As per our response to the previous comment, we have followed this suggestion when revising the final paragraph of the discussion.

Figures 1 and 2: In addition to these, plots of the mean and maybe standard deviation (or uncertainty) spatially, derived from the surviving ensemble members could also be a helpful way to summarize your results.

We agree that this would be valuable information. The standard deviation is the same for both Figures 1 and 2, as the only difference between the two figures is that the Bedmap2 topography has been subtracted from the simulated topography in Figure 2. We have therefore removed the bottom-right panel from Figure 2 (which simply replicates the same panel in Figure 1), and replaced it with two new panels showing the ensemble mean and ensemble standard deviation of the model error.

Supplemental Material: I am not sure myself, but is it appropriate to include the PISM user's guide as supplemental material? It already has a copyright and its own set of authors.

We consider that it is helpful to the reader for us to provide the version of the manual that documents the precise version of the model described in the manuscript.

The PISM manual is distributed as part of the model source code and is updated with each new release. It must also be compiled. As such, the only way that the reader could otherwise obtain the correct version of the manual would be for them to download the precise version of the model source code and to compile the documentation themselves. Under the circumstances, we think it is helpful for us to provide them with a PDF.

In regard to the copyright issue, the PISM source code (including the manual) is distributed under the terms of the GNU General Public License. This license grants legal permission to copy, distribute and/or modify the source code so long as the copies are distributed under the same terms. Geoscientific Model Development is distributed under a Creative Commons Attribution 4.0 License, which also permits copying, distribution and modification of content. Including the PISM manual as Supplementary Material is therefore permitted.