Response to D.J. McNeall

We are grateful to Dr. D.J. McNeall for the valuable comments and suggestions. We have made a concerted effort to address the concerns raised in the report. Please find below our detailed responses to the comments.

Major Comments:

1. A slightly more comprehensive summary of the emulator technique in the main text would be welcome:
We have added the following description for a more comprehensive summary of our emulation technique:

Because the principal components are uncorrelated, we can emulate each principal component separately. Our emulator consists of all these independent Gaussian processes. Although our emulator operates in the principal component space, we can reconstruct the ice thickness profile that corresponds to the emulated principal components (see the Supporting Information for details). Note that our likelihood formulation automatically penalizes the components with lower explained variation.

2. I would suggest acceptance of the paper, conditional on a test set of ensemble members of a sizeable fraction of the ensemble - perhaps at least a third, chosen at random. If computational effort is not a consideration, I would recommend a leave-one-out or leave-n-out test across the entire ensemble.

We agree that cross-validation using a sizeable fraction of the ensemble provides a better test for the performance of our approach. We have conducted leave-one-out cross-validation across the entire ensemble members and added discussion on the results in the Supporting Information as follows:

To investigate (i) whether the perfect model experiment results shown in the main text are sensitive to the values of input parameters assumed as the synthetic truth, and (ii) whether the prediction intervals for ice volume projections generated from our method have the right coverage, we have conducted leave-one-out cross-validation across all input parameter settings in the ensemble. In other words, we have repeated the same perfect model experiment described in the previous sections for all 100 possible different synthetic truths. We summarize the cross-validation results for emulation and calibrated projections in Figure S4 and Figure S5, respectively … The plots in Figure S5 show that the prediction intervals generated from our approach achieve the nominal coverage level only when the modern ice volume generated by the synthetic truth is close enough to the observed volume (i.e. within 10% of the observed value). The width of the prediction interval also varies considerably across the different assumed truths. Therefore, consistent with the findings in McNeal et al., 2013, selection of the assumed truth affects the calibration performance.

Figure S4 and S5 are included at the end of this letter as well. Since our design points in the parameter space are quite sparse, leave-one-out cross-validation is rigorous enough to test the

performance of our calibration approach.

Minor Comments:

1. There appears little justification of the use of 10 PCs in the emulator. What procedure was used to choose the 10 PCs, and why was 10 chosen as "good enough"?

We choose the 10 PCs to have explained variation of 90%. Although it was not originally explained in the manuscript, we have already confirmed that our results are robust against choice of number of principal components, by looking at the results based on more than 10 PCs. We have added the following short note on this at the end of Section 3 in the Supporting Information:

Using 10 principal components captures more than 90% of the variation in the model output, and we have confirmed that using more than 10 principal components does not significantly improve the emulation accuracy in cross-validation.

2. The "future work" section describes the aims of the authors to extend the methodology to the full two-dimensional thickness map of the ice sheet, rather than the one-dimensional thickness profile. Given the apparent availability of model data (as compared to observations), why did this work use only thickness profiles?

We apologize for not explicitly explaining the challenges related to emulation and calibration using 2-dimensional thickness maps. The main challenge involves modeling high-dimensional spatial data containing many zeros. To our knowledge this is an open problem in both computer model calibration and spatial modeling. One possible solution is using truncated Gaussian processes, which however requires dealing with a large number of latent variables, as many as the number of zeroes in the model output. For the SICPOLIS ensemble that we use here, we need to deal with about 600,000 latent variables and to our knowledge no current approaches can handle this properly. We have updated our description of computational challenges for calibration using the full two-dimensional ice thickness grid as follows:

Direct emulation of the full two-dimensional ice thickness grid is prohibitively expensive, due to (i) the cost of performing operations on large covariance matrices (see the Supporting Information and Chang et al., 2013, for details) and (ii) the need to model spatial processes that contain many zeros, which poses non-trivial computational and inferential challenges.

3-1. Clearly, leaving a discrepancy term out when discrepancy was added to the synthetic data, will result in a mis-specified probability distribution for the input parameters (and subsequent predictions of the ice sheet). The authors have missed a trick here – It would be very useful to show, comprehensively across the ensemble, how much error a mis-specified discrepancy term adds to predictions.

We have diagnosed the effect of including the discrepancy term on ice volume change projections and included the following discussion on the results in Section 6 of the Supporting Information:

Another important observation is that including the discrepancy term reduces the overconfidence that occurs when the synthetic truths are outside of the 90-110% range. The prediction intervals are overconfident when the synthetic truth is outside of this range because the coverage is consistently less than 95%. Including the discrepancy term reduces this issue in some degree since it makes the actual coverage closer to the nominal coverage when the synthetic truth yields the modern ice volume that is within at most 70% of the observed volume. However, this correction effect is not sufficient to make the prediction intervals achieve the nominal coverage.

3-2 It might also be worth demonstrating how much uncertainty a well-specified-but-uncertain discrepancy term adds to the predictions, and to the identifiability of the input parameters.

Our discrepancy term, constructed based on kernel convolution, is a well-specified-but-uncertain discrepancy term that is designed to capture a large scale model-observation discrepancy. Note that using an overly flexible discrepancy process leads to a serious identifability issues between the discrepancy process and the input parameters, and our discrepancy term using kernel basis with pre-specified range and smoothness parameters is one way to mitigate this issue while maintaining enough flexibility of the discrepancy process (Chang et al. 2014). We have added the following short note on this in the Supporting Information:

Fixing the range parameter not only reduces the computational cost for likelihood computation but also improves the identifiability between the input parameters and the discrepancy process.

4. Figure 1. could show the entire ensemble (perhaps greyed out), and highlight the subset of ensemble members.

We have incorporated your suggestions in Figure 1. Please see the revised figure below.

5. The accuracy of the emulator as demonstrated in figure 2. is impressive. Again, it would be useful to show how this varies across the entire ensemble. There are ideas for doing this using similar PC emulation techniques for one dimensional data in Challenor et al (2010), and McNeall (2008).

We have added Figure S3 below to the Supporting Information that shows the leave-one-out cross-validation results across the entire ensemble. We have also included the following discussion in Section 6 of the Supporting Information.

The results in Fig. S3 show that our emulator can predict the model output reasonably well across all input parameter settings. The predicted ice volume thickness profiles are concentrated around the diagonal line that connects the lower left and the upper right corners of the plot, and hence the emulator can predict the model output reasonably well for most input parameter settings. Note that leave-one-out cross-validation is already rigorous enough in our case due to the sparsity of the design points (100 points in 5-dimensional space) for the input parameters in our ensemble. We have also conducted leave-10-out cross-validation for emulation and the results are essentially the same (not shown).

6. If the authors are to extend the testing of the probabilistic methodology across the ensemble, a graphical representation of the strength of interactions between parameters - summarised across the entire ensemble- would be most welcome. The pairs plots as used show this nicely for a single ensemble member, but are not appropriate for large ensembles.

We have added the discussion below and Fig. S5 in the Supporting Information that summarize the interactions between the input parameters across the entire ensemble using the distributions of the rank (Spearman) correlations.

The cross-validation results allow us to examine the interaction between input parameters across all possible choices of the synthetic truth. We have computed the rank correlations between the input parameters across all 100 ensemble members and summarized their distributions in Figure S6. From the shapes of the densities we can identify five pairs of parameters that tend to be more negatively correlated: (i) the flow factor and the snow PDD factor, (ii) the flow factor and the geothermal heat flux, (iii) the basal sliding factor and the ice PDD factor, (iv) the geothermal heat flux and the ice PDD factor, and (v) the ice PDD factor and the snow PDD factor.

Figure S6 is also included at the end of this letter.

References

Chang, W., Haran, M., Olson, R., and Keller, K.: Fast dimension-reduced climate model calibration, Ann. Appl. Stat., accepted, 2014.

McNeall, D. J., Challenor, P. G., Gattiker, J. R., and Stone, E. J.: The potential of an observational data set for calibration of a computationally expensive computer model, Geosci. Model Dev. Discuss., 6, 2369–2401, doi:10.5194/gmdd-6-2369-2013, 2013.

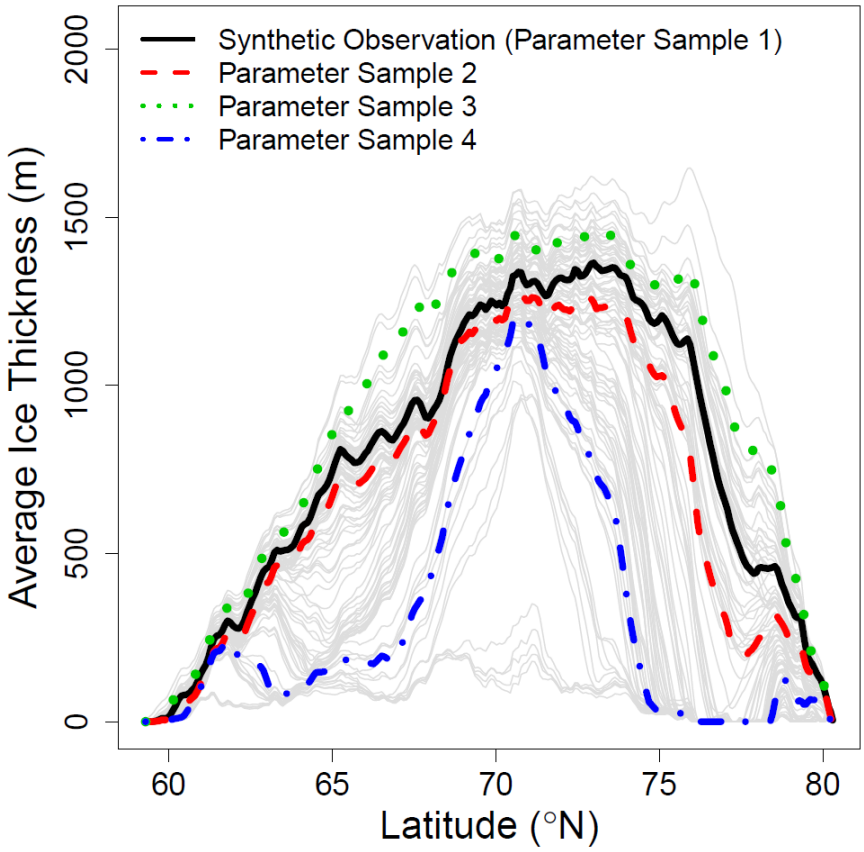

Fig. 1. Profiles of zonal mean ice thicknesses from four different evaluations of the ice sheet model SICOPOLIS (Greve, 1997; Greve et al., 2011). The solid black curve represents model run #67 from Applegate et al. (2012), which we take to be the synthetic truth for our perfect model experiments. The other curves represent examples of model runs used to construct the emulator: one run produces a zonal mean ice thickness curve similar to the synthetic observations (dashed red curve), another is generally too thick (dotted green curve), and a third is generally too thin (dot-dashed blue curve). As expected, our probability model assigns a greater posterior probability to the model run represented by the red curve than to the model runs represented by the blue and green curves. All the other model runs that are not highlighted above are represented as grey curves.

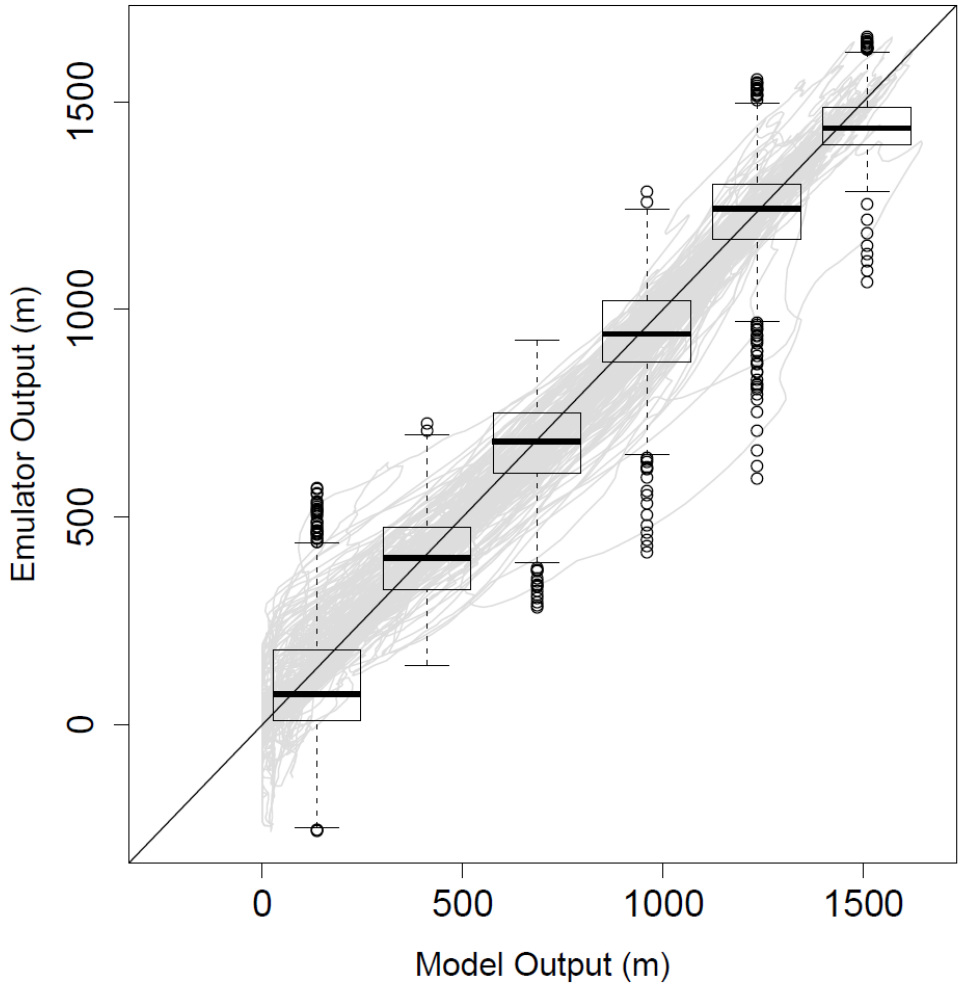

Fig. S4. Leave-one-out cross-validation results for the emulation performance. Each grey curve shows the comparison of zonal mean ice thickness transects from the model output and that from the emulator output for each parameter setting. Each boxplot shows the distribution of emulator output for each of the evenly spaced bins that span the range of true model output. In spite of the fact that our design points for parameter settings are quite sparse (100 runs in 5-dimensional space) most of the curves are concentrated around 1:1 line connecting the lower left and upper right corners of the plot, indicating that our emulator can reconstruct the original model output reasonably well across the input parameter settings.

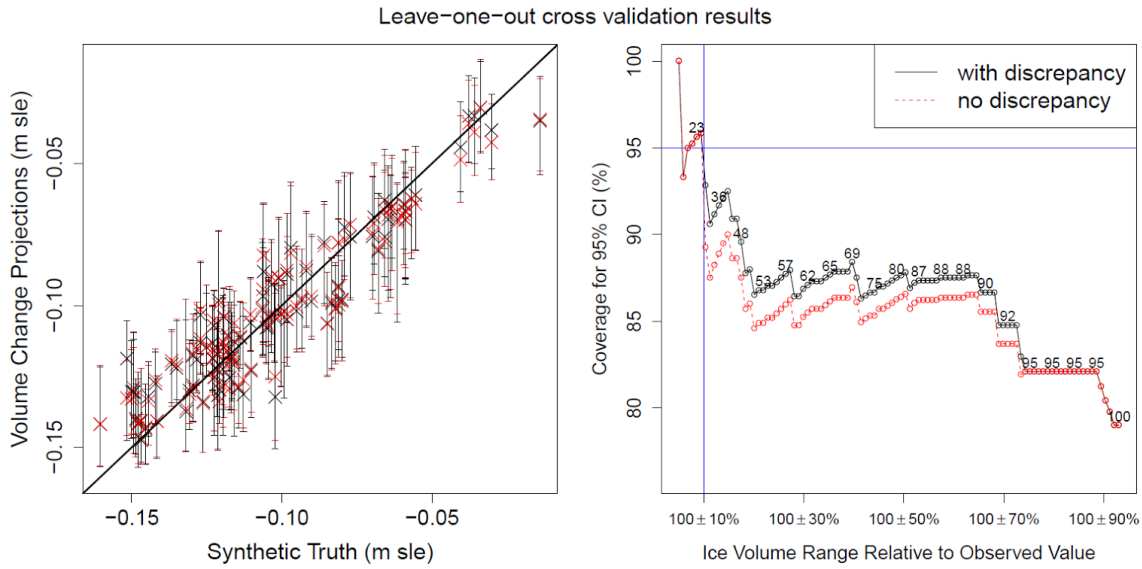

Fig. S5. Leave-one-out cross-validation results for ice volume change projections across all 100 input parameter settings as the synthetic truth. The left penal shows 95% prediction intervals for ice volume change projections across all 100 perfect model experiments conducted for cross-validation. If the interval covers the 1:1 line connecting the lower left and upper right corners of the plot, the 95% prediction interval includes the ice volume projection given by the synthetic truth. The right penal shows the coverage of those prediction intervals as a function of allowed range for the ice volume in 2005 AD relative to the observed ice volume. “”The numbers above the solid black line show how many synthetic truths fall into the given ice volume range. The plot shows that (i) the credible intervals achieve the nominal coverage level only for the “realistic” synthetic truths with modern ice volume within 10% of the observed ice volume, and (ii) the discrepancy term reduces overconfidence issues for the synthetic truths that are not within the 10% range.

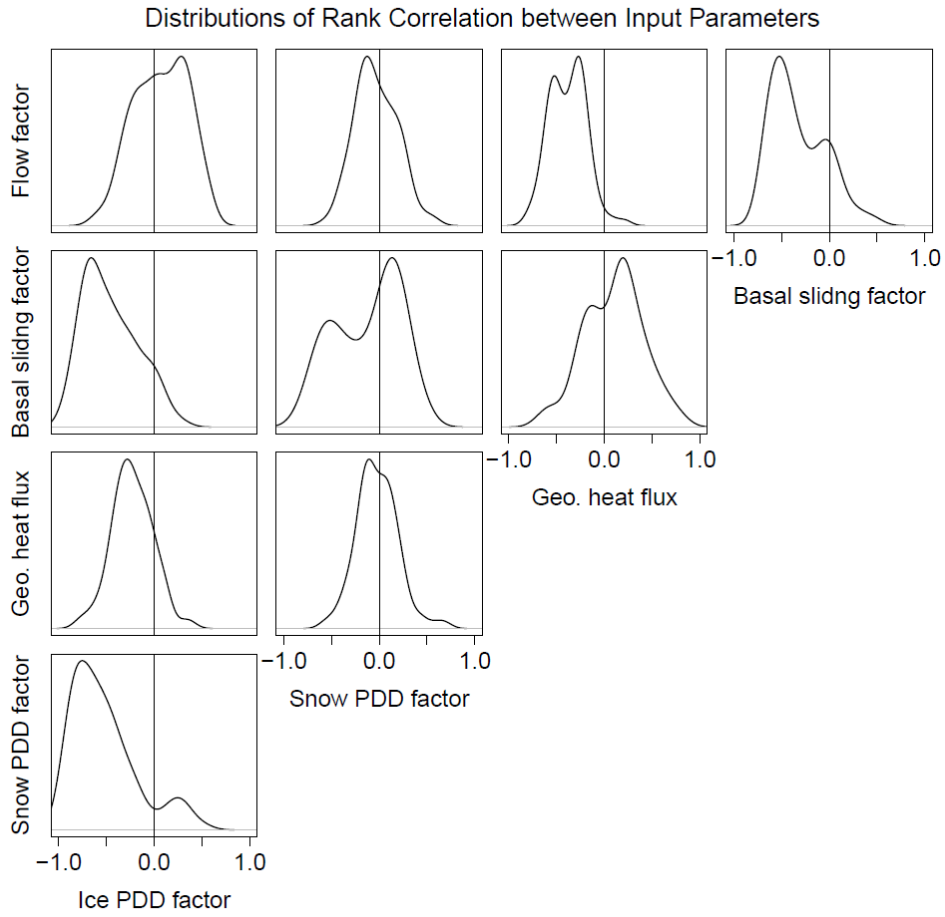

Fig. S6. Summary of interactions between input parameters computed from leave-one-out cross-validation. Each panel shows the distribution of the rank correlation between two input parameters across all synthetic truths in our leave-one-out cross-validation. Five pairs of input parameters, (i) the flow factor and the snow PDD factor, (ii) the flow factor and the geothermal heat flux, (iii) the basal sliding factor and the ice PDD factor, (iv) the geothermal heat flux and the ice PDD, and (v) the ice PDD factor and the snow PDD factor are tend to be more negatively correlated comparing to the other pairs of parameters.

Response to T.L. Edwards

We are grateful to Dr. T.L. Edwards for the valuable comments and suggestions. We have made a concerted effort to address the concerns raised in the report. Please find below our detailed responses to the comments.

Major Comments:

1) "(a) some imprecise and unsupported statements, and missing discussion points, which I would like to see addressed; (b) unnecessarily poor justification of using synthetic rather than real observations"

Thank you for the very detailed suggestions for addressing these issues. We have incorporated most of the suggestions in our manuscript (see the Specific Comments section below for more details).

2) "The paper's conclusions would also be more robust and substantial if the leave-one-out testing were repeated for all ensemble members, as is (reasonably) standard, or at least for more than three (10 or 20 might seem a reasonable minimum number to me)."

We now have included leave-one-out cross-validation results across the entire model runs and added the following discussion on the results in the Supporting Information:

To investigate (i) whether the perfect model experiment results shown in the main text are sensitive to the values of input parameters assumed as the synthetic truth, and (ii) whether the prediction intervals for ice volume projections generated from our method have the right coverage, we have conducted leave-one-out cross-validation across all input parameter settings in the ensemble. In other words, we have repeated the same perfect model experiment described in the previous sections for all 100 possible different synthetic truths. We summarize the cross-validation results for emulation and calibrated projections in Figure S4 and Figure S5, respectively … The plots in Figure S5 show that the prediction intervals generated from our approach achieve the nominal coverage level only when the modern ice volume generated by the synthetic truth is close enough to the observed volume (i.e. within 10% of the observed value). The width of the prediction interval also varies considerably across the different assumed truths. Therefore, consistent with the findings in McNeal et al., 2013, selection of the assumed truth affects the calibration performance.

Figure S4 and S5 are also included at the end of this letter.

3) "Another weakness is awareness of the relevant literature, including results from IPCC AR5 and the ice2sea project (See http://www.ice2sea.eu/programme/published-papers for references) and the UK UQ community."

We have included citations for AR5, ice2sea project, and the UK UQ community by following your suggestions (see below for details).

4) "Finally, it may be because I read the SI a while after the main paper but I found much of Sections 4 and 5 in the SI difficult to understand. I hope the brain dump of points I found confusing and suggested improvements given at the end of this review are useful."

We appreciate the useful suggestions for improving the Supporting Information. We have addressed most of the raised issues (see below for details).

Specific Comments:

1. Scientific points

1) 1906/12 I would argue you can make probabilistic projections without calibration (i.e. present prior density, if no observations available).
2) 1906/13 You don't use observational data but synthetic observations...

We have slightly changed the sentence as follows:

This method is an important step toward calibrated probabilistic projections of ice sheet contributions to sea level rise, in that it uses data-model fusion to learn about parameter values.

3) 1907/15 "primarily" - No, see IPCC (2013) sea level chapter: projections are primarily based on regional climate models, ice sheet models and glacier models

We updated the sentence to read, "Present estimates of future sea level rise are often derived from semi-empirical extrapolations of tide gauge data..." Many would argue that ice sheet models are still missing key processes and/or require calibration against data, and that semi-empirical models may do a better job of capturing the relevant effects until the ice sheet models catch up.

4) 1907/25 "spatial distribution" - and flow

We have added the word "flow" in the sentence.

5) 1913/16 "identify" -> present / describe / test (not sufficiently novel to warrant identify)

We have substituted "identify" with "present" here.

6) 1915/24 This is not a good justification for not using observed geometry, and doesn't match the scope of the paper. For me the justification is that you want to test the ability to retrieve the original parameter values - which you do - rather than make calibrated projections of the future. Whether the model gets the observed geometry right or not is not relevant to the stated aims of the paper. Your statement that model limitations will "cause problems" is both ill-defined (do you mean discrepancies too large to have any effect on the posterior? or more difficult to construct a statistical model of the discrepancy?) and not supported (by showing the ensemble thicknesses against the observed to demonstrate that they disagree substantially, or performing the calibration with the observations to show it has little effect). See also comment below on p1921.

Thank you for pointing out this point. To make our justification simple and clear, we have clarified the reason for not using observed geometry as follows:

Consistent with McNeall et al. (2013), we match the emulator estimates to assumed-true model output instead of observed ice thickness values (Bamber et al., 2001, 2013) because a perfect

model experiment is more suitable to achieve our main objectives, studying and demonstrating the performance of our probabilistic calibration method.

7) 1917/2-8 These lines do not describe passing of check #1: lines 9-13 do. (Suggest new paragraph at line 13 and reordering previous lines to reflect this.)

We have relocated a few sentences as below to incorporate your suggestions:

Aggregating the ice thicknesses to their zonal means allows easy visual comparison of different emulator-estimated ice thickness vectors to the assumed-true model realization (black curve, Fig. 1). The emulator, as trained on 99 of the model realizations from the Applegate et al. (2012) ensemble, successfully recovers the ice thicknesses from the left-out model realization (Fig. 2) when given the parameter combination for that left-out model realization as input. Differences between the assumed-true and emulated zonally-averaged ice thickness vectors are minor. Thus, our methods pass check #1, above.
Similarly, the conditional posterior density functions (Fig. 3) have maxima near the assumed-true parameter values. Parameter combinations yielding zonally-averaged ice thickness curves that lie close to the assumed-true model realization (e.g., the red curve in Fig. 1) are more likely (more probable based on the posterior distribution) than those with curves that lie farther from the assumed-true values (blue and green curves in Fig. 1). We do not expect that the modes of the marginal posterior density functions (Fig. 4b) will fall exactly at the assumed-true parameter values, because summing over one or more dimensions often moves the marginal mode away from the maximum of the multidimensional probability density function. In any case, the maximum posterior probability is close to the assumed-true parameter combination. Thus, our methods pass check #2, above. Some of the two-dimensional marginal probability density functions (Fig. 4b) show multiple modes and bands of high probability extending across the two-dimensional fields; we discuss the significance of these features below.

8) 1917/27 There *is* clustering around the best estimate for ice PDD: this should be mentioned.

We have added this to the text as follows:

As in Applegate et al. (2012), the "successful" design points show no clustering around the assumed-true parameter values, except around the true value of the ice PDD factor.

9) 1918/6 and Fig, 5 How do you obtain 95% intervals for the Applegate windowing method? By retaining those within 9.5%?

The 95% intervals were computed by simply looking at the 2.5th and 97.5th percentiles. I have added the following very short description for this.

The 95% probable interval produced by our methods is much smaller than that estimated by computing the 2.5th and the 97.5th percentiles of the volume change values selected by the 10% volume filter used in Applegate et al. (2012).

10) 1918/7 It doesn't only reflect the utility of spatial information: it also reflects the choice of window size (it could have been 5%), and your Bayesian statistical modeling choices.

We have added the following clarification for this point.

This reflects the utility of spatial information and our probabilistic calibration approach in reducing projection uncertainties as compared to the windowing approach in Applegate et al. (2012).

11) 1920/3 "incorrect" -> "non-optimal" or similar; also, some modes look like they have similar density, so worth mentioning there may not be a unique best estimate theta*.

We have substitute “incorrect” by “non-optimal”. We have also mentioned that multiple modes imply there are possibly many “good” estimates for □*.

If the surface has multiple "peaks" (i.e. regions of parameter space that are more plausible, given observations, than their surroundings), gradient descent methods can converge to a point which produces a better match to the data than any adjacent point, but is nevertheless far from the "best" parameter combination.

12) 1920/10 "true" -> "best", as you use in the SI, because you are not talking about retrieving the synthetic observation parameters here.

We have changed the wording from “true” to “best”.

13) 1920/12 There isn’t always wide variation: see e.g. ice2sea’s Shannon et al. (2013), Edwards et al. (2014b)...

We updated this sentence to read, "This problem may partly explain the wide variation in projections of sea level rise from the ice sheets, as made with state-of-the-art ice sheet models (Bindschadler et al., 2013; cf. Shannon et al., 2013; Edwards et al., 2014b)..."

14) 1920/14 And potentially also differences in spin-up method, if by "reproduced the modern ice sheet equally well" you mean only the topography (rather than dH/dt, velocities, ice temperature etc).

The reviewer makes a good point; we updated the sentence to read, "... even if the models had similar structures and reproduced the modern ice sheet topography and ice thicknesses equally well..."

15) 1920/25 "generally too thick" - show ensemble and obs in Figure? (as per comment above on p1915)

We appreciate the suggestion, but we believe that pointing readers to the Fig. 7 in Applegate et al. is enough for making this point.

16) 1920/27 "difficulties" - see ice2sea results for improvements in spin-up methods (holding to modern geometry, relaxation, using SMB corrections through the model simulations to account for remaining errors) - as per comment on p1908 in Section 2 below.

We updated this sentence to read, "... other ice sheet modeling experiments have similar difficulties in reproducing the modern ice sheet (e.g., Stone et al., 2010; Greve et al., 2011; Nowicki et al., 2013, their

Fig. 2; cf. Edwards et al., 2014a)."

17) 1921/2 Here is a bit more detail on why you are not using observations. Again I don't agree with this justification - couldn't you test the effects of a large discrepancy term with the synthetic observations? But if you do include this justification, add it earlier (p1915) too.

We have retracted this justification now.

18) Fig. 5 Are your kde bandwidths a bit small or are those bumps due to real physics?

Since our chain is well-mixed and the bandwidths are adequately selected, this is not a bandwidths issue—the bumps are due to the actual properties of the density function.

19) SI/55 "our experiences" - reference? leave-one-out validation? What is "very accurate"?

We have slightly revised this part as follows to provide more details.

However, according to our cross-validation experiments for various models including SICOPOLIS, the emulator based on this assumption usually provides an accurate approximation to the original model (see e.g., Chang et al 2014, Figure 2).

20) SI/170 Does the error rate in the cross-validation indicate some choices could be improved?

Due to the very irregular behavior of the ice volume change surface in the parameter space, the exponential covariance used here seems to provide the best emulator. Further improvement may be possible by, for example, introducing a non-separable covariance structure. However, this involves a significant effort in methodological development and the advantages of this added layer of complexity is unclear; we therefore leave this for future work.

21-1) SI-187 Do you use the old Bamber et al. geometry to pick the ensemble members because this was done in Applegate et al., or could you use the updated (2013) geometry instead?

The older Bamber data set (with updates by the seaRISE project) was used in the Applegate et al. (2012) ensemble, so it would be inconsistent to use the newer geometry here. We anticipate that the effects of incorporating the updated geometry in a new ensemble and model-data intercomparison would be fairly minor.

21-2) Why not do leave-one-out for each ensemble member, not just 3? (N.B. I may not be sympathetic to the argument "it takes 8 hours"... ;))

As we have described above, we have now conducted leave-one-out cross-validation across all possible choices for the synthetic truths and included the results in the Supporting Information.

22) SI-187 "essentially the same" - no, the effect of the calibration is much stronger! (maximum posterior density more than 2x greater) - why is this?

While conducting leave-one-out cross-validation across the all ensemble members, we have found that the performance of our calibration approach is sensitive to the assumed truth. We have deleted the

statement and added the following discussion on this.

The width of the prediction interval also varies considerably across the different assumed truths. Therefore, consistent with the findings in McNeal et al. (2013), selection of the assumed truth affects the calibration performance.

2. Literature

1) 1906 I think (of course I do...) it is relevant to cite Edwards et al. (2014a) and (2014b) The Cryosphere - probabilistic Greenland projections, calibrated in a Bayesian framework with spatial information from a regional climate model.

We agree that these papers are relevant. We have cited the two papers now.

2) 1906 IPCC (2013) replaces Meehl et al. (2007).

We have added a citation for IPCC (2013).

3) 1907 If citing SeaRISE, should cite ice2sea project (had much, if not more, model development) too.

We updated this sentence to read, "... such models have been the focus of intense development effort since the fourth Intergovernmental Panel on Climate Change assessment report (e.g., Bindschadler et al., 2013; Shannon et al., 2013; Edwards et al., 2014a)."

4) 1907 Why cite a palaeo ref for melting vs elevation?

We feel that citing Born and Nisancioglu (2012) is appropriate in this case. Although their study is primarily about the Greenland Ice Sheet during the Eemian, their study investigates the modern climatology of the ice sheet and presents a nice theoretical explanation of the elevation-melting feedback.

5) 1908 Projections from the ice2sea project use the observed geometry in one stage of the spin-up to reduce these errors. For example, in Edwards et al. (2014b) and Shannon et al. (2013) multi-model papers; see also ice2sea Greenland papers led by Goelzer, Gillet-Chaulet, Quiquet.

We inserted an additional paragraph that reads, "The above paragraphs discuss the case in which the ice sheet model is free to evolve to the state that is most consistent with the selected parameter combination, the bedrock topography and the climate (whether steady or varying). In such studies, parameters such as the basal sliding coefficient are held constant over the geographic area of the ice sheet. However, a number of recent studies (e.g., Shannon et al., 2013; Edwards et al., 2014b) have used an alternative approach in which the spatially-distributed basal sliding coefficients and/or surface mass balance fields are tuned so that the ice sheet model matches the observed modern geometry. This approach has several advantages; the simulated modern ice sheet is guaranteed to match the observed modern one, and the estimated basal sliding coefficients vary spatially, as is almost certainly the case for the real ice sheet. However, such studies are silent on interactions between parameters besides the basal sliding coefficient and surface mass balance, as we investigate here."

6) 1908 Cite Little et al. (2013, Nature Climate Change): a probabilistic study of Antarctica that uses a flow line model of the PIG (and extrapolation of observations elsewhere), calibrated with observations.

Please see our answer to 7) below.

7) 1908 As mentioned above, Edwards et al. (2014b) is a probabilistic projection for Greenland that uses a multi-model ice sheet ensemble and parameter perturbations calibrated using regional climate model data.

We have added the following discussion on Little et al. (2013) and Edwards et al. (2014b) as follows:

In a slightly different but relevant context, Little et al. (2013) and Edwards et al. (2014b) use Bayesian model averaging to assign scores to model runs in perturbed-parameter ensembles, but the scores in these methods are essentially based on RMSE for low-dimensional summaries of model output and therefore do not fully account for the spatial information in ice model output.

8) 1909 Update McNeall et al. reference - now accepted.

Thank you for pointing this out. This is updated now.

9) 1909 Edwards et al. (2014b) contains a _100 member perturbed parameter ensemble for one Greenland model, (technically PPEs for multiple models, albeit with N=3 per model for the others: : :).

We have added citation to Edwards et al (2014) accordingly.

10) SI / p3 Cite Kennedy and O'Hagan for emulation; ideally Goldstein / Rougier ref(s) too. If citing specific emulation applications, also include refs to the UK community (e.g. Sexton, Rougier, Williamson, McNeall, Lee) - or else remove Drignei et al. onwards and just cite emulation methods papers.

We have added the suggested citations as follows.

We emulate the ice sheet model output using Gaussian processes (GP), a fast method for probabilistic interpolation between existing model runs (Sacks et al. 1989; Kennedy and O'Hagan 2001; Higdon et al. 2008; Drignei et al. 2008; Rougier 2008; Bhat et al. 2012; Holden et al. 2010; Lee et al. 2011; Olson et al. 2012, 2013; McNeall et al. 2013; Williamson et al. 2013).

11) SI / p4 For theta* and discrepancy, again Goldstein and/or Rougier refs would be appropriate here.

We have added citation to Rougier (2007).

3. Clarity and other suggested improvements

1) It would be useful to have somewhere a clear summary of the main differences/ improvements c.f. McNeall et al.: i.e. a different model, plausible magnitude present and future projections rather than idealised/palaeo-simulations, a smaller ensemble, the differences in parameters and ranges, emulation

of higher dimensional output (latitudinal thickness PCs instead of scalar summaries), a Bayesian probabilistic calibration instead of a history matching approach. And anything else worth mentioning.

We feel that the main differences are already explained in detail in the previous version, except for the fact that the method used in McNeall et al. relies on historical mapping instead of probabilistic calibration. Please see our response to the next point.

2) 1909/9 May be worth saying explicitly that McNeall et al. study is not probabilistic.

We have added the following sentence:

Moreover, their calibration approach is based on "historical mapping" and does not provide probabilistic projections.

3) 1915/9-10 The observational discrepancy part is a little hard to understand: is it possible to illustrate it in a figure?

Please see the newly added plot below that illustrates the points made in 1915/9-10.

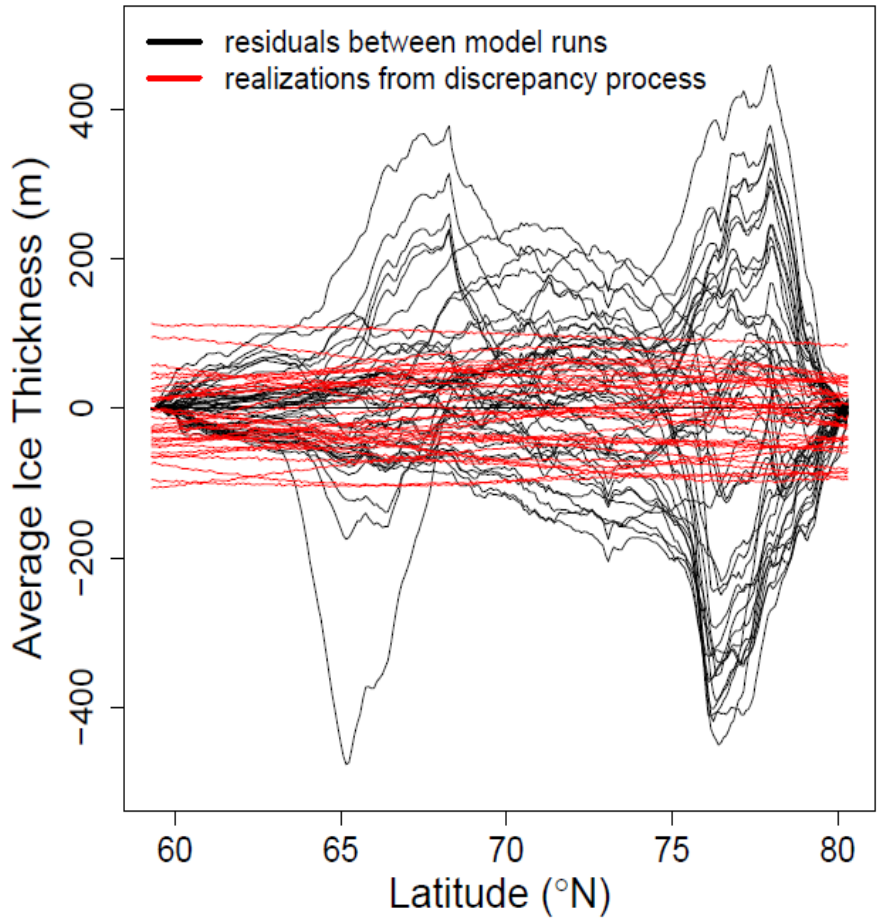

Fig. S1 Comparison between (i) residuals between the synthetic truth used in the main text (model run #67 in Applegate et al.) and other model runs (black solid curves) and (ii) 30 different realizations from the model for the simulated discrepancy (red solid curves). The residuals are computed by subtracting

the synthetic truth from each of the other model runs. For better display, we show only residual curves whose ranges are within (-500,500). It is easy to see that the black curves and red curves are generated from different processes, and therefore those two groups of curves can be separated by statistical inference (hence identifiable). The magnitudes of the simulated discrepancy processes are well within the range covered by the model runs (hence the posterior density of input parameters does not show too large variation).

4) 1918/5 Presumably you use the same windowing approach as Applegate et al. but with the synthetic (not observed) volume - probably a good idea to state this explicitly.

We have clarified this as follows.

For comparison, we also applied the windowing approach used by Applegate et al. (2012) to the model runs and the synthetic observation.

4. Supplementary Information

1) The SI would benefit from more explanation, and an assumption that the reader has not read the authors' previous work... It also needs a more thorough proof-reading.

We have made a concerted effort to address all the issues raised below. We hope that the Supplementary Information is easier to understand comparing to the previous version.

2) I am quite confused by the discrepancy modelling. Are the end of p4 and middle of p6 talking about the same thing? If so, I would move the latter forward and give the notation for the numerical choices (e.g. is phi_d 2100km? is kappa_d 2500m? how about the nugget of 1km?)

We apologize for the confusion here. The discrepancy term in the equation (S1) in p.4 is our model that is designed to capture the discrepancy process using kernel convolution (Higdon et al. 2008, Chang et al. 2014). The discrepancy process in p. 6 is the simulated discrepancy that is superimposed on the synthetic truth to construct the synthetic observation. We have modified the text in p. 6 to clarify this.

To make our experiment more realistic, the simulated discrepancy process is generated from a different model to the discrepancy term that we use in the equation (S1). The covariance function that we use for the Gaussian process model for the simulated discrepancy here is a squared exponential covariance having range of 2100 km, partial sill of 2500 m, and a nugget of 1 m. Our choice for the simulated discrepancy process is based on the following two general assumptions:…

3) Some further suggestions if you would like to make all of the SI as clear as the main manuscript and SI introduction:

- a plain English outline before/after the maths in each paragraph / section;

At the beginning of Section 2, 3, 4 and 5, we now have a paragraph that provides a motivation and/or an outline.

- an illustrative example of one emulator (either an idealised single model output, or the future

projections emulator) before showing the emulation of PCs;
- explanation of statistical terms for the benefit of numerical modellers: partial sills, nuggets, knots, identifiability, block updating, well-mixed chain (and what they mean in terms of assumptions e.g. is one of the nuggets the *emulator* discrepancy?);

Now we have added a paragraph that explains the basis of GP and the meaning of each parameter (see p. 4 in the SI).

However, we feel that explaining other terms may be too tangential and will require a lengthy description. We therefore choose not to elaborate on them.

- explanation of the subscripts y and d for the basis vectors, phi etc;

We have added the following explanation for this:

For the matrices and the statistical parameters used in the following sections, the subscript y indicates that a symbol is used for the emulation model, while the subscript d shows that a symbol is for the discrepancy model.

- a table summarising all parameters;

We have added a table summarizing our notation in the Supporting Information (see Table 1).

- explanation why kappa_y are re-estimated.

We have added the following short explanation.

This allows the emulator process to be re-scaled to better match the observational data.

- explanation how J is chosen / how much of the variation these PCs account for

We have included the discussion below to clarify this.

Using 10 principal components captures more than 90% of the variation in the model output, and we have confirmed that using more than 10 principal components does not significantly improve the emulation accuracy in cross-validation.

- ideally, explain how GP is different to linear regression emulation
- an illustrative figure would be useful

We have added the following comparison between GP and linear regression.

Unlike the linear regression, which requires specification of the mean function along with various statistical assumptions when dealing with highly nonlinear processes such as ice model outputs, the GP model can automatically handle such non-linearities indirectly via a relatively simple covariance function. Only required assumption for the GP model is that the model

output is a smoothly varying curve in the parameter space without too many abrupt changes.

As we explained here, coming up with a sensible emulator based on linear regression involves a major statistical modeling effort, we choose not to include an illustrative figure.

- translations of what the statistical model choices mean in terms of assumptions about the model

We have added the following discussion on using GP emulator.

The GP emulator approach yields a flexible approximation without requiring detailed physical information on the ice sheet model, unlike linear regression-based emulators (cf. piani et al., 2005). By interpolating existing model runs at different parameter settings, a GP emulator provides a reasonable approximation to the original model unless the model output abruptly changes in the input parameter space.

- could mention kriging, which may be familiar to the readers

I have added the following sentence.

Interpolation using GP emulator is essentially kriging in the input parameter space; the interpolator is a random process with a mean that the optimal interpolation between ice sheet model runs in terms of the expected mean squared error and a variance that quantifies the uncertainty of the interpolation.

- remove some unnecessary jargon, e.g. dispersed posterior density -> "has little effect" or similar

We have substituted “dispersed posterior density” with “likelihood then has little effect on the posterior distribution”.

- many mentions of perfect model experiment as if it was one part of the paper, but I think it is all of it? Also I find synthetic observations more clear than perfect model (I use perfect model for studies that do not include a discrepancy variance: : :)

We use the term perfect model experiment throughout the SI to avoid any misunderstanding about the results presented here. In other words, we would like to make sure that readers understand that no observations are used in making ice volume change projections even when they are looking up only a part of the SI.

- you talk about joint and marginal estimation - do you use both? (e.g. for the different 1D and 2D figures?)

We apologize for any confusion caused by our description, but we would also like to clarify that estimating the joint density and the marginal density are not separate procedures. The joint density obtained via MCMC automatically also gives us the marginal densities of the individual parameters -- we need only look at the samples corresponding to a single parameter at a time to obtain the corresponding marginal density estimates. Hence, the approach is to construct MCMC for the joint density, unless the joint density if of no interest and there is an analytical form of the marginal density of an individual parameter that avoids sampling the other parameters.

- explain definition of "error rate" in cross-validation (p8)

To validate the emulator constructed here, we have conducted leave-5%-out cross-validation. The mean error rate, computed by dividing the RMS by the overall mean, is around 16%.

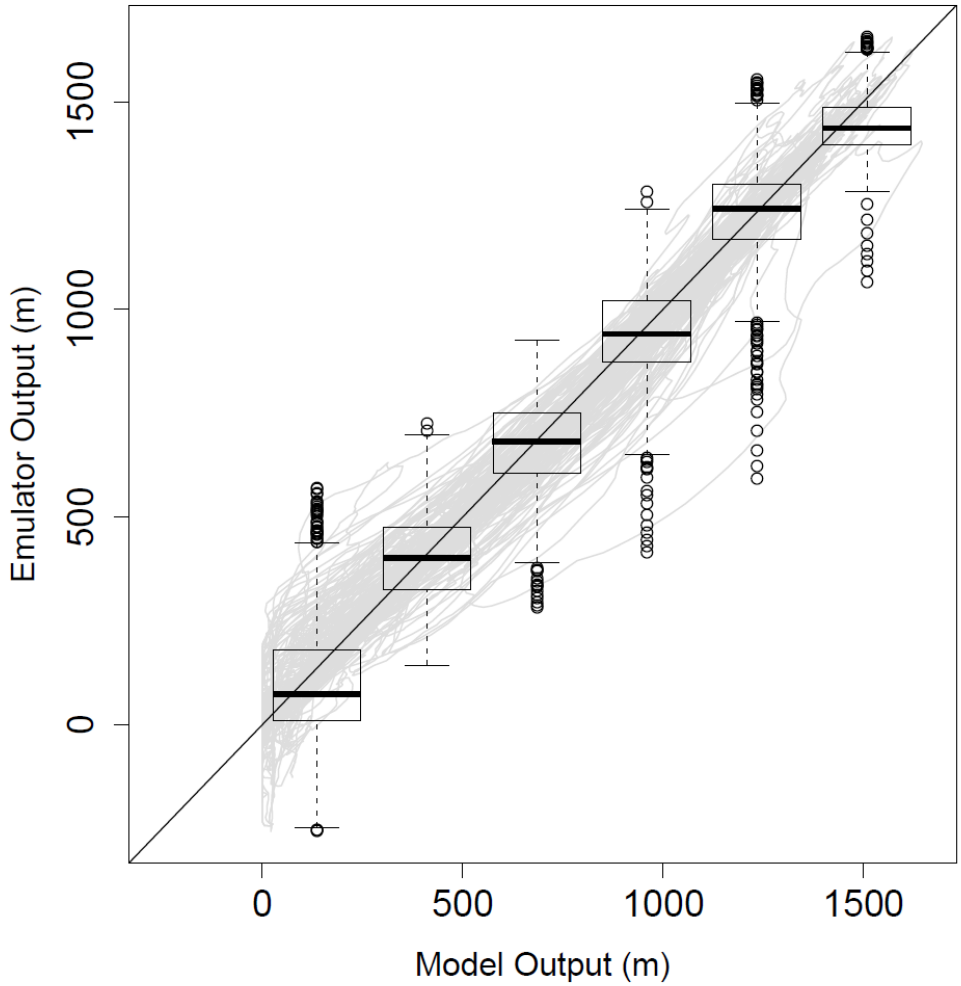

Fig. S4. Leave-one-out cross-validation results for the emulation performance. Each grey curve shows the comparison of zonal mean ice thickness transects from the model output and that from the emulator output for each parameter setting. Each boxplot shows the distribution of emulator output for each of the evenly spaced bins that span the range of true model output. In spite of the fact that our design points for parameter settings are quite sparse (100 runs in 5-dimensional space) most of the curves are concentrated around 1:1 line connecting the lower left and upper right corners of the plot, indicating that our emulator can reconstruct the original model output reasonably well across the input parameter settings.

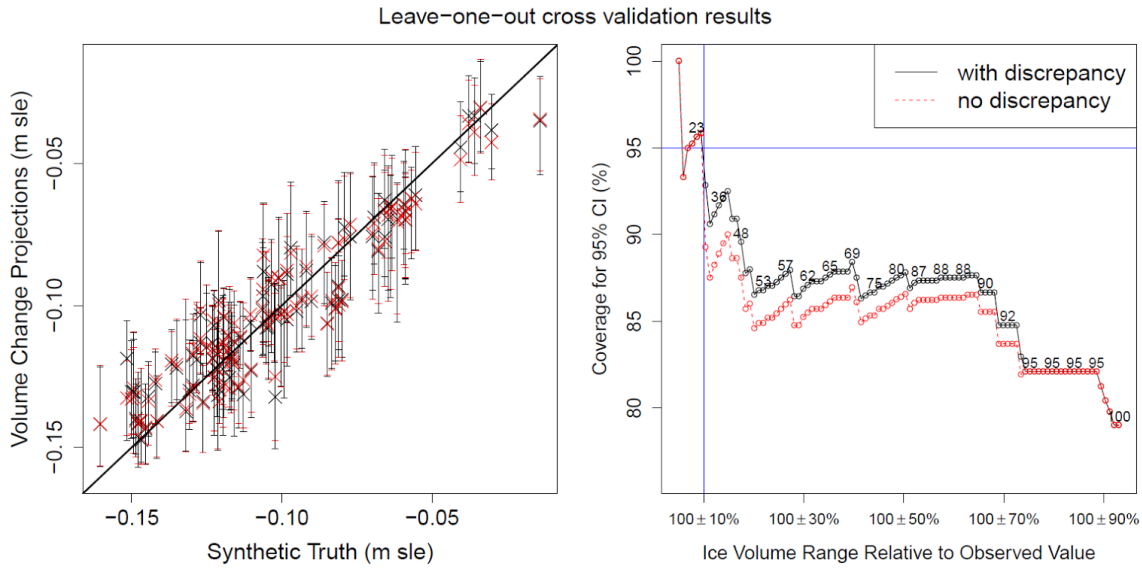

Fig. S5. Leave-one-out cross-validation results for ice volume change projections across all 100 input parameter settings as the synthetic truth. The left penal shows 95% prediction intervals for ice volume change projections across all 100 perfect model experiments conducted for cross-validation. If the interval covers the 1:1 line connecting the lower left and upper right corners of the plot, the 95% prediction interval includes the ice volume projection given by the synthetic truth. The right penal shows the coverage of those prediction intervals as a function of allowed range for the ice volume in 2005 AD relative to the observed ice volume. As going from left to right, the synthetic truths used in computing the coverages include more "unrealistic" ones in terms of modern ice volume. The numbers above the solid black line show how many synthetic truths fall into the given ice volume range. The plot shows that (i) the credible intervals achieve the nominal coverage level only for the "realistic" synthetic truths with modern ice volume within 10% of the observed ice volume, and (ii) the discrepancy term reduces overconfidence issues for the synthetic truths that are not within the 10% range.