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.

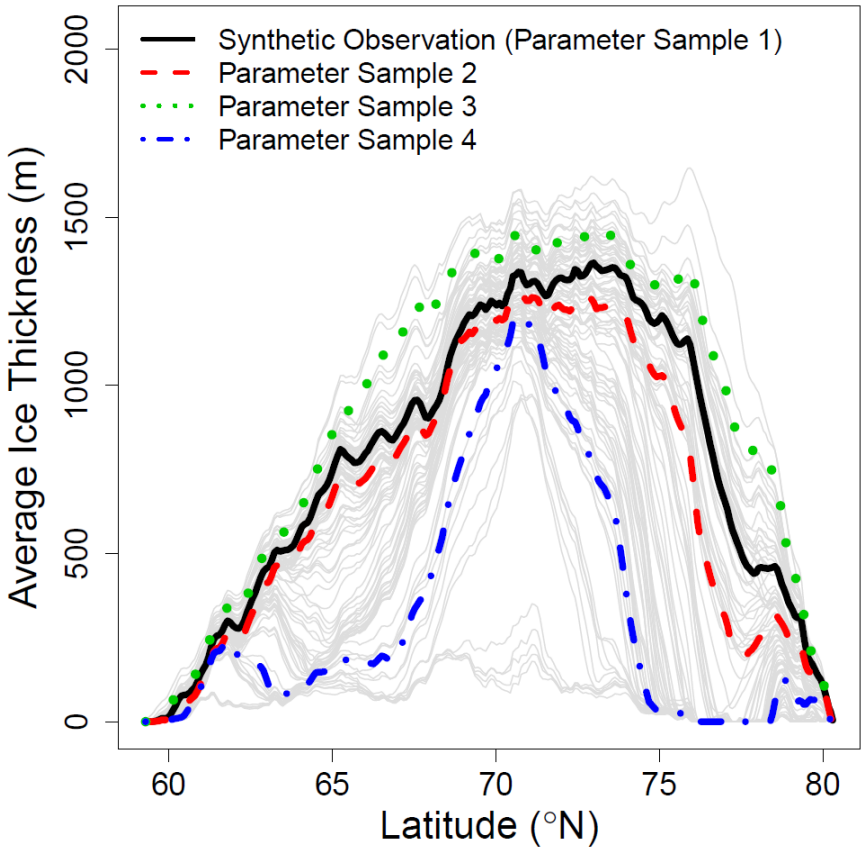

Fig. 1. Profiles of zonal mean ice thicknesses from four different evaluations of the ice sheet model SICOPOLIS (Greve, 1997; Greve et al., 2011). The solid black curve represents model run #67 from Applegate et al. (2012), which we take to be the synthetic truth for our perfect model experiments. The 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.

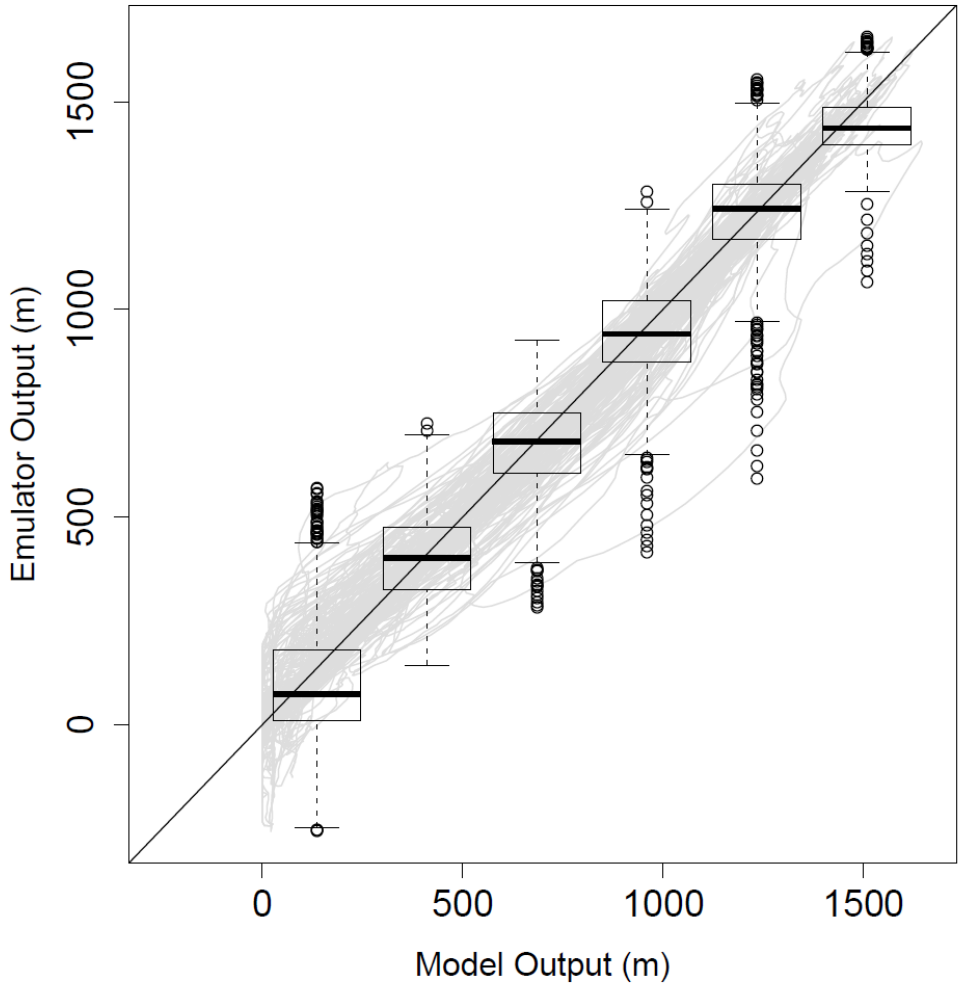

Fig. S4. Leave-one-out cross-validation results for the emulation performance. Each grey curve shows the comparison of zonal mean ice thickness transects from the model output and that from the emulator output for each parameter setting. Each boxplot shows the distribution of emulator output for each 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.

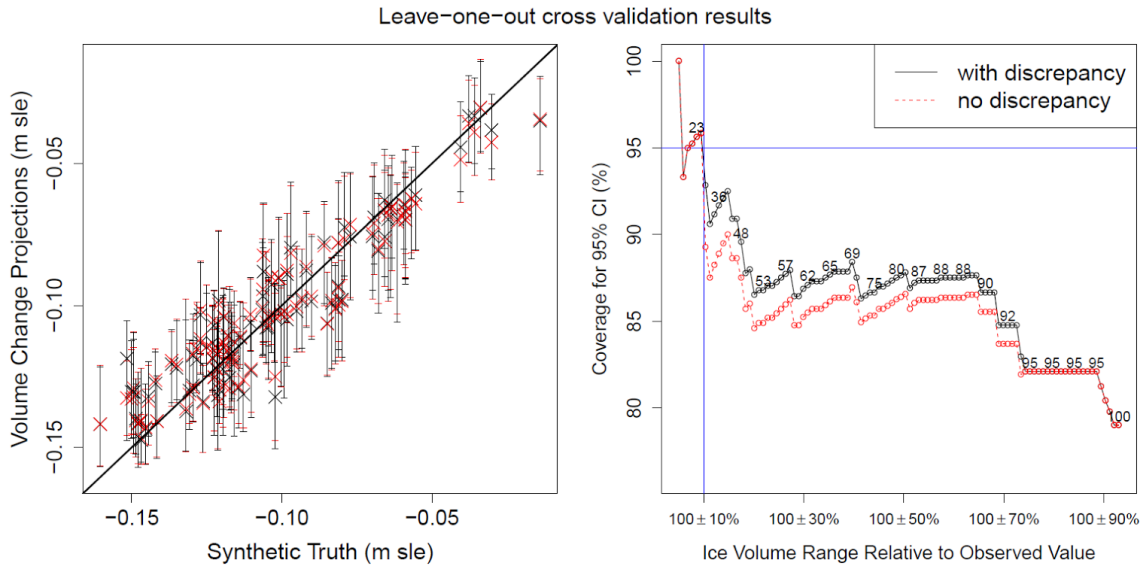

Fig. S5. Leave-one-out cross-validation results for ice volume change projections across all 100 input parameter settings as the synthetic truth. The left penal shows 95% prediction intervals for ice volume change projections across all 100 perfect model experiments conducted for cross-validation. If 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.

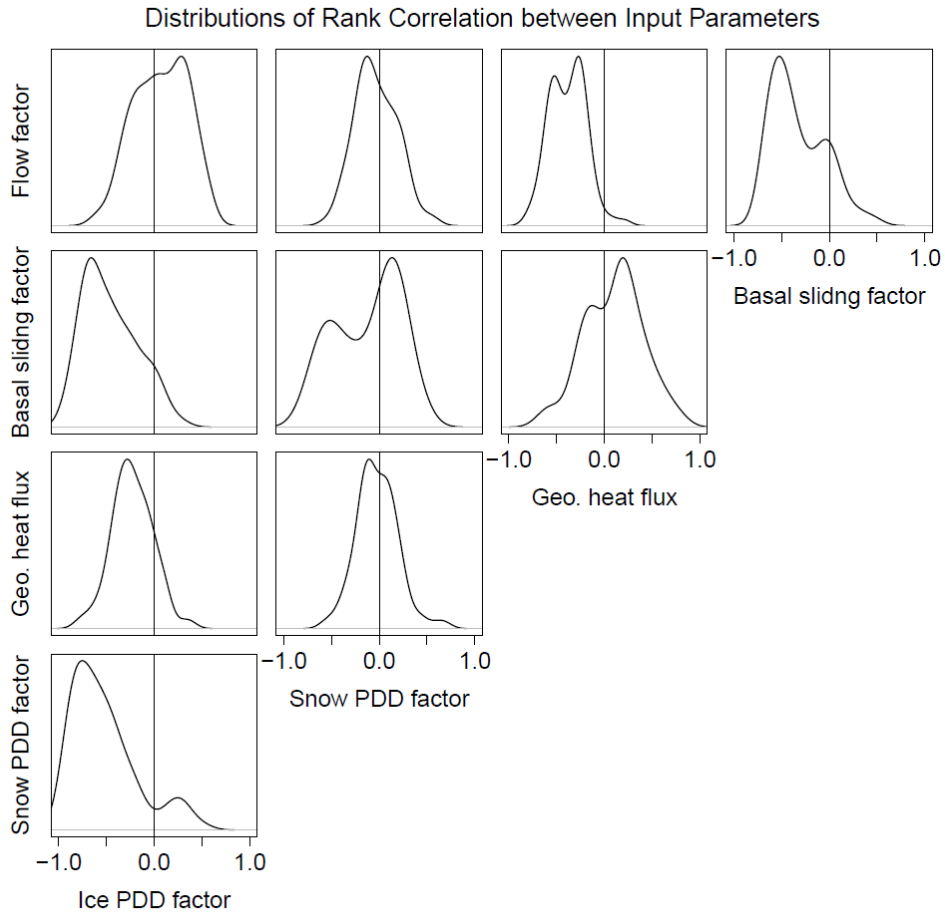

Fig. S6. Summary of interactions between input parameters computed from leave-one-out cross-validation. Each panel shows the distribution of the rank correlation between two input parameters across all synthetic truths in our leave-one-out cross-validation. Five pairs of input parameters, (i) the flow 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.