Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties

W. Chang¹, P. J. Applegate², M. Haran¹, K. Keller³

¹Department of Statistics, Pennsylvania State University, University Park, PA 16802
²Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA 16802
³Department of Geosciences, Pennsylvania State University, University Park, PA 16802

Correspondence to: W. Chang (wonchang@psu.edu)

Abstract
Computer models of ice sheet behavior are important tools for projecting future sea level rise. The simulated modern ice sheets generated by these models differ markedly as input parameters are varied. To ensure accurate ice sheet mass loss projections, these parameters must be constrained using observational data. Which model parameter combinations make sense, given observations? Our method assigns probabilities to parameter combinations based on how well the model reproduces the Greenland Ice Sheet profile. We improve on the previous state of the art by accounting for spatial information, and by carefully sampling the full range of realistic parameter combinations, using statistically rigorous methods. Specifically, we estimate the joint posterior probability density function of model parameters using Gaussian process-based emulation and calibration. This method is an important step toward probabilistic projections of ice sheet contributions to sea level rise, in that it uses observational data to learn about parameter values. This information can, in turn, be used to make projections while taking into account various sources of uncertainty, including parametric uncertainty, data-model discrepancy, and spatial correlation in the error structure. We demonstrate the utility of our method using a perfect model experiment, which shows that many different parameter combinations can generate similar modern ice sheet profiles. This result suggests that the large divergence of projections from different ice sheet models is
partly due to parametric uncertainty. Moreover, our method enables insight into ice sheet processes represented by parameter interactions in the model.

1 Introduction

Accurate projections of future sea level rise are important for present-day adaptation decisions. Global mean sea level has risen 0.2-0.3 m over the last two to three centuries (e.g. Church and White, 2006; Jevrejeva et al., 2008), and this rise is expected to continue in the future (Meehl et al., 2007). A significant fraction of world population and built infrastructure lies near present-day sea level, and these people and resources are at risk from sea level rise. Projections of sea level rise with sound characterization of the associated uncertainties can inform the design of risk management strategies (e.g., Lempert et al., 2012).

Here, we focus on the Greenland Ice Sheet component of future sea level rise, as estimated by ice sheet models. Computer models of ice sheet behavior make up an important member of a suite of methods for projecting sea level rise. Enhanced mass loss from the Greenland Ice Sheet is just one component of overall sea level rise, which also includes contributions from the Antarctic Ice Sheets, small glaciers, thermal expansion of ocean water, and the transfer of water stored on land to the oceans. However, the Greenland Ice Sheet is a large potential contributor to sea level rise, and also a highly uncertain one; if this ice sheet were to melt completely, sea level would rise by about 7 m (Bamber et al., 2001, 2013; Lemke et al., 2007), and both the rate of ice loss and its final magnitude are uncertain (Lenton et al., 2008). Present estimates of future sea level rise are derived primarily from semi-empirical extrapolations of tide gauge data (e.g., Rahmstorf, 2007; Grinsted et al., 2009; Jevrejeva et al., 2012) and expert assessments of future ice sheet behavior (e.g., Pfeffer et al., 2008; Bamber and Aspinall, 2013). Ice sheet models complement these methods, in that they provide internally-consistent representations of the processes that are important to the growth and decay of ice sheets. Although imperfect, 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).

To yield accurate projections, ice sheet models must be started from an initial condition that resembles the real ice sheet as closely as possible, both in terms of the spatial distribution of ice and the temperature distribution within the ice body. Ice flow is driven primarily by thickness and surface slope (e.g., Alley et al., 2010), and warm ice deforms more easily than
cold ice. Similarly, the melt rate of a patch of the ice sheet's surface is strongly sensitive to its
elevation (Born and Nisancioglu, 2012). Thus, errors in the initial condition used for ice sheet
model projections will lead to inaccuracies in simulated future ice distributions and sea level
rise contributions. In practice, all models include simplifications that also affect projection
accuracy (e.g., Kirchner et al., 2011), perhaps more than initial condition errors. However,
matching the modern ice sheet is a frequently-recurring theme in the literature (e.g., Ritz et al.,
1997; Greve, 1997; Huybrechts, 2002; Stone et al., 2010; Greve et al., 2011; Pollard and
DeConto, 2012).

The initial condition used in ice sheet models is a function of input parameter values, as well
as the spinup method. Because the thermal field within the ice sheet is incompletely known,
most modeling studies perform an initialization to bring the simulated ice sheet to a state that
is consistent with the present-day climatology (e.g., Stone et al., 2010), climate model output
(e.g., Fyke et al., 2011), or climate history estimated from ice cores (e.g., Applegate et al.,
2012). Most models allow the simulated ice sheet's surface topography to evolve during the
spinup period; thus, the estimated initial condition usually does not exactly match the
observed ice sheet topography (Bamber et al., 2001, 2013). For example, many studies obtain
a simulated modern Greenland ice sheet that is larger than expected (e.g. Heimbach et al.,
2008; Stone et al., 2010; Robinson et al., 2010; Vizcaino et al., 2010; Greve et al., 2011; cf.
Bamber et al., 2001, 2013). Ice sheet models have many uncertain parameters that affect the
softness of the ice, the speed of basal sliding, and the intensity of surface melting, among
other processes (Ritz et al., 1997; Hebeler et al., 2008; Stone et al., 2010; Fitzgerald et al.,
2011; Applegate et al., 2012). Adjusting these parameters changes the simulated modern ice
sheet (Stone et al., 2010; Applegate et al., 2012).

Despite the importance of achieving a good match between ice sheet model output and the
present-day ice geometry, it remains unclear how to use data on the modern ice sheet to assess
the relative plausibility of different model runs. The root-mean-squared error (RMSE) is
sometimes used for this purpose (e.g., Greve and Otsu, 2007; Stone et al., 2010). However, it
is unclear how to translate the RMSE values from a set of model runs into probabilistic
projections of ice volume change, as required for sea level studies. Using a probability model
that accounts for various uncertainties, as we do here, helps overcome this limitation.

Recent work by McNeall et al. (2013) and Gladstone et al. (2012) partly addresses this
challenge. McNeall et al. (2013) train a statistical emulator (e.g., Sacks et al., 1989; Kennedy
Only when the statistical emulators help overcome this dimensionality problem; however, some method for assigning plausibility scores to the emulator output is also needed.

Here, we address these challenges using a Bayesian framework that combines data, models, and prior beliefs about model input parameter values. Like McNeall et al. (2013), we train an
emulator on an ensemble of ice sheet model runs. However, we build on their work by using
an explicit likelihood function, and by incorporating information from a north-south profile of
average ice thicknesses. Specifically, we use a Gaussian process emulator to estimate the first
10 principal components of the zonal mean ice thickness profile, following a recent climate
model calibration study (Chang et al., 2013). Further, we perform a perfect model experiment
to investigate the interactions between input parameters. Our approach recovers the correct
parameter values and projected ice volume changes from an "assumed-true" model realization,
and the multi-dimensional probability density function displays expected physical interactions
(Section 1.2.1, below). These interactions were not evident from the simple analysis
employed by Applegate et al. (2012, their Fig. 1).

The paper proceeds as follows. In the remainder of the Introduction, we describe the
ensemble that we use to train the emulator. In Section 2, we outline our method for using a
Gaussian process emulator to estimate the principal components of the zonally-averaged ice
thicknesses, and the setup of our perfect model experiment. Section 3 presents the results of
the perfect model experiment. In Section 4, we conclude by pointing out the implications of
our work, as well as its limitations and potential directions for future research.

1.1 The ensemble

We train our emulator with a 100-member perturbed-parameter ensemble described in
Applegate et al. (2012). This ensemble uses the three-dimensional ice sheet model
SICOPOLIS (Greve, 1997; Greve et al., 2011). Each model run spans the period from
125,000 years ago (125 ka BP) to 3500, driven by surface temperature and sea level histories
derived from geologic data (Imbrie et al., 1984; Dansgaard et al., 1993; Johnsen et al., 1997)
and forced into the future with an asymptotic warming to ~5 °C above present values.
SICOPOLIS is a shallow ice-approximation model, meaning that it neglects longitudinal
stresses within the ice body (Kirchner et al., 2011). Like most ice sheet models, it also
includes many simplifications in calculating the surface mass balance, notably through its use
of the positive degree-day method for relating surface temperatures to melting (Braithwaite,
1995; Calov and Greve, 2005; van der Berg et al., 2011). These simplifications improve
SICOPOLIS' computational efficiency relative to higher-order or full-Stokes models (e.g.,
Seddik et al., 2012), allowing it to be run repeatedly over 10^5-yr time scales.
The parameter combinations in the Applegate et al. (2012) ensemble were chosen by Latin hypercube sampling (McKay et al., 1979), following the earlier work of Stone et al. (2010). Latin hypercube sampling distributes points throughout parameter space more efficiently than Monte Carlo methods (Urban and Fricker, 2010). In their experiment, Applegate et al. (2012) varied the ice flow enhancement factor, the ice and snow positive degree-day factors, the geothermal heat flux, and the basal sliding factor (Ritz et al., 1997; cf. Stone et al., 2010; Fitzgerald et al., 2011). These parameters control the softness of ice, the rapidity with which the ice sheet's surface lowers at a given temperature, the amount of heat that enters the base of the ice sheet, and the speed of sliding at a given stress (see Applegate et al., 2012 for an explanation of how each parameter affects model behavior).

McNeall et al. (2013) trained their emulator using a perturbed-parameter ensemble of ice sheet model runs published by Stone et al. (2010). Key differences between the Applegate et al. (2012) ensemble and the Stone et al. (2010) ensemble involve the parameters varied in the ensembles and the processes included in the simulations. Stone et al. (2010) varied the lapse rate instead of the basal sliding factor adjusted by Applegate et al. (2012). The model used by Stone et al. (2010; Glimmer v. 1.0.4; see Rutt et al., 2009) neglects basal sliding, a process included in the SICOPOLIS runs presented by Applegate et al. (2012).

The results presented by Applegate et al. (2012) suggest that widely diverging ice sheet model parameter values yield comparable modern ice sheets, but substantially different sea level rise projections. Applegate et al. (2012) assessed the plausibility of their model runs by comparing the simulated ice volumes in 2005 to the estimated modern ice volume (Bamber et al., 2001; Lemke et al., 2007); those runs that yielded modern ice volumes within 10% of the estimated value were kept. These plausible runs yielded a range of future sea level rise projections that was ~75% of the median estimate.

Moreover, the parameter combinations that agree well with the modern ice volume constraint are widely distributed over parameter space. With the exception of the ice positive degree-day factor, where only values less than ~15 mm day\(^{-1}\) OC\(^{-1}\) satisfy the ice volume constraint, no pattern emerges from the distribution of the successful runs through parameter space. McNeall et al. (2013) make a similar point using their own results. Statistically, this inability to learn about the plausibility of various parameter combinations given observations is termed an "identifiability problem."
1.2 Expected interactions among model input parameters

The apparently-structureless distribution of successful runs through parameter space (Applegate et al., 2012, their Fig. 1) may stem from interactions among the parameters. The parameters can be loosely grouped into those that control the ice sheet's surface mass balance (the ice and snow positive degree-day factors) and those that control ice movement (the ice flow enhancement factor, the basal sliding factor, and the geothermal heat flux). Either group of parameters can cause mass loss from the ice sheet to be high or low, given fixed values of the parameters in the other group. For example, a high ice positive degree-day factor should be associated with a low snow positive degree-day factor to produce the same amount of melt as a model run with more moderate values of both parameters. This interaction is bounded, however, because the maximum snow positive degree-day factor is much lower than the maximum value for ice; also, at the peak of the ablation season, there is no snow left on the lower parts of the ice sheet, so the ice positive degree-day factor dominates over part of the year. Similarly, the same ice velocities can be produced by either a high flow enhancement factor and a low basal sliding factor, or the reverse. Basal sliding can be a much faster process than ice flow, so this parameter interaction is also bounded. However, basal sliding operates only where the bed is thawed, and the geothermal heat flux likely controls the fraction of the bed that is above the pressure melting point.

The relatively small number of design points in the ensemble presented by Applegate et al. (2012) hinders mapping of the interactions among parameters over their five-dimensional space. Coherent mapping requires many more design points, but performing these additional runs with the full ice sheet model is impractical because of the model's high computational cost. This problem suggests a need for a computationally efficient emulator to fill the gaps in parameter space between the existing model runs.

2 Methods

As described above, our goals are 1) to identify a method for quantifying the agreement between ice sheet model output and observations that incorporates spatial information, 2) to characterize the interactions among input parameters, and 3) to produce illustrative projections of sea level rise from the Greenland Ice Sheet based on synthetic data. In this section, we provide an outline of our methods for achieving these goals; fuller descriptions appear in Chang et al. (2013) and in the Supplementary Information.
We accomplish goal #1 through identifying a statistical model that results in a likelihood function. This statistical model compares ice sheet model output and observations to evaluate the plausibility of a vector of model input parameter values $\theta$ while accounting for systematic discrepancies between the model output and the observations. The likelihood function for the ice thickness observations, denoted by $Z$, is based on the additive model

$$Z = Y(\theta) + \delta + \epsilon,$$

(1)

where $Y(\theta)$ is the ice thickness output from SICOPOLIS model at the vector of input parameter values $\theta$, $\delta$ is the discrepancy between model output and observations caused by structural problems in the model, and $\epsilon$ is the independent and identically distributed observational noise. More details of the statistical model, and the resulting likelihood function, are given in the Supplementary Information.

To achieve goal #2, we perform a "leave-one-out" perfect model experiment with a Gaussian process emulator, a computationally-cheap surrogate for the full ice sheet model. As described above, the model output $Y(\theta)$ is available only at a relatively small number of points in parameter space, and therefore it is necessary to build an emulator that approximates the model output $Y(\theta)$ at any given $\theta$.

Direct emulation of the full two-dimensional ice thickness grid is prohibitively expensive, due to the cost of performing operations on large covariance matrices (see the Supplementary Information and Chang et al., 2013, for details). To mitigate this computational challenge, we take the mean of each row in the ice thickness grid, thereby obtaining a 264-element vector of zonally-averaged ice thicknesses for each ice sheet model run. We then apply principal component analysis to these mean ice thickness vectors. The magnitudes of the first 10 principal components suffice to recover the mean ice thickness vectors. Because the principal components are uncorrelated, we can construct a separate emulator for the magnitude of each principal component. Note that our likelihood formulation automatically penalizes the components with lower explained variation.

Next, we train the emulator on all but one of the model runs. We refer to the output (specifically, the zonal mean ice thickness profile and the ice volume change projection) from this left-out model run as our "assumed truth." We examined the robustness of our methods by repeating our experiments using different model runs as the left-out "assumed truth;" see the Supplementary Information.
Before using the mean ice thickness profile from our assumed-true model run in our perfect model experiment, we contaminate it with spatially-correlated errors. These spatially-correlated errors reflect the discrepancies that we would expect to see between model output and data in a "real" calibration experiment, due to missing or parameterized processes in the model. In particular, we use spatially-correlated errors with a moderate magnitude (standard deviation of 50 m) and a large-scale spatial trend to represent a situation in which (i) the ice sheet model has reasonable skill in reproducing the observed spatial pattern of modern ice thickness, and (ii) the discrepancy pattern is notably different from patterns generated by the ice sheet model and is therefore statistically identifiable (see the Supporting Information for more details). Note that results from any probabilistic calibration method, including our approach, can be uninformative if condition (i) is not met, or subject to serious bias if condition (ii) is not met.

We then use Markov chain Monte Carlo (MCMC) to estimate the joint posterior probability distribution over the five-dimensional input parameter space. MCMC is a well-established (Hastings, 1970), but complex, statistical technique; Brooks et al. (2011) provide a book-length treatment. Briefly, the Metropolis-Hastings algorithm used in MCMC constructs a sequence of parameter combinations, each of which is chosen randomly from the region of parameter space surrounding the last point. Candidate parameter combinations are accepted if the posterior probability of the new point is greater than at the previous one, or with a certain probability determined by the Metropolis-Hastings acceptance ratio otherwise. If the candidate point is rejected, another candidate point is chosen at random according to a proposal distribution. 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 we expect that the simplifications involved in constructing the ice sheet model (e.g., Kirchner et al., 2011) will cause problems in matching the modeled ice sheet to observed ice thicknesses. The candidate points that are retained by the MCMC algorithm approximate the posterior probability distribution of the input parameter space. The candidate points from this algorithm therefore reflect various characteristics of the posterior distribution, including the marginal distributions of each of the parameters separately and their joint distributions. Hence, we can use MCMC to summarize what we have learned about the parameters from the model and observations while accounting for various uncertainties and prior information.
Finally, to achieve goal #3, we use a separate Gaussian process emulator to interpolate between the ice volume change projections from all the model runs in the original ensemble (Applegate et al., 2012), except the assumed-true realization. When applied to the sample of the model input parameters that we obtained from Markov chain Monte Carlo, this emulator yields a sample of ice volume changes, and thus sea level rise contributions, between 2005 and 2100. We then used kernel density estimation to compute the probability density of the projected sea level rise contributions. It should be noted that these projections are based on synthetic data (not real observations), and do not represent "real" projections of Greenland Ice Sheet mass loss over this century.

3 Results

Besides helping to diagnose interactions among ice sheet model parameters, our perfect model experiment allows us to test our overall procedure. We carry out several checks.

1) If the trained emulator is given the parameter settings from the left-out model realization, it should produce a close approximation to the actual output from that realization.

2) The maximum of the multidimensional posterior probability function from our Markov chain Monte Carlo analysis should lie close to the parameter settings from the left-out model realization.

3) The mode of the probability density function of ice loss projections should be close to the ice loss projection from the assumed-true model realization.

As detailed below, our methods pass all three of these checks.

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). 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). Thus, our methods pass check #1, above.

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. Similarly, the conditional posterior density functions (Fig. 3) have maxima near the assumed-true parameter values. 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.

For comparison, we also produced scatterplots of parameter combinations as projected onto two-dimensional slices through the five-dimensional parameter space (Fig. 4a), following Applegate et al. (2012, their Fig. 1). As in Applegate et al. (2012), the "successful" design points show no clustering around the assumed-true parameter values.

Our method also successfully recovers the ice volume loss produced by the assumed-true model realization (Fig. 5; see also Figs. S3, S4), reflected by the close correspondence between the mode of the probability density function produced by our methods and the vertical black line. Thus, our methods pass check #3, listed above. As previously noted, these projections are based on synthetic data; they are not "real" projections of Greenland Ice Sheet mass loss. For comparison, we also applied the windowing approach used by Applegate et al. (2012) to the model runs. The 95% probable interval produced by our methods is much smaller than that estimated by Applegate et al. (2012), reflecting the utility of spatial information in reducing projection uncertainties.

The prior density for the ice volume loss was constructed by assuming that all 99 design points used to train our emulator are equally likely. Interestingly, a uniform prior for the input parameters results in a skewed and multimodal prior distribution for the volume loss, indicating that the function that maps input parameters to projected ice volume changes is highly non-linear and not smooth. These characteristics also cause a small offset between the assumed-true projection and the mode of the posterior density. The marginal plots for the volume loss projection surfaces are shown in Figure S1 in the supporting material.
4 Discussion

As explained above, our goals for this work were to identify an objective function for matching ice sheet models to spatially-distributed data (especially ice thicknesses), map interactions among model input parameters, and develop methods for projecting future ice sheet mass loss, with well-characterized uncertainties. We demonstrated that our emulator reproduces a vector of zonally-averaged ice thicknesses from a given model run when trained on other members from the same ensemble (Fig. 2). We further showed that the emulator can recover the appropriate parameter combinations for an assumed-true model realization in a perfect model experiment (Figs. 3, 4b). Finally, we produced illustrative projections of Greenland Ice Sheet mass loss, based on synthetic data (Fig. 5; see also Figs. S3, S4). As noted above, our projections are for illustration only, and do not represent "real" projections of future Greenland Ice Sheet mass loss.

The utility of our approach becomes clear in comparing the marginal posterior probability density functions (Fig. 4a) and projections (red probability density functions and boxplots in Figs. 5, S3, and S4) to results from simpler methods (Fig. 4b; blue boxplots in Figs. 5, S3, and S4; Applegate et al., 2012). In Figure 4b, there are distinct modes in the marginal densities, indicating regions of parameter space that are more consistent with the assumed truth. These modes are absent in the simpler graphic (Fig. 4a). Similarly, the 95% probable interval of sea level rise contributions is narrower using our methods than if a simple windowing approach is applied (Fig. 5; see also Figs. S3, S4). Our results also show the importance of including the discrepancy term ($\delta$ in Eqn. 1) for recovering the appropriate parameter settings in our perfect model experiments (Fig. S2). If we leave this discrepancy term out, the marginal posterior density functions for each parameter clearly miss the true value.

The parameter interactions identified in this experiment are generally consistent with intuition (see Section 1.2.1 for descriptions of anticipated parameter interactions). Figure 4 shows inclined bands of high marginal posterior probability in the ice positive degree-day vs. snow positive degree-day, geothermal heat flux vs. ice flow factor, and basal sliding factor vs. flow factor panels. As expected, there are tradeoffs among each of these parameter pairs; for example, a low ice positive degree-day factor must be combined with a high snow positive degree-day factor to produce a reasonable match to the assumed truth. Somewhat surprisingly, the tradeoff between the geothermal heat flux and the ice flow factor is much stronger than that between the geothermal heat flux and the basal sliding factor. The geothermal heat flux
affects both ice deformation (which is temperature-sensitive) and basal sliding (which operates only where there is liquid water at the ice-bed interface). We hypothesize that the geothermal heat flux has a stronger effect on ice flow than basal sliding because ice deformation happens over a much larger fraction of the ice sheet's basal area than does sliding.

Multiple modes appear in the two-dimensional marginal density plots (Fig. 4), implying that standard methods for tuning of ice sheet models may converge to "incorrect" parameter combinations. Ice sheet models are commonly tuned by manually adjusting one parameter at a time until the simulated modern ice sheet resembles the real one (e.g., Greve et al., 2011). This procedure is an informal variant of so-called gradient descent methods, which search for optimal matches between models and data by moving down a continuous surface defined by the model's input parameters, the objective function, and the data. If the surface has multiple "peaks," 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 "true" parameter combination. 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 (cf. Bindschadler et al., 2013): even if the models had similar structures and reproduced the modern ice sheet equally well, we would still expect their future projections to diverge because of differences in input parameter choice.

### 4.1 Cautions and future directions

In this paper, we specifically avoid giving "real" projections of future Greenland Ice Sheet volume change, for two reasons. First, we match only a two-dimensional profile of zonally-averaged ice thicknesses from an assumed-true model run, rather than the two-dimensional grid of observed ice thicknesses (Bamber et al., 2001, 2013; see also McNeall et al., 2013). Second, the ensemble of ice sheet model runs (Applegate et al., 2012) that we use to calibrate our emulator has several important limitations, including the relative simplicity of the model used to generate the ensemble and the synthetic climate scenario used to drive the ensemble members into the future. Most importantly, this ensemble produced simulated modern ice sheets that are generally too thick in the southern part of Greenland and too thin in the northern part of the island (Applegate et al., 2012, their Fig. 7); 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). Thus, the emulator discrepancy
term, used to correct for differences between model output and observations, would need to be large, and the effects of this large discrepancy term on the projections are difficult to assess. The long-term goal of this work is to compare ice sheet model runs to actual data, thereby resulting in probabilistic projections of future ice sheet mass loss. To achieve this goal, we plan to expand our method to treat the full, two-dimensional ice thickness grid and take advantage of other spatially-distributed data sets (e.g., surface velocities; Joughin et al. 2010), and to generate new ice sheet model ensembles that overcome the limitations explained above.

5 Conclusions

In this paper, we presented an approach for probabilistic calibration of ice sheet models using spatially-resolved ice thickness information. Specifically, we constructed a probability model for assigning posterior probabilities to individual ice sheet model runs, and we used a Gaussian process emulator to interpolate between existing ice sheet model simulations. We reduced the dimensionality of the emulation problem by reducing profiles of mean ice thicknesses to their principal components. Finally, we showed how the posterior probabilities from the model calibration exercise can be used to make projections of future sea level rise from the ice sheets. In a perfect model experiment where the "true" parameter settings and future contributions of the ice sheet to sea level rise are known, our methods successfully recovered these values. The posterior probability density function that resulted from this experiment shows tradeoffs among parameters and multiple modes. The tradeoffs are consistent with physical expectations, whereas the multiple modes may indicate that commonly-applied methods for tuning ice sheet models can lead to calibration errors.

Acknowledgements

We thank R. Greve for distributing his ice sheet model SICOPOLIS freely on the Web (http://sicopolis.greveweb.net/), and N. Kirchner for help in setting up and using the model. R. Alley and D. Pollard provided helpful comments on a draft of the manuscript. This work was partially supported by the US Department of Energy, Office of Science, Biological and Environmental Research Program, Integrated Assessment Program, Grant No. DE-SC0005171; by the US National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507; and
by the Penn State Center for Climate Risk Management. Any opinions, findings, and
conclusions expressed in this work are those of the authors, and do not necessarily reflect the
views of the National Science Foundation or the Department of Energy.

Author contributions

WC designed the emulator, carried out the analyses, and wrote the first draft of the
Supplementary Information. PJA wrote the first draft of the body text and supplied the
previously-published ice sheet model runs (available online at
http://bolin.su.se/data/Applegate-2011). WC, PJA, MH, and KK jointly designed the research
and edited the paper text.
References


**Figure 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.
Figure 2. Comparison of zonal mean ice thickness transects from the assumed-true model run (#67 from Applegate et al., 2012) and that generated by the trained emulator at the same parameter combination as used in the assumed-true model run. In the top panel, the assumed-true profile is shown by a solid black line, and the emulator output is shown by a dashed red curve with circles. In the lower panel, each point stands for an individual latitude location. The red circles in the top panel fall almost exactly on top of the black curve, and the points in the lower panel fall almost exactly on a 1:1 line connecting the lower left and upper right corners of the plot. Thus, the emulator successfully recovers the ice thicknesses from an assumed-true model realization when trained on the other model runs from the same ensemble.
Figure 3. Prior (dashed red curves) and posterior (solid black curves) probability density functions of each input parameter, assuming that all the other parameters are held fixed at their assumed-true values. The vertical lines indicate the assumed-true values of the individual parameters.
**Figure 4.** Comparison between an exploratory data analysis, following Applegate et al. (2012), and the results of our probabilistic calibration. (a) Scatterplots of parameter settings used to train the emulator, as projected onto two-dimensional marginal spaces. Red dots, parameter settings resulting in simulated modern ice volumes within 10% of the synthetic truth (model run #67 of Applegate et al. 2012); blue crosses, parameter settings that yield ice volumes more than 10% larger or smaller than the synthetic truth. (b) Two-dimensional marginal posterior densities of all pairs of input parameters. Several of the marginal posterior density maps show inclined bands of higher probability, indicating interactions among parameters; other panels show multiple modes, representing potential "traps" for tuning of ice sheet models using simpler methods. See text for discussion.
Figure 5. Illustrative (not "real") ice volume change projections between 2005 and 2100, based on three different methods: i) the prior density of the input parameters (dashed green line); ii) parameter settings that pass the 10% ice volume filter used by Applegate et al. (2012) (solid blue line); and iii) the posterior density computed by our calibration approach (solid red line). The vertical line shows the ice volume change projection for the assumed-true parameter setting. The horizontal lines and the parentheses on them represent the range and the 95% prediction intervals, respectively; the crosses indicate the median projection from each method. The width of the 95% projection interval from our methods is narrower than if simpler methods are applied (blue boxplot; Applegate et al., 2012). Similar results are obtained if different model runs from the ensemble are left out (see Figs. S3 and S4). See text for discussion. m sle, meters of sea level equivalent.
1. Gaussian process emulator for principal components

In this section, we outline our statistical approach for ice sheet model emulation using Gaussian process (GP) models and principal component (PC) analysis (often referred to as empirical orthogonal functions, EOFs). Our approach follows Chang et al. (2013) in that we summarize the ice sheet model runs as PCs and calibrate the ice sheet parameters based on GP emulators for PCs. Our description of methods below therefore also closely follows the notation and description in Chang et al. (2013). By decomposing spatial patterns into a small number of variables representing important characteristics of model runs, our approach drastically increases computational efficiency without causing significant information loss.

We denote the number of model runs by $p$ and the number of spatial locations spatial locations by $n$. For the SICOPOLIS model output (from Applegate et al. 2012) we use here, $p = 99$ and $n = 264$. We let $Y(\theta, s)$ denote the ice thickness from the ice sheet model at a parameter setting $\theta = (\theta_1, \ldots, \theta_5)^T$ and a spatial location $s$. We let $s_1, \ldots, s_n$ be the spatial locations of the model grid points and $Y(\theta) = (Y(\theta, s_1), \ldots, Y(\theta, s_n))$ be the vector
of model output at a parameter setting $\theta$. Let $\theta_1, \ldots, \theta_p$ be the vectors of input parameters for our model. $Y$ is an $n \times p$ matrix of the ice sheet model output where its rows correspond to spatial locations and columns to parameter settings, i.e.

$$Y = \begin{pmatrix} Y(\theta_1, s_1), & Y(\theta_2, s_1), & \ldots, & Y(\theta_p, s_1) \\ Y(\theta_1, s_2), & Y(\theta_2, s_2), & \ldots, & Y(\theta_p, s_2) \\ \vdots & \vdots & \ddots & \vdots \\ Y(\theta_1, s_n), & Y(\theta_2, s_n), & \ldots, & Y(\theta_p, s_n) \end{pmatrix}.$$

Similarly, $Z(s)$ denotes the observed ice sheet thickness at a location $s$, and $Z = (Z(s_1), \ldots, Z(s_n))^T$ is the $n \times 1$ vector of the observational data.

2. Principal component analysis for model output

The first step is summarizing the model output by principal component analysis. Following the standard procedure of principal component analysis, the column means are subtracted from each element in the corresponding columns such that each column is centered on zero. We apply singular value decomposition to this centered output matrix to find the scaled principal basis vectors $k_1 = \sqrt{\lambda_1}e_1, \ldots, k_p = \sqrt{\lambda_p}e_p$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ and $e_1, \ldots, e_p$ are ordered eigenvalues and their eigenvectors respectively. Each eigenvalue represents the explained variation for the corresponding principal component. We keep only the first $J \ll p$ PCs with the largest explained variation (i.e. the largest eigenvalues) to minimize the information loss due to dimension reduction. The principal components for model output can be computed by

$$Y^R = (K^T_yK_y)^{-1}K^T_yY = (Y^R_1 \ldots Y^R_J)^T$$

where $K_y = (k_1, \ldots, k_J)$ is the principal basis matrix. $Y^R_i = (Y^R_i(\theta_1), \ldots, Y^R_i(\theta_p))^T$ is the $p \times 1$ vector of the $i$th principal components, and $Y^R_i(\theta_j)$ is the $i$th principal component at the parameter setting $\theta_j$. The resulting matrix $Y^R$ is the summarized output matrix with
rows for PCs and columns for parameter settings. The procedure reduces the size of the data from $n \times p$ to $J \times p$.

3. Gaussian process emulator

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; Higdon et al. 2008; Drignei et al. 2008; Holden et al. 2010; Bhat et al. 2012; Olson et al. 2012, 2013). 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). Moreover, in addition to its optimality in interpolating smoothly varying functions, the method enables a natural quantification of uncertainty. The interpolator is essentially 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.

Because the principal components are uncorrelated with each other by construction, we can model each of them separately using independent GPs. Note that this basically ignores the dependence between the principal components that is not captured by the covariances. However, according to our experiences for various models including SICOPOLIS, the emulator based on this assumption usually provides a very accurate approximation to the original model that is being emulated. We model each $Y^R_i$ using a GP with mean zero and covariance determined by the following squared exponential covariance function:

$$ Cov(Y^R_i(\theta_j), Y^R_i(\theta_k); \zeta_i, \kappa_{y,i}, \phi_i) = \zeta_i 1(\theta_j = \theta_k) + \kappa_{y,i} \exp \left(-\sum_{l=1}^{5} \left(\frac{\theta_{jl} - \theta_{kl}}{\phi_{il}}\right)^2\right), $$

where $\zeta_i, \kappa_{y,i}, \phi_{i1}, \ldots, \phi_{i5} > 0$ are covariance parameters, $\theta_{jl}$ is the $l$th element of $\theta_j$, and $1(\cdot)$ is the index function. The covariance parameters $(\zeta_1, \kappa_{1,y}, \phi_{11}, \ldots, \phi_{15}), \ldots, (\zeta_J, \kappa_{y,J}, \phi_{J1}, \ldots, \phi_{J5})$ are estimated by maximum likelihood estimation (MLE). Our emulator, denoted by $J \times 1$ vector-valued function $\eta(\theta, Y^R)$, is the predictive distribution of PCs at an untried param-
eter setting $\theta$ defined by the fitted GPs. Using the PC emulator, we can also emulate the original model transect by computing $K_y \eta(\theta, Y^R)$.

Note that our approach allows significant improvements in computational efficiency. Without any dimension reduction, the computational cost for a single likelihood evaluation scales as $O(n^3 p^3)$, which corresponds to a few hours of computing time. Thus, application of any numerical methods requiring repeated evaluation of the likelihood function is computationally prohibitive if no dimensional reduction is performed. Our approach decreases the computational complexity to $O(Jp^3)$, and this is a reduction from $3.18 \times 10^{14}$ flops to $1.56 \times 10^8$ flops in our case. The computing time reduces to less than a second for a single likelihood evaluation.

4. Model parameter calibration

In this section, we formulate the probability model for calibration using the PC emulator constructed above and explain the inference procedure for the model parameters using Markov chain Monte Carlo (MCMC).

We assume that the observational dataset is emulator output contaminated by model discrepancy and observational error;

$$Z = K_y \eta(\theta^*, Y^R) + K_d \nu + \epsilon;$$

(S1)

where $\theta^*$ is the best fit input parameter setting (Bayarri et al. 2007) for the observational data, and $\epsilon \sim N(0, \sigma^2 I_n)$ is the observational error with variance $\sigma^2 > 0$. $K_d \nu$ is the model-observation discrepancy picking up systematic differences between the model and the observations (cf. Bayarri et al. 2007; Bhat et al. 2012), where $K_d$ is a kernel basis matrix relating the spatial locations $s_1, \ldots, s_n$ to $J_d$ knot locations $a_1, \ldots, a_{J_d}$, and $\nu \sim N(0, \kappa_d I_{J_d})$ is the vector of knot processes, a set of random variables assigned to each of the knot locations with variance $\kappa_d > 0$. Our choice for the kernel function is an exponential covariance given
by
\[{K_d}_{ij} = \exp\left(-\frac{|s_i - a_j|}{\phi_d}\right),\]
with \(\phi_d > 0\). The variance parameter \(\kappa_d\) is subject to inference, and the correlation parameter \(\phi_d\) is pre-specified by expert judgment. In our implementation, we choose \(\phi_d\) as 5% of the maximum distance between the spatial locations on the model grid to yield a sufficiently flexible discrepancy pattern. Note that the kernel basis often needs to be substituted by its scaled principal basis (eigenvectors) to improve identifiability. See Chang et al. (2013) for a more detailed discussion. We used the 30 leading principal basis for \(K_d\) in our implementation. We apply a similar dimension reduction described in the previous section to find \(Z_R\), a summary of the observed transect as follows:

\[Z_R = (K^T K)^{-1} K^T Z,\] (S2)

and therefore the model for \(Z_R\) can be written as

\[Z_R \sim N\left(\begin{pmatrix} \mu_\eta \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\eta & 0 \\ 0 & \kappa_d I_d \end{pmatrix} + \sigma^2 (K^T K)^{-1}\right),\]

where \(\mu_\eta\) and \(\Sigma_\eta\) are the mean and covariance, respectively, of the emulator \(\eta(\theta^*, Y^R)\).

The parameters to be estimated in the calibration model are the ice sheet model input parameters \(\theta^*\), the discrepancy parameter \(\kappa_d\), and the observational error variance \(\sigma^2\). We also re-estimate the partial sill parameters \(\kappa_y = (\kappa_{y,1}, \ldots, \kappa_{y,J})\) for the emulator (Bayarri et al. 2007; Bhat et al. 2012; Chang et al. 2013). We define the posterior density based on the likelihood function given by (S2) denoted by \(\ell(Z_R| \theta^*, \kappa_y, \kappa_d, \sigma^2, Y^R)\) and some standard prior specifications denoted by \(f(\theta^*), f(\kappa_y), f(\kappa_d), \) and \(f(\sigma^2)\) (Higdon et al. 2008; Chang et al. 2013). Each of the input parameters in \(\theta^*\) receives a flat prior on a broad range determined by model ensemble design and physical knowledge. The observational error variance \(\sigma^2\) and the variance for the discrepancy \(\kappa_d\) have non-informative inverse-gamma priors with small shape parameters. We specify somewhat informative priors for \(\kappa_{y,1}, \ldots, \kappa_{y,J}\) by specifying a large shape parameter in order to avoid numerical instability and identifiability issues.
The posterior distribution resulting from the above model is

$$\pi(\theta^*, \kappa_y, \kappa_d, \sigma^2 | Z^R, Y^R) \propto \ell(Z^R | \theta^*, \kappa_y, \kappa_d, \sigma^2, Y^R) f(\theta^*) f(\kappa_y) f(\kappa_d) f(\sigma^2),$$

where

$$\ell(Z^R | \theta^*, \kappa_y, \kappa_d, \sigma^2, Y^R) \propto \left| \Sigma_\eta + K^T K \sigma^2 \right|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} Z^R^T (\Sigma_\eta + K^T K \sigma^2)^{-1} Z^R \right).$$

For each \( i \), we set \( a_{y,i} = 50 \) and choose \( b_{y,i} \) such that the mode of the prior density \( b_{y,i} / (a_{y,i}+1) \) coincides with the MLE of \( \kappa_{y,i} \) computed in the emulation stage. For other parameters, we impose vague priors by setting \( a_d = 2, b_d = 3, a_\sigma = 2, \) and \( b_\sigma = 3 \).

The synthetic observations used in our perfect model experiment are constructed by superimposing a random error generated from a Gaussian process model on the assumed true ice sheet status (run # 67). The covariance function that we use for the Gaussian process model 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 discrepancy process is based on the following two general assumptions: (i) the discrepancy is statistically identifiable from the emulator process, and (ii) SICOPOLIS has an enough skill to reproduce the observed ice profile. (i) is related to the value of the range parameter, which controls the effective distance at which two spatial locations are uncorrelated. To ensure that the discrepancy process is identifiable from the emulator process, we set the range parameter to be very large (80% of the spatial range of the model output) so that the discrepancy operates in a different spatial scale to the emulator process. (ii) is related to the value of the partial sill, which defines the magnitude of the discrepancy. Here we let the value of the partial sill to be reasonably small to simulate the situation that the structural error is not large.
and therefore SICOPOLIS can reproduce the observed ice profile reasonably well. Note that calibration based on any framework including our approach can become problematic if any of the assumptions are violated; if the discrepancy process operates in a similar spatial scale to the emulator process (i.e. (i) does not hold), the discrepancy causes identifiability issues and hence introduces a significant bias in the calibration result. If the magnitude of the discrepancy is too large (i.e. (ii) does not hold) compared to the variation between model outputs, the calibration results will become essentially non-informative (i.e. resulting in a very dispersed posterior density). Note that these are common issues for most existing calibration methods in general.

Based on the pseudo observations, we infer the parameters using the MCMC sample from the above posterior distribution obtained by the Metropolis-Hastings algorithm (cf. Higdon et al. 2009). In particular, we infer the input parameters in $\theta^*$ by investigating their marginal density $\pi(\theta^*|Z^R, Y^R)$. In our perfect model experiment, we obtained 300,000 draws using block updating when estimating the full joint density of all five parameters. The computing time takes about eight hours on a single high-performance core. For inference on individual input parameter, only 30,000 draws using block updating is sufficient. In both cases, we confirmed that the Monte Carlo chain is well-mixed by comparing the densities of the first half of the chain with the entire chain. We find the probability density of the input parameters via kernel density estimation for the MCMC sample. The estimated density can be easily plotted for visual analysis as shown in Figures 3 and 4. Note that ignoring the spatially correlated discrepancy results in a notably biased calibration results in our perfect model experiment. See Figure S2 for a comparison of posterior densities with and without the discrepancy term.

5. Ice volume change projection based on calibrated parameters
One important purpose of parameter calibration is making better projections for the future ice sheet mass loss. Making future projections based on calibration results requires a function that relates input parameter values $\theta^*$ to future changes in ice sheet volume. In our illustrative example, the variable that we want to project is the ice volume change from present to 2100 in meters of sea level equivalence. For each model run, we compute the ice volume change by subtracting the current ice volume from the future ice volume. We then obtain a 5-dimensional surface of ice volume change by interpolation between those computed changes.

Among many possible choices for the interpolator, we use the Gaussian process emulator similar to the model described in 3. More specifically, we fit a Gaussian process model for the ice volume change over the input parameter space with zero-mean and the covariance function

$$
\text{Cov}(\Delta v(\theta_j), \Delta v(\theta_k); \zeta^\text{vol}, \kappa^\text{vol}, \phi^\text{vol}_1) = \zeta^\text{vol}1(\theta_j = \theta_k) + \kappa^\text{vol}\exp\left(-\sum_{l=1}^5 \frac{|\theta_{jl} - \theta_{kl}|}{\phi^\text{vol}_l}\right),
$$

for any given design points $\theta_j$ and $\theta_k$ ($j, k = 1, \cdots, 100$), where $\Delta v(\theta)$ is the volume change at a parameter setting $\theta$, and $\zeta^\text{vol}, \kappa^\text{vol}, \phi^\text{vol}_1, \cdots, \phi^\text{vol}_5 > 0$ are the covariance parameters that need to be estimated via MLE. The resulting function can predict ice volume change at any given value of $\theta$ as the conditional mean given by the standard kriging approach (Cressie 1993). Figure S1 shows the marginal surface of the projection as a function of input parameters. To validate the emulator constructed here, we have conducted leave-5-percent-out cross validation and the mean error rate is around 16%; the error rate is a little higher than the heuristic upper limit for the generally acceptable emulation error (10%) due to the irregular behavior of the volume change surface.

We obtain a Monte Carlo sample of ice volume projections by supplying the posterior sample of the calibrated parameters to the interpolation function. Each element of the posterior sample is converted to ice volume change. The predictive density of the ice volume projection can be found by applying kernel density estimation. We find the prior density of the projections in the same manner; we convert the design points of the existing model runs
into the ice volume changes and compute the predictive density for it using kernel density estimation.

To investigate whether the perfect model experiment results shown in the main text are sensitive to the values of input parameters assumed as the synthetic truth, we have conducted perfect model experiments for additional parameter settings other than the one used in the manuscript. As illustrative examples, we below present the calibrated ice volume change projections (Figure S3 and S4) for two input parameter settings that result in the minimum and the maximum projected ice volume changes from 2005 to 2100 among the parameter settings that produce the modern ice volumes within 15 percent of the observed ice volume (Bamber et al. 2001). The results are essentially the same as the one presented in the manuscript; the densities of projected sea level rise peak around the true sea level rise values and the projection uncertainties have been significantly reduced comparing to the simpler method by Applegate et al. (2012).

6. Summary

We describe an ice sheet model calibration approach based on PCs of the model output and the observational data. We build a GP emulator for the PCs of the model output as a fast approximation to the ice sheet model. The calibration model links the observed PCs with the input parameters using the GP emulator while taking the systematic discrepancy into account. We infer the input parameters along with other statistical parameters in the calibration model using MCMC. Combined with projections generated by the ice sheet model, the resulting posterior density of the parameters provide calibrated probabilistic projections of the future ice sheet volume changes.
REFERENCES


Higdon, D., J. Gattiker, B. Williams, and M. Rightley, 2008: Computer model calibration using high-dimensional output. 103 (482), 570–583.


List of Figures

S1 Surfaces of ice volume change projections between 2005 and 2100 projected onto marginal spaces of all pairs of input parameters. Many local maxima and minima are scattered around the parameter space, indicating that the surfaces behave very irregularly and exhibit highly nonlinear relationship with the input parameters. m sle, meters of sea level equivalent.

S2 Comparison between calibration results with and without the discrepancy term $K_d\nu$ in the calibration model in (S1). In each panel, we tried to learn each of the parameters while fixing the other parameters at their assumed-true values. The prior densities are assumed to be uniform over a broad range (dashed red lines). While the posterior densities computed by including the discrepancy term in the model (solid black curves) pick up the true parameter values without notable biases, the posterior densities without the discrepancy term (solid blue curves) cannot recover the true values.

S3 Illustrative (not “real”) ice volume change projections between 2005 and 2100 for model run #23 in Applegate et al. (2012), based on three different methods: i) the prior density of the input parameters (dashed green line); ii) parameter settings that pass the 10% ice volume filter used by Applegate et al. (2012) (solid blue line); and iii) the posterior density computed by our calibration approach (solid red line). The model run has the smallest projected ice volume change from 2005 to 2100 among the model runs that yield modern ice volume within 15% of the observed modern ice volume. The vertical line shows the ice volume change projection for the assumed-true parameter setting. The horizontal lines and the parentheses on them represent the range and the 95% prediction intervals, respectively; the crosses indicate the median projection from each method.
The same comparison as Figure S3 for the model run #91 in Applegate et al. (2012), which results in the largest projected ice volume change from 2005 to 2100 among the model runs that yields the modern ice volume within 15% of the observed volume.
Fig. S1. Surfaces of ice volume change projections between 2005 and 2100 projected onto marginal spaces of all pairs of input parameters. Many local maxima and minima are scattered around the parameter space, indicating that the surfaces behave very irregularly and exhibit highly nonlinear relationship with the input parameters. m sle, meters of sea level equivalent.
Fig. S2. Comparison between calibration results with and without the discrepancy term $K_{d,v}$ in the calibration model in (S1). In each panel, we tried to learn each of the parameters while fixing the other parameters at their assumed-true values. The prior densities are assumed to be uniform over a broad range (dashed red lines). While the posterior densities computed by including the discrepancy term in the model (solid black curves) pick up the true parameter values without notable biases, the posterior densities without the discrepancy term (solid blue curves) cannot recover the true values.
Fig. S3. Illustrative (not “real”) ice volume change projections between 2005 and 2100 for model run #23 in Applegate et al. (2012), based on three different methods: i) the prior density of the input parameters (dashed green line); ii) parameter settings that pass the 10% ice volume filter used by Applegate et al. (2012) (solid blue line); and iii) the posterior density computed by our calibration approach (solid red line). The model run has the smallest projected ice volume change from 2005 to 2100 among the model runs that yield modern ice volume within 15% of the observed modern ice volume. The vertical line shows the ice volume change projection for the assumed-true parameter setting. The horizontal lines and the parentheses on them represent the range and the 95% prediction intervals, respectively; the crosses indicate the median projection from each method.
**Fig. S4.** The same comparison as Figure S3 for the model run #91 in Applegate et al. (2012), which results in the largest projected ice volume change from 2005 to 2100 among the model runs that yields the modern ice volume within 15% of the observed volume.