SIAM/ASA J. Uncertainty Quantification
Vol. 10, No. 2, pp. 717–744

Quantifying Spatio-Temporal Boundary Condition Uncertainty for the North American Deglaciation*

James M. Salter†, Daniel B. Williamson‡, Lauren J. Gregoire§, and Tamsin L. Edwards¶

Abstract. Ice sheet models are used to study the deglaciation of North America at the end of the last ice age (past 21,000 years), so that we might understand whether and how existing ice sheets may reduce or disappear under climate change. Though ice sheet models have a few parameters controlling physical behavior of the ice mass, they also require boundary conditions for climate (spatio-temporal fields of temperature and precipitation, typically on regular grids and at monthly intervals). The behavior of the ice sheet is highly sensitive to these fields, and there is relatively little data from geological records to constrain them as the land was covered with ice. We develop a methodology for generating a range of plausible boundary conditions, using a low-dimensional basis representation of the spatio-temporal input. We derive this basis by combining key patterns, extracted from a small ensemble of climate model simulations of the deglaciation, with sparse spatio-temporal observations. By jointly varying the ice sheet parameters and basis vector coefficients, we run ensembles of the Glimmer ice sheet model that simultaneously explore both climate and ice sheet model uncertainties. We use these to calibrate the ice sheet physics and boundary conditions for Glimmer by ruling out regions of the joint coefficient and parameter space via history matching. We use binary ice/no ice observations from reconstructions of past ice sheet margin position to constrain this space by introducing a novel metric for history matching to binary data.

Key words. history matching, calibration, emulation, dimension reduction, ice sheet model, uncertainty quantification

AMS subject classifications. 62P12, 60G15

DOI. 10.1137/21M1409135

1. Introduction.

The last deglaciation, involving the melting of the North American ice sheet, occurred from the last glacial maximum around 21 thousand years ago (ka) onwards (Carlson and Clark, 2012). By 6 ka, the ice sheet had almost disappeared from North America. The feedback between past climate and ice sheet melt is poorly understood (Ivanovic et al., 2016, 2018), with uncertainty in how much sea level rise can be attributed to ice melt caused by rapid warming events (Carlson and Clark, 2012), so ice sheet models are used to study the

* Received by the editors April 1, 2021; accepted for publication (in revised form) February 1, 2022; published electronically June 29, 2022.
https://doi.org/10.1137/21M1409135
Funding: This work was supported by the EPSRC-funded Past Earth Network grant EP/M008363/1 and the Isaac Newton Institute for Mathematical Sciences, Cambridge, Uncertainty Quantification program (EPSRC grant EP/K032208/1).

† Department of Mathematics, University of Exeter, Exeter, EX4 4QF, UK (j.m.salter@exeter.ac.uk).
‡ College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QF, UK, and Alan Turing Institute, London, NW1 2DB, UK (d.williamson@exeter.ac.uk).
§ Leeds University, Leeds, LS2 9JT, UK (l.j.gregoire@leeds.ac.uk).
¶ King's College London, London, WC2R 2LS, UK (tamsin.edwards@kcl.ac.uk).

deglaciation (see, e.g., Gregoire et al., (2012, 2016); Patton et al. (2017)). If the mechanisms that led to rapid warming and ice sheet melt in the past could be better understood, then these might be used to constrain predictions of future climate and ice sheet changes, improving their accuracy and reducing their uncertainty. However, climate is the largest source of uncertainty in modeling past ice sheet evolution (Seguinot et al., 2014; Charbit et al., 2007), and this source of uncertainty is challenging to characterize.

Ice sheet models comprise sets of partial differential equations (PDEs), containing parametrizations of physical processes that control the evolution of ice sheets. Thus ice sheet models include a number of parameters that control the flow and melt of ice sheets, but are only loosely constrained by observations. These parameters can be varied, with the output of the model being the evolution of the ice sheet extent, thickness, and flow over time. As well as these parameters that control the ice sheet behavior, simulating ice sheet evolution requires information about the climate, which controls the surface mass balance of an ice sheet (the balance between the accumulation of snow and the melting of snow and ice at the surface of the ice sheet). Although surface mass balance depends on processes that occur at small spatial and temporal scales (hourly, 1 km or less) (Noël et al., 2016), ice sheet models typically parameterize it as a function of monthly mean temperature and precipitation (e.g., the positive degree day method) (Reeh, 1989). Such parametrizations are useful for simulating the evolution of large continental scale ice sheets over long time scales. Yet the deglaciation occurs over millennia, so the monthly mean temperature and precipitation fields (the "boundary condition" to the ice sheet model) have an extremely high dimension: for a 48×37 spatial field (resolution of FAMOUS (Smith et al., 2008)) varying monthly for 15,000 years (we focus here on 21 ka to 6 ka), this requires around 320 million values, for temperature and precipitation separately.

In order to have confidence in the model, the inputs must be tuned ("calibrated") so that the output matches historical observations (as paleo quantities cannot be directly measured, we use the term "observations" to refer to reconstructions obtained from proxies) of the ice sheet (Hourdin et al., 2017). To successfully calibrate such a computer model, it is important to vary both the input parameters and boundary conditions. Due to the dimension of the boundary conditions needed for ice sheet models, it is attractive to use the output of climate models (Gregoire et al., 2016). However, running even a low-resolution climate model for the entire deglaciation is very expensive, requiring supercomputer time, and runs that are available may contain biases to temperature records for the particular time period and region of interest. Accurate past climate boundary conditions are needed to achieve a realistic deglaciation, and biases in the global climate model output may result in the ice sheet model being unable to reproduce historical observations of the ice sheet (Charbit et al., 2007; Seguinot et al., 2014). The space of potential boundary conditions has an extremely high dimension, with the limited number of climate model runs forming only a small subset of this space, not necessarily containing the part of space that is consistent with temperature records. Therefore, an alternative method is required to properly evaluate the effect of uncertain boundary conditions on the model output.

To tune computationally expensive models, the uncertainty quantification (UQ) field uses probabilistic calibration (Kennedy and O'Hagan, 2001; Higdon et al., 2008) and history matching (Craig et al., 1996, 2001), using statistical models ("emulators") that can be evaluated

quickly in place of the expensive computer model. Given the output of the computer model at a small number of settings of the inputs, an emulator is used to predict the output at unseen parameter settings, with an uncertainty on the prediction. Emulating and then calibrating computer models has been performed extensively (Williamson et al., 2013; Chang et al., 2014; Holden et al., 2015; Salter et al., 2019; Edwards et al., 2018), but the uncertainty due to the boundary conditions is not always considered. Pollard et al. (2016) and Chang et al. (2016) apply emulation and calibration to an ice sheet model, varying inputs relating to the ice sheet physics, but fixing the boundary condition using the output of another climate model.

Methods for reducing high-dimensional input spaces have been developed; e.g., Liu and Guillas (2017) model the spatial input (bathymetry) of a tsunami model via a stochastic partial differential equation model, before using gradient-based kernel dimension reduction prior to building emulators. The bathymetry has 3,200 dimensions, orders of magnitude smaller than the boundary conditions required for the deglaciation, with the bathymetry observed at locations across the whole spatial domain. In our application, the geological temperature observations are sparse, both spatially and temporally, with no observations over North America, the region for which we are studying the evolution of the ice sheet. The majority of observations are hundreds of kilometers apart, and ocean based, so that an attempt to use a purely stochastic process-based model for the boundary conditions may have problems setting appropriate correlations between the sparse observations and overcoming the biases caused by this inhomogeneous spatial distribution, with some regions having few records (e.g., interior of continents).

In this paper, we develop a novel method that enables us to define more plausible boundary conditions, using a low-dimensional representation of the full boundary condition input. This method exploits physical spatio-temporal structure in existing low-resolution climate model ensembles, while retaining enough flexibility to overcome the biases in these models. By varying a small number of coefficients that control the basis representation, we efficiently generate past climates that are consistent with observations. We can then better explore the uncertainty in the ice sheet model output, by jointly varying the ice sheet parameters and the boundary conditions, then searching for combinations that lead to output consistent with ice sheet observations using emulation and calibration (history matching).

Section 2 provides an overview of the Glimmer ice sheet model we use to simulate the North American ice sheet, and of the types of paleo-data that are available. Section 3 outlines the statistical methods from the UQ literature we use to analyze the output of Glimmer. Section 4 gives our framework for modeling and calibrating boundary conditions, with the boundary condition model for Glimmer fitted in section 5. Section 6 provides additional history matching methodology required to compare ice sheet thickness with binary observations. Section 7 emulates model output over three waves of history matching and shows how boundary condition uncertainty for Glimmer is reduced, with discussion in section 8.

2. The Glimmer ice sheet model and paleo observations.

Glimmer is a fast three-dimensional ice sheet model (Rutt et al., 2009) which has been used to simulate the past evolution of the North American ice sheet (Gregoire et al., 2012) and the past and future evolution of the Greenland ice sheet (Lunt et al., 2008; McNeall et al., 2013; Stone et al., 2010). Glimmer is a 3D thermomechanical model that simulates the flow of ice based on the

Table 1
The input parameters for Glimmer, and their prior ranges.

Input	Range
Flow factor, f	$[1, 10]$
Geothermal heat flux, G	$[0.02, 0.09] W m^{-2}$
Basal sliding, B_sed	$[0.5, 20] mm\, yr^{-1} Pa^{-1}$
Mantle relaxation time, τ	$[300, 9000]$ years
Positive degree day factor (snow), pdd_s	$[0.002, 0.006] m\, d^{-1} K^{-1}$
Positive degree day factor (ice), pdd_i	$[0.007, 0.02] m\, d^{-1} K^{-1}$
Lapse rate, LR	$[4, 8.2] K\, km^{-1}$

so-called Shallow Ice Approximation. This makes it a very fast model particularly suited to simulating large ensembles of the evolution of large continental ice sheets such as the North American ice sheet, over multimillennial time scales. Glimmer includes an isostasy model that simulates postglacial rebound of the solid Earth as the ice sheet shrinks. Here, we model the evolution of the North American ice sheet during the last deglaciation (21,000–7,000 years ago) following the same setup as in Gregoire et al. (2016). The domain is defined by a Cartesian grid covering North America and Greenland at 40 km horizontal resolution (194 x 150 cells) with 10 vertical levels. The model uses a positive degree day (PDD) scheme to simulate ice sheet surface mass balance, driven by monthly mean temperature and precipitation fields input every month.

There are numerous uncertain parameters in Glimmer that control the flow and melt of the ice sheet. Previous work has identified 7 parameters (Table 1), here referred to as $\mathbf{x}$, that have the strongest effect on the output of the model, controlling aspects of the ice sheet such as basal sliding and lapse rate (Hebeler et al., 2008; Gregoire, 2010; Gregoire et al., 2016). The simulations are run in two phases: a spin-up phase to build up the initial ice sheet through the last glacial cycle, and the main deglaciation phase following the method of Gregoire et al. (2016). The spin-up phase starts at 120 ka (the last interglacial period) with present-day ice sheets (i.e., no ice in North America), and we build up the ice sheet through the last glacial cycle by using a widely used method of climate-index forcing. This method consists of interpolating the climate fields between a warm interglacial climate (represented by present day observations) and a cold glacial climate at 21 ka using a temporally varying temperature index based on a Greenland ice core record (full details of the method can be found in Gregoire et al. (2016)). During the deglaciation phase, our period of interest, the simulation starts from the spin-up simulation at 21 ka and runs until 6 ka, when North America was ice free. Over this period, the climate boundary condition input is simply the monthly temperature and precipitation fields, typically from a transient run of a low-resolution global climate model (GCM), e.g., FAMOUS (Smith et al., 2008). This method allows us to produce a range of initial conditions that are compatible with the input climate at 21 ka and with the model parameters. Figure 1 gives an example of the spatial output of Glimmer, at 21 ka, with darker blue representing a thicker ice sheet, and the observed extent of the ice sheet (the ice margin) given by the red line (described in section 2.2).

Low resolution GCMs, such as FAMOUS, are the only type of climate models that have the necessary complexity and speed to simulate the evolution of ice sheets over millennia (Gregory

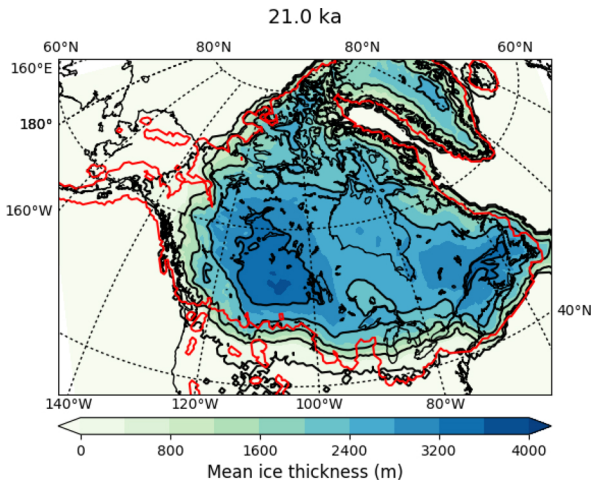

Figure 1. *An example of the ice thickness output given by Glimmer at the start of the deglaciation (21 ka). The red line shows the observed extent of the ice sheet.*

et al., 2012; Gregoire et al., 2015) and that are fast enough for uncertainty quantification. FAMOUS has been successfully used to determine the cause of the largest abrupt sea level rise of the geological past (Gregoire et al., 2012, 2016), to better understand ice sheet and ocean processes involved in abrupt climate changes (Smith and Gregory, 2009; Hawkins et al., 2011; Roberts et al., 2014; Jackson et al., 2017), and to simulate ocean biogeochemistry (Williams et al., 2013, 2014). Thus the usefulness and utility of the FAMOUS climate models has already been demonstrated through a great variety of studies. Performing this work with high-resolution regional climate models specifically developed for studying ice sheets is currently computationally impossible for uncertainty quantification in the past.

There are several problems with simply using GCM output as the Glimmer boundary condition. There are only a limited number of modeling groups who have been able to simulate the last deglaciation with climate models. Two groups have used intermediate complexity climate models of fairly high resolution (Menviel et al., 2011; Roche et al., 2011), while two other groups have used GCMs, with simulations taking from 4 months to 2 years to complete (Liu et al., 2009; Gregoire et al., 2012). The GCMs have their own input parameters, which have not been tuned with temperatures through the deglaciation. For example, the simulations carried out with the CCSM3 model (Liu et al., 2009) are tuned to modern-day climate, while simulations with the FAMOUS GCM were tuned to simulate both modern and glacial climates, but not the period in between (Gregoire et al., 2011).

Therefore, the boundary conditions currently used for Glimmer may not adequately match geological observations, nor does their spread capture uncertainty in those boundary conditions. There may be biases, both locally and globally, spatially and temporally, over the required paleo time scales, as paleo observations of temperature have not been used in the tuning process, and hence important warming and cooling periods may not be captured by the GCM. Important locations (e.g., where the ice sheet covers) may not be modeled well (likely been tuned to global metrics on paleo time scales, if at all). Such inadequacies may

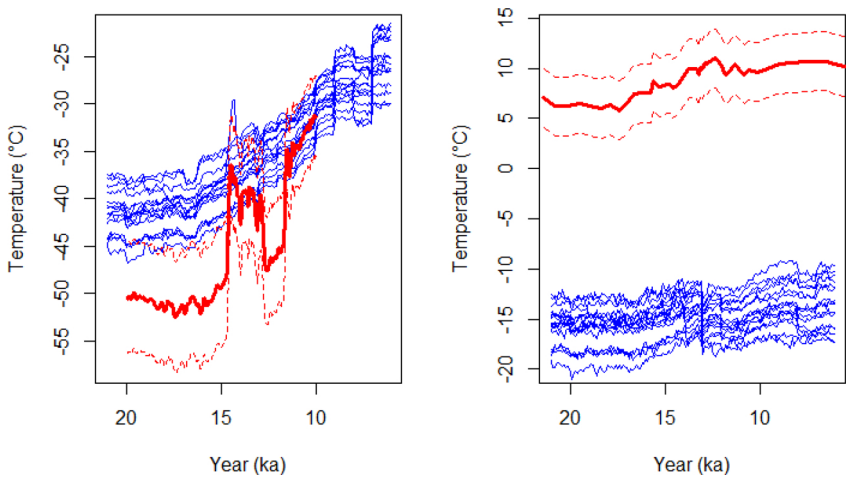

Figure 2. *The time series of temperature for each of the* 16 *ensemble members (blue) and the observed temperatures (red), averaged spatially for Greenland (left) and Alaska (right), with* 95 *% observation error uncertainty given by the dotted lines. Some of the abrupt changes in temperature seen for every ensemble member are due to the climate model's ice sheet topography being updated every* 1,000 *years.*

result in inaccurate simulations of the deglaciation when used to force Glimmer, leading to difficulties in studying the effect that rapid warming and cooling events have on ice sheets.

To illustrate this difference between the model and geological observations, we consider an ensemble of 16 GCM runs that have been used as boundary conditions for Glimmer: 15 FAMOUS simulations (Gregoire et al., 2011), and 1 TRACE simulation (Liu et al., 2009). Each ensemble member consists of monthly temperature and precipitation fields from 21 ka onwards, with FAMOUS on a 48×37 spatial grid ($7.5° \times 3.5°$), and TRACE a 96×48 grid ($3.75° \times 3.75°$). Figure 2 compares the ensemble temperatures (averaged across the nearest grid boxes) to observations in Greenland and Alaska (two of the closest observed spatial locations to the North American ice sheet). This plot shows that the GCM output is biased away from the observations in important locations. In Greenland, the ensemble is too warm at the start of the deglaciation, with the majority of runs not falling within the 95% error bounds on the observations. The runs generally fail to replicate the rapid warming around 14.7 ka, and none capture the rapid cooling around 13 ka. In Alaska, the temporal pattern is correct, but the ensemble is 15°C cooler than the observations, a difference of 10 standard deviations.

2.1. Temperature observations. In addition to the temperature observations for Greenland and Alaska, we have observations at other spatio-temporal locations, but none over the North American ice sheet (Shakun et al., 2012; Buizert et al., 2014).

Figure 3 shows 20 spatial locations where temperature observations are available. These observations are reconstructed from various different sources, for example, ice cores (Greenland), Mg/Ca, $U_{37}^{k'}$, and microfossils, and have varying degrees of uncertainty (see Shakun et al. (2012)): e.g., for microfossils, 1 standard deviation is given as 1.5°C, as in Alaska, while 1 standard deviation for the Greenland ice core varies between 2.1 and 3.2 through time. Shakun et al. (2012) give additional spatial locations, but we use this subset, ignoring sources considered less reliable and those in close proximity to chosen observations. Our subset con-

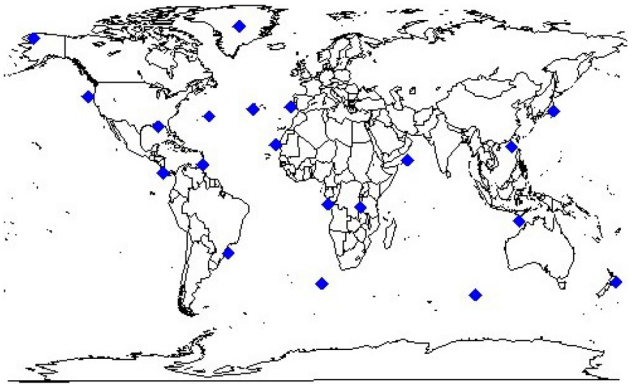

Figure 3. *Map showing the locations of observed temperatures.*

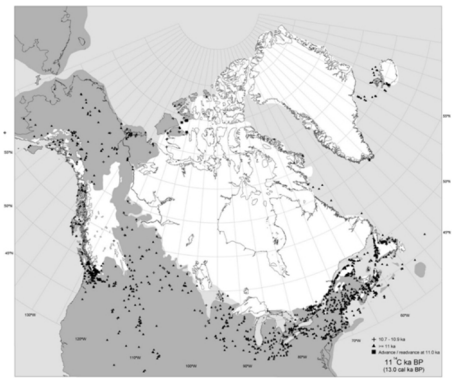

Figure 4. *The ice extent at* 13 *ka.*

tains those observations closest to North America, while also having well-spaced points around the world.

The geological records are also irregular in time: in the 20 locations in Figure 3, the number of time points with an observation varies from only 26 to over 100 (e.g., the Alaska observations plotted in Figure 2 are based on 43 observed time points). These are infrequent compared to the number of time points in the boundary condition, $15{,}000 \times 12 = 180{,}000$. Furthermore, the observations have a temporal uncertainty, with this generally increasing as we move further back in time, where 1 standard deviation can be 400 years or more.

2.2. Ice sheet observations. Glimmer gives ice thickness (in meters) as an output. The most robust constraints on past ice sheet evolution are reconstructions of the extent through time. For the North American ice sheet, the latest reconstruction is from Dyke (2004), an example of which is shown in Figure 4. This dataset provides estimates of ice margin position at quasi-regular 500–1000 year intervals through the deglaciation, based on expert interpretation of compilations of geological data that date the presence or absence of ice. We therefore have a set of maps showing the presence or absence of ice in each gridbox at given times throughout the deglaciation.

We also have estimates of the volume of the North American ice sheet through the deglaciation, inferred from ensembles of ice sheet models constrained with data on volume and extent (Tarasov et al., 2012). We use the spatio-temporal patterns of ice extent and the time series of ice volume to calibrate Glimmer within a history matching framework.

3. Emulation and history matching. In addition to having a high-dimensional boundary condition as an input, Glimmer's ice thickness output, f, is spatio-temporal, so we may require multivariate emulation methods to model $f(\mathbf{x})$. We vectorize the output so that $f(\cdot)$ is a vector of length ℓ, the number of outputs, and define an ensemble as $\mathbf{F} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$, where design $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ consists of input parameter settings from space $\mathcal{X}$.

Spatial output can be emulated via a low-dimensional basis representation, often the basis given by the singular value decomposition (SVD) of the (centered) model output (Higdon et al., 2008; Wilkinson, 2010; Sexton et al., 2011; Chang et al., 2014; Salter et al., 2019):

$$\mathbf{F}_{\boldsymbol{\mu}}^T = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^T, \tag{3.1}$$

where the ith column of $\mathbf{F}_{\boldsymbol{\mu}}$ is given by $f(\mathbf{x}_i) - \boldsymbol{\mu}$ for ensemble mean $\boldsymbol{\mu}$. The columns of $\boldsymbol{\Gamma}$ form a basis for $\mathbf{F}_{\boldsymbol{\mu}}$, and we project output onto this basis via

$$\boldsymbol{\omega}(\mathbf{x}) = (\boldsymbol{\Gamma}^T\mathbf{W}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{W}^{-1}(f(\mathbf{x}) - \boldsymbol{\mu}), \tag{3.2}$$

for positive definite variance matrix $\mathbf{W}$. The original field is reconstructed as

$$f(\mathbf{x}) = \boldsymbol{\Gamma}\boldsymbol{\omega}(\mathbf{x}) + \boldsymbol{\mu} + \boldsymbol{\epsilon}(\mathbf{x}), \tag{3.3}$$

with error $\boldsymbol{\epsilon}(\mathbf{x})$ ($= \mathbf{0}$ for $\mathbf{x} \in \mathbf{X}$) and $\boldsymbol{\omega}(\mathbf{x}) = (\omega_1(\mathbf{x}), \ldots, \omega_n(\mathbf{x}))^T$.

The first $q \ll n$ basis vectors explain the majority of the variability in $\mathbf{F}_{\boldsymbol{\mu}}$; hence $\boldsymbol{\Gamma}$ is truncated to give basis $\boldsymbol{\Gamma}_q = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q)$, where $\boldsymbol{\gamma}_i$ is the ith column of $\boldsymbol{\Gamma}$. Univariate emulators (commonly Gaussian process-based) for the q coefficients given by projection onto $\boldsymbol{\Gamma}_q$ are constructed:

$$\omega_i(\mathbf{x}) \sim \mathrm{GP}(m_i(\mathbf{x}), K_i(\mathbf{x}, \mathbf{x})), \quad i = 1, \ldots, q, \tag{3.4}$$

for mean function $m_i(\cdot)$ and covariance function $K_i(\cdot, \cdot)$ (typically the squared exponential).

When emulating and calibrating a spatial field, we select an optimal basis via a rotation of the SVD (or weighted SVD) basis (Salter et al., 2019). This allows patterns from the observations, which may not be present in the truncated basis $\boldsymbol{\Gamma}_q$, yet may appear as a linear combination of several low-eigenvalue vectors of $\boldsymbol{\Gamma}$, to be incorporated into the calibration basis, ensuring that the correct directions of output space can be searched (i.e., so that the basis choice doesn't guarantee that we conclude that the model cannot represent the observations). The quality of a basis can be assessed via

$$\begin{aligned}\mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}, \mathbf{z}) &= \|\mathbf{z} - \mathbf{r}(\mathbf{z})\|_{\mathbf{W}} = (\mathbf{z} - \mathbf{r}(\boldsymbol{\Gamma}, \mathbf{z}))^T\mathbf{W}^{-1}(\mathbf{z} - \mathbf{r}(\boldsymbol{\Gamma}, \mathbf{z})),\\ \mathbf{r}(\boldsymbol{\Gamma}, \mathbf{z}) &= \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{W}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{W}^{-1}\mathbf{z},\end{aligned}$$

for positive definite $\mathbf{W}$, basis $\boldsymbol{\Gamma}$, and "reconstruction error" $\mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}, \mathbf{z})$, representing the difference between the observations, $\mathbf{z}$, and their reconstruction, $\mathbf{r}(\boldsymbol{\Gamma}, \mathbf{z})$, given by the subspace defined by $\boldsymbol{\Gamma}$.

3.1. History matching. History matching is a method for calibrating the input parameters of a computer model, removing regions of parameter space that are unlikely to lead to output consistent with observations (Craig et al., 1996; Vernon et al., 2010). Given observations, $\mathbf{z}$, of a physical system, $\mathbf{y}$, represented by computer model $f(\cdot)$, we assume that

$$\mathbf{y} = f(\mathbf{x}^*) + \boldsymbol{\eta}, \quad \mathbf{z} = \mathbf{y} + \mathbf{e}, \tag{3.5}$$

where $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{e})$ is the observation error, $\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$ is the discrepancy between the "best" input of the computer model, $\mathbf{x}^*$, and the true system, and the terms in (3.5) are independent (Kennedy and O'Hagan, 2001). Alternative models for calibration and discrepancy have been proposed in, e.g., Tuo and Wu (2016); Plumlee (2017); Gu and Wang (2018).

Given $\boldsymbol{\Sigma}_\mathbf{e}$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and an emulator for $f(\mathbf{x})$ with expectation $\mathrm{E}[f(\mathbf{x})]$ and variance $\mathrm{Var}[f(\mathbf{x})]$, the implausibility of $\mathbf{x}$ is

$$\mathcal{I}(\mathbf{x}) = (\mathbf{z} - \mathrm{E}[f(\mathbf{x})])^T (\mathrm{Var}[f(\mathbf{x})] + \boldsymbol{\Sigma}_\mathbf{e} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}})^{-1} (\mathbf{z} - \mathrm{E}[f(\mathbf{x})]). \tag{3.6}$$

The space of not implausible runs ("not ruled out yet" (NROY) space) is

$$\mathcal{X}_{NROY} = \{\mathbf{x} \in \mathcal{X} | \mathcal{I}(\mathbf{x}) < b\}, \tag{3.7}$$

for a bound, b, used to rule out implausible settings of $\mathbf{x}$. For univariate $f(\cdot)$, Pukelsheim's 3-sigma rule (Pukelsheim, 1994) is used so that $b = 3^2$; for multivariate $f(\cdot)$, b is often set conservatively as the 99.5% value of the chi-squared distribution with ℓ (the number of outputs) degrees of freedom (Vernon et al., 2010). By designing new ensembles within NROY space, history matching can be performed iteratively over multiple waves ("refocusing") (Vernon et al., 2010; Salter and Williamson, 2016; Williamson et al., 2017). For multiple emulated outputs, the "jth maximum implausibility measure" can be used (Craig et al., 1997):

$$\mathcal{I}_{jM}(\cdot) = \max_i(\{\mathcal{I}_i(\cdot)\} \backslash \{\mathcal{I}_M(\cdot), \mathcal{I}_{2M}(\cdot), \ldots, \mathcal{I}_{(j-1)M}(\cdot)\}), \; \mathcal{I}_M(\cdot) = \max_i \mathcal{I}_i(\cdot), \tag{3.8}$$

i.e., $\mathcal{I}_{jM}(\mathbf{x})$ is the jth highest implausibility in the set $\{\mathcal{I}_i(\mathbf{x})\}$.

The output of Glimmer is ice thickness, which is not directly comparable to the binary observations of ice extent; hence we require an extension to the standard history matching methodology (section 6). Due to the high dimension of the boundary condition input, we cannot simply include it as part of $\mathbf{x}$ and apply the standard techniques described above.

4. A framework for calibrating boundary conditions. We now introduce our framework for modeling and calibrating spatio-temporal boundary conditions with sparse observations, as in paleoclimate problems. When observations are dense, it is possible to fit a spatio-temporal Gaussian Markov random field to the observations using INLA (Liu and Guillas, 2017), varying the parameters of this process to generate boundary conditions. However, when observations are sparse, such a process will revert towards its prior mean as the distance increases from an observed point. The majority of our observations are ocean based, so that a Gaussian process model may also be biased across North America where the ice sheet is. Instead, our approach involves defining a structured mean function that inherits physical patterns from GCM output, conditioned on geological temperature observations.

In general, we extend the formulation of the previous section so that the boundary condition is explicitly accounted for, and we are calibrating

$$\mathbf{z} = f(\mathbf{x}^*, \mathbf{T}(\mathbf{c}^*)) + \mathbf{e} + \boldsymbol{\eta}, \tag{4.1}$$

where the set of inputs in (3.5) has been augmented with boundary condition $\mathbf{T}(\mathbf{c})$, required for forcing $f(\cdot)$. $\mathbf{T}(\cdot)$ might represent a temperature field, itself parametrized by a vector of parameters $\mathbf{c}$ in space $\mathcal{C}$. In this way, our set of calibration parameters is $(\mathbf{x}, \mathbf{c}) \in \mathcal{X} \times \mathcal{C}$, and we proceed by emulating $f(\mathbf{x}, \mathbf{T}(\mathbf{c}))$ and using the emulator to jointly calibrate the model parameters, $\mathbf{x}$, and boundary condition, $\mathbf{T}(\mathbf{c})$, to observations, with NROY space of the form

$$(\mathcal{X} \times \mathcal{C})_{NROY} = \{(\mathbf{x}, \mathbf{c}) \in \mathcal{X} \times \mathcal{C} | \mathcal{I}(\mathbf{x}, \mathbf{c}) < b\},$$

(we will define $\mathcal{I}(\mathbf{x}, \mathbf{c})$ for binary observations in section 6).

The problem therefore becomes one of establishing the mapping $\mathbf{T}(\mathbf{c})$. Generally for Glimmer, the boundary condition has been the output of a GCM, runs of which have their own input parameters which could be interpreted as $\mathbf{c}$ in (4.1). However, while it is theoretically possible to couple the models in such a way and calibrate the model parameters jointly, in reality climate models are extremely expensive to run, particularly across the temporal domain in our application. The required $\mathbf{T}(\mathbf{c})$ will not be available at enough choices of $\mathbf{c}$ (for example, there were only 15 runs of GCM FAMOUS on this time scale at the time of writing), while it is also not clear whether such models can give a plausible match to geological observations.

Instead, we model the high-dimensional $\mathbf{T}(\mathbf{c})$, with the requirements that it is (a) fast to evaluate, (b) a plausible representation of geological observations, and (c) parametrized by a small number of coefficients $\mathbf{c}$. In doing so, we aim to ensure that the expensive ice sheet model is forced with temperature fields that may be realistic, while being able to sample and construct such fields efficiently and flexibly. Restricting the size of $\mathbf{c}$ has the goal of minimizing the number of calibration parameters and makes emulation and calibration more possible than if the emulators had a full $\mathbf{T}$ as an input.

4.1. Modeling the boundary condition. Due to the generally high dimension of boundary conditions, we exploit dimension reduction to construct the parametrized $\mathbf{T}(\mathbf{c})$. Let spatio-temporal boundary condition $\mathbf{T}$ have dimension $\ell_s \ell_t$, where ℓ_s and ℓ_t are the number of spatial and temporal dimensions, respectively. We model $\mathbf{T}$ as

$$\mathbf{T} \mid \mathbf{c} \sim \text{MVN}(h(\mathbf{c}), \boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_t), \qquad h(\mathbf{c}) = \boldsymbol{\mu} + \sum_{j=1}^{n_t} c_j^t \mathbf{t}_j + \sum_{j=1}^{n_s} c_j^s \mathbf{s}_j, \tag{4.2}$$

where $h(\mathbf{c})$ represents our parametrization of $\mathbf{T}$, dependent on the spatio-temporal field, $\boldsymbol{\mu}$, sets of basis vectors $\mathbf{t} = (\mathbf{t}_1, \ldots, \mathbf{t}_{n_t})$ and $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_{n_s})$, controlled by coefficient vector $\mathbf{c} = (\mathbf{c}^t, \mathbf{c}^s) \in \mathcal{C} \subset \mathbb{R}^{n_t + n_s}$, and $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_t$ are spatial and temporal variance matrices. We describe our method for generating $\mathbf{t}, \mathbf{s}$ and finding $\boldsymbol{\Sigma}_t, \boldsymbol{\Sigma}_s$ in section 4.2.

By jointly varying $\mathbf{c}$ and inputs $\mathbf{x}$, and running Glimmer at these choices, boundary condition uncertainty can be explored, as with ordinary calibration exercises. To fit $\mathbf{t}$ and $\mathbf{s}$, and to select $\mathcal{C}$, etc., we have observations of $\mathbf{T}$ at sparse spatio-temporal locations, modeled as

$$\mathbf{z}_T \sim \mathrm{MVN}(\mathbf{T}, \boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_{t'}), \tag{4.3}$$

for a different temporal variance matrix $\boldsymbol{\Sigma}_{t'}$ (for computational reasons, see section 4.3).

4.2. Fitting the boundary condition model. This section addresses fitting the model from (4.2), finding $\mathbf{T} \mid \mathbf{c}$ by defining spatio-temporal $(\boldsymbol{\mu}, \mathbf{t}, \mathbf{s})$ such that $\mathbf{T} \mid \mathbf{c}$ captures uncertainty in the sparse observations $\mathbf{z}_T$ and is suitable for use in calibration (4.1). So that $\mathbf{T}$ is physically plausible, we exploit an ensemble of GCM model output and force it towards observational records, with spatial and temporal patterns preserved in the representation of $\mathbf{T}$, combined with out-of-sample checks to avoid overfitting.

For the GCM ensemble, we use the $n = 16$ FAMOUS and TRACE GCM runs discussed in section 2, denoted by $\boldsymbol{\mathcal{T}} = (\tau_1, \ldots, \tau_n)$ (each run depends on its own set of input parameters, omitted for clarity as we only require the output fields). The runs in $\boldsymbol{\mathcal{T}}$ are generally not consistent with temperature records (Figure 2), and when these records lie outside the span of the climate ensemble (likely where $n \ll \ell_s \ell_t$) setting $\mathbf{t}$ equal to $\boldsymbol{\mathcal{T}}$ itself, or to its leading eigenvectors, may not allow temperature records to be adequately captured. On the other hand, a huge number of important physical constraints (amounting to spatio-temporal correlations) are present in the GCM output, so basing our $\mathbf{T}$ on a decomposition of $\boldsymbol{\mathcal{T}}$ is attractive.

We therefore incorporate steps to better exploit the information within $\boldsymbol{\mathcal{T}}$ given the observations $\mathbf{z}_T$, finding $(\boldsymbol{\mu}, \mathbf{t}, \mathbf{s})$ as follows:

1. Fix $\boldsymbol{\mu} = mean(\boldsymbol{\mathcal{T}})$.
2. Select $\mathbf{t}$ using a temporal decomposition at key observed location(s) (section 4.2.1).
3. Select $\mathbf{s}$ as a time-invariant spatial field to correct for remaining/induced biases (section 4.2.2).

This approach enables us to first ensure that, where possible, we extract an accurate temporal pattern from $\boldsymbol{\mathcal{T}}$ (e.g., the distinctive pattern of rapid changes in Greenland), described via set $\mathbf{t}$. Given this, the spatially derived vectors $\mathbf{s}$ are selected by trading off the ability of the model to capture $\mathbf{z}_T$ in other spatial locations. The resulting model should be able to represent observations more accurately than $\boldsymbol{\mathcal{T}}$, while accounting for uncertainty: producing interpretable basis vectors, and giving a flexible set of possible boundary conditions by varying $\mathbf{c}$ appropriately.

4.2.1. Fitting $\mathbf{t}$. We aim to find a linear combination of $\boldsymbol{\mathcal{T}}_{\boldsymbol{\mu}} = \boldsymbol{\mathcal{T}} - \boldsymbol{\mu}$ such that the observed time series at chosen spatial locations is reproducible by the basis $\mathbf{t}$ for $n_t < n$, where this may not be true for individual members of $\boldsymbol{\mathcal{T}}$. By taking linear combinations, we retain some physicality and smoothness from the climate model output, while the coefficients controlling the vectors in $\mathbf{t}$ may be interpretable. Overall, we find $\mathbf{t}$ such that

(i) $\boldsymbol{\mathcal{T}}_{\boldsymbol{\mu}}^T = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^T$;
(ii) $\boldsymbol{\Lambda}^* = \mathrm{argmin}_{\boldsymbol{\Lambda}} \mathcal{R}_{\boldsymbol{\Sigma}_s \otimes \boldsymbol{\Sigma}_{t'}}(\boldsymbol{\Gamma}\boldsymbol{\Lambda}, \mathbf{z}_T - \boldsymbol{\mu}), \quad \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T = \mathbb{I}_n, \quad \boldsymbol{\Lambda} \in \mathbb{R}^{n \times n_t}$;
(iii) $\mathbf{t}_j = \boldsymbol{\mathcal{T}}_{\boldsymbol{\mu}}[\mathbf{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}^*]_{\cdot j}, \quad j = 1, \ldots, n_t$.

Ignoring steps (ii) and (iii), and setting $\mathbf{t}$ as the first n_t vectors of $\boldsymbol{\Gamma}$, there is no guarantee of consistency with $\mathbf{z}_T$, and hence the optimization step of (ii) and (iii) is generally required.

In this application, the main issues with this approach are the sparsity of $\mathbf{z}_T$ and the dimension of the full field. To overcome these, we restrict the above optimization to a spatial

location (or locations) where observations are available, finding a rotation matrix $\mathbf{\Lambda}^*$ and hence basis $\mathbf{t}$ such that the error in the $\mathbf{W} = \mathbf{\Sigma}_s \otimes \mathbf{\Sigma}_{t'}$ norm is minimized at this observed location(s) alone. If we restrict $\boldsymbol{\mathcal{T}}$ to a single observed location, then the eigenvectors are ℓ_t-vectors rather than $\ell_s\ell_t$. The resulting optimized vectors remain a linear combination of $\boldsymbol{\mathcal{T}}$ and can be extrapolated to the full spatio-temporal domain to give the required full basis vectors $\mathbf{t}$ (see section **SM1.1** in the supplementary material).

The quality of the fit of $\mathbf{t}$ to $\mathbf{z}_T$ is restricted by the subspace defined by the full GCM ensemble; however, the optimization step has the benefit that only n_t (usually 2 or 3) coefficients need to be specified instead of n, and the possible linear combinations are naturally restricted to the subspace of the ensemble that best matches observations at the chosen spatial location(s) (thanks to $\mathbf{\Lambda}^*$).

4.2.2. Fitting s. Selecting $\mathbf{t}$ based on a subset of locations may restrict other regions of the spatial domain of $h(\mathbf{c})$ to subspaces that are inconsistent with other available observations, or our physical judgments. To remove some of these inadequacies, we add time-invariant spatial vectors $\mathbf{s}$ (we assume that we have already captured the majority of temporal variability through $\mathbf{t}$).

Conditional on the previous model components $(\boldsymbol{\mu}, \mathbf{t})$, we construct spatial quantities (by averaging temporally) that contain the unexplained ensemble variability, $\boldsymbol{\mathcal{T}}_\epsilon$, and remaining biases to the observations, $\mathbf{z}_\epsilon$:

(i) $\boldsymbol{\mathcal{T}}_\epsilon = \frac{1}{\ell_t}\sum_{j=1}^{\ell_t}(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mu}} - \mathbf{r}(\mathbf{t}, \boldsymbol{\mathcal{T}}_{\boldsymbol{\mu}}))\Big|_{t=j}, \quad \boldsymbol{\mathcal{T}}_\epsilon \in \mathbb{R}^{\ell_s \times n},$

(ii) $\mathbf{z}_\epsilon = \frac{1}{\ell_t}\sum_{j=1}^{\ell_t}(\mathbf{z}_T - \boldsymbol{\mu} - \mathbf{r}(\mathbf{t}, \mathbf{z}_T))\Big|_{t=j}, \quad \mathbf{z}_\epsilon \in \mathbb{R}^{\ell_s},$

where $\mathbf{r}(\mathbf{t}, \boldsymbol{\mathcal{T}}_{\boldsymbol{\mu}}), \mathbf{r}(\mathbf{t}, \mathbf{z}_T)$ are the reconstructions (with basis $\mathbf{t}$) of the centered climate ensemble, $\boldsymbol{\mathcal{T}}_{\boldsymbol{\mu}}$, and observations, $\mathbf{z}_T$, respectively, and $\cdot|_{t=j}$ denotes the restriction of vectorized spatio-temporal quantities to time $t = j$ (similarly, $\cdot|_{s=i}$ restricts to the ith spatial location).

We now find $\mathbf{s}$ analogously to $\mathbf{t}$, finding a linear combination such that the resulting fields minimize the difference between observations and the basis reconstruction:

(iii) $\boldsymbol{\mathcal{T}}_\epsilon^T = \mathbf{U}_\epsilon \mathbf{\Sigma}_\epsilon \mathbf{\Gamma}_\epsilon^T$;

(iv) $\mathbf{\Psi}^* = \operatorname{argmin}_{\mathbf{\Psi}} \mathcal{R}_{\mathbf{\Sigma}_s}(\mathbf{\Gamma}_\epsilon \mathbf{\Psi}, \mathbf{z}_\epsilon), \quad \mathbf{\Psi}\mathbf{\Psi}^T = \mathbb{I}_n, \quad \mathbf{\Psi} \in \mathbb{R}^{n \times n_s}$;

(v) $\mathbf{s}_j = \mathbf{1}_{\ell_t} \otimes \boldsymbol{\mathcal{T}}_\epsilon[\mathbf{\Psi}^*]_{\cdot j}, \quad j = 1, \ldots, n_s,$

with the optimized spatial fields repeated across time to provide spatial corrections via $\mathbf{1}_{\ell_t}$, an ℓ_t-vector of 1's.

As with the choice of $\mathbf{t}$, we have to deal with sparsity, and hence the optimization in (iv) is performed at a subset of spatial locations rather than the full $\mathbf{z}_\epsilon$, with left out observed locations reserved for validation checks to protect against overfitting. Section **SM1.2** provides the restriction to a subset of locations explicitly.

4.2.3. Variance matrices. In (4.2) and (4.3), we have variance matrices $\mathbf{\Sigma}_s, \mathbf{\Sigma}_t$, and $\mathbf{\Sigma}_{t'}$. The common spatial variance $\mathbf{\Sigma}_s$ is set using the temperature output of a different GCM (CanAM4, von Salzen et al. (2013)). For temporal variance $\mathbf{\Sigma}_t$, we impose correlations between close time points using a squared exponential covariance function. So that the observation error is uncorrelated in time, we set

$$\boldsymbol{\Sigma}_{t'} = \sigma_t^2 \mathbb{I}_{\ell_t},$$

with a fixed variance multiplier σ_t^2 over time for tractability. The different types of reconstructions of past temperature have varying errors (discussed in section 2.1); hence we set σ_t^2 on the order of the largest error across the records.

4.3. Simulating boundary conditions. Using $\mathbf{t}$ and $\mathbf{s}$, (4.2) gives a model for $\mathbf{T} \mid \mathbf{c}$ that we can use to generate a boundary condition. To do so, we sample a value of $\mathbf{c}$ and condition on observations $\mathbf{z}_T$. This is a sparsely observed vector, and we denote observed and missing entries of this vector by $\mathbf{z}_{T,obs}$ and $\mathbf{z}_{T,miss}$, respectively. By first integrating out missing observations $\mathbf{z}_{T,miss}$,

$$\begin{aligned} \pi(\mathbf{T} \mid \mathbf{z}_{T,obs}, \mathbf{c}) &= \int \pi(\mathbf{T}, \mathbf{z}_{T,miss} \mid \mathbf{z}_{T,obs}, \mathbf{c}) d\mathbf{z}_{T,miss} \\ &= \int \pi(\mathbf{T} \mid \mathbf{z}_{T,miss}, \mathbf{z}_{T,obs}, \mathbf{c}) \pi(\mathbf{z}_{T,miss} \mid \mathbf{z}_{T,obs}, \mathbf{c}) d\mathbf{z}_{T,miss}, \end{aligned} \tag{4.4}$$

we define the final boundary condition at $\mathbf{c}$ as the expectation of (4.4),

$$\mathbf{T}(\mathbf{c}) = \mathrm{E}[\mathbf{T} \mid \mathbf{z}_T, \mathbf{c}] = h(\mathbf{c}) + vec(\boldsymbol{\Sigma}_t(\boldsymbol{\Sigma}_{t'} + \boldsymbol{\Sigma}_t)^{-1}(\mathbf{z}_T - h(\mathbf{c}))^T),$$

derived in section **SM1**. We could instead sample from (4.4), but we use the expectation here so that deterministic emulators can be built.

Several assumptions are required to ensure that (4.4) is tractable. The Kronecker structure of the variance (equations (4.2), (4.3)) allows efficient inversion of an $\ell_s\ell_t \times \ell_s\ell_t$ matrix, needed for conditioning $\mathbf{T}$ on observations; this is important as we have $\ell_s = 48 \times 37$ and $\ell_t = 15000 \times 12$. The missing entries of the sparse observation vector are integrated over to avoid the loss of the Kronecker structure. Finally, either a common spatial or temporal matrix is required in (4.2) and (4.3) so that the variance matrices can be summed and maintain the Kronecker structure (see section **SM1** for full details of the calculations performed here).

4.4. Prior boundary condition space. Given $(\boldsymbol{\mu}, \mathbf{t}, \mathbf{s})$, a bounded space, $\mathcal{C}$, is defined so that Glimmer can only be run at plausible boundary conditions $\mathbf{T}(\mathbf{c})$. The resulting $\mathcal{C}$ should contain values of $\mathbf{c}$ where $h(\mathbf{c})$ (the prior mean of $\mathbf{T}$) is sufficiently similar to observations, so that we would wish to use such a $\mathbf{T}$ to force a deglaciation simulation with Glimmer.

A natural way to do this is within the history matching framework, although, unlike standard applications of history matching (see, e.g., Vernon et al. (2010); Williamson et al. (2015)), an emulator is not required to compare a generated boundary condition $h(\mathbf{c})$ to observations. By varying $\mathbf{c}$ and evaluating $h(\mathbf{c})$ in (4.2), we calculate an implausibility for spatial location i as

$$\begin{aligned} &\mathcal{I}_i(\mathbf{c}) = ((\mathbf{z} - h(\mathbf{c}))|_{s=i})^T([\boldsymbol{\Sigma}_s]_{ii}\boldsymbol{\Sigma}_{t'})^{-1}((\mathbf{z} - h(\mathbf{c}))|_{s=i}), \quad i \in S_T, \\ &\hat{\mathcal{I}}_i(\mathbf{c}) = 3\mathcal{I}_i(\mathbf{c})/b_i, \quad b_i = \chi^2_{\ell_i, 0.995}, \end{aligned}$$

where there are ℓ_i observed temperatures at location i (where the set S_T contains indices relating observed spatial locations to the full output), and the implausibility is scaled to have a common bound. Spatial dependence is ignored between locations due to their sparsity,

and the error variance at spatial location i is given by $\mathbf{\Sigma}_{t'}$ multiplied by the ith diagonal entry of $\mathbf{\Sigma}_s$. With the sparsity of spatial observations, only a small number ($\ll \ell_s$) of these implausibilities are available.

Sampling $\mathbf{c}$ from a prior $\pi(\mathbf{c})$, the coefficient space is defined as

$$\mathcal{C} = \{\mathbf{c} \sim \pi(\mathbf{c}) \,|\, \hat{\mathcal{I}}_{jM}(\mathbf{c}) < 3\},$$

with $\mathbf{c}$ ruled out if the boundary condition temperature for $\geq j$ of observed locations is implausible, and with $\mathcal{I}_{jM}(\mathbf{c})$ as defined in (3.8). Here we set $\pi(\mathbf{c})$ in an attempt to avoid unnaturally extreme temperatures in unobserved locations, although this judgment could be formalized within the above framework by including an implausibility for global mean/min/max temperature as an additional constraint, with a wider prior $\pi(\mathbf{c})$ then possible.

4.4.1. Calibrating $(\mathbf{x}, \mathbf{c})$. Our aim is to calibrate $(\mathbf{x}, \mathbf{c})$ using ice sheet observations. Having used $\mathbf{z}_{T,obs}$ to derive $\mathcal{C}$, we view $\pi(\mathbf{c} \mid \mathbf{z}_{T,obs}) \sim \text{Unif}(\mathbf{c} \in \mathcal{C})$, and our samples $\mathbf{T} \mid \mathbf{z}_{T,obs}, \mathbf{c}$ (equation (4.4)) are samples from the joint distribution $\pi(\mathbf{T}, \mathbf{c} \mid \mathbf{z}_{T,obs})$:

$$\pi(\mathbf{T}, \mathbf{c} \mid \mathbf{z}_{T,obs}) = \pi(\mathbf{T} \mid \mathbf{z}_{T,obs}, \mathbf{c})\pi(\mathbf{c} \mid \mathbf{z}_{T,obs}).$$

Given ensembles of Glimmer where $\mathbf{x}$ and $\mathbf{c}$ are varied, we further constrain our distribution for the full boundary condition $\mathbf{T}$ by emulating and history matching using observations of the ice sheet. The small number of parameters in $\mathbf{c}$, compared to the dimension of $\mathbf{T}$, makes this a tractable calibration problem, and we can apply methods discussed in section 3.

5. Fitting $\mathbf{T}$ for Glimmer. Glimmer is run by taking 100-year averages of GCM output; hence we average across 100-year periods instead of fitting the model using monthly GCM data, reducing the overall dimension of the boundary condition by a factor of 100, to around 3 million, and giving more consistency with the geological observations (typically on a scale of ≥ 100 years). After initially fitting a single $\mathbf{T}(\mathbf{c})$, we fit separate models to give the flexibility required to capture the variability in the observations throughout the 15,000 years. The temporal domain is split into three intervals (21–15 ka, 15–13 ka, and 13–6 ka), with the short second interval containing the key rapid warming in Greenland, and a separate basis is chosen for each. We address smoothing between time periods and seasonality in sections **SM1.3** and **SM1.4**, but consider these concerns secondary to achieving a model for the overall warming and cooling patterns.

5.1. Basis vectors. We first construct $\mathbf{t}$ as in section 4.2.1, comparing the basis reconstructions with the observations in Greenland (one of the closest locations to the North American ice sheet in the observational data). The rapid warming and cooling between 15 ka and 11 ka is not generally represented in the GCM boundary condition, so that finding $\mathbf{T}(\mathbf{c})$ that can capture this, if possible, is important. We therefore aim to find a model that replicates past Greenland temperatures accurately. For each time period, taking the leading two eigenvectors was sufficient to represent the observed time series, resulting in a total of 6 basis vectors and coefficients at this step. Figure 5 shows the temporal component of $\mathbf{t}$ in Greenland.

Given $\mathbf{t}$, we applied the method from section 4.2.2 to select the spatial corrections, $\mathbf{s}$, again separately for each time period. The resulting spatial fields are shown in Figure 6,

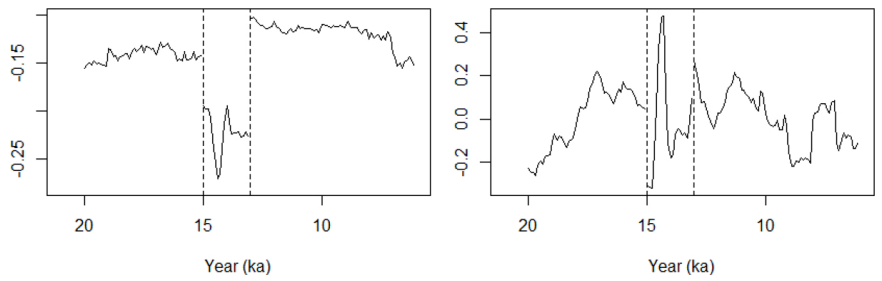

Figure 5. *The basis vectors* $\mathbf{t}$ *in Greenland, from which we can calculate coefficients for the (centered) ensemble members. Left: first basis vector for each time period. Right: second basis vector for each time period.*

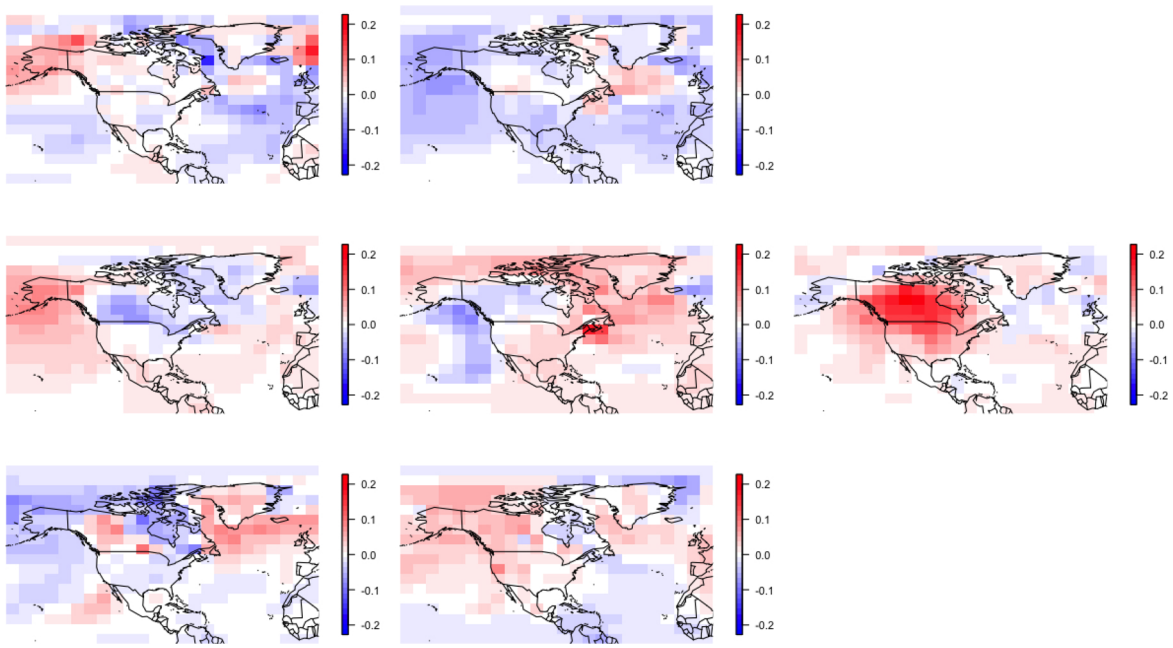

Figure 6. *Spatial basis vectors for the* 3 *time periods (top:* 21 *ka–*15 *ka; middle:* 15 *ka–*13 *ka; bottom:* 13 *ka–*6 *ka), given using the method in section* 4.2.2*, zoomed in on North America. Each vector is reasonably smooth spatially and has some effect over North America.*

with $n_s = 2, 3$ and 2, respectively. Each pattern is reasonably smooth and reduces the biases between the observations and the previously fitted components of the boundary condition model. Initially two vectors were selected for each time period, trading off ability to capture the observations with minimizing the number of coefficients to vary. A third was selected for the second time period to give extra control over the temperature in North America (without this addition, generated temperatures here for this time period were uniformly too cold, with little variability given through changing $\mathbf{c}$). In total, we have 13 coefficients that generate a full spatio-temporal boundary condition.

Adding more basis vectors beyond n_t or $n_s = 2$ may give more accurate models for the boundary condition, but there is a trade-off between accuracy and minimizing the number

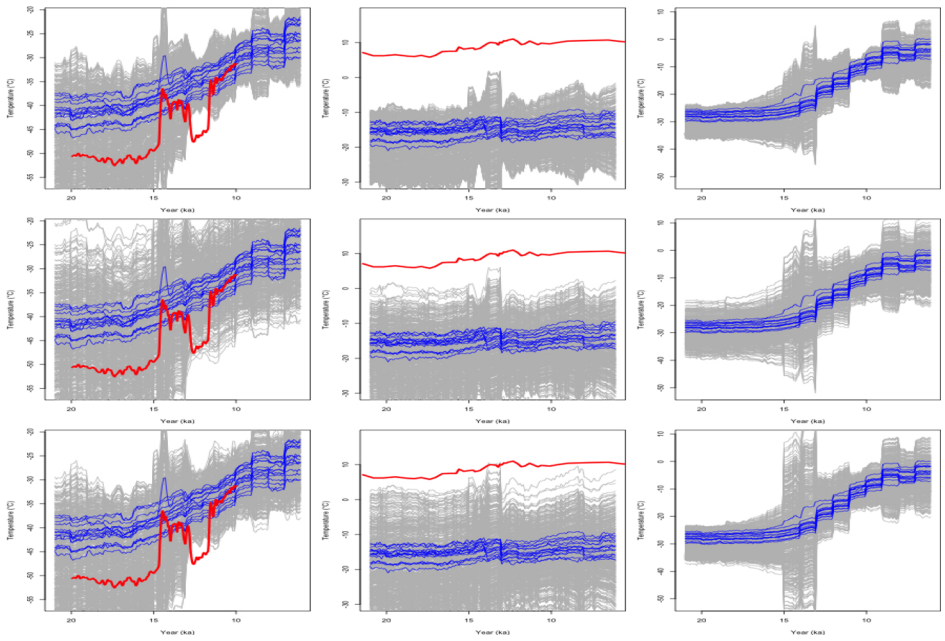

Figure 7. *The observed temperatures (red), climate ensemble (blue), and boundary conditions (gray) given by sampling* 500 *sets of coefficients for different model choices, in Greenland (left), Alaska (middle), and North America (right). Row* 1*: chosen* $\mathbf{t}$; *row* 2*: adding the leading spatial vectors; row* 3*: adding the chosen* $\mathbf{s}$.

of input parameters required. As the dimension of the input space increases, samples from this space become more sparse, and it may be more difficult to identify signals from the various parameters when emulating. The default choice here of 2 gives the combination of keeping the total number of calibration parameters relatively small, while giving a model that is flexible enough to generate a range of boundary conditions that are broadly consistent with observations.

5.2. Validation. Prior to defining coefficient space $\mathcal{C}$ to remove poor choices of $\mathbf{c}$, we explore the range of boundary conditions generated by our representation. The 3 rows of Figure 7 compare the following boundary conditions, with 500 sets of coefficients sampled in each case:

1. Using $(\boldsymbol{\mu}, \mathbf{t})$, all spatial coefficients set to 0.
2. Using $(\boldsymbol{\mu}, \mathbf{t})$, with leading spatial eigenvectors added (no optimisation for $\mathbf{s}$).
3. Using $(\boldsymbol{\mu}, \mathbf{t})$, with optimised $\mathbf{s}$.

In this application, because we set $\mathbf{t}$ using the leading eigenvectors in Greenland rather than optimizing, the first option is roughly equivalent to using the leading spatio-temporal vectors themselves (not exactly the same due to the extrapolation from a single location, but similar as in both cases we are restricted to the subspace spanned by $\boldsymbol{\mathcal{T}}$).

The main benefit of adding $\mathbf{t}$ and generating boundary conditions from coefficients is that, in each of the three plotted locations, the output is more varied than the original ensemble (top row of Figure 7). The observations (red) in Greenland are generally within the range of temperatures given by the 500 samples. In Alaska, however, a large bias between the

observations and the modeled boundary condition remains, demonstrating the need for the second basis selection step. Alternatively, we could use Alaska together with Greenland when selecting $\mathbf{t}$, but this would result in a trade-off in the quality of fit for Greenland. Additionally, the main difference between the GCM and observations in Alaska is an overall temperature bias, rather than in the temporal pattern, which is more suited to being corrected by the spatially chosen basis vectors $\mathbf{s}$.

In the leading eigenvectors of the spatially averaged ensemble, the weightings for locations in Alaska (and North America generally) are relatively low, so that any increase in temperature in Alaska results in much larger changes elsewhere (row 2); e.g., while the boundary conditions in the second row are closer to the Alaska observations than when only $\mathbf{t}$ was used in row 1, this causes higher temperatures in Greenland, which is less consistent with the observations. By applying the optimization step, we find an $\mathbf{s}$ with larger weightings in Alaska relative to Greenland, so that the temperature in Alaska can be increased without causing large biases at other observed locations (row 3).

5.3. Coefficient space. The coefficients used to generate boundary conditions in Figure 7 had not been constrained, so that a wide range of boundary conditions were possible, the majority of which are extremely biased in one of the plotted locations or elsewhere. Using the chosen basis vectors, we defined coefficient space $\mathcal{C}$ as in section 4.4 (shown in Figure **SM1**), sampling 500 values of $(\mathbf{x}, \mathbf{c}) \in \mathcal{X} \times \mathcal{C}$ using a Latin hypercube to give the wave 1 design, $(\mathbf{X} \times \mathbf{C})^{(1)}$. We generated boundary condition $\mathbf{T}$ via (4.4) and ran Glimmer with input $(\mathbf{x}, \mathbf{T})$ to give the wave 1 ensemble, $\mathbf{F}^{(1)}$.

Figure 8 shows the boundary conditions given by $(\mathbf{X} \times \mathbf{C})^{(1)}$ (gray lines) for Greenland, Alaska, and the center of North America. Compared to Figure 7, there is less variability allowed, but by exploring sets of coefficients on our basis vectors, we have found boundary conditions that produce temperatures that are relatively consistent with the observations in Greenland and Alaska, and more accurate than the climate ensemble, with the rapid warming and cooling periods captured.

Initially, our model has colder temperatures than the climate ensemble in North America, but offers a wider range throughout, with the GCM runs contained within this spread from 14 ka onwards. This colder initial temperature may be accurate, given the climate ensemble is known to be too warm in Greenland. In some of the ensemble members, there is an unrealistic downward jump in temperature around 15 ka in North America. We have left these in to establish whether we can rule out such runs when we history match Glimmer's output.

Overall, our model allows the boundary condition to be varied more than in any previous study.

6. History matching to binary data. We have spatial binary observations for the ice extent, $\mathbf{z}^b$, at certain time points, whereas Glimmer's output is ice thickness. Chang et al. (2016) (binary output, binary observations) and Chang et al. (2019) (binary and thickness output, thickness observations) provide extensions to Bayesian calibration problems with binary data, and Sung et al. (2020) provide an alternative calibration approach for univariate binary output. Our application is slightly different, in that we have thickness output to be compared to binary observations, and rather than applying the binary-only method (Chang et al., 2016),

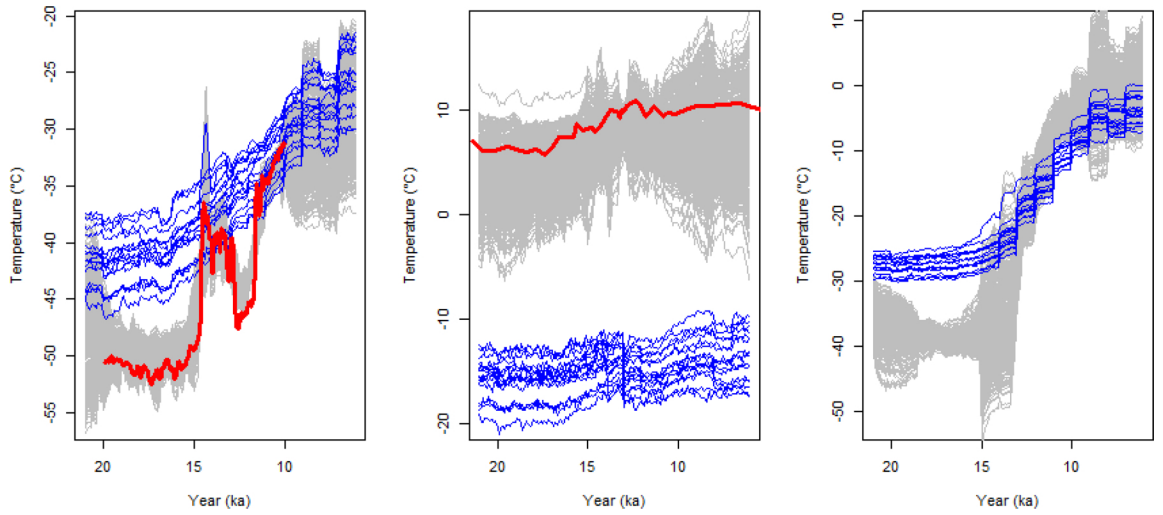

Figure 8. *The observed temperatures (red), climate ensemble (blue), and wave* 1 *(gray) boundary conditions in Greenland (left), Alaska (middle), and North America (right).*

we allow the thickness to be incorporated into the emulation and calibration process, prior to comparison to the observations.

Gregoire et al. (2016) ruled out runs by counting the number of boxes that are misclassified, based on setting a threshold for model thickness that can be treated as "no ice." The number of misclassified boxes was calculated for only the observed runs of Glimmer. Here, we define a similar measure using the emulated thickness, allowing the entire input space to be searched, as is standard in history matching.

6.1. General formulation. We assume that binary observations, $\mathbf{z}^b$, are the binary representation of the modeled latent process, $f(\mathbf{x}, \mathbf{c})$ (in our case, ice thickness), at input $\mathbf{x}^*$ (given observation error $\mathbf{e}$ and discrepancy $\boldsymbol{\eta}$, as in section 3.1):

$$\mathbf{z}^b = \mathbb{1}^b(f(\mathbf{x}^*, \mathbf{c}^*) + \mathbf{e} + \boldsymbol{\eta}),$$

where a general ℓ-dimensional spatial field $f(\mathbf{x}, \mathbf{c})$ is converted to binary via

$$f^b(\mathbf{x}, \mathbf{c}) = \mathbb{1}^b(f(\mathbf{x}, \mathbf{c})), \quad [f^b(\mathbf{x}, \mathbf{c})]_i = \begin{cases} 0 & \text{if} \quad [f(\mathbf{x}, \mathbf{c})]_i \le T^b, \\ 1 & \text{otherwise,} \end{cases} \qquad i = 1, \dots, \ell, \tag{6.1}$$

for a threshold of T^b. Given emulators for the spatial field, we account for the uncertainty in the binary representation at $\mathbf{x}$ by drawing m samples from the emulator posterior and converting each to binary via (6.1). We assess the distance between $\mathbf{z}^b$ and sample j, $f_j^b(\mathbf{x}, \mathbf{c})$, via the number of misclassified grid boxes, as in the usual ice sheet modeling approach (Gregoire et al., 2016). Sample implausibility $\mathcal{I}_j^b(\mathbf{x}, \mathbf{c})$ is given by

$$\mathcal{I}_j^b(\mathbf{x}, \mathbf{c}) = (\mathbf{z}^b - f_j^b(\mathbf{x}, \mathbf{c}))^T(\mathbf{z}^b - f_j^b(\mathbf{x}, \mathbf{c})), \tag{6.2}$$

and letting $\mathcal{I}^b(\mathbf{x},\mathbf{c}) = (\mathcal{I}_1^b(\mathbf{x},\mathbf{c}),\ldots,\mathcal{I}_m^b(\mathbf{x},\mathbf{c}))$, we can define NROY space as

$$(\mathcal{X}\times\mathcal{C})_{NROY} = \{(\mathbf{x},\mathbf{c}) \in \mathcal{X}\times\mathcal{C} | g(\mathcal{I}^b(\mathbf{x},\mathbf{c})) \leq N^b\},$$

where g is a function of the samples of misclassified grid boxes at $(\mathbf{x},\mathbf{c})$ (e.g., the mean, minimum, or 5th percentile, equivalent to requiring at least 5% chance that the true error is $\leq N^b$), so that $(\mathbf{x},\mathbf{c})$ is ruled out if it is unlikely that the number of misclassified grid boxes falls below a threshold, N^b.

How to select N^b and $g(\cdot)$ is problem dependent. For N^b, we use a similar rule of thumb as in Gregoire et al. (2016), where simulations with an extent error of $> 23\%$ were discarded, with this tolerance chosen based on comparing Glimmer output to observations, and judging which are acceptable. We use 25% as a baseline for the acceptable error, although for the regions considered in section 7, this is not always suitable, as it can lead to ruling out either all or none of $\mathcal{X}\times\mathcal{C}$; hence we adjust N^b in these cases, so that we can search for improvements. This may seem a large inconsistency between the model output and observations, but we deliberately selected regions where there is a large mismatch between model output we've seen and the truth, in an attempt to find whether these inconsistencies can be reduced. Ideally, setting N^b would incorporate information about model discrepancy, as we may know that the model can't match in some way.

To select $g(\cdot)$, we use an ensemble of Glimmer output, comparing the number of misclassified boxes for the observed output at $(\mathbf{x},\mathbf{c})$ to the emulator samples in (6.2), and we use $g(\cdot)$ from the above options that is most predictive of the truth across the ensemble.

7. Calibrating Glimmer. Given the fitted boundary condition model for producing plausible temperatures, and the wave 1 ensemble (Figure 8), we now emulate and calibrate the ice sheet parameter and boundary condition inputs of Glimmer. We do not explicitly fit a model for the precipitation boundary condition, instead using monthly mean precipitation from the standard FAMOUS simulation as in Gregoire et al. (2012) for consistency with previous work.

We performed 3 waves of history matching, iteratively ruling out space that was inconsistent with ice sheet observations, using both the volume and extent of the ice sheet as a proxy for the thickness output of Glimmer. Instead of emulating the full spatio-temporal output, at each wave we selected different aspects to be emulated and removed clearly unphysical behaviors (a benefit of history matching is that it does not require an accurate emulator for every output at once). Table 2 summarizes the outputs emulated, and at which waves, as well as the implausibility measures and bounds that were used. For the ice sheet volumes, the error variances were estimated from Tarasov et al. (2012).

For the volume at a single time point, the emulator was a stationary Gaussian process. We spatially emulate the thickness by calculating the SVD basis of the centered ensemble (as in (3.1)), finding an optimal rotation if required, and then emulating the coefficients on this basis ((3.2) and (3.4)), as described in section 3. The emulated coefficients reconstruct fields of ice thickness, which are converted to binary representations via (6.1), and compared to ice extent observations as in (6.2). We set the ice presence threshold as $T^b = 10$ to account for observational error in the ice extent. This choice is uncertain and could in future be a distribution.

Table 2

Information for outputs used in history matching. "Waves" indicates at which waves the output was emulated, "Impl" gives which implausibility was used (with $g(\cdot)$ when required), ℓ is the dimension of the output, σ_e^2 is the observation error variance for the volumes, and b_{id} is the history matching bound for output id.

Output	*id*	Waves	Impl	$g(\cdot)$	ℓ	σ_e^2	b_{id}
21 ka volume	*vol*21	1,2	$\mathcal{I}$		1	4	3^2
21 ka southwest region	*sw*21	1	$\mathcal{I}^b$	*min*	1116		0.25ℓ
21 ka central region	*ce*21	1	$\mathcal{I}^b$	*mean*	868		0.025ℓ
14 ka volume	*vol*14	2,3	$\mathcal{I}$		1	1.416	3^2
14 ka region	*reg*14	2	$\mathcal{I}^b$	*min*	1176		0.33ℓ
10 ka volume	*vol*10	2	$\mathcal{I}$		1	0.279	3^2
10 ka region	*reg*10	2	$\mathcal{I}^b$	*mean*	1066		0.25ℓ
6 ka region	*reg*6	2	$\mathcal{I}^b$	*mean*	1271		0.25ℓ

7.1. Wave 1. At wave 1, only 21 ka was considered, as the output for later years is dependent on the initial ice sheet, and due to the wide range of behaviors possible by varying the boundary condition coefficients, the initial ice sheet is often clearly implausible. The top half of Figure 9 shows the average ice extent at 21 ka across the wave 1 ensemble, and the volume through time, with a large spread of potential volumes at 21 ka. We built emulators for the volume at 21 ka (leave-one-out cross-validation in Figure **SM3**), and two separate regions of spatial output, indicated in Figure 9, covering regions where there was a difference between the ensemble and the extent reconstruction at 21 ka: the center of the ice sheet, where there should be ice, and the southwest, where Glimmer's ice coverage generally extends too far.

For the binary implausibility, Figure **SM2** considers the choice of $g(\cdot)$ for these two regions, comparing the implausibility given by the emulators to the truth across the wave 1 ensemble. For the central region, $g(\cdot) = \text{mean}(\cdot)$ ruled out space accurately (using the 5th percentile or minimum failed to rule out several runs that poorly matched $\mathbf{z}^b$, as the distribution of $\mathcal{I}^b$ was often bimodal for this region, with the latent process close to T^b), whereas $g(\cdot) = \min(\cdot)$ was more accurate for the southwest region.

At wave 1, we required each output to be not implausible, defining NROY space as

$$(\mathcal{X} \times \mathcal{C})^{(1)}_{NROY} = \{(\mathbf{x}, \mathbf{c}) \in \mathcal{X} \times \mathcal{C} \,|\, \mathcal{I}_i(\mathbf{x}, \mathbf{c}) < b_i, i = (vol21, sw21, ce21)\},$$

for bounds b_i given in Table 2. For the volume, we have $b_{vol21} = 3^2$, as is standard in univariate history matching. For the southwest region, we use a 25% mismatch (as discussed in section 6). For the central region, this was too high, and in order to rule out runs that generally look implausible, we set the bound at 2.5% of the spatial dimension ($\ell = 868$ for this region). $(\mathcal{X} \times \mathcal{C})^{(1)}_{NROY}$ consists of 5.4% of the original space, illustrated in Figure **SM4**. We sampled 500 points from $(\mathcal{X} \times \mathcal{C})^{(1)}_{NROY}$, chosen via a combination of ensuring we had a space-filling sample from $(\mathcal{X} \times \mathcal{C})^{(1)}_{NROY}$ and including parameter settings leading to the lowest implausibilities for each emulated output, and ran Glimmer to obtain the wave 2 ensemble, $\mathbf{F}^{(2)}$.

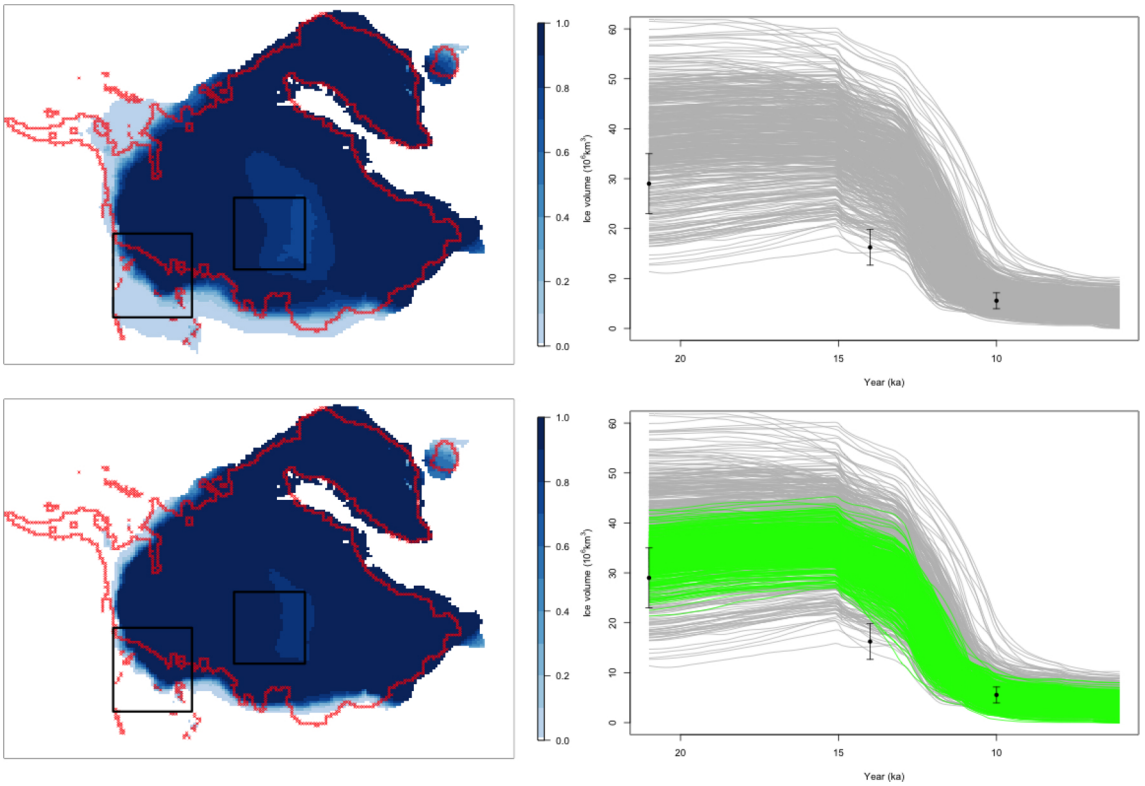

Figure 9. *Left: the proportion of wave* 1 *ensemble members containing ice in a grid box at* 21 *ka, with red crosses indicating the ice extent, and the wave* 1 *emulated regions shown by the two boxes (top), with the wave* 2 *ensemble below. Right: the volume of wave* 1 *ensemble members in time, with observations (with error) shown in black (top), with the wave* 2 *ensemble added in green in the lower plot.*

The lower half of Figure 9 compares waves 1 and 2. Runs where there was significantly too much ice at the southwest edge of the ice sheet in wave 1 have now been ruled out, so that the output is more consistent with the observed extent at 21 ka. There are also fewer runs in the wave 2 ensemble with an opening in the center of the ice sheet, although all parameter settings that lead to this have not yet been ruled out. The lower right plot shows that we have ruled out runs with a significantly too high or low volume at 21 ka, with the spread of possible volumes at wave 2 (green) much smaller than at wave 1. Despite this improvement, we still generally do not satisfy the volume constraint at 14 ka.

7.2. Waves 2 and 3. At wave 2, we used metrics throughout the deglaciation (21 ka, 14 ka, 10 ka, 6 ka), allowing coefficients controlling the boundary condition in each of the three time periods to be constrained. These time points were chosen to give a coverage of the entire deglaciation. The spatial regions were chosen by selecting regions towards the edge of the ice sheet, where the ensemble generally does not match observations, but where there is a range of behaviors in the ensemble (left halves of Figures **SM5**, **SM6**, and **SM7**). We also emulate the volume at 21 ka again, as there are runs in the wave 2 ensemble that have too much ice at this time, and also emulate the volume at 14 ka, where the model runs generally have too

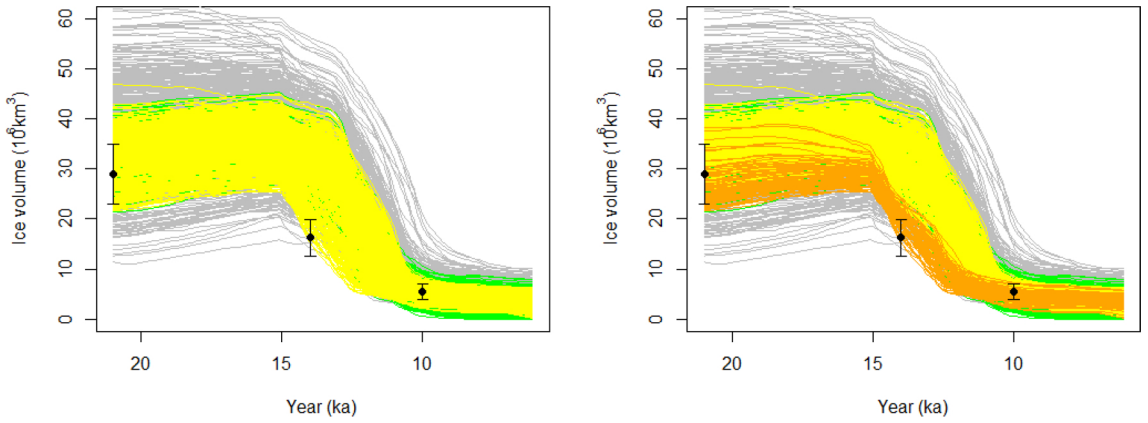

Figure 10. *Ice sheet volume for the wave* 1 *(gray), wave* 2 *(green), and wave* 3 *(yellow) ensembles, with the observations and observational error shown in black. The orange runs added in the right plot are those not ruled out after wave* 3.

much ice, and 10 ka (already several runs that satisfy this). Validation plots for the volume emulators at 21 ka and 14 ka are shown in Figure **SM3**.

At this wave, we found that requiring all 6 of the emulated outputs to be consistent with the observations was an extremely strong constraint, with it challenging to match the observed volume at 14 ka while simultaneously satisfying all other constraints. We do not know whether it is actually possible for Glimmer to satisfy all 6 of these constraints at the same time; hence to allow us to further explore this, we relax our definition of NROY space and keep $(\mathbf{x}, \mathbf{c})$ where at least 4 of the 6 constraints are satisfied (with $\hat{\mathcal{I}}$ the scaled implausibility as in (4.4)):

$$(\mathcal{X} \times \mathcal{C})^{(2)}_{NROY} = \{(\mathbf{x}, \mathbf{c}) \in (\mathcal{X} \times \mathcal{C})^{(1)}_{NROY} \,|\, \hat{\mathcal{I}}_{4M}(\mathbf{x}, \mathbf{c}) < 3\}.$$

After wave 2, NROY space is 1.1% of the full space (Figure **SM8**). We obtain a wave 3 ensemble by sampling from $(\mathcal{X} \times \mathcal{C})^{(2)}_{NROY}$ using the same considerations as at the previous wave. Figures **SM5**, **SM6**, and **SM7** show how the ice extent has changed between waves 2 and 3 for the emulated regions, and the left half of Figure 10 plots the volume for each of the three ensembles, with wave 3 added in yellow.

At wave 2, few runs matched the volume observation at 14 ka, with the majority of ice sheets retreating too slowly. In the wave 3 ensemble, while the spread of volumes at 21 ka is extremely similar to that of wave 2, a number of these runs do now exhibit ice sheets that melt quickly enough, with several runs satisfying the 14 ka volume constraint. The wave 3 ensemble also has a narrower, more accurate range of volumes at 10 ka than at wave 2.

With no more resources for running Glimmer, we used the volume at 14 ka to give a final NROY space, as this was a difficult constraint to satisfy. Emulating using the wave 3 ensemble, the wave 3 NROY space is

$$(\mathcal{X} \times \mathcal{C})^{(3)}_{NROY} = \{(\mathbf{x}, \mathbf{c}) \in (\mathcal{X} \times \mathcal{C})^{(2)}_{NROY} \,|\, \mathcal{I}_{vol14}(\mathbf{x}, \mathbf{c}) < 3^2\},$$

and consists of 0.06% of the original space (Figure **SM9**).

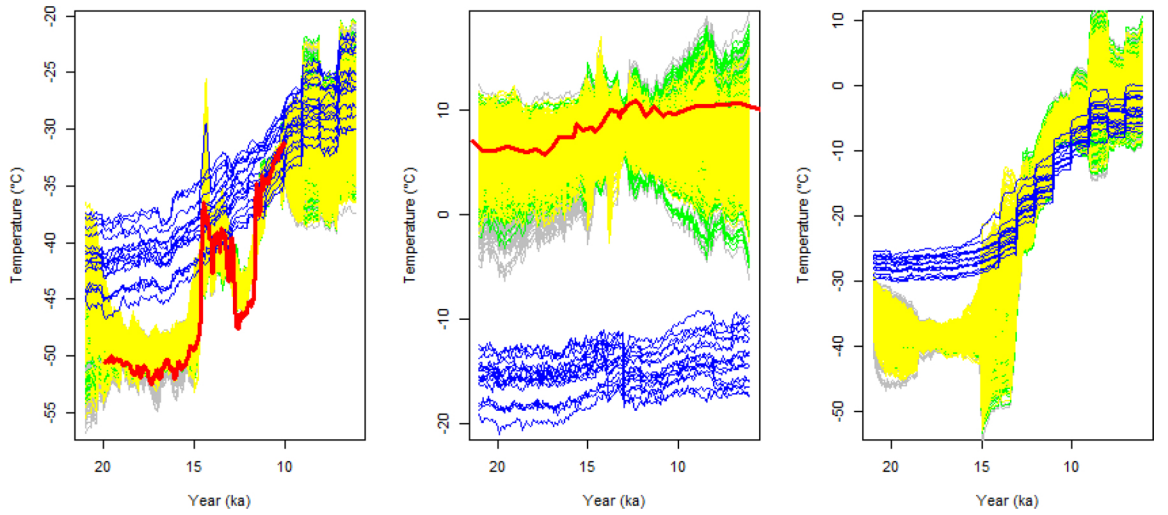

Figure 11. *The observed temperatures (red), climate ensemble (blue), wave 1 (gray), wave 2 (green), and wave 3 (yellow) boundary conditions in Greenland (left), Alaska (middle), and North America (right).*

The right half of Figure 10 shows the ensemble members that lie in this final NROY space, colored orange. As this space was defined using only an emulator for the volume at 14 ka, the ensemble members that are not ruled out generally match the observed time series of ice melt, with retreat occurring rapidly enough prior to 14 ka, which was an issue in the first two ensembles.

7.3. Boundary condition uncertainty. Our initial goal was to calibrate the boundary condition of Glimmer; hence we now consider how the space of possible boundary conditions has evolved due to history matching.

Figure 11 shows the simulated boundary conditions for each of the three ensembles, compared to the observations and climate ensemble. In each location, we have started to reduce the spread of plausible temperatures; e.g., for each, we have ruled out the coldest temperatures in the first time period. Even though the original dimension of the input was over 300 million, we can quantify and constrain uncertainty in the boundary condition using only 13 coefficients to represent it.

Figure 12 shows the spread of boundary conditions for the wave 3 ensemble members that were not ruled out after wave 3 (orange lines), compared to the previous waves. When we only consider these runs, we see that the range of boundary condition temperatures that are possible in each location has been reduced more substantially than at previous waves. In Greenland, the warmest peaks around 13 ka have been ruled out, as well as the coldest initial temperatures. In both Alaska and North America, the spread of temperatures in the first time period has been reduced further. All of the boundary conditions featuring the large, unrealistic, downward temperature shifts in North America have been ruled out at wave 3, leaving temperature profiles that generally increase through time, as expected. History matching has enabled us to rule out clearly unphysical boundary conditions.

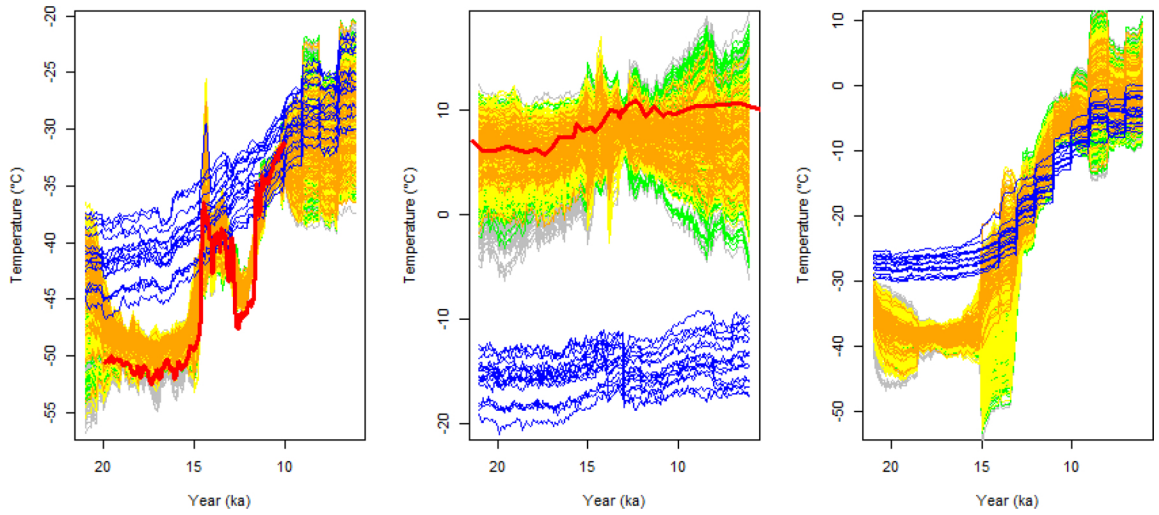

Figure 12. *The observed temperatures (red), climate ensemble (blue), wave* 1 *(gray), wave* 2 *(green), wave* 3 *(yellow), and the not ruled out wave* 3 *(orange) boundary conditions in Greenland (left), Alaska (middle), and North America (right).*

The final range of temperatures in North America (over the ice sheet) has been reduced from the initial space of boundary conditions that we allowed. The range of not implausible boundary conditions is similar to the range given by the climate ensemble, but with a different warming trajectory. At the start of the deglaciation, our boundary condition allows colder temperatures. At the end, we initially had a spread of possible temperatures covering the climate ensemble, but ruled out most of the colder boundary conditions, suggesting that the climate ensemble may be too cold after 10 ka.

Overall, we have found a subset of the initial boundary condition space, distinct from the original climate ensemble, that allows Glimmer to reproduce the ice sheet volume through the deglaciation (Figure 10).

8. Discussion. We have developed a framework for calibrating high-dimensional boundary conditions that have sparse observational data. Our method allows a range of plausible boundary conditions to be efficiently generated, given a small ensemble of climate model runs and sparse observations. The resulting boundary conditions are more historically accurate, and more varied, than the small number of GCM runs that have been used to force Glimmer previously, removing large known biases between the boundary conditions and geological temperature observations. The parameterized low-dimensional form allows calibration of the boundary conditions and computer model parameters to be performed jointly, quantifying the uncertainty in the output due to each, not possible without a boundary condition model and some form of dimension reduction.

Performing three waves of history matching on the Glimmer ice sheet model reduced the range of possible temperatures compared to the prior, while also allowing the interaction of the boundary conditions with the ice sheet parameters to be explored. The reduction in the spread of ice sheet volumes, and the improvement in the ice extents, shows the success of this

method. At wave 3, we have started to identify runs with a melting trajectory consistent with observations of the deglaciation (not present at wave 1). With further ensembles of Glimmer, we would explore this space to identify runs with ice extents that are more consistent with observations through the deglaciation.

Our method for history matching binary observations successfully identified runs with ice extent more consistent with the observations than in the initial ensemble, although there was still a reasonably large discrepancy between the wave 3 ensemble and the truth. To identify whether this is due to not yet finding the region of the input space that leads to output most consistent with ice extent reconstructions, or that this is structural error, would require further waves of history matching. Modeling the precipitation boundary condition may give further improvements, although this problem suffers from an even greater lack of geological observations.

Selecting an appropriate low-dimensional basis for the boundary condition is important. We were restricted by the small size of the GCM ensemble, hence the approach of splitting selection into two steps, but with a larger ensemble it may be possible to directly select spatio-temporal vectors that match historical observations, due to the higher number of degrees of freedom available.

An extension of the binary history matching method presented here would be to consider the probability that the latent thickness process matches the observations in each grid box, rather than using the fixed threshold to convert ice thickness to binary. This would likely improve matching in regions where the emulated thickness is relatively low, and hence close to the threshold for the presence of ice.

Enabling model performance to be explored under a range of realistic boundary conditions may lead to improved future development of ice sheet models, with a better understanding of which processes cannot be represented currently, and where discrepancies with observations lie. In general, our modeling framework can be combined with expert elicitation, for example, to explore the effect that certain patterns or changes to the boundary conditions have on the ice sheet.

Ultimately, improved ice sheet models that more accurately simulate the past will help with understanding the deglaciation. Reducing the uncertainty in the past temperatures required to give a realistic deglaciation may eventually help to improve future projections, with these being used as out-of-sample constraints for the output of global climate models run into the future.

REFERENCES

C. Buizert, V. Gkinis, J. P. Severinghaus, F. He, B. S. Lecavalier, P. Kindler, M. Leuenberger, A. E. Carlson, B. Vinther, V. Masson-Delmotte, et al. (2014), *Greenland temperature response to climate forcing during the last deglaciation*, Science, 345, pp. 1177–1180.

A. E. Carlson and P. U. Clark (2012), *Ice sheet sources of sea level rise and freshwater discharge during the last deglaciation*, Rev. Geophys., 50, RG4007.

W. Chang, P. J. Applegate, M. Haran, and K. Keller (2014), *Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: Toward projections of ice mass loss with uncertainties*, Geosci. Model Dev. Discuss., 7, pp. 1905–1931.

W. Chang, M. Haran, P. Applegate, and D. Pollard (2016), *Calibrating an ice sheet model using high-dimensional binary spatial data*, J. Amer. Statist. Assoc., 111, pp. 57–72.

W. Chang, B. A. Konomi, G. Karagiannis, Y. Guan, and M. Haran (2019), *Ice Model Calibration Using Semi-Continuous Spatial Data*, preprint, https://arxiv.org/abs/1907.13554.

S. Charbit, C. Ritz, G. Philippon, V. Peyaud, and M. Kageyama (2007), *Numerical reconstructions of the Northern Hemisphere ice sheets through the last glacial-interglacial cycle*, Climate Past, 3, pp. 15–37.

P. S. Craig, M. Goldstein, J. C. Rougier, and A. H. Seheult (2001), *Bayesian forecasting for complex systems using computer simulators*, J. Amer. Statist. Assoc., 96, pp. 717–729.

P. S. Craig, M. Goldstein, A. Seheult, and J. Smith (1996), *Bayes linear strategies for matching hydrocarbon reservoir history*, in Bayesian Statistics, 5, Oxford University Press, pp. 69–95.

P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith (1997), *Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes linear strategies for large computer experiments*, in Case Studies in Bayesian Statistics, Springer, pp. 37–93.

A. S. Dyke (2004), *An outline of North American deglaciation with emphasis on central and northern Canada*, in Quaternary Glaciations: Extent and Chronology, Vol. 2, Elsevier, pp. 373–424.

T. L. Edwards, M. Brandon, G. Durand, N. R. Edwards, N. R. Golledge, P. B. Holden, I. Nias, A. Payne, C. Ritz, and A. Wernecke (2018), *Revisiting Antarctic ice loss due to marine ice cliff instability*, Nature, 566, pp. 58–64.

L. J. Gregoire (2010), *Modelling the Northern Hemisphere Climate and Ice Sheets During the Last Deglaciation*, Ph.D. thesis, University of Bristol.

L. J. Gregoire, B. Otto-Bliesner, P. J. Valdes, and R. Ivanovic (2016), *Abrupt Bølling warming and ice saddle collapse contributions to the meltwater pulse a rapid sea level rise*, Geophys. Res. Lett., 43, pp. 9130–9137.

L. J. Gregoire, A. J. Payne, and P. J. Valdes (2012), *Deglacial rapid sea level rises caused by ice-sheet saddle collapses*, Nature, 487, pp. 219–222.

L. J. Gregoire, P. J. Valdes, and A. J. Payne (2015), *The relative contribution of orbital forcing and greenhouse gases to the North American deglaciation*, Geophys. Res. Lett., 42, pp. 9970–9979.

L. J. Gregoire, P. J. Valdes, A. J. Payne, and R. Kahana (2011), *Optimal tuning of a GCM using modern and glacial constraints*, Climate Dynam., 37, pp. 705–719.

J. Gregory, O. Browne, A. Payne, J. Ridley, and I. Rutt (2012), *Modelling large-scale ice-sheet-climate interactions following glacial inception*, Climate Past, 8, pp. 1565–1580.

M. Gu and L. Wang (2018), *Scaled Gaussian stochastic process for computer model calibration and prediction*, SIAM/ASA J. Uncertain. Quantif., 6, pp. 1555–1583, https://doi.org/10.1137/17M1159890.

E. Hawkins, R. S. Smith, L. C. Allison, J. M. Gregory, T. J. Woollings, H. Pohlmann, and B. De Cuevas (2011), *Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport*, Geophys. Res. Lett., 38, L10605.

F. Hebeler, R. S. Purves, and S. S. Jamieson (2008), *The impact of parametric uncertainty and topographic error in ice-sheet modelling*, J. Glaciol., 54, pp. 899–919.

D. Higdon, J. Gattiker, B. Williams, and M. Rightley (2008), *Computer model calibration using high-dimensional output*, J. Amer. Statist. Assoc., 103, pp. 570–583.

P. B. Holden, N. R. Edwards, P. H. Garthwaite, and R. D. Wilkinson (2015), *Emulation and interpretation of high-dimensional climate model outputs*, J. Appl. Stat., 42, pp. 2038–2055.

F. Hourdin, T. Mauritsen, A. Gettelman, J.-C. Golaz, V. Balaji, Q. Duan, D. Folini, D. Ji, D. Klocke, Y. Qian, et al. (2017), *The art and science of climate model tuning*, Bull. Amer. Meteorol. Soc., 98, pp. 589–602.

R. Ivanovic, L. Gregoire, M. Kageyama, D. Roche, P. Valdes, A. Burke, R. Drummond, W. R. Peltier, and L. Tarasov (2016), Transient climate simulations of the deglaciation 21-9 thousand years before present (version 1)-PMIP4 core experiment design and boundary conditions, Geosci. Model Dev., 9, pp. 2563–2587.

R. Ivanovic, L. Gregoire, A. Wickert, and A. Burke (2018), *Climatic effect of Antarctic meltwater overwhelmed by concurrent Northern hemispheric melt*, Geophys. Res. Lett., 45, pp. 5681–5689.

L. Jackson, R. S. Smith, and R. Wood (2017), *Ocean and atmosphere feedbacks affecting AMOC hysteresis in a GCM*, Climate Dyn., 49, pp. 173–191.

M. C. Kennedy and A. O'Hagan (2001), *Bayesian calibration of computer models*, J. R. Stat. Soc. B Stat. Methodol., 63, pp. 425–464.

X. Liu and S. Guillas (2017), *Dimension reduction for Gaussian process emulation: An application to the influence of bathymetry on tsunami heights*, SIAM/ASA J. Uncertain. Quantif., 5, pp. 787–812, https://doi.org/10.1137/16M1090648.

Z. Liu, B. Otto-Bliesner, F. He, E. Brady, R. Tomas, P. Clark, A. Carlson, J. Lynch-Stieglitz, W. Curry, E. Brook, et al. (2009), *Transient simulation of last deglaciation with a new mechanism for Bølling-Allerød warming*, Science, 325, pp. 310–314.

D. J. Lunt, G. L. Foster, A. M. Haywood, and E. J. Stone (2008), *Late Pliocene Greenland glaciation controlled by a decline in atmospheric CO 2 levels*, Nature, 454, pp. 1102–1105.

D. McNeall, P. Challenor, J. Gattiker, and E. Stone (2013), *The potential of an observational data set for calibration of a computationally expensive computer model*, Geosci. Model Dev., 6, pp. 1715–1728.

L. Menviel, A. Timmermann, O. E. Timm, and A. Mouchet (2011), *Deconstructing the last glacial termination: The role of millennial and orbital-scale forcings*, Quat. Sci. Rev., 30, pp. 1155–1172.

B. Noël, W. Jan Van De Berg, H. Machguth, S. Lhermitte, I. Howat, X. Fettweis, and M. R. Van Den Broeke (2016), *A daily, 1 km resolution data set of downscaled Greenland ice sheet surface mass balance (1958-2015)*, Cryosphere, 10, pp. 2361–2377.

H. Patton, A. Hubbard, K. Andreassen, A. Auriac, P. L. Whitehouse, A. P. Stroeven, C. Shackleton, M. Winsborrow, J. Heyman, and A. M. Hall (2017), *Deglaciation of the Eurasian ice sheet complex*, Quat. Sci. Rev., 169, pp. 148–172.

M. Plumlee (2017), *Bayesian calibration of inexact computer models*, J. Amer. Statist. Assoc., 112, pp. 1274–1285.

D. Pollard, W. Chang, M. Haran, P. Applegate, and R. DeConto (2016), *Large ensemble modeling of the last deglacial retreat of the West Antarctic Ice Sheet: Comparison of simple and advanced statistical techniques*, Geosci. Model Dev., 9, pp. 1697–1723, https://doi.org/10.5194/gmd-9-1697-2016.

F. Pukelsheim (1994), *The three sigma rule*, Amer. Statist., 48, pp. 88–91.

N. Reeh (1989), *Parameterization of melt rate and surface temperature on the Greenland ice sheet*, Polarforschung, 59, pp. 113–128.

W. H. Roberts, P. J. Valdes, and A. J. Payne (2014), *Topography's crucial role in Heinrich Events*, Proc. Natl. Acad. Sci. USA, 111, pp. 16688–16693.

D. M. Roche, H. Renssen, D. Paillard, and G. Levavasseur (2011), *Deciphering the spatio-temporal complexity of climate change of the last deglaciation: A model analysis*, Climate Past, 7, pp. 591–602.

I. C. Rutt, M. Hagdorn, N. Hulton, and A. Payne (2009), *The Glimmer community ice sheet model*, J. Geophys. Res. Earth Surf., 114, F02004.

J. M. Salter and D. Williamson (2016), *A comparison of statistical emulation methodologies for multi-wave calibration of environmental models*, Environmetrics, 27, pp. 507–523.

J. M. Salter, D. B. Williamson, J. Scinocca, and V. Kharin (2019), *Uncertainty quantification for computer models with spatial output using calibration-optimal bases*, J. Amer. Statist. Assoc., 114, pp. 1800–1814.

J. Seguinot, C. Khroulev, I. Rogozhina, A. P. Stroeven, and Q. Zhang (2014), *The effect of climate forcing on numerical simulations of the Cordilleran ice sheet at the Last Glacial Maximum*, Cryosphere, 8, pp. 1087–1103.

D. M. Sexton, J. M. Murphy, M. Collins, and M. J. Webb (2011), *Multivariate probabilistic projections using imperfect climate models Part* I*: Outline of methodology*, Climate Dyn., 38, pp. 2513–2542.

J. D. Shakun, P. U. Clark, F. He, S. A. Marcott, A. C. Mix, Z. Liu, B. Otto-Bliesner, A. Schmittner, and E. Bard (2012), *Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation*, Nature, 484, pp. 49–54.

R. S. Smith and J. M. Gregory (2009), *A study of the sensitivity of ocean overturning circulation and climate to freshwater input in different regions of the North Atlantic*, Geophys. Res. Lett., 36, L15701.

R. S. Smith, J. M. Gregory, and A. Osprey (2008), *A description of the FAMOUS (version XDBUA) climate model and control run*, Geosci. Model Dev., 1, pp. 53–68.

E. Stone, D. Lunt, I. Rutt, and E. Hanna (2010), *Investigating the sensitivity of numerical model simulations of the modern state of the Greenland ice-sheet and its future response to climate change*, Cryosphere, 4, pp. 397–417.

C.-L. Sung, Y. Hung, W. Rittase, C. Zhu, and C. Wu (2020), *Calibration for computer experiments with binary responses and application to cell adhesion study*, J. Amer. Statist. Assoc., 115, pp. 1664–1674.

L. Tarasov, A. S. Dyke, R. M. Neal, and W. R. Peltier (2012), *A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling*, Earth Planet. Sci. Lett., 315, pp. 30–40.

R. Tuo and C. F. J. Wu (2016), *A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties*, SIAM/ASA J. Uncertain. Quantif., 4, pp. 767–795, https://doi.org/10.1137/151005841.

I. Vernon, M. Goldstein, and R. G. Bower (2010), *Galaxy formation: A Bayesian uncertainty analysis*, Bayesian Anal., 5, pp. 619–669.

K. von Salzen, J. F. Scinocca, N. A. McFarlane, J. Li, J. N. Cole, D. Plummer, D. Verseghy, M. C. Reader, X. Ma, M. Lazare, and L. Solheim (2013), *The Canadian fourth generation atmospheric global climate model (CanAM4). Part* I*: Representation of physical processes*, Atmos. Ocean, 51, pp. 104–125.

R. D. Wilkinson (2010), *Bayesian calibration of expensive multivariate computer experiments*, in Large-Scale Inverse Problems and Quantification of Uncertainty, L. T. Biegler, et al., eds., John Wiley, pp. 195–216.

J. Williams, R. Smith, P. Valdes, B. Booth, and A. Osprey (2013), *Optimising the FAMOUS climate model: Inclusion of global carbon cycling*, Geosci. Model Dev., 6, pp. 141–160.

J. Williams, I. Totterdell, P. Halloran, and P. Valdes (2014), *Numerical simulations of oceanic oxygen cycling in the FAMOUS Earth-System model: FAMOUS-ES, version* 1.0, Geosci. Model Dev., 7, pp. 1419–1431.

D. Williamson, A. T. Blaker, C. Hampton, and J. Salter (2015), *Identifying and removing structural biases in climate models with history matching*, Climate Dyn., 45, pp. 1299–1324.

D. Williamson, M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki (2013), *History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble*, Climate Dyn., 41, pp. 1703–1729.

D. B. Williamson, A. T. Blaker, and B. Sinha (2017), *Tuning without over-tuning: Parametric uncertainty quantification for the NEMO ocean model*, Geosci. Model Dev., 10, pp. 1789–1816.