The Annals of Applied Statistics
2020, Vol. 14, No. 2, 605–634
https://doi.org/10.1214/19-AOAS1305

A FAST PARTICLE-BASED APPROACH FOR CALIBRATING A 3-D MODEL OF THE ANTARCTIC ICE SHEET

BY BEN SEIYON LEE[1,*], MURALI HARAN[1,†], ROBERT W. FULLER[2,‡], DAVID POLLARD[3] AND KLAUS KELLER[2,§]

[1]*Department of Statistics, Pennsylvania State University,* [*]*skl5261@psu.edu;* [†]*mharan@stat.psu.edu*
[2]*Department of Geosciences, Pennsylvania State University,* [‡]*rwf136@psu.edu;* [§]*klaus@psu.edu*
[3]*Earth and Environmental Systems Institute, Pennsylvania State University, pollard@essc.psu.edu*

We consider the scientifically challenging and policy-relevant task of understanding the past and projecting the future dynamics of the Antarctic ice sheet. The Antarctic ice sheet has shown a highly nonlinear threshold response to past climate forcings. Triggering such a threshold response through anthropogenic greenhouse gas emissions would drive drastic and potentially fast sea level rise with important implications for coastal flood risks. Previous studies have combined information from ice sheet models and observations to calibrate model parameters. These studies have broken important new ground but have either adopted simple ice sheet models or have limited the number of parameters to allow for the use of more complex models. These limitations are largely due to the computational challenges posed by calibration as models become more computationally intensive or when the number of parameters increases.

Here, we propose a method to alleviate this problem: a fast sequential Monte Carlo method that takes advantage of the massive parallelization afforded by modern high-performance computing systems. We use simulated examples to demonstrate how our sample-based approach provides accurate approximations to the posterior distributions of the calibrated parameters. The drastic reduction in computational times enables us to provide new insights into important scientific questions, for example, the impact of Pliocene era data and prior parameter information on sea level projections. These studies would be computationally prohibitive with other computational approaches for calibration such as Markov chain Monte Carlo or emulation-based methods. We also find considerable differences in the distributions of sea level projections when we account for a larger number of uncertain parameters. For example, based on the same ice sheet model and data set, the 99th percentile of the Antarctic ice sheet contribution to sea level rise in 2300 increases from 6.5 m to 13.1 m when we increase the number of calibrated parameters from three to 11. With previous calibration methods, it would be challenging to go beyond five parameters. This work provides an important next step toward improving the uncertainty quantification of complex, computationally intensive and decision-relevant models.

1. Introduction. How much will the Antarctic ice sheet contribute to future sea level rise? The geological records suggest that ice sheets can quickly contribute considerable amounts to global sea level rise (Deschamps et al. (2012)), in some cases up to 58 m (Fretwell et al. (2012)). Projections of future sea level rise depend on deeply uncertain projections of the Antarctic ice sheet's (AIS) mass loss (Le Bars, Drijfhout and de Vries (2017), Wong, Bakker and Keller (2017), Le Cozannet, Manceau and Rohmer (2017)). Close to eight percent of the current global population is threatened by a five meter rise in sea level (Nicholls, Tol and

Received March 2019; revised October 2019.
Key words and phrases. Antarctic ice sheet model, computer model calibration, paleoclimate, sequential Monte Carlo, uncertainty quantification.

Vafeidis (2008)), and 13 percent of the global urban population is threatened by a 10 meter sea level rise (McGranahan, Balk and Anderson (2007)). Quantifying and characterizing the long-term behavior of the Antarctic ice sheet is hence a key input to the design of coastal risk-management strategies (cf. Garner and Keller (2018), Sriver et al. (2018), Oppenheimer and Alley (2016)).

Ice sheet models rely on poorly constrained parameters, and recent studies show that uncertainty in model parameters results in highly uncertain projections of sea level change (Stone et al. (2010), Applegate et al. (2012), Fitzgerald et al. (2012), Collins (2007)), thereby affecting climate risk decision-making (O'Neill et al. (2006), Hannart et al. (2013)). Recent studies have addressed parametric uncertainty via calibration studies using modern observations, but these are either limited to simple ice sheet models (Ruckert et al. (2017), Fuller, Wong and Keller (2017)) or a small number of model parameters (Chang et al. (2016b), Edwards et al. (2019), Schlegel et al. (2018)). Numeric solvers have been used to infer the field of basal sliding parameters from satellite observations (Isaac, Stadler and Ghattas (2015), Isaac et al. (2015)).

Ice-sheet models vary in complexity, and the key drivers of computational cost are the spatial and temporal resolutions. Simpler models (cf. Bakker, Applegate and Keller (2016), Shaffer (2014)) have short computer model run times on the order of a few seconds, but they may oversimplify or even exclude important physical processes. More complex models (cf. DeConto and Pollard (2016), Larour et al. (2012), Greve (1997), Rutt et al. (2009)) can better represent key ice dynamics and typically run at higher spatiotemporal resolutions. However, they require longer model run times. Here, we use a relatively complex ice-sheet model, the Pennsylvania State University 3D ice-sheet model (PSU3D-ICE) (Pollard and DeConto (2012)), but with considerably coarser resolution than in previous work, so that each set of simulations for this study takes on the order of 10 to 15 minutes of wall time.

Past studies calibrate simpler models with many model parameters using Markov chain Monte Carlo (MCMC) (cf. Ruckert et al. (2017), Bakker, Applegate and Keller (2016), Petra et al. (2014)); these approaches are effective in the context of computationally inexpensive models (model run times of a few seconds) and, hence, do not extend to the kind of models we consider in this manuscript. Some studies have employed emulation-calibration methods (Sansó, Forest and Zantedeschi (2008), Liu, Bayarri and Berger (2009), Bhat et al. (2010)) to calibrate computer models with long run times, but these approaches are applicable to only a small number of parameters. For computer models with longer run times and a large number of model parameters, emulation-calibration can be computationally prohibitive because building an accurate emulator requires a very large set of training data (Bastos and O'Hagan (2009), Maniyar, Cornford and Boukouvalas (2007)).

We propose calibrating an ice sheet model which: (1) accounts for important physical processes, (2) includes several key parameters to analyze and quantify parametric uncertainty and (3) expands the calibration dataset to the Pliocene. For this study, the Antarctic ice sheet model runs at a spatial resolution of 80 km and temporal resolution of eight years, which is a compromise between preserving reasonable accuracy of physical simulations vs. maintaining a feasible model run time. We estimate that current rigorous methods for calibrating this model via MCMC would take roughly on the order of years of wall time. We investigate methods for calibration that are amenable to heavy parallelization and computationally efficient, thereby reducing the computational wall time from years to hours. We find that these methods are broadly applicable to computer models with a moderate model run time (six seconds to 15 minutes) and a moderate number of model parameters (five to 20), based on available computing resources. While this does not cover more complex models or larger number of parameters, our methods are applicable to many scientifically important and policy-relevant computer models.

Studying the Antarctic ice sheet's future behavior motivates the need for a computationally efficient approach for computer model calibration. We turn to sequential Monte Carlo methods (cf. Doucet, Godsill and Andrieu (2000), Del Moral, Doucet and Jasra (2006), Chopin (2002)), building upon particle-based methods for computer model calibration (Higdon et al. (2008), Kalyanaraman et al. (2016)). Our approach builds upon an adaptive tempering schedule and an adaptive mutation stage (Jasra et al. (2011)) which have been used for Bayesian variable selection (Schäfer and Chopin (2013)), Bayesian model comparison (Zhou, Johansen and Aston (2016)) and estimating initial conditions of the Navier–Stokes system of equations (Kantas, Beskos and Jasra (2014), Llopis et al. (2018)).

By using massive parallelization in a high-performance computing environment, we obtain a dramatic speed-up over current MCMC-based calibration methods, roughly reducing wall time by a factor of 3000. We also limit expensive computer model runs by imposing stopping rules and adaptive sampling techniques. We provide practical guidelines designed to: (1) reduce total wall time, (2) limit the number of expensive computer model runs and (3) simplify implementation for the user. Our computationally efficient calibration approach is readily applicable to many computer models for which rigorous calibration may be currently infeasible.

We note that we focus on a "static" system where all observations are available at once; hence, there is only one posterior distribution of interest, which we approximate using our particle-based approach. The PSU3D-ICE model is dissipative where it evolves to a single constant steady state for a given set of parameter values and external forcing (Willems (1972)). Unlike choatic systems, such as global weather models, "microscopic" changes in the initial states do not change the results; in other words, there is no "butterfly effect" (Lorenz (1972)). We use our approach to calibrate the PSU3D-ICE model (DeConto and Pollard (2016)) using paleoclimate data and modern observational records. Previous work focuses on calibrating the PSU3D-ICE model using fewer parameters (Chang et al. (2016b), Edwards et al. (2019)) or surrogate models using limited training data (Chang et al. (2016a)). Using our new method, we show that the information regarding the extent of the Antarctic ice sheet in the Pliocene era strongly influences parametric and projection uncertainty. We find that using improved geological data and analysis to characterize the Antarctic ice sheet's contribution to sea level rise in the Pliocene can bring about considerably sharper sea level projections for future centuries.

The paper is structured as follows. In Section 2 we provide an overview of the ice-sheet model (PSU3D-ICE). In Section 3 we describe the model calibration framework and discuss challenges with current calibration methods. We propose our fast particle-based approach for computer model calibration in Section 4. In Section 5 we demonstrate the application of our method to a simulated example. In Section 6 we apply our method to the PSU3D-ICE model and report our scientific conclusions. We end with caveats and directions for future research in Section 7.

2. Computer model description and observations. In this section we provide background information for the PSU3D-ICE Antarctic ice-sheet model (DeConto and Pollard (2016)) as well as the paleoclimate records and modern observations used to calibrate the model.

2.1. *The PSU3D-ICE model.* The PSU3D-ICE model simulates the long-term dynamics of continental ice sheets. It has previously been applied to past and future variations of the Antarctic ice sheet (Pollard and DeConto (2009, 2012), Pollard, DeConto and Alley (2015), Pollard et al. (2016), Pollard, Gomez and Deconto (2017)). Slow ice deformation under its own weight is modeled by scaled dynamical equations for internal shear, horizontal

stretching and basal sliding. Other variables and processes include internal ice temperatures, bedrock deformation beneath the ice load, surface snowfall and melting, oceanic melting beneath floating ice shelves and calving of ice into the ocean (Pollard and DeConto (2012)). A recently proposed mechanism called marine ice-cliff instability (MICI) that can drastically attack ice in marine basins, involving hydrofracturing due to surface liquid water and structural failure of tall ice cliffs, is included here (Pollard, DeConto and Alley (2015), DeConto and Pollard (2016)). Note that this mechanism has recently been questioned (Edwards et al. (2019), Golledge et al. (2019)).

For the simulations in this study, a polar stereographic grid spanning Antarctica is used with a horizontal resolution of 80 kilometers (km) which yields a model run time of approximately 10 to 15 minutes for each set of past and future simulations described below. This is a considerably coarser spatial resolution than previous continental-scale applications which have used resolutions of 10 to 40 km. However, sensitivity tests with the model show reasonable independence of results with model resolution, due to the grid-independent parameterization of important subgrid processes, such as grounding-line flux and cliff failure (Pollard, DeConto and Alley (2015)). Those tests and the reasonable agreement in additional limited offline tests at 80 km vs. finer resolutions indicate that the coarser resolution is adequate for this study.

We evaluate the PSU3D-ICE model over three separate time periods. As in previous ensemble work with this model (Chang et al. (2016a, 2016b), Pollard et al. (2016), DeConto and Pollard (2016)), the time periods are selected to include major ice-sheet variations that stringently test the model and have at least some paleo data to provide useful quantitative constraints. The three time periods are: (1) a single episode of high sea level rise during the warm mid-Pliocene (which extended roughly from 3.2 to 2.6 million years before present); (2) the last interglacial period around 125,000 to 115,000 years ago, at the start of the last Pleistocene glacial-interglacial cycle when global climate was slightly warmer than today, the major Northern Hemispheric ice sheets were most recently absent prior to the modern interglacial period, Greenland was smaller and the West Antarctic ice sheet may have undergone major collapse; and (3) the last deglacial period from the Last Glacial Maximum about 20,000 years ago to the present and then 5000 years into a warmer future. In Figure 1 we present 1500 model simulations from the PSU3D-ICE model for all three time periods as well as projections until year 2500. We describe the three model simulations below.

To represent a single high sea level episode during the warm mid-Pliocene era (roughly 3.2 to 2.6 million years before present), we initialize the ice-sheet model to modern conditions and run the model forward for 5000 years. As described in previous Pliocene applications (Pollard, DeConto and Alley (2015), Pollard, Gomez and Deconto (2017)), atmospheric climatic forcing is provided by the RegCM3 regional climate model (Pal et al. (2007)) adapted for polar regions and driven by the GENESIS v3 global climate model (Alder et al. (2011)). The atmospheric carbon dioxide concentration is set to at 400 parts per million by volume (ppmv), and a warm austral summer orbit is specified. We use oceanic temperatures from the modern World Ocean Atlas database (Levitus et al. (2012)), with a $+2°C$ uniform perturbation added to represent mid-Pliocene ocean warming. Atmospheric monthly cycles of surface air temperature and precipitation are used to compute melting and annual mass balance on the ice-sheet surface, and oceanic temperatures are used to compute basal melting under floating ice shelves (Pollard, DeConto and Alley (2015)).

For the Last Interglacial (LIG) we initialize the ice-sheet model to modern conditions and run the model from 130,000 to 120,000 years before present (130 ka to 120 ka). As described in DeConto and Pollard (2016), LIG climates are specified as uniform perturbations to modern climatology (Le Brocq, Payne and Vieli (2010) for atmosphere and Levitus et al. (2012) for ocean). The atmospheric and ocean temperature perturbations vary stepwise in

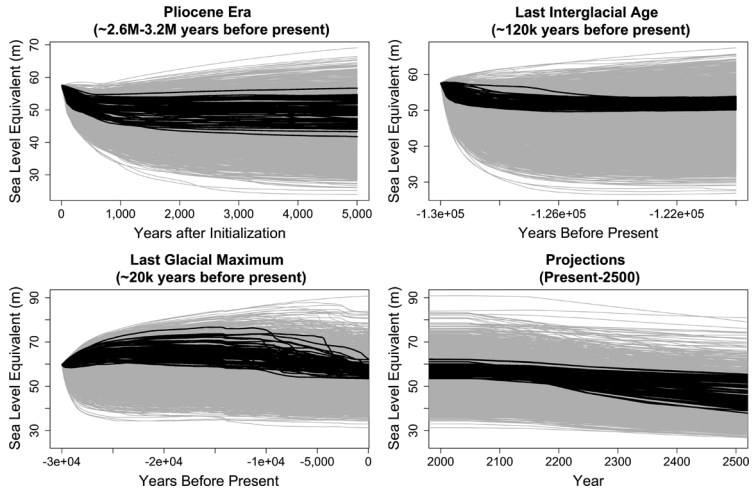

FIG. 1. *Time series of* 1500 *simulated model output from the PSU3D-ICE model where each model run corresponds to a line. Data are generated using* 1500 *parameter sets from the prior distribution. The y-axis denotes the Antarctic ice sheet's contribution to sea level change in meters* (*m*). *We approximate the present as year* 1950. *Model simulations that have a non-zero likelihood are denoted by black lines and runs that have a zero likelihood are displayed in light gray.* (*Top left*) *Model output for the Pliocene era model run where the x-axis denotes years after initialization.* (*Top right*) *Model output for the last interglacial age where the x-axis denotes years before the present.* (*Bottom left*) *Model output for the Last Glacial Maximum where the x-axis denotes years before the present.* (*Bottom right*) *Model projections for* 2000–2500 *where the x-axis represents the year.*

time. From 130 ka to 125 ka, they are +1.97°C and +1.70°C, respectively. From 125 ka to 120 ka, they are +1.41°C and +1.51°C, respectively.

The Last Glacial Maximum, modern, and future eras are simulated in one continuous run, over the last 40,000 years through the last deglacial period to modern, and extended 5000 years into the future. As described in Pollard et al. (2016), the model is initialized appropriately at 40 ka (40,000 years before present, or BP, relative to 1950 AD) from a previous long-term run. Atmospheric forcing is supplied using a modern climatological Antarctic dataset (Le Brocq, Payne and Vieli (2010)) with uniform cooling perturbations applied proportional to a deep sea-core δ^{18}O record (Pollard and DeConto (2009, 2012)). Oceanic forcing is supplied from a coupled atmosphere-ocean general circulation model (AOGCM) simulation of the last 20,000 years (Liu et al. (2009)). After reaching present day, each run is extended for 5000 years with atmospheric and oceanic forcing as described in DeConto and Pollard (2016), for the Representative Concentration Pathway (RCP) 8.5 scenario of future greenhouse gas emissions and concentrations (Meinshausen et al. (2011)), often called "business as usual." Atmospheric temperatures and precipitation are obtained by appropriately weighting previously saved simulations of the RegCM3 regional climate model for particular carbon dioxide levels, and oceanic temperatures are supplied from an archived transient NCAR global model simulation (Shields and Kiehl (2016)).

After each model run we extract the pertinent model output, specifically, the Antarctic ice sheet's contribution to sea level change (m), total ice volume (km^3) and total grounded ice area (km^2). We then compare this to the corresponding paleo- or modern observational records. In this study we examine 11 model parameters considered to be important in modeling the behavior of the Antarctic ice sheet—OCFACMULT, OCFACMULTASE, CRHSHELF, CRHFAC, ENHANCESHEET, ENHANCESHELF, FACEMELTRATE, TAUASTH, CLIFFVMAX, CALVLIQ and CALVNICK. Detailed descriptions of each parameter are provided in the Supplementary Material (Lee et al. (2020)).

We note that this is a much larger number of parameters than typically considered for models with such detailed dynamics. The ice-sheet model has many more parameters than the 11

chosen here. The values for many of them are reasonably well established in the glaciological literature, resulting from published work over the last several decades applying similar models to the Antarctic ice sheet. Those parameters mostly involve terrestrial processes (i.e., where ice is grounded on bedrock) that are constrained directly or indirectly by observational data of the modern ice sheet, and/or laboratory ice physics, such as the rheology of ice, ice streaming vs. shearing flow, basal sliding coefficients and modern ice distribution and thicknesses. The 11 parameters chosen here can have large effects on the results, but are not well constrained by modern observations because they apply to processes: (1) that have occurred in the past and expected in the future, but are not active today, or (2) are undergoing rapid change in recent decades. Examples of (1) are basal sliding coefficients for bedrock in modern ocean regions where grounded ice advanced during past glacial maxima and the timescale of bedrock rebound under varying ice loads. Examples of (2) are coefficients for oceanic melting at the base of floating ice shelves and oceanic melting at vertical ice fronts. A subset of these parameters have been used in more limited ensembles with this model (Chang et al. (2016b, 2016a), Pollard et al. (2016), Pollard, Gomez and Deconto (2017)), but here the 11 parameters constitute the bulk of important yet relatively unconstrained parameters in the model.

2.2. *Paleoclimate records and modern observations*. For the paleoclimate records we use the Antarctic ice sheet's contribution to sea level change in the following eras: Pliocene (~2.6–3.2 million years before present), the Last Interglacial Age (~125,000 to 115,000 years before present), and the Last Glacial Maximum (~20,000 years before present). We specify the Antarctic ice sheet's contribution to sea level change in terms of global mean sea level equivalents (SLE) relative to the modern ice sheet, thereby correctly allowing for marine ice grounded below sea level. The base units are meters (m). We adopt the following ranges for the paleoclimate records which account for considerable uncertainty in published estimates (cf. Kopp et al. (2009), Dutton et al. (2015)): (1) 5 m to 25 m for the Pliocene (Naish et al. (2009), Rovere et al. (2014), Cook et al. (2013)); (2) 3.5 m to 7.5 m for the Last Interglacial Age (Fuller, Wong and Keller (2017), DeConto and Pollard (2016)); and (3) -5 m to -15 m for the Last Glacial Maximum (Ruckert et al. (2017), Pollard et al. (2016)).

Modern observations include total volume and grounded area of the Antarctic ice sheet as well as 10 spatial locations that currently have ice present. Units for total volume and total grounded ice area are cubic kilometers (km^3) and square kilometers (km^2), respectively. Observations come from the Bedmap2 dataset (Fretwell et al. (2012)) which provide the most recent gridded maps of ice surface elevation, bedrock elevation and ice thickness. The Bedmap2 maps are generated using multiple sources, including satellite altimetry, airborne and ground radar surveys and seismic sounding.

3. Model calibration framework. In this section we describe the general computer model calibration framework. In computer model calibration key computer model parameters are estimated by comparing the computer model output and observational data (cf. Chang et al. (2016a), Kennedy and O'Hagan (2001), Bayarri et al. (2007), Bhat et al. (2010)). Calibration methods also account for key sources of uncertainty, such as model-observation discrepancy and observational error (Kennedy and O'Hagan (2001), Bayarri et al. (2007), Brynjarsdóttir and O'Hagan (2014)). We describe a model for output in the form of spatial data as this directly relates to our simulated data example in Section 5; a time series version of this applies to the PSU3D-ICE model in Section 6.

Let $Y(s, \theta)$ be the computer model output at the spatial location $s \in \mathcal{S} \subseteq \mathbb{R}^2$ and the parameter setting $\theta \in \Theta \subseteq \mathbb{R}^d$. $\mathcal{S}$ is the spatial domain of the process, and Θ is the parameter space of the computer model with integer d being the number of input parameters.

$\mathbf{Y} = (Y(s_1, \theta_i), \ldots, Y(s_n, \theta_i))^T$ is the computer model output at parameter setting θ_i and spatial locations $(s_1, \ldots, s_n)$. $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))^T$ is the observed spatial process at locations $(s_1, \ldots, s_n)$.

We model the observational data Z as follows

$$Z = Y(\theta) + \delta + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ is independently and identically distributed observational error and δ is a systemic data-model discrepancy term. The discrepancy δ is modeled as a zero-mean Gaussian process, where $\delta \sim N(0, \Sigma_\delta(\xi_\delta))$. This discrepancy term is essential for parameter calibration (Bhat et al. (2010), Bayarri et al. (2007)) and ignoring it may yield biased and overconfident estimates and projections (Brynjarsdóttir and O'Hagan (2014)). $\Sigma_\delta(\xi_\delta)$ is the spatial covariance matrix between spatial points $s_1, \ldots, s_n$ with covariance parameters ξ_δ. We set standard prior distributions for the model parameters, θ, and observational error variance, σ_ϵ^2. On the other hand, informative priors are necessary for the discrepancy term's covariance parameters ξ_δ. Then, we infer θ, σ^2 and ξ_δ by sampling from the posterior distribution, $\pi(\theta, \sigma_\epsilon^2, \xi_\delta | Z)$, via Markov chain Monte Carlo (MCMC).

Challenges with computer model calibration. We focus on a specific class of computer models, characterized by: (1) a moderate run time (six seconds to 15 minutes), and (2) moderately large parameter space (five to 20 parameters). The modified PSU3D-ICE Antarctic ice sheet model (Section 2) fits the specifications for this class of computer models. Several other important models that can potentially be modified to fit within this class are single column atmospheric models (Bony and Emanuel (2001), Dal Gesso and Neggers (2018), Gettelman et al. (2019)), simplified earth systems models (Monier et al. (2013)), hydrological soil moisture models (Sorooshian, Duan and Gupta (1993), Liang et al. (1994)) and integrated multisector models for human and earth dynamics (Kim et al. (2006)).

The calibration framework requires running the computer model once for each iteration of the MCMC algorithm. Subject to overall calibration wall times, MCMC-based calibration methods are well suited to computer models that run very quickly, typically under six seconds per model run. The PSU3D-ICE model takes approximately 10 to 15 minutes per run on a single 2.3-GHz Intel Xeon E5-2697V4 (Broadwell) processor. We estimate that a standard MCMC-based calibration approach for this would take on the order of 2.9 years to approximate the posterior distribution $\pi(\theta | Z)$.

Surrogate methods such as Gaussian process-based emulators are well suited to computer models with long run times. A good design is important for building accurate surrogates. Dense sampling schemes, such as full factorial or fractional factorial designs, capture higher order interactions; however, running the computer model at each of the design points is costly. Space-filling designs, such as the Latin Hypercube Design (McKay, Beckman and Conover (2000), Steinberg and Lin (2006), Stein (1987)), or adaptive experimental designs (Chang et al. (2016a), Gramacy and Apley (2015), Urban and Fricker (2010), Queipo et al. (2005)) use fewer design points but may possibly generate low-fidelity surrogate models by ignoring higher order interactions among inputs (Liu and Guillas (2017)). Since the PSU3D-ICE model exhibits nonlinear dependencies among input parameters (Pollard and DeConto (2012)), we would be limited to six or fewer parameters using standard emulation-calibration techniques (with our available computing resources).

4. Fast particle-based calibration. In this section we present a fast particle-based method to calibrate computers models with moderate model run time (six seconds to 15 minutes) and a moderate number of model parameters (five to 20). We begin with a description of a sequential sampling-importance-resampling algorithm. Then, we present modifications

to the algorithm designed to improve computational efficiency. We examine advantages and limitations of our approach. Finally, we discuss tuning mechanisms for our method and provide practical guidelines.

4.1. *Sequential sampling-importance-resampling with mutation.* We propose a series of sampling-importance-resampling with mutation operations which includes evolving importance and target distributions. The objective is to efficiently approximate a target distribution using a swarm of evolving particles. Our approach falls under the umbrella of sequential Monte Carlo algorithms (Del Moral, Doucet and Jasra (2006), Doucet, Godsill and Andrieu (2000), Liu and West (2001)) which have gained wide practical use (cf. Kantas et al. (2015), Papaioannou, Papadimitriou and Straub (2016), Kalyanaraman et al. (2016), Jeremiah et al. (2011), Morzfeld et al. (2018)). In particular, we build upon the iterated batch importance sampling (IBIS) (Chopin (2002), Crisan and Doucet (2000)) method.

Sampling-importance-resampling. Sampling-importance-resample (Gordon, Salmond and Smith (1993), Doucet, de Freitas and Gordon (2001)) is a sampling method used to approximate a target distribution $\pi(\theta)$ using samples from an importance distribution $q(\theta)$. Suppose we want to estimate $\mu = E_\pi[g(\theta)]$. Given $q(\theta) > 0$ whenever $g(\theta)\pi(\theta) > 0, \forall\theta \in \Theta$, we observe that $E_\pi[g(\theta)] = E_q[g(\theta)w(\theta)]$, where $w(\theta) = \frac{\pi(\theta)}{q(\theta)}$ is the importance weight and $\sum_{i=1}^N w(\theta_i) = 1$. The importance sampling estimator is $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^N g(\theta_i)w(\theta_i)$ and $\hat{\mu}_n \to \mu$ with probability 1 by the strong law of large numbers. For target distributions with an unknown normalizing constant, such as the posterior distribution of the model calibration parameters $\pi(\theta|Z)$, the importance weights $w(\theta_i)$, must be normalized.

An extension of importance sampling is sampling-importance-resampling, which provides an approximation of a target distribution via samples from an importance distribution and corresponding importance weights (Gordon, Salmond and Smith (1993)). The target distribution $\pi(\theta)$, is approximated by the empirical distribution of the samples $\hat{\pi}(\theta)$ and their corresponding normalized weights $\tilde{w}(\theta_i)$'s,

$$\pi(\theta) \approx \hat{\pi}(\theta) = \sum_{i=1}^{N} \tilde{w}(\theta_i)\delta(\theta_i),$$

where $\delta(\theta_i)$ is the Dirac measure that puts unit mass at θ_i and $\sum_{i=1}^N \tilde{w}(\theta_i) = 1$.

Poor choices of importance distributions may yield inaccurate approximations of the target distribution (Doucet, Godsill and Andrieu (2000)) due to weight degeneracy and sample impoverishment. As a result, the bulk of the resampled particles, θ_i, do not reside in the high-probability regions of $\pi(\theta)$. Weight degeneracy occurs when almost all of the samples drawn the importance function have near-zero importance weights, leaving just a few samples with any significant weights. Multinomial resampling using the normalized importance weights $\tilde{w}(\theta_i)$ can combat weight degeneracy by eliminating the particles with very small important weights and replicating those with higher weights (Gordon, Salmond and Smith (1993), Doucet, Godsill and Andrieu (2000)). After resampling, we reset all of the importance weights $w(\theta_i)$ to $1/N$ and replace the weighted empirical distribution $\hat{\pi}(\theta)$ with an unweighted empirical distribution $\ddot{\pi}(\theta)$,

$$\ddot{\pi}(\theta) = \frac{1}{N}\sum_{i=1}^{N} N_i\delta(\theta_i),$$

where N_i is the number of replicates corresponding to particle θ_i and $\sum_{i=1}^N N_i = N$.

Weight degeneracy can lead to sample impoverishment where a small subset of particles θ_i's are heavily replicated in the resampling step; hence, few unique particle remain. The

unweighted/resampled empirical distribution $\ddot{\pi}(\theta)$ may poorly approximate the true target distribution $\pi(\theta)$. To alleviate sample impoverishment, mixture approximations (Gordon, Salmond and Smith (1993)) or kernel smoothing methods (Liu and West (2001)) can mutate or rejuvenate the replicated particles. However, these methods may not scale well to high-dimensional target distributions (Doucet, Godsill and Andrieu (2000)).

An alternative method mutates the replicated particles with samples from $K(\theta_i^{(t-1)})$, the Metropolis–Hastings transition kernel (Gilks and Berzuini (2001)), whose stationary distribution is also the target distribution $\pi(\theta)$. Here we run J Metropolis–Hastings updates for each particle θ_i, for $i = 1, \ldots, N$. Other mutation schemes use genetic algorithms (Zhu et al. (2018)) or different transition kernels, $K(\cdot)$ (Papaioannou, Papadimitriou and Straub (2016), Murray, Lee and Jacob (2016)). The length of the Markov chain, J, will be short and dependent on computing resources. We set the jth sample drawn via MCMC as the mutated particle $\tilde{\theta}_i$. Since $\tilde{\theta}_i \sim \pi(\theta)$, the resulting empirical distribution $\tilde{\pi}(\theta)$ approximates the target distribution $\pi(\theta)$:

$$\pi(\theta) \approx \tilde{\pi}(\theta) = \sum_{i=1}^{N} \tilde{\theta}_i \delta(\tilde{\theta}_i).$$

Even with the mutation step, sampling-importance-resampling may incur large computational costs. Poor choices of importance distributions may result in extreme sample impoverishment, due to the large discrepancy between the importance and target distribution. Here, the mutation stage typically requires very long (and costly) chains of the Metropolis–Hastings algorithm to move the particles into the high-probability regions of the target distribution (Li et al. (2014)).

Sequential sampling-importance-resampling. Our fast particle-based approach addresses the limitations noted above. We propose a series of intermediate posterior distributions $\pi_t(\theta|Z)$, for $t = 1, \ldots, T$ which will act as importance and target distributions. Existing methods use intermediate posterior distributions for parameter estimation of static systems (Chopin (2002), Papaioannou, Papadimitriou and Straub (2016), Nguyen (2014)), uncertainty quantification for chemical processes (Kalyanaraman et al. (2016)) and calculating maximum-likelihood estimates for hierarchical systems (Lele, Dennis and Lutscher (2007)).

Intermediate posterior distributions can be generated using likelihood tempering (Chopin (2002), Neal (2001), Liang and Wong (2001)). For each intermediate posterior distribution $\pi_t(\theta|Z)$, the likelihood component is a fractional power of the original likelihood $p(Z|\theta)$. The tth intermediate posterior distribution, $\pi_t(\theta)$, is generated as follows:

$$\pi_t(\theta|Z) \propto p(Z|\theta)^{\gamma_t} p(\theta), \tag{2}$$

where γ_t's are determined according to a schedule where $\gamma_0 = 0 < \gamma_1 < \cdots < \gamma_T = 1$.

For cycle $t = 1$, we set the importance distribution to be the prior distribution, $p(\theta)$, and the target distribution to be the first intermediate posterior distribution, $\pi_1(\theta|Z)$. For cycle t the importance distribution is $\pi_{t-1}(\theta|Z)$, and the target distribution is $\pi_t(\theta|Z)$. Note that the likelihood incorporation schedule need not be uniform. For instance, more of the likelihood can be incorporated into the earlier intermediate posterior distributions.

Finally, we mutate the particles via short runs of the Metropolis–Hastings algorithm where the stationary distribution is $\pi_t(\theta)$, the tth intermediate posterior distribution. Note that the importance and target distributions are consecutive (tth and $t + 1$th) intermediate posterior distributions, so there is considerable overlap between the high-probability regions of the two distributions. Convergence results for this family of Sequential Monte Carlo algorithms are provided in Crisan and Doucet (2000), Beskos et al. (2016) and Giraud and Del Moral (2017).

4.2. *Stopping criterion.* We present a stopping rule that controls the number of Metropolis–Hastings updates within the mutation step. This provides an automatic heuristic determining when to stop the mutation stage, and it can also eliminate unnecessary computer model runs. The stopping rule is based on the Bhattacharyya distance (Bhattacharyya (1946)), $D_B(p, q)$, which measures the similarity between two distributions, $p(\theta)$ and $q(\theta)$. We first evaluate the stopping criterion after $2k$ Metropolis–Hastings updates; if the criterion is not met, then we reevaluate after k subsequent updates.

Consider $\theta_t^{i,k}$, the ith particle, or parameter setting, after the kth mutation step of the Metropolis–Hastings algorithm during cycle number t. Let $\boldsymbol{\theta}_t^k = \{\theta_t^{1,k}, \ldots, \theta_t^{n,k}\}$ denote the set of parameters $\theta_t^{i,k}$'s. Let $h(\theta_t^{i,k})$ be the target metric of interest evaluated at parameter setting $\theta_t^{i,k}$, in this case, the Antarctic ice sheet contribution to sea level change in 2100. Let $\mathbf{h}(\boldsymbol{\theta}_t^k) = \{h(\theta_t^{1,k}), \ldots, h(\theta_t^{n,k})\}$ denote the set of target metrics $h(\theta_t^{i,k})$'s.

At mutation update $2k$ we partition the range spanned by two sets of target metric samples—$\mathbf{h}(\boldsymbol{\theta}_t^k)$ and $\mathbf{h}(\boldsymbol{\theta}_t^{2k})$—into m nonoverlapping blocks of equal width. Then, we compute the real-valued Bhattacharyya distance $D_B(\mathbf{h}(\boldsymbol{\theta}_t^k), \mathbf{h}(\boldsymbol{\theta}_t^{2k})) = -\ln(\sum_{i=1}^n \sqrt{p_i q_i})$ where p_i and q_i are the proportion of samples, from $\mathbf{h}(\boldsymbol{\theta}_t^k)$ and $\mathbf{h}(\boldsymbol{\theta}_t^{2k})$, respectively, that lie within the ith partition. The mutation step proceeds until $D_B(h(\theta_k), h(\theta_{2k})) < \epsilon_B$, the stopping criterion. If the stopping criterion is not fulfilled, we run k additional Metropolis–Hastings updates and evaluate the stopping criterion at iterations $3k$ and $2k$. We repeat this until the stopping criterion is met. We obtain the threshold ϵ_{BD} through a Monte Carlo simulation run prior to the calibration study. Section 4.4 discusses tuning for k, ϵ_B, and m.

4.3. *Adaptive incorporation schedule.* In equation 2 we introduce a standard incorporation schedule $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_T)$. In the standard implementation the user must select the total number of sampling-importance-resampling cycles (T) and the likelihood incorporation increments γ_t for $t = (0, \ldots, T)$. Past research proposed novel methods to adaptive choose the incorporation schedule, γ_t, yet maintain a constant number of cycles, T (Nguyen (2014), Kalyanaraman et al. (2016)). Here, we introduce an adaptive incorporation schedule that automatically determines both the total number of sampling-importance-resampling cycles, T, and incorporation schedule, $\boldsymbol{\gamma}$. Introducing the adaptive incorporation schedule into the particle-based calibration framework provides computational and practical benefits by: (1) reducing the number of computer model evaluations, (2) decreasing the overall calibration wall times and (3) simplifying implementation for the user.

The adaptive incorporation schedule proceeds as follows. On initialization we set the initial incorporation increment γ_0 to 0. We draw the initial set of particles θ_0 from $\pi_0(\theta|Z) \propto L(\theta|Z)^0 p(\theta) = p(\theta)$, the prior distribution of model parameters. For cycle $t = 1, 2, 3, \ldots$, we calculate the full likelihood $L(\theta_{t-1}^{(i)}|Z)$ for $i = 1, \ldots, N$ where $\theta_{t-1}^{(i)}$ denotes the parameter samples from the previous cycle $t - 1$. For computational efficiency, we reuse the likelihood evaluations from the previous cycle. Next, we find the optimal γ_t that returns an effective sample size (ESS) of $\text{ESS}_{\text{thresh}}$ or a sample size closest to $\text{ESS}_{\text{thresh}}$: $\gamma_t = \text{argmin}_\gamma\{(\text{ESS}_\gamma - \text{ESS}_{\text{thresh}})^2\}$, where $\gamma \in (\gamma_{\min}, 1 - \gamma_{t-1})$, $\gamma_{\min}$ is a previously set minimum incorporation value, $\text{ESS}_\gamma = \sum_{i=1}^N \frac{1}{w_t^{(i)2}}$ and $w_t^{(i)} \propto L(\theta_t^{(i)}|Z)^\gamma$. Note that we can lower computational costs by evaluating the full likelihood $L(\theta_0^{(i)}|Z)$ only once before the optimization.

We stop the scheduling algorithm when $\sum_{i=1}^t \gamma_t = 1$. This occurs when the entire likelihood has been incorporated, and the target distribution has evolved to the full posterior distribution $\pi(\theta|Z)$. Note at each cycle t, we set the incorporation increment (γ_t) to be between $\gamma_{\min}$ and $1 - \sum_{i=1}^t \gamma_t$. In Section 4.4 we describe how to set the minimum incorporation increment $\gamma_{\min}$ and the threshold effective sample size, $\text{ESS}_{\text{thresh}}$.

Adaptive likelihood incorporation schedule:

1. Initialization: At $t = 0$, set $\gamma_0 = 0$.
2. When $t > 0$ and $\sum_{i=1}^{t-1} \gamma_i < 1$:
   - Compute $L(\theta_{t-1}^{(i)}|Z)$ for $i = 1, \ldots, N$
   - Set $\gamma_t = \text{argmin}_\gamma\{(\text{ESS}_\gamma - \text{ESS}_{\text{thresh}})^2\}$, where $\text{ESS}_\gamma = \sum_{i=1}^{N} \frac{1}{w_t^{(i)2}}$, $w_t^{(i)} \propto L(\theta_t^{(i)}|Z)^\gamma$, and $\gamma \in (\gamma_{\min}, 1 - \gamma_{t-1})$.
   - $\gamma_{\min}$ is a predetermined minimum incorporation value
3. When $t > 0$ and $\sum_{i=1}^{t-1} \gamma_i = 1$: Stop calibration

4.4. *Tuning the algorithm.* Much of Algorithm 1 is automated. However, the user needs to choose: (1) the total number of particles, N, (2) the number of Metropolis–Hastings updates run before checking the stopping criterion, K, (3) the minimum incorporation $\gamma_{\min}$ and (4) the effective sample size threshold $\text{ESS}_{\text{thresh}}$. (1) and (2) should be set based on the amount of available computational resources, but our simulation study results favor having more particles N than longer Metropolis–Hastings updates K. We chose 2015 particles, which requires 56 nodes with 36 processors per node; thereby leaving one processor to execute master tasks. We set the reference length k for the Metropolis–Hastings updates to be 7. Based on simulation experiments, the empirical distribution of particles stabilize after 10 to 15 updates. In this study we set the floor for the incorporation increment, $\gamma_{\min}$, to be 0.1, so that at each cycle the weights for the importance sampling step is at least $L(\theta|Z)^{0.1}$.

The automated likelihood tempering schedule (Section 4.3) ensures that the effective sample size (ESS) of the final particles does not fall below a predetermined threshold $\text{ESS}_{\text{thresh}}$. For moderate-dimensional parameter spaces (five to 20), the effective sample size is important as it is an indicator of the discrepancy between the true target distribution and the particle-based empirical distribution (cf. Doucet, de Freitas and Gordon (2001), Gordon, Salmond and Smith (1993)). A low ESS suggests that only a few particles have any significant weight, and it is often indicative of weight degeneracy and a poor approximation of the target distribution (Kong (1992)). A suitable ESS can be obtained by minimizing ρ, the second moment of the Radon–Nikodym derivative between the target and the proposal distribution (Whiteley, Lee and Heine (2016), Kong (1992)), generating more sophisticated proposal distributions via implicit sampling (Morzfeld et al. (2015)) and examining distances between target and proposal distributions within an intrinsic dimension (Agapiou et al. (2017)). Other research (Martino, Elvira and Louzada (2017)) points to alternative definitions of the ESS than the traditional method based on the variance of the weights (Liu and Chen (1998)). In this study we utilize the common definition of ESS (Kong (1992)), which is based on the variance of the importance weights, and we set $\text{ESS}_{\text{thresh}} = \frac{N}{2}$, which is the typical threshold used by many sequential Monte Carlo methods (Del Moral, Doucet and Jasra (2006)) prior to resampling.

We obtain ϵ_{BD} as follows. Prior to running Algorithm 1, we obtained samples of a target metric (Antarctic Ice Sheet contribution to sea level rise in 2100) from an initial survey of computer model runs. Let μ and σ^2 denote the sample mean and variance of the target metric mentioned above. We generate a collection of B samples of size n, denoted as $\mathbf{x} = \{x_1, \ldots, x_B\}$. Here, $x_b \sim \mathcal{N}(\mu, \sigma^2)$, with μ and σ^2 previously defined. Let $x_{\text{base}} \sim \mathcal{N}(\mu, \sigma^2)$ be a baseline sample for calculating the Bhattacharyya distance. We calculate $D_B(x_b, x_{\text{base}})$ for $b = 1, \ldots, B$ and set ϵ_{BD} to be the 0.975 quantile. In this study we chose $B = 1000$ and the number of partitions $m = 200$.

We calibrate the PSU3D-ICE model using Cheyenne (Computational and Information Systems Laboratory (2017)), a 5.34-petaflops high-performance computer operated by the Na-

Algorithm 1: Fast particle-based calibration

Data: Z
Initialization:
Draw $\theta_0^{(i)} \sim p(\theta)$ for particles $i = 1, \ldots, N$.
Set $w_0^{(i)} = 1/N$, $\gamma_0 = 0$, and K;
for *cycles* $t = 1, \ldots, T$ **do**
1. Compute full likelihood:
Calculate $L(\theta_{t-1}^{(i)}|Z)$ for $i = 1, \ldots, N$;
2. Select optimal likelihood incorporation increment γ_t:
Set $\gamma_t = \text{argmin}_\gamma\{(\text{ESS}_{\gamma_t} - \text{ESS}_{\text{thresh}})^2\}$, where $\gamma \in (\gamma_{\min}, 1 - \sum_{i=1}^{t-1} \gamma_{t-1})$
Note: $\text{ESS}_{\gamma_t} = \sum_{i=1}^{N} \frac{1}{w_t^{(i)2}}$ and $w_t^{(i)} \propto L(\theta_t^{(i)}|Z)^{\gamma_t}$;
3. Compute importance weights:
$w_t^{(i)} \propto w_{t-1}^{(i)} \times L(\theta_t^{(i)}|Z)^{\gamma_t}$;
4. Resample particles:
Draw $\theta_t^{(i)}$ from $\{\theta_{t-1}^{(1)}, \ldots, \theta_{t-1}^{(N)}\}$ with probabilities $\propto \{w_t^{(1)}, \ldots, w_t^{(N)}\}$;
5. Set intermediate posterior distribution:
Set $\pi_t(\theta|Z) \propto L(\theta_i|Z)^{\tilde{\gamma}} \pi(\theta)$, where $\tilde{\gamma} = \sum_{j=1}^{t} \gamma_j$;
6. Mutation:
Using each particle $(\theta_t^{(1)}, \ldots, \theta_t^{(N)})$ as the initial value, run N chains of an MCMC algorithm with target distribution $\pi_t(\theta|Z)$ for $2K$ iterations;
7. Check stopping criterion:
Compute $\delta_B = D_B(h(\theta_t^K), h(\theta_t^{2K}))$;
if $\delta_B < \epsilon_B$, **then**
Set $\theta_t^{(i)} = \theta_t^{(i),2K}$;
else
Run K additional updates and re-evaluate stopping criterion
Continue until stopping criterion is met
8. Stop when full likelihood is incorporated;
if $\sum_{i=1}^{N} \gamma_t = 1$, **then**
End Algorithm
else
Reset weights: $w_t^{(i)} = 1/N$ for particles $i = 1, \ldots, N$;
Set $t = t + 1$, and return to Step 1.

tional Center for Atmospheric Research (NCAR). Parallelized operations, such as calculating importance weights and mutation, proceed via message passing interface (MPI). To limit communication costs, we build the ice sheet model and load the relevant datasets separately on each processor.

4.5. *Computational advantages and limitations*. We take advantage of the embarrassingly parallel nature of the importance sampling and mutation steps to reduce wall time. In our approach, the Metropolis–Hastings updates in the mutation stage are the primary drivers of computational cost. To address this cost, we propose an automated stopping rule for the mutation stage. We also introduce an adaptive likelihood incorporation schedule that automatically selects an efficient number of sampling-importance-resampling cycles. The stopping rule and adaptive likelihood incorporation schedule simplifies implementation for the user (due to automation) and reduces the number of computer model runs needed for calibration.

Our approach is a viable alternative to existing calibration methods which may be computationally infeasible. MCMC-based calibration methods using the computer model is computationally prohibitive due to the sequential nature of MCMC algorithms. Emulation-calibration methods, while efficient for expensive computer models, do not easily scale to problems with many parameters (say more than five or six for this model). Also, multiple-try MCMC methods (Liu, Liang and Wong (2000)), a mixture of importance sampling and MCMC, may incur large costs because several parallel processes must be initialized and terminated at each iteration of the MCMC chain. Multiple-try MCMC may experience slow mixing, especially when the Markov chain moves to the low-probability regions of the target distribution distribution (Martino (2018)).

While our method has many computational advantages, we note that the heavy parallelization in our approach requires access to and the ability to work with high performance computing resources. Given our current computing resources, our method is ideally suited to models that run between six seconds and 15 minutes. For models with longer run times, the computational costs remain prohibitive. MCMC algorithms may be feasible and simpler to implement for models with shorter run times. As is the case with parallel computing methods, communication costs must be small relative to the computer model run times; otherwise, we would not reap the benefits of our approach.

5. Simulated example and results. In this section we calibrate a simple computer model using three different methods. We simulate a data set of size $n = 300$ where the spatial locations s_i for $i = 1, \ldots, n$ are in the unit domain $[0, 1]^2$. We generate the data via a modified version of the computer model presented in Bayarri et al. (2007). We construct a simple computer model as follows:

$$Y(s_i, \theta) = 5 \times \exp\{-\theta(\text{lat}_i \times \text{lon}_i)\},$$

where $Y(s_i, \theta)$ is a real-valued computer model output at model parameter setting θ and at a spatial location specified by lat_i and lon_i which represent the latitude and longitude of the ith location. The true process includes a data-model discrepancy term $\delta(s_i)$ which is defined as $\delta(s_i) = -1.5 \times (\text{lat}_i \times \text{lon}_i)$ and i.i.d. observational error $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. For this example we set $\theta = 1.7$ and $\sigma_\epsilon^2 = 0.5$. To generate the observational data, $Z(s_i)$, we combine the computer model output $Y(s_i, \theta)$, the data-model discrepancy, $\delta(s_i)$ and the observational error, ϵ_i, as follows:

$$Z(s_i) = Y(s_i, \theta) + \delta(s_i) + \epsilon_i.$$

We model the observations as

$$Z(s_i) = 5 \times \exp\{-\theta(\text{lat}_i \times \text{lon}_i)\} + \delta(s_i) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ are the i.i.d. observational errors. Since the actual form of the discrepancy term is unknown, we model the discrepancy, $\delta(s_i)$, as a zero-mean Gaussian process, $\delta(s) \sim \mathcal{GP}(0, \Sigma_\delta(\xi_\delta))$, where ξ_δ is a vector containing the covariance parameters. To allow for some roughness of the process between spatial locations, we choose an exponential covariance function $\Sigma_\delta(\xi_\delta) = \sigma_\delta^2 \exp(-\frac{|s_i - s_j|}{\phi_\delta})$ with $\xi_\delta = (\phi_\delta, \sigma_\delta^2)$. To complete the Bayesian framework, we use the prior distributions $\theta \sim \mathcal{N}(0, 100)$, $\sigma_\epsilon^2 \sim \mathcal{IG}(2, 2)$, $\phi_\delta \sim \mathcal{U}(0.01, 1.5)$ and $\sigma_\delta^2 \sim \mathcal{IG}(2, 2)$.

We compare results from three calibration methods: (1) MCMC-based, (2) standard particle-based and (3) adaptive particle-based. In the MCMC-based method, we generated 100,000 samples from $\pi(\theta, \phi_\delta, \sigma_\delta^2, \sigma_\epsilon^2 | Z)$ via the Metropolis–Hastings algorithm. Next, the standard and adaptive particle-based calibration methods use $N = 2000$ particles to approximate $\pi(\theta, \phi_\delta, \sigma_\delta^2, \sigma_\epsilon^2 | Z)$. For the standard particle-based method, we set the total number of cycles to be 10 and establish a uniform likelihood incorporation $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{10})$,

TABLE 1
Simulated example calibration results for three calibration approaches: (1) *Adaptive particle-based*; (2) *Standard particle-based*; *and* (3) *MCMC-based*. *All three approaches yield comparative results*

	θ	ϕ_δ	σ_δ^2	σ_ϵ^2
Adaptive Particle-Based (Est)	2.04	1.22	0.78	0.44
Adaptive Particle-Based (95% CI)	(1.05, 3.14)	(0.83, 1.50)	(0.36, 1.32)	(0.36, 0.52)
Standard Particle-Based (Est)	2.04	1.22	0.80	0.44
Standard Particle-Based (95% CI)	(1.03, 3.11)	(0.81, 1.50)	(0.32, 1.33)	(0.35, 0.51)
MCMC-Based (Est)	2.04	1.21	0.79	0.44
MCMC-Based (95% CI)	(1.06, 3.17)	(0.80, 1.50)	(0.34, 1.33)	(0.36, 0.52)

where $\gamma_t = 0.1$ for $t = 1, \ldots, 10$. We run $K = 100$ Metropolis–Hastings updates for each mutation cycle. In the adaptive particle-based calibration approach, our algorithm automatically chose four cycles with a likelihood incorporation schedule $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0.100, 0.148, 0.2743, 0.4777)$ using the adaptive likelihood incorporation schedule. For each mutation step we run batches of $K = 10$ Metropolis–Hastings updates until the stopping criterion is met.

All three methods yield comparable calibration results (see Table 1); however, our adaptive particle-based approach exhibits a considerable speedup in computation. For the model parameter, θ, calibration via MCMC (the "gold standard") provides estimate $\hat{\theta}_{\text{mcmc}} = 2.04$ and 95% credible interval bounds $(1.06, 3.17)$. Similarly, the standard particle-based approach generates estimate $\hat{\theta}_{\text{std}} = 2.04$ with 95% credible interval bounds $(1.03, 3.11)$, and the adaptive particle-based approach yields estimate $\hat{\theta}_{\text{adapt}} = 2.04$ with 95% credible interval bounds $(1.05, 3.14)$. The adaptive particle-based approach has considerably shorter wall times due to fewer computer model evaluations. To illustrate, the adaptive approach requires just $4 \times 20 = 80$ sequential computer model runs, as opposed to $10 \times 200 = 2000$ runs for the standard particle-based approach and 100,000 for the MCMC-based approach.

6. Application to the PSU3D-ICE model. Here, we provide specifics for calibrating the PSU3D-ICE model and discuss how our method provides key computational benefits over existing calibration approaches. We also summarize results from a comparative analysis of three calibration methods within the context of the PSU3D-ICE model. The efficiency of our computational approach allows us to study the effect of observations from the Pliocene on parameter calibration and projections of sea level rise and also enables us to conduct a prior sensitivity analysis.

6.1. *Calibrating PSU3D-ICE*. We calibrate 11 model parameters using both paleoclimate records and modern observations from satellite imagery (Section 2). For the paleoclimate records, modern volume and modern grounded ice area, we use independent truncated normal distributions. The upper and lower ranges for the truncated normal likelihood functions are based on domain area expertise and past studies (Section 2.2).

We calibrate the PSU3D-ICE model using five observations: (1) Z_{plio}, the Antarctic ice sheet's contribution to sea level change (m) in the Pliocene, (2) Z_{lig}, contribution in the Last Interglacial Age (m), (3) Z_{lgm}, contribution in the Last Glacial Maximum (m), (4) Z_{vol}, the Antarctic ice sheet's total ice volume in the modern era (km^3) and (5) Z_{area}, total grounded ice area in the modern era (km^2). We also use observations of ice occurrence taken at 10 strategic point in the Antarctic ice sheet. Here, $Z_{\text{spat}} = (Z_{\text{spat},1}, \ldots, Z_{\text{spat},10})$. All 10 locations have ice presence; so, $Z_{\text{spat},i} = 1$ for locations $i = 1, \ldots, 10$.

Likelihood. For the observational dataset, $\mathbf{Z} = (Z_{\text{plio}}, Z_{\text{lig}}, Z_{\text{lgm}}, Z_{\text{vol}}, Z_{\text{area}}, Z_{\text{spat},1}, \ldots, Z_{\text{spat},10})$, we define a likelihood function using truncated normal distributions and indicator functions. For the modern volume, modern total grounded area and paleoclimate records, we use independent truncated normal distributions as the observational model. $TN(\mu, \sigma^2, \alpha, \beta)$ denotes a truncated normal distributions with the mean (μ), variance (σ^2), upper bound (α) and lower bound (β):

$$
\begin{aligned}
Z_{\text{plio}} &\sim \mathcal{TN}\big(\mu = Y(\theta)_{\text{plio}}, \sigma^2 = 30^2, \alpha = Y(\theta)_{\text{plio}} - 10, \beta = Y(\theta)_{\text{plio}} + 10\big),\\
Z_{\text{lig}} &\sim \mathcal{TN}\big(\mu = Y(\theta)_{\text{lig}}, \sigma^2 = 10^2, \alpha = Y(\theta)_{\text{lig}} - 2, \beta = Y(\theta)_{\text{lig}} + 2\big),\\
Z_{\text{lgm}} &\sim \mathcal{TN}\big(\mu = Y(\theta)_{\text{lgm}}, \sigma^2 = 20^2, \alpha = Y(\theta)_{\text{lgm}} - 5, \beta = Y(\theta)_{\text{lig}} + 5\big),\\
Z_{\text{vol}} &\sim \mathcal{TN}\big(\mu = Y(\theta)_{\text{vol}}, \sigma^2 = 1.6 \times 10^{15}, \alpha = Y(\theta)_{\text{vol}} - 2.5 \times 10^{15},\\
&\qquad \beta = Y(\theta)_{\text{vol}} + 2.5 \times 10^{15}\big),\\
Z_{\text{area}} &\sim \mathcal{TN}\big(\mu = Y(\theta)_{\text{area}}, \sigma^2 = 0.6 \times 10^{12}, \alpha = Y(\theta)_{\text{area}} - 1.5 \times 10^{12},\\
&\qquad \beta = Y(\theta)_{\text{area}} + 1.5 \times 10^{12}\big).
\end{aligned}
$$

The second set of observations are binary occurrences of ice at 10 strategically placed points on the Antarctic ice sheet (Lee et al. (2020)). For these observations we use indicator functions as the observational model as follows:

$$
Z_{\text{spat}} \sim \prod_{i=1}^{10} \mathbb{I}\big(Y(\theta)_{\text{spat},i} = Z_{\text{spat},i}\big),
$$

where $Y(\theta)_{\text{spat},i}$ denotes the model spatial output for a model run using parameters θ.

Priors. We set the prior distributions for the 11 model parameters based on expert knowledge. Five model parameters—CALVNICK, TAUASTH, CALVLIQ, CLIFFVMAX, FACEMELTRATE—have uniform prior distributions. Here, $\theta \sim U(\alpha, \beta)$, where α and β denote the upper and lower bounds of the uniform distribution. The prior distributions are as follows:

- $\theta_{\text{CALVNICK}} \sim \mathcal{U}(0, 2)$
- $\theta_{\text{TAUASTH}} \sim \mathcal{U}(1000, 5000)$
- $\theta_{\text{CALVLIQ}} \sim \mathcal{U}(0, 200)$
- $\theta_{\text{CLIFFVMAX}} \sim \mathcal{U}(0, 12{,}000)$
- $\theta_{\text{FACEMELTRATE}} \sim \mathcal{U}(0, 20)$

Six parameters—OCFACMULT, OCFACMULTASE, CRHSHELF, ENHANCESHEET, ENHANCESHELF, CRHFAC—have log-uniform prior distributions. Here, $\theta \sim LU(\text{base}, \alpha, \beta)$, which implies $\log_{\text{base}}(\theta) \sim U(\alpha, \beta)$ where α and β denote the upper and lower bounds of the uniform distribution. The prior distributions are as follows:

- $\log_{10}(\theta_{\text{OCFACMULT}}) \sim \mathcal{U}(-0.5, 0.5)$,
- $\log_{10}(\theta_{\text{OCFACMULTASE}}) \sim \mathcal{U}(0, 1)$,
- $\log_{10}(\theta_{\text{CRHSHELF}}) \sim \mathcal{U}(-7, -4)$,
- $\log_{10}(\theta_{\text{ENHANCESHEET}}) \sim \mathcal{U}(-1, 1)$.
- $\log_{0.3}(\theta_{\text{ENHANCESHELF}}) \sim \mathcal{U}(-1, 1)$,
- $\log_{10}(\theta_{\text{CRHFAC}}) \sim \mathcal{U}(-2, 2)$.

We can estimate the data-model discrepancy as an additive model bias, $\alpha \in \mathbb{R}$, such that our observational model (1) is modified to be $Z = Y(\theta) + \alpha + \epsilon$. For observations that are discontinuous in time, past ice-sheet calibration studies (Edwards et al. (2019), Williamson et

al. (2013), Ruckert et al. (2017)) model the discrepancy term as a tolerance to the observation measurement error which follows the zero-mean Gaussian process framework provided in Kennedy and O'Hagan (2001). For the PSU3D-ICE model we find that calibration with and without the discrepancy term yields very similar results.

6.2. *Computational benefits of our approach*. Our adaptive particle-based approach greatly reduces calibration wall times compared to using an all-at-once random-walk Metropolis–Hastings algorithm as in past ice-sheet calibration studies (cf. Ruckert et al. (2017), Bakker, Applegate and Keller (2016), Petra et al. (2014)). Our fast calibration approach had a total wall-time of ~ 6.5 hours and evolved 2015 particles for an effective sample size (ESS) of 1533. For the MCMC-based calibration approach, it would be computationally prohibitive to generate a large enough sample with a similar ESS. Instead, we estimate the time to generate an ESS of 1533. We ran the Metropolis–Hastings algorithm for 12 days to generate 1500 samples. We calculated the effective sample size per hour (Jones et al. (2006)) for each model parameter and then projected the time required to obtain an ESS of 1533, the ESS from the particle-based approach. It would require 12 to 18 months running the Metropolis–Hastings algorithm to generate the same ESS as our particle-based approach.

The computing times are based on the PSU3D-ICE model run at 80 km spatial resolution and an adaptive temporal resolution with a baseline timestep of eight years. Run times are for the NCAR Cheyenne HPC system with 2.3-GHz Intel Xeon E5-2697V4 Broadwell processors. Note that in practice, computation times for the particle-based methods can be slightly higher due to initialization and communications costs inherent to parallelized computing. Reduction of initialization and communication costs is an active area of research with novel methods in development (Ballard, Siefert and Hu (2016), Fan et al. (2018)). Note that the computation times for the MCMC-based approach are quite optimistic as we initialized the Markov chain and set the proposal distribution using all samples generated from our particle-based approach. In general, MCMC algorithms would not have access to these particles and would, therefore, likely require even more iterations of the MCMC algorithm to achieve the desired ESS.

6.3. *Comparisons to other calibration approaches*. We conduct a comparative study between our particle-based calibration approach and competing emulation-calibration methods (see Supplementary Material (Lee et al. (2020)) for details). We calibrate the PSU3D-ICE model using three methods:

1. A low-dimensional emulation-calibration approach: This approach varies only three parameters—OCFACMULT, CALVLIQ and CLIFFVMAX—and fixes the remaining eight parameters at scientifically justified values provided by our expert on ice sheets (DP). We include this approach because reducing the number of parameters is a common way to address computational challenges associated with calibration with long model run times (e.g., Edwards et al. (2019), Chang et al. (2014), Sacks et al. (1989)). We chose these three parameters because they are considered to be important in modeling the long-term evolution of the Antarctic ice sheet (Edwards et al. (2019), DeConto and Pollard (2016)). We train a Gaussian process emulator using 512 design points and use the squared exponential covariance function to represent the dependence between the design points. For the experimental design we use a full factorial design with eight equally spaced points for each model parameter.
2. A high-dimensional emulation-calibration approach: This approach calibrates all 11 selected parameters of the PSU3D-ICE model. We train a Gaussian process emulator using 512 design points generated via Latin hypercube design (LHC). Similar to the low-dimensional case, we use a squared exponential covariance function to model the dependence between design points. Emulation and calibration details are provided in the Supplementary Material (Lee et al. (2020)).

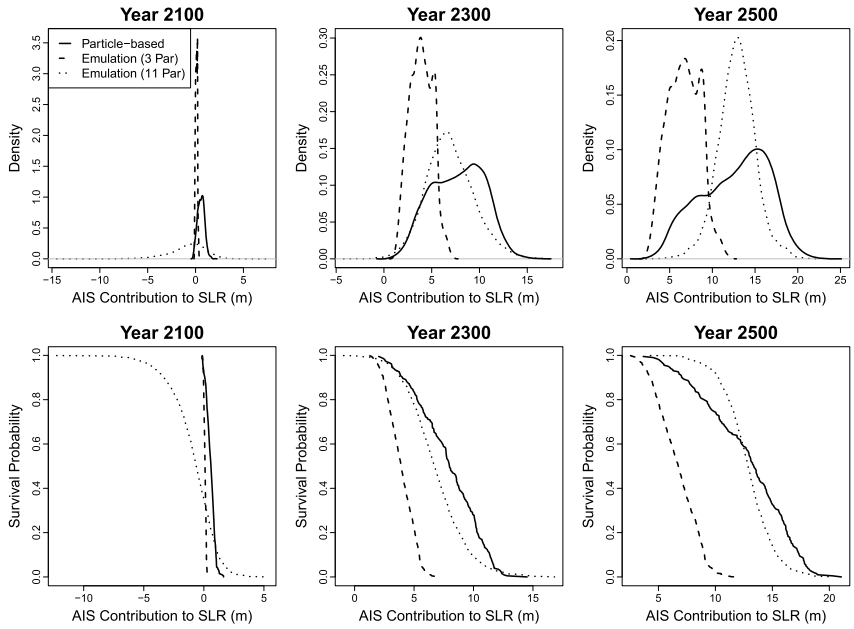

FIG. 2. (*Top Panel*) *Posterior densities of the projected Antarctic ice sheet's contribution to sea level change in* 2100, 2300, *and* 2500 *using the adaptive particle-based approach* (*solid line*), *emulation-calibration with three parameters* (*dashed line*), *and emulation-calibration with* 11 *parameters* (*dotted line*). (*Bottom Panel*) *Empirical survival functions of the projected Antarctic ice sheet's contribution to sea level change in* 2100, 2300, *and* 2500 *using the adaptive particle-based approach* (*solid line*), *emulation-calibration with three parameters* (*dashed line*), *and emulation-calibration with* 11 *parameters* (*dotted line*). *Three-parameter emulation results in sharper densities centered on distinctively lower point estimates. The* 11*-parameter emulation-calibration approach results in highly uncertain projections.*

3. Our particle-based approach: We use our heavily parallelized particle-based approach to calibrate all 11 selected parameters.

For the first method, we find that by fixing eight of 11 parameters we greatly constrain the parameter space and, thereby, underestimate the parametric uncertainty underlying the ice sheet model. Projections for the Antarctic sea level contribution in 2100–2500 are much lower and overconfident compared to those from our particle-based approach (Figure 2). For the second method, the limited amount of design points (training data) generates an inaccurate surrogate model as shown by the large out-of-sample cross-validated mean squared prediction error (Supplementary Material (Lee et al. (2020))). This calls into question the parameter estimates as well as the resulting projections. As shown in Figure 3, the second approach produces extremely sharp posterior distributions for two key model parameters, CLIFFVMAX and TAUASTH, which is inconsistent with the parameter estimates from the particle-based approach.

Figure 4 compares the posterior densities of projections and hindcasts for the three-parameter emulation-calibration approach and our 11-parameter particle-based method. Note that the three-parameter emulation-calibration approach (striped blue shading) underestimates the tail-area risk, or the 99th% quantile, for sea level projections compared to our approach (striped red shading). By calibrating more parameters, we can expect the tail-area risk to increase by a factor of 74 in 2100 and 65 in 2300.

The three-parameter emulation-calibration required 1.5 minutes to fit the Gaussian process emulator using 12 processors on the Cheyenne HPC system and $\sim$1.5 hours to generate 500k samples via MCMC from the posterior distribution. The 11-parameter emulation-calibration

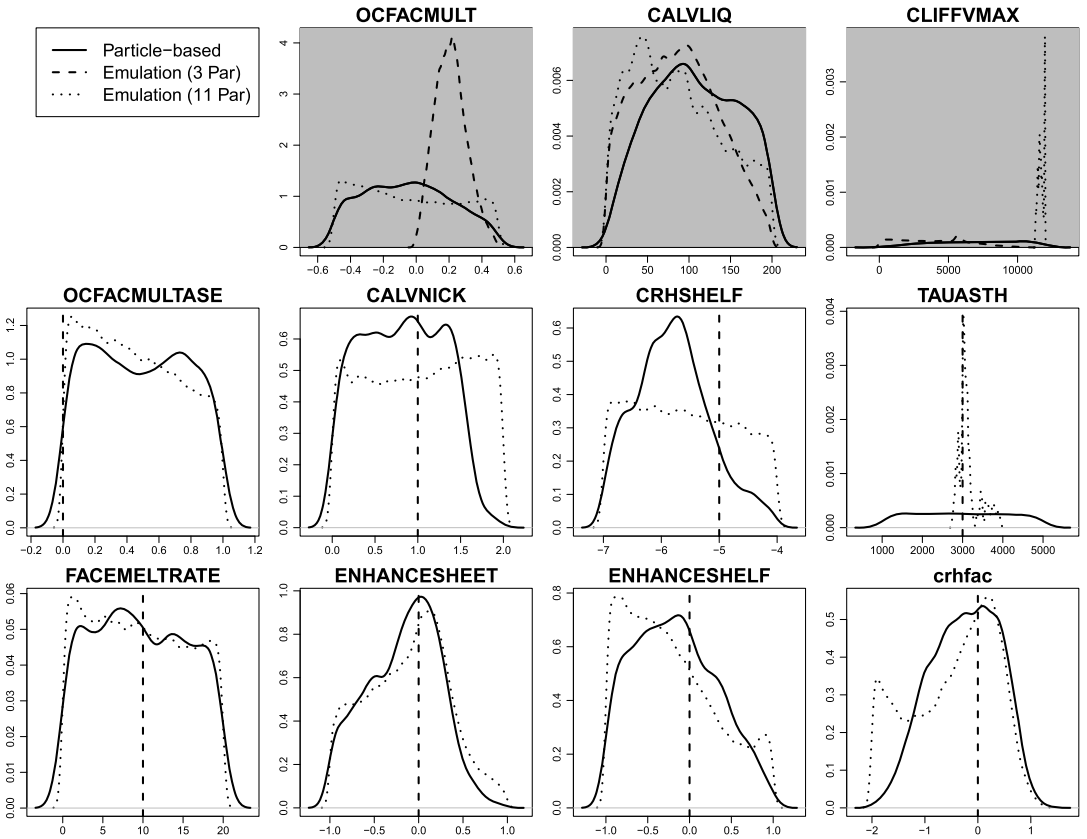

FIG. 3. *Posterior densities of model parameters using the adaptive particle-based approach* (*solid line*), *emulation-calibration with three parameters* (*dashed line*) *and emulation-calibration with* 11 *parameters* (*dotted line*). *Three-parameter emulation-calibration experiments use model parameters OCFACMULT, CALVLIQ and CLIFFVMAX. The* 11*-parameter emulation-calibration experiment include all model parameters. Shaded panels denote parameters used in the three-parameter emulation-calibration experiment. Both emulation-based approaches result in sharper posterior densities than the particle-based approach for a subset of the model parameters.*

required 10 minutes to fit the emulator using 12 processors on the Cheyenne HPC system and ∼1.5 hours to generate 500k samples via MCMC from the posterior distribution.

6.4. *The effect of deep time observations on projections*. Calibration can be improved by considering an important source of uncertainty, the state of the Antarctic ice sheet during the Pliocene era (Dolan et al. (2018), Salzmann et al. (2013), Dutton et al. (2015)). There is some evidence that the Antarctic ice sheet experienced fluctuations in volume during the Pliocene era (Naish et al. (2009)). Other studies suggest that at peak warming episodes during the Pliocene era, the Antarctic ice sheet had a lower volume, contributing to higher sea level rise (Cook et al. (2014), Dolan et al. (2011), Dowsett and Cronin (1990), Pollard and DeConto (2009), Pollard, DeConto and Alley (2015), De Boer, Stocchi and Van De Wal (2014)). However, the maximum Antarctic ice retreat and sea level rise contribution during the Pliocene remains largely uncertain (Dutton et al. (2015), Rovere et al. (2014)).

We examine whether the width of the Pliocene observation windows (5 m to 25 m, 5 m to 10 m, 10 m to 25 m) has an influence on sea level projections and parameter estimation. (See Supplementary Material (Lee et al. (2020)) for details on how these windows affect the likelihood function.) Our results demonstrate that information regarding the nature of the Antarctic ice sheet during the Pliocene era has a strong influence on sea level projections. Figure 5 illustrates how the posterior densities for two key model parameters (CALVLIQ

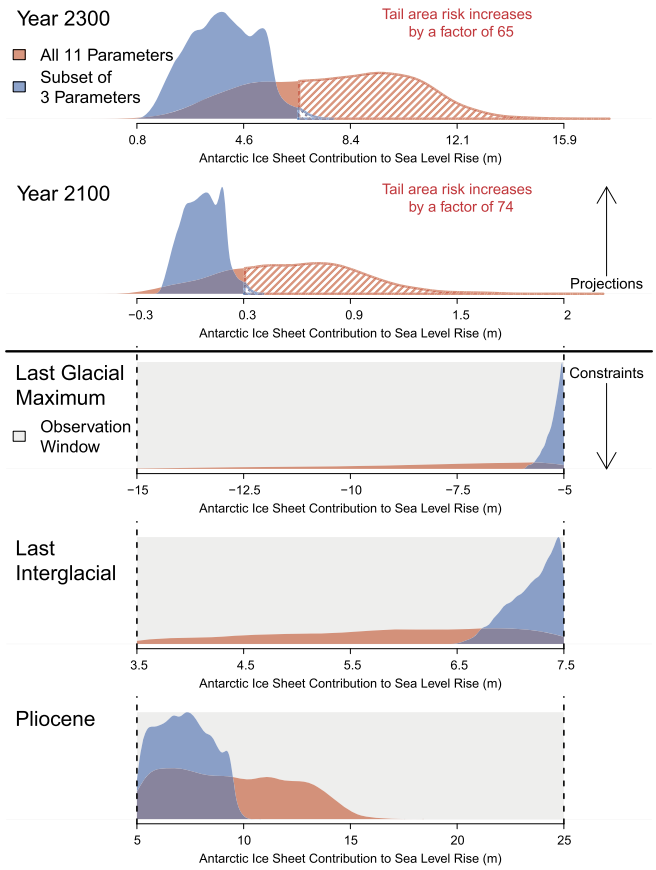

FIG. 4. *Antarctic ice Sheet contribution to sea level rise in the Pliocene* (*bottom panel*), *Last Interglacial Age* (*fourth panel*), *Last Glacial Maximum* (*third panel*), 2100 (*second panel*), *and* 2300 (*first panel*). *Red shading denotes the posterior densities for each time period and projections after calibrating* 11 *parameters using our fast particle-based approach. Blue shading denotes the posterior densities after calibrating three parameters using emulation-calibration. The light gray shading represents the observational constraints for the Last Glacial Maximum*, *Last Interglacial Age*, *and Pliocene. The striped red and striped blue shading represents the* 99*%th percent quantile for the* 11*-parameter approach and three-parameter approach*, *respectively.*

and CLIFFVMAX) differ under the three Pliocene windows. Both parameters influence ice dynamics inherent to marine cliff instability (MICI)—hydrofracturing due to surface melt (CALVLIQ) and structural failure of tall ice cliffs (CLIFFVMAX). As shown in Figure 6, increasing the Pliocene window from the range 5 m to 10 m to the range 10 m to 25 m requires more aggressive MICI (larger values of these parameters), hence resulting in higher projections of sea level rise (e.g., exceeding 3 m in 2300). If we are very uncertain about the Pliocene (represented by a very large window of 5 m to 25 m), the resulting sea level projections in 2300 also become highly uncertain (95% credible interval of 1.2 m to 12.4 m), compared to projections from narrower windows of 5 m to 10 m (95% credible interval of 1.2 m to 11.5 m) or 10 m to 25 m (95% credible interval of 3.0 m to 12.9 m).

The experiments using low (5 m to 10 m) and high (10 m to 25 m) Pliocene windows utilized subsets of the samples generated from the main calibration, and the corresponding subsamples had an effective sample size (ESS) of 891 and 642, respectively.

6.5. *Sensitivity to model parameter priors.* Calibration results may exhibit sensitivity to the choice of the model parameters' prior distributions (cf. Jackson et al. (2015), Reese et al. (2004)), especially for sparse observational records. This constitutes an important source

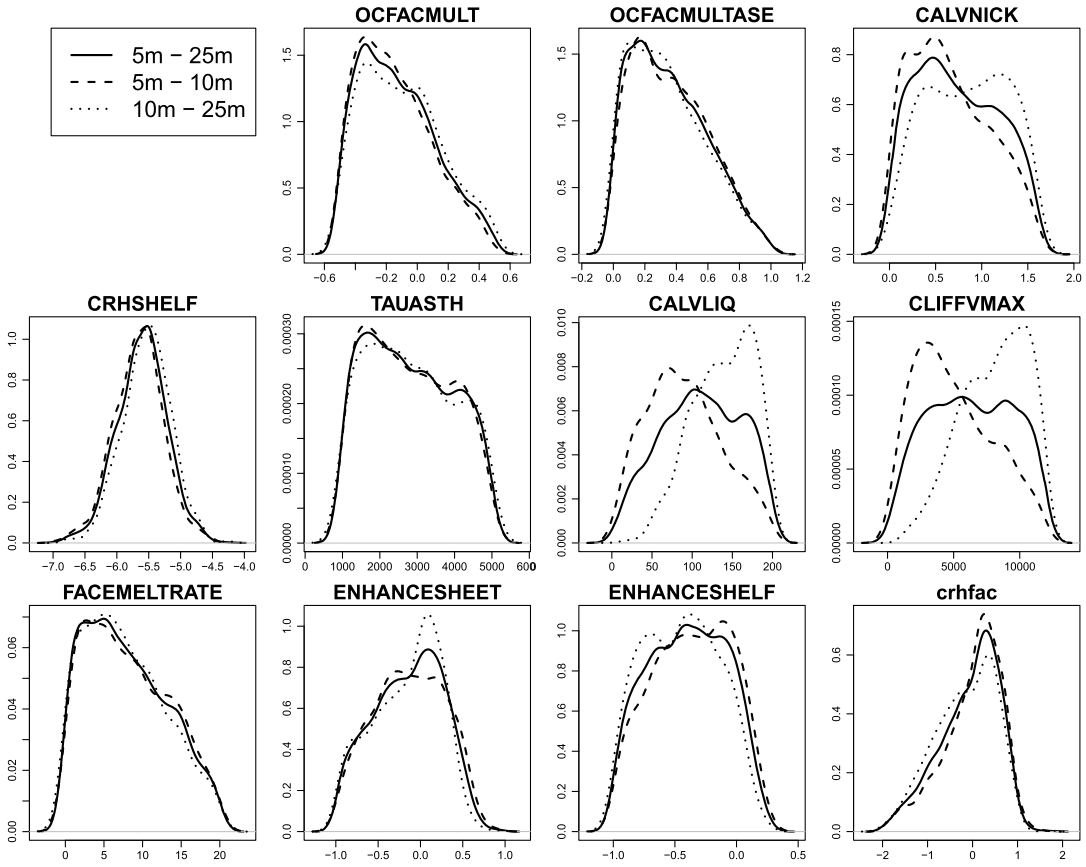

FIG. 5. *Posterior densities of model parameters for calibration using a wide Pliocene window of* 5 *m to* 25 *m* (*solid line*), *low window of* 5 *m to* 10 *m* (*dashed line*), *and a high window of* 10 *m to* 25 *m* (*dotted line*). *There is noticeable change in the densities for three model parameters—CALVNICK*, *CALVLIQ*, *and CLIFFVMAX*.

of second-order, or deep uncertainty, an important factor in the design of risk management strategies (Keller and McInerney (2008)). To examine prior sensitivity, we calibrate the ice sheet model using two sets of prior distributions which are in the form of uniform or log-uniform distributions. One set of priors has a much wider range (large difference between upper and lower bounds) than the other. The much wider ranges represent physically possible parameter values that do not violate any fundamental physical laws, and the narrower ranges represent values that yield reasonable model behavior found in many years of unstructured tuning by the model developers (Pollard and DeConto (2012)). We provide additional details in the Supplementary Material (Lee et al. (2020)).

The choice of prior distributions has a notable effect on parameter estimates (Figure 7) and sea level projections (Figure 8 and Table 2). Note that constraining the model parameters a priori may underestimate sea level projections. However, overly wide prior distributions may permit physically unrealistic outcomes. Hence, it is important to carefully construct prior distributions based on domain area expertise, as we have in this manuscript. In particular, changing the prior on the parameter CLIFFVMAX—wastage rate for unstable marine ice cliffs—can have a strong impact on projections. For a prior range of 0 km/year to 12 km/year, the 95% credible interval for the Antarctic ice sheet's contribution to sea level rise in 2300 is 1.2 m to 12.4 m. A wider prior range of 0 km/year to 600 km/year results in considerably higher projection uncertainty denoted by a 95% credible interval of 0.7 m to 21.0 m.

For the experiment using the wide priors, our particle-based calibration approach utilized 2015 particles to obtain an effective sample size (ESS) of 1583.

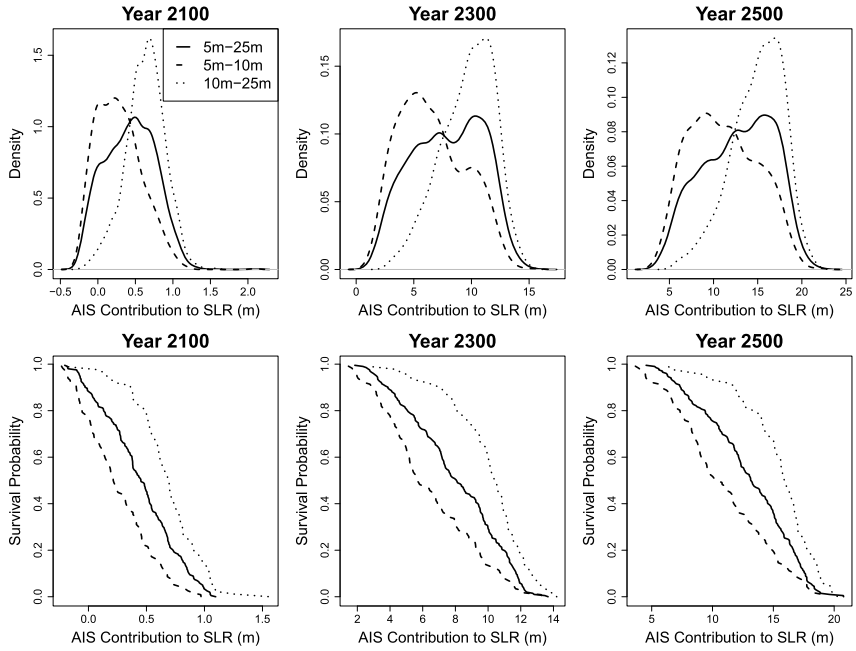

FIG. 6. (*Top Panel*) *Posterior densities of the projected Antarctic ice sheet's contribution to sea level change in* 2100, 2300, *and* 2500 *for calibration using a wide Pliocene window of* 5 *m to* 25 *m* (*solid line*), *low window of* 5 *m to* 10 *m* (*dashed line*), *and a high window of* 10 *m to* 25 *m* (*dotted line*). (*Bottom Panel*) *Empirical survival function of the projected Antarctic ice sheet's contribution to sea level change in* 2100, 2300, *and* 2500 *for calibration using a wide Pliocene window of* 5 *m to* 25 *m* (*solid line*), *low window of* 5 *m to* 10 *m* (*dashed line*), *and a high window of* 10 *m to* 25 *m* (*dotted line*). *constraining the Pliocene windows yield sharper projections of sea level rise. The higher window results in considerably higher projections than the lower window.*

The wider range for CLIFFVMAX explores a fundamental uncertainty in MICI—the rate at which very tall ice cliffs will disintegrate back into the ice sheet interior. If grounding lines retreat into the interior of deep Antarctic basins, the exposed ice cliffs will be taller than any observed today, and the wastage velocities (CLIFFVMAX) could conceivably be much greater than the approximately 12 km per year observed today at the ice fronts of major Greenland glaciers (which might not even be approximate analogs for MICI, being driven instead mainly by buoyant calving; Murray et al. (2015)). The bimodal character of the posterior densities in the top panels of Figure 8 for 2300 and 2500 are due to the very large CLIFFVMAX range. The upper peak centered on around 20 m is produced by CLIFFVMAX values of approximately 100 km per year and above, which produce collapse of almost all marine ice in both East and West Antarctica. The lower peak centered on around 5 m occurs for many lower CLIFFVMAX values, for which the more vulnerable West Antarctic ice sheet collapses, but marine basins in East Antarctica do not retreat.

TABLE 2
Antarctic ice sheet's projected contribution to sea level change in 2100–2500 *after calibration using narrow and wide prior distributions*

Prior	Year 2100	Year 2200	Year 2300	Year 2400	Year 2500
Narrow	0.4 (−0.3, 1)	3.8 (0.1, 6.7)	7.9 (1.2, 12.4)	10.6 (2.5, 15.5)	12.8 (3.7, 18.5)
Wide	1.8 (−0.4, 5.5)	10 (−0.2, 19.5)	13.9 (0.7, 21)	15.5 (1.8, 21.8)	16.6 (3.1, 22.5)

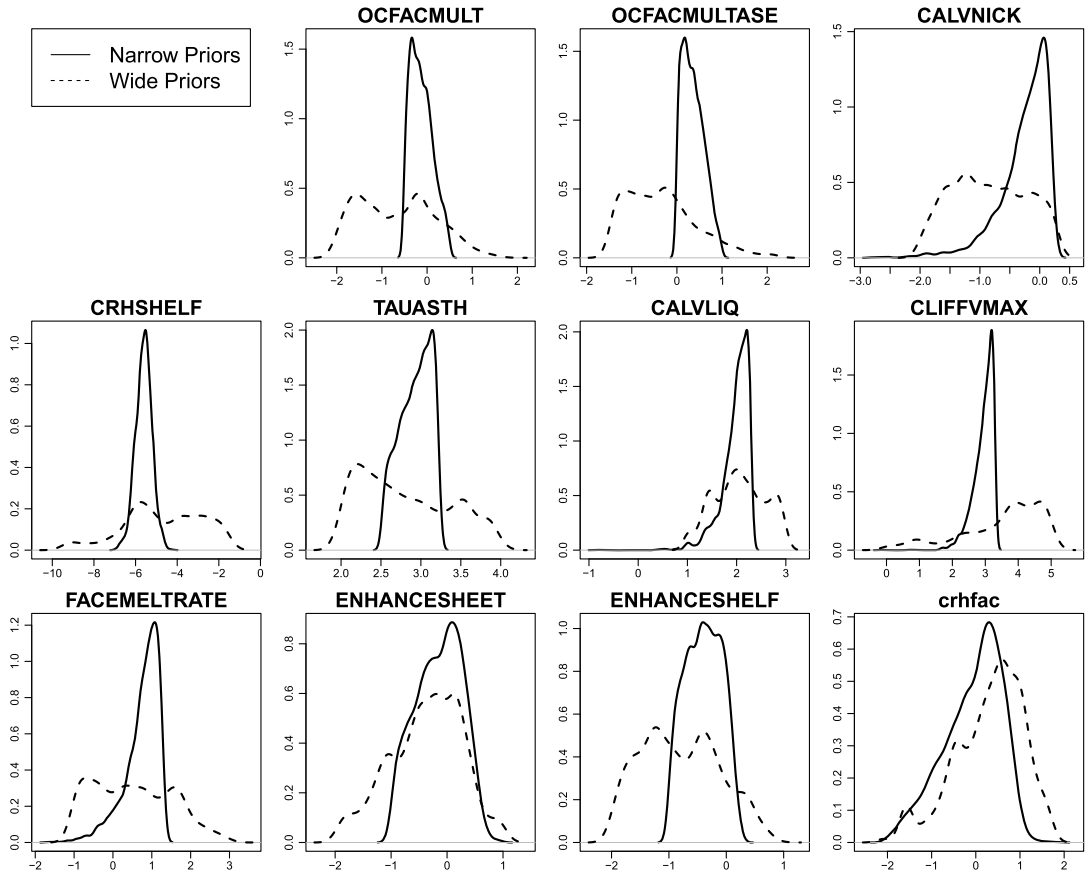

FIG. 7. *Posterior densities of model parameters using expert prior distributions* (*solid lines*) *and wider expert prior distributions* (*dashed lines*). *The dissimilarity of posterior distributions indicate that calibration results are highly sensitive to the choice of prior distributions.*

7. Discussion.

7.1. *Summary.* We present a novel particle-based approach to calibrate the 80 km resolution PSU3D-ICE model. We show that our approach provides good approximations and drastically reduces overall calibration wall times by heavily parallelizing the sequential Monte Carlo algorithm and carefully tuning the algorithm to further reduce the number of sequential model evaluations. Our algorithm is applicable to a broad class of models that have a moderate run time (given our computing resources, between a few seconds and several minutes) and a moderate number of model parameters (in our case between five and 20).

We use this new method to assess the impacts of neglecting parametric uncertainties on sea level projections. Emulation-calibration methods using fewer parameters yield lower and more overconfident projections of sea level rise than using more parameters through the particle-based calibration approach. This method includes the recent study of Edwards et al. (2019), who found that the important mechanism of marine ice cliff instability (MICI) is not necessary to capture past variations. In this case, future sea level projections are considerably lower. In contrast, our new approach that accounts for more parametric uncertainties suggests that MICI may still be important and future sea level projections may be much higher, especially considering potential Pliocene windows. Using emulation-calibration in a high-dimensional parameter space induces considerable emulator-model discrepancy and can result in large projection uncertainties. Our method utilizes the actual ice sheet model,

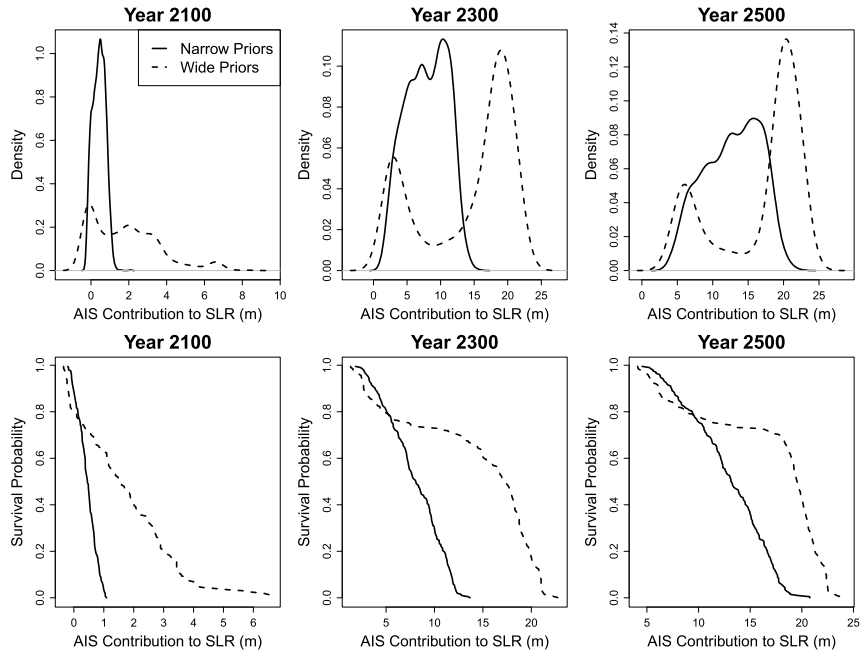

FIG. 8. (*Top Panel*) *Posterior densities of the projected Antarctic ice sheet's contribution to sea level change in* 2100, 2300, *and* 2500 *using expert prior distributions* (*solid lines*) *and wider expert prior distributions* (*dashed lines*). (*Bottom Panel*) *Empirical survival function of the projected Antarctic ice sheet's contribution to sea level change in* 2100, 2300, *and* 2500 *using expert prior distributions* (*solid lines*) *and wider expert prior distributions* (*dashed lines*). *For wide prior distributions*, *projections for future sea level rise is higher and more uncertain*, *and there exists bi-modality in the projections' posterior predictive distribution*.

thereby preserving the highly nonlinear ice dynamics as well as the complex interactions between model parameters. This has clear policy-relevant implications because projections from ice sheet models inform economic and engineering assessments (cf. Sriver et al. (2018), Diaz and Keller (2016), Johnson, Fischbach and Ortiz (2013)).

Our approach enables calibration experiments that were computationally prohibitive using current calibration methods. First, assuming different ranges of Pliocene era sea level constraints (low vs. high) results in markedly different characterizations of parametric uncertainty and projections of sea level rise over the next five centuries. These results suggest that improved geological data from the Pliocene can help better quantify the model parameters central to marine ice cliff instability (MICI) and improve sea level projections. Second, calibration results are highly sensitive to the choice of prior distributions. By overconstraining the prior distributions (in particular by not allowing very fast cliff disintegration rates), we can mischaracterize parametric uncertainty and drastically underestimate future sea level changes.

7.2. *Caveats*. Our conclusions are subject to the usual caveats that also point to promising and policy relevant research directions. Key methodological caveats include that our calibration approach may not scale well to computer models with long model run times (> 15 minutes), high-dimensional input spaces (> 20 parameters) or a combination of both. For high-dimensional input spaces, our approach would require: (1) a large number of particles to sensibly approximate the target distribution, (2) longer mutation stages to move the particles into the high-probability regions and (3) prohibitively large amount of computational resources to implement our approach. Our approach may not be suitable for computer models that use multiple processors for a single model run. Selecting an appropriate number of

particles remains an open question. Past theoretical work (Crisan and Doucet (2000)) states that using more particles yields better approximations of the target distributions. Here, we set the total particle count with respect to the available resources.

A number of caveats apply to our scientific findings. Using the PSU3D-ICE model at a coarser resolution than previous studies (DeConto and Pollard (2016), Chang et al. (2016a, 2016b), Pollard et al. (2016)) is admittedly a compromise between physical fidelity and run-time feasibility. At coarser resolutions, complex ice processes may not properly coalesce due to the spatial constraints. However, through a simulated example we found that a single model run at high-resolution (40 km) ran nearly eight times longer than one at at coarser resolution (80 km). Replicating this calibration study at sharper spatial resolutions (40 km to 10 km) is a natural and worthwhile extension of this study. Promising avenues for future work would include incorporating parallel MCMC approaches such as multiple-try Metropolis (Liu, Liang and Wong (2000)) or "emcee" samplers (Goodman and Weare (2010)) to reduce computer model runs in the mutation stage. Finally, the likelihood functions for the paleoclimate records may heavily influence calibration results. We have shown how the choice of expert priors influence calibration, but the influence of likelihood functions remains unexamined.

Acknowledgments. We would like to thank Don Richards, Daniel Gilford, Bob Kopp, Kelsey Ruckert, Vivek Srikrishnans, Robert Ceres, Kristina Rolph, Mahkameh Zarekarizi and Casey Hegelson for useful discussions. This work was partially supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Program, Earth and Environmental Systems Modeling, MultiSector Dynamics, Contract No. DE-SC0016162 and by the National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507. This study was also cosupported by the Penn State Center for Climate Risk Management. We would like to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Department of Energy, the National Science Foundation or other funding entities. Any errors and opinions are, of course, those of the authors. We are not aware of any real or perceived conflicts of interest for any authors.

Author contributions. All authors codesigned the overall study. BSL and MH formulated the statistical method. BSL wrote the computer code for calibration, designed the Pliocene window analysis and wrote the first draft of the manuscript. RF integrated the calibration method into the Cheyenne high perfomance computing system. MH edited the text. KK designed the comparative methods analysis and edited the text. DP provided code and data for the 80 km resolution PSU3D-ICE model, designed the prior sensitivity study and edited the text.

SUPPLEMENTARY MATERIAL

Supplement to "Supplement: A fast particle-based approach for calibrating a 3-D model of the Antarctic ice sheet" (DOI: 10.1214/19-AOAS1305SUPP; .pdf). Additional details on model parameters, priors, emulation-calibration details, fundamental equations for the PSU3D-ICE model, and simulated examples.

REFERENCES

AGAPIOU, S., PAPASPILIOPOULOS, O., SANZ-ALONSO, D. and STUART, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statist. Sci.* **32** 405–431. MR3696003 https://doi.org/10.1214/17-STS611

ALDER, J. R., HOSTETLER, S. W., POLLARD, D. and SCHMITTNER, A. (2011). Evaluation of a present-day climate simulation with a new coupled atmosphere-ocean model GENMOM. *Geosci. Model Dev.* **4** 69–83.

APPLEGATE, P. J., KIRCHNER, N., STONE, E. J., KELLER, K. and GREVE, R. (2012). An assessment of key model parametric uncertainties in projections of Greenland ice sheet behavior. *Cryosphere* **6** 589–606.

BAKKER, A. M., APPLEGATE, P. J. and KELLER, K. (2016). A simple, physically motivated model of sea-level contributions from the Greenland ice sheet in response to temperature changes. *Environmental Modelling & Software* **83** 27–35.

BALLARD, G., SIEFERT, C. and HU, J. (2016). Reducing communication costs for sparse matrix multiplication within algebraic multigrid. *SIAM J. Sci. Comput.* **38** C203–C231. MR3507096 https://doi.org/10.1137/15M1028807

BASTOS, L. S. and O'HAGAN, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics* **51** 425–438. MR2756478 https://doi.org/10.1198/TECH.2009.08019

BAYARRI, M. J., BERGER, J. O., CAFEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R. J., PAULO, R., SACKS, J. et al. (2007). Computer model validation with functional output. *Ann. Statist.* **35** 1874–1906. MR2363956 https://doi.org/10.1214/009053607000000163

BESKOS, A., JASRA, A., KANTAS, N. and THIERY, A. (2016). On the convergence of adaptive sequential Monte Carlo methods. *Ann. Appl. Probab.* **26** 1111–1146. MR3476634 https://doi.org/10.1214/15-AAP1113

BHAT, K. S., HARAN, M., GOES, M. and CHEN, M. (2010). Computer model calibration with multivariate spatial output: A case study. *Frontiers of Statistical Decision Making and Bayesian Analysis* 168–184.

BHATTACHARYYA, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā* **7** 401–406. MR0018387

BONY, S. and EMANUEL, K. A. (2001). A parameterization of the cloudiness associated with cumulus convection; evaluation using TOGA COARE data. *J. Atmos. Sci.* **58** 3158–3183.

BRYNJARSDÓTTIR, J. and O'HAGAN, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Probl.* **30** 114007, 24. MR3274591 https://doi.org/10.1088/0266-5611/30/11/114007

CHANG, W., HARAN, M., OLSON, R. and KELLER, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Stat.* **8** 649–673. MR3262529 https://doi.org/10.1214/14-AOAS733

CHANG, W., HARAN, M., APPLEGATE, P. and POLLARD, D. (2016a). Calibrating an ice sheet model using high-dimensional binary spatial data. *J. Amer. Statist. Assoc.* **111** 57–72. MR3494638 https://doi.org/10.1080/01621459.2015.1108199

CHANG, W., HARAN, M., APPLEGATE, P. and POLLARD, D. (2016b). Improving ice sheet model calibration using paleoclimate and modern data. *Ann. Appl. Stat.* **10** 2274–2302. MR3592057 https://doi.org/10.1214/16-AOAS979

CHOPIN, N. (2002). A sequential particle filter method for static models. *Biometrika* **89** 539–551. MR1929161 https://doi.org/10.1093/biomet/89.3.539

COLLINS, M. (2007). Ensembles and probabilities: A new era in the prediction of climate change. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **365** 1957–1970.

COMPUTATIONAL AND INFORMATION SYSTEMS LABORATORY (2017). Cheyenne: HPE/SGI ICE XA System (University Community Computing). Boulder, CO: National Center for Atmospheric Research. doi:10.5065/D6RX99HX. https://doi.org/10.5065/D6RX99HX

COOK, C. P., VAN DE FLIERDT, T., WILLIAMS, T., HEMMING, S. R., IWAI, M., KOBAYASHI, M., JIMENEZ-ESPEJO, F. J., ESCUTIA, C., GONZÁLEZ, J. J. et al. (2013). Dynamic behaviour of the East Antarctic ice sheet during Pliocene warmth. *Nature Geoscience* **6** 765–769.

COOK, C., HILL, D., VAN DE FLIERDT, T., WILLIAMS, T., HEMMING, S., DOLAN, A., PIERCE, E., ESCUTIA, C., HARWOOD, D. et al. (2014). Sea surface temperature control on the distribution of far-traveled Southern Ocean ice-rafted detritus during the Pliocene. *Paleoceanography* **29** 533–548.

CRISAN, D. and DOUCET, A. (2000). Convergence of sequential Monte Carlo methods. Technical Report CUED/F-INFENG/TR381, Signal Processing Group, Department of Engineering, University of Cambridge.

DAL GESSO, S. and NEGGERS, R. (2018). Can we use single-column models for understanding the boundary layer cloud-climate feedback? *J. Adv. Model. Earth Syst.* **10** 245–261.

DE BOER, B., STOCCHI, P. and VAN DE WAL, R. (2014). A fully coupled 3-D ice-sheet-sea-level model: Algorithm and applications. *Geosci. Model Dev.* **7** 2141–2156.

DECONTO, R. M. and POLLARD, D. (2016). Contribution of Antarctica to past and future sea-level rise. *Nature* **531** 591–597. https://doi.org/10.1038/nature17145

DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. MR2278333 https://doi.org/10.1111/j.1467-9868.2006.00553.x

DESCHAMPS, P., DURAND, N., BARD, E., HAMELIN, B., CAMOIN, G., THOMAS, A. L., HENDERSON, G. M., OKUNO, J. and YOKOYAMA, Y. (2012). Ice-sheet collapse and sea-level rise at the Bølling warming 14,600 years ago. *Nature* **483** 559.

DIAZ, D. and KELLER, K. (2016). A potential disintegration of the West Antarctic ice sheet: implications for economic analyses of climate policy. *Am. Econ. Rev.* **106** 607–11.

DOLAN, A. M., HAYWOOD, A. M., HILL, D. J., DOWSETT, H. J., HUNTER, S. J., LUNT, D. J. and PICKERING, S. J. (2011). Sensitivity of Pliocene ice sheets to orbital forcing. *Palaeogeography, Palaeoclimatology, Palaeoecology* **309** 98–110.

DOLAN, A. M., DE BOER, B., BERNALES, J., HILL, D. J. and HAYWOOD, A. M. (2018). High climate model dependency of Pliocene Antarctic ice-sheet predictions. *Nat. Commun.* **9** 2799. https://doi.org/10.1038/s41467-018-05179-4

DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*. *Stat. Eng. Inf. Sci.* 3–14. Springer, New York. MR1847784 https://doi.org/10.1007/978-1-4757-3437-9_1

DOUCET, A., GODSILL, S. and ANDRIEU, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10** 197–208.

DOWSETT, H. J. and CRONIN, T. M. (1990). High eustatic sea level during the middle Pliocene: evidence from the southeastern US Atlantic Coastal Plain. *Geology* **18** 435–438.

DUTTON, A., CARLSON, A. E., LONG, A. J., MILNE, G. A., CLARK, P. U., DECONTO, R., HORTON, B. P., RAHMSTORF, S. and RAYMO, M. E. (2015). SEA-LEVEL RISE. Sea-level rise due to polar ice-sheet mass loss during past warm periods. *Science* **349** aaa4019. https://doi.org/10.1126/science.aaa4019

EDWARDS, T. L., BRANDON, M. A., DURAND, G., EDWARDS, N. R., GOLLEDGE, N. R., HOLDEN, P. B., NIAS, I. J., PAYNE, A. J., RITZ, C. et al. (2019). Revisiting Antarctic ice loss due to marine ice-cliff instability. *Nature* **566** 58–64. https://doi.org/10.1038/s41586-019-0901-4

FAN, W., YU, W., XU, J., ZHOU, J., LUO, X., YIN, Q., LU, P., CAO, Y. and XU, R. (2018). Parallelizing sequential graph computations. *ACM Trans. Database Syst.* **43** Art. 18, 39. MR3892830 https://doi.org/10.1145/3282488

FITZGERALD, P., BAMBER, J., RIDLEY, J. and ROUGIER, J. (2012). Exploration of parametric uncertainty in a surface mass balance model applied to the Greenland ice sheet. *Journal of Geophysical Research: Earth Surface* **117**.

FRETWELL, P., PRITCHARD, H., VAUGHAN, D., BAMBER, J., BARRAND, N., BELL, R., BIANCHI, C., BINGHAM, R., BLANKENSHIP, D. et al. (2012). Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. *The Cryosphere Discussions* **6** 4305–4361.

FULLER, R. W., WONG, T. E. and KELLER, K. (2017). Probabilistic inversion of expert assessments to inform projections about Antarctic ice sheet responses. *PLoS ONE* **12** e0190115. https://doi.org/10.1371/journal.pone.0190115

GARNER, G. G. and KELLER, K. (2018). Using direct policy search to identify robust strategies in adapting to uncertain sea-level rise and storm surge. *Environmental Modelling & Software* **107** 96–104.

GETTELMAN, A., TRUESDALE, J. E., BACMEISTER, J. T., CALDWELL, P. M., NEALE, R. B., BOGENSCHUTZ, P. A. and SIMPSON, I. R. (2019). The single column atmosphere model version 6 (SCAM6): Not a scam but a tool for model evaluation and development. *J. Adv. Model. Earth Syst.* **0**.

GILKS, W. R. and BERZUINI, C. (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 127–146. MR1811995 https://doi.org/10.1111/1467-9868.00280

GIRAUD, F. and DEL MORAL, P. (2017). Nonasymptotic analysis of adaptive and annealed Feynman–Kac particle models. *Bernoulli* **23** 670–709. MR3556789 https://doi.org/10.3150/14-BEJ680

GOLLEDGE, N. R., KELLER, E. D., GOMEZ, N., NAUGHTEN, K. A., BERNALES, J., TRUSEL, L. D. and EDWARDS, T. L. (2019). Global environmental consequences of twenty-first-century ice-sheet melt. *Nature* **566** 65–72. https://doi.org/10.1038/s41586-019-0889-9

GOODMAN, J. and WEARE, J. (2010). Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci.* **5** 65–80. MR2600822 https://doi.org/10.2140/camcos.2010.5.65

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-Radar and Signal Processing* **140** 107–113. IET.

GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395 https://doi.org/10.1080/10618600.2014.914442

GREVE, R. (1997). Application of a polythermal three-dimensional ice sheet model to the Greenland ice sheet: Response to steady-state and transient climate scenarios. *J. Climate* **10** 901–918.

HANNART, A., GHIL, M., DUFRESNE, J.-L. and NAVEAU, P. (2013). Disconcerting learning on climate sensitivity and the uncertain future of uncertainty. *Clim. Change* **119** 585–601.

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 https://doi.org/10.1198/016214507000000888

ISAAC, T., STADLER, G. and GHATTAS, O. (2015). Solution of nonlinear Stokes equations discretized by high-order finite elements on nonconforming and anisotropic meshes, with application to ice sheet dynamics. *SIAM J. Sci. Comput.* **37** B804–B833. MR3418230 https://doi.org/10.1137/140974407

ISAAC, T., PETRA, N., STADLER, G. and GHATTAS, O. (2015). Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet. *J. Comput. Phys.* **296** 348–368. MR3348506 https://doi.org/10.1016/j.jcp.2015.04.047

JACKSON, C. H., JIT, M., SHARPLES, L. D. and DE ANGELIS, D. (2015). Calibration of complex models through Bayesian evidence synthesis: A demonstration and tutorial. *Medical Decision Making* **35** 148–161.

JASRA, A., STEPHENS, D. A., DOUCET, A. and TSAGARIS, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scand. J. Stat.* **38** 1–22. MR2760137 https://doi.org/10.1111/j.1467-9469.2010.00723.x

JEREMIAH, E., SISSON, S., MARSHALL, L., MEHROTRA, R. and SHARMA, A. (2011). Bayesian calibration and uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and sequential Monte Carlo samplers. *Water Resour. Res.* **47**.

JOHNSON, D. R., FISCHBACH, J. R. and ORTIZ, D. S. (2013). Estimating surge-based flood risk with the coastal Louisiana risk assessment model. *Journal of Coastal Research* **67** 109–126.

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478 https://doi.org/10.1198/016214506000000492

KALYANARAMAN, J., KAWAJIRI, Y., LIVELY, R. P. and REALFF, M. J. (2016). Uncertainty quantification via Bayesian inference using sequential Monte Carlo methods for CO2 adsorption process. *AIChE Journal* **62** 3352–3368.

KANTAS, N., BESKOS, A. and JASRA, A. (2014). Sequential Monte Carlo methods for high-dimensional inverse problems: A case study for the Navier–Stokes equations. *SIAM/ASA J. Uncertain. Quantificat.* **2** 464–489. MR3283917 https://doi.org/10.1137/130930364

KANTAS, N., DOUCET, A., SINGH, S. S., MACIEJOWSKI, J. and CHOPIN, N. (2015). On particle methods for parameter estimation in state-space models. *Statist. Sci.* **30** 328–351. MR3383884 https://doi.org/10.1214/14-STS511

KELLER, K. and MCINERNEY, D. (2008). The dynamics of learning about a climate threshold. *Clim. Dyn.* **30** 321–332.

KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398 https://doi.org/10.1111/1467-9868.00294

KIM, S. H., EDMONDS, J., LURZ, J., SMITH, S. J. and WISE, M. (2006). The objECTS framework for integrated assessment: Hybrid modeling of transportation. *Energy J.* 63–91.

KONG, A. (1992). A note on importance sampling using standardized weights. University of Chicago, Dept. of Statistics, Tech. Rep, 348.

KOPP, R. E., SIMONS, F. J., MITROVICA, J. X., MALOOF, A. C. and OPPENHEIMER, M. (2009). Probabilistic assessment of sea level during the last interglacial stage. *Nature* **462** 863–867. https://doi.org/10.1038/nature08686

LAROUR, E., SEROUSSI, H., MORLIGHEM, M. and RIGNOT, E. (2012). Continental scale, high order, high spatial resolution, ice sheet modeling using the Ice Sheet System Model (ISSM). *Journal of Geophysical Research: Earth Surface* **117**.

LE BARS, D., DRIJFHOUT, S. and DE VRIES, H. (2017). A high-end sea level rise probabilistic projection including rapid Antarctic ice sheet mass loss. *Environ. Res. Lett.* **12** 044013.

LE BROCQ, A. M., PAYNE, A. J. and VIELI, A. (2010). An improved Antarctic dataset for high resolution numerical ice sheet models (ALBMAP v1). *Earth System Science Data* **2** 247–260.

LE COZANNET, G., MANCEAU, J.-C. and ROHMER, J. (2017). Bounding probabilistic sea-level projections within the framework of the possibility theory. *Environ. Res. Lett.* **12** 014012.

LEE, B. S., HARAN, M., FULLER, R. W., POLLARD, D. and KELLER, K. (2020). Supplement to "A fast particle-based approach for calibrating a 3-D model of the Antarctic ice sheet." https://doi.org/10.1214/19-AOAS1305SUPP.

LELE, S. R., DENNIS, B. and LUTSCHER, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* **10** 551–563.

LEVITUS, S., ANTONOV, J. I., BOYER, T. P., BARANOVA, O. K., GARCIA, H. E., LOCARNINI, R. A., MISHONOV, A. V., REAGAN, J., SEIDOV, D. et al. (2012). World ocean heat content and thermosteric sea level change (0–2000 m). *Geophys. Res. Lett.* **39** 1955–2010.

LI, T., SUN, S., SATTAR, T. P. and CORCHADO, J. M. (2014). Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. *Expert Systems with Applications* **41** 3944–3954.

LIANG, F. and WONG, W. H. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Statist. Assoc.* **96** 653–666. MR1946432 https://doi.org/10.1198/016214501753168325

LIANG, X., LETTENMAIER, D. P., WOOD, E. F. and BURGES, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res., Atmos.* **99** 14415–14428.

LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. MR2486241 https://doi.org/10.1214/09-BA404

LIU, J. S. and CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93** 1032–1044. MR1649198 https://doi.org/10.2307/2669847

LIU, X. and GUILLAS, S. (2017). Dimension reduction for Gaussian process emulation: An application to the influence of bathymetry on tsunami heights. *SIAM/ASA J. Uncertain. Quantificat.* **5** 787–812. MR3686816 https://doi.org/10.1137/16M1090648

LIU, J. S., LIANG, F. and WONG, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* **95** 121–134. MR1803145 https://doi.org/10.2307/2669532

LIU, J. and WEST, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice*. *Stat. Eng. Inf. Sci.* 197–223. Springer, New York. MR1847793

LIU, Z., OTTO-BLIESNER, B., HE, F., BRADY, E., TOMAS, R., CLARK, P., CARLSON, A., LYNCH-STIEGLITZ, J., CURRY, W. et al. (2009). Transient simulation of last deglaciation with a new mechanism for Bølling–Allerød warming. *Science* **325** 310–314.

LLOPIS, F. P., KANTAS, N., BESKOS, A. and JASRA, A. (2018). Particle filtering for stochastic Navier–Stokes signal observed with linear additive noise. *SIAM J. Sci. Comput.* **40** A1544–A1565. MR3807951 https://doi.org/10.1137/17M1151900

LORENZ, E. N. (1972). Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas? Presented at the 139th Meeting of the AAAS. Available at http://eaps4.mit.edu/research/Lorenz/Butterfly_1972.pdf. Accessed: 2019-08-04.

MANIYAR, D., CORNFORD, D. and BOUKOUVALAS, A. (2007). Dimensionality reduction in the emulator setting. Technical report, Neural Computing Research Group, University of Aston.

MARTINO, L. (2018). A review of multiple try MCMC algorithms for signal processing. *Digit. Signal Process.* **75** 134–152. MR3765591 https://doi.org/10.1016/j.dsp.2018.01.004

MARTINO, L., ELVIRA, V. and LOUZADA, F. (2017). Effective sample size for importance sampling based on discrepancy measures. *Signal Process.* **131** 386–401.

MCGRANAHAN, G., BALK, D. and ANDERSON, B. (2007). The rising tide: Assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment and Urbanization* **19** 17–37.

MCKAY, M. D., BECKMAN, R. J. and CONOVER, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **421** 55–61.

MEINSHAUSEN, M., SMITH, S. J., CALVIN, K., DANIEL, J. S., KAINUMA, M., LAMARQUE, J.-F., MATSUMOTO, K., MONTZKA, S., RAPER, S. et al. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Clim. Change* **109** 213.

MONIER, E., SCOTT, J., SOKOLOV, A., FOREST, C. and SCHLOSSER, C. (2013). An integrated assessment modelling framework for uncertainty studies in global and regional climate change: The MIT IGSM-CAM (version 1.0). *Geoscientific Model Development Discussions* **6**.

MORZFELD, M., TU, X., WILKENING, J. and CHORIN, A. J. (2015). Parameter estimation by implicit sampling. *Commun. Appl. Math. Comput. Sci.* **10** 205–225. MR3402350 https://doi.org/10.2140/camcos.2015.10.205

MORZFELD, M., DAY, M. S., GROUT, R. W., PAU, G. S. H., FINSTERLE, S. A. and BELL, J. B. (2018). Iterative importance sampling algorithms for parameter estimation. *SIAM J. Sci. Comput.* **40** B329–B352. MR3769709 https://doi.org/10.1137/16M1088417

MURRAY, L. M., LEE, A. and JACOB, P. E. (2016). Parallel resampling in the particle filter. *J. Comput. Graph. Statist.* **25** 789–805. MR3533638 https://doi.org/10.1080/10618600.2015.1062015

MURRAY, T., SELMES, N., JAMES, T. D., EDWARDS, S., MARTIN, I., O'FARRELL, T., ASPEY, R., RUTT, I., NETTLES, M. et al. (2015). Dynamics of glacier calving at the ungrounded margin of Helheim Glacier, southeast Greenland. *J Geophys Res Earth Surf* **120** 964–982. https://doi.org/10.1002/2015JF003531

NAISH, T., POWELL, R., LEVY, R., WILSON, G., SCHERER, R., TALARICO, F., KRISSEK, L., NIESSEN, F., POMPILIO, M. et al. (2009). Obliquity-paced Pliocene West Antarctic ice sheet oscillations. *Nature* **458** 322–328. https://doi.org/10.1038/nature07867

NEAL, R. M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. MR1837132 https://doi.org/10.1023/A:1008923215028

NGUYEN, T. L. T. (2014). Sequential Monte-Carlo sampler for Bayesian inference in complex systems. PhD thesis, Lille 1.

NICHOLLS, R. J., TOL, R. S. and VAFEIDIS, A. T. (2008). Global estimates of the impact of a collapse of the West Antarctic ice sheet: an application of FUND. *Clim. Change* **91** 171.

O'NEILL, B. C., CRUTZEN, P., GRÜBLER, A., DUONG, M. H., KELLER, K., KOLSTAD, C., KOOMEY, J., LANGE, A., OBERSTEINER, M. et al. (2006). Learning and climate change. *Climate Policy* **6** 585–589.

OPPENHEIMER, M. and ALLEY, R. B. (2016). How high will the seas rise? *Science* **354** 1375–1377. https://doi.org/10.1126/science.aak9460

PAL, J. S., GIORGI, F., BI, X., ELGUINDI, N., SOLMON, F., GAO, X., RAUSCHER, S. A., FRANCISCO, R., ZAKEY, A. et al. (2007). Regional climate modeling for the developing world: the ICTP RegCM3 and RegCNET. *Bull. Am. Meteorol. Soc.* **88** 1395–1410.

PAPAIOANNOU, I., PAPADIMITRIOU, C. and STRAUB, D. (2016). Sequential importance sampling for structural reliability analysis. *Structural Safety* **62** 66–75.

PETRA, N., MARTIN, J., STADLER, G. and GHATTAS, O. (2014). A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM J. Sci. Comput.* **36** A1525–A1555. MR3233941 https://doi.org/10.1137/130934805

POLLARD, D. and DECONTO, R. M. (2009). Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458** 329–332. https://doi.org/10.1038/nature07809

POLLARD, D. and DECONTO, R. (2012). Description of a hybrid ice sheet-shelf model, and application to Antarctica. *Geosci. Model Dev.* **5** 1273.

POLLARD, D., DECONTO, R. M. and ALLEY, R. B. (2015). Potential Antarctic ice sheet retreat driven by hydrofracturing and ice cliff failure. *Earth and Planetary Science Letters* **412** 112–121.

POLLARD, D., GOMEZ, N. and DECONTO, R. M. (2017). Variations of the Antarctic ice sheet in a coupled ice sheet-Earth-sea level model: Sensitivity to viscoelastic Earth properties. *Journal of Geophysical Research: Earth Surface* **122** 2124–2138.

POLLARD, D., CHANG, W., HARAN, M., APPLEGATE, P. and DECONTO, R. (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** 1697–1723.

QUEIPO, N. V., HAFTKA, R. T., SHYY, W., GOEL, T., VAIDYANATHAN, R. and TUCKER, P. K. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Sciences* **41** 1–28.

REESE, C. S., WILSON, A. G., HAMADA, M., MARTZ, H. F. and RYAN, K. J. (2004). Integrated analysis of computer and physical experiments. *Technometrics* **46** 153–164. MR2060014 https://doi.org/10.1198/004017004000000211

ROVERE, A., RAYMO, M. E., MITROVICA, J., HEARTY, P. J., O'LEARY, M. and INGLIS, J. (2014). The Mid-Pliocene sea-level conundrum: glacial isostasy, eustasy and dynamic topography. *Earth and Planetary Science Letters* **387** 27–33.

RUCKERT, K. L., SHAFFER, G., POLLARD, D., GUAN, Y., WONG, T. E., FOREST, C. E. and KELLER, K. (2017). Assessing the impact of retreat mechanisms in a simple Antarctic ice sheet model using Bayesian calibration. *PLoS ONE* **12** e0170052.

RUTT, I. C., HAGDORN, M., HULTON, N. and PAYNE, A. (2009). The Glimmer community ice sheet model. *Journal of Geophysical Research: Earth Surface* **114**.

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. With comments and a rejoinder by the authors. MR1041765

SALZMANN, U., DOLAN, A. M., HAYWOOD, A. M., CHAN, W.-L., VOSS, J., HILL, D. J., ABE-OUCHI, A., OTTO-BLIESNER, B., BRAGG, F. J. et al. (2013). Challenges in quantifying Pliocene terrestrial warming revealed by data–model discord. *Nature Climate Change* **3** 969–974.

SANSÓ, B., FOREST, C. E. and ZANTEDESCHI, D. (2008). Inferring climate system properties using a computer model. *Bayesian Anal.* **3** 1–37. MR2383247 https://doi.org/10.1214/08-BA301

SCHÄFER, C. and CHOPIN, N. (2013). Sequential Monte Carlo on large binary sampling spaces. *Stat. Comput.* **23** 163–184. MR3016936 https://doi.org/10.1007/s11222-011-9299-z

SCHLEGEL, N.-J., SEROUSSI, H., SCHODLOK, M. P., LAROUR, E. Y., BOENING, C., LIMONADI, D., WATKINS, M. M., MORLIGHEM, M. and VAN DEN BROEKE, M. R. (2018). Exploration of Antarctic ice sheet 100-year contribution to sea level rise and associated model uncertainties using the ISSM framework. *Cryosphere* **12** 3511–3534.

SHAFFER, G. (2014). Formulation, calibration and validation of the DAIS model, a simple Antarctic ice sheet model sensitive to variations of sea level and ocean subsurface temperature. *Geosci. Model Dev.* **7** 1803–1818.

SHIELDS, C. A. and KIEHL, J. T. (2016). Atmospheric river landfall-latitude changes in future climate simulations. *Geophys. Res. Lett.* **43** 8775–8782.

SOROOSHIAN, S., DUAN, Q. and GUPTA, V. K. (1993). Calibration of rainfall-runoff models: Application of global optimization to the Sacramento soil moisture accounting model. *Water Resour. Res.* **29** 1185–1194.

SRIVER, R. L., LEMPERT, R. J., WIKMAN-SVAHN, P. and KELLER, K. (2018). Characterizing uncertain sea-level rise projections to support investment decisions. *PLoS ONE* **13** e0190641. https://doi.org/10.1371/journal.pone.0190641

STEIN, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29** 143–151. MR0887702 https://doi.org/10.2307/1269769

STEINBERG, D. M. and LIN, D. K. J. (2006). A construction method for orthogonal Latin hypercube designs. *Biometrika* **93** 279–288. MR2278083 https://doi.org/10.1093/biomet/93.2.279

STONE, E., LUNT, D., RUTT, I., HANNA, E. et al. (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** 397–417.

URBAN, N. M. and FRICKER, T. E. (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Comput. Geosci.* **36** 746–755.

WHITELEY, N., LEE, A. and HEINE, K. (2016). On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli* **22** 494–529. MR3449791 https://doi.org/10.3150/14-BEJ666

WILLEMS, J. C. (1972). Dissipative dynamical systems. I. General theory. *Arch. Ration. Mech. Anal.* **45** 321–351. MR0527462 https://doi.org/10.1007/BF00276493

WILLIAMSON, D., GOLDSTEIN, M., ALLISON, L., BLAKER, A., CHALLENOR, P., JACKSON, L. and YAMAZAKI, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dyn.* **2** 1703–1729.

WONG, T. E., BAKKER, A. M. and KELLER, K. (2017). Impacts of Antarctic fast dynamics on sea-level projections and coastal flood defense. *Clim. Change* **144** 347–364.

ZHOU, Y., JOHANSEN, A. M. and ASTON, J. A. D. (2016). Toward automatic model comparison: An adaptive sequential Monte Carlo approach. *J. Comput. Graph. Statist.* **25** 701–726. MR3533634 https://doi.org/10.1080/10618600.2015.1060885

ZHU, G., LI, X., MA, J., WANG, Y., LIU, S., HUANG, C., ZHANG, K. and HU, X. (2018). A new moving strategy for the sequential Monte Carlo approach in optimizing the hydrological model parameters. *Adv. Water Resour.* **114** 164–179.