The Annals of Applied Statistics
2014, Vol. 8, No. 2, 649–673
DOI: 10.1214/14-AOAS733

FAST DIMENSION-REDUCED CLIMATE MODEL CALIBRATION AND THE EFFECT OF DATA AGGREGATION[1]

BY WON CHANG, MURALI HARAN, ROMAN OLSON AND KLAUS KELLER

Pennsylvania State University

How will the climate system respond to anthropogenic forcings? One approach to this question relies on climate model projections. Current climate projections are considerably uncertain. Characterizing and, if possible, reducing this uncertainty is an area of ongoing research. We consider the problem of making projections of the North Atlantic meridional overturning circulation (AMOC). Uncertainties about climate model parameters play a key role in uncertainties in AMOC projections. When the observational data and the climate model output are high-dimensional spatial data sets, the data are typically aggregated due to computational constraints. The effects of aggregation are unclear because statistically rigorous approaches for model parameter inference have been infeasible for high-resolution data. Here we develop a flexible and computationally efficient approach using principal components and basis expansions to study the effect of spatial data aggregation on parametric and projection uncertainties. Our Bayesian reduced-dimensional calibration approach allows us to study the effect of complicated error structures and data-model discrepancies on our ability to learn about climate model parameters from high-dimensional data. Considering high-dimensional spatial observations reduces the effect of deep uncertainty associated with prior specifications for the data-model discrepancy. Also, using the unaggregated data results in sharper projections based on our climate model. Our computationally efficient approach may be widely applicable to a variety of high-dimensional computer model calibration problems.

1. Introduction. Computer models play an important role in understanding complex physical processes in modern science and engineering. They are particularly important in climate science where computer models, complex deterministic systems used to model the Earth System, are used both to study climate phenomena as well as make projections about the future. A major source of uncertainty in climate projections is due to uncertainties about model parameters. Parameter calibration involves characterizing our knowledge about a model parameter by using observational data. Here we use calibration to refer to a statistical method that summarizes information about a parameter in terms of a probability distribution. In this distribution parameter values that generate output more compatible with

Received October 2013; revised February 2014.

[1]Supported by NSF through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507.

Key words and phrases. Climate model, calibration, Gaussian process, principal components, high-dimensional spatial data.

observational data are assigned higher probabilities than parameters less compatible with observations. Calibration of the parameters using observational data is hence one avenue to reduce the uncertainty in future projections. A number of issues and challenges arise when performing statistical calibration of model parameters. Because each run of the computer model is computationally expensive, computer model output is typically obtained for a relatively small sample of parameter values. Furthermore, the model output at each parameter setting may be high dimensional and in the form of spatial fields. A sound statistical approach to this problem needs to simultaneously address the spatial dependence in the data and model outputs, account for various sources of uncertainty and remain computationally efficient. Computational efficiency is key in order to utilize all the relevant observations at the appropriate scale; previous methods for climate model calibration have relied on heavy data aggregation, thereby potentially discarding valuable information.

The scientific problem motivating our statistical analysis is the projection of the future state of the North Atlantic meridional overturning circulation (AMOC) in response to anthropogenic climate change. The AMOC is a large-scale ocean circulation that transports cold and dense water equatorward in the deep North Atlantic, and warm and salty water poleward in the upper layers of the North Atlantic. The AMOC might show a persistent weakening in response to anthropogenic forcing. Because the AMOC plays an important role in heat and carbon transport, an AMOC weakening is projected to have considerable impacts on climate, and, in response, on natural and human systems [cf. Alley et al. (2007), Keller et al. (2005, 2007)]. We use previously published perturbed physics ensemble runs [Sriver et al. (2012)] of the University of Victoria Earth System Climate Model (UVic ESCM) [Weaver et al. (2001)] to set up the calibration problem. Specifically, the runs model transient behavior of the climate system over the years 1800–2100. Each run starts from a control climate, obtained by running the model from the same initial condition to equilibrium at preindustrial conditions. Vertical ocean mixing is important in projecting the AMOC [Wunsch and Ferrari (2004)], but most of the mixing occurs on scales below that of the UVic ESCM, hence, mixing is "parameterized" [cf. Goes et al. (2010), Schmittner et al. (2009), Weaver et al. (2001)] using a "vertical background diffusivity" (K_{bg}). The AMOC projections depend on the K_{bg} parameter values [e.g., Goes et al. (2010)]. The value of K_{bg} is uncertain; it therefore needs to be calibrated using observations of the climate that are informative about K_{bg} [cf. Bhat et al. (2012), Goes et al. (2010)].

Here, we calibrate K_{bg} using observations of ocean potential temperature from the World Ocean Atlas 2009 [Antonov et al. (2010), Locarnini et al. (2010)]. The World Ocean Atlas is a gridded data product generated by interpolation of instrumental observations. The observational data at irregularly distributed locations are interpolated onto a regular grid by Barnes interpolation [Barnes (1964)]. The parameter K_{bg} affects the depth of the oceanic pycnocline [Gnanadesikan (1999)] and the AMOC [Bryan (1987), Goes et al. (2010)]. As a consequence, models with

different K_{bg} values are expected to result in different ocean temperature distributions. Ocean temperatures are therefore informative about K_{bg}. Note that neither data assimilation nor calibration play any role in producing the World Ocean Atlas data set and the data product does not depend on any assumptions about vertical diffusivity.

Both the UVic ESCM output and the observational data are spatial data sets with more than 60,000 spatial locations. Of particular interest is how data aggregation affects the calibration result for K_{bg}. Often observations of climate and climate model outputs are 3-D spatial fields. When the spatial data sets are large it is common practice to aggregate them into 1-D or 2-D patterns [Bhat et al. (2012), Drignei, Forest and Nychka (2008), Forest, Stone and Sokolov (2008), Goes et al. (2010), Olson et al. (2012), Sansó and Forest (2009), Schmittner et al. (2009)] either to avoid computational issues or because the skill of the models at higher resolution may not always be trusted.

An important and interesting question is what information, if any, is lost by this data aggregation. Aggregating data for model calibration may increase or decrease the model parameter uncertainty depending on the relative importance of several processes. Data aggregation may lead to information loss which would result in larger uncertainties about the calibrated parameters. On the other hand, data aggregation can potentially reduce the magnitude of model errors, for instance, due to unresolved variability or structural errors in the climate models, which could in turn lead to smaller uncertainties about the calibrated parameters. We hypothesize for the scientific questions and models we consider here that data aggregation may lead to considerable loss of information resulting in increased uncertainties about model parameters. Increased uncertainties about parameters propagates to increased uncertainty in climate projections, which can impact risk- and decision-analysis.

We adopt a Gaussian process-based approach to the calibration problem [Kennedy and O'Hagan (2001), Sacks et al. (1989)]. Gaussian processes provide flexible statistical interpolators or "emulators" of the computer model across various parameter settings and are therefore attractive for climate model calibration [cf. Bhat et al. (2012), Sansó and Forest (2009)]. Unfortunately, the likelihood evaluations involved in fitting such models can become prohibitive with high-dimensional spatial data due to the expensive matrix operations involved. Current approaches for high-dimensional computer model calibration can reduce the computational burden and make likelihood evaluation feasible for moderately large data sets (spatial fields observed at a few thousand locations) or data sets that are on a regular and complete grid [Bayarri et al. (2007), Bhat et al. (2012), Higdon et al. (2008)]. However, to our knowledge, no current calibration approach can overcome the computational challenge of dealing with large spatial data sets (more than tens of thousands of data points) on an incomplete grid. Here an incomplete grid refers to one with a large number of missing points (about 40%).

The impact of data aggregation on climate model calibration is a largely unanswered question due to the inability of existing methods to analyze large spatial data sets of both computer model output and observations. Throughout this manuscript we will use "large" to refer to data sets that comprise over tens of thousands of spatial observations. Here we develop a computationally efficient approach that handles large data sets. This approach gives us the freedom to carry out a careful study of the effects of data aggregation, for example, comparing calibration based on unaggregated three-dimensional data with calibration based on aggregated two-dimensional or one-dimensional data. Our approach also enables one to investigate the interaction between data aggregation and data-model discrepancies and errors when inferring computer model parameters. In our simulated examples, we have shown that the method can handle complicated model-observation discrepancy processes without sacrificing computational efficiency.

The remainder of this paper is organized as follows. In Section 2 we provide a description of the data set. In Section 3 we describe our two-stage framework for climate model calibration and the associated computational challenges. In Section 4 we propose a general model calibration approach in a reduced-dimensional space that uses a combination of principal components and a basis representation to overcome computational challenges. In Section 5 we provide implementation details and in Section 6 we discuss the results from simulated examples and real data. We conclude this paper with caveats and future directions in Section 7.

2. Data description. Our goal is to build an emulator based on spatial output from UVic ESCM and to calibrate vertical ocean diffusivity (K_{bg}) using ocean potential temperature data. The UVic ESCM runs are 3-dimensional patterns of the mean ocean potential temperature over 1955–2006 at 250 parameter settings. The parameters controlling model outputs are vertical ocean diffusivity (K_{bg}), anthropogenic aerosol scaling factor (A_{scl}) and climate sensitivity (C_{s}). Note that we converted longwave radiation feedback factor, which is one of the original input parameters for UVic, into C_{s} using a simple spline fit. We refer to Sriver et al. (2012) for the design points and details of the ensemble runs.

To avoid problems related to model artifacts and sparse sampling, we excluded data beyond 60°N and 80°S and 3000 m in depth [Bhat et al. (2012), Key et al. (2004), Schmittner et al. (2009)]. UVic ESCM outputs are on a 77 (latitude) × 100 (longitude) × 13 (depth) grid, but the number of locations that have nonmissing observations is 65,595. The missing values occur because there is no ocean at the locations in the UVic ESCM representation. At each grid point we compute a temporal mean over the time period of 1955–2006 to average out the effect of unresolved internal variability.

The observational data are on a 180 (latitude) × 360 (longitude) × 33 (depth) grid, and we remap this observed data into the UVic model grid using a linear interpolation using only the nearby points. See Figures A1 and A2 in the supplementary material [Chang et al. (2014)] for comparison between the UVic model

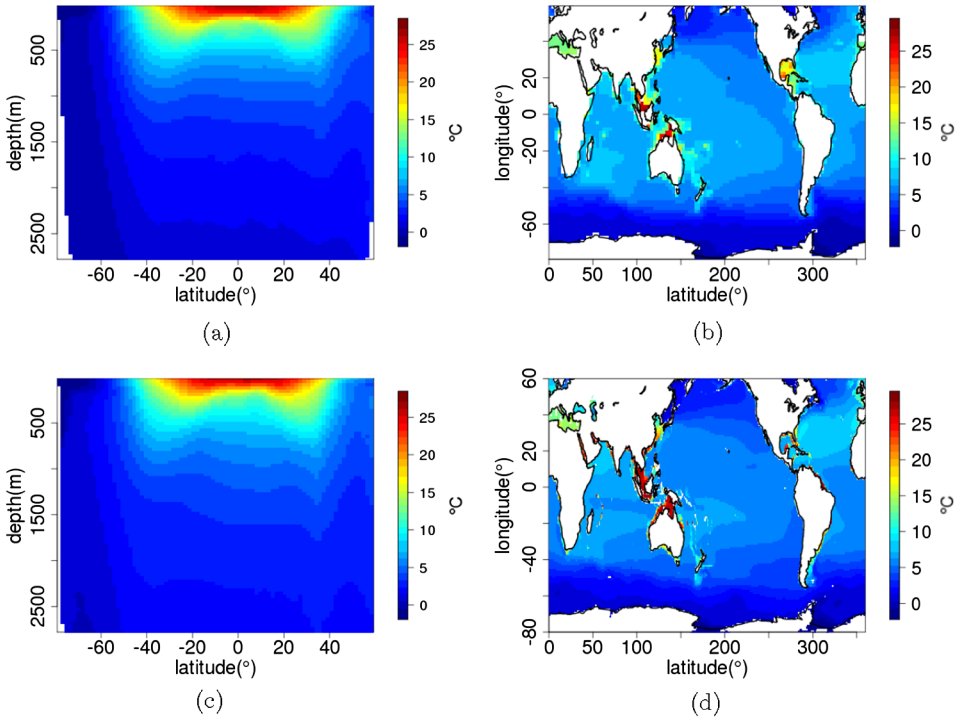

FIG. 1. *Plots of ocean temperature patterns averaged over* 1955–2006. *A UVic ESCM run* (*the first row*) *and the observational data from World Ocean Atlas* 2009 (*the second row*). *The left column shows the latitude–depth profiles of zonal mean, and the right column displays the longitude–latitude profiles of vertical mean. Note that the model run shown here is an example of* 250 *model runs described in Section* 2. (a) *and* (b) *Latitude–depth pattern* ($K_{\text{bg}} = 0.2$, $A_{\text{scl}} = 1.5$, $C_{\text{s}} = 3.976$). (c) *and* (d) *Latitude–depth pattern* (*observational data*).

grid and the observational data grid. This results in a relatively small reduction to 61,051 data points. The model output locations are also adjusted accordingly. We convert the observed in situ temperature field into the potential temperature field in order to (i) have the same measurement unit with UVic ESCM output and (ii) adjust the effect of pressure on ocean temperature. We obtain potential temperature from the in situ temperature [Locarnini et al. (2010)] and salinity fields [Antonov et al. (2010)] using the UNESCO equation of state [UNESCO (1981)] following Bryden (1973) and Fofonoff (1977). During the conversion procedure, we assume a simplified ocean pressure field varying as a function of latitude and depth [Lovett (1978)]. As we do for the UVic ESCM output, we compute a temporal mean over 1955–2006 at each location. Figure 1 shows examples of UVic ESCM model runs and the converted observational data.

3. Model calibration framework. Our computer model calibration framework consists of two stages, (i) model emulation and (ii) parameter calibration

[Bayarri et al. (2007), Bhat et al. (2012)]. First, we construct an emulator that interpolates computer model outputs at different parameter settings using Gaussian random fields [Sacks et al. (1989)]. This can be viewed as statistical interpolation or "kriging" [Cressie (1994)] in the computer model parameter space. Second, we infer the computer model parameters by relating observational data to computer model output using the emulator, while considering observational error and allowing for systematic discrepancies between the model and observations [Kennedy and O'Hagan (2001)]. Note that this two-stage approach has some advantages over fully Bayesian methods that combine the two stages into a single inferential step. By constructing an emulator solely based on computer model output [Bhat et al. (2012), Liu, Bayarri and Berger (2009), Rougier (2008)], this two-stage approach ensures that inference in the emulation stage is not contaminated by model discrepancy and observational error. In addition, separating the emulation stage from calibration provides an easier way to diagnose the accuracy of an emulator. Furthermore, computations are faster and parameter identifiability problems are reduced.

Let $Y(\mathbf{s}, \boldsymbol{\theta})$ denote the computer model output at the spatial location $\mathbf{s} =$ (longitude, latitude, depth) $\in \mathcal{S} \subseteq \mathbb{R}^3$ and the model parameter setting $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\mathcal{S}$ is the spatial domain of the process and $\boldsymbol{\Theta}$ is the computer model parameter space, typically a subset of a unidimensional or multidimensional Euclidean space. In our calibration problem, $\boldsymbol{\Theta} \subset \mathbb{R}^3$ since there are three input parameters. Furthermore, $Z(\mathbf{s})$ is the corresponding observation at the spatial location $\mathbf{s}$. Since each run of the climate model is computationally expensive, we can obtain computer model outputs only for a relatively small number of design points p. We denote these design points in the parameter space by $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p \in \boldsymbol{\Theta}$. Let $\mathbf{Y}_i \in \mathbb{R}^n$ be the computer model output at each parameter setting $\boldsymbol{\theta}_i$ for $i = 1, \ldots, p$. Each computer model output $\mathbf{Y}_i = (Y(\mathbf{s}_1, \boldsymbol{\theta}_i), \ldots, Y(\mathbf{s}_n, \boldsymbol{\theta}_i))^T$ is a spatial process observed at n different spatial locations $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$. In our calibration problem, $n = 61{,}501$ and $p = 250$. Let $\mathbf{Y}$ be the vector of concatenated computer model outputs such that $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_p)^T$. We denote the observed spatial process at n locations by $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$. Note that we assume that the locations for each model output and observation data are the same. If they are different, one can interpolate either of them depending on which one has a higher resolution. Our objective is to infer the parameter $\boldsymbol{\theta}$ by combining information from $\mathbf{Z}$ and $\mathbf{Y}$.

3.1. *Two-stage emulation and calibration.* We first outline our general framework for emulation and calibration.

Model emulation using Gaussian random fields. As described in Bhat et al. (2012), a standard approach to approximate the climate model output is using a Gaussian process such that

$$\mathbf{Y} \sim N\big(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\xi}_y)\big),$$

with a $np \times b$ covariate matrix $\mathbf{X}$ and a vector of regression coefficient $\boldsymbol{\beta}$. The covariates in $\mathbf{X}$ are the spatial locations (e.g., latitude, longitude and depth) and the climate parameters. The covariate matrix $\mathbf{X}$ contains all the spatial coordinates and the parameter settings used to define the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\xi}_y)$. The vector $\boldsymbol{\xi}_y$ contains all the parameters determining the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\xi}_y)$. In our application, since the mean term $\mathbf{X}\boldsymbol{\beta}$ is set to $\mathbf{0}$ (see Section 4.2.1 for more details), we fit a Gaussian random field to $\mathbf{Y}$ by finding the maximum likelihood estimate (MLE) of only $\boldsymbol{\xi}_y$, denoted by $\hat{\boldsymbol{\xi}}_y$.

The fitted Gaussian random field defines the probability model for the computer model output at any location $\mathbf{s} \in \mathcal{S}$ and parameter setting $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Therefore, the Gaussian process model provides a predictive distribution of computer model output at any untried value of $\boldsymbol{\theta}$ given the existing output $\mathbf{Y}$ [Sacks et al. (1989)]. We denote the resulting interpolated process by $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y})$ and call it an emulator process. This approach automatically provides a quantification of interpolation uncertainty.

Model calibration using Gaussian random field model. Once an emulator $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y})$ is available, we model the observational data $\mathbf{Z}$,

$$\mathbf{Z} = \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}) + \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is an independently and identically distributed observational error and $\boldsymbol{\delta}$ is a data-model discrepancy term. The discrepancy $\boldsymbol{\delta}$ is also modeled as a Gaussian process, thus, $\boldsymbol{\delta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_d(\boldsymbol{\xi}_d))$ with a spatial covariance matrix $\boldsymbol{\Sigma}_d(\boldsymbol{\xi}_d)$ between the locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and a vector of covariance parameters $\boldsymbol{\xi}_d$. The details regarding the specification of the covariance function are provided in Section 4.2.2. This discrepancy term is crucial for parameter calibration [cf. Bayarri et al. (2007), Bhat, Haran and Goes (2010)]. Note that this problem is ill posed without any prior information for $\boldsymbol{\xi}_d$, so an informative prior is necessary. Our inference for $\boldsymbol{\theta}$, $\boldsymbol{\xi}_d$ and σ^2 is based on their resulting posterior distribution.

3.2. *Challenges with high-dimensional data.* High-dimensional data sets pose considerable computational challenges due to the expensive likelihood function calculations that involve high-dimensional matrix computations. For instance, in the calibration problem described in Section 5, the dimensionality of the model output and the observational data is $n = 984$ in the 2-D case and $n = 61{,}051$ in the 3-D case, with $p = 250$. The latter example involves prohibitive computations with naïve implementations (discussed and explained in Section A1 in the supplementary material [Chang et al. (2014)]). For instance, with n-dimensional climate model outputs at p different parameter settings, evaluation of the likelihood function requires $\mathcal{O}(n^3p^3)$ operations. Therefore, numerical methods such as Newton–Raphson or MCMC algorithms become infeasible.

4. Model calibration with high-dimensional spatial data. We develop a dimension reduction approach based on spatial basis functions to increase computational efficiency. Spatial basis functions can map high-dimensional data into a low-dimensional space [Bayarri et al. (2007)] and find a representation of the probability model that results in lower computational cost for likelihood evaluations [Bhat et al. (2012), Higdon et al. (2008)]. Since there may be a trade-off between parsimony and accurate inference, it is crucial to find a set of spatial basis functions that gives a computationally feasible likelihood formulation without considerable loss of information. Below, we review drawbacks to the current approaches in the context of high-dimensional spatial data and propose a new approach to overcome these limitations.

4.1. *Current approaches*. Various methods have been introduced to overcome computational challenges with models for high-dimensional spatial data. These methods may be roughly grouped into the following three categories: low-rank representations, likelihood approximations and sparse covariance approximations. Low-rank representation methods such as kernel convolution [Higdon (1998)], Gaussian predictive process [Banerjee et al. (2008)] and fixed rank kriging [Cressie and Johannesson (2008)] approximate spatial processes using a set of basis functions and typically reduce the computational costs by using algorithms for patterned covariance matrices, for instance, the Sherman–Morrison–Woodbury formula. Likelihood approximation methods substitute the expensive likelihood function with a relatively inexpensive approximation such as the Whittle likelihood [Fuentes (2007)] or composite likelihood [Caragea and Smith (2006), Eidsvik et al. (2013), Stein, Chi and Welty (2004), Vecchia (1988)]. Sparse covariance approximation methods such as covariance tapering [Furrer, Genton and Nychka (2006)], Gaussian Markov random field approximations [Lindgren, Rue and Lindström (2011), Simpson, Lindgren and Rue (2012)] and lattice kriging [Nychka et al. (2013)] introduce sparsity in the covariance or precision matrix, thereby allowing for fast computations using sparse matrix algorithms.

A few different approaches to climate model calibration with multivariate computer model outputs have been developed in recent years, including Bayarri et al. (2007), Higdon et al. (2008) and Bhat et al. (2012). These approaches, however, are not readily applicable to the 3-D model output and observations we consider here due to the following reasons. First, in spite of the gains in computational efficiency, likelihood evaluations remain computationally prohibitive. The computational cost of a single likelihood evaluation in the emulation step in Bhat et al. (2012) scales as $\mathcal{O}(nJ^2)$ where J is the number of knots for the kernel basis. In Higdon et al. (2008), the computational cost scales with $\mathcal{O}(p^3 J_y^3)$ where J_y is the number of principal components used to represent the data. For the 3-D calibration problem we consider, n is 61,051 and J should be more than $3(K_v) \times 3(A_{\mathrm{scl}}) \times 3(C_{\mathrm{s}}) \times 3$ (depth) $\times$ 20 (longitude) $\times$ 15 (latitude) $= 24{,}300$ to ensure the number of knots to be at least three and greater than 20% of the design points for each dimension. The

number of principal components J_y needs be at least 20 to have more than 90% of explained variance. Second, the transformation based on the basis matrix may not be applicable to two- or three-dimensional spatial data. Using a wavelet transformation [Bayarri et al. (2007)] requires the same dyadic (a power of 2) number of data points for each spatial dimension, and the data need to be on a regular grid without missing values; irregular data and missing values are common in climate model calibration problems. In addition, conducting Bayesian inference on the joint posterior distribution may pose difficulties, both computationally as well as in terms of prior specification and identifiability issues. For example, Higdon et al. (2008) requires estimating $4 \times J_y + 1$ parameters, which translates to an 81-dimensional distribution for the 3-D case in Section 5.

4.2. *Reduced-dimensional model calibration*. Our method to overcome the aforementioned challenges relies on (i) representing the spatial field using a principal component basis and (ii) emulating each principal component separately. Instead of using a principal component basis to reduce the complexity of matrix computation as in Higdon et al. (2008), we use it to map the computer model outputs into a low-dimensional space and construct an emulator in that space directly. Since the principal components are uncorrelated by construction, we can build the emulator by constructing a 1-dimensional Gaussian process for each principal component in parallel. Fitting Gaussian random fields for each principal component requires estimation of only five parameters (see below). The likelihood evaluations involve covariance matrices of size $p \times p$. These features allow us to construct the emulator in a computationally efficient and highly automated manner. Moreover, since the principal component transformation can be applied to nondyadic spatial data with irregular locations, it has a broader range of application than wavelet transformations. In the calibration step, we develop an approach to map the observational data into a low-dimensional space.

4.2.1. *Computer model emulation*. The first step of this approach is to find the basis matrix for computer model output. We consider the computer model outputs as an n-dimensional data set with p replications and find the principal component basis. Let $\mathbf{M}$ denote the $p \times n$ matrix storing the computer model outputs $\mathbf{Y}_1, \ldots, \mathbf{Y}_p$ such that

$$\mathbf{M} = \begin{pmatrix} \mathbf{Y}_1^T \\ \vdots \\ \mathbf{Y}_p^T \end{pmatrix}. \tag{4.1}$$

Following the standard process of finding principal components, we first preprocess the computer model outputs to make the column means of the matrix $\mathbf{M}$ all 0's. Applying singular value decomposition (SVD), we find the scaled eigenvectors $\mathbf{k}_1 = \sqrt{\lambda_1}\mathbf{e}_1, \ldots, \mathbf{k}_p = \sqrt{\lambda_p}\mathbf{e}_p$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ are the ordered eigenvalues and $\mathbf{e}_1, \ldots, \mathbf{e}_p$ are the corresponding eigenvectors of the covariance matrix

of $\mathbf{M}$, where $J_y \ll p$ is the number of principal components that we decide to use in the emulator. One can choose the number of principal components by looking at the proportion of explained variation given by $\frac{\sum_{i=1}^{J_y}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$. We define the basis matrix for computer model output by $\mathbf{K}_y = (\mathbf{k}_1, \ldots, \mathbf{k}_{J_y})$.

For each parameter setting $\boldsymbol{\theta}_i$ $(i = 1, \ldots, p)$, the first J_y principal components $\mathbf{Y}_i^R = (Y_{i1}^R, \ldots, Y_{iJ_y}^R)^T$ are computed as

$$\mathbf{Y}_i^R = \left(\mathbf{K}_y^T\mathbf{K}_y\right)^{-1}\mathbf{K}_y^T\mathbf{Y}_i.$$

Let $\mathbf{Y}^R = (\mathbf{Y}_1^R, \ldots, \mathbf{Y}_p^R)^T$, hence, each element of this matrix $\{\mathbf{Y}^R\}_{ij} = Y_{ij}^R$ is the jth principal component at the ith computer model parameter setting. Since the columns in $\mathbf{K}_y$ are orthogonal, the principal components found here are uncorrelated to each other and this leads us to a parallelized emulator construction that is similar to the wavelet transformation approach in Bayarri et al. (2007). For each jth principal component across the parameter settings (i.e., for each jth column of $\mathbf{Y}^R$), we construct a Gaussian random field with the squared exponential covariance function such that

$$\operatorname{Cov}\left(Y_{kj}^R, Y_{lj}^R\right) = \kappa_{y,j}\exp\left(-\sum_{i=1}^{3}\frac{|\theta_{ik}-\theta_{il}|^2}{\phi_{y,ij}^2}\right) + \zeta_{y,j}1(\boldsymbol{\theta}_k = \boldsymbol{\theta}_l) \tag{4.2}$$

with partial sill $\kappa_{y,j}$, nugget $\zeta_{y,j}$ and range parameters $\boldsymbol{\phi}_{y,j} = (\phi_{y,1j}, \phi_{y,2j}, \phi_{y,3j})^T$. Leave-10-percent-out cross-validation experiments with 50 different randomly generated parameter configurations indicate that the squared exponential covariance shows a better fit than alternatives such as the exponential covariance (Figure A3 in the supplementary material [Chang et al. (2014)]). Note that this choice of covariance function may yield a fitted parameter field that is too regular, resulting in imprecise estimation of the range parameters $\phi_{y,ij}$, $i = 1, \ldots, 3$, $j = 1, \ldots, J_y$. However, our purpose here is not to estimate the range parameters precisely; the range parameters depend on the distance in the parameter space, which is a somewhat arbitrary notion. Note also that the mean term of each Gaussian process used here is set to be zero, since each of the principal components has zero mean across the parameter settings.

We denote the collection of emulator parameters for the jth principal component by $\boldsymbol{\xi}_{y,j} = (\kappa_{y,j}, \zeta_{y,j}, \boldsymbol{\phi}_{y,j})^T$. One can construct the emulator by finding the MLE $\hat{\boldsymbol{\xi}}_{y,j}$ for each j separately. The emulator $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R)$ is the collection of predictive processes of J_y principal components at $\boldsymbol{\theta}$ defined by the covariance function (4.2) and the MLEs $\hat{\boldsymbol{\xi}}_{y,1}, \ldots, \hat{\boldsymbol{\xi}}_{y,J_y}$. Note that even though we construct the emulator in terms of the principal components, we can make a projection $\mathbf{Y}^*$ in the original space at a new parameter setting $\boldsymbol{\theta}^*$ by computing

$$\mathbf{Y}^* = \mathbf{K}_y\boldsymbol{\eta}\left(\boldsymbol{\theta}^*, \mathbf{Y}^R\right).$$

To summarize, the emulation step uses the data $\mathbf{Y}_1^R, \ldots, \mathbf{Y}_{J_y}^R$ of dimension p and computes MLEs $\hat{\boldsymbol{\xi}}_{y,1}, \ldots, \hat{\boldsymbol{\xi}}_{y,J_y}$. Hence, the computational cost is reduced from $\mathcal{O}(n^3 p^3)$ to $\mathcal{O}(J_y p^3)$ when compared to a naïve approach. The resulting fitted model is then used for the calibration step as described in the following section.

4.2.2. *Computer model calibration.* Using $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R)$, the emulator for the principal components, we reformulate the model for observational data in (3.1) as

$$\mathbf{Z} = \mathbf{K}_y \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R) + \mathbf{K}_d \boldsymbol{\nu} + \boldsymbol{\varepsilon},$$

where $\mathbf{K}_d \boldsymbol{\nu}$ is a kernel convolution representation [Higdon (1998)] of the discrepancy $\boldsymbol{\delta}$. $\boldsymbol{\nu}$ is a vector of independent and identically distributed Normal random variates at $J_d \ll n$ locations, $\boldsymbol{\nu} \sim N(\mathbf{0}, \kappa_d \mathbf{I}_{J_d})$. $\mathbf{a}_1, \ldots, \mathbf{a}_{J_d} \in \mathcal{S}$. The variance parameter κ_d determines the magnitude of discrepancy, and the range parameters $\phi_{d,1}, \phi_{d,2} > 0$ specify the bandwidth of kernels. We define the kernel basis by

$$(\mathbf{K}_d)_{ij} = \exp\left(-\frac{g(s_{1i}, s_{2i}, a_{1j}, a_{2j})}{\phi_{d,1}} - \frac{|s_{3i} - a_{3j}|}{\phi_{d,2}}\right), \tag{4.3}$$

where s_{ki} and a_{kj} are the kth elements of $\mathbf{s}_i$ and $\mathbf{a}_j$, respectively. Our choice for the knot points $\mathbf{a}_1, \ldots, \mathbf{a}_{J_d}$ are on a grid of 15.6 degrees in latitude, 36 degrees in longitude and 429 m in depth. The design of these points does not affect the resulting process unless chosen to be too sparse. The geodesic distance function measures the great circle distance between two points on the Earth's surface. The function $g(s_{1i}, s_{2i}, a_{1j}, a_{2j})$ is given by

$$r \arccos\big(\sin(s_{2i}) \sin(a_{2j}) + \cos(s_{2i}) \cos(a_{2j}) \cos|s_{1i} - a_{1j}|\big),$$

where r is the radius of Earth (6378 km). By following Higdon et al. (2008), the range parameters are prespecified by scientific expert judgment; this reduces computations and identifiability issues (see Section 5 for specification of these parameters). The kernel function in (4.3) yields a valid covariance under geodesic distance since it is strictly positive definite on a sphere [Gneiting (2013)]. We assumed separability for distance along the surface and distance along the depth. The resulting process is approximately twice differentiable [Zhu and Wu (2010)], which produces a reasonable model for discrepancy. Even though the discrepancy model implies an isotropic discrepancy process [Higdon (2002)], the resulting process is flexible enough to capture the general trend in the discrepancy.

Instead of using the model (4.3) directly, we conduct calibration with reduced-dimensional data for computational efficiency. Let $\mathbf{Z}^R$ be a reduced version of the original data such that

$$\mathbf{Z}^R = (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{Z} = \begin{pmatrix} \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R) \\ \boldsymbol{\nu} \end{pmatrix} + (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \boldsymbol{\varepsilon},$$

where $\mathbf{K} = (\mathbf{K}_y \ \ \mathbf{K}_d)$. The probability model of $\mathbf{Z}^R$ is

$$\mathbf{Z}^R \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\eta}} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} & \mathbf{0} \\ \mathbf{0} & \kappa_d \mathbf{I}_{J_d} \end{pmatrix} + \sigma^2 (\mathbf{K}^T \mathbf{K})^{-1}\right), \tag{4.4}$$

where $\boldsymbol{\mu}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ are the mean and covariance given by the emulator $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R)$. It is often helpful to apply singular value decomposition to $\mathbf{K}_d$ and use the first $J_d^{PC} \ll J_d$ eigenvectors $\mathbf{K}_d^{PC}$ in place of $\mathbf{K}_d$ to find $\mathbf{Z}^R$. In addition to the obvious computational advantage, this often results in better inference since it corresponds to a regularized estimate given by ridge regression [see Hastie, Tibshirani and Friedman (2009), page 66]; this was corroborated by our extensive simulation studies.

Note that the term $\sigma^2(\mathbf{K}^T\mathbf{K})^{-1}$ in (4.4) automatically adjusts the contribution of each principal component to the calibration result. This can be illustrated by considering the model without the discrepancy, and the variance in the likelihood function is simply $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} + \sigma^2(\mathbf{K}_y^T\mathbf{K}_y)^{-1}$. Since $(\mathbf{K}_y^T\mathbf{K}_y)^{-1}$ is a diagonal matrix and its ith diagonal element is the reciprocal of the ith eigenvalue, $(\mathbf{K}_y^T\mathbf{K}_y)^{-1}$ inflates the variance of principal components with small eigenvalues. Therefore, the principal components with smaller explained variance will have less effect on the calibration result.

We now briefly examine the covariance structure implied by our model. Using the leading J_y principal components, the covariance between computer model outputs at two different spatial and parametric coordinates $(\mathbf{s}_1, \boldsymbol{\theta}_1)$ and $(\mathbf{s}_2, \boldsymbol{\theta}_2)$ can be written as

$$\begin{aligned}\operatorname{Cov}\big(Y(\mathbf{s}_1, \boldsymbol{\theta}_1), Y(\mathbf{s}_2, \boldsymbol{\theta}_2)\big) &\approx \operatorname{Cov}\left(\sum_{i=1}^{J_y} \sqrt{\lambda_i}\mathbf{e}_i(\mathbf{s}_1) w_i(\boldsymbol{\theta}_1), \sum_{j=1}^{J_y} \sqrt{\lambda_j}\mathbf{e}_j(\mathbf{s}_2) w_j(\boldsymbol{\theta}_2)\right) \\ &= \sum_{i=1}^{J_y} \lambda_i \mathbf{e}_i(\mathbf{s}_1)\mathbf{e}_i(\mathbf{s}_2) \operatorname{Cov}\big(w_i(\boldsymbol{\theta}_1), w_i(\boldsymbol{\theta}_2)\big),\end{aligned}$$

where $\mathbf{e}_i(\cdot)$ is the ith eigenfunction satisfying

$$\int \operatorname{Cov}\big(Y(\theta_1, \mathbf{s}_1), Y(\theta_2, \mathbf{s}_2)\big)\mathbf{e}_i(\mathbf{s}_2)\, d\mathbf{s}_1 = \lambda_i \mathbf{e}_i(\mathbf{s}_2) \operatorname{Cov}\big(w_i(\boldsymbol{\theta}_1), w_i(\boldsymbol{\theta}_2)\big),$$

with the corresponding eigenvalue λ_i. We let $w_i(\cdot)$ denote the Gaussian process of the ith principal component with the covariance function defined in (4.2). The leading eigenfunctions give the best approximation among all possible orthogonal bases since it minimizes the total mean square error [Jordan (1961)]. Since we can assume different covariance functions for each principal component process, our model can yield a nonseparable space-parameter covariance function. In contrast, if we were to assume separability such that

$$\operatorname{Cov}\big(Y(\boldsymbol{\theta}_1, \mathbf{s}_1), Y(\boldsymbol{\theta}_2, \mathbf{s}_2)\big) = C_{\mathrm{s}}(\mathbf{s}_1, \mathbf{s}_2) C_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

for some positive definite covariance functions C_s and C_θ, the covariance function for the ith principal component process becomes

$$\operatorname{Cov}\big(w_i(\boldsymbol{\theta}_1), w_i(\boldsymbol{\theta}_2)\big) = C_\theta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\lambda_i.$$

The detailed derivation is provided in Section A2 in the supplementary material [Chang et al. (2014)]. The separability assumption therefore results in a restrictive covariance structure such that the correlation functions for all principal component processes are the same. Hence, even though our reduced dimensional approach utilizes a covariance that is easy to specify, it provides a richer class of covariance functions than a separable covariance structure. Our cross-validation studies show that our assumption is adequate for emulating the computer model (see Section 5 for details).

Priors. We estimate the joint density of $\boldsymbol{\theta}$, κ_d and σ^2 using the Metropolis–Hastings algorithm. Following Bayarri et al. (2007), we allow for additional flexibility by estimating the partial sill parameters $\kappa_{y,1}, \ldots, \kappa_{y,J_y}$ for the emulator. Prior specification for the parameters in the observational model is straightforward. The discrepancy variance κ_d and the observational error variance σ^2 receive inverse-gamma priors with small shape parameter values. The prior for each parameter is a uniform distribution over a broad range or determined by scientific knowledge. In order to stabilize the inference, we put an informative prior to encourage $\kappa_{y,1}, \ldots, \kappa_{y,J_y}$ to vary around their estimated values in the emulation stage. See Section 5 for more details about prior specifications for our problem.

Computing. The computational costs may be summarized as follows:

(1) Find the basis matrix $\mathbf{K}_y = (\sqrt{\lambda_1}\mathbf{e}_1, \ldots, \sqrt{\lambda_{J_y}}\mathbf{e}_{J_y})$ by computing the singular value decomposition of $\mathbf{M}$ in (4.1). This computation is of order $\mathcal{O}(n^3)$, but needs to be done only once.

(2) Compute $\mathbf{Y}_R$ where its ith row is the transpose of $(\mathbf{K}_y^T\mathbf{K}_y)^{-1}\mathbf{K}_y^T\mathbf{Y}_i$.

(3) Construct a Gaussian random field for each column of $\mathbf{Y}_R$ by finding the MLE $\hat{\boldsymbol{\xi}}_{y,i}$ for each $i = 1, \ldots, J_y$. The computational cost is of order $\mathcal{O}(J_y p^3)$ for each likelihood evaluation.

(4) Compute $\mathbf{Z}_R = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{Z}$. The computational complexity of this step is $\mathcal{O}((J_y + J_d)^3)$.

(5) Using Metropolis–Hastings, draw an MCMC sample of $\boldsymbol{\theta}$, σ^2, κ_d and $\kappa_{y,1}, \ldots, \kappa_{y,J_y}$ from the joint posterior distribution based on the model in (4.4). The computational cost for each likelihood evaluation is of order $\mathcal{O}((J_y + J_d)^3)$. The overall cost of our implementation is $\mathcal{O}(pJ_y^3)$ for the emulation step and $\mathcal{O}((J_y + J_d)^3)$ for the calibration step.

5. Implementation details. We apply our method to data at three different aggregation levels. In the 1-D case, we compute the vertical means at $n = 13$ different depth points [Goes et al. (2010)]. In the 2-D case, the zonal means are computed at $n = 984$ latitude and depth points [Bhat et al. (2012)]. We use the original

pattern without any aggregation in the 3-D case ($n = 61{,}051$). The number of principal components is determined to have more than 90% of the explained variation. The number of components is 5 for the 1-D, 10 for the 2-D and 20 for the 3-D case. We also tried using 10 principal components for the 1-D, 20 for the 2-D and 30 for the 3-D case to have more than 95% explained variation, but did not find any improvement in the calibration result.

We use all 250 design points in the parameter space to build the emulator. We conducted leave-10-percent-out cross-validation and the results show that our emulator can predict the model output precisely. For the 2-D (latitude–depth) case, for instance, the emulator based on principal components can reproduce well the spatial pattern at any given parameter setting. More specifically, we randomly held out 25 model runs from the model output and then predicted these "hold outs" based on the remaining 225 model runs using our emulator. An example of our results is shown in Figure 2, which indicates that the predicted output and the original output are essentially indistinguishable. Other cross-validation results, including for the 3-D case, are similar. In addition, we found that the root mean squared error was very small relative to the scale of the data. We also tested the prediction performance of our emulator for the principal components using uncorrelated standardized prediction errors [Bastos and O'Hagan (2009)] (see Section A3 in the supplementary material [Chang et al. (2014)] for details). The graphical diagnostics in Figure A4 in the supplementary material [Chang et al. (2014)] show that our emulators predict the processes of leading principal components reasonably well.

We fix C_{s} and A_{scl} at the default values of the UVic ESCM in the calibration stage and make an inference only for K_{bg}. The default values are 1 for A_{scl} and 3.819 for C_{s}. One may choose to integrate out these two parameters, but since the ocean temperature field lacks strong information about A_{scl} and C_{s}, their estimated posterior densities are overly dispersed. This introduces unnecessary bias in the estimate of K_{bg} due to the highly nonlinear relationship between climate parameters [Olson et al. (2012)], thus, we decided not to integrate out those two parameters.

Following Bhat et al. (2012), we assume a flat prior with a broad range for K_{bg}, from 0.05 to 0.55. The variance for the observational error (σ^2) and the model discrepancy (κ_d) receive inverse-Gamma priors, and we denote them by $\mathrm{IG}(a_v, b_v)$ and $\mathrm{IG}(a_z, b_z)$. We set the shape parameters for them to be $a_v = 2$ and $a_z = 2$. To check the sensitivity of our approach to prior specifications, we tried four different combinations, (2, 2), (2, 100), (100, 2) and (100, 100) for b_v and b_z. The emulator variances $\kappa_{y,1}, \ldots, \kappa_{y,J_y}$ also receive inverse-Gamma priors with a shape parameter of 5. The scale parameters are determined to have modes at their estimated values in the emulation stage. Because of parameter identifiability problems which in turn affected computation, we fixed the range parameter for depth at 3000 m and

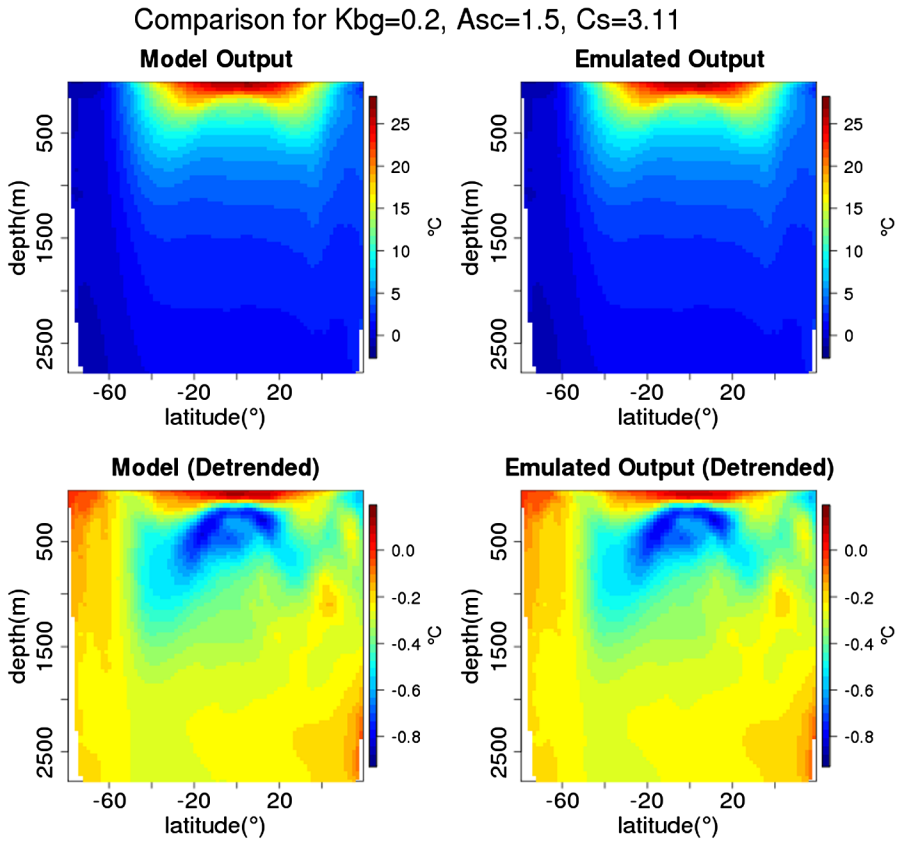

FIG. 2. *An example of our leave-*10*-percent-out cross-validation experiment for* 2*-D case* (*latitude–depth patterns*). *The first row shows the comparison between the raw original output and the raw emulated output. The second row shows the same comparison using detrended outputs, which are computed by subtracting the mean across the parameter settings at each location.*

for surface as 4800 km. The knot points $(\mathbf{a}_1, \ldots, \mathbf{a}_{J_d})$ are on a grid of 15.6 degrees in latitude, 36 degrees in longitude and 429 m in depth. The design of these points does not affect the resulting process unless chosen to be too sparse. We found that a wide range of the different settings for these parameters gave the same inference result for K_{bg} and, hence, our particular choices did not affect the results. The number of knot locations for the discrepancy kernel is 800 in the 3-D case, 80 in the 2-D case and 13 in the 1-D case. The number of principal components used for the discrepancy is 200 in the 3-D case, 20 in the 2-D case and 5 in the 1-D case. The number of principal components was determined using standard practice—by ensuring that at least 95% of the variability in the data was explained in each case. In order to check the robustness of our results, we tried different numbers of principal components. For example, when we increased the number of principal

components to 300 in the 3-D case, 30 in the 2-D case and 8 in the 1-D case we found that we obtained virtually identical calibration results.

Finally, we note that we ran an MCMC algorithm with 25,000 iterations for the calibration step. We carefully checked our results by comparing summaries (e.g., posterior density estimates) based on the first 15,000 runs with those obtained from the entire 25,000 runs and verified that our MCMC-based estimates are reliable.

6. Results.

6.1. *Computational benefit*. The biggest challenge in the considered analysis is the computational cost of evaluating the likelihood function in the 3-D case, which requires dealing with 61,051-dimensional spatial data sets. To our knowledge, current approaches cannot address this problem with reasonable computational effort (discussed below). In the emulation stage, the required number of principal components is about 20 for the 3-D case for reasonable accuracy and this means we still need to invert a $(p \times J_y) \times (p \times J_y) = 5000 \times 5000$ covariance matrix for the likelihood evaluation in the emulation stage if one uses the formulation due to Higdon et al. (2008). Moreover, the number of parameters to be estimated is $20 \times 4 + 1 = 81$ ($J_y \times$ the number of parameters for each Gaussian process emulator + one nugget parameter), and this also possibly increases computational cost significantly. The method of Bhat et al. (2012) requires multiplication of a $J \times n = 24{,}300 \times 61{,}501$ matrix into another $n \times J = 61{,}501 \times 24{,}300$ matrix in the likelihood evaluation, and this makes the likelihood evaluation computationally prohibitive.

The approximate computational time in emulation stage for the 1-D, 2-D and 3-D cases for different methods are illustrated in Figure 3. Our approximation was derived as follows:

$$\text{(computing time for our approach)} \times \frac{\text{(complexity for a method)}}{\text{(complexity for our approach)}}.$$

The computing time for our approach is for the PORT routine in R [David (1990), Gay (1983)] run on a system with Intel Xeon E5450 Quad-Core 3.0 GHz (without parallelization). Note that this approximated computing time for other approaches is optimistic; we believe that they will probably take longer than indicated. We use this to suggest that even when viewing the other approaches' cost optimistically, our method provides dramatically reduced computational time. We also note here that the computing time for our approach can be further reduced by parallelization. We describe some experimental results on computing time reduction using parallel computing in Section A4 in the supplementary material [Chang et al. (2014)].

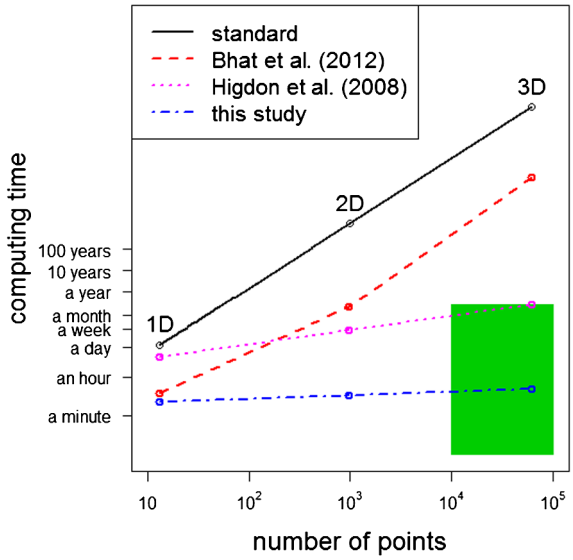

FIG. 3. *Comparison of computational costs for the emulation step between the current approaches and the new approach. The green box near the bottom right corner shows computing times that are practical, ranging from one second to three months.*

6.2. *The effect of data aggregation on climate model calibration.* In order to study the effect of data aggregation on climate model calibration, we conducted a study with pseudo-observational data. The simulated data are generated as follows:

(1) Choose the 3-D pattern of UVic ESCM output with $K_{\mathrm{bg}} = 0.2$, $A_{\mathrm{scl}} = 1.5$ and $C_{\mathrm{s}} = 3.976$ as the synthetic truth. The values for A_{scl} and C_{s} were selected based on the fact that they were the closest parameter values to the default values for the UVic model; the value for K_{bg} was obtained from the posterior mode from previously published work [cf. Bhat et al. (2012)].

(2) Compute three different 3-D patterns of residuals between the observational data (Wold Ocean Atlas 2009 data described above) and the UVic model outputs with $K_{\mathrm{bg}} = 0.1$, $K_{\mathrm{bg}} = 0.2$ and $K_{\mathrm{bg}} = 0.3$. The values for A_{scl} and C_{s} are the same as in step 1. Average them over each location to compute a pseudo-residual. This is a more realistic and challenging residual than one obtained by simulation from a simple error model, for example, a realization from a Gaussian process model. For brevity, we describe here just this most challenging case; our methods worked even better in terms of posterior variance when error processes were assumed to be simpler.

(3) Superimpose the pseudo-residual on the synthetic truth to construct pseudo-observational data in 3-D.

(4) Aggregate the 3-D pseudo-observational data into 2-D and 1-D by integrating the ocean temperature with respect to water volume.

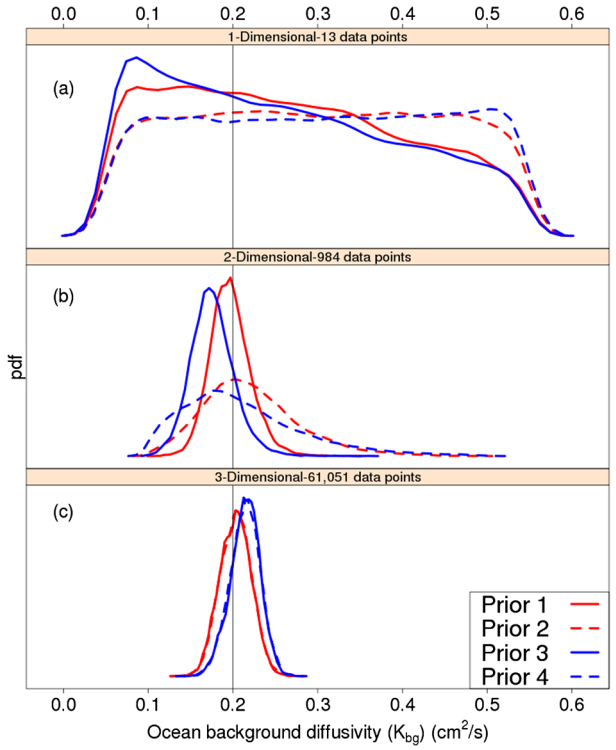

FIG. 4. *Prior sensitivity test in the simulated example. Calibration of* K_{bg} *value based on*: (a) 1-*D depth profile*, (b) 2-*D latitude–depth pattern*, (c) 3-*D nonaggregated data. Each line represents posterior density from four different priors*: ($b_v = 2, b_z = 2$) (*Prior* 1, *solid red line*), ($b_v = 2, b_z = 100$) (*Prior* 2, *dashed red line*), ($b_v = 100, b_z = 2$) (*Prior* 3, *solid blue line*) *and* ($b_v = 100, b_z = 100$) (*Prior* 4, *dashed blue line*). *The solid vertical line represents the true value of* K_{bg} *in the synthetic truth.* b_v *and* b_z *are hyperparameters of* κ_d *and* σ^2, *respectively, in* (4.4).

The calibration results based on this simulated example are shown in Figure 4. The sensitivity test indicates that the posterior distribution of K_{bg} in the 1-D and the 2-D cases relies on the specification of priors. This deep uncertainty is drastically reduced when the full data set (3-D) is used. A comparison result based on the real data from Ocean Atlas 2009 is shown in Figure 5. As in the simulated example, the calibration results based on the 3-D data are more robust to the prior specification.

In addition to the experiment above, we also examined the effect of random sampling of spatial locations, which may be considered a reasonable alternative to averaging over particular dimensions. As an illustrative example we randomly chose 1300 grid points from the 3-D locations using simple random sampling and calibrated K_{bg} based on only the selected points. By repeating the same experiment with 10 different random samples, we found that calibration results are substantially different across the samples (see Figure A5 in the supplementary material [Chang et al. (2014)]). These results indicate that sampling of spatial locations

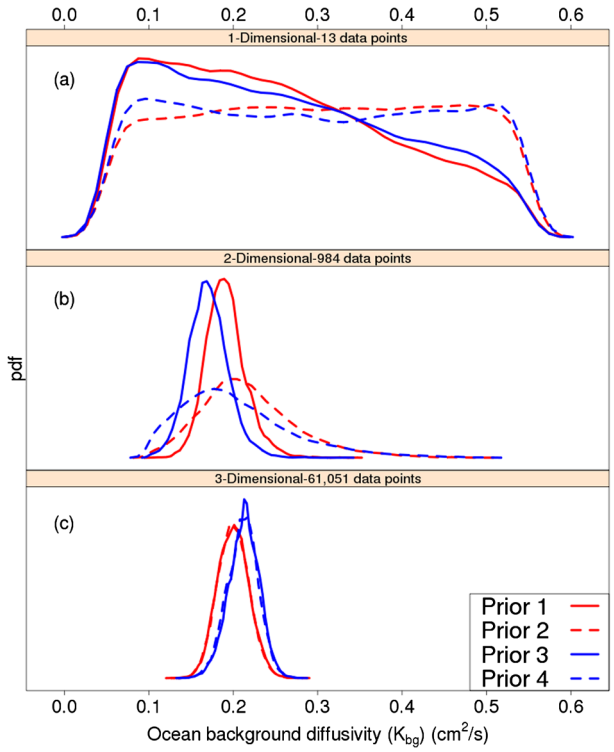

FIG. 5. *Prior sensitivity test using observational data from the World Ocean Atlas* 2009. *Calibration results based on*: (a) 1-*D depth profile*, (b) 2-*D latitude–depth pattern*, (c) 3-*D nonaggregated data*. *Each line represents posterior density from four different priors*: ($b_v = 2, b_z = 2$) (*Prior* 1, *solid red line*), ($b_v = 2, b_z = 100$) (*Prior* 2, *dashed red line*), ($b_v = 100, b_z = 2$) (*Prior* 3, *solid blue line*) *and* ($b_v = 100, b_z = 100$) (*Prior* 4, *dashed blue line*), *where* b_v *and* b_z *are hyperparameters of* κ_d *and* σ^2, *respectively*, *in* (4.4).

may introduce additional sampling errors to the calibration results; using all available data points is therefore desirable.

Using the full pattern of the 3-D data has important benefits, as it drastically reduces the deep uncertainty due to different prior specifications. We hypothesize that this is because the full nonaggregated spatial patterns contain more information about both the observational error and the discrepancy. In order to reflect the uncertainty due to prior choice to our density estimate for K_{bg}, we show in Figure 6 the posterior distributions when the prior is assumed to be with equal probability any one of the 4 priors considered, along with the resulting AMOC projections. We define AMOC projection as the annual maximum value of the meridional overturning streamfunction in the Atlantic between 0° and 70°N. The corresponding projections for AMOC change between 1970 to 1999 mean and 2070 to 2099 mean indicate that the unaggregated pattern gives a much narrower 95% predictive interval than the aggregated ones. Therefore, data aggregation increases the deep

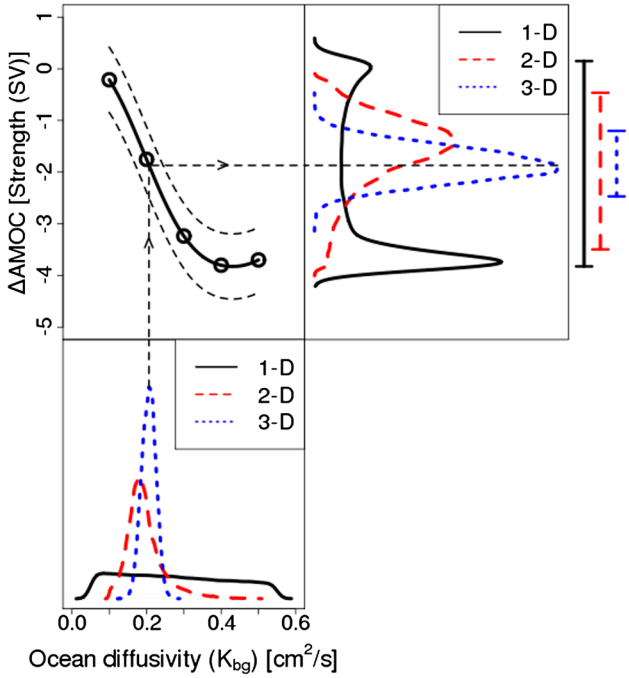

FIG. 6. *Combined posterior densities of* K_{bg} *from different prior specifications* (*lower left*), *the relationship between* K_{bg} *and projected AMOC change of the* 2070–2099 *mean from the* 1970–1999 *mean* (*upper left*), *and the resulting AMOC change projections* (*upper right*) *using* 1-*D* (*solid black line*), 2-*D* (*dashed red line*) *and* 3-*D* (*dotted blue line*) *data with their* 95% *credible intervals* (*bars at the right*).

uncertainty surrounding AMOC projection, and using unaggregated data reduces uncertainty regarding the future behavior of the AMOC.

7. Discussion.

7.1. *Summary.* We have considered here the problem of calibrating a climate model parameter, K_{bg}, in an Earth System Model of Intermediate complexity by using spatial observations of the potential temperature. In order to study the effects on both calibration and climate projections of using unaggregated versions of the data, we develop new methods that are computationally tractable for calibration with high-dimensional spatial data sets. Using our methods, we show that utilizing 3-D spatial data reduces the uncertainty about K_{bg} and is more robust to various prior specifications than calibration based on 2-D or 1-D aggregated versions of the data. The results suggest that using unaggregated data is valuable for reducing deep uncertainty associated with different priors. We note that we have tested our method in several other real data calibration problems and have obtained similar

results. For example, we carried out calibration for K_{bg} using a completely different set of observations, CFC-11 [Bhat et al. (2012)], and found again that using the 3-D pattern sharpened our inference about K_{bg} when compared to 2-D patterns.

We have demonstrated here that our computer model calibration approach is computationally efficient even when dealing with high-dimensional data. By exploiting the orthogonality of a principal components decomposition of the data, this method can keep the computational cost affordable for high-dimensional data with more than 60,000 spatial locations and 250 parameter settings. In addition, our simulated examples show that our approach can handle complicated model-observation discrepancies. The method can be easily extended to allow for calibration with multiple tracers—we can simply consider the variance–covariance matrix for all tracers and use its principal components to build an emulator. For the 2-D case, the posterior densities from using a separable covariance structure are not very different from our PCA-based approach; in fact, the bias appears to be slightly smaller for our approach (Figure A6 in the supplementary material [Chang et al. (2014)]). Moreover, our method results in a sharper density, which is an important criterion for calibration performance [cf. Gneiting, Balabdaoui and Raftery (2007)]. For larger data, it is more challenging to devise a simple but flexible approach that scales as well as the PCA-based method. We note here that our approach can be applied to even larger data sets. In climate science, data sets consisting of millions of data points are common. Our calibration approach in principle applies immediately to such data. A potential computational bottleneck is the SVD. However, we note that the SVD needs to be performed only once and we sidestep issues related to memory management since we do not have to store the large matrices. Furthermore, high-dimensional SVD is an active area of research and, hence, computationally efficient approaches are being developed [cf. Baglama and Reichel (2005), Halko, Martinsson and Tropp (2011)].

In the context of computer model calibration for making climate projections in general, we find that spatial data aggregation appears to generally lead to larger uncertainties (Figure 6), which is consistent with what we had hypothesized about the model and observations. This implies that the effect of information loss on the calibrated parameter is more important than the effect of reduction in model errors by aggregation. From a climate projections and decision-making perspective, the projections we obtain can be used as an input to integrated assessment or economic models; reduced uncertainties may therefore have tangible implications. Our method is immediately available for use with other climate models and observational products. By virtue of unlocking the full wealth of previously untapped information in large three-dimensional data sets, our method has a strong potential to improve projections of a host of policy-relevant climate variables.

7.2. *Caveats*. A general issue with principal components is also worth considering in this context: the principal components for the computer model outputs are selected based on explained variation and, thus, there is no guarantee that these

leading principal components carry the most important information about the climate parameters. However, our extensive study of the effect of changing the number of principal components suggests that this is not problematic in our context. Our results are consistent with the recent theoretical results in Artemiou and Li (2009) that suggest that there is a low probability that other (nonleading) principal components will have a strong correlation with the climate parameters. We hypothesize that our principal components-based approach does not lose valuable information about the climate parameters. In the discrepancy model, one important simplifying assumption is the separability between surface and depth effects. Our simulated example shows that the separability assumption provides a good approximation to the realistic discrepancy process. Nonseparable covariance function that combines geodesic distance and Euclidean distance remains as an avenue of ongoing and alive research and the subject of future work. Furthermore, our study of calibration with simulated examples shows that even though the number of K_{bg} settings at which the model is run is relatively sparse, there is enough information to reliably calibrate K_{bg} based on our emulator.

Our study is also subject to the usual caveats with respect to scientific conclusions. First, we ignore the interpolation uncertainty when we compute the density of AMOC projection based on the density of K_{bg}. Second, the result is based on a single data set and, thus, we cannot fully evaluate the effect of structural uncertainty due to the model-observation discrepancy and unresolved natural variability cannot be accounted for; this variability could impact conclusions as well [Olson et al. (2013)]. These caveats, of course, apply to almost all existing approaches to climate model calibration and projection.

Acknowledgments. The authors are very grateful to an anonymous referee, an Associate Editor and the Editor for very detailed and helpful comments and suggestions that have greatly improved this manuscript. All views, errors and opinions are solely that of the authors.

Author contributions. Won Chang and Murali Haran formulated the statistical method. Won Chang wrote all the computer code for emulation and calibration and wrote the first draft of the manuscript. Murali Haran edited the text. Klaus Keller designed the AMOC projection study and edited the text. Roman Olson supplied the previously-published UVic ESCM model runs and processed observational data from World Ocean Atlas 2009.

SUPPLEMENTARY MATERIAL

Supplement to "Fast dimension-reduced climate model calibration and the effect of data aggregation" (DOI: 10.1214/14-AOAS733SUPP; .pdf). Further technical details and supplementary figures can be found in this supplementary material.

REFERENCES

ALLEY, R., BERNTSEN, T., BINDOFF, N. L., CHEN, Z., CHIDTHAISONG, A., FRIEDLINGSTEIN, P., GREGORY, J., HEGERL, G., HEIMANN, M., HEWITSON, B., JOOS, F., JOUZEL, J., KATTSOV, V., LOHMANN, U., MANNING, M., MATSUNO, T., MOLINA, M., NICHOLLS, N., OVERPECK, J., QIN, D., RAGA, G., RAMASWAMY, V., REN, J., RUSTICUCCI, M., SOLOMON, S., SOMERVILLE, R., STOCKER, T. F., STOTT, P., WHETTON, P., WOOD, R. A. and WRATT, D. (2007). Climate Change 2007: The Physical Science Basis: Summary for Policymakers: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC.

ANTONOV, J. I., SEIDOV, D., BOYER, T. P., LOCARNINI, R. A., MISHONOV, A. V., GARCIA, H. E., BARANOVA, O. K., ZWENG, M. M. and JOHNSON, D. R. (2010). World Ocean Atlas 2009. In *NOAA Atlas NESDIS* 69 (S. Levitus, ed.). *Salinity* **2** 184. US Government Printing Office, Washington, DC.

ARTEMIOU, A. and LI, B. (2009). On principal components and regression: A statistical explanation of a natural phenomenon. *Statist. Sinica* **19** 1557–1565. MR2589197

BAGLAMA, J. and REICHEL, L. (2005). Augmented implicitly restarted Lanczos bidiagonalization methods. *SIAM J. Sci. Comput.* **27** 19–42 (electronic). MR2201173

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 825–848. MR2523906

BARNES, S. L. (1964). A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteorol.* **3** 396–409.

BASTOS, L. S. and O'HAGAN, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics* **51** 425–438. MR2756478

BAYARRI, M. J., BERGER, J. O., CAFEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R. J., PAULO, R., SACKS, J. and WALSH, D. (2007). Computer model validation with functional output. *Ann. Statist.* **35** 1874–1906. MR2363956

BHAT, K. S., HARAN, M. and GOES, M. (2010). Computer model calibration with multivariate spatial output: A case study. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M. H. Chen, P. Müller, D. Sun, K. Ye and D. K. Dey, eds.) 168–184. Springer, New York.

BHAT, K. S., HARAN, M., OLSON, R. and KELLER, K. (2012). Inferring likelihoods and climate system characteristics from climate models and multiple tracers. *Environmetrics* **23** 345–362. MR2935569

BRYAN, F. (1987). Parameter sensitivity of primitive equation ocean general circulation models. *J. Phys. Oceanogr.* **17** 970–985.

BRYDEN, H. L. (1973). New polynomials for thermal expansion, adiabatic temperature gradient and potential temperature of sea water. *Deep-Sea Res. Oceanogr. Abstr.* **20** 401–408.

CARAGEA, P. and SMITH, R. L. (2006). Approximate likelihoods for spatial processes. Preprint.

CHANG, W., HARAN, M., OLSON, R. and KELLER, K. (2014). Supplement to "Fast dimension-reduced climate model calibration and the effect of data aggregation." DOI:10.1214/14-AOAS733SUPP.

CRESSIE, N. (1994). Models for spatial processes. *Statistical Methods for Physical Science* **28** 93–124.

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 209–226. MR2412639

DAVID, M. G. (1990). Usage summary for selected optimization routines. Technical Report 153, AT&T Bell Laboratories Technology Computing Science, Murray Hill, NJ.

DRIGNEI, D., FOREST, C. E. and NYCHKA, D. (2008). Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling. *Ann. Appl. Stat.* **2** 1217–1230. MR2655656

EIDSVIK, J., SHABY, B., REICH, B., WHEELER, M. and NIEMI, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *J. Comput. Graph. Statist.* **23** 295–315. MR3215812

FOFONOFF, N. P. (1977). Computation of potential temperature of seawater for an arbitrary reference pressure. *Deep-Sea Res.* **24** 489–491.

FOREST, C. E., STONE, P. H. and SOKOLOV, A. P. (2008). Constraining climate model parameters from observed 20th century changes. *Tellus A* **60** 911–920.

FUENTES, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *J. Amer. Statist. Assoc.* **102** 321–331. MR2345545

FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261

GAY, D. M. (1983). Algorithm 611. Subroutines for unconstrained minimization using a model/trust-region approach. *ACM Trans. Math. Software* **9** 503–524. MR0791980

GNANADESIKAN, A. (1999). A simple predictive model for the structure of the oceanic pycnocline. *Science* **283** 2077–2079.

GNEITING, T. (2013). Strictly and nonstrictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. MR3102554

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 243–268. MR2325275

GOES, M., URBAN, N. M., TONKONOJENKOV, R., HARAN, M., SCHMITTNER, A. and KELLER, K. (2010). What is the skill of ocean tracers in reducing uncertainties about ocean diapycnal mixing and projections of the atlantic meridional overturning circulation? *J. Geophys. Res.-Oceans* **115** C10042.

HALKO, N., MARTINSSON, P. G. and TROPP, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53** 217–288. MR2806637

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HIGDON, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat.* **5** 173–190.

HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer, London. MR2059819

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

JORDAN, K. L. (1961). Discrete representations of random signals. Technical Report 378, Massachusetts Institute of Technology, Cambridge, MA.

KELLER, K., HALL, M., KIM, S.-R., BRADFORD, D. F. and OPPENHEIMER, M. (2005). Avoiding dangerous anthropogenic interference with the climate system. *Clim. Change* **73** 227–238.

KELLER, K., DEUTSCH, C., HALL, M. G. and BRADFORD, D. F. (2007). Early detection of changes in the North Atlantic meridional overturning circulation: Implications for the design of ocean observation systems. *J. Climate* **20** 145–157.

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

KEY, R. M., KOZYR, A., SABINE, C. L., LEE, K., WANNINKHOF, R., BULLISTER, J. L., FEELY, R. A., MILLERO, F. J., MORDY, C. and PENG, T. H. (2004). A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). *Glob. Biogeochem. Cycles* **18** GB4031.

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 423–498. MR2853727

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

LOCARNINI, R. A., MISHONOV, A. V., ANTONOV, J. I., BOYER, T. P., GARCIA, H. E., BARANOVA, O. K., ZWENG, M. M. and JOHNSON, D. R. (2010). World Ocean Atlas 2009. In *NOAA Atlas NESDIS* 68 (S. Levitus, ed.). *Temperature* **1** 184. US Government Printing Office, Washington, DC.

LOVETT, J. R. (1978). Merged seawater sound-speed equations. *J. Amer. Statist. Assoc.* **63** 1713–1718.

NYCHKA, D., BANDYOPADHYAY, S., HAMMERLING, D., LINDGREN, F. and SAIN, S. (2013). A multi-resolution Gaussian process model for the analysis of large spatial data sets. NCAR Technical Note NCAR/TN-504+STR. DOI:10.5065/D6VH5KTJ.

OLSON, R., SRIVER, R., GOES, M., URBAN, N. M., MATTHEWS, H. D., HARAN, M. and KELLER, K. (2012). A climate sensitivity estimate using Bayesian fusion of instrumental observations and an Earth System model. *J. Geophys. Res.-Atmos.* **117** D04103.

OLSON, R., SRIVER, R., HARAN, M., CHANG, W., URBAN, N. M. and KELLER, K. (2013). What is the effect of unresolved internal climate variability on climate sensitivity estimates? *J. Geophys. Res.-Atmos.* **118** 4348–4358.

ROUGIER, J. (2008). Comment on article by Sansó et al. [MR2383247]. *Bayesian Anal.* **3** 45–56. MR2383249

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

SANSÓ, B. and FOREST, C. (2009). Statistical calibration of climate system properties. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 485–503. MR2750089

SCHMITTNER, A., URBAN, N. M., KELLER, K. and MATTHEWS, D. (2009). Using tracer observations to reduce the uncertainty of ocean diapycnal mixing and climate–carbon cycle projections. *Glob. Biogeochem. Cycles* **23** 737–750.

SIMPSON, D., LINDGREN, F. and RUE, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics* **23** 65–74. MR2873784

SRIVER, L., NATHAN, M., OLSON, R. and KELLER, K. (2012). Toward a physically plausible upper bound of sea-level rise projections. *Clim. Change* **115** 893–902.

STEIN, M. L., CHI, Z. and WELTY, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 275–296. MR2062376

UNESCO (1981). Tenth Report of the joint panel on oceanographic tables and standards. Technical Reports on Marine Science 36, UNESCO.

VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. R. Stat. Soc. Ser. B* **50** 297–312. MR0964183

WEAVER, A. J., EBY, M., WIEBE, E. C., BITZ, C. M., DUFFY, P. B., EWEN, T. L., FANNING, A. F., HOLLAND, M. M., MACFADYEN, A. and MATTHEWS, H. D. (2001). The UVic Earth System climate model: Model description, climatology, and applications to past, present and future climates. *Atmos.-Ocean.* **39** 361–428.

WUNSCH, C. and FERRARI, R. (2004). Vertical mixing, energy, and the general circulation of the oceans. *Annu. Rev. Fluid Mech.* **36** 281–314.

ZHU, Z. and WU, Y. (2010). Estimation and prediction of a class of convolution-based spatial nonstationary models for large spatial data. *J. Comput. Graph. Statist.* **19** 74–95. MR2654401

W. CHANG
M. HARAN
DEPARTMENT OF STATISTICS
PENN STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: wonchang@psu.edu
mharan@stat.psu.edu

R. OLSON
K. KELLER
DEPARTMENT OF GEOSCIENCES
PENN STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: rzt2-wrk@psu.edu
klaus@psu.edu