ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)
68, *Part* 1, *pp.* 51–78

Computer model calibration with large non-stationary spatial outputs: application to the calibration of a climate model

Kai-Lan Chang and Serge Guillas

University College London, UK

[Received June 2017. Revised July 2018]

Summary. Bayesian calibration of computer models tunes unknown input parameters by comparing outputs with observations. For model outputs that are distributed over space, this becomes computationally expensive because of the output size. To overcome this challenge, we employ a basis representation of the model outputs and observations: we match these decompositions to carry out the calibration efficiently. In the second step, we incorporate the non-stationary behaviour, in terms of spatial variations of both variance and correlations, in the calibration. We insert two integrated nested Laplace approximation–stochastic partial differential equation parameters into the calibration. A synthetic example and a climate model illustration highlight the benefits of our approach.

Keywords: Gaussian process; Integrated nested Laplace approximation–stochastic partial differential equation; Matérn fields; Uncertainty quantification

1. Introduction

Complex computer models are widely used in various fields of science and technology to mimic complex physical systems. Computer model calibration involves comparing the simulations of a complex computer model with the physical observations of the process being simulated. Increasingly, computer model outputs are in the form of spatial fields, particularly in environmental sciences. This poses a particular challenge to the calibration method.

The class of models that we consider in this paper is computer models with parametric inputs of reasonable dimension (say below 20), and outputs distributed over two dimensions over the plane or the sphere. This is unlike the formulation of Kennedy and O'Hagan (2001), which is usually applied to scalar outputs. Our motivations come from climate modelling. Climate scientists compare model outputs at a certain relevant altitude distributed over the sphere, typically over a grid (along latitude and longitude), with a spatial data set of observations at the same altitude.

In this paper, we develop our Bayesian calibration technique based on the framework of Kennedy and O'Hagan (2001): we approximate the expensive computer model by a Gaussian process (GP). This formulation has proven to be effective in a wide range of applications. However, the GP calibration is computationally expensive for large model output spaces (cubic complexity in the number of output points that are used to fit the GP due to the Cholesky decomposition). Therefore several attempts to tackle this issue in the context of times series

Address for correspondence: Kai-Lan Chang, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK.
E-mail: ucakkac@ucl.ac.uk

 0035–9254/19/68051
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.

of outputs or spatial outputs have been made either by using truncated basis representations of model outputs to reduce dimension (Bayarri *et al.*, 2007; Higdon *et al.*, 2008; Chang *et al.*, 2014; Holden *et al.*, 2015), or by using a separable covariance function over space and tuning parameters to build a theoretical emulator for multivariate outputs (Rougier, 2008; Bhat *et al.*, 2010). We provide here a solution that makes use of an adequate representation of the spatial outputs using Gaussian fields (GFs).

GFs play an important role in spatial statistics. The traditional approach is to specify a GF through its covariance function. Another approach is to use the class of Gaussian Markov random fields, which are discretely indexed GFs. The Markov property yields a sparse precision matrix, so efficient numerical algorithms can be employed. Lindgren *et al.* (2011) showed that the Gaussian Markov random field representation can be constructed explicitly by using a certain form of stochastic partial differential equation (SPDE) which has a GF with Matérn covariance as its solution. The representation employs piecewise linear basis functions, and Gaussian weights with Markov dependences determined by the finite element method over a triangulation of the domain. This technique can deal with large spatial data sets and naturally accounts for non-stationarity. Our paper combines the strengths of the calibration formulation with a truncated basis, and the SPDE-defined scale and precision parameterization to deal with large-scale spatial outputs, and still provides a compromise with computational feasibility to employ a fully Bayesian approach.

1.1. Challenge in Bayesian calibration

Among existing approaches of using the basis representation of model outputs, dimension reduction is carried out mostly by data-driven basis functions, i.e. principal components (PCs), also known as empirical orthogonal functions; see Higdon *et al.* (2008). Data-driven basis functions offer a computationally efficient approach to adapt the outputs. For computer model calibration of spatial outputs, this approach ignores the nature of the spatial dependence of the outputs, treating spatial data as a multivariate vector.

Since the dimension of the input space for known input parameters is 2 (the location in space), we could employ the usual calibration framework (Kennedy and O'Hagan, 2001). However, this framework can deal with only a few thousand output points at these input locations. But climate models produce outputs over large regular grid cells; for example our climate model uses a grid of $n = 96 \times 144 = 13824$ cells, and this is at a rather coarse choice of resolution. We calibrate four input parameters with $r = 100$ runs; thus the number of computer runs r, multiplied by the output size n, creates a data matrix that is too large to fit a GP, which is an impossible task for a fully Bayesian calibration (cubic complexity in the total number of output points to fit the GP). Hence our approach aims to reduce the large amount of model outputs with a smaller basis representation that makes use of the spatial dependence to extract key pieces of information, instead of using all the output cells. Our approach involves transforming a large scalar output over space into a much smaller set of scalars by using a spherical harmonics representation and the SPDE technique.

1.2. Atmospheric chemistry model output

We consider that an atmospheric chemistry model discretizes the Earth's surface into a three-dimensional grid of cells over time, which can be characterized by horizontal (latitude and longitude), vertical (altitude or pressure level) and temporal resolutions. The output in each cell is parameterized by complex mathematical equations that describe the chemistry species in it and the physical circulation through it. The four-dimensional interactions of climate dynamics

are currently beyond our scope for the calibration. Our paper focuses only on the horizontal variations. Our practical interest is to tune, and quantify uncertainty in, climate experiments. The 'Whole atmosphere community climate model' (WACCM) is a general circulation model of the middle and upper atmosphere. The WACCM is an extension of the National Center for Atmospheric Research 'Community Earth system model'. Many parameterizations of physical processes have to be set to run the WACCM, resulting in potential concerns about error growth (Liu *et al*., 2009).

To describe the general framework, let $\eta(\mathbf{s}_i, \boldsymbol{\theta}_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, r$, be the r-runs model outputs measured at n locations. Here we refer to $m = n \times r$ as the total number of outputs in the simulations. We choose a design made of combinations of input values, and we impose distributional prior assumptions on the inputs. The aim of calibration is to estimate the best input setting $\boldsymbol{\theta}^*$ to match outputs to observations, and to investigate the discrepancy between observations and optimized outputs. Note that, in terms of the calibration framework of Kennedy and O'Hagan (2001), our experiment does not have 'known variable parameters'; output cells are prescribed as a resolution in the climate model, and thus the spatial variations in different model runs are completely differentiated by calibration inputs $\boldsymbol{\theta}$.

For each single run, the WACCM simulates output over a grid of $n = 96 \times 144 = 13824$ cells. We explore the zonal wind outputs over the sphere, varying according to four gravity wave (GW) input parameters with $r = 100$ runs, to calibrate the GW parameters. The number of computer runs r, multiplied by the output size n, is too large to fit a GP to the computer model and thus challenges the fully Bayesian calibration to be performed.

1.3. The propagation of gravity waves

In climate modelling, the GWs parameterization aims to reduce zonal mean wind biases. Small modification of parameterized GWs can have large effects by improving the propagation pathways of the Rossby waves (Alexander and Sato, 2015). GWs also play a dominant role in driving the quasi-biennial oscillation (QBO), which is a dynamic process of zonal mean zonal winds from eastward to westward in the tropical stratosphere. GWs, which are also called small-scale atmospheric waves, generate a wide range of short horizontal wavelengths from mesoscale to thousands of kilometres (Ern *et al*., 2014), and an even wider range of the processes impacted by GWs (turbulence scales to planetary scales) (Liu *et al*., 2014). It is thus a challenge to simulate numerically all small waves and their cumulated effects that contribute to the QBO pattern based on global observations (Alexander *et al*., 2010; Geller *et al*., 2013; Yu *et al*., 2017).

1.4. Outline of this paper

We propose to use a fixed spatial basis, like Bayarri *et al*. (2007) did by employing a wavelets basis to describe functional model outputs. Our approach is also related to recent multiresolution methods on spatial data (Nychka *et al*., 2002, 2015; Ilyas *et al*., 2017). With a fixed basis, we can easily compare model outputs with observations over space. In addition, the use of a fixed basis facilitates the quantification of the non-stationarity across space in the SPDE model.

In Section 2 we present our approach in detail. We employ a truncated basis representation, such as a B-splines decomposition or spherical harmonics transforms, to capture the output features spatially. We then explore how parameters in an SPDE model can explicitly quantify the non-stationarity of the spatial field (Bolin and Lindgren, 2011; Blangiardo and Cameletti, 2015; Zammit-Mangion *et al*., 2015; Liu *et al*., 2016): we extend our approach by including spatially varying scale and precision parameters in an SPDE model in our calibration framework. We

then apply these techniques to a synthetic example in Section 3 and our real climate experiment in Section 4. Finally, in Section 5, we discuss potential improvements to our approach.

2. Methods

To address the challenges of the uncertainty quantification at both global and local scales, and to maintain computational feasibility for the Bayesian calibration, we pursue a sophisticated effort that approximates the spatial variations effectively and efficiently. In Section 2.1, we adopt a reduced rank spatial basis representation to capture the large-scale spatial variability. In the next step in Section 2.2, we review the spatial modelling technique through the SPDE approach and highlight its strength in capturing local spatial structures. We then combine these two approaches into the calibration framework in Section 2.3 and provide guidance for the implementation in Section 2.4.

2.1. Basis representation for the model output

In this section, we decompose spatial outputs and observations onto a basis of real-valued basis functions, such as *B*-splines or spherical harmonics. We parsimoniously represent these surfaces and construct a methodology for the calibration that makes use of the coefficients in these representations. We follow the Bayesian calibration setting of Kennedy and O'Hagan (2001). Let $\boldsymbol{\theta}$ be the calibration parameters. The output $\eta(\cdot)$ is computed at inputs $(\mathbf{s}, \boldsymbol{\theta})$ in an m-point experimental design, where $m = n \times r$ means r computer runs measured at n locations. The output $\eta(\mathbf{s}, \boldsymbol{\theta})$ is an approximation of the reality $y^{\mathrm{R}}(\mathbf{s})$. The discrepancy between the simulator and the reality at the spatial locations is denoted $\delta(\mathbf{s})$. The field data or observations $y^{\mathrm{F}}(\mathbf{s})$ of the reality are collected at a number of locations $\mathbf{s}$ in an n-point spatial design (here a simple grid) and are subject to a normal observation error $\epsilon(\mathbf{s})$ with a constant variance across locations. The measurement locations for observations and outputs can be different, since the methodology accommodates such variation. The main equation is

$$y^{\mathrm{F}}(\mathbf{s}) = y^{\mathrm{R}}(\mathbf{s}) + \epsilon(\mathbf{s}) = \eta(\mathbf{s}, \boldsymbol{\theta}^*) + \delta(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{1}$$

This formulation includes both parameter uncertainty and model discrepancy; however, it is difficult to distinguish the uncertainty in the calibration parameters from discrepancy in real applications due to lack of identifiability (Brynjarsdóttir and O'Hagan, 2014). Note that output cells are prescribed as a model resolution; the uncertainties in η are completely determined by $\boldsymbol{\theta}$. We use a set of spatial basis functions $\{\psi_z\}$, where z is an integer that represents the index number within the ordered basis, to decompose each run of model output over space. Precisely, for the N_ηth level of expansion and for each run j,

$$\eta(\mathbf{s}, \boldsymbol{\theta}_j) = \sum_{z=1}^{N_\eta} c_z^{\mathrm{M}}(\boldsymbol{\theta}_j)\psi_z(\mathbf{s}) \qquad j = 1, \ldots, r.$$

We assume that the approximation error in this representation is ignorable (i.e. we expect that, the more bases, the lower the approximation error). The coefficients $\{c_z^{\mathrm{M}}\}$ represent the surface features at different levels of expansion. Similarly to Nychka *et al.* (2015), we conjecture that different spatial basis functions will be valid for this representation, such as the Wendland family (Wendland, 2004) that was used in Nychka *et al.* (2015) or popular spline-based approaches (Wood, 2003; Williamson *et al.*, 2012; Chakraborty *et al.*, 2013; Bowman and Woods, 2016; Chang *et al.*, 2017). The observations can be written as (with an associated approximation error ignored)

$$y^{\mathrm{F}}(\mathbf{s}) = \sum_{z=1}^{N_y} c_z^{\mathrm{F}} \psi_z(\mathbf{s}).$$

The physical spaces of both model outputs and observations are transformed into a functional space that is spanned by the fixed basis. Since the aim is to calibrate the spatial outputs, we also assume that the reality $y^{\mathrm{R}}(\mathbf{s})$, the discrepancy function $\delta(\mathbf{s})$ and the measurement errors $\epsilon(\mathbf{s})$ can be represented by similar basis representations, albeit with more levels of variation than model outputs:

$$y^{\mathrm{R}}(\mathbf{s}) = \sum_{z=1}^{N_y} c_z^{\mathrm{R}} \psi_z(\mathbf{s}),$$
$$\delta(\mathbf{s}) = \sum_{z=1}^{N_y} c_z^{\delta} \psi_z(\mathbf{s}),$$
$$\epsilon(\mathbf{s}) = \sum_{z=1}^{N_y} c_z^{\epsilon} \psi_z(\mathbf{s}).$$

Indeed, the computer model does not include all possible physical processes that affect the measurements. Hence, the spatial outputs from the computer simulation should be relatively smoother than the observations. Therefore we assume a larger number of basis functions, N_y, in the observations (automatically as well in the discrepancy and error functions) than for model outputs (N_η, $N_\eta \leqslant N_y$). In the formulation of the calibration algorithm, we introduce coefficients $\{c_z^{\mathrm{M}} | N_\eta < z \leqslant N_y\}$, all set to be $\mathbf{0}$. Indeed, we can then use the same number of basis functions N_y to decompose y^{F} and η. Then matching the coefficients in equation (1) yields

$$c_z^{\mathrm{F}} = c_z^{\mathrm{R}} + c_z^{\epsilon} = c_z^{\mathrm{M}}(\boldsymbol{\theta}^*) + c_z^{\delta} + c_z^{\epsilon}, \qquad z = 1, \ldots, N_y. \tag{2}$$

Hence, only the relatively smooth variations of the computer model match the variations in observations. At this point we seek to capture only the large-scale variability derived from calibration parameters; local structures will be accounted for in Section 2.2. The weights for the measurement errors, c_z^{ϵ}, are assumed to follow $N(0, 1/\lambda_\epsilon)$.

2.1.1. Gaussian process for the transformed coefficients

The GP assumption is imposed on each coefficient $c_z^{\mathrm{M}}(\boldsymbol{\theta})$, $z = 1, \ldots, N_y$, of mean 0 and with a covariance function

$$\mathrm{cov}\{c_z^{\mathrm{M}}(\boldsymbol{\theta}), c_{z'}^{\mathrm{M}}(\boldsymbol{\theta}')\} = \frac{1}{\lambda_\eta} I_{zz'} \prod_{k=1}^{q} \rho_{\eta k}^{2^{\gamma_{\eta k}} |\theta_k - \theta_k'|^{\gamma_{\eta k}}}, \tag{3}$$

where $I_{zz'}$ is the Kronecker delta ($I_{zz'} = 1$ if $z = z'$ and $I_{zz'} = 0$ otherwise), q is the dimension of $\boldsymbol{\theta}$, λ_η controls the marginal precision of $\eta(\cdot, \cdot)$ and $\boldsymbol{\rho}_\eta$ controls the strength of the dependence in each of the pairs of $\boldsymbol{\theta}$. To simplify the complexity and because the computer model response to input tunes is nearly smooth and continuous, it is generally reasonable to assume that $\boldsymbol{\gamma}_\eta = 2$ (Sacks *et al.*, 1989; Higdon *et al.*, 2004; Linkletter *et al.*, 2006). Note that the coefficients $\{c_z^{\mathrm{M}}\}$ must be scaled to the unit hypercube; otherwise this covariance model is not appropriate. This reparameterization of the square exponential covariance leads to a smooth and infinitely differentiable representation for the model output (Stein, 1999). In addition, coefficients that are associated with the same basis ψ_z form a block in the covariance structure, and we assume that the correlation between different indices z is 0. Hence the rN_y-vector $\mathbf{c}^{\mathrm{M}}$ has a multivariate

normal prior with mean 0 and a covariance matrix with $r \times r$ N_y-blocks in the diagonal, and the off-diagonal blocks are zero matrices.

The strong assumption of independence of the coefficients, through different blocks in the covariance, may not be fully justifiable in all real applications. Indeed, it is possible that certain physical properties propagate across multiple scales (but, even in that case, it may not constitute a large proportion of the variation). However, this assumption leads to a great computational advantage in terms of forming a block diagonal covariance model in the GP model. Traditionally a GP fitting involves a complexity of $O(m^3) = O(n^3 r^3)$ and a storage cost of $O(m^2) = O(n^2 r^2)$. In our approach the complexity and cost of our model are $O(N_y^3 r^3)$ and $O(N_y^2 r^2)$, where $N_y \ll n$. The block diagonal assumption further reduces the complexity and cost to $O(N_y r^3)$ and $O(N_y r^2)$. In a simulation study we discuss how this assumption is a compromise between fidelity and complexity.

The decomposed discrepancy term c_z^δ quantifies the inadequacy between the simulator and reality in the functional domain. We assume that each c_z^δ follows a normal distribution of mean 0 and with a covariance function

$$\operatorname{cov}(c_z^\delta, c_{z'}^\delta) = \frac{1}{\lambda_\delta} I_{zz'}. \tag{4}$$

There is no conceptual difference in the model bias between our setting and another setting that relies on a projection onto a basis (e.g. the PC approach), but there are differences in the ability to pin down the biases concretely and adequately. Indeed, our approach allows the bias to represent complex ranges of variations (due to its expression in a basis and the addition of non-stationarity in what follows in this paper). Note that, among existing studies identifying climate model biases, most of the biases display a systematic tendency (either underestimation or overestimation) across certain regions (Jun *et al.*, 2008; Lamarque *et al.*, 2013; Wang *et al.*, 2014; Williamson *et al.*, 2015) and thus a non-stationarity feature is desirable.

All the unknown parameters in the algorithm require specified prior distributions which represent uncertainty about the values of these parameters. The following choices are made for the priors.

(a) To represent our vague knowledge about calibration parameters, we specify a uniform prior distribution over each of the calibration parameter intervals.
(b) To model the correlation parameters $\rho_{\eta_k}, k = 1, \ldots, q$, a beta(1, 0.1) distribution is used, which conservatively places most of its prior mass on values of $\boldsymbol{\rho}_\eta$ near 1 (indicating an insignificant effect).
(c) Gamma prior distributions are used for each of the precision parameters λ_η, λ_δ and λ_ϵ. Specifically, we use priors $\lambda_\eta \sim$ GAM(5, 5) (with expectation 1 due to standardization of the responses), $\lambda_\delta \sim$ GAM(1, 0.01) (with expectation around 10% of the standard deviation SD of the standardized responses) and $\lambda_\epsilon \sim$ GAM(1, 0.003) (with expectation around 5% of the SD of the standardized responses).

2.1.2. The posterior distributions

In this stage, all the r-run model outputs and observations, measured over an n-grid of cells, are reduced to transformed coefficients. Denote the joint $(r+1)N_y$ data vector $\mathbf{D} = (c^{\mathrm{F}}, c^{\mathrm{M}})$. The sampling likelihood for the full data is then

$$L(\mathbf{D}|\boldsymbol{\theta}, \lambda_\eta, \boldsymbol{\rho}_\eta, \lambda_\delta, \Sigma_\epsilon) \propto |\Sigma_{\mathbf{D}}|^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{D}^{\mathrm{T}} \Sigma_{\mathbf{D}}^{-1} \mathbf{D})\}, \tag{5}$$

where

$$\Sigma_{\mathbf{D}} = \Sigma_{\eta} + \begin{pmatrix} \Sigma_{\epsilon} + \Sigma_{\delta} & 0 \\ 0 & 0 \end{pmatrix},$$

in which Σ_{ϵ} is the $N_y \times N_y$ observation covariance matrix, Σ_{η} is obtained for each pair of $(r+1)N_y$ simulation inputs through equation (3) corresponding to $\mathbf{D}$ and Σ_{δ} is an $N_y \times N_y$ matrix obtained for each pair of N_y input through the instances of equation (4) that correspond to the coefficients c^{F}. Let $\pi(\boldsymbol{\theta})$ be the joint prior distribution for the (unknown) calibration vector $\boldsymbol{\theta}$. The resulting posterior density has the form

$$\pi(\boldsymbol{\theta}, \lambda_{\eta}, \boldsymbol{\rho}_{\eta}, \lambda_{\delta} | \mathbf{D}) \propto L(\mathbf{D} | \boldsymbol{\theta}, \lambda_{\eta}, \boldsymbol{\rho}_{\eta}, \lambda_{\delta}, \Sigma_{\epsilon}) \, \pi(\boldsymbol{\theta}) \, \pi(\lambda_{\eta}) \, \pi(\boldsymbol{\rho}_{\eta}) \, \pi(\lambda_{\delta}), \tag{6}$$

which can be explored via a Markov chain Monte Carlo (MCMC) technique, for which we employ a Metropolis–Hastings algorithm. The calibrated vector is then denoted by $\boldsymbol{\theta}^* = \arg\max_{\theta} \pi(\boldsymbol{\theta}, \lambda_{\eta}, \boldsymbol{\rho}_{\eta}, \lambda_{\delta} | \mathbf{D})$. We implement a Metropolis–Hastings algorithm to produce the realization of the posterior. Metropolis updates are used for the correlation and the calibration parameters with a uniform proposal distribution centred at the current value of the parameter. The precision parameters are sampled by using Hastings updates with a uniform proposal distribution centred at the current value of the parameter (Higdon *et al.*, 2008). This eventually yields draws from the posterior distribution by repeatedly accepting and rejecting a choice of move in the parameter space.

The specification of the covariance structures for the truncated basis representation is a mathematical challenge: finding explicit expressions for the covariance is difficult (Jun and Stein, 2008). There is an alternative way to model complex spatial covariance structures efficiently with the added bonus of a suitable depiction of the non-stationarity structure into our calibration algorithm: the SPDE approach. We introduce it in the next section.

2.2. Spatial modelling through the stochastic partial differential equation approach

Traditional models in spatial statistics build an approximation of the entire underlying random field. They are usually specified through the covariance function of the latent field. To assess uncertainties in the spatial interpolation over the whole spatial domain, we cannot build models only for the discretely located observations or model outputs; we need to build an approximation of the entire underlying stochastic process defined on the spatial field. We consider statistical models for which the unknown functions are assumed to be realizations of a Gaussian random spatial process. The conventional fitting approach spatially interpolates values as linear combinations of the original observed locations, and this constitutes the spatial kriging predictor.

Because of the fixed underlying covariance structure, this approach requires more sophisticated treatments to take into consideration non-stationarity (Stein, 2005; Jun and Stein, 2008; Yue and Speckman, 2010; Kleiber and Nychka, 2012; Gramacy and Apley, 2015). A different computational approach was introduced by Lindgren *et al.* (2011), in which random fields were expressed as a weak solution to an SPDE, with explicit links between the parameters of the SPDE model and the Matérn covariance function. In this section we review some of the main concepts in spatial modelling through the SPDE approach.

It may seem contradictory to make use of the SPDE approach since it seemingly only captures local structures, and climate model outputs display smooth variations. However, the SPDE approach, especially the non-stationarity version, can translate these smooth variations of the model outputs (and of observations) into a statistical description of the variations across space that efficiently characterizes the spatial behaviour (through the scale and precision parameters). Spatially distributed observations will still display more erratic behaviour than model outputs,

but the SPDE approach will allow the calibration to be steered only by the parameters that are associated with the smoothest components.

2.2.1. *Matérn covariance and the link to stochastic partial differential equations*

The Matérn function is a flexible covariance structure and is widely used in spatial statistics (Stein, 2005; Jun and Stein, 2007, 2008; Gneiting *et al.*, 2010; Genton and Kleiber, 2015). The choice of covariance is not that important indeed for calibration parameters (Kennedy and O'Hagan, 2001) but, for the outputs (across location inputs), the choice of covariance is essential, as we show. The shape parameter $\nu > 0$, the scale parameter $\kappa > 0$ and the marginal precision $\tau^2 > 0$ parameterize it:

$$\operatorname{cov}(\mathbf{h}) = \frac{2^{1-\nu}}{(4\pi)^{d/2}\,\Gamma(\nu + d/2)\kappa^{2\nu}\tau^2}\kappa\|\mathbf{h}\|^{\nu} K_{\nu}\,\kappa\|\mathbf{h}\|, \qquad \mathbf{h} \in \mathbb{R}^d,$$

where $\mathbf{h}$ denotes the difference between any two locations s and s', $\mathbf{h} = s - s'$ and K_ν is the modified Bessel function of the second kind of order ν.

We denote by $Y(\mathbf{s})$ the observations (or the spatially distributed model outputs) for a latent spatial field $X(\mathbf{s})$, with a Matérn covariance structure. We assume zero mean Gaussian noise $\mathcal{W}(\mathbf{s})$, with a constant variance σ_s^2: $Y(\mathbf{s}) = X(\mathbf{s}) + \mathcal{W}(\mathbf{s})$. Thus, according to Whittle (1963), the latent field $X(\mathbf{s})$ is the solution of a stationary SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2}\tau X(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \tag{7}$$

where Δ is the Laplace operator. We explain in the next subsection how the analysis of this SPDE can be carried out by the finite element method. The regularity (or smoothness) parameter ν essentially determines the order of differentiability of the fields. The link between the Matérn field and the SPDE is given by $\alpha = \nu + d/2$, which makes explicit the relationship between dimension and regularity for fixed α. On more general manifolds than $\mathbb{R}^d$, such as the sphere (Chang, Guillas and Fioletov, 2015), the direct Matérn representation is not easy to implement (for example, Matérn covariance with great circle distance is valid only at $\nu \in (0, 0.5]$ (Gneiting, 2013)), but the SPDE formulation provides a natural generalization, and the ν-parameter will keep its meaning as the quantitative measure of regularity. Instead of defining Matérn fields by the covariance function, Lindgren *et al.* (2011) used the solution of the SPDE as a definition, and it is much easier and flexible to do so. This definition also facilitates non-stationary extensions by allowing the SPDE parameters κ and τ in equation (7) to vary with space; hence they are denoted $\kappa(\cdot)$ and $\tau(\cdot)$ respectively.

2.2.2. *Stochastic partial differential equation model construction*

We estimate the SPDE parameters and supply uncertainty information about the spatial fields by using the *integrated nested Laplacian approximation* (INLA) framework, which is available as an R package (`http://www.r-inla.org/`) (Lindgren and Rue, 2015; Rue *et al.*, 2017). The models that are implemented in the INLA–SPDE framework are built on a basis representation (triangulation over the spatial domain): $X(\mathbf{s}) = \Sigma_{i=1}^{M} \varphi_i(\mathbf{s}) w_i$, where $\{w_i\}$ are the stochastic weights chosen so that the distribution of the functions $X(\mathbf{s})$ approximates the distribution of solutions to the SPDE on the space, and $\varphi_i(\mathbf{s})$ are a piecewise linear basis with compact support (i.e. finite elements) to obtain a Markov structure, and to preserve it when conditioning on local observed locations. The Markov property yields a sparse precision matrix, so efficient numerical algorithms can be employed for large spatial data. The projection of the SPDE onto the basis

representation is chosen by a finite element method. The finite element method represents a general class of techniques for the approximate solution to partial differential equations. The piecewise linear basis functions defined by a triangulation of the spatial domain enable us to evaluate the precision matrix of the latent field explicitly. As a result, $X(\mathbf{s})$ follows a normal distribution with mean 0, and the precision matrix can be explicitly expressed as a combination of the piecewise linear basis functions weighted by κ and τ (which means that κ and τ have a joint influence on the marginal variances of the latent field). Then $X(\mathbf{s})$ can be generated continuously as approximate solutions to the SPDE.

For the WACCM output domain, the triangulation is simply built on regularly gridded cells. Note that the triangulation can be made adaptive to the irregularly distributed spatial data (Cameletti *et al.*, 2013). The default value in the INLA algorithm is $\alpha = 2$, but $0 \leqslant \alpha < 2$ are also available, though yet to be completely tested (Lindgren and Rue, 2015). So, with a two-dimensional manifold (e.g. $\mathbb{R}^2$ and $\mathbb{S}^2$), the smoothness parameter ν must be fixed at 1 because of the relationship $\alpha = \nu + d/2$. The strength of this SPDE technique enables us to quantify the level of non-stationarity by employing spatial basis representations for both κ and τ (i.e. these quantities are constants in a stationary field). With a focus on the calibration, let $\kappa^{\mathrm{M}}(\mathbf{s}, \boldsymbol{\theta})$ and $\tau^{\mathrm{M}}(\mathbf{s}, \boldsymbol{\theta})$ be the scale and precision parameters in an SPDE model used to approximate the model outputs. To obtain basic identifiability, $\kappa^{\mathrm{M}}(\mathbf{s}, \boldsymbol{\theta})$ and $\tau^{\mathrm{M}}(\mathbf{s}, \boldsymbol{\theta})$ are taken to be positive, and their logarithm can be decomposed as

$$\log\{\kappa^{\mathrm{M}}(\mathbf{s}, \boldsymbol{\theta}_j)\} = \sum_{z=1}^{N_\kappa} \kappa_z^{\mathrm{M}}(\boldsymbol{\theta}_j)\psi_z(\mathbf{s}),$$

$$\log\{\tau^{\mathrm{M}}(\mathbf{s}, \boldsymbol{\theta}_j)\} = \sum_{z=1}^{N_\tau} \tau_z^{\mathrm{M}}(\boldsymbol{\theta}_j)\psi_z(\mathbf{s}), \qquad j = 1, \ldots, r.$$

Each basis function is evaluated at output cells and observed locations. The coefficients $\{\kappa_z^{\mathrm{M}}\}$ and $\{\tau_z^{\mathrm{M}}\}$ represent local variances and correlation ranges (Bolin and Lindgren, 2011; Lindgren *et al.*, 2011; Fuglstad *et al.*, 2015). For simplicity, we call these coefficients 'SPDE parameters' in the calibration. In the next section we introduce how to incorporate the SPDE parameters in calibration to enhance the prediction accuracy.

2.3. Combining stochastic partial differential equation modelling and calibration

A reduced rank approach was often used to ease the computational issue in large spatial data sets (Banerjee *et al.*, 2008; Cressie and Johannesson, 2008; Furrer and Sain, 2009; Katzfuss and Cressie, 2011). To reduce and summarize a spatial field properly, both global and local scale dependences need to be well captured and represented. To do so, a two-steps approximation was developed by combining the reduced rank representation and sparse matrix techniques, to account for global and local structures respectively (Stein, 2007; Sang and Huang, 2012). We follow the same idea of using a reduced rank representation to capture global scale variability (described in Section 2.1), whereas, instead of tapering the covariance matrix into a sparse matrix, we use the INLA–SPDE technique to represent small-scale variability. In this section we describe the details of our extension by including the SPDE-defined scale and precision parameters in the Bayesian calibration.

As $\{\kappa_z^{\mathrm{M}}(\boldsymbol{\theta})\}$ and $\{\tau_z^{\mathrm{M}}(\boldsymbol{\theta})\}$ can quantify the non-stationarity and derivative information in the spatial process, we now include these two types of coefficient in our technique (combined with $\{c_z^{\mathrm{M}}(\boldsymbol{\theta})\}$ in the previous section, and vectorized all coefficients as a scalar). Then our approach represents the observations and model input–output relationship as

$$y^{\mathrm{F}}(s_1),\ldots,y^{\mathrm{F}}(s_n) \overset{\text{transform}}{\longrightarrow} c_1^{\mathrm{F}},\ldots,c_{N_y}^{\mathrm{F}},\kappa_1^{\mathrm{F}},\ldots,\kappa_{N_\kappa}^{\mathrm{F}},\tau_1^{\mathrm{F}},\ldots,\tau_{N_\tau}^{\mathrm{F}},$$

$$\eta(s_1,\theta_1),\ldots,\eta(s_n,\theta_1) \overset{\text{transform}}{\longrightarrow} c_1^{\mathrm{M}}(\theta_1),\ldots,c_{N_y}^{\mathrm{M}}(\theta_1),\kappa_1^{\mathrm{M}}(\theta_1),\ldots,\kappa_{N_\kappa}^{\mathrm{M}}(\theta_1),\tau_1^{\mathrm{M}}(\theta_1),\ldots,\tau_{N_\tau}^{\mathrm{M}}(\theta_1),$$

$$\vdots$$

$$\eta(s_1,\theta_r),\ldots,\eta(s_n,\theta_r) \overset{\text{transform}}{\longrightarrow} c_1^{\mathrm{M}}(\theta_r),\ldots,c_{N_y}^{\mathrm{M}}(\theta_r),\kappa_1^{\mathrm{M}}(\theta_r),\ldots,\kappa_{N_\kappa}^{\mathrm{M}}(\theta_r),\tau_1^{\mathrm{M}}(\theta_r),\ldots,\tau_{N_\tau}^{\mathrm{M}}(\theta_r)$$

where $N_y+N_\kappa+N_\tau \ll n$. The aim is to combine the SPDE parameters as non-stationary information for the implementation of the calibration algorithm, and to model all coefficients jointly with the GP assumption. We also assume that the three types of coefficient are independent. To describe the formulation of the design matrix, let $\{z_1,z_2,z_3|z_1=1,\ldots,N_y;z_2=1,\ldots,N_\kappa;z_3=1,\ldots,N_\tau\}$ be the indices that are used to represent each triplet of coefficients. The calibration formulation is hence

$$\begin{pmatrix} c_{z_1}^{\mathrm{F}} \\ \kappa_{z_2}^{\mathrm{F}} \\ \tau_{z_3}^{\mathrm{F}} \end{pmatrix} = \begin{pmatrix} c_{z_1}^{\mathrm{M}}(\boldsymbol{\theta}) \\ \kappa_{z_2}^{\mathrm{M}}(\boldsymbol{\theta}) \\ \tau_{z_3}^{\mathrm{M}}(\boldsymbol{\theta}) \end{pmatrix} + \begin{pmatrix} c_{z_1}^{\delta} \\ \kappa_{z_2}^{\delta} \\ \tau_{z_3}^{\delta} \end{pmatrix} + \begin{pmatrix} c_{z_1}^{\epsilon} \\ \kappa_{z_2}^{\epsilon} \\ \tau_{z_3}^{\epsilon} \end{pmatrix}.$$

Thus there are $(N_y+N_\kappa+N_\tau)$-blocks of coefficients corresponding to each combination of $\boldsymbol{\theta}_j$, $j=1\ldots,r$, in the covariance matrix. The GP assumption is imposed on each coefficient $(c_{z_1,j}^{\mathrm{M}},\kappa_{z_2,j}^{\mathrm{M}},\tau_{z_3,j}^{\mathrm{M}})^{\mathrm{T}}$ with mean 0 and covariance function

$$\operatorname{cov}\{(c_{z_1}^{\mathrm{M}}(\boldsymbol{\theta}),\kappa_{z_2}^{\mathrm{M}}(\boldsymbol{\theta}),\tau_{z_3}^{\mathrm{M}}(\boldsymbol{\theta}))^{\mathrm{T}},(c_{z_1'}^{\mathrm{M}}(\boldsymbol{\theta}'),\kappa_{z_2'}^{\mathrm{M}}(\boldsymbol{\theta}'),\tau_{z_3'}^{\mathrm{M}}(\boldsymbol{\theta}'))^{\mathrm{T}}\} = \frac{1}{\lambda_\eta}\prod_{i=1}^{3} I_{z_iz_i'}\prod_{k=1}^{q}\rho_{\eta k}^{4(\theta_k-\theta_k')^2},$$

where $I_{z_iz_i'}=1$ if $z_i=z_i'$ and $I_{z_iz_i'}=0$ otherwise. In other words, these three types of coefficient $\{c_{z_1}^{\mathrm{M}},\kappa_{z_2}^{\mathrm{M}},\tau_{z_3}^{\mathrm{M}}\}$ have a joint multivariate normal prior distribution with mean 0, and a covariance structure forming a block diagonal matrix:

$$\begin{pmatrix} c_{z_1}^{\mathrm{M}} \\ \kappa_{z_2}^{\mathrm{M}} \\ \tau_{z_3}^{\mathrm{M}} \end{pmatrix} \sim N\left[\mathbf{0},\begin{pmatrix} \operatorname{cov}\{c_{z_1}^{\mathrm{M}}(\boldsymbol{\theta}),c_{z_1'}^{\mathrm{M}}(\boldsymbol{\theta}')\} & 0 & 0 \\ 0 & \operatorname{cov}\{\kappa_{z_2}^{\mathrm{M}}(\boldsymbol{\theta}),\kappa_{z_2'}^{\mathrm{M}}(\boldsymbol{\theta}')\} & 0 \\ 0 & 0 & \operatorname{cov}\{\tau_{z_3}^{\mathrm{M}}(\boldsymbol{\theta}),\tau_{z_3'}^{\mathrm{M}}(\boldsymbol{\theta}')\} \end{pmatrix}\right].$$

The elements in each block are also block diagonal matrices. The model discrepancy term in the functional space follows a GP assumption defined in equation (4). All the prior assumptions that were discussed in the previous section remain unchanged. Thus the sampling likelihood (5) and the posterior distribution (6) still hold in this case. Overall, we decompose the model outputs into a basis via the coefficients $\{c^{\mathrm{M}}\}$, and estimate the SPDE parameters $\{\kappa^{\mathrm{M}},\tau^{\mathrm{M}}\}$ in the latent field through a regression onto these basis functions. We are essentially fitting a GP model with $\{c^{\mathrm{M}}\}$ for the regression mean structure and $\{\kappa^{\mathrm{M}},\tau^{\mathrm{M}}\}$ for the parameters of a Matérn covariance function.

2.4. Guidance for the number of basis functions

In real applications, we often do not know whether the calibrated values work until actually performing a validation. It can be computationally challenging to find the optimized orders for the combination of N_y, N_κ and N_τ. Similarly to most truncated basis representations, we choose the number of basis functions *post hoc*. We provide the following model selection guidelines.

(a) The basis representation for the mean structure of model outputs plays a dominant role in the algorithm. Typically we cannot expect to calibrate a global process only through a local structure. Therefore N_y usually needs to be greater than $N_\kappa + N_\tau$.
(b) Calibration with only one of the coefficients κ or τ cannot improve the analysis. The reason is the fact that κ and τ represent a spatial process jointly being tacitly assumed. Recall that the Matérn function is controlled by the smoothness parameter ν, the scale parameter κ and the precision parameter τ. The parameter ν is fixed by $\alpha = \nu + d/2$ in connection with the SPDE; thus the approximated spatial process depends on κ and τ jointly. Both κ and τ need to be included to reflect the full variation in the spatial field.

In this paper we use spherical harmonics (SHs) as our primary investigation. The SHs represent the wave features at different scales on the sphere (Bolin and Lindgren, 2011; Jun and Stein, 2008). For calibration, it seems unnecessary in general to approximate the spatial processes with very high order expansions of SHs to fit each run of model output best. The main requirement is to extract sufficient and meaningful information about the calibration parameters from the variations in the SH coefficients that could be attributed to variations in the inputs. To ensure that this requirement is met, a simple validation is to increase the basis number and to recalibrate the model. In case the results have no statistically significant effects, then the number is sufficiently large. Muir and Tkalčić (2015) utilized the corrected Akaike information criterion AIC to choose an optimal maximum order of expansion for irregular data on the sphere in a hierarchical Bayesian setting. The results show that the third–fifth orders of expansion in SHs are generally a turning point from fast to slow reduction in AIC in terms of balancing explanatory power with simplicity (although not the smallest AIC). In all these approaches, the choice of the number of basis vectors is currently *post hoc*. We reckon that the third or fourth order of SH transform for capturing large-scale variability, along with a lower order of SPDE non-stationary information to account for local structure, as a good start in practical application.

3. Simulation study: non-stationary field

To illustrate the methodology, this synthetic example simulates a non-stationary field on the sphere, with an anisotropic property (the spatial correlation depends on latitude), to demonstrate how including the parameters in the SPDE can enhance the GP calibration in such situations. We illustrate how the parameters in an SPDE technique can be incorporated in our calibration algorithm to model non-stationarity over a spherical domain. With $n = 10 \times 10$ regularly spaced locations in latitude L and longitude l, and $r = 50$ computer runs according to a maximin Latin hypercube design for the calibration inputs, the function with three calibration parameters ($q = 3$) is set to

$$f(\tilde{\mathbf{s}}, \boldsymbol{\theta}) = (0.5 s_1^2 + \theta_1 s_2 s_3) \begin{cases} \theta_2 s_2 & \text{if } L > \pi/2, \\ \theta_3 \exp(-s_3 - s_1) & \text{if } L \leqslant \pi/2, \end{cases} \quad (\theta_1, \theta_2, \theta_3) \in [0,1]^3, \qquad (8)$$

where the true values for $(\theta_1, \theta_2, \theta_3)$ are set to (0.5, 0.2, 0.8) and $(s_1, s_2, s_3) = (\cos(l)\sin(L), \sin(l)\sin(L), \cos(L))$ are spherical co-ordinates. We create a non-stationary spatial field by introducing different structures in the northern and southern hemispheres, where θ_1 is a global calibration parameter, and (θ_2, θ_3) are local variates. In this example the local structures are designed to be larger than the global structure: $\exp(-s_3 - s_1)$ has stronger variation than s_2, and both of them have a larger variation than $s_2 s_3$ (see the magnitude of variation in each component from Fig. 1).

First, we perform the spherical harmonics transform (SHT) onto observations y^{F} and each computer run η_j, $j = 1, \ldots, 50$, and then carry out the calibration on the coefficients. In total, we

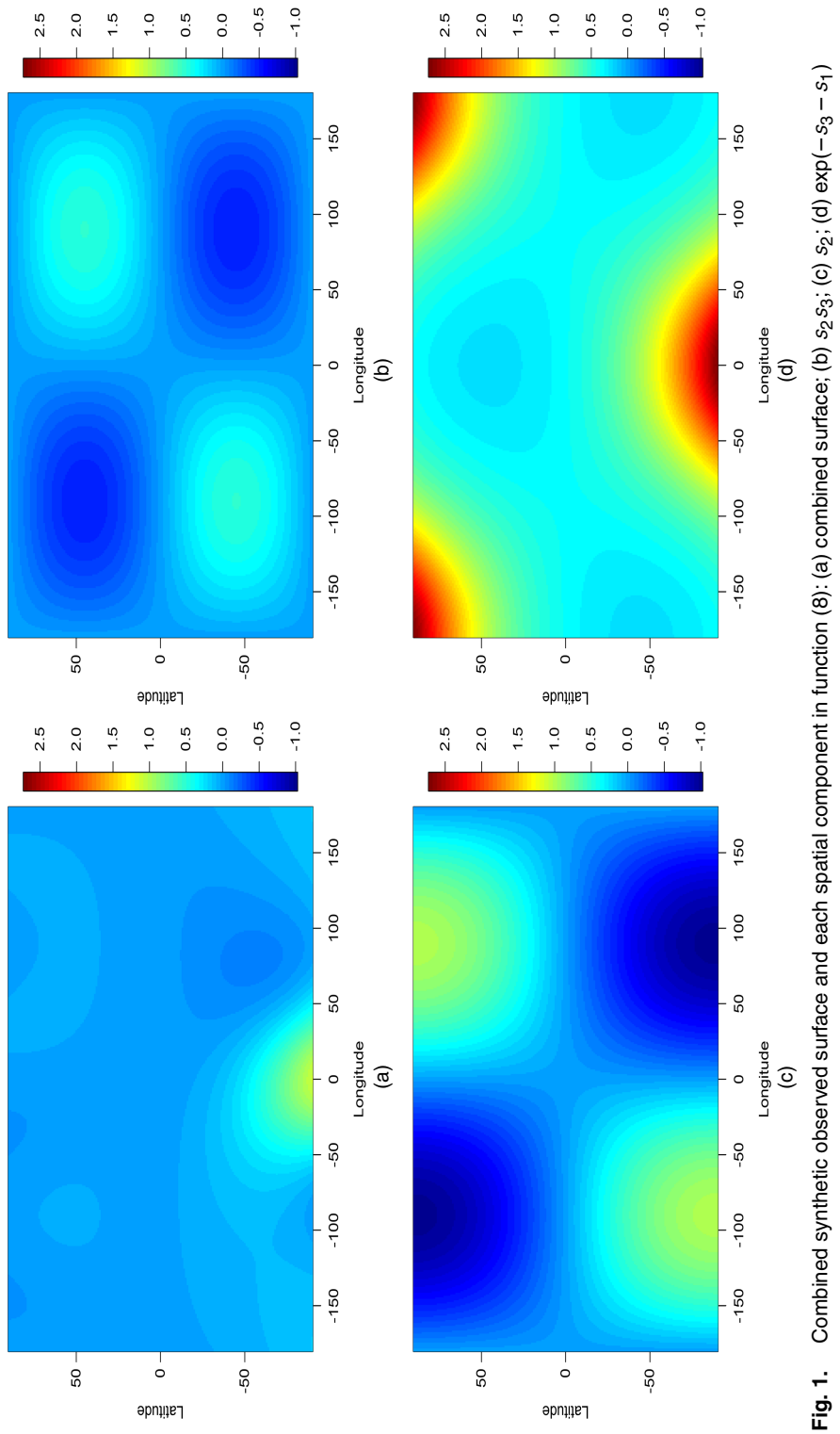

Fig. 1. Combined synthetic observed surface and each spatial component in function (8): (a) combined surface; (b) $s_2 s_3$; (c) s_2; (d) $\exp(-s_3 - s_1)$

estimate 13 models with different numbers of expansion order. The results of using the fourth–seventh orders of the SHT are shown in the first part of Table 1 (strategies A–D). We can see that the global calibration parameter θ_1 is estimated well. However, even though the convergence of an MCMC chain can be established for θ_2 and θ_3, the posterior means are underestimated. According to the root-mean-square error (RMSE) between assumed and predicted observations, which can be written as

$$\mathrm{RMSE} = \sqrt{\left[\frac{\sum_{i=1}^{n}\{f(s_i, \boldsymbol{\theta}^*) - f(s_i, \boldsymbol{\theta}^{\mathrm{post}})\}^2}{n}\right]},$$

an increase in the expansion order cannot improve the results. This underestimation can be viewed as a deficiency to capture local variations through a global mean structure: the variation that is created by these two parameters will be obscured and distorted by the variation from θ_1.

To understand the role of the SPDE parameters in the calibration, we then perform a calibration using only the coefficients $\{\kappa^{\mathrm{M}}, \tau^{\mathrm{M}}\}$. Under the same priors and algorithm, the posterior mean and SD of the first three orders of the expansion for κ^{M} and τ^{M} are shown in the second part of Table 1 (strategies E–G). Even though the calibration does not fully succeed (and should not without matching original outputs to observations but only SPDE information), the result in the second-order expansion for κ^{M} and τ^{M} seems informative as the posterior modes are close to the true values. The first two orders of the expansion surface for κ^{F} and τ^{F} for the observations are shown in Fig. 2. It is difficult to interpret the features of $\kappa(s)$ and $\tau(s)$ directly. However, from Figs 2(c) and 2(d) we can see that a strong north-east–south-west flow in $\kappa^{\mathrm{F}}(s)$ matches the pattern in Figs 1(b) and 1(c), and a high anticorrelation between $\tau^{\mathrm{F}}(s)$ (inverse precision) and the y^{F}-surface.

For the next step, we infer $\{c^{\mathrm{M}}, \kappa^{\mathrm{M}}, \tau^{\mathrm{M}}\}$ jointly with the GP model. We combined the coefficients in strategies A–C (coefficients for the mean structure) and strategies E and F (coefficients for the SPDE parameters). The results are presented in the third part of Table 1. We can see that, with the SPDE information included, we achieve an improvement in the

Table 1. Posterior mean and SD for $(\theta_1, \theta_2, \theta_3)$ in function (8), RMSE and number of coefficients (right-hand column) under different orders of SHT for $\{\eta, \kappa, \tau\}$ per model run†

Strategy	η	κ	τ	θ_1 *(=0.5)*	θ_2 *(=0.2)*	θ_3 *(=0.8)*	*RMSE*	$N_y + N_\kappa + N_\tau$
A	4	—	—	0.505 (0.050)	0.188 (0.048)	0.762 (0.038)	92	15
B	5	—	—	0.498 (0.053)	0.179 (0.062)	0.746 (0.050)	132	21
C	6	—	—	0.477 (0.062)	0.166 (0.079)	0.705 (0.069)	237	28
D	7	—	—	0.488 (0.112)	0.198 (0.127)	0.695 (0.119)	257	36
E	—	1	1	0.579 (0.158)	0.148 (0.068)	0.620 (0.200)	431	6
F	—	2	2	0.560 (0.097)	0.189 (0.078)	0.740 (0.089)	145	12
G	—	3	3	0.785 (0.078)	0.442 (0.155)	0.858 (0.054)	433	20
H	4	1	1	0.452 (0.097)	0.071 (0.049)	0.495 (0.037)	755	21
I	5	1	1	0.495 (0.044)	0.133 (0.049)	0.498 (0.032)	737	27
J	6	1	1	0.356 (0.050)	0.135 (0.052)	0.686 (0.119)	322	34
K	4	2	2	0.553 (0.068)	0.225 (0.108)	0.771 (0.109)	80	27
L	5	2	2	0.529 (0.068)	0.179 (0.107)	0.794 (0.098)	28	33
M	6	2	2	0.537 (0.066)	0.171 (0.110)	0.789 (0.083)	39	40

†RMSE was multiplied by 10^3 to illustrate the magnitude.

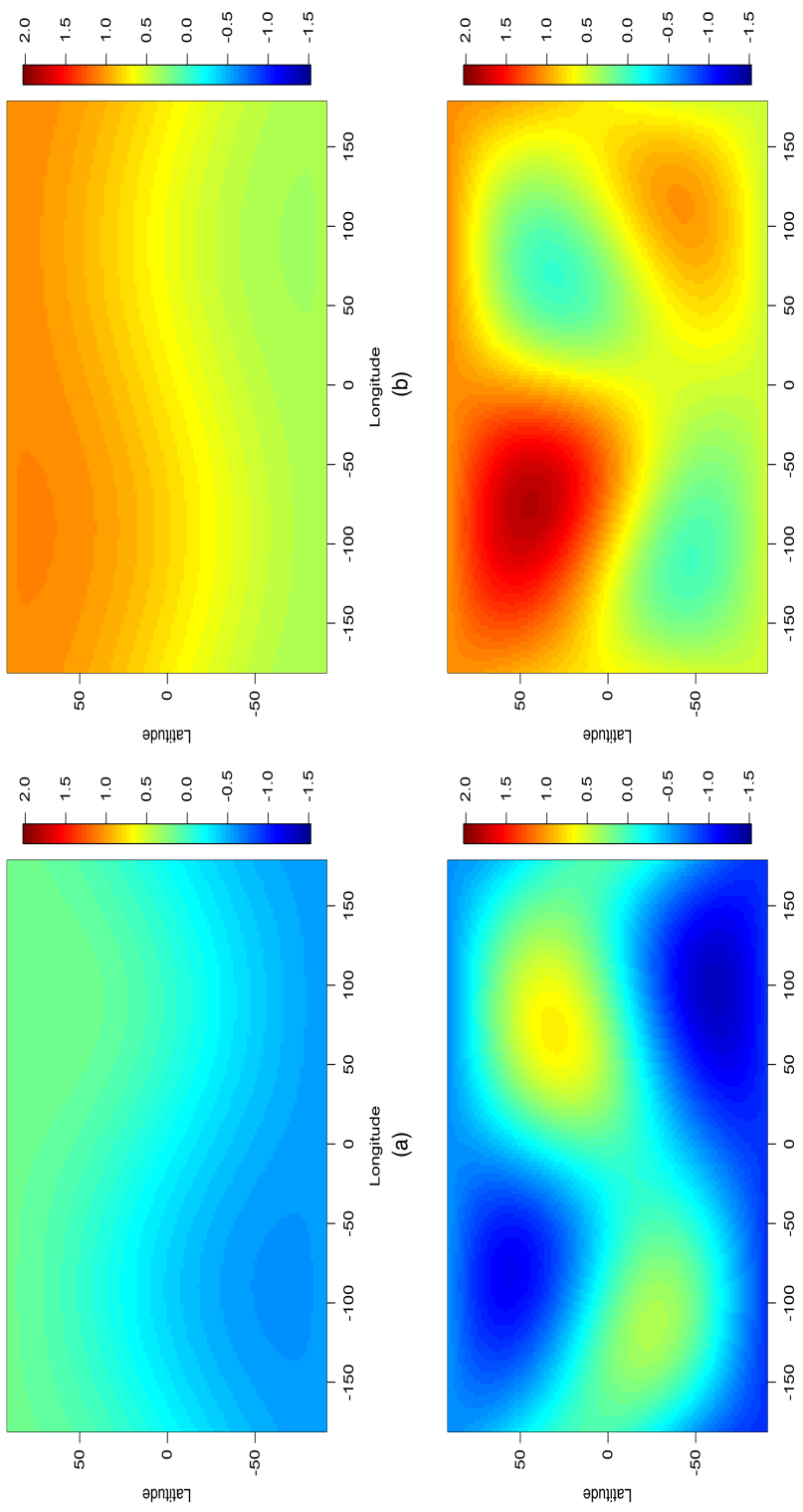

Fig. 2. First- and second-order expansion surfaces for $_{\kappa}^{F}$ and $_{\tau}^{F}$ for observations y^{F}: (a) first-order $_{\kappa}^{F}$; (b) first-order $_{\tau}^{F}$; (c) second-order $_{\kappa}^{F}$; (d) second-order $_{\tau}^{F}$

calibration by combining local structure with a global process. For example, strategies C and K have a similar number of coefficients, but the combination with the SPDE increases the estimated accuracy in θ_2 and θ_3. The similar case of strategies D and L also supports the use of SPDE information. Strategy L uses a smaller number of coefficients, while achieving an improvement in terms of increased accuracy in θ_3 and reduced RMSE. Nevertheless, only in the case of the second-order expansion do the SPDE parameters help; the first-order expansion cannot achieve a good result in this example. This demonstrates that the non-stationarity is rather complex. From these findings we thus acknowledge that the SPDE technique enables us to identify the local feature from the global spatial process in the calibration. Therefore we highlight that, when we cannot make an improvement in the accuracy of estimation by increasing the basis number into the mean structure, the SPDE technique can serve as a valuable alternative.

We provide more illustrations of the flexibility of our approach under various situations:

(a) calibration with irregularly spaced outputs over the plane by using a *B*-spline basis;
(b) investigation of the connection between the accuracy of calibration and the number of computer runs r, and between the accuracy of calibration and the orders and modes of SHs;
(c) comparison of our approach and the empirical orthogonal functions approach and original Kennedy and O'Hagan (2001) framework, in the on-line supplemental material for the interested reader.

4. Application to the 'Whole atmosphere community climate model' experiments

A series of WACCM runs with the component set prescribed sea ice, data ocean and specified chemistry, with horizontal resolution $1.9^\circ \times 2.5^\circ$ and 66 vertical levels were simulated from January 1st, 2000. The GW parameterizations in the WACCM depend on four inputs.

(a) cbias ($\theta_1 \in [-5, 5]$): anisotropy of the source spectrum, e.g. -5 m s^{-1}; the spectrum has a stronger westward component, with the centre of the spectrum at 5 m s^{-1} westward. Note that the default simulation in the WACCM is isotropic (i.e. cbias $= 0$). An anisotropic GW source has been long reckoned to have potential to improve the middle atmosphere circulation compared with an isotropic source (Medvedev *et al.*, 1998; Hamilton, 2013; Chunchuzov *et al.*, 2015).
(b) effgw ($\theta_2 \in [0.05, 0.3]$), the efficiency factor, measures the GW intermittency.
(c) flatgw ($\theta_3 \in [1, 3]$) controls the momentum flux of the parameterized waves at the launch levels.
(d) launlvl ($\theta_4 \in [50, 700]$) are the launch levels of the waves.

The values of GW inputs $\boldsymbol{\theta}$ are generated by a maximin Latin hypercube design (but scaled to be $[0, 1]^4$). We simulated $r = 100$ runs for 2 months. The first month was discarded as a spin-up period (Eyring *et al.*, 2016). Each output was computed over 96 latitudes and 144 longitudes, so the total output size is $n \times r = 96 \times 144 \times 100 = 1382400$. We perform the calibration for the WACCM, either against synthetic (but with added non-stationary observation errors) or real observations, to validate our approach fully.

4.1. Calibration against synthetic observations

4.1.1. Model set-up

To illustrate our methodology, we compare the zonal wind simulations $\eta(\mathbf{s}_i, \boldsymbol{\theta}_j)$, where $\mathbf{s}_i$, $i = 1, \ldots, 96 \times 144$, are the latitude and longitude on the spherical domain, and $j = 1, \ldots, 100$ is the

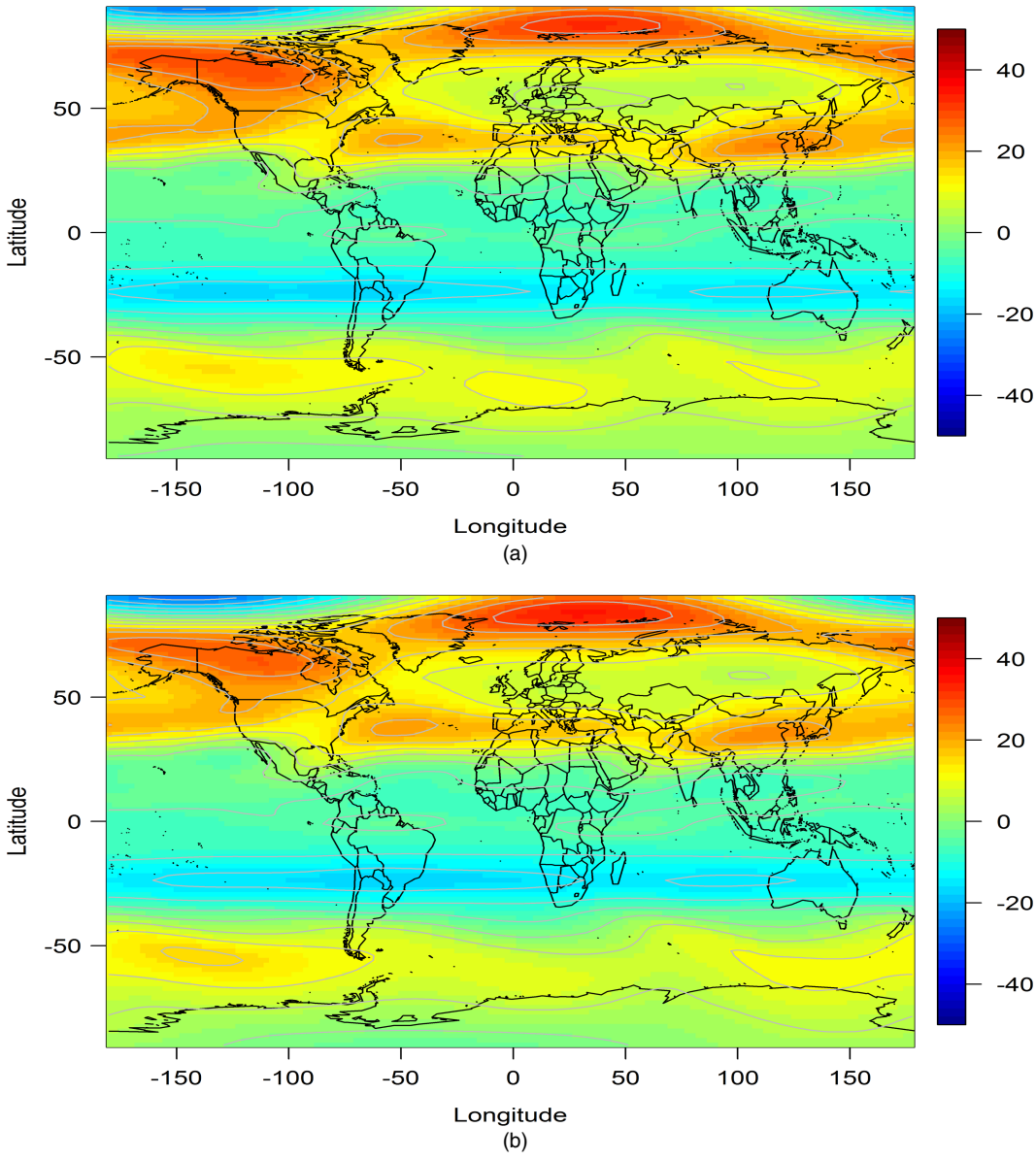

Fig. 3. (a) Zonal wind standard output η^* and (b) assumed observed surface with noise y^{F} and discrepancy added to the zonal wind standard output (30 mb, February 2000)

index of the runs, with the WACCM's standard outputs (i.e. default simulation), instead of actual observations. Therefore we know the true GW parameter values and can validate our method. Let $\eta^*(\mathbf{s}_i)$ be the zonal wind surface from WACCM standard output. To account for possible observation error and a lack of physics in the model (discrepancy), and thus to evaluate the robustness of our method, we add smooth noise to $\eta^*(\mathbf{s}_i)$ by assuming that the observations are given by

$$y^{\mathrm{F}}(\tilde{\mathbf{s}}_i) = \eta^*(\tilde{\mathbf{s}}_i) + \frac{\sigma_{\eta^*}}{5} s_1 + \frac{1}{2} s_2 s_3,$$

where $\tilde{\mathbf{s}}_i = (s_1, s_2, s_3)$ are the spherical co-ordinates, and $\sigma_{\eta^*} = 11.14$ is the SD of η^*. Figs 3(a) and 3(b) show the zonal wind surfaces from standard outputs and synthetic observations at 30 mb, February 2000.

As for the computational issue, in practice it is difficult to deal with a size of model output beyond moderately large (say of the order of 2000 responses). Here we have $r = 100$ computer runs; therefore we seek to decompose each model output with about 20 coefficients. We represent observations and model discrepancies by using third- and fourth-order SHTs for model outputs and observations respectively. This allows enough flexibility. We report the two strategies A with or B without including first-order SPDE non-stationary information. We also report two other strategies that use five (strategy C) or 10 (strategy D) PCs (with 95.8% and 97.9% respectively of the variation explained) to decompose the model outputs and observations (see the algorithm in the on-line supplemental material).

4.1.2. Prediction accuracy

The posterior modes of each strategy are shown in Table 2. Both strategies A and B calibrate θ_2 well and slightly overestimate θ_4. The inclusion of SPDE parameters in strategies A *versus* B not only increases the accuracy of the posterior mode for θ_1 but also estimates very closely θ_3, which is a difficult task as the true value lies on the lower bound. The quantification of the anisotropic velocity in a large spatial process is a difficult problem (Large *et al.*, 2001; Lauritzen *et al.*, 2015). The improvement in accuracy of the estimation of θ_1 confirms the value of using the SPDE technique in the calibration since the non-stationarity allows the amount of flexibility that is required to identify more clearly the value of θ_1.

Unfortunately, the MCMC runs do not converge for strategies C and D; hence their posterior modes are uninterpretable (but we report them nevertheless). This result can be expected. Our GW parameterization aims to reduce zonal wind bias at the tropics associated with the QBO; however, the principal mode of variability in our model outputs occurs across the northern hemisphere (in East Asia to be precise), where the influence of the GW is indirect (see the next section for further discussion in comparison with real observations). The PC decomposition will focus on the variability in the northern hemisphere compared with the tropics. Recent studies suggest that a PC-based approach tends to cause a 'terminal case analysis' in climate modelling (Salter *et al.*, 2018), which means that there is no set of parameters that can allow the model to mimic reality. For this reason the PC-based approach is not appropriate for our calibration setting.

Fig. 4 shows a concrete example of such lack of convergence in a synthetic example (see the on-line supplementary material). This is a comparison of the MCMC sample paths of the

Table 2. Posterior mode of GW parameters on the rescaled [0, 1] range†

Strategy	*cbias* ($\theta_1^* = 0.5$)	*effgw* ($\theta_2^* = 0.56$)	*flatgw* ($\theta_3^* = 0$)	*launlvl* ($\theta_4^* = 0.2308$)
A (SH—non-stationary SPDE)	**0.435**	**0.547**	**0.060**	**0.276**
B (SH—stationary SPDE)	0.361	**0.561**	0.281	**0.282**
C (5 PCs)	*0.538*	*0.396*	*0.741*	*0.523*
D (10 PCs)	*0.639*	*0.082*	*0.466*	*0.908*

†The MCMC algorithm did not converge in cases C and D, so these estimates (in italics) are unreliable. Valid calibrations are highlighted in bold.

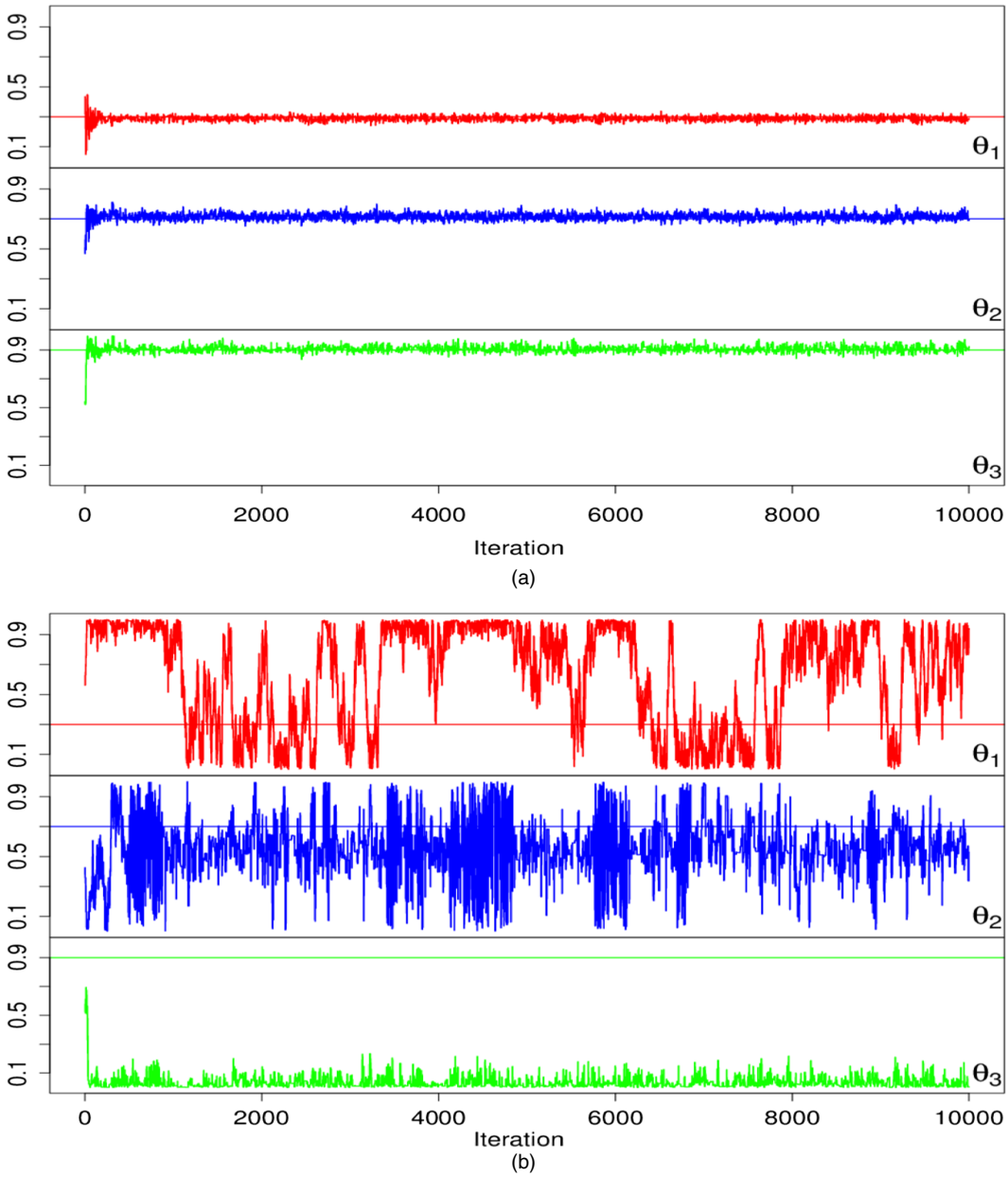

Fig. 4. MCMC paths of θ over the design space $[0,1]^3$ for our approach and PC-based algorithms (a synthetic case is given in the on-line supplementary material) (———, true values): (a) second SHT (nine coefficients); (b) nine empirical orthogonal functions representation

calibration parameters by second-order SHs (nine coefficients) and nine-PCs representation. We can see that, for all calibration parameters in the SHs approach, convergence occurred after roughly 500 iterations, whereas chains do not converge in the PC approach.

Figs 5(a) and 5(b) show the boxplots of the marginal posterior distributions for the ρ_ηs for strategies A and B, which control the dependence strength in each pair of $\boldsymbol{\theta}$s in the GP model.

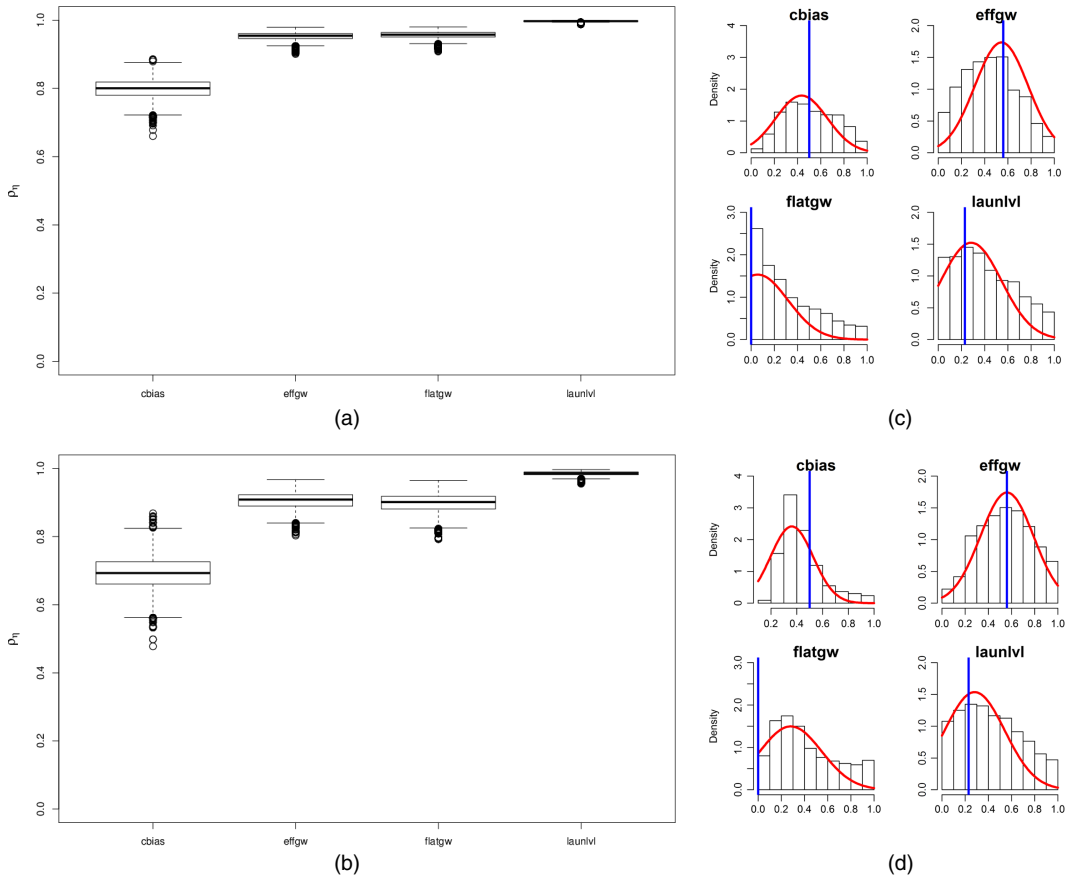

Fig. 5. (a), (b) Boxplots of the marginal posterior distribution for correlation parameters ρ_η for strategies A and B respectively and (c), (d) marginals for the posterior distribution of the GW parameters θ for strategies A and B respectively, I, true values (30 mb, February 2000)

The posterior density of ρ_4 converges to 1 which indicates a very weakly significant effect for θ_4. The marginal posterior densities for each $\boldsymbol{\theta}$ are displayed in Figs 5(c) and 5(d). Our approach provides a good compromise between computational feasibility and fidelity to the data by using only parsimonious representations. The results suggest that our technique on calibration of global scale outputs is effective.

4.2. Calibration against real observations

4.2.1. Posterior sampling

The final step is to carry out the calibration against real observations. We use zonal wind data obtained from the *European Centre for Medium-Range Weather Forecasts* 40-year reanalysis data archive ERA. We focus on the altitude of 1 mb, as the outputs in low altitudes are less sensitive to GW parameterizations and match the observations well already. Figs 6(a) and 6(b) show the ERA observations and zonal wind surfaces from standard outputs at 1 mb, February 2000. Under the same settings as described in the previous section, Fig. 6(c) shows the MCMC paths for three chains, with 6000 iterations, corresponding respectively to the calibration

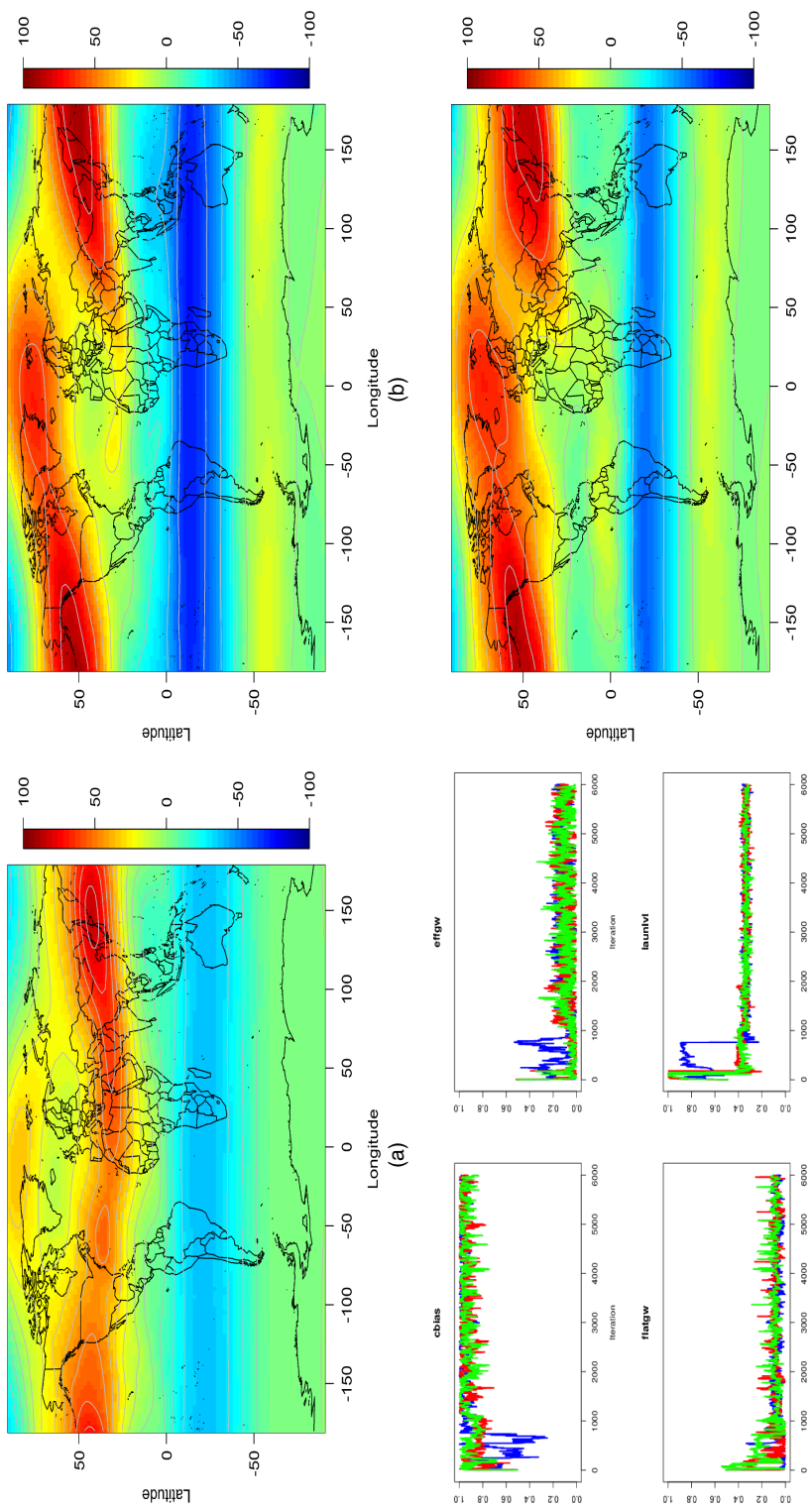

Fig. 6. Zonal wind from (a) the ERA data and (b) WACCM standard output; (c) MCMC paths for three chains (values are scaled to be $[0, 1]^4$; for details of the parameters and their uncertainty ranges refer to the main text) and (d) zonal wind generated by posterior from the calibration (1 mb, February 2000)

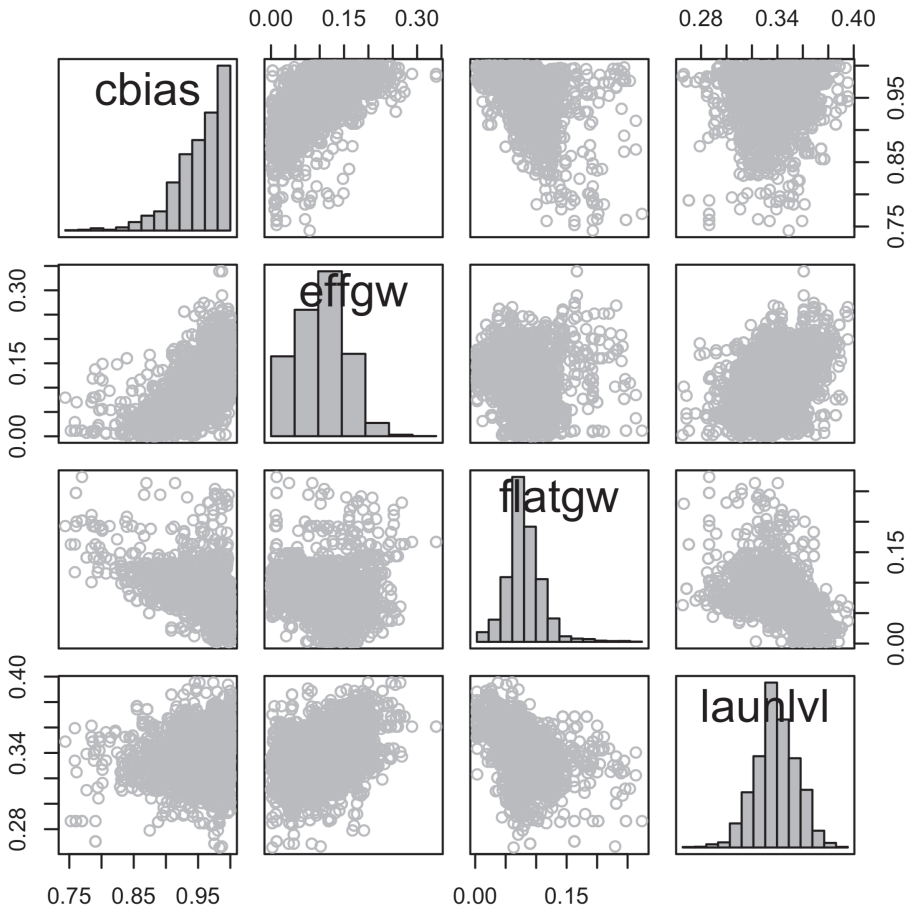

Fig. 7. Density of the posterior calibration parameters for zonal wind simulation (1 mb, February 2000)

parameters. The convergence of the MCMC chain can be established for the parameters θ_2, θ_3 and θ_4, with posterior modes 0.107 (SD = 0.051), 0.081 (SD = 0.029) and 0.339 (SD = 0.018) in the [0,1] scale respectively. The posterior mode of θ_1 lies within the upper bound. We then use posterior modes for these paths, collected as input values for the validation of the WACCM. The calibrated output displayed in Fig. 6(d) shows an RMSE of 18.15, which is a percentage improvement of 14.99% over the standard output (the RMSE between ERA observations and standard output is 21.35).

The resulting histograms for the calibration parameters, with the first 1000 iterations dropped as they are reckoned to be burn-in, are shown in Fig. 7. As expected from the MCMC plot, a normal distribution can be established for θ_2, θ_3 and θ_4. The distribution of θ_1 shown is skewed against the upper bound. It means that the possible calibrated value may lie outside the boundary. Since θ_1 represents the anisotropic velocity of zonal wind (the model default is assumed to be isotropic), the results suggest that we would need a more eastward component. It seems that this is a spurious effect of the simplicity in the parameterization. Indeed, to avoid losing the westward components and to acknowledge the physical reality, it may be helpful to have a 'bimodal' spectrum (Arfeuille *et al.*, 2013; Zhu *et al.*, 2017), with one peak in the eastward direction and another in the westward, and these two components do not have

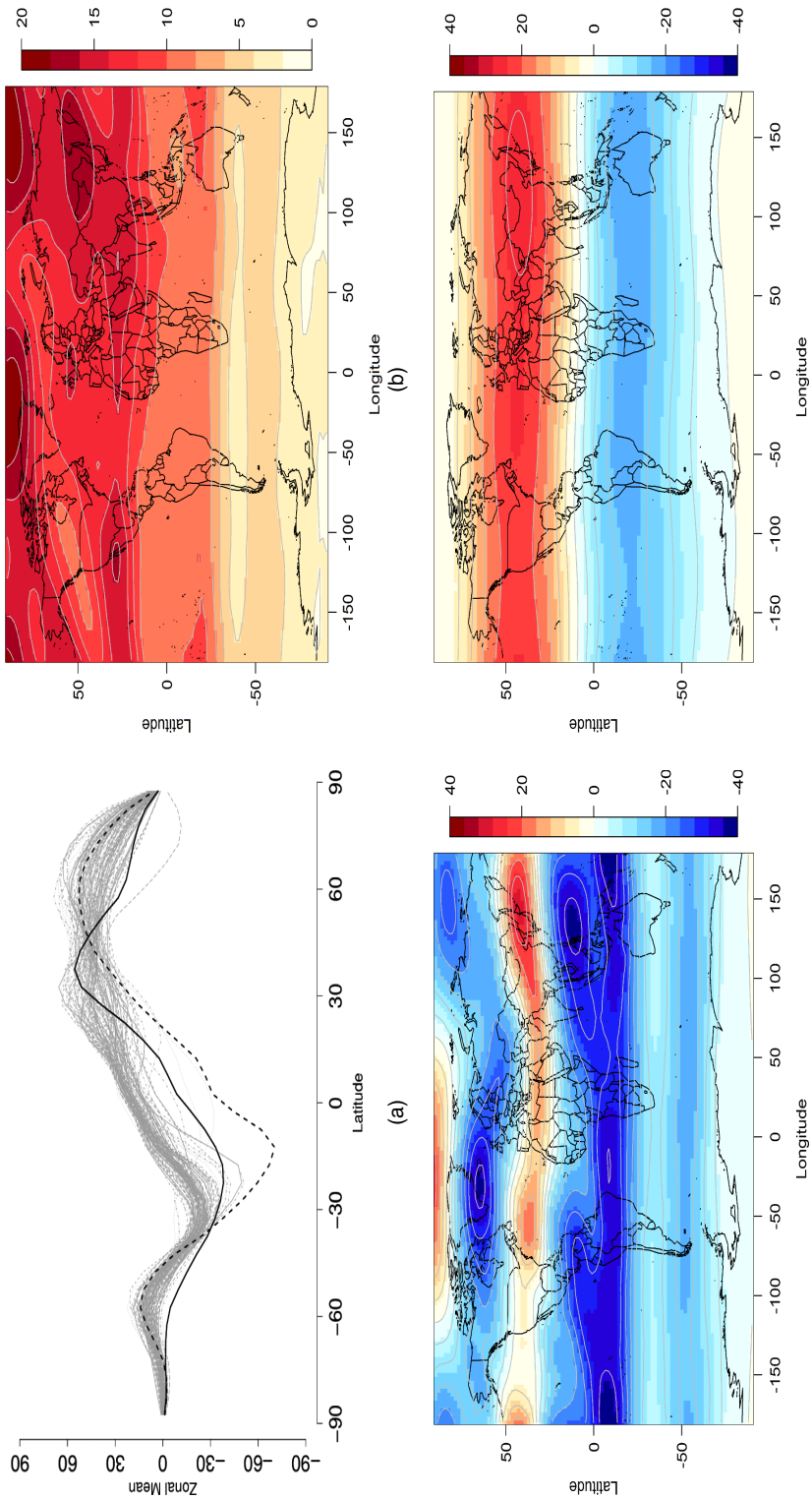

Fig. 8. (a) Zonal means of observations (———), standard (— — —) and model outputs (· · · · · ·), (b) grid-by-grid SDs map across model runs, (c) differences between observations and mean structure of model outputs and (d) model mean discrepancies map (1 mb, February 2000)

to be the same. Indeed, uneven amplitudes of the QBO easterly and westerly phases have often been observed in previous studies (Naujokat, 1986; Garcia *et al.*, 1997; Ern *et al.*, 2008). GW schemes are currently under development within the National Center for Atmospheric Research WACCM working group to improve the representation of the QBO and to help to fix the cold pole problem (Garcia *et al.*, 2017). Further development will enable us to have more flexible GWs schemes, but it is beyond the scope of the present setting for the climate simulation.

4.2.2. Model discrepancy and uncertainty

To assess the model uncertainty, Fig. 8(a) shows the zonal means calculated over every 5° belt of observations (black full curve), standard outputs (black broken curve) and each run of model output (grey dotted curves): note that the zonal means over the tropics are high compared with the observations and standard outputs. The input value of θ_3 in the standard output is at the lower border of the parameter range; this may produce relatively extreme behaviour over the tropics in our model runs. Fig. 8(b) represents the grid-by-grid SDs map across model outputs. We can see that the spatial process is clearly anisotropic and highly latitude dependent; the uncertainties are concentrated over the northern hemisphere, and little significant variability can be found over the southern hemisphere. Fig. 8(c) compares the differences between observations and mean structure of model outputs in each cell (with respect to 100 Latin hypercube designs), i.e. $\delta_{\text{initial}}(\mathbf{s}) = y^{\text{F}}(\mathbf{s}) - \bar{\eta}(\mathbf{s}, \boldsymbol{\theta})$, where $\bar{\eta}(\mathbf{s})$ are the output means over space. Fig. 8(c) provides potential features of model discrepancy over space (albeit not the true discrepancy). As expected, the model tends to overestimate the values over the tropics, which matches the pattern in Fig. 8(a). Besides, this surface seems to match the pattern in Fig. 8(b). The largest model bias (apart from the tropics) and variability both occur over north-east Asia and the North Pole. Fig. 8(d) shows the posterior mean discrepancies surface in the sense of $\delta^*(\mathbf{s}) = y^{\text{R}}(\mathbf{s}) - \eta(\mathbf{s}, \boldsymbol{\theta}^*)$. Our calibration reduces the bias (i.e. overestimation) over the tropics, as well as bias (i.e. underestimation) over the North Pole, whereas the bias over north-east Asia remains.

4.2.3. Validation

We use the mode from each posterior distribution to simulate 5 years (two QBO cycles) of zonal wind output. Fig. 9(a) shows monthly RMSEs at 1 mb globally, from 2000 to 2004. The overall averaged RMSE for the standard and calibrated outputs are 24.51 and 22.99 respectively, which are a small improvement. Indeed, our inertial GW scheme is designed to reduce the zonal wind bias over the tropics; we should not expect that our calibration will improve model simulations globally. We thus investigate RMSEs over the tropics over the same period. The RMSE trends are shown in Fig. 9(b). The overall averaged RMSE over the tropics for the standard and calibrated outputs are 26.64 and 17.87 respectively. Therefore the improvement is more significant over the tropics, with percentage improvement 32.9%. Simulations by our calibrated outputs outperform the standard code in 51 months out of the 60 months. The calibration of the WACCM with real observations over the whole output domain (i.e. including across altitudes) constitutes another level of complexity that needs joint scientific and statistical expertise. It is currently under investigation but is beyond the scope of this paper. Indeed, observations are scarce at these altitudes and show features that require specific understanding of the upper atmosphere dynamics before being used for calibration, and over many years of simulation for an adequate comparison.

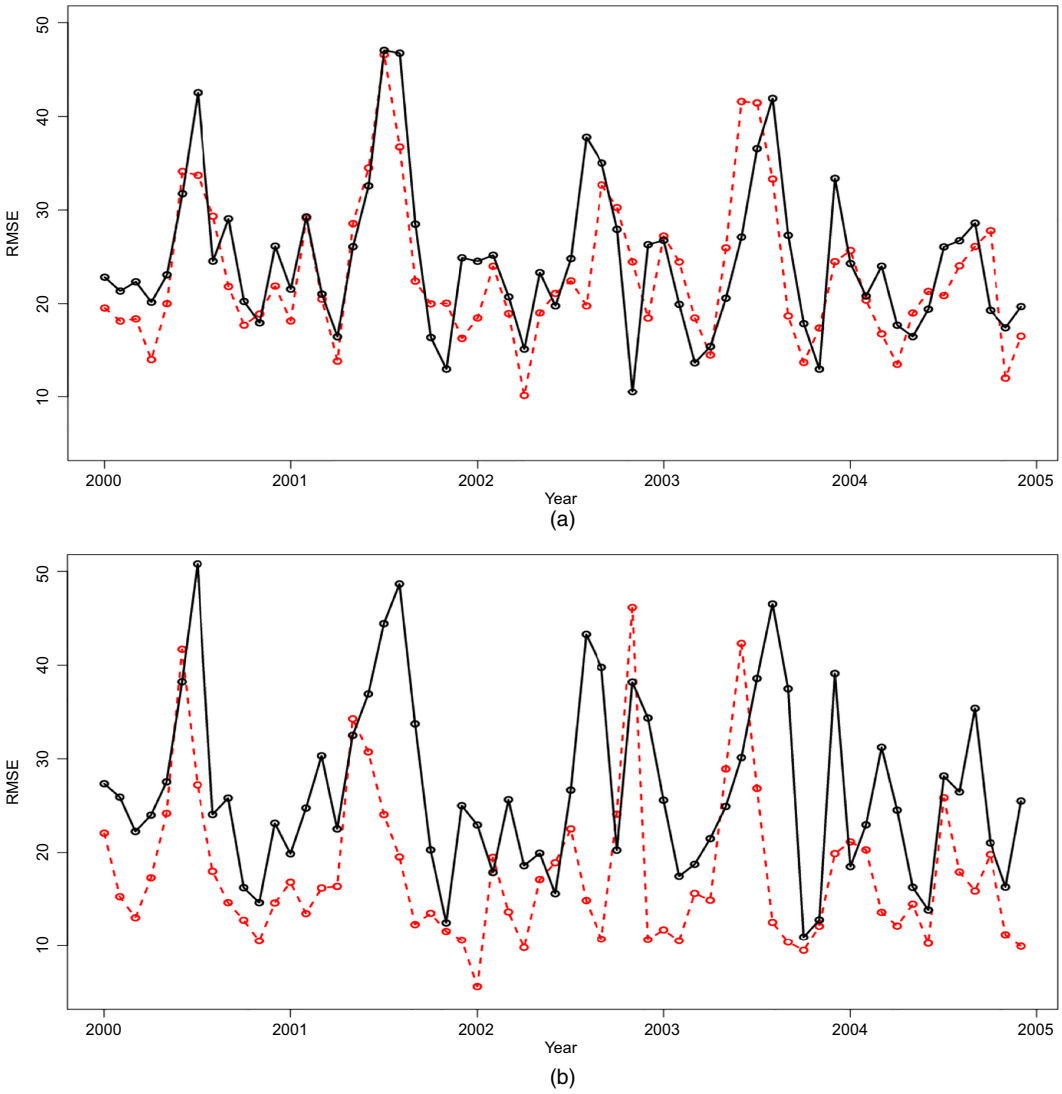

Fig. 9. Monthly RMSE trends between the ERA observations and standard outputs (———) or calibrated outputs (– – –), from 2000 to 2004: (a) global RMSEs; (b) tropical RMSEs

5. Conclusion and discussion

Our approach improved the calibration of large-scale computer model outputs distributed over space, parsimoniously, by using bases representations for the mean structures of the spatial surfaces. In addition, the INLA–SPDE approach was used to decompose its parameters characterizing non-stationarity over the same bases to improve calibration. The synthetic and real examples confirm the ability of our approach to perform calibration efficiently and accurately. Our method was inspired by the wavelets method of Bayarri *et al.* (2007), but with a different type of outputs: spatial *versus* time series. We can expect that the spherical wavelet decomposition may also be a possible alternative basis representation on the spatial domain, whenever appropriate (e.g. for sharp variations).

Another advantage of using the SH basis, compared with data-driven bases such as PCs, is that sequential design is allowed (Beck and Guillas, 2016), because the basis elements will not change, and model runs are obtained at the same grids or scattered locations. In this study we illustrate our technique on a specific horizontal output from the WACCM simulator. The SHT of model outputs can also be extended to time varying processes. As noted by Jones (1963), if a random field on a sphere varies with time, the representation becomes $\eta(\mathbf{s}, t) = \Sigma_{k=0}^{\infty} \Sigma_{h=-k}^{k} c_{k,h}(t)\psi_{k,h}(\mathbf{s})$, where $c_{k,h}(t)$ is an ordinary one-dimensional stochastic process. The set of all $c_{k,h}(t)$ form an infinite dimensional stochastic process. Theoretically we can represent model outputs in space–time settings with such representations. Nevertheless, in climate or chemistry–transport simulations, we often encounter not only outputs in time and horizontal resolution, but also in vertical resolution. Therefore extensions to four-dimensional correlations are needed, but they must maintain the computational tractability.

In our approach the covariance matrix is formulated as a block diagonal structure. We could relax this assumption and then adopt the block composite likelihood approach to accelerate the algorithm (Chang, Haran, Olson and Keller, 2015). Unfortunately, this approach covers only the stationary case (though it could be extended). Our approach naturally and efficiently models non-stationarity in space. Furthermore, there are cases where our approach is computationally more efficient than that of Chang, Haran, Olson and Keller (2015). Indeed, if m is large, their computational cost is about $O(\Sigma_{i=1}^{B} m_i^3)$, where $\Sigma_{i=1}^{B} m_i = m$ (depends on the number and size of blocks m_i), whereas our cost is $O(N_y^3 r^3)$, which is lower in many, but not all, applications. Since our climate experiment involves direct input–output projection, another potential extension of our approach is to combine recent developments in the Bayesian treed calibration technique, which partitions input space into subregions where our reduced rank approach can be applied, to accelerate the calibration further (Karagiannis *et al.*, 2017; Konomi *et al.*, 2017).

Acknowledgements

We thank Dr Hanli Liu (National Center for Atmospheric Research) for technical support in running the chemistry–climate model WACCM. SG gratefully acknowledges support through Natural Environment Research Council grant 'Probability, uncertainty and risk in the natural environment', NE/J017434/1.

References

Alexander, M. J., Geller, M., McLandress, C., Polavarapu, S., Preusse, P., Sassi, F., Sato, K., Eckermann, S., Ern, M., Hertzog, A., Kawatani, Y., Pulido, M., Shaw, T. A., Sigmond, M., Vincent, R. and Watanabe, S. (2010) Recent developments in gravity-wave effects in climate models and the global distribution of gravity-wave momentum flux from observations and models. *Q. J. R. Meteorol. Soc.*, **136**, 1103–1124.

Alexander, M. J. and Sato, K. (2015) Gravity wave dynamics and climate: an update from the SPARC gravity wave activity. *SPARC Newslett.*, **44**, 9–13.

Arfeuille, F., Luo, B., Heckendorn, P., Weisenstein, D., Sheng, J., Rozanov, E., Schraner, M., Brönnimann, S., Thomason, L. and Peter, T. (2013) Modeling the stratospheric warming following the Mt. Pinatubo eruption: uncertainties in aerosol extinctions. *Atmos. Chem. Phys.*, **13**, 11221–11234.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J. and Walsh, D. (2007) Computer model validation with functional output. *Ann. Statist.*, **35**, 1874–1906.

Beck, J. and Guillas, S. (2016) Sequential design with mutual information for computer experiments (MICE): emulation of a tsunami model. *J. Uncertnty Quant.*, **4**, 739–766.

Bhat, K. S., Haran, M. and Goes, M. (2010) Computer model calibration with multivariate spatial output: a case study. *Front. Statist. Decsin Makng Baysn Anal.*, 168–184.

Blangiardo, M. and Cameletti, M. (2015) *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Chichester: Wiley.

Bolin, D. and Lindgren, F. (2011) Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Statist.*, **5**, 523–550.
Bowman, V. E. and Woods, D. C. (2016) Emulation of multivariate simulators using thin-plate splines with application to atmospheric dispersion. *J. Uncertnty Quant.*, **4**, 1323–1344.
Brynjarsdóttir, J. and O'Hagan, A. (2014) Learning about physical parameters: the importance of model discrepancy. *Inv. Probl.*, **30**, article 114007.
Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2013) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Adv. Statist. Anal.*, **97**, 109–131.
Chakraborty, A., Mallick, B. K., McClaren, R. G., Kuranz, C. C., Bingham, D., Grosskopf, M. J., Rutter, E. M., Stripling, H. F. and Drake, R. P. (2013) Spline-based emulators for radiative shock experiments with measurement error. *J. Am. Statist. Ass.*, **108**, 411–428.
Chang, K.-L., Guillas, S. and Fioletov, V. E. (2015) Spatial mapping of ground-based observations of total ozone. *Atmos. Measmnt Tech.*, **8**, 4487–4505.
Chang, W., Haran, M., Olson, R. and Keller, K. (2014) Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Statist.*, **8**, 649–673.
Chang, W., Haran, M., Olson, R. and Keller, K. (2015) A composite likelihood approach to computer model calibration with high-dimensional spatial data. *Statist. Sin.*, **25**, 243–259.
Chang, K.-L., Petropavlovskikh, I., Cooper, O. R., Schultz, M. G. and Wang, T. (2017) Regional trend analysis of surface ozone observations from monitoring networks in eastern North America, Europe and East Asia. *Elem. Sci. Anth.*, **5**, 1–22.
Chunchuzov, I., Kulichkov, S., Perepelkin, V., Popov, O., Firstov, P., Assink, J. and Marchetti, E. (2015) Study of the wind velocity-layered structure in the stratosphere, mesosphere, and lower thermosphere by using infrasound probing of the atmosphere. *J. Geophys. Res. Atmos.*, **120**, 8828–8840.
Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.
Ern, M., Ploeger, F., Preusse, P., Gille, J., Gray, L., Kalisch, S., Mlynczak, M., Russell, J. and Riese, M. (2014) Interaction of gravity waves with the QBO: a satellite perspective. *J. Geophys. Res. Atmos.*, **119**, 2329–2355.
Ern, M., Preusse, P., Krebsbach, M., Mlynczak, M. and Iii, J. R. (2008) Equatorial wave analysis from SABER and ECMWF temperatures. *Atmos. Chem. Phys.*, **8**, 845–869.
Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J. and Taylor, K. E. (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscient. Modl Devlpmnt*, **9**, 1937–1958.
Fuglstad, G.-A., Lindgren, F., Simpson, D. and Rue, H. (2015) Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statist. Sin.*, 115–133.
Furrer, R. and Sain, S. R. (2009) Spatial model fitting for large datasets with applications to climate and microarray problems. *Statist. Comput.*, **19**, 113–128.
Garcia, R. R., Dunkerton, T. J., Lieberman, R. S. and Vincent, R. A. (1997) Climatology of the semiannual oscillation of the tropical middle atmosphere. *J. Geophys. Res. Atmos.*, **102**, 26019–26032.
Garcia, R. R., Smith, A. K., Kinnison, D. E., de la Cámara, Á. and Murphy, D. J. (2017) Modification of the gravity wave parameterization in the Whole Atmosphere Community Climate Model: motivation and results. *J. Atmos. Sci.*, **74**, 275–291.
Geller, M. A., Alexander, M. J., Love, P. T., Bacmeister, J., Ern, M., Hertzog, A., Manzini, E., Preusse, P., Sato, K., Scaife, A. A. and Zhou, T. (2013) A comparison between gravity wave momentum fluxes in observations and climate models. *J. Clim.*, **26**, 6383–6405.
Genton, M. G. and Kleiber, W. (2015) Cross-covariance functions for multivariate geostatistics. *Statist. Sci.*, **30**, 147–163.
Gneiting, T. (2013) Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, **19**, 1327–1349.
Gneiting, T., Kleiber, W. and Schlather, M. (2010) Matérn cross-covariance functions for multivariate random fields. *J. Am. Statist. Ass.*, **105**, 1167–1177.
Gramacy, R. B. and Apley, D. W. (2015) Local Gaussian process approximation for large computer experiments. *J. Computnl Graph. Statist.*, **24**, 561–578.
Hamilton, K. (2013) *Gravity Wave Processes: Their Parameterization in Global Climate Models*. New York: Springer Science and Business Media.
Higdon, D. M., Gattiker, J. R., Williams, B. and Rightley, M. (2008) Computer model calibration using high dimensional output. *J. Am. Statist. Ass.*, **103**, 570–5833.
Higdon, D. M., Kennedy, M., Cavendish, J. C., Cafeo, J. A. and Ryne, R. D. (2004) Combining field data and computer simulations for calibration and prediction. *SIAM J. Scient. Comput.*, **26**, 448–466.
Holden, P. B., Edwards, N. R., Garthwaite, P. H. and Wilkinson, R. D. (2015) Emulation and interpretation of high-dimensional climate model outputs. *J. Appl. Statist.*, **42**, 2038–2055.
Ilyas, M., Brierley, C. M. and Guillas, S. (2017) Uncertainty in regional temperatures inferred from sparse global observations: application to a probabilistic classification of El Niño. *Geophys. Res. Lett.*, **44**, 9068–9074.
Jones, R. H. (1963) Stochastic processes on a sphere. *Ann. Math. Statist.*, **34**, 213–218.

Jun, M., Knutti, R. and Nychka, D. W. (2008) Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? *J. Am. Statist. Ass.*, **103**, 934–947.
Jun, M. and Stein, M. L. (2007) An approach to producing space–time covariance functions on spheres. *Technometrics*, **49**, 468–479.
Jun, M. and Stein, M. L. (2008) Nonstationary covariance models for global data. *Ann. Appl. Statist.*, **2**, 1271–1289.
Karagiannis, G., Konomi, B. A. and Lin, G. (2017) On the Bayesian calibration of expensive computer models with input dependent parameters. *Spatl Statist.*, to be published.
Katzfuss, M. and Cressie, N. (2011) Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *J. Time Ser. Anal.*, **32**, 430–446.
Kennedy, M. and O'Hagan, A. (2001) Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc.* B, **63**, 425–464.
Kleiber, W. and Nychka, D. W. (2012) Nonstationary modeling for multivariate spatial processes. *J. Multiv. Anal.*, **112**, 76–91.
Konomi, B. A., Karagiannis, G., Lai, K. and Lin, G. (2017) Bayesian Treed Calibration: an application to carbon capture with AX sorbent. *J. Am. Statist. Ass.*, **112**, 37–53.
Lamarque, J.-F., Shindell, D. T., Josse, B., Young, P., Cionni, I., Eyring, V., Bergmann, D., Cameron-Smith, P., Collins, W. J., Doherty, R., Dalsoren, S., Faluvegi, G., Folberth, G., Ghan, S. J., Horowitz, L. W., Lee, Y. H., MacKenzie, I. A., Nagashima, T., Naik, V., Plummer, D., Righi, M., Rumbold, S. T., Schulz, M., Skeie, R. B., Stevenson, D. S., Strode, S., Sudo, K., Szopa, S., Voulgarakis, A. and Zeng, G. (2013) The Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): overview and description of models, simulations and climate diagnostics. *Geoscient. Modl Devlpmnt*, **6**, 179–206.
Large, W. G., Danabasoglu, G., McWilliams, J. C., Gent, P. R. and Bryan, F. O. (2001) Equatorial circulation of a global ocean climate model with anisotropic horizontal viscosity. *J. Phys. Oceanog.*, **31**, 518–536.
Lauritzen, P. H., Bacmeister, J. T., Callaghan, P. F. and Taylor, M. A. (2015) NCAR global model topography generation software for unstructured grids. *Geoscient. Modl Devlpmnt*, **8**, 3975–3986.
Lindgren, F. and Rue, H. (2015) Bayesian spatial and spatiotemporal modelling with R-INLA. *J. Statist. Softwr.*, **63**, 1–25.
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 (with discussion). *J. R. Statist. Soc.* B, **73**, 423–498.
Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. M. and Ye, K. Q. (2006) Variable selection for Gaussian process models in computer experiments. *Technometrics*, **48**, 478–490.
Liu, X., Guillas, S. and Lai, M.-J. (2016) Efficient spatial modelling using the SPDE approach with bivariate splines. *J. Computnl Graph. Statist.*, **25**, 1176–1194.
Liu, H., McInerney, J., Santos, S., Lauritzen, P., Taylor, M. and Pedatella, N. (2014) Gravity waves simulated by high-resolution Whole Atmosphere Community Climate Model. *Geophys. Res. Lett.*, **41**, 9106–9112.
Liu, H., Sassi, F. and Garcia, R. (2009) Error growth in a whole atmosphere climate model. *J. Atmos. Sci.*, **66**, 173–186.
Medvedev, A., Klaassen, G. and Beagley, S. (1998) On the role of an anisotropic gravity wave spectrum in maintaining the circulation of the middle atmosphere. *Geophys. Res. Lett.*, **25**, 509–512.
Muir, J. and Tkalčić, H. (2015) A method of spherical harmonic analysis in the geosciences via hierarchical Bayesian inference. *Geophys. J. Int.*, **203**, 1164–1171.
Naujokat, B. (1986) An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics. *J. Atmos. Sci.*, **43**, 1873–1877.
Nychka, D. W., Bandyopadhyay, S., Hammerling, D., Lindgren, F. and Sain, S. (2015) A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Computnl Graph. Statist.*, **24**, 579–599.
Nychka, D. W., Wikle, C. and Royle, J. A. (2002) Multiresolution models for non-stationary spatial covariance functions. *Statist. Modllng*, **2**, 315–331.
Rougier, J. (2008) Efficient emulators for multivariate deterministic functions. *J. Computnl Graph. Statist.*, **17**, 827–843.
Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. (2017) Bayesian computing with INLA: a review. *Rev. Statist. Appl.*, **4**, 395–421.
Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989) Design and analysis of computer experiments. *Statist. Sci.*, **4**, 409–423.
Salter, J. M., Williamson, D. B., Scinocca, J. and Kharin, V. (2018) Uncertainty quantification for spatio-temporal computer models with calibration-optimal bases. *Preprint arXiv:1801.08184*. Department of Mathematics, University of Exeter, Exeter.
Sang, H. and Huang, J. Z. (2012) A full scale approximation of covariance functions for large spatial data sets. *J. R. Statist. Soc.* B, **74**, 111–132.
Stein, M. L. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer Science and Business Media.
Stein, M. L. (2005) Space–time covariance functions. *J. Am. Statist. Ass.*, **100**, 310–321.
Stein, M. L. (2007) Spatial variation of total column ozone on a global scale. *Ann. Appl. Statist.*, **1**, 191–210.

Wang, C., Zhang, L., Lee, S.-K., Wu, L. and Mechoso, C. R. (2014) A global perspective on CMIP5 climate model biases. *Nat. Clim. Change*, **4**, 201–205.
Wendland, H. (2004) *Scattered Data Approximation*. Cambridge: Cambridge University Press.
Whittle, P. (1963) Stochastic processes in several dimensions. *Bull. Int. Statist. Inst.*, **40**, 974–994.
Williamson, D., Blaker, A. T., Hampton, C. and Salter, J. (2015) Identifying and removing structural biases in climate models with history matching. *Clim. Dyn.*, **45**, 1299–1324.
Williamson, D., Goldstein, M. and Blaker, A. (2012) Fast linked analyses for scenario-based hierarchies. *Appl. Statist.*, **61**, 665–691.
Wood, S. N. (2003) Thin plate regression splines. *J. R. Statist. Soc.* B, **65**, 95–114.
Yu, C., Xue, X., Wu, J., Chen, T. and Li, H. (2017) Sensitivity of the quasi-biennial oscillation simulated in WACCM to the phase speed spectrum and the settings in an inertial gravity wave parameterization. *J. Adv. Modlng Earth Syst.*, **9**, 389–403.
Yue, Y. and Speckman, P. L. (2010) Nonstationary spatial Gaussian Markov random fields. *J. Computtnl Graph. Statist.*, **19**, 96–116.
Zammit-Mangion, A., Rougier, J., Bamber, J. and Schön, N. (2015) Resolving the Antarctic contribution to sea-level rise: a hierarchical modelling framework. *Environmetrics*, **25**, 245–264.
Zhu, Y., Toon, O. B., Lambert, A., Kinnison, D. E., Bardeen, C. and Pitts, M. C. (2017) Development of a polar stratospheric cloud model within the Community Earth System Model: Assessment of 2010 Antarctic winter. *J. Geophys. Res. Atmos.*, **122**, 10418–10438.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for "Computer model calibration with large nonstationary spatial outputs: application to the calibration of a climate model"'.