Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Uncertainty quantification for computer models with spatial output using calibration-optimal bases

James M. Salter, Daniel B. Williamson, John Scinocca & Viatcheslav Kharin

To cite this article: James M. Salter, Daniel B. Williamson, John Scinocca & Viatcheslav Kharin (2018): Uncertainty quantification for computer models with spatial output using calibration-optimal bases, Journal of the American Statistical Association, DOI: 10.1080/01621459.2018.1514306

To link to this article: https://doi.org/10.1080/01621459.2018.1514306

Uncertainty quantification for computer models with spatial output using calibration-optimal bases

James M. Salter*
Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK.
and
Daniel B. Williamson
Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK.
and
John Scinocca
Canadian Centre for Climate Modelling and Analysis
Victoria, Canada.
and
Viatcheslav Kharin
Canadian Centre for Climate Modelling and Analysis
Victoria, Canada.

August 14, 2018

The authors gratefully acknowledge support from EPSRC fellowship No. EP/K019112/1 and support from the Canadian Network for Regional Climate and Weather Processes (CNRCWP), funded by the Natural Science and Engineering Research Council (NSERC Grant 433915-2012). The authors would also like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the Uncertainty Quantification programme where work on this paper was undertaken (EPSRC grant no EP/K032208/1). We would also like to thank Yanjun Jiao for managing our ensembles of CanAM4.

Abstract

The calibration of complex computer codes using uncertainty quantification (UQ) methods is a rich area of statistical methodological development. When applying these techniques to simulators with spatial output, it is now standard to use principal component decomposition to reduce the dimensions of the outputs in order to allow Gaussian process emulators to predict the output for calibration. We introduce the 'terminal case', in which the model cannot reproduce observations to within model discrepancy, and for which standard calibration methods in UQ fail to give sensible results. We show that even when there is no such issue with the model, the standard decomposition on the outputs can and usually does lead to a terminal case analysis. We present a simple test to allow a practitioner to establish whether their experiment will result in a terminal case analysis, and a methodology for defining calibration-optimal bases that avoid this whenever it is not inevitable. We present the optimal rotation algorithm for doing this, and demonstrate its efficacy for an idealised example for which the usual principal component methods fail. We apply these ideas to the CanAM4 model to demonstrate the terminal case issue arising for climate models. We discuss climate model tuning and the estimation of model discrepancy within this context, and show how the optimal rotation algorithm can be used in developing practical climate model tuning tools.

Keywords: Climate models, tuning, history matching, Bayesian calibration, rotation

1 Introduction

The design and analysis of computer experiments, now part of a wider cross-disciplinary endeavour called 'Uncertainty Quantification' or 'UQ', has a rich history in statistical methodological development as far back as the landmark paper by Sacks et al. (1989). The calibration of computer simulators, a term reserved for methods that locate simulator input values with outputs that are consistent with physical observations (the inverse problem), is a well studied problem in statistical science, with Kennedy and O'Hagan's Bayesian approach based on Gaussian processes the most widely used (Kennedy and O'Hagan, 2001).

The essence of the statistical approach to calibration is to combine a formal statistical model relating the computer simulator to real-world processes for which we have partial observations (Kennedy and O'Hagan, 2001; Goldstein and Rougier, 2009; Williamson et al., 2013), with a statistical representation of the relationship between inputs and outputs of the simulator based, typically, on Gaussian processes (Haylock and O'Hagan, 1996).

Extensions for computer simulators with spatio-temporal output have centred around projecting the output onto a basis and adapting calibration methods to the lower-dimensional projections of these fields. Though wavelets (Bayarri et al., 2007) and B-splines (Williamson et al., 2012) have been tried, the approach due to Higdon et al. (2008), based on the principal components of the simulator output, has become the default method. Statistical methodological developments in UQ have built on principal component methods (e.g. Wilkinson (2010); Chang et al. (2014, 2016)), and they have seen wide application, particularly in the analysis of climate models (Sexton et al., 2011; Chang et al., 2014; Pollard et al., 2016).

What statisticians term calibration is referred to as 'tuning' in the climate modelling community, a process that has a huge influence on the projections made by each modelling centre and by the Intergovernmental Panel on Climate Change (Stocker et al., 2013). Each modelling centre submits integrations of their climate model for 4 different forcing scenarios

(known as Representative Concentration Pathways) to each phase of the Coupled Model Intercomparison Project (Meehl et al., 2000), with the input parameters of the model 'tuned' prior to submission so that the model output compares favourably with certain key observations. The resulting integrations, and not the simulators themselves, are what most climate scientists call 'climate models' (i.e. simulators are not considered to be functions of these now fixed parameters). These integrations are used to discover physical mechanisms (Scaife et al., 2012), projected trends (Screen and Williamson, 2017), drivers of variability (Collins et al., 2010) and future uncertainty to aid policy making (Harris et al., 2006).

Despite the application of UQ methods to the calibration of 'previous-generation' climate models, referred to in the papers above and many others, UQ is not used for tuning within any of the major climate modelling centres (Hourdin et al., 2017). Instead, climate model parameters are often explored individually and tuning done by hand and eye, with the parameters changed, and the new run either accepted or rejected based on heuristic comparison with the current 'best' integration. Different descriptions of these processes are offered by Mauritsen et al. (2012); Williamson et al. (2017); Hourdin et al. (2017).

This lack of uptake of state-of-the-art statistical methodology for calibration amongst some of the world's most important computer simulators should give us pause for thought. The 'off-the-shelf' methodology, Bayesian calibration with principal components, is widely used elsewhere, well published, and is applied to many lower resolution climate models within the climate science literature. Is the lack of uptake a communication issue, or are there features of our methodology that mean it doesn't scale up well to climate simulators?

In this paper we show how the terminal case, wherein a simulator cannot be satisfactorily calibrated, manifests in the inference of standard UQ methodologies. We then demonstrate that even when there is a good solution to the inverse problem, the use of standard basis representations of spatial output (e.g. principal components across the design) can and regularly do lead to the terminal case and incorrect inference. We develop a simple test to

see whether an analysis will lead to the terminal case before performing the calibration and, when the terminal case is not guaranteed, provide a methodology for finding an optimal basis for calibration, via a basis rotation. The efficacy of our methodology is demonstrated through application to an idealised example, and its relevance to climate model tuning through application to the calibration of the atmosphere of the current Canadian climate model, CanAM4.

In Section 2, we review UQ methodologies for calibration and present the terminal case for scalar model output. Section 3 reviews the standard approach to handling spatial output and demonstrates the implications of the terminal case for these methods through an idealised example. Section 4 presents novel methods for finding optimal bases for calibration that overcome the terminal case issues and demonstrates the efficacy of calibrating with optimal bases for our example. In Section 5 we see that standard approaches always lead to terminal analyses in CanAM4, and show how our optimal basis methodology can be used in the process of climate model tuning. Section 6 contains discussion.

2 Calibration methodologies and the terminal case

We consider a computer simulator to be a vector-valued function $f(\boldsymbol{\theta}, \mathbf{x})$, with input parameters $\boldsymbol{\theta}$ that we wish to estimate/constrain, and 'control' or 'forcing' parameters, $\mathbf{x}$, both of which can be altered to perform computer experiments. For example, $\mathbf{x}$ might represent future CO_2 concentrations in a climate model. $f(\cdot, \mathbf{x})$ simulates a physical system $\mathbf{y}(\mathbf{x})$, and we have access to measurements or observations $\mathbf{z}$, of part or all of $\mathbf{y}$. The goal of calibration methods is to use $\mathbf{z}$ to learn about $\boldsymbol{\theta}$. In what follows we remove the control parameters, $\mathbf{x}$, to simplify the notation, as they are not involved in calibration, but in subsequent prediction.

The two statistical methodologies for calibration that we focus on here are Bayesian (or

probabilistic) calibration (Kennedy and O'Hagan, 2001; Higdon et al., 2008), and history matching with iterative refocussing (Craig et al., 1996; Vernon et al., 2010; Williamson et al., 2017). Both begin with the same type of assumption, namely that there exists a best input setting, $\boldsymbol{\theta}^*$, so that

$$\mathbf{y} = f(\boldsymbol{\theta}^*) + \boldsymbol{\eta}, \qquad \mathbf{z} = \mathbf{y} + \mathbf{e} \tag{1}$$

for mean-zero independent observation errors, $\mathbf{e}$, and model discrepancy, $\boldsymbol{\eta}$ (though history matching differs in only requiring uncorrelated terms in (1) rather than independent terms).

Both methods require an emulator, usually a Gaussian process representation of function $f(\boldsymbol{\theta})$, trained using runs $\mathbf{F} = (f(\boldsymbol{\theta}_1), ..., f(\boldsymbol{\theta}_n))$ based on design $\mathbf{X} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$. For scalar $f(\cdot)$, the general model is

$$f(\boldsymbol{\theta})|\boldsymbol{\beta}, \boldsymbol{\phi} \sim \mathrm{GP}\left(\boldsymbol{\beta}^T g(\boldsymbol{\theta}), R(|\boldsymbol{\theta} - \boldsymbol{\theta}'|; \boldsymbol{\phi})\right), \tag{2}$$

where $g(\boldsymbol{\theta})$ is a vector of specified regressors, $\boldsymbol{\beta}$ their coefficients, and $R(|\boldsymbol{\theta}-\boldsymbol{\theta}'|; \boldsymbol{\phi})$ a weakly stationary covariance function with parameters $\boldsymbol{\phi}$. The model is completed by specifying a prior on the parameters, $\pi(\boldsymbol{\beta}, \boldsymbol{\phi})$, and posterior inference given $\mathbf{F}$ follows naturally with

$$f(\boldsymbol{\theta})|\mathbf{F}, \boldsymbol{\beta}, \boldsymbol{\phi} \sim \mathrm{GP}(m^*(\boldsymbol{\theta}), R^*(\cdot, \cdot; \boldsymbol{\phi}))$$

with

$$\begin{aligned} m^*(\boldsymbol{\theta}) &= \boldsymbol{\beta}^T g(\boldsymbol{\theta}) + \mathbf{K}(\boldsymbol{\theta})\mathbf{V}^{-1}\left(\mathbf{F} - \boldsymbol{\beta}^T g(\mathbf{X})\right), \quad \mathbf{K}(\boldsymbol{\theta}) = R(\boldsymbol{\theta}, \mathbf{X}; \boldsymbol{\phi}), \\ R^*(\boldsymbol{\theta}, \boldsymbol{\theta}'; \boldsymbol{\phi}) &= R(\boldsymbol{\theta}, \boldsymbol{\theta}'; \boldsymbol{\phi}) - \mathbf{K}(\boldsymbol{\theta})\mathbf{V}^{-1}\mathbf{K}(\boldsymbol{\theta}')^T, \quad \mathbf{V} = R(\mathbf{X}, \mathbf{X}; \boldsymbol{\phi}). \end{aligned}$$

There are many variants on emulation, with some practitioners preferring no regressors

(Chen et al., 2016), different types of correlation function (including no correlation) (Kaufman et al., 2011; Salter and Williamson, 2016), and different priors, $\pi(\boldsymbol{\beta}, \boldsymbol{\phi})$, with some leading to partially analytic posterior inference (Haylock and O'Hagan, 1996). As history matching only requires posterior means and variances of the emulator, Bayes linear analogues are sometimes used (Vernon et al., 2010). Generalisations to multivariate Gaussian processes are natural (Conti and O'Hagan, 2010), and we address the difficulty with high dimensional output from Section 3 onwards.

2.1 Probabilistic calibration

Though the underlying statistical model and the emulator are similar for both history matching and probabilistic calibration, the assumptions placed upon $\boldsymbol{\theta}^*$, and the resulting inference, are quite different. Probabilistic calibration places a prior on $\boldsymbol{\theta}^*$, $\pi(\boldsymbol{\theta}^*)$, and a Gaussian process prior for the discrepancy, $\boldsymbol{\eta} \sim \text{GP}(0, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$, before deriving the posterior $\pi(\boldsymbol{\theta}^*, \boldsymbol{\eta}|\mathbf{F}, \mathbf{z})$, and marginalising for $\boldsymbol{\theta}^*$. The discussion of Kennedy and O'Hagan (2001), and the later paper by Brynjarsdóttir and O'Hagan (2014), argue that lack of identifiability between $\boldsymbol{\theta}^*$ and $\boldsymbol{\eta}$ mean that strong prior information on $\boldsymbol{\eta}$ or $\boldsymbol{\theta}^*$ is essential for effective probabilistic calibration to be possible.

2.2 History matching and iterative refocussing

Note that, given a discrepancy variance, probabilistic calibration must still give a posterior $\pi(\boldsymbol{\theta}^*|\mathbf{F}, \mathbf{z})$ that integrates to 1, thus predetermining an analysis that will point to some region of parameter space Θ as being 'most likely'. This can be undesirable in some application areas, as often the goal is to find out if the simulator *can* get 'close enough' to the observations, so that experiments predicting the future can be trusted. Climate model tuning is a good example of this, where part of the goal in tuning is to find out whether

it is the choice of parameters, or the parameterisation itself, that is leading model bias (Mauritsen et al., 2012; Hourdin et al., 2017).

The method of history matching and iterative refocussing allows the question of whether the model is fit for purpose to be answered as part of the calibration exercise, by altering the problem from one of looking for the best input directly, to one of trying to rule out regions of Θ that could not contain $\boldsymbol{\theta}^*$. A model unfit for purpose would have all of Θ ruled out. The method defines an implausibility measure, $\mathcal{I}(\boldsymbol{\theta})$, with

$$\mathcal{I}(\boldsymbol{\theta}) = (\mathbf{z} - \mathrm{E}[f(\boldsymbol{\theta})])^T (\mathrm{Var}(\mathbf{z} - \mathrm{E}[f(\boldsymbol{\theta})]))^{-1} (\mathbf{z} - \mathrm{E}[f(\boldsymbol{\theta})]), \tag{3}$$

where the expectations and variances of $f(\boldsymbol{\theta})$ are derived from the Gaussian process emulator description above, and are conditioned on the runs $\mathbf{F}$. If $\mathcal{I}(\boldsymbol{\theta})$ exceeds a threshold, T, that value of $\boldsymbol{\theta}$ is considered implausible and ruled out, thus defining a membership function for a subspace Θ' of Θ that is Not Ruled Out Yet (NROY), with $\Theta' = \{\boldsymbol{\theta} \in \Theta : \mathcal{I}(\boldsymbol{\theta}) \leq T\}$. The choice of T will be problem dependent, though typically, if $\mathbf{z}$ is one-dimensional, Pukelsheim's three sigma rule (Pukelsheim, 1994) is used to set $T = 9$ (Craig et al., 1996; Williamson et al., 2015). For ℓ-dimensional $\mathbf{z}$, Vernon et al. (2010) define $T = \chi^2_{\ell,0.995}$, the 99.5th percentile of the χ^2-distribution with ℓ degrees of freedom, or a conservative T can be derived through Chebysev's inequality.

A key principle behind history matching is its iterative nature. Following an initial set of runs, a 'wave' of history matching is conducted, leading to a certain percentage of Θ being ruled out. A new wave can then be designed within NROY space, and the procedure repeated, refocussing the search for possible $\boldsymbol{\theta}^*$ (Williamson et al., 2017).

Discrepancy and observation error variances, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}$, are important in both prob-

abilistic calibration and history matching. For the latter, equation (1) leads to

$$\mathrm{Var}(\mathbf{z} - \mathrm{E}[f(\boldsymbol{\theta})]) = \mathrm{Var}[f(\boldsymbol{\theta})] + \boldsymbol{\Sigma}_{\boldsymbol{\eta}} + \boldsymbol{\Sigma}_{\mathbf{e}}$$

in equation (3), whilst a Normal assumption on $\mathbf{e}$ in calibration means $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}$ appear in the likelihood.

In this paper, we focus on optimal spatial calibration for both types of methodology, as the issues we shall identify in Section 3 apply equally to both, though manifest in different ways, as we shall illustrate now with our discussion of the terminal case.

2.3 The terminal case

Consider a computer simulator, $f(\boldsymbol{\theta})$, a discrepancy variance assessment $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and an observation error variance $\boldsymbol{\Sigma}_{\mathbf{e}}$, where both variance matrices are positive definite. We define the terminal case to occur when $\mathcal{I}(\boldsymbol{\theta}) > T$, for T as above and for a perfect emulator, so that, in equation (3), $\mathrm{E}[f(\boldsymbol{\theta})] = f(\boldsymbol{\theta})$ and $\mathrm{Var}[f(\boldsymbol{\theta})] = 0$ for all $\boldsymbol{\theta}$. So, from a history matching perspective, the terminal case occurs when the model is too far from the observations at every point in parameter space according to the model discrepancy. Hence, all of Θ is ruled out, and the modellers must reconsider their simulator, or their error tolerance.

Within a probabilistic calibration framework, the terminal case implies a prior-data conflict so that, in some sense, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ has been 'misspecified' or the expert is 'wrong'. Lack of identifiability requires informative expert judgement for discrepancy (Brynjarsdóttir and O'Hagan, 2014), yet the difficulty in providing such judgements for complex computer simulators (Goldstein and Rougier, 2009) may mean that the terminal case would occur quite often in practice. It is therefore important to see how such prior-data conflict would manifest.

Figure 1 shows 20 steps of an iterative probabilistic calibration of a 1d $f(\boldsymbol{\theta})$ that we can

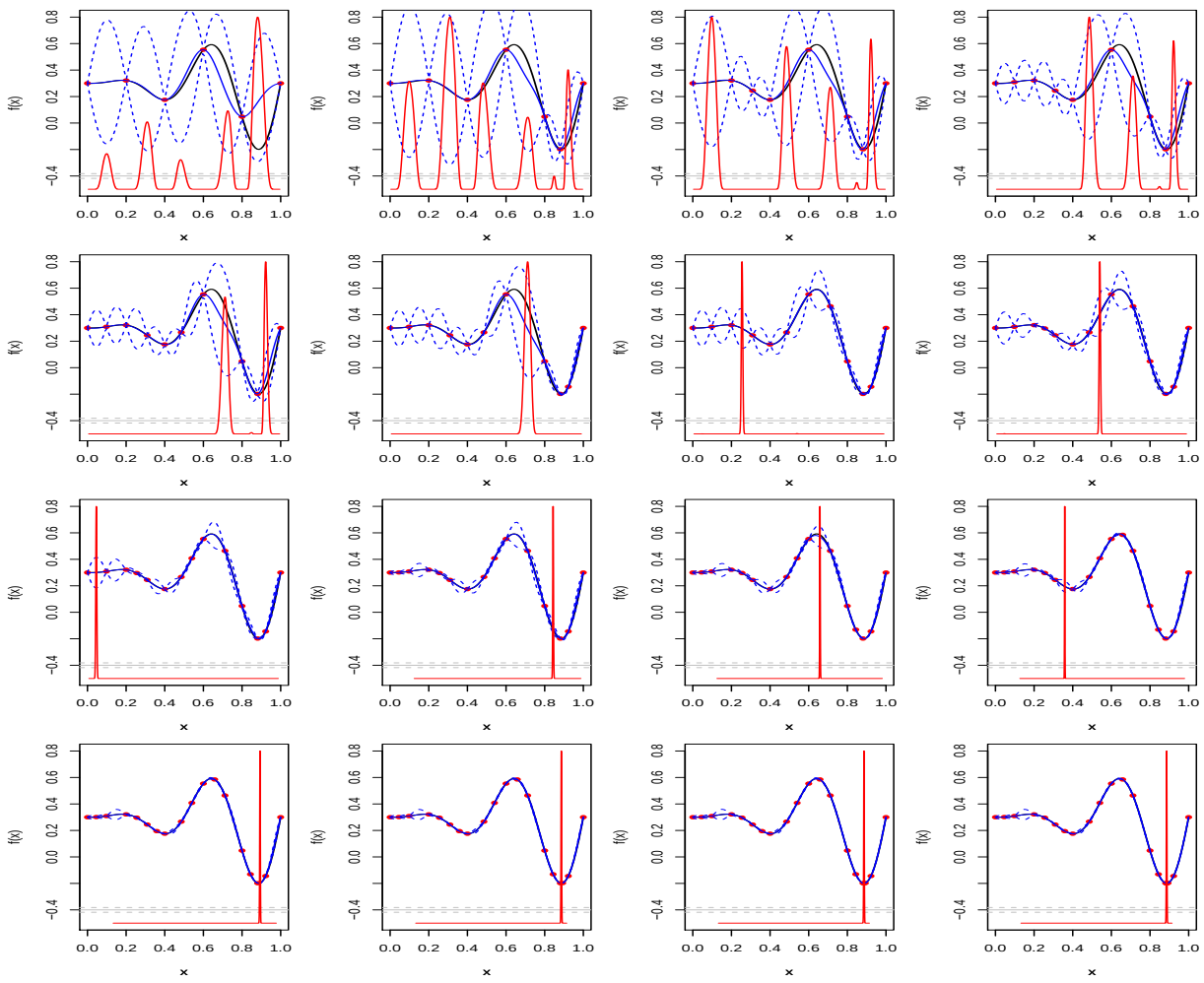

Figure 1: Showing 20 steps of an iterative probabilistic calibration of a computer simulator (black line) to observations (solid grey line) with $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}$ misspecified (dashed grey lines ± 3 standard deviations). Observations of the model (red dots) iteratively taken at the MAP estimate for $\boldsymbol{\theta}^*$ following the fitting of a GP emulator (mean solid blue line, ± 2 standard deviations dashed blue lines), and the posterior distribution of $\boldsymbol{\theta}^*$ overlaid at each step (solid red line).

evaluate quickly (black line), to observations (solid grey line), with $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}$ misspecified (dashed grey lines ± 3 standard deviations) so that the true function does not come as close

to the observations as the expert judgement indicates. Starting with an equally spaced 6 point design, a Gaussian process emulator is fitted for fixed correlation length (the mean function is the solid blue line, 2 standard deviation intervals are given by the blue dashed lines), and the posterior distribution $\pi(\boldsymbol{\theta}^* \mid \mathbf{z}, \mathbf{F})$ overlaid (solid red line). We then evaluate $f(\boldsymbol{\theta})$ at the maximum a posteriori estimate for $\boldsymbol{\theta}^*$, refit our Gaussian process, and compute the new posterior over $\boldsymbol{\theta}^*$ to produce the next plot.

From panel 6 onwards, we see the issue with the terminal case for probabilistic calibration. Our posterior beliefs are highly peaked at one particular $\boldsymbol{\theta}$ value, yet evaluating the model there completely shifts the peak to a location for which we had near zero prior density. Each evaluation of the simulator, which for climate models may take weeks or months, shifts the posterior spike to an unexpected (a priori) part of parameter space. It is often not efficient to run expensive simulators, such as climate models, that require expert time to run and manage, one run at a time (Williamson, 2015). The scientists that manage jobs on supercomputers, for example, require batches of runs that can be run in parallel. However, batch designs could be even worse here. Guided by the posterior density at each point, batch designs would be the near equivalent of one point at the MAP estimate, simply shifting the peak of the posterior to somewhere as yet unsampled.

Eventually, as we see from the bottom 4 panels, posterior uncertainty in $f(\boldsymbol{\theta})$ is sufficiently reduced, and $\pi(\boldsymbol{\theta}^* \mid \mathbf{F}, \mathbf{z})$ settles on the 'least bad' value of $\boldsymbol{\theta}$, where $f(\boldsymbol{\theta})$ is closest to the observations (though around 30 standard deviations away). For simulators with input spaces of much higher dimensions (the climate models we work with have typically specified 10-30 parameters to focus on, though these would be a subset of several hundred), we are unlikely to ever be able to reduce emulator uncertainty to the extent that the posterior spike settles over the least bad parameter setting. Hence, under an iterative procedure such as this, we would continue to chase the best input throughout parameter space, constantly moving the spike as in a game of whack-a-mole, until we run out of resources.

Our illustration of the terminal case shows that though careful subjective prior information is required for model discrepancy in order to overcome the identifiability issues with the calibration model, if those judgements lead to a prior-data conflict via a terminal case, good calibration will not be possible, and it will take a great deal of resource (enough data to build a near perfect emulator everywhere) to discover this. It would seem more natural to first history match in order to check we are not in a terminal case, and, if not, perform a probabilistic calibration within NROY space as in Salter and Williamson (2016).

Whichever calibration method, or combination of them, is preferred, it is important to understand this terminal case, as we shall show that even for models that can reproduce observations exactly, tractable methods for calibrating high dimensional output can result in a terminal case analysis.

3 Calibration with spatial output

For spatial fields, the most common approach to emulation and calibration involves projecting the model output onto a low-dimensional basis, $\mathbf{\Gamma}$, and emulating the coefficients, so that fewer emulators are required (Bayarri et al., 2007; Higdon et al., 2008; Wilkinson, 2010; Sexton et al., 2011) (although alternatives, such as emulating every grid box individually, have been applied, e.g. by Gu et al. (2016)).

Writing the model output $f(\boldsymbol{\theta}_i)$ as a vector of length ℓ, so that $\mathbf{F}$ has dimension $\ell \times n$, the singular value decomposition (SVD) is used to give n eigenvectors that can be used as basis vectors (equivalently, finding the principal components) (Higdon et al., 2008; Wilkinson, 2010; Sexton et al., 2011; Chang et al., 2014, 2016). For the size of model output typically explored using these methods, $\mathbf{\Gamma}$ will not be of full rank as $n << \ell$. This means that while $\mathbf{F}$ can be represented exactly by projection onto $\mathbf{\Gamma}$, general ℓ-dimensional fields will not have a perfect representation on $\mathbf{\Gamma}$. As the majority of the variability in $\mathbf{F}$ is usually

explained by only the first few eigenvectors, the basis is truncated after q vectors, giving a basis $\mathbf{\Gamma}_q = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q)$ of dimension $\ell \times q$, often chosen so that more than 95% of $\mathbf{F}$ is explained by $\mathbf{\Gamma}_q$. Various rules-of-thumb are used dependent on the problem, e.g. Higdon et al. (2008) truncate after 99%, while Chang et al. (2014) use 90%.

In order to emulate the model, the runs are first centred by subtracting their mean, $\boldsymbol{\mu}$, from each column of $\mathbf{F}$, giving the centred ensemble $\mathbf{F}_{\boldsymbol{\mu}}$ (we use the term ensemble to mean the collection of runs, as is common in the study of climate models). $\mathbf{F}_{\boldsymbol{\mu}}$ is then projected onto the basis $\mathbf{\Gamma}_q$, giving q coefficients associated with each parameter choice:

$$\mathbf{c}(\boldsymbol{\theta}_i) = (\mathbf{\Gamma}_q^T \mathbf{\Gamma}_q)^{-1} \mathbf{\Gamma}_q^T (f(\boldsymbol{\theta}_i) - \boldsymbol{\mu}). \tag{4}$$

Given q coefficients, a field of size ℓ is reconstructed via

$$f(\boldsymbol{\theta}_i) = \mathbf{\Gamma}_q \mathbf{c}(\boldsymbol{\theta}_i) + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{5}$$

with $\boldsymbol{\epsilon} = \mathbf{0}$ for $\boldsymbol{\theta}_i \in \mathbf{X}$. Emulators for the coefficients of the first q SVD basis vectors are then built:

$$c_i(\boldsymbol{\theta}) \sim \mathrm{GP}(m_i^*(\boldsymbol{\theta}), R_i^*(\boldsymbol{\theta}, \boldsymbol{\theta}; \boldsymbol{\phi})), \quad i = 1, \ldots, q. \tag{6}$$

Given these emulators, calibration can either be performed using the entire ℓ-dimensional output, with emulator expectations and variances transformed to the ℓ-dimensional space of the original field (Wilkinson, 2010), or on its q-dimensional basis representation, with the observations projected onto this basis (Higdon et al., 2008).

Calibration (via either history matching or probabilistic calibration) requires an informative prior process model for the spatial discrepancy, $\boldsymbol{\eta}$. This could be a stationary process defined through a simple covariance function over the output dimensions, though a richer class of non-stationary process defined via kernel convolution is often used (Hig-

don, 1998; Chang et al., 2014, 2016). These approaches specify a number of knots over the spatial field and define discrepancy to be a mixture of kernels around each of these knots. As with any calibration problem, however, strong prior information for discrepancy processes is essential to overcome identifiability issues, as discussed in Section 2.3. The way to include this information has been to fix the correlation parameters of the kernels and to have an infomative prior for their variances. With such a prior, a terminal case analysis is just as possible as for the 1D example we presented earlier.

Suppose that the prior on the process is strong enough to overcome identifiability issues and is such that *we don't have a terminal case*. When using a basis emulator to calibrate $f(\cdot)$, we may artificially induce a terminal case analysis, as reconstructions from coefficients on the basis are restricted to a q-dimensional subspace of ℓ-dimensional space. Further, it will not be clear whether our analysis implies that the model is incapable of reproducing $\mathbf{z}$, or that this was due to a poor basis choice. The SVD basis chooses the q-dimensional subspace that explains the maximum amount of variability in $\mathbf{F}$ with the fewest number of basis vectors. This choice does not guarantee that important directions in $\mathbf{F}$ that are consistent with $\mathbf{z}$ are preserved.

3.1 Illustrative example

We illustrate this problem with an idealised example of a 6 parameter function $f(\boldsymbol{\theta})$ (detailed in Section S1), with output given over a 10×10 grid. Observations, $\mathbf{z}$, are given by a known input parameter setting, $f(\boldsymbol{\theta}^*)$, with $\mathrm{N}(0, \boldsymbol{\Sigma}_\mathbf{e})$ observation error added (given in (S2), with $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ defined in (S3)), so that a calibration exercise should be able to identify $\boldsymbol{\theta}^*$. In our example, the great majority of the input space leads to output that is biased away from $\mathbf{z}$: the proportion of input space leading to output consistent with $\mathbf{z}$ is around 0.01%.

The first panel of Figure 2 shows the observations, $\mathbf{z}$, with a strong signal on the main

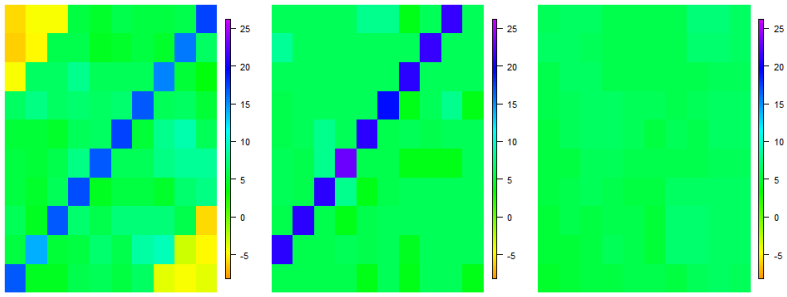

Figure 2: Left: the observations, $\mathbf{z}$, for our function. Centre: the ensemble mean. Right: the reconstruction of $\mathbf{z}$ using the truncated SVD basis.

diagonal. The second panel is the mean of the output field over a maximin Latin hypercube sample of size 60 in Θ (i.e. the mean of ensemble $\mathbf{F}$). The strong signal in the ensemble is a biased version of $\mathbf{z}$. In a climate context, this is analogous to the Gulf Stream being observed in the incorrect place in model output.

We calculate the SVD basis $\mathbf{\Gamma}$ as described above. Over 95% of the ensemble variability is explained by projection onto the first four basis vectors, which we refer to as the 'truncated basis', $\mathbf{\Gamma}_4$. If we project $\mathbf{z}$ onto this basis and reconstruct the original field using these coefficients, via equations (4) and (5), we obtain the field given by the third panel of Figure 2: the distinctive pattern found in $\mathbf{z}$ has been lost. That is, spatial calibration with $\mathbf{\Gamma}_4$ would ultimately rule out parameter space that contained the true coefficients due to poor reconstruction, suggesting that, for reconstructions of the field using $\mathbf{\Gamma}_4$, we are in the terminal case.

Fitting Gaussian process emulators to the coefficients given by projection of $\mathbf{F}_{\boldsymbol{\mu}}$ onto the four basis vectors, the expectation and variance at $\boldsymbol{\theta}$ is given by

$$\mathrm{E}[\mathbf{c}(\boldsymbol{\theta})] = (\mathrm{E}[c_1(\boldsymbol{\theta})], \ldots, \mathrm{E}[c_4(\boldsymbol{\theta})])^T, \quad \mathrm{Var}[\mathbf{c}(\boldsymbol{\theta})] = diag(\mathrm{Var}[c_1(\boldsymbol{\theta})], \ldots, \mathrm{Var}[c_4(\boldsymbol{\theta})]).$$

To probabilistically calibrate or history match on the original field, we require $\mathrm{E}[f(\boldsymbol{\theta})]$ and $\mathrm{Var}[f(\boldsymbol{\theta})]$ in terms of the coefficient emulators. These are

$$\mathrm{E}[f(\boldsymbol{\theta})] = \boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\boldsymbol{\theta})], \quad \mathrm{Var}[f(\boldsymbol{\theta})] = \boldsymbol{\Gamma}_q \mathrm{Var}[\mathbf{c}(\boldsymbol{\theta})] \boldsymbol{\Gamma}_q^T + \boldsymbol{\Gamma}_{-q} \boldsymbol{\Sigma}_{-q} \boldsymbol{\Gamma}_{-q}^T$$

where $\boldsymbol{\Gamma}_{-q}$ contains the discarded basis vectors, and $\boldsymbol{\Sigma}_{-q}$ is a diagonal matrix with the associated eigenvalues as the diagonal elements (Wilkinson, 2010).

Calibrating in the coefficient subspace requires projection of $\mathbf{z}$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}$ onto $\boldsymbol{\Gamma}_4$. For example, the implausibility in (3) on the coefficients becomes

$$\tilde{\mathcal{I}}(\boldsymbol{\theta}) = (\boldsymbol{\Gamma}_q^T \mathbf{z} - \mathrm{E}[\mathbf{c}(\boldsymbol{\theta})])^T (\mathrm{Var}[\mathbf{c}(\boldsymbol{\theta})] + \boldsymbol{\Gamma}_q^T \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\Gamma}_q + \boldsymbol{\Gamma}_q^T \boldsymbol{\Sigma}_{\mathbf{e}} \boldsymbol{\Gamma}_q)^{-1} (\boldsymbol{\Gamma}_q^T \mathbf{z} - \mathrm{E}[\mathbf{c}(\boldsymbol{\theta})]). \quad (7)$$

Using the 0.995 value of the chi-squared distribution with 100 degrees of freedom to history match via (3), we rule out the whole parameter space, Θ, and so we are in the terminal case. Hence probabilistic calibration gives peaked prediction at the incorrect value of $\boldsymbol{\theta}^*$, consistent with the description given in Section 2.3 (see SM section S1.1, Figures S3, S4).

By history matching on the coefficients instead, using (7), and setting T using the chi-squared distribution with 4 degrees of freedom, we find an NROY space consisting of 3.8% of Θ. However, we rule out 58% of the parameter space that was consistent with $\mathbf{z}$, as the important directions for comparing the model to observations are not contained in $\boldsymbol{\Gamma}_4$.

Whether we are calibrating on the original field, or on the coefficients, the 'best' result we are able to find is that given by the reconstruction of $\mathbf{z}$ with $\boldsymbol{\Gamma}_4$, given in the final panel of Figure 2. On the field, we are in the terminal case. On the coefficients, we are attempting to find runs that give coefficients that lead to this reconstruction, regardless of what happens in the directions we are interested in (i.e. the main diagonal pattern). Henceforth, we choose to focus on calibration on the field, as it compares all aspects of the

observed output to the model, rather than a few summaries of it.

4 Optimal basis selection

For calibration, there are two main requirements for a basis, $\mathbf{B}$, representing high dimensional output: being able to represent $\mathbf{z}$ with $\mathbf{B}$ (a feature not guaranteed by principal component methods), and retaining enough signal in the chosen subspace to enable accurate emulators to be built for the basis coefficients (as principal components do).

A natural method for satisfying the first goal is to minimise the error given when the observations are reconstructed using $\mathbf{B}$. Define the reconstruction error, $\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z})$ via

$$\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z}) = \|\mathbf{z} - \mathbf{B}(\mathbf{B}^T\mathbf{W}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{W}^{-1}\mathbf{z}\|_{\mathbf{W}}. \tag{8}$$

where $\|\mathbf{v}\|_{\mathbf{W}} = \mathbf{v}^T\mathbf{W}^{-1}\mathbf{v}$ is the norm of vector $\mathbf{v}$, and $\mathbf{W}$ is an $\ell \times \ell$ positive-definite weight matrix. By setting $\mathbf{W} = \boldsymbol{\Sigma}_{\mathbf{e}} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, $\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z})$ is analogous to (3), and is the implausibility when we know the basis coefficients exactly (so that the emulator variance is 0).

As $\mathbf{W}$ will not generally be a multiple of the identity matrix, the SVD projection from (4) is not appropriate for $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$. Therefore, (4) becomes

$$\mathbf{c}(\boldsymbol{\theta}_i) = (\mathbf{B}^T\mathbf{W}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{W}^{-1}(f(\boldsymbol{\theta}_i) - \boldsymbol{\mu}), \tag{9}$$

with this projection minimising the error in $\|\cdot\|_{\mathbf{W}}$ (Section S2), hence the definition of the reconstruction error in (8).

We present everything in full generality for positive definite $\mathbf{W}$. Therefore, $\mathbf{B}$ is an orthonormal basis if $\mathbf{B}^T\mathbf{W}^{-1}\mathbf{B} = \mathbb{I}_n$. A basis with this property can be obtained using

generalised SVD (Jolliffe, 2002), with $\mathbf{W} = \mathbb{I}_\ell$ giving the usual SVD decomposition:

$$\mathbf{F}_{\boldsymbol{\mu}}^T = \mathbf{U}\mathbf{D}\mathbf{B}^T, \quad \mathbf{U}^T\mathbf{U} = \mathbb{I}_n,\, \mathbf{B}^T\mathbf{W}^{-1}\mathbf{B} = \mathbb{I}_n.$$

As a measure of whether emulators can be built, we use the proportion of variability explained by projection of the ensemble onto each basis vector $\mathbf{b}_k$, $\mathcal{V}_k(\mathbf{B}, \mathbf{F}_{\boldsymbol{\mu}})$, with

$$\mathcal{V}_k(\mathbf{B}, \mathbf{F}_{\boldsymbol{\mu}}) = \frac{\sum_{j=1}^n \|\mathbf{b}_k(\mathbf{b}_k^T\mathbf{W}^{-1}\mathbf{b}_k)^{-1}\mathbf{b}_k^T\mathbf{W}^{-1}(f(\boldsymbol{\theta}_j) - \boldsymbol{\mu})\|_{\mathbf{W}}}{\sum_{j=1}^n \|f(\boldsymbol{\theta}_j) - \boldsymbol{\mu}\|_{\mathbf{W}}}. \tag{10}$$

The proportion of ensemble variability explained by $\mathbf{B}$, $\mathcal{V}(\mathbf{B}, \mathbf{F}_{\boldsymbol{\mu}})$, is

$$\mathcal{V}(\mathbf{B}, \mathbf{F}_{\boldsymbol{\mu}}) = \frac{\sum_{j=1}^n \|\mathbf{B}(\mathbf{B}^T\mathbf{W}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{W}^{-1}(f(\boldsymbol{\theta}_j) - \boldsymbol{\mu})\|_{\mathbf{W}}}{\sum_{j=1}^n \|f(\boldsymbol{\theta}_j) - \boldsymbol{\mu}\|_{\mathbf{W}}}. \tag{11}$$

The SVD basis maximises $\mathcal{V}_k(\mathbf{B}, \mathbf{F}_{\boldsymbol{\mu}})$ for each k, given the previous vectors and subject to orthogonality.

Prior to building emulators and performing calibration for a given basis, we can assess whether we are in the terminal case or not. For history matching threshold T, if $\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z}) > T$ then we are in the terminal case on $\mathbf{B}$, and would even rule out values of $\boldsymbol{\theta}^*$ that reproduce $\mathbf{z}$ exactly. If $\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z}) > T$ for some $\{\mathbf{B}, \mathbf{W}\}$, we may view $\mathbf{W}$ as having been misspecified, as in the terminal case described in Section 2.3. However, we may also have under-explored the output dimension of $f(\cdot)$, so that $\mathbf{B}$ does not allow us to get close enough to $\mathbf{z}$. We revisit this test in the context of optimal basis choice in Section 4.2.

Figure 3 compares $\mathcal{V}(\cdot, \mathbf{F}_{\boldsymbol{\mu}})$ and $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$ for the example of Section 3. We refer to plots of this type as VarMSE plots. The red line represents $\mathcal{R}_{\mathbf{W}}(\mathbf{B}_k, \mathbf{z})$, and the blue line shows $\mathcal{V}(\mathbf{B}_k, \mathbf{F}_{\boldsymbol{\mu}})$, for each truncated basis, $\{\mathbf{B}_k\}_{k=1}^n$. The vertical dotted line indicates where the basis is truncated if we wish to explain 95% of the ensemble variability, and the horizontal

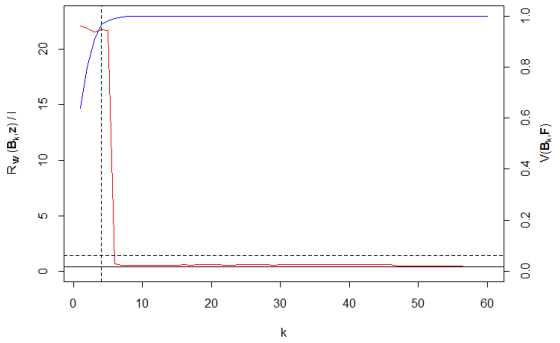

Figure 3: A plot showing how the reconstruction error (red) and proportion of ensemble variability explained (blue) change as the SVD basis is increased in size, for $\mathbf{W} = \boldsymbol{\Sigma}_{\mathbf{e}} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.

dotted line represents the history matching bound, T. The solid horizontal line is equal to $\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z})$. For the SVD basis in our example, we see that $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$ is large (compared to T) until $k = 6$, where the error decreases below the threshold, indicating that the sixth basis vector contains patterns that are important for explaining $\mathbf{z}$. As further basis vectors are added, $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$ continues to decrease, suggesting that patterns relevant for representing $\mathbf{z}$ are in fact included in $\mathbf{B}$ for this example. However, the later basis vectors explain low percentages of the variability in the ensemble, with the low signal to noise ratio of projected coefficients making accurate emulation impossible. If we could emulate the coefficients for the 5th and 6th basis vectors, we would more accurately represent $\mathbf{z}$, although this rapid decrease in the reconstruction error is a feature of our example, rather than a general property of the SVD basis, and therefore we still truncate at 95% for illustrative purposes.

The SVD basis aims to maximise the blue line for each basis vector added, whereas, for calibration, we require the red line to be below T. The problem of basis selection for calibration is one of trading off these two requirements, reducing $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$ while ensuring that each $\mathcal{V}_k(\cdot, \mathbf{F}_{\boldsymbol{\mu}})$ is large enough to enable emulators to be built. Given that the full SVD basis may contain information and patterns that allow the observations to be more

accurately represented, the information contained in this basis may be combined in such a way that the resulting basis is suitable for calibration, with important low-order patterns blended with those that explain more of the ensemble variability.

4.1 Rotating a basis

Performing a rotation of an ensemble basis $\mathbf{B}$ using an $n \times n$ rotation matrix, $\mathbf{\Lambda}$, rearranges the signal from the ensemble, potentially allowing the new truncated basis to be a better representation of $\mathbf{z}$. A general $n \times n$ rotation matrix $\mathbf{\Lambda}$ can be defined by composing $n(n-1)/2$ matrices that give a rotation by an angle around each pair of dimensions (Murnaghan, 1962). Our goal is to find $\mathbf{\Lambda}$ such that $\mathbf{B\Lambda}$ minimises $\mathcal{R}_{\mathbf{W}}((\mathbf{B\Lambda})_q, \mathbf{z})$, subject to constraints on $\mathcal{V}_k(\cdot, \mathbf{F}_{\boldsymbol{\mu}})$ that allow the projected coefficients to be emulated.

To directly define a rotation matrix $\mathbf{\Lambda}$ via optimisation requires a large number of angles to be found, even when the ensemble size is small. Instead, we take an iterative approach, selecting new basis vectors sequentially while minimising $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$ at each step, in such a way that guarantees that the resulting basis is an orthogonal rotation of the original basis.

Given $p < n$ basis vectors, $\mathbf{B}_p = (\mathbf{b}_1, \ldots, \mathbf{b}_p)$, we define the 'ensemble residual' as

$$\mathbf{F}_{\epsilon} = \mathbf{F}_{\boldsymbol{\mu}} - \mathbf{B}_p(\mathbf{B}_p^T\mathbf{W}^{-1}\mathbf{B}_p)^{-1}\mathbf{B}_p^T\mathbf{W}^{-1}\mathbf{F}_{\boldsymbol{\mu}}$$

This represents the variability in the ensemble not explained by $\mathbf{B}_p$. Define the 'residual basis', $\mathbf{B}_{\epsilon}$, to be the matrix containing the right singular vectors of $\mathbf{F}_{\epsilon}$. The residual basis gives basis vectors that explain the remaining variability in $\mathbf{F}_{\boldsymbol{\mu}}$, given vectors $\mathbf{B}_p$.

4.2 The optimal rotation algorithm

Given an orthogonal basis $\mathbf{B}$ for $\mathbf{F}_{\boldsymbol{\mu}}$ with dimension $\ell \times n$; a positive definite $\ell \times \ell$ weight matrix $\mathbf{W} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}} + \boldsymbol{\Sigma}_{\mathbf{e}}$; a vector $\mathbf{v}$, where v_i is the minimum proportion of the ensemble variability to be explained by the i^{th} basis vector; the total proportion of ensemble variability to be explained by the basis v_{tot}; and a bound T (usually that implied by history matching, $T = \chi^2_{\ell,0.995}$), we find an optimal basis for performing calibration as follows:

1. If $\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z}) > T$, stop and revisit the specification of $\mathbf{W}$, or add more runs to $\mathbf{F}_{\boldsymbol{\mu}}$. Else set $k = 1$.

2. Let $\boldsymbol{\Gamma}^*_k = (\boldsymbol{\gamma}^*_1, \ldots, \boldsymbol{\gamma}^*_{k-1}, \mathbf{B}\boldsymbol{\lambda}_k)$ and set

$$\boldsymbol{\lambda}^*_k = argmin_{\boldsymbol{\lambda}_k} \mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}^*_k, \mathbf{z})$$

   such that $\mathcal{V}_k(\boldsymbol{\Gamma}^*_k, \mathbf{F}_{\boldsymbol{\mu}}) \geq v_k$. Define the new normalised vector as

$$\boldsymbol{\gamma}^*_k = \frac{\mathbf{B}\boldsymbol{\lambda}^*_k}{\sqrt{\|\mathbf{B}\boldsymbol{\lambda}^*_k\|_{\mathbf{W}}}},$$

   and set $\boldsymbol{\Gamma}^*_k = (\boldsymbol{\gamma}^*_1, \ldots, \boldsymbol{\gamma}^*_{k-1}, \boldsymbol{\gamma}^*_k)$.

3. Find the residual basis given $\boldsymbol{\Gamma}^*_k$, $\mathbf{B}^k_\epsilon$, and form the orthogonal rank n basis

$$\boldsymbol{\Gamma}^* = (\boldsymbol{\Gamma}^*_k, [\mathbf{B}^k_\epsilon]_{n-k}).$$

4. Define $q \geq k$ as the minimum value satisfying $\mathcal{V}(\boldsymbol{\Gamma}^*_q, \mathbf{F}_{\boldsymbol{\mu}}) \geq v_{tot}$, where $\boldsymbol{\Gamma}^*_q$ represents the first q columns of $\boldsymbol{\Gamma}^*$. If $\mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}^*_q, \mathbf{z}) < T$, then stop, and return $\boldsymbol{\Gamma}^*_q$ as the truncated basis for calibration. Else, set $k = k + 1$ and $\mathbf{B} = [\mathbf{B}^k_\epsilon]_{n-k}$, and return to step 2.

Prior to applying the algorithm, we must specify an initial basis, $\mathbf{B}$, a weight matrix, $\mathbf{W}$, and the parameters v_{tot} and $\mathbf{v}$ to control the amount of variability explained by each basis vector. We use the SVD basis (with respect to $\mathbf{W}$) for $\mathbf{B}$, however other choices are possible, e.g. we could apply Gram-Schmidt to the ensemble itself and rotate this, or apply a different scaling to the SVD basis.

At each step, our algorithm selects the linear combination of a given basis that minimises $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$, subject to explaining a given percentage of ensemble variability, and given any previously selected basis vectors. If the defined truncation $\mathbf{\Gamma}_q^*$ satisfies $\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q^*, \mathbf{z}) < T$, then the algorithm terminates, as standard residual variance maximising basis vectors no longer lead to a terminal case analysis. We allow a basis to be identified that satisfies our two goals: we do not rule out $\mathbf{z}$, and have coefficients that can be emulated, if $\mathbf{v}$ is set appropriately. To optimise for $\boldsymbol{\lambda}_k$, we use simulated annealing (Yang Xiang et al., 2013), although any optimisation scheme that converges could be used.

The check in step 1 of our algorithm is due to the following result (proved in S2):

Result 1 (Invariance of $\mathcal{R}_{\mathbf{W}}(\cdot, \cdot)$ to rotation)**.** *For a rotation matrix* $\mathbf{\Lambda}$ *of dimension* $k \times k$*, and a set of basis vectors* $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$*, we have*

$$\mathcal{R}_{\boldsymbol{W}}(\boldsymbol{B}_k, \boldsymbol{z}) = \mathcal{R}_{\boldsymbol{W}}(\boldsymbol{B}_k \mathbf{\Lambda}, \boldsymbol{z}), \quad k = 1, \ldots, n \tag{12}$$

Regardless of the rotation that is applied to $\mathbf{B}$, we cannot reduce the reconstruction error below that given by the full basis originally. However, because the SVD basis is always truncated prior to this minimum value being reached, we can search for a rotation that rearranges the information from the SVD basis in such a way that satisfies

$$\mathcal{R}_{\mathbf{W}}((\mathbf{B}\mathbf{\Lambda})_q, \mathbf{z}) << \mathcal{R}_{\mathbf{W}}(\mathbf{B}_q, \mathbf{z}), \tag{13}$$

incorporating important, potentially low-order, patterns into the q basis vectors that we emulate. Hence step 1 of the algorithm provides an important test as to whether our ensemble and uncertainty assessment, $(\mathbf{F}, \mathbf{W})$, are sufficient to avoid a terminal case analysis, and shows when a rotation exists, up to the choice of $\mathbf{v}$.

Theorem 1. $\mathbf{\Gamma}^*$ *in step 3 of the optimal rotation algorithm is an orthogonal rotation of* $\boldsymbol{B}$.

The results and proofs required, and the proof of Theorem 1 itself, are found in Section S2. Given that $\mathbf{B}$ passes step 1 of the algorithm, existence of an optimal rotation depends on the choice of $\mathbf{v}$:

Theorem 2. *At the* k^{th} *iteration of the optimal rotation algorithm, given an orthogonal* $\mathbf{\Gamma}^*_{k-1}$ *that satisfies* $\mathcal{V}_j(\mathbf{\Gamma}^*_{k-1}, \boldsymbol{F_\mu}) \geq v_j$, $j = 1, \ldots, k-1$, $\exists\, \boldsymbol{\gamma}^*_k \perp \mathbf{\Gamma}^*_{k-1}$ *with* $\mathcal{V}(\boldsymbol{\gamma}^*_k, \boldsymbol{F_\mu}) \geq v_k$ *and* $\mathcal{R}_{\boldsymbol{W}}(\mathbf{\Gamma}^*_k, \boldsymbol{z}) \leq \mathcal{R}_{\boldsymbol{W}}(\tilde{\mathbf{\Gamma}}_k, \boldsymbol{z}) \leq \mathcal{R}_{\boldsymbol{W}}(\mathbf{\Gamma}^*_{k-1}, \boldsymbol{z})$, *for* $\mathbf{\Gamma}^*_k = (\mathbf{\Gamma}^*_{k-1}, \boldsymbol{\gamma}^*_k), \tilde{\mathbf{\Gamma}}_k = (\mathbf{\Gamma}^*_{k-1}, \tilde{\boldsymbol{\gamma}}_k)$, *and* $\mathcal{V}(\tilde{\boldsymbol{\gamma}}_k, \boldsymbol{F_\mu}) \geq v_k \,\forall \tilde{\boldsymbol{\gamma}}_k \perp \mathbf{\Gamma}^*_{k-1} \iff \mathcal{V}_1(\boldsymbol{B}^{k-1}_\epsilon, \boldsymbol{F_\mu}) \geq v_k$. *In this case the algorithm converges to* $\boldsymbol{\gamma}^*_k$.

Proof. By construction, $\mathcal{V}_1(\mathbf{B}^{k-1}_\epsilon, \mathbf{F_\mu}) = \max_j \mathcal{V}_j(\mathbf{B}^{k-1}_\epsilon, \mathbf{F_\mu}) = \max \mathcal{V}(\boldsymbol{\epsilon}, \mathbf{F_\mu})\ \forall\, \boldsymbol{\epsilon} \in span\{\mathbf{F}^{k-1}_\epsilon\}$.
Hence if $\mathcal{V}_1(\mathbf{B}^{k-1}_\epsilon, \mathbf{F_\mu}) < v_k$, $\nexists \boldsymbol{\gamma}^*_k = \mathbf{B}^{k-1}_\epsilon \boldsymbol{\lambda}_k$ such that $\mathcal{V}(\boldsymbol{\gamma}^*_k, \mathbf{F_\mu}) \geq v_k$.
If $\mathcal{V}_1(\mathbf{B}^{k-1}_\epsilon, \mathbf{F_\mu}) \geq v_k \implies \exists \boldsymbol{\gamma}^*_k = \mathbf{B}^{k-1}_\epsilon \boldsymbol{\lambda}_k$ with

i) $\mathcal{V}(\boldsymbol{\gamma}^*_k, \mathbf{F_\mu}) \geq v_k$,

ii) $\boldsymbol{\gamma}^*_k \perp \mathbf{\Gamma}^*_{k-1}$ (by Theorem 1),

iii) $\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}^*_k, \mathbf{z}) \leq \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}^*_{k-1}, \mathbf{z})$: let $\mathbf{c}^*_{k-1} = ((\mathbf{\Gamma}^*_{k-1})^T \mathbf{W}^{-1} \mathbf{\Gamma}^*_{k-1})^{-1} (\mathbf{\Gamma}^*_{k-1})^T \mathbf{W}^{-1} \mathbf{z}$ and $\mathbf{c}^*_k = ((\mathbf{\Gamma}^*_k)^T \mathbf{W}^{-1} \mathbf{\Gamma}^*_k)^{-1} (\mathbf{\Gamma}^*_k)^T \mathbf{W}^{-1} \mathbf{z}$ be the coefficients given by projecting $\mathbf{z}$ onto $\mathbf{\Gamma}^*_{k-1}$ and $\mathbf{\Gamma}^*_k$ respectively. Let $\mathbf{c}^* = (\mathbf{c}^*_{k-1}, 0)$, then

$$\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}^*_{k-1}, \mathbf{z}) = \|\mathbf{z} - \mathbf{\Gamma}^*_{k-1}\mathbf{c}^*_{k-1}\|_{\mathbf{W}} = \|\mathbf{z} - \mathbf{\Gamma}^*_k \mathbf{c}^*\|_{\mathbf{W}} \geq \|\mathbf{z} - \mathbf{\Gamma}^*_k \mathbf{c}^*_k\|_{\mathbf{W}} = \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}^*_k, \mathbf{z}).$$

as by construction $\mathbf{c}^*_k$ minimises the reconstruction error in the $\mathbf{W}$ norm.

Finally, $\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_k^*, \mathbf{z}) \leq \mathcal{R}_{\mathbf{W}}(\tilde{\mathbf{\Gamma}}_k, \mathbf{z}) \, \forall \, \tilde{\boldsymbol{\gamma}}_k = \mathbf{B}_\epsilon^{k-1} \tilde{\boldsymbol{\lambda}}_k$ with $\mathcal{V}(\tilde{\boldsymbol{\gamma}}_k, \mathbf{F}_{\boldsymbol{\mu}}) \geq v_k$ (by convergence of the optimiser, e.g. Aarts and Van Laarhoven (1985) for simulated annealing). □

In practice, when applying our algorithm to high dimensional model output, we have found that only a small number (three or fewer) of iterations have been required, hence $\mathbf{v}$ often has a low dimension. The values of $\mathbf{v}$ required will depend on the problem, with a different approach required when a small number of vectors explain the majority of the ensemble, compared to when a large proportion of the variability is spread across many SVD basis vectors. In the former case, the values of $\mathbf{v}$ may be relatively high, whilst in the latter they can be lower, relative to the proportion explained by the equivalent SVD basis vectors. A reasonable approach is to initially set $\mathbf{v}$ as half of the proportion explained by the corresponding SVD basis vectors, reducing these further if the resulting q is too large. As Theorem 2 shows, it is possible to set $\mathbf{v}$ in such a way that the algorithm is unable to find a suitable basis. If we cannot find a k^{th} basis vector that satisfies the variability constraint, given $\mathbf{\Gamma}_{k-1}^*$, then a basis doesn't exist for this choice of $\mathbf{v}$, and the specification needs revisiting: either v_k needs to be decreased, or an earlier constraints needs relaxing.

The choice v_{tot} is also a concern for the standard UQ approaches based on principal components. In our experience, using similar rules (e.g. 95% or 99%) to the SVD applications leads to 0-2 extra basis vectors required.

In an application, it may be desirable to include certain physical patterns, deemed to be important, in our basis $\mathbf{B}$, which may not lie within the subspace defined by $\mathbf{F}_{\boldsymbol{\mu}}$. In this case, if we have p selected physical vectors, $\mathbf{B}_p = (\mathbf{b}_1, \ldots, \mathbf{b}_p)$, combining these with the first $n-p$ vectors of the residual basis will not necessarily explain all of the variability in $\mathbf{F}_{\boldsymbol{\mu}}$. The algorithm may be applied to the $n+p$ vectors given by the physical vectors and the full residual basis, giving a rotation of this space rather than of $\mathbf{F}_{\boldsymbol{\mu}}$. As truncation occurs after the majority of variability of $\mathbf{F}_{\boldsymbol{\mu}}$, v_{tot}, is explained, the resulting truncated basis, while not

strictly a rotation of the subspace defined by $\mathbf{F}_{\boldsymbol{\mu}}$, will exhibit similar qualities, and may be superior for representing $\mathbf{z}$, if important physical patterns can be emulated when combined with signal from the ensemble.

To perform the algorithm with basis vectors from outside the subspace defined by $\mathbf{F}_{\boldsymbol{\mu}}$, rather than finding linear combinations of the residual basis at step $k > 1$, $\mathbf{B} = (\mathbf{B}_p, \mathbf{B}_\epsilon)$ is used at each step, with orthogonality imposed after each new basis vector has been selected, via Gram-Schmidt (as by Result S3, applying Gram-Schmidt does not affect $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$).

4.3 Idealised example continued

We now apply the optimal rotation algorithm to the example of Section 3. We set $\mathbf{v} = (0.4, 0.1, 0.1)$, $v_{tot} = 0.95$, and $\mathbf{B}$ as the SVD basis, with the projection of (4) used for consistency with Section 3.1, to show that rotation fixes the described problems. One iteration of the algorithm finds a basis that satisfies $\mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}_q^*, \mathbf{z}) < T$, with $q = 5$ (i.e. we need the first 4 vectors of the residual basis so that $\boldsymbol{\Gamma}_q^*$ explains at least 95% of $\mathbf{F}_{\boldsymbol{\mu}}$).

The reconstruction of $\mathbf{z}$ with this basis, and associated VarMSE plot, are shown in Figure 4. Now, our basis allows us to accurately represent $\mathbf{z}$ with the leading vectors, as the important patterns from low-order eigenvectors have been combined with the leading patterns (hence an additional vector being required to explain more than 95% of $\mathbf{F}_{\boldsymbol{\mu}}$).

Performing history matching as before, and using the reconstructions of the original fields rather than the coefficients, we find that 31.5% of Θ is now in NROY space (Figure S5). Performing our previous check on the accuracy of the match, we find that no runs consistent with $\mathbf{z}$ have been ruled out.

As we are no longer in the terminal case, we perform probabilistic calibration on the field. The posterior densities found by calibrating on $\boldsymbol{\Gamma}_q^*$ are shown in Figure S3, with the average simulator output given by samples from this posterior in the first plot of Figure

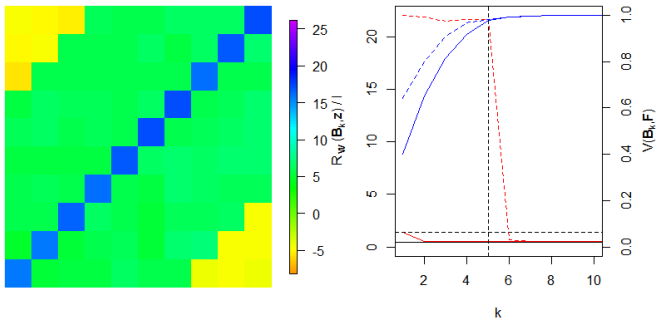

Figure 4: The reconstruction of $\mathbf{z}$ using the truncated basis $\mathbf{\Gamma}_q^*$, and the VarMSE plot for this basis, with the truncated SVD basis given by the red and blue dotted lines.

5. While the samples here are not consistent with $\mathbf{z}$, as the off-diagonal is too strong, we have been able to identify runs where there is signal on the main diagonal. This is because the rotated basis allows for this direction of the output space to be searched. The limited signal in the important directions from $\mathbf{F}$ has been extracted and used to guide calibration.

We continue the calibration by running a new design within NROY space. This new design should contain more signal in the direction of $\mathbf{z}$, and hence it should be possible to find a rotation that reduces $\mathcal{R}_{\mathbf{W}}(\cdot, \mathbf{z})$ further than at the previous wave. We select 60 points from the wave 1 NROY space and run $f(\cdot)$ at these points to give the wave 2 ensemble. We perform a rotation, and emulate and calibrate using the wave 2 ensemble. History matching reduces NROY space to 3.1% of Θ (Figure S7). If we instead perform probabilistic calibration, with zero density assigned to regions outside of the wave 1 NROY space, we find the average output field in the 2nd plot of Figure 5 (posteriors in Figure S6).

These results represent a large improvement over performing only one wave. We have ruled out the majority of Θ, allowing future runs to be focussed in this region. Probabilistic calibration is more accurate, with samples containing a strong diagonal, as with $\mathbf{z}$.

Repeating the process, our wave 3 ensemble contains patterns more consistent with $\mathbf{z}$ than in previous waves, and hence the truncated SVD basis does not rule out the re-

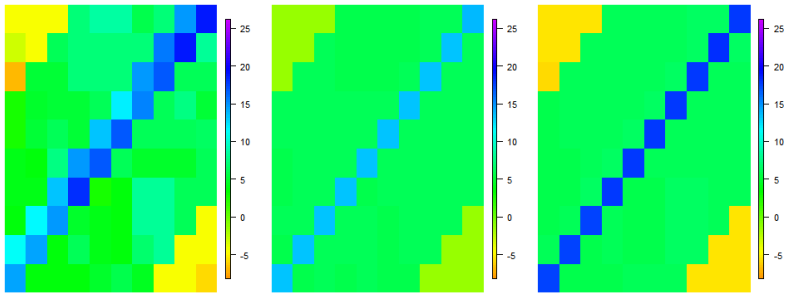

Figure 5: The mean output $f(\boldsymbol{\theta})$ for samples of $\boldsymbol{\theta}$ from the probabilistic calibration posterior, for the wave 1 rotated basis, the wave 2 rotated basis and the wave 3 SVD basis.

construction of $\mathbf{z}$, and no rotation is required. Following emulation for this basis, history matching leads to an NROY space consisting of 2% of Θ (Figure S8). Probabilistic calibration (in the wave 2 NROY space) gives the average output in Figure 5 (posteriors in Figure S6), showing that our samples are now consistent with $\mathbf{z}$.

5 Application to tuning climate models

In this section, we will demonstrate that optimal rotation is an important and necessary tool if attempting to perform UQ for climate model tuning. However, as we will discuss, climate model tuning is not a solved problem, and it would be of limited value to simply show how calibration with our method can lead to an improvement in one output field over the standard methods, without necessarily improving the whole model or addressing the concerns of the community. We will motivate our discussion using the current Canadian atmosphere model, CanAM4.

CanAM4 is an Atmospheric Global Climate Model, which integrates the primitive equations on a rotating sphere employing a spherical-harmonic spatial discretization truncated triangularly at total wavenumber 63 (T63), with 37 vertical levels spanning the troposphere

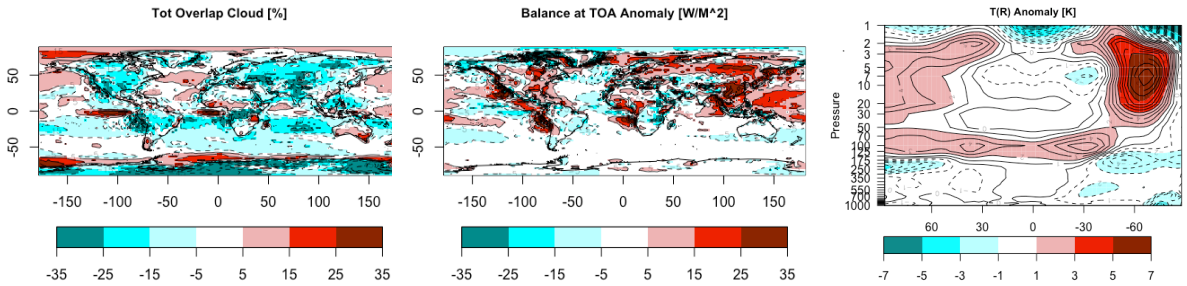

Figure 6: The standard CanAM4 anomaly field for a) CLTO, b) RTMT and c) TA.

and stratosphere (von Salzen et al., 2013). CanAM4 has a large number of adjustable, 'free', parameters of which 13 will be varied here. The climatological influence of each set of free parameters is determined from 5-year 'present-day' integrations with prescribed sea-surface temperatures and sea-ice. Model output is represented on the 'linear' 128×64 Gaussian grid corresponding to the model's T63 spectral resolution.

There are many output fields that must be checked for consistency with the observed climate when tuning the parameters of a climate model (in the case of CanAM4 there are more than 20). Here we focus on just 3 2D fields: vertical air temperature (TA), the top of the atmosphere radiative balance (RTMT, Wm^{-2}) and the cloud overlap percentage (CLTO). For RTMT and CLTO, the output is given over a longitude-latitude grid, so that $\ell = 8192$. TA is the temperature averaged over longitude for each latitude and vertical pressure level so that $\ell = 2368$. There is also a temporal aspect to the output, with monthly values for every grid box; however, we remove this here by averaging over 5 years of June, July, August (JJA) output.

When evaluating and tuning the model, spatial anomaly plots are routinely examined to see how the model compares with observations. An anomaly plot shows the difference between the model and the observations with a blue-white-red colour scale set such that blue is 'too negative', white is 'alright' and red is 'too positive'. So, for example, in a

temperature anomaly plot, red areas show where the model is too warm (for the modellers) compared to observations. Figure 6 shows anomaly plots for CLTO, RTMT, and TA for the standard configuration of CanAM4, with the colour scales representing the standard colours used by the modellers when tuning the model.

A goal of tuning is to try to reduce or remove biases that are visible from these plots. Yet equally important is to learn which biases cannot be removed simply by adjusting the model parameters. This is the search for 'structural errors' in the model (what statisticians would call model discrepancies). Structural errors indicate that there are flaws with individual parametrisations, or with the way they interact, that cannot be fixed by tuning. Understanding what these structural errors are so that they might be addressed either as part of this phase of development or for the next is one of the major goals of tuning (Williamson et al., 2015). However, joint estimation of model discrepancy variances and model parameters is not possible without strong prior information (Brynjarsdóttir and O'Hagan, 2014) due to lack of identifiability.

When working with CanAM4 then, our goal is to use history matching with a 'tolerance to error' discrepancy variance (Williamson et al., 2015, 2017) that aims to reduce the size of NROY space, so that, ultimately, in a calibration exercise we have strong prior information about $\boldsymbol{\theta}^*$ and some structured information on discrepancy. A formal methodology for achieving this is beyond the scope of this paper. However, we will demonstrate that optimal rotation is a crucial component for any attempt of this nature.

We designed 62 runs of CanAM4, varying 13 parameters and using a k-extended Latin Hypercube (Williamson, 2015). Figure 7 shows VarMSE plots for the output fields CLTO, RTMT and TA for this ensemble. The weight matrix $\mathbf{W}$ used for the reconstructions represents our tolerance to model error (we discussed its correspondence to model discrepancy ($\mathbf{W} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}} + \boldsymbol{\Sigma}_{\mathbf{e}}$) in Section 4), and the red lines in these plots represent 2 alternatives based on interpreting the colour scales pertaining to the white regions in Figure 6. We assume

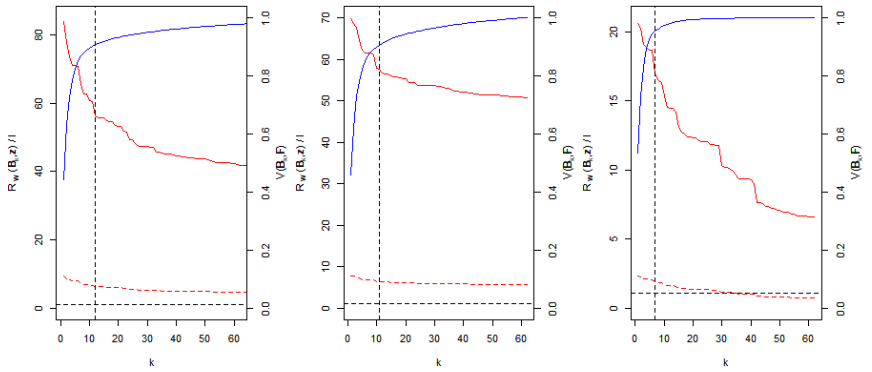

Figure 7: VarMSEplots for a) CLTO b) RTMT and c) TA, with $\mathbf{W}$ based on 1SD (dotted line) and 3SD (solid line). The dotted horizontal line indicates T.

that the white region represents 3 standard deviations (solid red line) and 1 standard deviation (dashed red line), and set a diagonal $\mathbf{W}$ accordingly. The solid red lines on each plot indicate that we have a terminal case analysis under the small model discrepancy.

The larger discrepancy indicates a terminal case in CLTO and RTMT, and that 35 basis vectors would be enough to adequately reconstruct TA. However, the blue line in the TA plot shows that there is so little ensemble signal on the basis vector coefficients after arguably 20 (or fewer) basis vectors, that calibration on 35 basis vectors is not possible. If discrepancy were increased (an operation that involves scaling the red line until it lies below the bound T represented by the dashed horizontal line in the plots), all 3 panels demonstrate that the reconstruction error under SVD decreases too slowly, so that a large number of basis vectors, each with coefficients that are increasingly difficult to emulate due to the decreasing ensemble signal, would be required to avoid a terminal case analysis.

Suppose model discrepancy $\mathbf{\Sigma_\eta} >> \mathbf{\Sigma_e}$ so that we can consider $\mathbf{W} = \mathbf{\Sigma_\eta}$ in the following. In order to use optimal rotation, we require $\mathbf{W}$ such that $\mathcal{R}_{\mathbf{W}}(\mathbf{B}, \mathbf{z}) < T$, which is not true under our specification above for RTMT and CLTO. If we really believed our $\mathbf{\Sigma_\eta}$ represented the climate model's ability to reproduce observed climate, then this indicates

that we need a larger ensemble in order to explore the model's variability. In that case, it may be desirable to follow a procedure like the one we present here to design these runs.

In our case, we believe it is clear that we have misspecified model discrepancy. In fact, we used a place-holder tolerance to error, so this analysis indicates that we are not tolerant enough to model error (at this stage). To explore model discrepancy, we first perform a rotation under the $\mathbf{W}$ given above, using the algorithm without step 1 in order to find $\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q^*, \mathbf{z})$ as close to the reconstruction error of the untruncated SVD basis as possible, for small q and whilst retaining emulatability by setting $\mathbf{v} = (0.35, 0.1, 0.1)$ (as 3 rotated vectors is enough). Given this rotation, we then set

$$\mathbf{\Sigma}_{\boldsymbol{\eta}} = \frac{\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q^*, \mathbf{z})}{b}\mathbf{W}, \quad b = \chi^2_{\ell,j} \tag{14}$$

where $j < 0.995$ is a tuning parameter. This ensures that when reconstructed with the new basis, the observations will not be ruled out, and hence we can identify an NROY space likely to contain runs as consistent with $\mathbf{z}$ as possible, given the limited information we have with 62 ensemble members. This has the effect of 'scaling' the reconstruction error for the rotated basis seen in Figure 8 below the horizontal dotted line at the point the basis is truncated. For our fields, we set $j = 0.95$ for RTMT, and $j = 0.68$ for the others (as $j = 0.95$ ruled out all of Θ for CLTO and TA).

We define NROY space as runs where $\boldsymbol{\theta}$ is not ruled out using (3) for each of TA, CLTO and RTMT. This NROY space consists of 0.9% of Θ. We then design and run a new 50-member ensemble within this NROY space (discussed in Salter (2017), Section 6.3.5).

Upon inspection of the TA field for this wave 2 ensemble, we observe that every run contains the previously found strong warm bias in the Southern Hemisphere (Figure S9). As our optimal basis choice permitted the search for runs not containing this structural bias, these results are evidence that this may be a structural error. In practice, how much

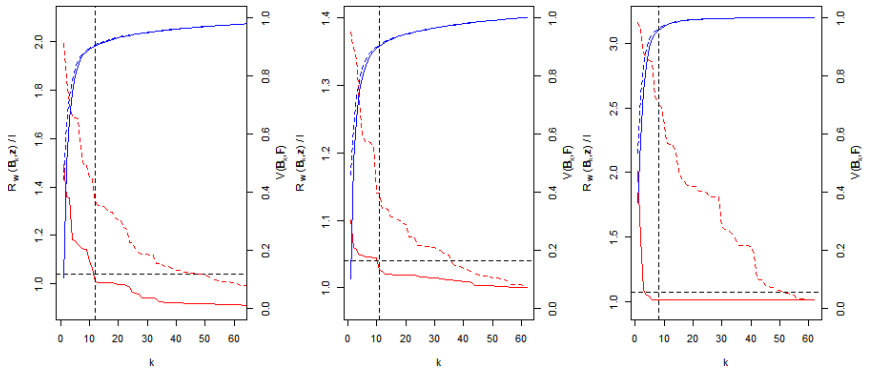

Figure 8: VarMSEplots for CLTO, RTMT and TA, with $\mathbf{W} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, with rotated basis (solid lines) and SVD basis (dotted lines). The dotted horizontal line indicates T.

evidence is required before the modellers are convinced that a particular bias is structured or not is a climate modelling decision. Certainly, we could repeat our wave 1 procedure within the current NROY space and run a wave 3 and so on. This has the benefit of increasing the density of points in Θ and the accuracy of emulators in key regions of Θ, thus insuring against possible 'spikes' in the model input space that would correct the bias.

Assuming our modellers were convinced to treat this feature as a structural bias, we demonstrate an approach to include this information within the iterative calibration procedure. We first revisit the specification of the TA discrepancy, selecting the region with the common warm bias shown in Figure S10, deemed to be a structural error, and increasing $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ for the grid boxes in this region. To do this, we set $\mathbf{W}$ as a diagonal matrix with 100 for the grid boxes in this region, and 1s elsewhere on the diagonal (note this is one possible choice. We might, instead, increase the correlation between these gridboxes in $\mathbf{W}$ in addition).

We re-define the wave 1 NROY space so that it only depends on CLTO and RTMT (consisting of 41.4% of Θ), and then include NROY wave 1 runs with the wave 2 ensemble when rotating and building emulators for wave 2. For TA, the optimal rotation algorithm

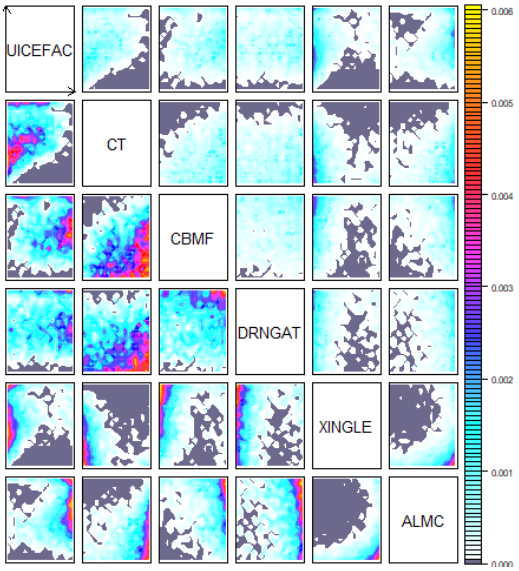

Figure 9: Plot showing the composition of the wave 2 NROY space for 6 of the parameters of CanAM4. For each cell of a pairwise plot, a large sample is taken over the remaining parameters and the proportion of space that is not ruled out is calculated. The lower left gives the same plots, with scales set for each individual plot to show more structure.

is applied using the newly-designed $\mathbf{W}$, with the discrepancy $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ defined via (14), to ensure that $\mathbf{z}$ is not ruled out ($\mathbf{W}$ reflected our beliefs about the structure, not the magnitude, of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$). History matching using the wave 2 bases and emulators leads to an NROY space containing 0.03% of Θ. Plots illustrating this NROY space for six of the more active parameters are shown in Figure 9. We see that the regions with the greatest density of points in NROY space are generally found towards the edges of the parameter ranges. From the lower left plots, it is easier to identify relationships between some of the parameters, e.g. CBMF generally needs to be high while UICEFAC needs to be towards the centre of its allowable range.

The calibration of climate models, or even simply the search for structural biases, is a massive undertaking, and a full tuning is well beyond the scope of this paper. Each

small ensemble of CanAM4 required 2 days of super-computer time to run and, in reality, the modellers routinely check over 20 spatial fields (and a host of other metrics) when tuning the model. UQ can help with the tuning procedure in providing tools (emulators) that allow Θ to be explored much more quickly than is currently possible at the modelling centres. However, as this application has demonstrated, using the off-the-shelf methods based on the SVD basis does not work for tuning in general. It did not work for the 3 fields we showed here, nor have the authors ever found climate model output for which the problems we identified here were not present. Our application demonstrates the optimal rotation algorithm as an effective solution to quickly find bases without these issues for calibration.

6 Discussion

In this paper, we highlighted the issue of terminal case analyses for the calibration of computer models. A terminal case analysis occurs when the prior assessment of model discrepancy is inconsistent with the computer simulator's ability to mimic reality, and leads either to useless and incorrect posterior distributions (using the fully Bayesian procedure) or ruling out all of parameter space (using history matching). We showed that even when the prior assessment of model discrepancy is not inconsistent with the ability of the simulator, dimension reduction of spatial output using the ensemble-derived principal components (the SVD basis) often leads to a terminal case analysis.

We proposed a rotation of the SVD basis to highlight and incorporate important low-signal patterns that may be contained in the original SVD basis, giving a new basis that avoids the terminal case when this is possible. We then presented an efficient algorithm for optimal rotation, guaranteeing to avoid the terminal case when the model discrepancy allows, whilst ensuring enough signal on leading basis vectors to permit the fitting of

emulators. We proved that our algorithm gives a valid rotation of the original basis, and established a fast test to see whether a given ensemble of model runs and discrepancy specification automatically leads to a terminal case analysis prior to rotation. Our methods are presented for models with spatial output, however, if basis methods were to be used for more general high dimensional output (e.g. spatio-temporal), the optimal rotation approach would not change if, for example, PCA were taken over the entire spatio-temporal output for the design, as in Higdon et al. (2008).

We demonstrated the efficacy of our method using an idealised application, and showed that it scaled up to the important case of spatial output for state-of-the-art climate models. Our application highlighted the issue of the terminal case for climate model analyses, and showed the problems with using SVD in practice. We applied history matching for 2 waves to CanAM4 and showed how, combined with optimal rotation, we can begin to distinguish between what the modellers term 'structural errors' and 'tuning errors'.

A purely methodological UQ approach for tuning climate models does not exist. It may be tempting, for UQ practitioners who are not familiar with climate models, to claim that calibration of computer simulators is a 'solved' problem and that 'all' that is required is for the modellers to specify their model discrepancy. We believe that the challenge for model tuning lies in the understanding of this elusive quantity. For the statistical community, rather than focussing on developing comprehensive methods for calibrating climate models automatically, this should mean engaging with modellers to develop robust tools and methods to help identify and understand these errors. This type of approach would have obvious implications for tuning, but would also feed into model development as it becomes better understood which parameters control various biases, and therefore which parameterisations need particular attention during the next development cycle.

References

Aarts, E. H. and P. J. Van Laarhoven (1985). Statistical cooling: A general approach to combinatorial optimization problems. *Philips J. Res. 40*(4), 193–226.

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

Brynjarsdóttir, J. and A. O'Hagan (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems 30*(11), 114007.

Chang, W., P. J. Applegate, M. Haran, and K. Keller (2014). Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties. *Geoscientific Model Development Discussions 7*(2), 1905–1931.

Chang, W., M. Haran, P. Applegate, and D. Pollard (2016). Calibrating an ice sheet model using high-dimensional binary spatial data. *Journal of the American Statistical Association 111*(513), 57–72.

Chen, H., J. L. Loeppky, J. Sacks, W. J. Welch, et al. (2016). Analysis Methods for Computer Experiments: How to Assess and What Counts? *Statistical Science 31*(1), 40–60.

Collins, M., S.-I. An, W. Cai, A. Ganachaud, E. Guilyardi, F.-F. Jin, M. Jochum, M. Lengaigne, S. Power, A. Timmermann, et al. (2010). The impact of global warming on the tropical Pacific Ocean and El Niño. *Nature Geoscience 3*(6), 391–397.

Conti, S. and A. O'Hagan (2010). Bayesian emulation of complex multi-output and dynamic computer models. *Journal of statistical planning and inference 140*(3), 640–651.

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

Goldstein, M. and J. Rougier (2009). Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference 139*(3), 1221–1239.

Gu, M., J. O. Berger, et al. (2016). Parallel partial Gaussian process emulation for computer models with massive output. *The Annals of Applied Statistics 10*(3), 1317–1347.

Harris, G., D. Sexton, B. Booth, M. Collins, J. Murphy, and M. Webb (2006). Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. *Climate Dynamics 27*(4), 357–375.

Haylock, R. and A. O'Hagan (1996). On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. *Bayesian statistics 5*, 629–637.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics 5*(2), 173–190.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association 103*(482).

Hourdin, F., T. Mauritsen, A. Gettelman, J.-C. Golaz, V. Balaji, Q. Duan, D. Folini, D. Ji, D. Klocke, Y. Qian, et al. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society 98*(3), 589–602.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Kaufman, C. G., D. Bingham, S. Habib, K. Heitmann, and J. A. Frieman (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *The Annals of Applied Statistics 5*(4), 2470–2492.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464.

Mauritsen, T., B. Stevens, E. Roeckner, T. Crueger, M. Esch, M. Giorgetta, H. Haak, J. Jungclaus, D. Klocke, D. Matei, et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems 4*(3).

Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer (2000). The coupled model intercomparison project (CMIP). *Bulletin of the American Meteorological Society 81*(2), 313–318.

Murnaghan, F. D. (1962). *The unitary and rotation groups*, Volume 3. Spartan books.

Pollard, D., W. Chang, M. Haran, P. Applegate, and R. DeConto (2016). Large ensemble modeling of the last deglacial retreat of the West Antarctic Ice Sheet: comparison of simple and advanced statistical techniques. *Geoscientific Model Development 9*(5), 1697–1723.

Pukelsheim, F. (1994). The three sigma rule. *The American Statistician 48*(2), 88–91.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statistical science*, 409–423.

Salter, J. M. (2017). *Uncertainty quantification for spatial field data using expensive computer models: refocussed Bayesian calibration with optimal projection.* Ph. D. thesis, University of Exeter.

Salter, J. M. and D. Williamson (2016). A comparison of statistical emulation methodologies for multi-wave calibration of environmental models. *Environmetrics 27*(8), 507–523.

Scaife, A. A., T. Spangehl, D. R. Fereday, U. Cubasch, U. Langematz, H. Akiyoshi, S. Bekki, P. Braesicke, N. Butchart, M. P. Chipperfield, et al. (2012). Climate change projections and stratosphere–troposphere interaction. *Climate Dynamics 38*(9-10), 2089–2097.

Screen, J. A. and D. Williamson (2017). Ice-free Arctic at 1.5 [deg] C? *Nature Climate Change 7*(4), 230–231.

Sexton, D. M., J. M. Murphy, M. Collins, and M. J. Webb (2011). Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *Climate dynamics 38*(11-12), 2513–2542.

Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, B. Bex, and B. Midgley (2013). IPCC, 2013: climate change 2013: the physical science basis. contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change.

Vernon, I., M. Goldstein, and R. G. Bower (2010). Galaxy formation: a Bayesian uncertainty analysis. *Bayesian Analysis 5*(4), 619–669.

von Salzen, K., J. F. Scinocca, N. A. McFarlane, J. Li, J. N. Cole, D. Plummer, D. Verseghy, M. C. Reader, X. Ma, M. Lazare, et al. (2013). The Canadian fourth generation atmospheric global climate model (CanAM4). Part I: representation of physical processes. *Atmosphere-Ocean 51*(1), 104–125.

Wilkinson, R. D. (2010). Bayesian calibration of expensive multivariate computer experiments. *Large-Scale Inverse Problems and Quantification of Uncertainty, Ser. Comput. Stat., edited by LT Biegler et al*, 195–216.

Williamson, D. (2015). Exploratory ensemble designs for environmental models using k-extended Latin Hypercubes. *Environmetrics 26*(4), 268–283.

Williamson, D., A. T. Blaker, C. Hampton, and J. Salter (2015). Identifying and removing structural biases in climate models with history matching. *Climate Dynamics 45*(5-6), 1299–1324.

Williamson, D., M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate dynamics 41*(7-8), 1703–1729.

Williamson, D., M. Goldstein, and A. Blaker (2012). Fast linked analyses for scenario-based hierarchies. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(5), 665–691.

Williamson, D. B., A. T. Blaker, and B. Sinha (2017). Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development 10*(4), 1789.

Yang Xiang, S. Gubian, B. Suomela, and J. Hoeng (2013). Generalized Simulated Annealing for Efficient Global Optimization: the GenSA Package for R. *The R Journal Volume 5/1, June 2013*.