International Journal for Uncertainty Quantification, x(x): 1–34 (2021)

EFFICIENT CALIBRATION FOR HIGH-DIMENSIONAL COMPUTER MODEL OUTPUT USING BASIS METHODS

James M. Salter[1,*] & Daniel B. Williamson[1,2]

[1]*Department of Mathematics, University of Exeter, Exeter, UK*

[2]*Alan Turing Institute, London, UK*

*Address all correspondence to: James M. Salter, Department of Mathematics, University of Exeter, Exeter, UK, E-mail: j.m.salter@exeter.ac.uk

Original Manuscript Submitted: 5th July 2021; Final Draft Received:

Calibration of expensive computer models using emulators for high-dimensional output fields can become increasingly intractable with the size of the field(s) being compared to observational data. In these settings, dimension reduction is attractive, reducing the number of emulators required to mimic the field(s) by orders of magnitude. By comparing to popular independent emulation approaches that fit univariate emulators to each grid cell in the output field, we demonstrate that using a basis structure for emulation, aside from the clear computational benefits, is essential for obtaining coherent draws that can be compared with data or used in prediction. We show that calibrating on the subspace spanned by the basis is not generally equivalent to calibrating on the full field (the latter being generally infeasible owing to the large number of matrix inversions required for calibration and the size of the matrices on the full field). We then present a projection that allows accurate calibration on the field for exactly the cost of calibrating in the subspace, by projecting in the norm induced by our uncertainties in observations and model discrepancy and given a one-off inversion of a large matrix. We illustrate the benefits of our approach and compare with standard univariate approaches for emulating and calibrating the high dimensional ice sheet model Glimmer.

KEY WORDS: *Uncertainty quantification; Dimension reduction; History matching; Emulation; Basis rotation*

2152–5080/17/$35.00

1. INTRODUCTION

A computer model, $f(\cdot)$, is a representation of a real-world process, given by a set of equations and parametrisations, that takes a vector of inputs $\mathbf{x}$, and returns an output. This output may be a single value, a spatial field, a time series, or a combination of these across multiple different fields (e.g. climate models [1]). Computer models often represent complex processes, and may require long running times on expensive supercomputers. It is therefore only possible to evaluate the model at a small sample of values from the input space, hence statistical models ('emulators') are often used as a proxy, giving predictions and uncertainty for the output at values of $\mathbf{x}$ for which the outputs are unseen [2,3].

Using an emulator of $f(\cdot)$, observations of the real-world process can be used to calibrate the inputs, $\mathbf{x}$, of the computer model. This can be done either probabilistically, with a distribution given for the best setting of the input parameters ('Bayesian calibration', [4]), or via history matching [5–7]. History matching (HM) uses a Bayes Linear approach, with only expectations and variances required, and instead of returning a distribution, rules out regions of the input parameter space that are inconsistent with the observations, based on an implausibility measure and a threshold for removing runs. Alternative approaches have been developed with the aim of overcoming identifiability problems between model parameters and model discrepancy, in both a frequentist (as in [8,9]) and Bayesian (e.g., [10,11] setting. In this article, our focus is on HM, although there are parallels with probabilistic calibration as in [4], with the HM implausibility similar to the negative log likelihood.

High-dimensional computer model output has several different forms, requiring different approaches in order to emulate the output. For example, time series output often lends itself to an autoregressive approach [12,13], whilst spatial fields are often projected onto a low-dimensional basis given by the principal components of the output [3, 14], or some other optimally-selected basis for calibration [15]. In these cases, emulators are then fitted for the coefficients in the reduced space. This reduced basis approach may be used for temporal (e.g. [16]) or spatio-temporal (e.g. [3]) output with few or no adjustments required. The low-dimensional basis method is attractive because it reduces the computation required, particularly when the field dimension is very large, whilst retaining the output correlation structure through the basis vectors. Such an emulation approach has been used in a number of different fields, including a multitude of climate model applications [15,17,18], ice sheet modelling [19], electro-physiology [20], and experiments imploding steel cylinders [3].

Alternatively, every grid box or time point can be emulated individually [21–24]. As the number of emulators to be built scales with the size of the output, Gu & Berger set common regressors for the mean function, and fix the correlation parameters, across all grid boxes [23]. Validating emulators for thousands of grid boxes may be a challenge, and only an automated approach to this is generally feasible.

Given a set of emulators for the model output, how best to overcome the problem of high-dimensionality when calibrating these fields to data is not clear, owing to the requirement of inverting an $(\ell \times \ell)$-dimensional matrix, for ℓ the dimension of the observed field. One approach is to use emulated coefficients to reconstruct the original field, and compare this to the observations themselves [3,17,20], although for high ℓ and non-diagonal variance matrices, this becomes increasingly intractable. This is also an issue with the 'emulate every output' approach, where a summary of the full output is often used for calibration for the same computational reasons [24]. Instead, all quantities defined over the field can be projected onto a low-dimensional basis, with the representation of the observations compared to emulated output on the basis [14,18,19], with fast calculations in this reduced subspace.

In this paper, we compare 'every grid box' and basis methods for emulation and their subsequent use in calibration, demonstrating computational savings in terms of evaluating predictions and history matching, and explore other desirable properties of the basis emulators such as the physical coherency of posterior draws, which can be beneficial even when ℓ is small enough for other approaches to be tractable. Given emulators with a basis structure, we provide an efficient way to calculate the implausibility over the original field, so that history matching high-dimensional fields is tractable, only requiring repeated evaluations of the inexpensive subspace implausibility, without losing any information from the field implausibility, and allowing all quantities to be specified on the physically-interpretable field level. We demonstrate the importance of projection in the 'correct' norm, with the best input and general ordering according to the distance metric sensitive to the projection method. Due to the relationship between the implausibility and the likelihood, this result also has implications for probabilistic calibration if directly performed in a subspace.

Overall, the main contributions here are methodology of how to generally perform history matching for large observational fields (where past studies only provided a framework, but not an approach that would work for general, non-diagonal variance matrices), and a method for deriving a suitable threshold for ruling out implausible regions of parameter space in such cases. Additionally, the existing literature tends to pick either an 'emulate all' or 'basis decomposition' approach at the start of an emulation study. We compare the two options here, and offer advice for practitioners in the field as to the benefits and drawbacks of each.

Section 2 outlines emulation and history matching for high-dimensional fields. Section 3 compares general properties of basis and univariate emulators. Section 4 considers history matching in the projected space, and provides a fast method for calculating the full implausibility given a basis structure, and guidance on setting a pragmatic bound. Section 5 emulates and history matches an output of the Glimmer ice sheet model, with discussion in Section 6.

2. SPATIO-TEMPORAL HISTORY MATCHING

2.1 Emulation

Emulators are used in place of the computer model when it is costly or time-consuming to run, with Gaussian processes a popular choice [2,25,26]. Emulation depends on running the true model, $f(\cdot)$, at n settings $\mathbf{x} \in \mathcal{X}$, giving ensemble $\mathbf{F} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) \in \mathbb{R}^{\ell \times n}$, with $f(\mathbf{x}_i) \in \mathbb{R}^{\ell}$.

2.1.1 Univariate emulators

In this setting, each of the ℓ outputs of $f(\cdot)$, denoted by subscript i, is emulated as a Gaussian process, with:

$$f_i(\mathbf{x}) \sim \text{GP}(m_i(\mathbf{x}), R_i(\mathbf{x}, \mathbf{x}')), \quad i = 1, \ldots, \ell, \tag{1}$$

for mean function $m_i(\cdot)$ and covariance $R_i(\cdot, \cdot)$. These functions may be fitted individually for all ℓ outputs, allowing different terms in the mean function, and different correlation lengths [22,24,27], or, for computational convenience, a fixed set of regressors may be imposed across all ℓ outputs, with a single set of correlation lengths estimated [23]. The former approach offers greater flexibility, although is more time consuming.

The number of emulated outputs in the above applications is varied. In [27], the output is global cloud condensation nuclei, with a spatial grid of 8192 cells repeated for 12 months, requiring a total of $\ell = 98304$ emulators. [22] and [23] both emulate aspects of the spatial output of a simulator of volcanic pyroclastic flows, with the former using '$10^2 - 10^4$' emulators, and the latter emulating $\ell = 23040$ coordinates.

2.1.2 Basis emulation

For validation and computational purposes, low-dimensional representations of the output are commonly used, requiring significantly fewer emulators than a univariate approach. The high-dimensional data is projected onto a basis, often given by the principal components across the model runs (the Singular Value Decomposition (SVD) basis), and the coefficients on this basis are emulated [3,17–20].

To find the principal component (PC) basis, the ensemble mean, $\boldsymbol{\mu}$, given by averaging across the rows of $\mathbf{F}$, is subtracted from each column of $\mathbf{F}$, to give the centred ensemble, $\mathbf{F}_{\boldsymbol{\mu}}$. The basis, $\boldsymbol{\Gamma}$, is found via singular value decomposition as (and throughout $\boldsymbol{\Gamma}$ refers to this SVD basis):

$$\mathbf{F}_{\boldsymbol{\mu}}^T = \mathbf{U}\mathbf{D}\boldsymbol{\Gamma}^T. \tag{2}$$

The basis is truncated after the first q vectors $\boldsymbol{\gamma}_i \in \mathbb{R}^\ell$, such that this set is sufficient to explain a high percentage of the variability in $\mathbf{F}_{\boldsymbol{\mu}}$ (commonly, 90% or 95%, but problem dependent), resulting in basis $\boldsymbol{\Gamma}_q = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q) \in \mathbb{R}^{\ell \times q}$. Projection of an output field, $f(\mathbf{x})$, onto basis $\boldsymbol{\Gamma}_q$ is given by:

$$\mathbf{c}(\mathbf{x}_i) = (\boldsymbol{\Gamma}_q^T \mathbf{W}^{-1} \boldsymbol{\Gamma}_q)^{-1} \boldsymbol{\Gamma}_q^T \mathbf{W}^{-1} (f(\mathbf{x}_i) - \boldsymbol{\mu}), \tag{3}$$

for a positive definite weight matrix $\mathbf{W}$ that defines the norm of the space in which we perform the projection, with $\|\mathbf{v}\|_{\mathbf{W}} = \mathbf{v}^T \mathbf{W}^{-1} \mathbf{v}$ the norm of vector $\mathbf{v}$. For applications with the SVD/PC basis, this may be $\mathbf{W} = \mathbb{I}_\ell$ (L_2 projection), or based on the observation error and discrepancy variances (see Section 2.2) as in [3] and [15]. A set of coefficients is mapped back to the original field for prediction or calibration purposes via:

$$f(\mathbf{x}_i) = \boldsymbol{\mu} + \boldsymbol{\Gamma}_q \mathbf{c}(\mathbf{x}_i) + \boldsymbol{\epsilon}, \tag{4}$$

for error vector $\boldsymbol{\epsilon}$. If $q = n$, then $\boldsymbol{\epsilon} = \mathbf{0}$ for $\mathbf{x}_i \in \mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Emulators are built for the coefficients on the first q basis vectors,

$$c_i(\mathbf{x}) \sim \text{GP}(m_i(\mathbf{x}), R_i(\mathbf{x}, \mathbf{x}')), \quad i = 1, \ldots, q, \tag{5}$$

with $\text{E}[\mathbf{c}(\mathbf{x})] = (\text{E}[c_1(\mathbf{x})], \ldots, \text{E}[c_q(\mathbf{x})])^T$, the emulator expectation for each of the q basis vectors, and $\text{Var}[\mathbf{c}(\mathbf{x})] = diag(\text{Var}[c_1(\mathbf{x})], \ldots, \text{Var}[c_q(\mathbf{x})])$ the associated $q \times q$ variance matrix. We retrieve the ℓ-dimensional expectation and variance of $f(\mathbf{x})$ via:

$$\text{E}[f(\mathbf{x})] = \boldsymbol{\mu} + \boldsymbol{\Gamma}_q \text{E}[\mathbf{c}(\mathbf{x})], \quad \text{Var}[f(\mathbf{x})] = \boldsymbol{\Gamma}_q \text{Var}[\mathbf{c}(\mathbf{x})] \boldsymbol{\Gamma}_q^T + \text{Var}[\boldsymbol{\epsilon}], \tag{6}$$

where $\text{Var}[\boldsymbol{\epsilon}]$ contains the variance due to the discarded basis vectors, $\boldsymbol{\Gamma}_{-q}$ [17]. As truncation occurs after the majority of ensemble variability is explained, these removed directions generally have $\text{Var}[\boldsymbol{\epsilon}] << \boldsymbol{\Gamma}_q \text{Var}[\mathbf{c}(\mathbf{x})] \boldsymbol{\Gamma}_q^T$ and are often ignored.

An alternative to the common approach above, the 'Generalized probabilistic PCA (GPPCA)', is given in [28], for settings where the data is correlated, allowing the vectors to be correlated, which may be a more appropriate assumption in some applications. Instead of simply calculating the PCA basis across the data, the GPPCA basis is given by the maximum marginal likelihood estimate, after the Gaussian process prior on the basis coefficients $c_i(\mathbf{x})$ has been marginalised out.

2.1.3 Basis selection

The space of possible reconstructions, $\mathbf{\Gamma}_q\mathbf{c}(\cdot)$, is a q-dimensional surface in ℓ-dimensional space, restricted by the basis. We therefore need to ensure that the chosen basis has certain desirable properties: firstly, that $q << \ell$; secondly, that emulation is possible when the model output is projected onto the basis, i.e. that the coefficients are not dominated by noise and vary smoothly in the inputs, (so we cannot just use arbitrary patterns); and finally that the observations of the real field that we wish to calibrate to, $\mathbf{z} \in \mathbb{R}^\ell$, lie within its span (up to some error), so that we avoid guaranteeing that we rule out their basis representation. The truncated SVD basis satisfies the first two by its variance-maximising property, with the q leading basis vectors containing signal from the model output $\mathbf{F}$ that usually allows emulation of the coefficients. However, it does not use knowledge about $\mathbf{z}$, and as our setting is often one with small n, high ℓ, and a large, mostly unexplored parameter space $\mathcal{X}$, we won't generally by chance have output that is similar to observations.

To assess the quality of a basis chosen for calibration, for a given $\ell \times \ell$ positive definite matrix $\mathbf{W}$, [15] define the 'reconstruction error' of a basis $\mathbf{\Gamma}_q$ as:

$$\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z}) = \|\mathbf{z} - \mathbf{\Gamma}_q(\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{\Gamma}_q)^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}\|_{\mathbf{W}}. \tag{7}$$

This quantity represents the difference between $\mathbf{z}$ and its reconstruction on the calibration basis $\mathbf{\Gamma}_q$ (i.e. we project $\mathbf{z}$ onto $\mathbf{\Gamma}_q$ and then map back to the original ℓ-dimensional space), with respect to the weighting matrix $\mathbf{W}$ for $\|\mathbf{a}\|_{\mathbf{W}} = \mathbf{a}^T\mathbf{W}^{-1}\mathbf{a}$. This metric essentially describes how well the basis $\mathbf{\Gamma}_q$ represents $\mathbf{z}$, and is analogous to the history matching implausibility defined in the next section.

2.2 History matching

Given an emulator(s) for a computer model, it is often of interest to calibrate the input parameters (that have unknown values, and are not observed in field experiments) using observations of the real-world system represented by the model. History matching rules out settings of the input parameters, $\mathbf{x} \in \mathcal{X}$, that lead to computer model output, $f(\mathbf{x})$ (a vector of length ℓ), that are not consistent with observations, $\mathbf{z}$, given an error specification [5,7,29,30]. History matching uses a statistical model that links the true value of the system, $\mathbf{y}$, with the computer model, generally given by [4]:

$$\mathbf{z} = f(\mathbf{x}^*) \oplus \boldsymbol{\eta} \oplus \mathbf{e}, \tag{8}$$

where $\boldsymbol{\eta}$ (the discrepancy between the output given at the 'best' setting, $\mathbf{x}^*$, of $f(\cdot)$, and reality, $\mathbf{y}$) and $\mathbf{e}$ (the observation error) are uncorrelated (indicated by $\oplus$) mean-zero terms, with positive definite variance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}$ respectively, i.e. given suitable error matrices, the observations can be represented as the sum of a deterministic function representing reality, and a noise vector. Rather than requiring full distributions on $\boldsymbol{\eta}$ and $\mathbf{e}$, history matching only uses expectations and variances. When $f(\cdot)$ is expensive to run, it is replaced by an emulator (Section 2.1).

The implausibility, $\mathcal{I}(\mathbf{x})$, for a parameter setting $\mathbf{x}$ is defined as the Mahalanobis distance between $\mathbf{z}$ and the predictive expectation from an emulator for the computer model:

$$\mathcal{I}(\mathbf{x}) = (\mathbf{z} - \mathrm{E}[f(\mathbf{x})])^T (\mathrm{Var}(\mathbf{z} - \mathrm{E}[f(\mathbf{x})]))^{-1} (\mathbf{z} - \mathrm{E}[f(\mathbf{x})]), \tag{9}$$

where, under the model assumptions in (8), we have $\ell \times \ell$ variance matrix:

$$\mathrm{Var}(\mathbf{z} - \mathrm{E}[f(\mathbf{x})]) = \mathrm{Var}[f(\mathbf{x})] + \boldsymbol{\Sigma}_{\mathbf{e}} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}. \tag{10}$$

Large values of this distance indicate that it is implausible that $\mathbf{x} = \mathbf{x}^*$. Using $\mathcal{I}(\mathbf{x})$ and an emulator for $f(\mathbf{x})$ that gives some $\mathrm{E}[f(\mathbf{x})]$, $\mathrm{Var}[f(\mathbf{x})]$, 'Not Ruled Out Yet' (NROY) space contains all not implausible $\mathbf{x}$, defined as [31,32]:

$$\mathcal{X}_{NROY} = \{\mathbf{x} \in \mathcal{X} | \mathcal{I}(\mathbf{x}) < T\}, \tag{11}$$

for bound T. If $\mathbf{z} - \mathrm{E}[f(\mathbf{x})]$ is Normal, then $\mathcal{I}(\mathbf{x}) \sim \chi^2_\ell$, for ℓ the rank of (10), and bound T may be set using a quantile of this distribution (e.g. $T = \chi^2_{\ell,0.995}$, so that $P(\mathcal{I}(\mathbf{x}) < T) = 0.995$).

Given an NROY space, there are a number of possible next steps. History matching is often an iterative procedure, with new batches of expensive model runs performed for a new design within the current NROY space, emulators refined, and a new NROY defined, with the aim of zooming in on the appropriate region of parameter space [32–34]. Multiple iterations are often carried out, and performing Bayesian calibration within the resulting space may give more accurate results as the emulators should be improved (in terms of expectation and reduced variance) in the regions of space that are most consistent with the observations [26]. Additional metrics to match to can be introduced at later stages, e.g. we could initially identify the region of $\mathcal{X}$ that is consistent with some metric, then given this search for those consistent with other metrics. In [35], a model for temperature is history matched, with the resulting NROY space sampled from to provide a boundary condition for an ice sheet model, exploring both model parameter and boundary condition uncertainty.

By setting $\mathbf{W} = \boldsymbol{\Sigma}_\mathbf{e} + \boldsymbol{\Sigma}_\boldsymbol{\eta}$ in (7), $\mathcal{R}_\mathbf{W}(\boldsymbol{\Gamma}_q, \mathbf{z})$ is equivalent to $\mathcal{I}$ (equation (9)) if the emulator variance $\text{Var}[f(\mathbf{x})] = \mathbf{0}$, and hence we have that if $\mathcal{R}_\mathbf{W}(\boldsymbol{\Gamma}_q, \mathbf{z}) > T$, then the representation of $\mathbf{z}$ on the basis would be ruled out, regardless of whether the computer model can represent $\mathbf{z}$ (termed the 'terminal case'). [15] alleviates this problem by rotating the full basis, combining important (in terms of explaining $\mathbf{z}$) low order patterns with the leading vectors (to allow the construction of informative emulators), though any method that satisfies the described criteria can be used to find a suitable low-dimensional representation for emulation and calibration. The selection used by [15] has the following steps:

1. Calculate the SVD basis, $\boldsymbol{\Gamma}$, and truncate after q vectors such that 'enough' ensemble variability, given by some proportion v, is explained by $\boldsymbol{\Gamma}_q$.
2. If $\mathcal{R}_\mathbf{W}(\boldsymbol{\Gamma}_q, \mathbf{z}) < T$, then the truncated basis $\boldsymbol{\Gamma}_q$ is suitable.
3. If $\mathcal{R}_\mathbf{W}(\boldsymbol{\Gamma}, \mathbf{z}) \geq T$, then there is not enough information in the ensemble to adequately represent $\mathbf{z}$ under error specification $\mathbf{W}$, and hence more model runs are required, or a change in our tolerance to error.
4. If $\mathcal{R}_\mathbf{W}(\boldsymbol{\Gamma}_q, \mathbf{z}) > T$ and $\mathcal{R}_\mathbf{W}(\boldsymbol{\Gamma}, \mathbf{z}) \leq T$, find rotation matrix $\boldsymbol{\Lambda}^* \in \mathbb{R}^{n \times n}$ where:

$$\boldsymbol{\Lambda}^* = \arg\min_{\boldsymbol{\Lambda}} \mathcal{R}_\mathbf{W}((\boldsymbol{\Gamma}\boldsymbol{\Lambda})_{q*}, \mathbf{z})$$

   such that $(\boldsymbol{\Gamma}\boldsymbol{\Lambda})_{q*}$ explains at least proportion v of ensemble variability, and subject to ensuring leading vectors explain above some threshold (to aid emulation).
5. Replace $\boldsymbol{\Gamma}_q$ with $(\boldsymbol{\Gamma}\boldsymbol{\Lambda}^*)_{q*}$ in subsequent emulation and calibration.

In practice, the optimisation in step 4 can be performed iteratively to reduce the complexity of the problem.

3. PROPERTIES OF BASIS EMULATION

In this section, we compare the 'emulate every grid box' approach to a general basis emulation method, in terms of the theoretical improvement in efficiency and sampling, before applying both methods to a real application of emulating ice sheet thickness in Section 5. For the theoretical comparison, we assume that a suitable calibration basis $\boldsymbol{\Gamma}_q$ has been found, e.g. one that satisfies the requirements highlighted in Section 2.1.3, which may not always be a straightforward task. The focus of [15] was how to select such a basis, hence we refer readers to that paper for guidance on how one might approach this. The advantages described here are inherent in any basis coefficient emulation approach, so are present in e.g. [3,19,20] and others referenced earlier, even if not explicitly addressed in each article.

3.1 Efficiency

The computational advantages of having a basis structure in emulation manifest in several ways, including in emulator fitting, prediction and validation (and finally in calibration exercises, see Section 4).

A clear benefit is in requiring only q emulators instead of ℓ, even in the age of faster GPUs and greater parallelisation (any increase in computational power applies to both methods). The saving in time fitting emulators often allows a less automatic approach to be used, with an expert able to use their knowledge to fit better individual emulators. Additionally, it may be possible to more rigorously validate emulators when there are only q to consider. An automatic approach to validation, say with flags for when cross-validation or prediction from a validation set results in a large number of the true values lying outside a prediction interval, will catch a lot of the more obvious errors, but it can be difficult to diagnose all potential errors in such a way [36].

For example, consider the validation plot in the left of Figure 1, which shows predictions made at 100 validation points, for a particular emulated output (either one of thousands, or for a basis coefficient). Note that the form of the data and emulator are not important here, and this is only meant to be illustrative of a potential method of failure, and that there are many ways that one could find such a situation. Here, 5 of the 100 true values lie outside the 95% emulator predictions (shown by the red points), and so we might say that this validates well based on this statistic, however there is clearly an issue with underestimating the truth when x is low. It is feasible to check q plots (e.g. for cross-validation, prediction on a validation set, predictions against individual parameters) by hand, and hence it is much easier to diagnose problems such as non-stationarity (e.g. [37]) and to apply more time and effort to emulate such outputs more accurately.

A final benefit in terms of efficiency is when making predictions. We may wish to sample the input space and propagate uncertainty through the emulator (uncertainty analysis) for prediction of the field for its own sake or as a driving variable to another model for a large number of $\mathbf{x} \in \mathcal{X}$. This is substantially faster for q emulators compared to ℓ. When performing a multi-wave history match, a similarly large number of emulator evaluations are required and such computational savings become even greater as the number of waves increases. Similarly, if we wish to consider multiple high-dimensional fields (e.g. temperature, precipitation, sea level pressure etc. in a climate model), emulating each with a basis representation provides further savings.

Following the method of Gu & Berger, with a common set of terms in the mean function, the expectation can be calculated extremely efficiently for any choice of $\mathbf{x}$ [23]. However, the variance is more expensive (scales with ℓ) and is required at each $\mathbf{x}$ for calibration and uncertainty analysis, hence if we wish to perform such tasks there is still a large computational cost associated with finding the variance for the ℓ emulators, before the frequent $\ell \times \ell$ inversions.

3.2 Physical coherence of samples

When derived from a set of model output, the PC/SVD basis will encode correlations and smoothness from the output, and the patterns extracted may have physical meaning. Predictions from the basis emulator will then generally look like plausible, physically-coherent model output, with samples from the emulator posterior retaining this property. Although the expectation of the independent emulators may look extremely similar to that of the basis emulator (and in fact adding spatial dependence between outputs to the independent emulators does not affect the mean given the formulation in [23] (see Theorem 6.1)), any dependence across outputs has been ignored so that if we draw a realisation of the field at $\mathbf{x}$ from the emulator posterior, the resulting field may not be smooth, and is not a realistic representation of the model (unless the emulator variance is small compared to the magnitude of variability in the output). For example, if we are performing uncertainty analysis on a chain of models, where the output of one model (or an emulator of this output) is used as the input to the next, it is important to be able to run subsequent models with plausible outputs of the previous models, with samples rather than just the expectation required in order to properly propagate uncertainty through the chain.

More concretely, (6) gives the expectation and variance of $f(\mathbf{x})$ given emulators for the coefficients, and if we instead draw a sample $\mathbf{s}_q$ from the coefficient emulator, and obtain a sampled ℓ-dimensional field by mapping this as in (6):

$$\mathbf{s}_q(\mathbf{x}) \sim \mathrm{N}_q(\mathrm{E}[\mathbf{c}(\mathbf{x})], \mathrm{Var}[\mathbf{c}(\mathbf{x})]), \quad \mathbf{s}_\ell(\mathbf{x}) = \boldsymbol{\mu} + \boldsymbol{\Gamma}_q \mathbf{s}_q, \tag{12}$$

we see that all posterior samples are linear combinations of $\boldsymbol{\Gamma}_q$, encoding correlations and patterns from this whilst accounting for the emulator variance.

As a simple illustration of potential consequences here, consider a Monte Carlo estimate of the risk of exceeding a threshold in multiple locations. For example, an 'inundation event' might involve multiple co-located grid cells exceeding a threshold for a natural hazard model, and insurance companies might define such events in their policies or as part of regulatory compliance. To make the illustration as simple as possible, assume that we have two outputs, $f = (f_1, f_2)$, and represent them in two different ways: i) as independent $\mathrm{N}(0, 1)$ distributions (the every grid box method), and ii) as a bivariate Normal (the correlated emulator method) with covariance 0.9. We incur some loss if both outputs are greater than 1.5 (our 'inundation event'). Figure 1 (right) shows the distribution of $min(f_1, f_2)$ for the UV (blue) and multivariate (orange) approaches. The shaded areas highlight the exceedance probabilities (0.43% and 4.95% respectively - an order of magnitude difference).

This is an extreme example, and if we emulated such correlated outputs independently we would usually have

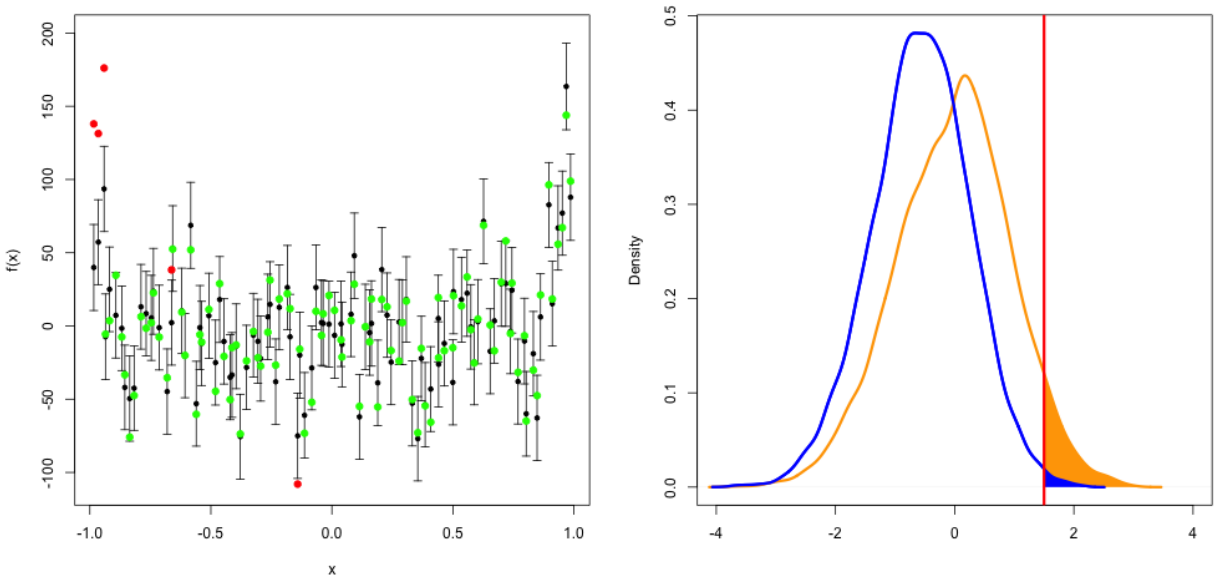

FIG. 1: Left: an example validation plot, with the true points coloured green if they lie within 95% prediction intervals, and red otherwise. Right: The distribution of the minimum across the two grid boxes, for the uncorrelated (blue) and correlated (orange) samples, with the red line at 1.5 indicating the threshold to be crossed.

similarly correlated expectations. However, this is not structurally encoded, and if we are considering interactions between many different outputs, then it is much more difficult to assess whether the correlation matters or not, particularly for uncertainty analysis. We revisit this idea for the ice sheet application in Section 5.2.

3.3 Ensemble assessment

It is trivial to identify the terminal case with a basis (Section 2.1.3), with this problem fixable by either selecting a better basis (rotation or similar), reassessing the observation error and model discrepancy in $\mathbf{W}$, or by obtaining more model runs (seeing better runs or patterns). Theoretically, the univariate approach has full degrees of freedom, and can produce $\mathbf{z}$ perfectly due to the independence of the emulators. In practice, this will not be the case if $\mathbf{z}$ lies outside of the span of the ensemble, with extrapolation likely to be required in multiple locations to find $\mathbf{z}$, hence we can only reassess our tolerance to error via $\boldsymbol{\Sigma}_\mathbf{e}$ and $\boldsymbol{\Sigma}_\boldsymbol{\eta}$ after performing the expensive task of building and sampling from ℓ independent emulators.

4. EFFICIENT HISTORY MATCHING

Given an emulator, a common next step is to search for inputs leading to output consistent with the real world.

We wish to calibrate using all available information, incorporating any knowledge about correlations from $\boldsymbol{\Sigma}_\mathbf{e}$,

$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and the model output into the resulting analysis. As ℓ increases, calculating $\mathcal{I}(\mathbf{x})$ (equation (9)), which does include all information about the ℓ-dimensional field, becomes more expensive, due to the necessary inversion of an $\ell \times \ell$ variance matrix that varies with $\mathbf{x}$ (up to $O(\ell^3)$ complexity). To history match, the implausibility must be evaluated thousands or millions of times, particularly if either several waves are performed, or if the resulting NROY space is small, so that it is difficult to sample from [33,38]. The implausibility is equivalent to the negative log likelihood (without the log determinant), and probabilistic calibration requires repeated evaluations of the likelihood within an MCMC sampler, resulting in the same computational problem.

For large ℓ, as with emulation it is attractive to apply a low-dimensional basis approach to calibration. Given a basis, $\boldsymbol{\Gamma}_q$, and emulators for the coefficients on these q basis vectors, we can history match in the subspace defined by $\boldsymbol{\Gamma}_q$, as has been performed extensively for probabilistic calibration [3,14,18,19]. We define the 'coefficient implausibility', analogous to (9) in the subspace, as:

$$\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}) = (\mathbf{c}(\mathbf{z}) - \mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\mathrm{Var}[\mathbf{c}(\mathbf{x})] + \mathrm{Var}[\mathbf{c}(\mathbf{e})] + \mathrm{Var}[\mathbf{c}(\boldsymbol{\eta})])^{-1}(\mathbf{c}(\mathbf{z}) - \mathrm{E}[\mathbf{c}(\mathbf{x})]), \tag{13}$$

where subscript $\mathbf{W}$ indicates that projection of ℓ-dimensional quantities is performed with respect to positive definite matrix $\mathbf{W}$, i.e. $\mathbf{z}$, $\boldsymbol{\Sigma}_{\mathbf{e}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ are projected onto basis $\boldsymbol{\Gamma}_q$ as follows (see [15] for proof that projection $\mathbf{P}_{\mathbf{W}}$ is optimal):

$$\mathbf{P}_{\mathbf{W}} = (\boldsymbol{\Gamma}_q^T\mathbf{W}^{-1}\boldsymbol{\Gamma}_q)^{-1}\boldsymbol{\Gamma}_q^T\mathbf{W}^{-1}, \quad \mathbf{c}(\mathbf{z}) = \mathbf{P}_{\mathbf{W}}\mathbf{z},$$
$$\mathrm{Var}[\mathbf{c}(\mathbf{e})] = \mathbf{P}_{\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{P}_{\mathbf{W}}^T, \quad \mathrm{Var}[\mathbf{c}(\boldsymbol{\eta})] = \mathbf{P}_{\mathbf{W}}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\mathbf{P}_{\mathbf{W}}^T.$$

The distance metric in (13) only requires inversions of $q \times q$ matrices for $q << \ell$, hence is significantly faster than evaluating $\mathcal{I}(\mathbf{x})$ in general. The covariance matrices are assumed to be fixed, and the resulting implausibility will be sensitive to their choice of structure. The discrepancy $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ is often the harder one to specify, however if an NROY space is empty then this may suggest that it has been misspecified (see [39] for treating discrepancy as tolerance to error).

4.1 Efficiently calculating implausibility

If the emulator for $f(\mathbf{x})$ has a basis structure, and we have equality between the $\mathbf{W}$ in the projection norm and the fixed variance matrices of the field implausibility, we can write the full implausibility $\mathcal{I}(\mathbf{x})$ such that only a single $\ell \times \ell$ matrix inversion is required. All of the variability due to $\mathbf{x}$ is evaluated within a $q \times q$ inversion, and NROY

space can then be defined with repeated evaluations of the fast $\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x})$.

Theorem 1. *For basis $\boldsymbol{\Gamma}_q$, and $\boldsymbol{W} = \boldsymbol{\Sigma}_{\boldsymbol{e}} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, we have:*

$$\mathcal{I}(\boldsymbol{x}) = \mathcal{R}_{\boldsymbol{W}}(\boldsymbol{\Gamma}_q, \boldsymbol{z}) + \tilde{\mathcal{I}}_{\boldsymbol{W}}(\boldsymbol{x}), \tag{14}$$

and hence can write

$$\mathcal{X}_{NROY} = \{\boldsymbol{x} \in \mathcal{X} | \tilde{\mathcal{I}}_{\boldsymbol{W}}(\boldsymbol{x}) < T - \mathcal{R}_{\boldsymbol{W}}(\boldsymbol{\Gamma}_q, \boldsymbol{z})\}, \tag{15}$$

with $T = \chi^2_{\ell, 0.995}$.

That is, if all quantities are projected with $\mathbf{W} = \boldsymbol{\Sigma}_{\mathbf{e}} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, then we can exactly evaluate $\mathcal{I}(\mathbf{x})$, as the sum of $\mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}_q, \mathbf{z})$, the reconstruction error of $\mathbf{z}$ on basis $\boldsymbol{\Gamma}_q$ (given in (7)), and $\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x})$, the $\mathbf{W}$-projected subspace implausibility at $\mathbf{x}$. The reconstruction error is fixed for all $\mathbf{x} \in \mathcal{X}$, requiring a one-off calculation, whilst $\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x})$ involves only q-dimensional multiplications, for small q. The proof (given in the Appendix) relies on the well-known Woodbury formula [40,41], also used for efficient calculations by [3] (for inverting the high-dimensional matrix in the calibration likelihood) and [42] (outer product emulation).

From this result, we see that it is critical to ensure that the chosen basis, $\boldsymbol{\Gamma}_q$, does not result in the terminal case: if $\mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}_q, \mathbf{z}) > T$, then the bound for $\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x})$ is negative, and so all $\mathbf{x} \in \mathcal{X}$ are ruled out and $\mathcal{X}_{NROY}$ is empty. If the observations can be represented perfectly by the basis ($\mathcal{R}_{\mathbf{W}}(\boldsymbol{\Gamma}_q, \mathbf{z}) = 0$), then $\mathcal{I}(\mathbf{x}) = \tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x})$, suggesting that the chi-squared bound with ℓ degrees of freedom is also appropriate in the q-dimensional subspace (depending on the rank of the variance matrix when the emulator variance matrix, which is of rank q, is included).

Projecting the relevant objects in $\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x})$ using e.g., L_2 projection instead of with the $\mathbf{W}$ norm will break the equality in (14). Although calculating the coefficient implausibility with L_2 projection can give similar results to $\mathbf{W}$ projection (as for some choices in Section 5.5), and hence will also have a strong correlation to the true $\mathcal{I}(\mathbf{x})$ in these cases, in general the correlation decreases as more structure is added to $\boldsymbol{\Sigma}_{\mathbf{e}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$. In general, to ensure that no information is lost from the full field implausibility, it is important to project using $\mathbf{W}$, as in (14).

Given that we have a method for calculating the ℓ-dimensional $\mathcal{I}(\mathbf{x})$ efficiently if we make the assumption of emulation via a basis representation, this gives strong motivation for using such a basis when we wish to calibrate, matching to the whole output field without great expense and without needing to form the $\ell \times \ell$ emulator variance matrices, whilst taking advantage of the other benefits of basis emulation discussed in Section 3.

4.2 History matching bound

When faced with a complex, high-dimensional output field, and a small number of runs, at the start of a history matching exercise the emulator variance may be large relative to the other sources of uncertainty, so that $\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}) < T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z})$ for all $\mathbf{x}$, and nothing is ruled out. In such a situation, using χ^2_q as the bound instead could be appropriate (there are q directions in the dominant emulator variance). However, lowering the bound from $T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z})$ risks ruling out runs that we should not. Suppose that for some $\tilde{\mathbf{x}} \in \mathcal{X}$ we have:

$$\text{Var}[f(\tilde{\mathbf{x}})] = 0, \quad \chi^2_{q,0.995} < \tilde{\mathcal{I}}_{\mathbf{W}}(\tilde{\mathbf{x}}) < T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z}), \tag{16}$$

so that we have no emulator variance at $\tilde{\mathbf{x}}$, and the implausibility at this point lies below the bound and $\tilde{\mathbf{x}}$ should not be ruled out. If the emulator variance is higher for other inputs in $\mathcal{X}$, we may have:

$$\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}) < T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z}) \quad \forall \mathbf{x} \in \mathcal{X}, \tag{17}$$

so that no space is ruled out with the usual bound. However, we should not use $\chi^2_{q,0.995}$ instead as this will incorrectly rule out $\tilde{\mathbf{x}}$.

A possible strategy for estimating a suitable bound $\tilde{T}$ is as follows: let $V_{min} = \min_{\mathbf{x}\in\mathcal{X}} \text{Var}[\mathbf{c}(\mathbf{x})]$ be the minimum emulator variance, and let $\mathcal{C}^*$ be the set containing all coefficients that are considered close enough to $\mathbf{z}$ in the subspace when there is zero emulator variance, i.e. coefficients that lie in the true NROY space, and set:

$$\begin{aligned} \mathcal{C}^* &= \{\mathbf{c} : (\mathbf{c}(\mathbf{z}) - \mathbf{c})^T (\mathbf{P}_{\mathbf{W}} \mathbf{W} \mathbf{P}_{\mathbf{W}}^T)^{-1} (\mathbf{c}(\mathbf{z}) - \mathbf{c}) < T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z})\}, \\ \tilde{T} &= max\{\max_{\mathbf{c}\in\mathcal{C}^*} (\mathbf{c}(\mathbf{z}) - \mathbf{c})^T (\mathbf{P}_{\mathbf{W}} \mathbf{W} \mathbf{P}_{\mathbf{W}}^T + V_{min})^{-1} (\mathbf{c}(\mathbf{z}) - \mathbf{c}), \chi^2_{q,0.995}\}, \\ \mathcal{X}_{NROY} &= \{\mathbf{x} \in \mathcal{X} | \tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}) < \tilde{T}\}. \end{aligned} \tag{18}$$

Calculating the distance metric for $\mathbf{c} \in \mathcal{C}^*$ with V_{min} gives a lower bound for the actual $\tilde{\mathcal{I}}_{\mathbf{W}}$ (increasing the emulator variance decreases $\tilde{\mathcal{I}}_{\mathbf{W}}$), hence $\tilde{T}$ ensures that we do not incorrectly rule out good model output.

As $V_{min} \to 0$ (either because there is no nugget in the emulator, or because we have a more accurate emulator at a later wave), then $\tilde{T} \to T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z})$. Therefore, this estimate is consistent with the theoretical result, with an adjustment using $\text{Var}[\mathbf{c}(\mathbf{x})]$ for practicality.

For the $\tilde{\mathbf{x}}$ example above, this method would result in no change to the bound as $V_{min} = 0$, so we would still be

unable to rule out any of $\mathcal{X}$, which may suggest that the tolerance to error is too large. In most cases, (18) will allow a larger percentage of parameter space to be ruled out in an accurate manner, without resorting to the q-dimensional assumption which can have negative consequences if there is a large degree of variability in the emulator variance.

4.3 Calibration efficiency for univariate emulators

With independent emulators for each grid box, there is no basis structure to emulator predictions, and the result from Theorem 1 does not apply. Although the $\ell \times \ell$ emulator variance $\text{Var}[f(\mathbf{x})]$ is diagonal due to the independence assumption, $\mathbf{\Sigma_e}$ and $\mathbf{\Sigma_\eta}$ will generally not be, hence there is an expensive inversion that varies with $\mathbf{x}$. If we assume that $\mathbf{W} = \mathbf{\Sigma_e} + \mathbf{\Sigma_\eta}$ is diagonal, then $\mathcal{I}$ can be calculated reasonably quickly, involving multiplication of ℓ-dimensional vectors at each $\mathbf{x}$. Given the fixed costs required for Theorem 1 ($\mathbf{W}^{-1}$, projections of $\mathbf{z}, \mathbf{\Sigma_e}, \mathbf{\Sigma_\eta}$), at $\mathbf{x}$ we only have q-dimensional quantities, and as $q << \ell$ this method will still generally be faster, despite the matrix inversion. Regardless, an uncorrelated error structure will not generally be a reasonable assumption, hence the flexibility afforded by the structure in Theorem 1 is beneficial.

Given emulators for every grid box, for efficiency we could instead use the univariate implausibility for each output individually, ignoring any correlations in the variance matrices, or match to global summaries or particular aspects of the output, as is commonly done for climate model output. For example, [24] history match to 9 regional summaries, and [43] use individual locations chosen to be representative of the full output. For correlated, smooth output these strategies should be successful, though using summaries can lose some of the original information and fail to rule out poor model runs, e.g. due to competing biases cancelling out, which should not be an issue with the basis structure and appropriate projection from Theorem 1. With such computer models generally being expensive, exploiting the information as much as possible at each wave can be important.

5. CASE STUDY: GLIMMER ICE SHEET MODEL

In this section, we apply independent and basis emulation methods to an ice sheet model. Section 5.1 compares the accuracy of the emulators for various metrics. Section 5.2 evaluates qualities of posterior samples. Section 5.3 compares the computational time required for emulating and calibrating with each method, Section 5.4 compares results for history matching and Section 5.5 considers sensitivity to the chosen projection method.

Glimmer is an ice sheet model that simulates the growth and retreat of ice sheets over North America [44]. The output here is spatio-temporal fields of ice thickness over a 194×150 spatial grid every 100 years from 21,000-6,000 years ago, covering the retreat of the North American ice sheet from the last glacial maximum to the end of the last ice

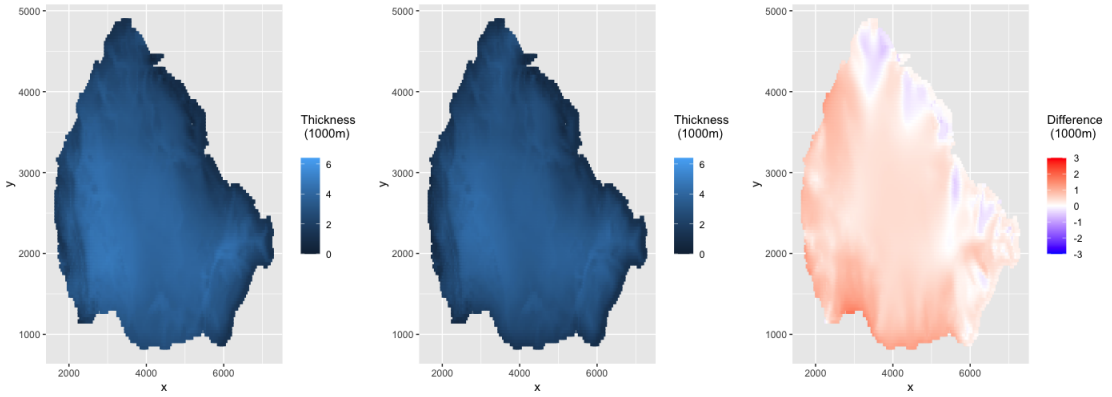

FIG. 2: Left: proxy observations. Centre: mean of the training data. Right: difference between the observations and the mean, all in 1000s of metres. The data is plotted on a projected grid (Lambert Azimuthal Equal Area projection centred at (45,-95)).

age. In this study, we consider only the ice thickness at 21,000 years ago, and use an ensemble $\mathbf{F} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ of $n = 500$ model runs from [35], where there are a total of 11 input parameters that may be active at the start of the deglaciation: 7 parameters controlling the physics of the ice sheet itself, and 4 parameters controlling the boundary condition temperature.

A high proportion of the 29,100 grid boxes always have zero ice thickness as they lie beyond the maximum ice sheet extent at 21,000 years ago. We filter out grid boxes, generally around the edges of the ice sheet, where there is commonly (for at least 20% of runs) no ice, resulting in an output field with dimension $\ell = 8922$. We remove ensemble members that contain different ice patterns, such as patches of zeros within the boundary of the ice sheet, and construct an $n = 250$-member ensemble where the ice extent is relatively consistent. In the vast majority of locations, thickness is significantly different from zero, so that negative predictions are not an issue and our emulator assumptions are valid, and this dataset is appropriate for comparing the two emulation methodologies. We split the filtered ensemble into training (100 runs, roughly $10 \times$ number of parameters, following [45]) and validation (150 runs) sets.

There are no geological observations of ice sheet thickness thousands of years in the past, only estimates of the extent of the ice sheet and its volume at various times. For modern ice sheets, maps of ice thickness exist (e.g. [46] for Greenland). So that we have a map of ice thickness to calibrate to, we select some $\mathbf{x}^*$ from the validation set, and take this known $f(\mathbf{x}^*)$ to be proxy observations $\mathbf{z}$. Figure 2 compares our proxy observations and the mean of the training data, showing that there is generally thicker ice at $\mathbf{x}^*$ than in the ensemble in general (red in the 3rd plot).

To calibrate, we require an error specification as in (8). We set $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = 0$ as by construction we know that there exists $\mathbf{x}^*$ such that $\mathbf{z}$ can be produced by the model, up to $\mathbf{e}$, and define $\boldsymbol{\Sigma}_{\mathbf{e}}$ using a squared exponential kernel based

on the distance between spatial locations (given by $\mathbf{s}_i = (s_{i1}, s_{i2})$), with $(i,j)^{th}$ entry:

$$\mathbf{\Sigma}_{\mathbf{e}}^{ij} = \sigma^2 \exp\{-\sum_{k=1}^{2} |s_{ik} - s_{jk}|^2 / \delta_k\}, \tag{19}$$

with correlation lengths $\boldsymbol{\delta} = (0.05, 0.05)$ and variance $\sigma^2 = 0.01$. This choice of $\boldsymbol{\delta}$ ensures that there is some correlation between errors at close spatial locations, which is a reasonable assumption for physical fields, whilst the variance has been chosen such that only a small region of the input space leads to output that is consistent with $\mathbf{z}$ (here, 1.2% of the 250 model runs lie in the true NROY space). We often find this to be the case for complicated computer models of physical systems (e.g. climate model output in [15]).

Overall, the approach to emulation and history matching (for the basis method) is summarised as:

- Calculate PC basis across training ensemble $\mathbf{F} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) \in \mathbb{R}^{\ell \times n}$ (as in (2));
- Apply basis selection algorithm if required, returning $\mathbf{\Gamma}_q$ where the 1st few vectors explain the majority of variability in $\mathbf{F}$ (Section 2.2);
- Project model runs $f(\mathbf{x}_i)$ onto chosen basis (as in (3));
- Emulate coefficients via Gaussian processes (as in (5));
- Calculate implausibility $\mathcal{I}(\mathbf{x})$ (using (14)), given $\mathbf{z}$, $\mathbf{\Sigma}_{\mathbf{e}}$, $\mathbf{\Sigma}_{\boldsymbol{\eta}}$, emulators.

5.1 Emulation

We emulate the model output using the 100 training runs, both by building emulators for every grid box (UV), and using a basis choice (BAS). For the basis $\mathbf{\Gamma}_q$ we consider both the PC basis and a rotated basis [15] (so that $\mathbf{z}$ is represented as well as possible given $\mathbf{\Sigma}_{\mathbf{e}}$ and the 100 training runs, as described in Section 2.2). These bases have $q = 4$ and $q = 5$ basis vectors respectively, both explaining greater than 90% of the variability in the training data (requiring 95% to be explained would require a further 9-10 basis vectors, and attempting to emulate these resulted in poor predictive ability. If this loss of information results in poor performance compared to UV, then the basis and emulation choices may need to be revisited).

Figure 3 compares the difference between $\mathbf{z}$ and its representation on the truncated PC basis (left) and the truncated rotated basis (right), with this anomaly reduced by the latter. There remain errors towards the edges of the ice sheet, but the smooth patterns in the centre of $\mathbf{z}$, that the truncated PC basis could not capture, are represented more

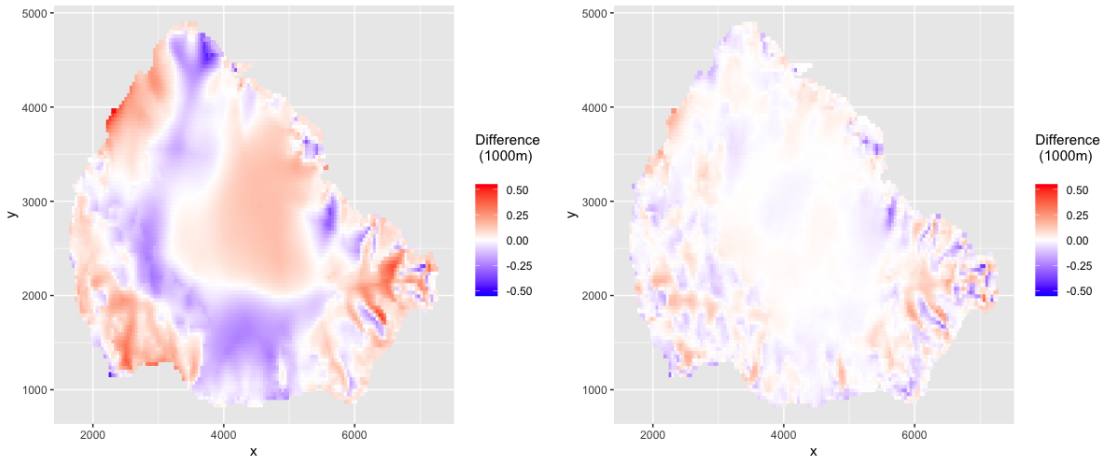

FIG. 3: The difference between the observations and their reconstruction in 1000s of metres, with the truncated PC basis (left) and the truncated rotated basis (right).

Type:	**UV**	**UV+**	**BAS**	**BAS+**	**PC**
Number of emulators	$\ell = 8922$	$\ell = 8922$	$q = 5$	$q = 5$	$q = 4$
RMSE, validation set	0.1510	0.1489	0.1616	0.1592	0.1628
RMSE, $\mathbf{x}^*$	0.1574	0.1554	0.1076	0.1187	0.1207
W error, validation set	0.9979	1.0735	0.9223	0.8701	0.8934
W error, $\mathbf{x}^*$	1.0153	1.2293	0.3816	0.4377	0.4450

TABLE 1: Validation summaries for different sets of emulators, for chosen best input $\mathbf{x}^*$. For the validation set, the reported RMSE and **W** error is the median across the 150 runs. The RMSE and **W** error are only comparable across the emulator types, not with each other.

accurately by the rotated version. Minimising the error is generally important, so that we can explain as much of the modelled process as possible via our emulators, and avoid assigning explainable patterns to random error or systematic bias, hence we mainly focus on the rotated choice, and this is the choice referred to as method BAS hereafter.

When constructing emulators, we use the RobustGaSP package [47] as it is fast at fitting large numbers of GPs. We fit emulators with both a constant mean and a linear mean, and allow the correlation lengths to vary across the univariate emulators. Doing so means that the expectation for the full field at $\mathbf{x}$ cannot be calculated as efficiently as in [23], but as we require the variance and implausibility metric thereafter (more expensive), and as calibration applications have generally allowed for this full flexibility, we allow this here. For the basis emulators, we fit emulators with a constant mean, a linear mean, and with an automatically chosen mean function (a luxury we can afford owing to the small number of GPs required). We use the default kernel choice in RobustGaSP, the Mateŕn 5/2 kernel.

We compare the performance of each set of emulators across the validation set and at $\mathbf{x}^*$, with Table 1 showing several summary statistics for each. The sets of emulators compared are UV (constant mean for each grid box), UV+ (the most accurate out of constant mean and linear for each grid box), BAS (constant mean for the basis coefficients),

BAS+ (best of constant, linear, and structured mean), and PC (constant mean only for this basis, as adding structure did not improve these metrics). In general, the UV+ and BAS+ options add some accuracy, but the constant mean approach performs relatively well in this example.

The table gives the median RMSE across the 150 validation runs and at $\mathbf{x}^*$, and the median error when scaled by $\mathbf{W}$. When we consider the RMSE, the two UV emulators generally outperform all basis options, with UV+ giving a slight improvement. At $\mathbf{x}^*$, the accuracy is substantially better for BAS, BAS+, and PC. When we instead consider the error in the $\mathbf{W}$ norm, all basis methods now outperform the UV options, both in terms of the validation set and at $\mathbf{x}^*$. It is unsurprising that the BAS methods perform better in the $\mathbf{W}$ norm, as these correlations are accounted for when we project onto the basis. Likewise, if we consider the errors independently, then it is reasonable to expect that the UV method will perform better, as it is unconstrained by correlations, and is attempting to fit as well as possible in each grid box, rather than trading off between related grid boxes. The UV+ method is worse than UV in the $\mathbf{W}$ norm as the predictions may change rapidly between adjacent grid boxes when switching from a constant mean to linear mean, and whilst in terms of L_2 error it is more accurate, this is not true if the errors are correlated.

At $\mathbf{x}^*$, the basis emulators are more accurate in this example, regardless of the basis choice, possibly aided by the correlated structure built into the emulators allowing more accurate extrapolation to this part of space. For the UV methods, we need to extrapolate simultaneously in several locations in order to get closer to $\mathbf{z}$, which is more difficult than when we have this structure built in.

Note that even if the coefficient emulators are perfect, the error at $\mathbf{x}^*$ will not be zero, unless $\mathbf{z}$ can be written exactly as a linear combination of the q chosen basis vectors. In practice, this will not be the case, and the most accurate the coefficient emulators can be is given by the reconstruction error (7). For the PC basis, this error is 0.395, whilst for the rotated basis it is 0.187. Hence, although the final row of Table 1 suggests a limited difference between the prediction at $\mathbf{x}^*$ between the two bases, there is more potential for improvement with the rotated basis (say, given additional model runs), whereas the PC basis is close to its theoretical minimum.

Figure 4 compares the two error metrics for the UV and BAS emulators, showing that when considering the RMSE (left), the error for BAS is generally slightly higher, with the converse true for the $\mathbf{W}$ error, with a correlation around 0.91 for each metric. The fact that there is relatively little difference between the two methods lends strength to the argument for a basis method when considering their subsequent usage.

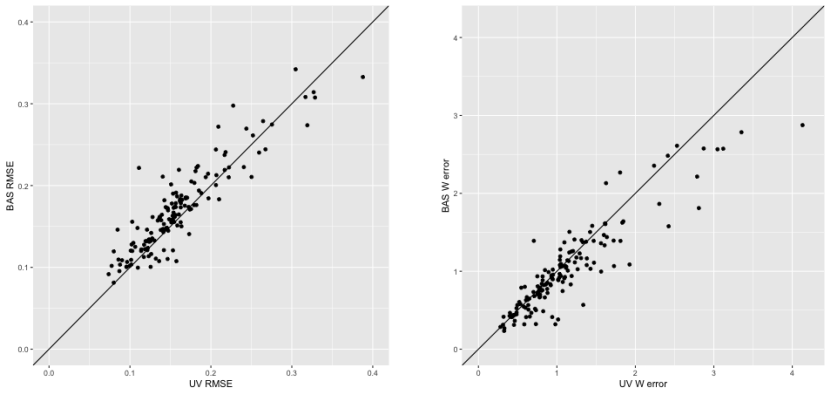

FIG. 4: The RMSE (left) and **W** error (right) across the validation set using the UV and BAS emulators.

5.2 Posterior sampling

Figure 5 compares emulator samples at $\mathbf{x}^*$ across two fixed latitudes (with the phenomenon exaggerated for a full 2D map), for the UV (left) and BAS (right) emulators. In each case, the emulator mean (blue) is relatively consistent with the truth (green), and is almost completely contained within the samples, with the BAS emulators generally more accurate for the top profile, whilst UV is better for the lower profile (particularly in the mean at the eastern end).

The main difference between the two plots is in the samples, with individual samples from BAS resembling the physical model output, where this is not true for the UV emulator due to the lack of correlation. History matching only requires posterior expectations and variances to calculate implausibilities, hence when this is our goal, the physical incoherence of samples is not necessarily a problem. If instead the goal is prediction (e.g. in a time series we want to predict the output at the next, unobserved timepoints), or we are attempting to infer the probability of significant ice loss or risk of collapse of an ice sheet in a region, it then becomes extremely important to account for correlations between locations.

Similarly as in Section 3.2, suppose that the average predicted thickness along the second ice thickness profile shown in Figure 5 is considered a reasonable proxy for overall changes in the ice sheet. According to samples from the UV emulator, the average relative thickness is between 0.31 and 0.38, whilst for the BAS samples the range is 0.23-0.51, so that the correlated samples show a chance of both higher and lower thicknesses than would be expected by the UV case. If a collapse is triggered at an average level of less than 0.3 or 0.25, the UV samples indicate that this cannot happen, whereas the conclusion from BAS is that there is a non-zero chance of the true response being at this level.

A similar range may be achieved by considering samples at individual grid boxes for the UV emulator, but how

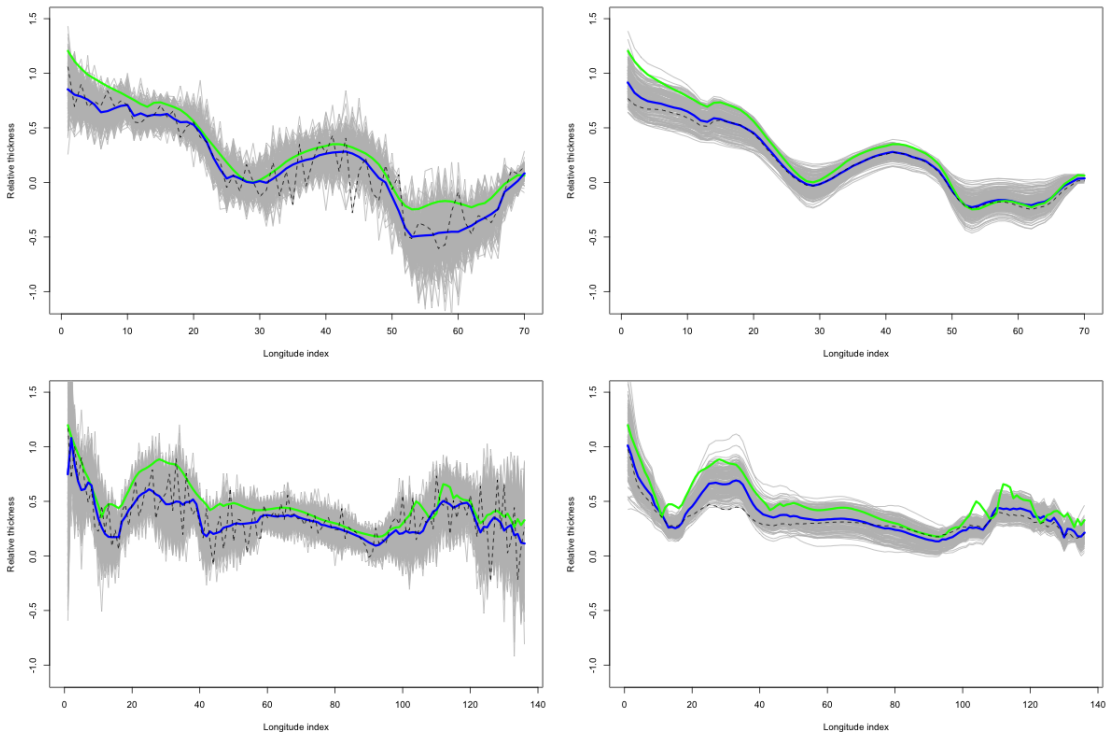

FIG. 5: Ice sheet profile at two latitudes for the UV (left) and BAS emulators (right), at $\mathbf{x}^*$, where the green line is the observed profile, the blue line is the emulator mean, and the grey lines are samples from the emulator variance. The black dotted line represents one particular sampled profile.

these relate to each other if multiple grid boxes across a region are important to the overall response is not clear.

5.3 Computational time

We compare the time taken to evaluate predictions and implausibilities for the UV and BAS methods for this $\ell = 8922$-dimensional example. We ignore the cost of fitting and validating the emulators themselves, as we could have used the faster method of [23], and instead focus on the post-emulation calculations.

To build a picture of NROY space (as in Section 5.4), 10^6 or more samples from the emulator posterior may be required, although this is dependent on the dimension of $\mathcal{X}$, the size of the NROY space (for small NROY spaces, we will require more samples to find a given number of acceptable points), and the number of waves already performed. The UV method requires ℓ emulator evaluations at each $\mathbf{x}$, compared to q for BAS, hence if we have 10^6 samples from $\mathcal{X}$, in our ice sheet example we need 5 million evaluations for BAS, compared to 8922 million for UV (a factor of 1784 greater). The number of evaluations required for BAS scales with Nq, for N the number of samples, instead of $N\ell$ for UV.

Samples	$\mathrm{E}[f], \mathrm{Var}[f]$	UV impl	**UV**	$\mathbf{W}^{-1}$	$\mathrm{E}[\mathbf{c}], \mathrm{Var}[\mathbf{c}]$	$\mathcal{I}(\mathbf{x})$	**BAS**
10^3	195.33	0.27	196	1422	0.11	4.02	1426
10^4	1714.45	2.88	1717	1422	1.01	4.47	1427
10^5	17017	31.35	17048	1422	10.01	6.76	1439
10^6	219334	317.26	219651	1422	129.02	27.93	1579

TABLE 2: Time (in seconds) to calculate emulator predictions and implausibilities for the UV and BAS methods, using a MacBook Pro with 8GB memory, 2.3 GHz Intel Core i5 processor.

Table 2 compares the computational time required by the UV and BAS methods when the number of samples from $\mathcal{X}$ increases. For UV, we evaluate $\mathrm{E}[f(\mathbf{x})]$ and $\mathrm{Var}[f(\mathbf{x})]$, and due to the lack of structure resulting in no fast method for calculating $\mathcal{I}$, resulting in an expensive inversion of a non-diagonal variance matrix at any $\mathbf{x}$ (as $(\mathrm{Var}[f(\mathbf{x})] + \boldsymbol{\Sigma}_{\mathbf{e}})^{-1}$ varies in $\mathbf{x}$), we calculate the univariate implausibility using the average expectation and variance across the ℓ outputs (calibrating using a summary as in [24]). For BAS, we need $\mathbf{W}^{-1}$, $\mathrm{E}[\mathbf{c}(\mathbf{x})], \mathrm{Var}[\mathbf{c}(\mathbf{x})]$ and $\mathcal{I}(\mathbf{x})$, with the implausibility using (14).

Overall, the BAS method is significantly faster for 10^5 or more samples, with little additional time required as the number of samples increases by a factor of 10. BAS requires a larger initial cost, with the one-off inversion of $\mathbf{W}$, but any subsequent matrix calculations (e.g. the reconstruction error) exploit stored quantities, and large savings are realised by the significantly fewer emulator evaluations required. Greater parallelisation would significantly improve the UV method, but such improvements would benefit the BAS method also, with it being difficult to surpass the efficient calculation of $\mathcal{I}$ by exploiting the basis structure.

Given emulator expectations and variances, the implausibility for BAS is also inexpensive to evaluate across a large sample, due to (14), with a million evaluations possible in under 30 seconds (including the time to calculate the reconstruction error, and project the required quantities onto the basis). For UV, even if we average the outputs at $\mathbf{x}$ and calculate a univariate implausibility metric, this implausibility calculation is slower when the number of samples is 10^5 or higher, due to averaging across large matrices.

As ℓ increases, in general there is a larger initial cost for basis methods due to the inversion of $\mathbf{W}$ (unless this is diagonal), but all subsequent calculations, sampling and inferences are relatively cheap given q is always small, whilst the number of emulator evaluations increases with ℓ in the UV case, leading to greater computation for any future tasks.

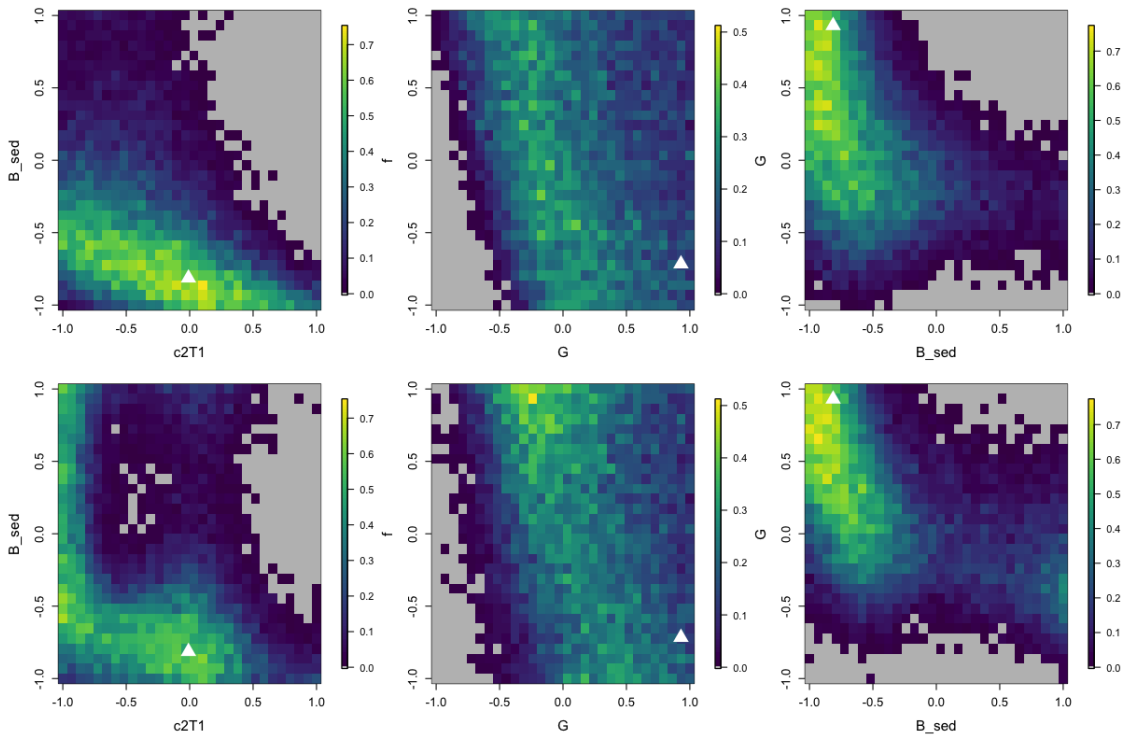

FIG. 6: Density plot of NROY space for selected pairs of input parameters, for the UV NROY space (top row) and the BAS+ NROY space (equivalent plots on the bottom row). Each pixel in a pairwise plot shows the proportion of runs that are in NROY space, averaged across the remaining parameters. Grey regions indicate parts of space where everything is ruled out. The white triangle indicates the location of $\mathbf{x}^*$, showing that this value has not been ruled out.

5.4 History matching

There are several barriers to a direct comparison between the two approaches when calibrating. Exploiting the basis structure in (14), even for $\ell = 8922$ it is extremely fast to evaluate the full ℓ-dimensional implausibility for millions of points, whereas if we aim to calculate $\mathcal{I}(\mathbf{x})$ without this structure (i.e. with the UV emulators), there is a large expense.

Due to this prohibitive expense for the UV emulators, as seen in Section 5.3, we can instead consider either a) calibrating to a global summary or b) assuming a diagonal $\mathbf{\Sigma_e}$, so that inversion is significantly faster (simply multiplying by a vector). Neither option gives a perfect comparison, as the change in assumptions about the error structure of the field alters the underlying 'true' space that we are searching for, and there is not a consistency between thresholds for the different methods. Therefore, we assume that the BAS approach (that we can calculate fully and efficiently) is correct, and make assumptions that allow an approximate approach to be compared to this. In practice, if the decision was made to fit univariate emulators, either of the above would be chosen, and error tolerances defined

with this in mind.

Using the BAS+ emulators, we calculate the full implausibility as in (14), evaluating the emulators and implausibilities for a space-filling sample from $\mathcal{X}$. As the emulator variance is large relative to $\mathbf{\Sigma_e}$, the standard threshold of $T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z})$ rules out none of $\mathcal{X}$. If we were to simply consider the coefficient implausibility instead and set the bound for ruling out space as $\chi^2_{q,0.995}$, then 0.21% of runs are considered to be not implausible. Given that 1.2% of the runs across the training and validation sets are known to lie in the true NROY space, and given the presence of high emulator variance at this wave, this is almost certainly a too strict definition of $\mathcal{X}_{NROY}$, and we risk ruling out suitable regions of $\mathcal{X}$.

To provide a pragmatic bound that should avoid incorrectly ruling out too much of $\mathcal{X}$, we use the algorithm outlined in (18). For each basis vector, the emulator contained a vector so that the minimum variance (V_{min}) is not equal to zero, and therefore the estimated bound will not be equal to the theoretical bound $T - \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z})$. The resulting bound is estimated as $\tilde{T} = 100.5$, leading NROY space to be given by:

$$\mathcal{X}_{NROY} = \{\mathbf{x} \in \mathcal{X} | \tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}) < 100.5\}, \tag{20}$$

substantially less strict than setting the bound as $\chi^2_{q,0.995} = 16.7$, and hence we avoid the mistake of ruling out space too aggressively. This NROY space consists of 16% of $\mathcal{X}$.

To compare with the UV+ emulators, we set all non-diagonal entries of $\mathbf{\Sigma_e}$ to 0. With errors assumed to be independent, the implausibility values will be higher, and the usual bound results in the vast majority of space being ruled out (and we may have set a different error tolerance initially if we truly did have independent errors). Instead of using the same bound, we set the bound as the 16^{th} percentile, so that the NROY space is the same size as in (20), and we can at least compare the structure of the two spaces (i.e. do we have similar ordering of points).

Figure 6 compares the proportion of space classified as $\mathcal{X}_{NROY}$ for the two examples, with pairs of input parameters plotted, with all other parameters randomly sampled and averaged across, for the UV approximation (top) and BAS implausibility (bottom). There are clear relationships between some of the parameters, with some particular combinations of parameters completely ruled out (grey regions). There are differences, unsurprising due to the change in assumption made about the error structure, but in general the same parts of space are being completely ruled out, and the highest density is often in a similar location.

We can further compare the NROY spaces as a whole by plotting (unnormalised) densities for each parameter, weighted by the exponential of the negative implausibility, with zero posterior density outside of $\mathcal{X}_{NROY}$, as in Figure

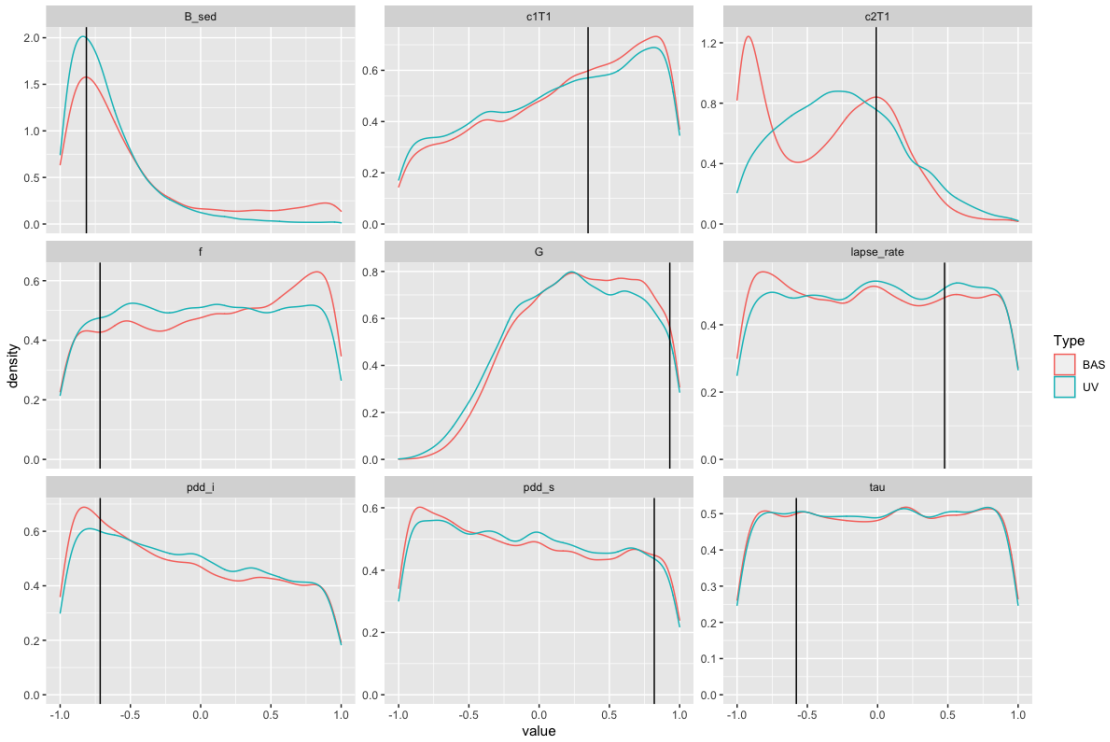

FIG. 7: Weighting the UV and BAS NROY spaces by $exp\{-\mathcal{I}(\mathbf{x})\}$, for nine of the model parameters, with the vertical line giving the location of $\mathbf{x}^*$.

7. Again, we see that in general the two methods are relatively similar, with peaked distributions around or close to $\mathbf{x}^*$ for some parameters (e.g. B_sed), whilst others are still relatively uniform at this stage. Biases in density away from $\mathbf{x}^*$ (given by the vertical line) does not imply a poor calibration, as it is possible that other regions of the parameter space, $\mathcal{X}$, lead to output consistent with $\mathbf{z}$. In reality, we would aim to refine the emulation and calibration by refitting emulators with the validation runs included, and/or running new ensembles of Glimmer, sampled from NROY space (and past examples have demonstrated that calibration accuracy can be improved after performing a few waves of history matching, e.g. [26]).

Approximating the error variance by forcing it to be diagonal, therefore, does not appear to have made a huge difference, given that we can set an appropriate threshold for ruling out space. This result is similar to how there was not a huge difference in the emulator accuracy for the two cases, but given that one choice is significantly faster whilst allowing greater flexibility in the error specification (at least whilst an $\ell \times \ell$ inversion is feasible as a one-off), this is likely the better choice in most cases.

5.5 Projection method

The results from calibrating using basis coefficients, with a measure as in (13), are sensitive to the choice of $\mathbf{W}$ (i.e. to the projection norm used). In this section, we consider two cases: i) setting $\mathbf{W} = \mathbf{\Sigma_e}$, for consistency with basis selection (Section 2.1.3); ii) setting $\mathbf{W} = \mathbb{I}_\ell$, i.e L_2 projection, so that $\mathbf{P}_{\mathbb{I}_\ell} = (\mathbf{\Gamma}_q^T\mathbf{\Gamma}_q)^{-1}\mathbf{\Gamma}_q^T$. It is not clear whether projection in a different norm will greatly affect the results, because $\mathbf{W} = \mathbf{\Sigma_e}$ is included in (13) regardless of the projection method (via the inverted variance term).

To demonstrate the potential difference, we calculate the coefficient implausibility for the 100 fields in the Glimmer training data using both projection methods (with $\text{Var}[\mathbf{c}(\mathbf{x})] = 0$, so that projection is the only difference), for different choices of $\mathbf{\Sigma_e}$ (as defined in equation (19)). We vary $\mathbf{\Sigma_e}$ by changing the correlation lengths $\boldsymbol{\delta}$, and allowing a non-constant variance across the output field, with the variance above an arbitrary latitude (with 8% of outputs above this line) set at a chosen σ_n^2, with $\sigma^2 = 0.01$ elsewhere as previously. Such an error variance represents how correlated the errors are, and potentially down-weights errors in the far north (i.e. we are more interested in representing $\mathbf{z}$ accurately elsewhere).

Table 3 gives the correlation between $\tilde{\mathcal{I}}_{\mathbf{\Sigma_e}}(\mathbf{x})$ and $\tilde{\mathcal{I}}_{\mathbb{I}_\ell}(\mathbf{x})$ as $\mathbf{\Sigma_e}$ is altered by changing ($\boldsymbol{\delta}$, σ_n^2), with $\boldsymbol{\delta} = 0.05$ and $\sigma_n^2 = 0.01$ equivalent to $\mathbf{\Sigma_e}$ from Section 5. Generally, as the degree of correlation in $\mathbf{W}$ increases, the strength of the relationship between the two implausibilities decreases. Similarly, there is a decrease in the correlation between the two metrics as the two variance multipliers become more different, although this is more evident for the less correlated $\mathbf{\Sigma_e}$ choices. This lack of correlation between the two options is despite the fact that they both include the new $\mathbf{W}$ as their variance term: the new $\mathbf{\Sigma_e}$ enters into $\tilde{\mathcal{I}}_{\mathbb{I}_\ell}(\mathbf{x})$ via the variance in every case.

There are a number of consequences of changing the projection choice. The ordering of points may differ between the the two metrics, resulting in different conclusions about $\mathbf{x}^*$, either in terms of its posterior distribution or the single value of $\mathbf{x}^*$ that minimises distance to $\mathbf{z}$. This can also cause changes in the size and composition of the resulting NROY space. Even when consistency between the two metrics is relatively high, there could still be changes in the ordering of points, and whether or not points should be ruled out. For the $\mathbf{\Sigma_e}$ used in the example in Section 5, there is an extremely strong correlation (≈ 1) whether we use L_2 or $\mathbf{W} = \mathbf{\Sigma_e}$ projection, but in general $\mathbf{W}$ can cause a difference in calibration results.

6. DISCUSSION

In this article, we have presented a method for efficiently emulating and history matching large output fields, without any loss of information, by using a basis structure when emulating, achieving calibration over the full field for the cost

δ, σ_n^2	**0.01**	**1**	**100**
0	1	0.956	0.955
0.05	1	0.989	0.941
0.5	0.257	0.466	0.433
1	0.522	0.350	0.457

TABLE 3: Correlations between the subspace implausibility with projection in L_2 and $\mathbf{W} = \mathbf{\Sigma_e}$, for different choices of $\mathbf{\Sigma_e}$, varying the correlation length δ and the variance multiplier σ_n^2 for the far north of the ice sheet.

of calibration in the subspace. Despite a relatively similar accuracy for the basis and univariate emulation approaches for a particular ice sheet example, the basis structure offers several advantages that make it a more suitable choice in many cases, including efficiency in terms of building and validating the smaller number of emulators (allowing more time for human input to improve emulation), efficiency in evaluating predictions and the implausibility metric for any $\mathbf{x} \in \mathcal{X}$, the physical-coherency of posterior samples from the emulator, and full flexibility in the definition of variance matrices $\mathbf{\Sigma_e}$ and $\mathbf{\Sigma_\eta}$.

Given a choice of basis for emulation, and the variance matrices required in history matching ($\mathbf{\Sigma_e}, \mathbf{\Sigma_\eta}$), we showed that the expensive, ℓ-dimensional implausibility metric over the original field can be calculated exactly for any $\mathbf{x}$ with a single large matrix inversion required. To do so requires an appropriate choice of projection norm ($\mathbf{W} = \mathbf{\Sigma_e} + \mathbf{\Sigma_\eta}$), and we showed the sensitivity of calibration in a subspace to the projection norm. Given this 'correct' choice, we have that the full implausibility is the sum of the reconstruction error of the truncated basis and the coefficient implausibility with this projection. For a new $\mathbf{x}$, we require only a $q \times q$ matrix inversion for $q << \ell$, so that the standard history matching metric is tractable for the millions of evaluations required in history matching. Without such a basis structure and consistent projection, this would not be possible.

Although calibration only requires the expectation and variance of the field, being able to draw realistic samples is a benefit of the basis emulator, not given by an independent emulation approach (and given the computational savings afforded by such a method, getting this 'for free' is a nice property to have). This is also beneficial for Bayesian calibration, where full distributions are used, rather than the expectations and variances as in history matching.

When history matching, we typically perform a multi-wave experiment, so that savings given when emulating a single field once are multiplied as we move to later waves, where to determine whether a point is in the current NROY space, we may need to evaluate its emulator expectation, and implausibility, at each of the previous waves. Climate models, in particular, have many different high-dimensional output fields, with the benefits of basis emulation and calibration increasing with the number of fields. Having $O(10)$ emulators per field, rather than thousands, enables more expert time to be spent on fitting each emulator and exploration of $\mathcal{X}$ to proceed more efficiently. Calculating $\mathcal{I}$

for each field is extremely fast (30 seconds for 1 million evaluations), whereas for UV emulators a summary would be needed to achieve this speed.

To apply a basis method, little extra work or knowledge is required, as standard univariate emulators can be fitted to basis coefficients, as in our application. Selecting an appropriate basis is therefore the main problem, and in many cases, an out-of-the-box method such as SVD (with a rotation when required) is fast and easy to apply, giving an intuitive spatial basis. Even if a summary is used for calibration, rather than the full field, so that the fast implausibility calculation demonstrated here is not required, building and evaluating fewer emulators gives computational savings, whilst yielding spatially-coherent samples from the emulator posterior.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from EPSRC fellowship No. EP/K019112/1, and 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).

APPENDIX A. PROOF OF THEOREM 1

We apply the Woodbury formula [40,41]:

$$(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1}+\mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}, \tag{A.1}$$

where $\mathbf{A}$ is an $\ell \times \ell$ matrix, $\mathbf{C}$ is a $q \times q$ matrix, $\mathbf{U}$ is an $\ell \times q$ matrix, and $\mathbf{V}$ is a $q \times \ell$ matrix.

To prove the result, we show that the difference between the field implausibility and the reconstruction error can be written as $\tilde{\mathcal{I}}_{\mathbf{W}}$. We first expand the field implausibility using the Woodbury formula, so that:

$$\begin{aligned}
\mathcal{I}(\mathbf{x}) &= (\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\boldsymbol{\Gamma}_q \mathrm{Var}[\mathbf{c}(\mathbf{x})]\boldsymbol{\Gamma}_q^T + \mathbf{W})^{-1}(\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]) \\
&= (\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T\{\mathbf{W}^{-1} - \mathbf{W}^{-1}\boldsymbol{\Gamma}_q(\mathrm{Var}[\mathbf{c}(\mathbf{x})]^{-1} + \boldsymbol{\Gamma}_q^T\mathbf{W}^{-1}\boldsymbol{\Gamma}_q)^{-1}\boldsymbol{\Gamma}_q^T\mathbf{W}^{-1}\}(\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]) \\
&= (\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T\mathbf{W}^{-1}(\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]) \\
&\quad - (\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\mathbf{W}^{-1}\boldsymbol{\Gamma}_q(\mathrm{Var}[\mathbf{c}(\mathbf{x})]^{-1} + \boldsymbol{\Psi})^{-1}\boldsymbol{\Gamma}_q^T\mathbf{W}^{-1})(\mathbf{z}-\boldsymbol{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]),
\end{aligned}$$

where $\boldsymbol{\Psi} = \boldsymbol{\Gamma}_q^T\mathbf{W}^{-1}\boldsymbol{\Gamma}_q$. Applying the Woodbury formula again, we have:

$$(\mathrm{Var}[\mathbf{c}(\mathbf{x})]^{-1} + \boldsymbol{\Psi})^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}(\mathrm{Var}[\mathbf{c}(\mathbf{x})] + \boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{\Psi}^{-1}. \tag{A.2}$$

Therefore, the field implausibility can be written as:

$$\begin{aligned}\mathcal{I}(\mathbf{x}) &= (\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])^T\mathbf{W}^{-1}(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])\\ &\quad-(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])^T\mathbf{W}^{-1}\mathbf{\Gamma}_q\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])\\ &\quad+(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])^T\mathbf{W}^{-1}\mathbf{\Gamma}_q\mathbf{\Psi}^{-1}(\mathrm{Var}[\mathbf{c}(\mathbf{x})]+\mathbf{\Psi}^{-1})^{-1}\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})]).\end{aligned} \tag{A.3}$$

By rewriting $\tilde{\mathcal{I}}_{\mathbf{W}}$ (from (13)),

$$\begin{aligned}\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}) &= (\mathbf{c}(\mathbf{z})-\mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\mathrm{Var}[\mathbf{c}(\mathbf{x})]+\mathrm{Var}[\mathbf{c}(\mathbf{e})]+\mathrm{Var}[\mathbf{c}(\boldsymbol{\eta})])^{-1}(\mathbf{c}(\mathbf{z})-\mathrm{E}[\mathbf{c}(\mathbf{x})])\\ &= (\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\mathrm{Var}[\mathbf{c}(\mathbf{x})]+\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}(\mathbf{\Sigma}_{\mathbf{e}}+\mathbf{\Sigma}_{\boldsymbol{\eta}})\mathbf{W}^{-1}\mathbf{\Gamma}_q\mathbf{\Psi}^{-1})^{-1}\times\\ &\quad(\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathrm{E}[\mathbf{c}(\mathbf{x})])\\ &= (\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\mathrm{Var}[\mathbf{c}(\mathbf{x})]+\mathbf{\Psi}^{-1})^{-1}(\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathrm{E}[\mathbf{c}(\mathbf{x})]),\end{aligned} \tag{A.4}$$

we have that the final line of (A.3) is the coefficient implausibility:

$$\begin{aligned}&(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])^T\mathbf{W}^{-1}\mathbf{\Gamma}_q\mathbf{\Psi}^{-1}(\mathrm{Var}[\mathbf{c}(\mathbf{x})]+\mathbf{\Psi}^{-1})^{-1}\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])\\ &= (\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\mathrm{Var}[\mathbf{c}(\mathbf{x})]+\mathbf{\Psi}^{-1})^{-1}\times\\ &(\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])\\ &= (\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathrm{E}[\mathbf{c}(\mathbf{x})])^T(\mathrm{Var}[\mathbf{c}(\mathbf{x})]+\mathbf{\Psi}^{-1})^{-1}(\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}\mathbf{z}-\mathrm{E}[\mathbf{c}(\mathbf{x})]).\end{aligned}$$

Hence, from (A.3), we have:

$$\begin{aligned}\mathcal{I}(\mathbf{x}) &= (\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])^T\mathbf{W}^{-1}(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])\\ &\quad-(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])^T\mathbf{W}^{-1}\mathbf{\Gamma}_q\mathbf{\Psi}^{-1}\mathbf{\Gamma}_q^T\mathbf{W}^{-1}(\mathbf{z}-\mathbf{\Gamma}_q\mathrm{E}[\mathbf{c}(\mathbf{x})])+\tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}).\end{aligned} \tag{A.5}$$

Next, we rewrite the reconstruction error by adding and subtracting $\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]$:

$$\begin{aligned}
\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z}) &= (\mathbf{z} - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z})^T \mathbf{W}^{-1} (\mathbf{z} - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) \\
&= (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] + \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z})^T \mathbf{W}^{-1} \times \\
&(\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] + \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) \\
&= (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]) + \\
&(\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z})^T \mathbf{W}^{-1} (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) + \\
&2(\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) \\
&= \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3.
\end{aligned} \tag{A.6}$$

$\mathcal{R}_1$ is already present in the decomposition of $\mathcal{I}(\mathbf{x})$ in (A.5). Using that:

$$\mathbb{I} = \mathbf{\Psi}^{-1} \mathbf{\Psi} = \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{\Gamma}_q, \tag{A.7}$$

we have:

$$\begin{aligned}
\mathcal{R}_2 &= (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z})^T \mathbf{W}^{-1} (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) \\
&= (\mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z})^T \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{\Gamma}_q (\mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) \\
&= (\mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z})^T \mathbf{\Psi} (\mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) \\
&= (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{z})^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Psi} \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{z}) \\
&= (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{z})^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{z}) \\
&= (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathcal{R}_3 &= 2(\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} (\mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})] - \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z}) \\
&= -2(\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} \mathbf{\Gamma}_q (\mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z} - \mathrm{E}[\mathbf{c}(\mathbf{x})]) \\
&= -2(\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} \mathbf{\Gamma}_q (\mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{z} - \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]) \\
&= -2(\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]).
\end{aligned}$$

Hence, from (A.6):

$$\mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z}) = (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]) - (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})])^T \mathbf{W}^{-1} \mathbf{\Gamma}_q \mathbf{\Psi}^{-1} \mathbf{\Gamma}_q^T \mathbf{W}^{-1} (\mathbf{z} - \mathbf{\Gamma}_q \mathrm{E}[\mathbf{c}(\mathbf{x})]),$$

and combining this with (A.5),

$$\mathcal{I}(\mathbf{x}) = \mathcal{R}_{\mathbf{W}}(\mathbf{\Gamma}_q, \mathbf{z}) + \tilde{\mathcal{I}}_{\mathbf{W}}(\mathbf{x}). \tag{A.8}$$

REFERENCES

1. von Salzen, K., Scinocca, J.F., McFarlane, N.A., Li, J., Cole, J.N., Plummer, D., Verseghy, D., Reader, M.C., Ma, X., Lazare, M., , The Canadian fourth generation atmospheric global climate model (CanAM4). Part I: representation of physical processes, *Atmosphere-Ocean*, 51(1):104–125, 2013.
2. Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P., Design and analysis of computer experiments, *Statistical science*, pp. 409–423, 1989.
3. Higdon, D., Gattiker, J., Williams, B., and Rightley, M., Computer model calibration using high-dimensional output, *Journal of the American Statistical Association*, 103(482), 2008.
4. Kennedy, M.C. and O'Hagan, A., Bayesian calibration of computer models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.
5. Craig, P.S., Goldstein, M., Seheult, A., and Smith, J., Bayes linear strategies for matching hydrocarbon reservoir history, *Bayesian statistics*, 5:69–95, 1996.
6. Williamson, D., Blaker, A.T., Hampton, C., and Salter, J., Identifying and removing structural biases in climate models with history matching, *Climate Dynamics*, 45(5-6):1299–1324, 2015.
7. Andrianakis, I., McCreesh, N., Vernon, I., McKinley, T.J., Oakley, J.E., Nsubuga, R.N., Goldstein, M., and White, R.G., Efficient history matching of a high dimensional individual-based hiv transmission model, *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):694–719, 2017.
8. Tuo, R. and Wu, C.J., Efficient calibration for imperfect computer models, *The Annals of Statistics*, 43(6):2331–2352, 2015.
9. Wong, R.K., Storlie, C.B., and Lee, T.C., A frequentist approach to computer model calibration, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):635–648, 2017.
10. Plumlee, M., Bayesian calibration of inexact computer models, *Journal of the American Statistical Association*, 112(519):1274–1285, 2017.
11. Gu, M. and Wang, L., Scaled gaussian stochastic process for computer model calibration and prediction, *SIAM/ASA Journal*

on Uncertainty Quantification, 6(4):1555–1583, 2018.

12. Liu, F. and West, M., A dynamic modelling strategy for Bayesian computer model emulation, *Bayesian Analysis*, 4(2):393–411, 2009.

13. Williamson, D. and Blaker, A.T., Evolving Bayesian emulators for structured chaotic time series, with application to large climate models, *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):1–28, 2014.

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

15. Salter, J.M., Williamson, D.B., Scinocca, J., and Kharin, V., Uncertainty quantification for computer models with spatial output using calibration-optimal bases, *Journal of the American Statistical Association*, 114(528):1800–1814, 2019.

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

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

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

19. Chang, W., Applegate, P.J., Haran, M., and Keller, K., 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, 2014.

20. Coveney, S., Corrado, C., Oakley, J.E., Wilkinson, R.D., Niederer, S.A., and Clayton, R.H., Bayesian Calibration of Electrophysiology Models Using Restitution Curve Emulators, *Frontiers in Physiology*, p. 1120, 2021.

21. Lee, L., Carslaw, K., Pringle, K., and Mann, G., Mapping the uncertainty in global ccn using emulation, *Atmospheric Chemistry and Physics*, 12(20):9739–9751, 2012.

22. Spiller, E.T., Bayarri, M., Berger, J.O., Calder, E.S., Patra, A.K., Pitman, E.B., and Wolpert, R.L., Automating emulator construction for geophysical hazard maps, *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):126–152, 2014.

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

24. Johnson, J.S., Regayre, L.A., Yoshioka, M., Pringle, K.J., Lee, L.A., Sexton, D.M., Rostron, J.W., Booth, B.B., and Carslaw, K.S., The importance of comprehensive parameter sampling and multiple observations for robust constraint of aerosol radiative forcing, *Atmospheric Chemistry and Physics*, 18(17):13031–13053, 2018.

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

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

27. Lee, L., Pringle, K., Reddington, C., Mann, G., Stier, P., Spracklen, D., Pierce, J., and Carslaw, K., The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei, *Atmospheric Chemistry and Physics*, 13(17):8879–8914, 2013.

28. Gu, M. and Shen, W., Generalized probabilistic principal component analysis of correlated data., *J. Mach. Learn. Res.*, 21:13–1, 2020.

29. Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K., 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, 2013.

30. Vernon, I., Liu, J., Goldstein, M., Rowe, J., Topping, J., and Lindsey, K., Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions, *BMC systems biology*, 12(1):1, 2018.

31. Vernon, I. and Goldstein, M., Bayes linear analysis of imprecision in computer models, with application to understanding galaxy formation, In *Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications*, pp. 441–450. Society for Imprecise Probability: Theories and Applications, 2009.

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

33. Andrianakis, I., Vernon, I.R., McCreesh, N., McKinley, T.J., Oakley, J.E., Nsubuga, R.N., Goldstein, M., and White, R.G., Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda, *PLoS computational biology*, 11(1):e1003968, 2015.

34. Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Fairhead, L., Diallo, F.B., and Volodina, V., Process-based climate model development harnessing machine learning: Ii. model calibration from single column to global, *Journal of Advances in Modeling Earth Systems*, 13(6):e2020MS002225, 2021.

35. Salter, J.M., Williamson, D.B., Gregoire, L.J., and Edwards, T.L., Quantifying spatio-temporal boundary condition uncertainty for the North American deglaciation, *SIAM/ASA Journal on Uncertainty Quantification*, Accepted, 2022.

36. Xu, W., Williamson, D.B., and Challenor, P., Local voronoi tessellations for robust multiwave calibration of computer models, *International Journal for Uncertainty Quantification*, 11(5), 2021.

37. Volodina, V. and Williamson, D., Diagnostics-driven nonstationary emulators using kernel mixtures, *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):1–26, 2020.

38. Williamson, D. and Vernon, I., Efficient uniform designs for multi-wave computer experiments, *arXiv preprint arXiv:1309.3520*, 2013.

39. Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., Rio, C., Audouin, O., Salter, J., Bazile, E., , Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement, *Journal of Advances in Modeling Earth Systems*, 13(3):e2020MS002217, 2021.
40. Woodbury, M.A., Inverting modified matrices, *Memorandum report*, 42(106):336, 1950.
41. Higham, N.J., *Accuracy and stability of numerical algorithms*, SIAM, 2002.
42. Rougier, J., Efficient Emulators for Multivariate Deterministic Functions, *Journal of Computational and Graphical Statistics*, 17(4):827–843, 2008.
43. Lee, L.A., Reddington, C.L., and Carslaw, K.S., On the relationship between aerosol model uncertainty and radiative forcing uncertainty, *Proceedings of the National Academy of Sciences*, 113(21):5820–5827, 2016.
44. Rutt, I.C., Hagdorn, M., Hulton, N., and Payne, A., The Glimmer community ice sheet model, *Journal of Geophysical Research: Earth Surface*, 114(F2), 2009.
45. Loeppky, J.L., Sacks, J., and Welch, W.J., Choosing the sample size of a computer experiment: A practical guide, *Technometrics*, 51(4):366–376, 2009.
46. Bamber, J.L., Layberry, R.L., and Gogineni, S., A new ice thickness and bed data set for the Greenland ice sheet: 1. Measurement, data reduction, and errors, *Journal of Geophysical Research: Atmospheres*, 106(D24):33773–33780, 2001.
47. Gu, M., Palomo, J., and Berger, J.O., RobustGaSP: Robust Gaussian Stochastic Process Emulation in R, *arXiv preprint arXiv:1801.01874*, 2018.