Variational Calibration of Computer Models

Sébastien Marmin, Maurizio Filippone

Variational Calibration of Computer Models

Sébastien Marmin* Maurizio Filippone*

October 26, 2018

Abstract

Bayesian calibration of black-box computer models offers an established framework to obtain a posterior distribution over model parameters. Traditional Bayesian calibration involves the emulation of the computer model and an additive model discrepancy term using Gaussian processes; inference is then carried out using MCMC. These choices pose computational and statistical challenges and limitations, which we overcome by proposing the use of approximate Deep Gaussian processes and variational inference techniques. The result is a practical and scalable framework for calibration, which obtains competitive performance compared to the state-of-the-art.

1 INTRODUCTION

The inference of parameters of expensive computer models from data is a classical problem in Statistics (Sacks et al., 1989). Such a problem is often referred to as *calibration* (Kennedy and O'Hagan, 2001), and the results of calibration are often of interest to draw conclusions on parameters that may have a direct interpretation of real physical quantities. There are many fundamental difficulties in calibrating expensive computer models, which we can distinguish between computational and statistical. Computational issues arise from the fact that traditional optimization and inference techniques require running the expensive computer model many times for different values of the parameters, which might be unfeasible within a given computational budget. Statistical limitations, instead, arise from the fact that computer models are abstractions of real processes, which might be inaccurate.

Building on previous work from Sacks et al. (1989), in their seminal paper, Kennedy and O'Hagan (2001) propose a statistical model, based on Gaussian processes (GPs; Rasmussen and Williams (2006)), which jointly tackles the problems above. In their model, which we will refer to as KOH, the output of a deterministic computer model is emulated through a GP estimated from a set of computer experiments; in this way, computational issues are bypassed by using the prediction of the GP for a given set of parameters in place of the computer model. Observations from the real process, also known as field observations, instead, are modeled as the sum of the output of the GP emulating the computer model and another GP, which accounts for the mismatch between the computer model and the real process of interest. The KOH model is treated in a Bayesian way, making it suitable for problems where quantification of uncertainty is an important requirement. This is often the case when one is interested in drawing conclusions on parameters of interest, making predictions for decision-making with specific cost associated to predictions, or when one is interested in iteratively improve the experimental design.

While the KOH model and inference make for an attractive and elegant framework to tackle quantification of uncertainty for calibration of expensive computer models, there are a number of limitations, which we aim to overcome with this work. From the modeling perspective, GPs are indeed flexible emulators, provided that a suitable covariance function is chosen. This may limit the class of functions that can be modeled compared to alternatives from, for instance, the literature of nonstationary GPs (e.g. Paciorek and Schervish (2003)) or, more recently, of Deep GPs (DGPs; Damianou and Lawrence (2013)). From the computational perspective, limitations are inherited from the scalability issues of GPs (Rasmussen and Williams, 2006), which form the building blocks of this model, and the use of Markov chain Monte Carlo (MCMC) (Neal, 1993) techniques to carry out inference.

This work aims to tackle these issues by proposing the use of recent developments in the GP and DGP literature and variational inference, to (i) extend the modeling capabilities of GPs in the emulation using DGPs; (ii) extend the original framework in Kennedy and O'Hagan (2001), by casting the model as a special case of a DGP; (iii) use techniques based on random feature expansions and stochastic variational inference, building on the work by Cutajar et al. (2017), to obtain a scalable framework for Bayesian calibration of computer models. We extensively

*Department of Data Science, EURECOM, France

validate our proposal, which we name V-CAL, on a variety of calibration problems, comparing with alternatives from the state-of-the-art. We demonstrate the flexibility and the scalability of V-CAL, as well as the ability to capture the uncertainty in model parameters and model mismatch.

2 RELATED WORK

The framework of Kennedy and O'Hagan (2001) and related calibration methodologies (e.g., Bayarri et al. (2007); Higdon et al. (2004); Goldstein and Rougier (2004)) have been extended and developed for applications in fields as diverse as climatology (Sans et al., 2008; Salter et al., 2018), environmental sciences (Larssen et al., 2006; Arhonditsis et al., 2007), biology (Henderson et al., 2009), and mechanical engineering (Williams et al., 2006), to name a few.

More generally, Bayesian calibration had a large echo in statistical analysis of computer experiments. Among many others, Wang et al. (2009) propose a Bayesian approach where the posterior of the field observation is expressed as the sum of two independent posterior distributions derived separately, representing the discrepancy and the computer model. Storlie et al. (2015) propose a calibration approach which can handle categorical inputs within a Bayesian smoothing spline ANOVA framework.

Identifiability issues with the KOH model were pointed out in the discussion of the paper of Kennedy and O'Hagan (2001). Such issues arise due to the over-parameterization of the model, whereby it is possible to confound the effects of calibration parameters and model discrepancy. Later, more work has been done on illustrating the lack of identifiability and proposing ways to mitigate it (Arendt et al., 2012; Brynjarsdóttir and O'Hagan, 2014). In the former, it is shown that, when the computer model has multiple responses, identifiability is improved by considering them jointly with multiple output GP models. In the latter, the authors show with simple examples the crucial importance of prior information, and they discuss how eliciting this may be difficult for arbitrary black-box computer models.

Tuo and Wu (2016) propose a proof of inconsistency of a simplified version of the KOH model for Bayesian calibration. As an alternative, calibration with convergence properties is performed by minimizing a loss function involving the discrepancy. More recently, Plumlee (2017) provide analytic guarantees on the use of discrepancy priors that reduce the suboptimal broadness of the posterior.

A large literature is devoted to the practicalities of numerically challenging applications. The KOH calibration model and its inference suffer from the well known computational burden of standard GP models. Higdon et al. (2008) use basis functions representations to tackle problems with high dimensional output. Gramacy et al. (2015) use local approximate GP modeling and calibrate parameters by solving a derivative-free maximization of a likelihood term. Pratola and Higdon (2016) handle large problems using a sum-of-trees regression for modeling jointly data from computer model and the real process.

Very recent approaches combine Bayesian and discrepancy-based loss minimization providing good performance on moderate-size data sets. Gu and Wang (2018) perform calibration within a Bayesian framework, by defining a prior distribution directly on the L_2 norm of the discrepancy. Finally, Wong et al. (2017); Xie and Xu (2018) sample from the posterior distribution over calibration parameters by minimizing the L_2 norm of a sample path of the discrepancy.

3 VARIATIONAL CALIBRATION

In this section we introduce V-CAL, which is based on the KOH calibration model, random feature expansions of GPs and DGPs, and Stochastic Variational Inference (SVI). We first review the KOH model for Bayesian calibration of computer models and random feature expansions of GPs and DGPs. We then present our contribution; see the supplementary material for an illustrative example describing how V-CAL works in detail.

3.1 Background on Bayesian Calibration

Consider prediction and uncertainty analysis of a physical phenomenon approximated by a computer model (often expensive to evaluate). Observations $Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^\top \in \mathbb{R}^{n \times d_{\text{out}}}$ are made over variable inputs $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ in a given set $\mathcal{D}_1 \subset \mathbb{R}^{d_1}$. For example, in a climatological context, Y could correspond to temperature measurements with respect to the latitude and the longitude (with $d_{\text{out}} = 1$ and $\mathcal{D}_1 = [-90, 90] \times [-180, 180[$). The computer model simulating the real phenomenon requires the calibration inputs $\boldsymbol{\theta} \in \mathcal{D}_2 \subset \mathbb{R}^{d_2}$, and, given these, it is a function of $\boldsymbol{x} \in \mathcal{D}_1$. Calibration inputs $\boldsymbol{\theta}$ determine which specific application we are reproducing (e.g., exchange rates determining the carbon cycle). We use $\boldsymbol{t} \in \mathcal{D}_2$ to denote particular values of calibration inputs, and $\boldsymbol{\theta}$ to denote the

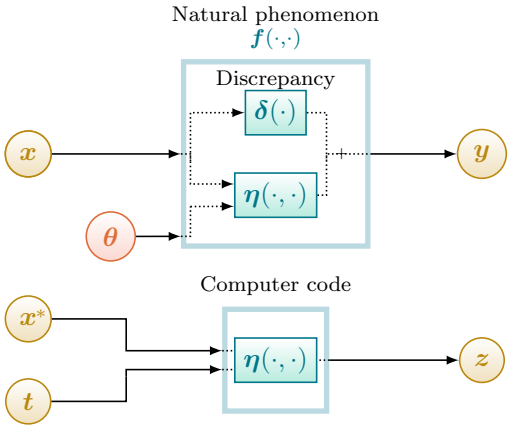

Figure 1: KOH Calibration Model.

true (unknown) values we are interested in inferring. In their Bayesian formulation, Kennedy and O'Hagan (2001) introduce a prior over $\boldsymbol{\theta}$, and they aim to characterize the posterior distribution over $\boldsymbol{\theta}$ given the data collected from the observations of the real process and runs of the computer model.

Beside the observations Y associated with X, the computer model is run at (possibly different) inputs $X^* = [\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_N^*]^\top$ with calibration inputs $T = [\boldsymbol{t}_1, \dots, \boldsymbol{t}_N]^\top$. Generally N is larger than n as running the computer model is easier (albeit computationally expensive) compared to performing a real world observation. We denote by $Z = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_N]^\top \in \mathbb{R}^{N \times d_{\text{out}}}$ the output of computer model evaluated at X^* and T.

We assume that Y and Z are drawn from some distributions $p(\boldsymbol{y}_i|\boldsymbol{f}_i)$ and $p(\boldsymbol{z}_j|\boldsymbol{\eta}_j^*)$, which determine the likelihood functions. The matrices $F = [\boldsymbol{f}_1, \dots, \boldsymbol{f}_n]^\top$ and $H^* = [\boldsymbol{\eta}_1^*, \dots, \boldsymbol{\eta}_N^*]^\top$ result from mapping $(\boldsymbol{x}_i, \boldsymbol{\theta})_{i=1,\dots,n}$ and $(\boldsymbol{x}_j^*, \boldsymbol{t}_j)_{j=1,\dots,N}$ through $\boldsymbol{f}$ and $\boldsymbol{\eta}$, respectively. The link between the computer model (with latent representation $\boldsymbol{\eta}$) and the real phenomenon (with latent representation $\boldsymbol{f}$) is modeled by

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}) = \boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{t}) + \boldsymbol{\delta}(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{\delta}$ represents the discrepancy between the computer model and the real process. Figure 1 illustrates the KOH calibration model.

In this framework, $\boldsymbol{\eta}$ and $\boldsymbol{\delta}$ are expressed as independent multidimensional GPs. In order to keep the notation uncluttered, we denote the GP covariance parameters for $\boldsymbol{\eta}$ and $\boldsymbol{\delta}$ with ψ and the input locations X, X^*, T with Ξ. The marginal likelihood is

$$\begin{aligned} p(Y, Z|\,\Xi, \psi) = \int & p(Y|\,H_{\boldsymbol{\theta}} + \Delta) p(Z|\,H^*) p(\Delta|\,X, \psi) \\ & p(H_{\boldsymbol{\theta}}, H^*|\,\boldsymbol{\theta}, \Xi, \psi) p(\boldsymbol{\theta}) \mathrm{d}H^* \,\mathrm{d}\Delta \,\mathrm{d}H_{\boldsymbol{\theta}} \,\mathrm{d}\boldsymbol{\theta}, \end{aligned}$$

where $H_{\boldsymbol{\theta}} = [\boldsymbol{\eta}(\boldsymbol{x}_1, \boldsymbol{\theta}), \dots, \boldsymbol{\eta}(\boldsymbol{x}_n, \boldsymbol{\theta})]^\top$ and $\Delta = [\boldsymbol{\delta}(\boldsymbol{x}_1), \dots, \boldsymbol{\delta}(\boldsymbol{x}_n)]^\top$. The nontrivial dependence of this integrand with respect to $\boldsymbol{\theta}$, which are the parameters of interest, makes inference over $\boldsymbol{\theta}$ intractable, and this requires approximations. We do this by employing an approximation of the GPs composing the model via random feature expansions, building on the work by Gal and Ghahramani (2016); Cutajar et al. (2017), and through variational inference techniques.

3.2 Generalized KOH Calibration Model

The original formulation of the KOH calibration model involves the use of GPs to emulate the computer model and to model the additive discrepancy. As pointed out by Kennedy and O'Hagan (2001), additive discrepancy is very specific and this formulation can be relaxed (as e.g. in Qian and Wu (2008)). We propose to do so by assuming that observations from the real process are obtained through the warping of the emulator, as follows

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}) = \boldsymbol{g}\left(\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{t}), \boldsymbol{x}\right). \tag{2}$$

Clearly, we retrieve the KOH formulation when the warping applies the identity to $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{t})$ and adds it to a GP on $\boldsymbol{x}$. We can therefore interpret the KOH model as a special case of a DGP, where the first layer implements the

emulation of the computer model and the second layer implements a particular type of warping to model the real process. Similarly to the original KOH model, its generalization allows one to reason about the mismatch between the computer model and the real process through the analysis of the warping function. We will illustrate examples in the experiments.

3.3 Random Features Expansions

In order to develop a practical and scalable calibration framework, we propose to approximate GPs using random feature expansions (Rahimi and Recht, 2008; Lázaro-Gredilla et al., 2010; Cutajar et al., 2017). This approximation turns GPs into Bayesian linear models with a set of basis functions determined by the choice of the GP covariance function, as discussed next.

Consider a GP prior with zero mean (without loss of generality) and a covariance function $k(\cdot, \cdot)$. The properties of the covariance function determine the characteristics of the GP prior. When the covariance function is Gaussian, Matérn or arc-cosine, to name a few, it is possible to show that draws from the GP prior are a linear combination of an infinite number of basis functions, with Gaussian-distributed weights (Neal, 1996; Rasmussen and Williams, 2006). In practice, denoting by $\boldsymbol{f}$ a draw from a GP evaluated at n inputs, this can be expressed as $\boldsymbol{f} = \Phi\mathbf{w}$, with $\mathbf{w} \sim \mathcal{N}(0, I)$ and infinite dimensional, and Φ the evaluations of the infinite basis functions at the n inputs. The covariance of $\boldsymbol{f}$ is readily obtained as

$$\text{cov}(\boldsymbol{f}) = \text{E}(\Phi\mathbf{w}\mathbf{w}^\top\Phi^\top) = \Phi\text{E}(\mathbf{w}\mathbf{w}^\top)\Phi^\top = \Phi\Phi^\top.$$

The idea of random feature expansions is to obtain tractable ways to truncate the infinite representation of GPs, introducing a finite set of basis functions and a finite set of weights for approximating efficiently $\boldsymbol{f}$. There are various ways to carry out the truncation; when the covariance is shift-invariant, it is possible to express the covariance as the Fourier transform of a positive measure, and this makes apparent the low-rank decomposition of the covariance considering a finite set of frequencies (Rahimi and Recht, 2008).

When using GPs in modeling problems with n observations, the truncation has the advantage of avoiding the need to solve algebraic operations with the covariance matrix, which generally cost $\mathcal{O}(n^3)$ operations. Instead, the truncation turns GPs into Bayesian linear models, for which inference can be done linearly in n and $\mathcal{O}(N_{\text{RF}}^3)$, where N_{RF} denotes the number of random features used in the truncation.

In this work, we propose to expand the two GPs $\boldsymbol{\eta}$ and $\boldsymbol{\delta}$ in the KOH calibration model using N_{RF} random features. Assuming that the GPs have a Gaussian covariance with precision A_η and A_δ and marginal variances σ_η^2 and σ_δ^2, we obtain:

$$\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{\Phi}_\eta(\Omega_\eta^{(1)}\boldsymbol{x} + \Omega_\eta^{(2)}\boldsymbol{\theta})^\top W_\eta, \tag{3}$$

$$\boldsymbol{\delta}(\boldsymbol{x}) = \boldsymbol{\Phi}_\delta(\Omega_\delta\boldsymbol{x})^\top W_\delta. \tag{4}$$

Here, the functions $\boldsymbol{\Phi}_\eta, \boldsymbol{\Phi}_\delta : \mathbb{R}^{N_{\text{RF}}} \to \mathbb{R}^{N_{\text{RF}}}$ consist in the element-wise application of sine and cosine functions, scaled by $\frac{\sigma_\eta}{\sqrt{N_{\text{RF}}/2}}$ and $\frac{\sigma_\delta}{\sqrt{N_{\text{RF}}/2}}$, respectively. The elements of W_η and W_δ, of size $N_{\text{RF}} \times d_{\text{out}}$, have i.i.d. standard normal priors, whereas the frequency matrices $\Omega_\eta = [\Omega_\eta^{(1)}, \Omega_\eta^{(2)}]$, Ω_δ, of size $N_{\text{RF}} \times (d_1 + d_2)$, $N_{\text{RF}} \times d_1$, have i.i.d. normal rows, with covariance dependent on the positive definite matrices A_η and A_δ; in particular, each row Ω_η is i.i.d. $\mathcal{N}(\mathbf{0}, A_\eta)$. Similar considerations apply to Ω_δ. Figure 2 represents the model (using a neural network-like diagram) according to Equations (1), (3), and (4).

Deep extension: It is possible to extend the proposed formulation letting $\boldsymbol{\eta}$ and/or $\boldsymbol{\delta}$ to be modeled as DGPs instead of GPs. This is straightforward, as the random feature approximation turns GPs into shallow Bayesian neural networks, so it is possible to approximate DGPs by stacking these approximate GPs, obtaining a Bayesian deep neural network. The deep extension is particularly useful when the emulator or the real process exhibit space-dependent behavior that are difficult to model by designing appropriate covariance functions. DGPs offer a way to learn such nonstationarities from data, so this is particularly appealing in such challenging applications. We will explore this possibility in the experiments.

3.4 Stochastic Variational Inference

In this work, we propose a general formulation based on variational inference techniques to approximate the posterior distribution over all model parameters, that is W_η, W_δ, and $\boldsymbol{\theta}$, noting that there might be cases where W_η, W_δ

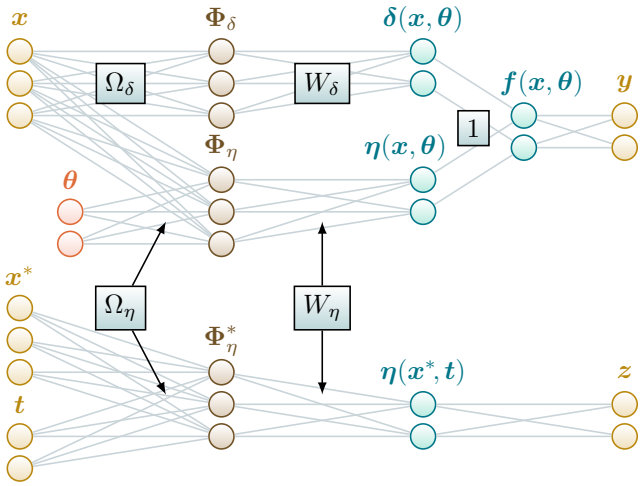

Figure 2: Neural Network representation of the proposed approximation to the KOH model. Equations (3) and (4) formulate two-layer modules for GP approximations of $\boldsymbol{\eta}$ and $\boldsymbol{\delta}$.

can be inferred analytically (Gaussian likelihoods). In particular, we introduce an approximation to the posterior $q(W_\eta, W_\delta, \boldsymbol{\theta})$, which we aim to make as close as possible to the actual posterior over these parameters. Using standard variational inference techniques, it is possible to derive a lower bound of the log-marginal likelihood as follows:

$$\mathcal{L} \geq \mathcal{E} - \mathrm{D}_{\mathrm{KL}}(q(W_\eta, W_\delta, \boldsymbol{\theta}) || p(W_\eta)p(W_\delta)p(\boldsymbol{\theta})), \tag{5}$$

where $\mathcal{E} = \mathbb{E}_{q(W_\eta, W_\delta, \boldsymbol{\theta})}[\log(p(Y, Z|\Xi, \psi, \Omega, W_\eta, W_\delta, \boldsymbol{\theta}))]$ and $\Omega = [\Omega_\eta, \Omega_\delta]$. With an expression for the lower bound of the marginal likelihood, we can now attempt to maximize it with respect to the parameters of $q(W_\eta, W_\delta, \boldsymbol{\theta})$. The lower bound contains two terms: the first ($\mathcal{E}$) is a model fitting term, whereas the second is a regularization term which penalizes approximations that deviate too much from the prior. This second D_{KL} term can be computed analytically when priors and approximate posteriors have particular forms (e.g., multivariate Gaussian).

There is an apparent complication in the fact that the first term in the lower bound depends on $q(W_\eta, W_\delta, \boldsymbol{\theta})$ through the expectation of the log-likelihood. However, this is usually bypassed by employing stochastic optimization using Monte Carlo with a finite set of samples from $q(W_\eta, W_\delta, \boldsymbol{\theta})$:

$$\mathcal{E} \approx \frac{1}{N_{\mathrm{MC}}} \sum_{i=1}^{N_{\mathrm{MC}}} \log\left(p(Y, Z|\Xi, \psi, \Omega, W_\eta^{(i)}, W_\delta^{(i)}, \boldsymbol{\theta}^{(i)})\right)$$

with $W_\eta^{(i)}, W_\delta^{(i)}, \boldsymbol{\theta}^{(i)} \sim q(W_\eta, W_\delta, \boldsymbol{\theta})$. The Monte Carlo approximation is unbiased, and so it is its derivative with respect to any of the parameters of $q(W_\eta, W_\delta, \boldsymbol{\theta})$. This means that we can employ stochastic gradient optimization to adapt the parameters of $q(W_\eta, W_\delta, \boldsymbol{\theta})$ to maximize the lower bound with guarantees to reach a local optimum of the objective (Robbins and Monro, 1951; Graves, 2011). The only precaution to take to make this viable, is to reparameterize the samples from $q(W_\eta, W_\delta, \boldsymbol{\theta})$ using the so-called reparameterization trick (Kingma and Welling, 2014); for instance, assuming a fully factorized Gaussian posterior over all parameters, the expression $\theta_\ell^{(i)} = \mu_\ell + \epsilon_\ell^{(i)}\sigma_\ell$ separates out the stochastic ($\epsilon_\ell^{(i)} \sim \mathcal{N}(0,1)$) and deterministic ($\mu_\ell$ and σ_ℓ) components in the way samples from the approximate posterior are generated. The same can be done for the other parameters of interest W_η and W_δ. In this way, the derivative with respect to the variational parameters (e.g. μ_ℓ and σ_ℓ) can be used for stochastic optimization.

Mini-batch-based learning and automatic differentiation: Part of the huge success of deep learning is due to the possibility to exploit mini-batch-based learning and automatic differentiation. The former allows to achieve scalability, as the model progressively learns by iteratively looking at subsets of data. The latter allows to tremendously simplify the implementation of complex models, as one has to implement the objective function of interest and automatic differentiation takes care of computing its derivatives based on the graph of computation and the chain rule. Traditional implementations and approximations of GPs do not allow for the use of mini-batch learning, because doing so ignores the covariance among observations, which is crucial for effective GP modeling.

The proposed GP and DGP approximation and SVI allow us to exploit mini-batch-based learning. When the likelihood factorizes across observations, the terms within the Monte Carlo approximation of $\mathcal{E}$, say $\mathcal{E}^{(i)}$, can be unbiasedly estimated by selecting m out of n terms in the set of indices $\mathcal{I}_m$ (Graves, 2011).

$$\tilde{\mathcal{E}}^{(i)} \approx \frac{n}{m} \sum_{j \in \mathcal{I}_m} \log \left(p(\boldsymbol{y}_j, \boldsymbol{z}_j | \Xi, \psi, \Omega, W_\eta^{(i)}, W_\delta^{(i)}, \boldsymbol{\theta}^{(i)}) \right)$$

This approximation introduces an extra level of stochasticity in the optimization, but it allows one to scale the learning of these models to virtually any number of observations; previous work has reported results on DGPs for 10^7 observations with a single-machine implementation (Cutajar et al., 2017).

Implementation Details Considering the large number of parameters to optimize, the learning procedure is divided in stages. We first focus on the computer model response: all parameters are fixed except the ones influencing the prediction of Z, i.e. σ_z, the means and variances of the components of W_η, the correlation length and variance of η. In the second stage all others parameters are freed for inferring Y and $\boldsymbol{\theta}$ jointly. Within each stage, we first optimize the means and variances of W, and then all parameters jointly with a smaller learning rate. The variational distributions are initialized equal to the priors.

4 EXPERIMENTS

In this section we extensively validate V-CAL. For each experiment the discrepancy structure (additive (1) or general (2)) will be specified. We will also test the model when using DGPs instead of GPs.

The experiments have the following setup. The likelihoods $p(y_i|f_i)$ and $p(z_j|\eta_j^*)$ are assumed Gaussian with variances σ_y^2 and σ_z^2 treated as hyperparmeters within ψ. All covariance functions of η and δ are Gaussian. The variational posteriors $q(W_\eta)q(W_\delta)q(\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$ are Gaussian.

4.1 Illustrative Example

We illustrate the variational calibration with one variable and one calibration input. As a first test, the prior and hyperparameters used to generate the data set are assumed to be known, with $\theta \sim \mathcal{N}(0,1)$, $\sigma_\eta = 1$, $A_\eta = \frac{1}{2}I$, $\sigma_\delta = \frac{2}{10}$, $A_\delta = \frac{1}{20}$. We choose locations for $N = 7$ computer runs and $n = 4$ observations from the real process in a space filling manner in $[0,1] \times [-\frac{5}{2}, \frac{5}{2}]$. The output vector $\boldsymbol{Z}$ of the computer model at $(\boldsymbol{x}_i^*)_{i=1,\dots,N}$ is sampled form its prior distribution. In order to determine the real observations $\boldsymbol{Y}$, we first sample $p(\theta)$ to get θ_{true}. Then the observation values are computed using Equation 1. The results of V-CAL are displayed in Figure 3. In the first and the third panels, we see that the posterior of θ obtained analytically by integrating out W_η and W_δ has its mass concentrated around the true value ≈ 0.8, where there is a (color) match between $\boldsymbol{Z}$ (the the dots) and $\boldsymbol{Y}$ (the lines). The variational posterior (blue line) offers a reasonable approximation of the true posterior.

4.2 Model Calibration in Cell Biology

We apply V-CAL to a biological application, which has been previously studied in Plumlee (2017) and Xie and Xu (2018). The output is the normalized current through ion channels of cardiac cells needed to maintain the membrane potential at -35 mV. The input variable x is the logarithm of the experiment time rescaled to $\mathcal{D}_1 = [0,1]$. The calibration inputs $\boldsymbol{\theta} \in \mathcal{D}_2 = [0,10]^3$ control a mathematical model $\eta_{\text{cell}}(\cdot, \boldsymbol{\theta})$ of the phenomenon proposed by Clancy and Rudy (1999). Here it is considered to be an expensive black box with $N = 300$ runs available, whereas the number of observations is $n = 19$. The runs are located in a space filling manner in $\mathcal{D}_1 \times \mathcal{D}_2$ (Latin hypercube sampling optimized with maximin distance criterion).

We compare V-CAL with additive and general discrepancy against four competitors. The method "L_2" is a simple minimization over $\boldsymbol{\theta}$ of the L_2 residual error $||\boldsymbol{Y} - \hat{\eta}_{\text{cell}}(\boldsymbol{X}, \boldsymbol{\theta})||$, where $\hat{\eta}$ is a surrogate model of η_{cell} given X^*, T and Z. Its minimization takes 30 seconds and the residual error is 1.31. This method is generally good for predicting observations from the real process, but it provides no quantification of uncertainty. The method KOH and PROJECTED refer to the implementation in `R` language of the KOH model and the Bayesian Projected Calibration of Xie and Xu (2018). Finally, ROBUST refers to the calibration of the `R` package `RobustCalibration` using scaled GPs (Gu and Wang, 2018).

In Table 3, we report for each method the mean squared error (MSE), $\mathbb{E}_{q(\boldsymbol{\theta})}(||\boldsymbol{Y} - \eta_{\text{cell}}(\boldsymbol{X}, \boldsymbol{\theta})||^2)$, where q represents the estimated posterior density of $\boldsymbol{\theta}$. All tuning parameters of the codes are left to default values, and

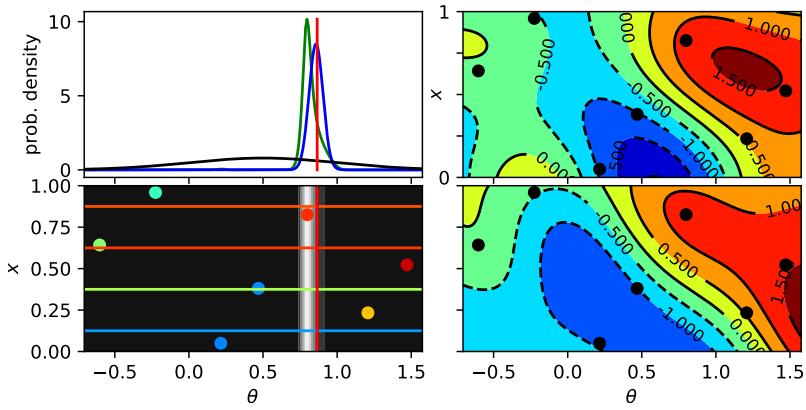

Figure 3: **Top-left:** the prior (black), the analytical posterior (green) and the variational posterior (blue) distributions of θ and the actual value used to generate $\boldsymbol{Y}$ (red). **Top-right:** the true response η used to generate the data set of this example and the locations of the computer runs (dots). **Bottom-left:** shows Y as horizontal lines in $\mathcal{D}_1 \times \mathcal{D}_2$. The color of the lines correspond to the values Y_i. The dots represent the computer runs Z. The grey level represent the posterior distribution of θ (also displayed in the top-left panel). **Bottom-right:** the posterior mean of η.

Table 1: Comparison of calibration error on the Cell Biology calibration problem.

METHOD	TIME	MSE
PROJECTED	3792	2.52
V-CAL additive	79	1.83
V-CAL general	93	1.55
KOH	3245	5.10
ROBUST	361	1.99

all methods are run on the same machine to ensure some fairness in reporting running time (laptop with 4×2.50 GHz cores). We see that the MSE values obtained by the methods PROJECTED, V-CAL and ROBUST are significantly lower than the MSE of KOH. The proposed V-CAL is the fastest among the competitors.

The posterior distributions over $\boldsymbol{\theta}$ obtained by the calibration methods we tested are reported in Figure 4. All methods yield a distribution concentrated around the L_2 minimizer (the red dot). However, the distributions are clearly not similar to each other (except for ROBUST and V-CAL). This could be explained by differences in the model formulations. Although we ensured that the covariance and mean functions are the same for all competing methods (Matérn with smoothness 5/2 and constant mean), there are several differences that cannot be matched. For instance, ROBUST has an additional step in the hierarchy of priors concerning the L_2 norm of the discrepancy. Moreover, the definition of the calibration parameters $\boldsymbol{\theta}$ itself differs among methods. In PROJECTED, $\boldsymbol{\theta}$ is a minimizer of a given stochastic process, while other methods follow the KOH definition. Also ROBUST performs a fully Bayesian inference including hyperparameters, while in the other methods, including ours, they are optimized. It would be straightforward to allow for a Bayesian treatment of the hyperparameters in V-CAL, but we leave this for future work. To visualize the results of the calibration process, in Figure 5 we overlay the observations from the real process with the responses of the computer model $\eta_{\text{cell}}(\cdot, \boldsymbol{\theta})$ when $\boldsymbol{\theta}$ is sampled from its posterior distribution. All the probabilistic method present a good fit while allowing for quantification of uncertainty in predictions, with larger uncertainty for models that account for the uncertainty in the hyperparameters. We see how the computer model output η is warped by g in V-CAL with general discrepancy (Equation (2)). In Figure 6 we display the expected derivative of the warping with respect to the computer model output, i.e. $\mathbb{E}\frac{\partial g(\cdot,x)}{\partial \eta}$, for three values of x. As the estimated values oscillate around one for every $x \in \mathcal{D}_1$, this model confirms that an additive discrepancy is a sensible assumption. When the estimated $g(\cdot, x)$ is exactly the identity, the general discrepancy boils down to an additive one. This figure also shows how the model with general discrepancy can adapt to data sets with space dependent behavior. Indeed in this test case the values of η have a very different distribution according to x. If

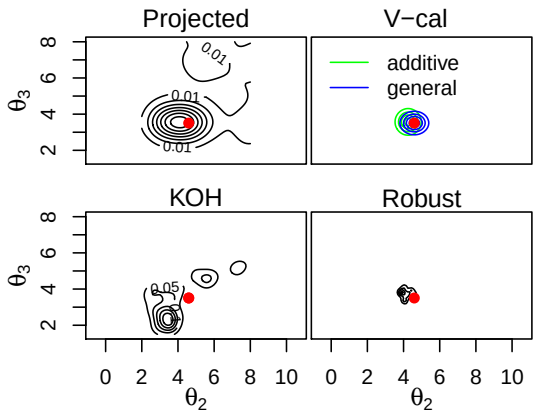

Figure 4: Posterior distributions of $\boldsymbol{\theta}$ from the calibration methods (integrated over θ_1 for visualization)

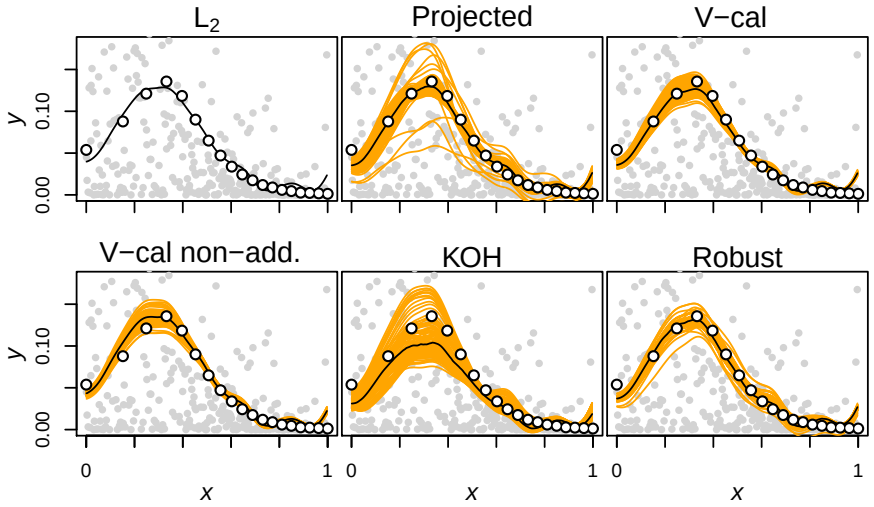

Figure 5: Samples of $\eta_{\text{cell}}(\cdot, \boldsymbol{\theta})$, with $\boldsymbol{\theta}$ drawn from its posterior. The grey dots represent the computer runs and the white dots the real observations.

x is around 0.2, the distribution of the computer runs is very asymmetric, with a heavy tail for high values, and very short for low value (see the grey dots in Figure 5). On the other hand for higher values of x, say higher than 0.6, the distribution of Z is more symmetric, and looks closer to a Gaussian distribution. This corresponds to the warping observed in Figure 6, where η gets its output warped and concentrated asymmetrically toward lower values for $x = 0.2$, while for $x = 0.6$ or 1, its Gaussian output is almost left untouched.

4.3 Model Calibration for Complex Response

We deal now with a case-study with locally non-smooth response of the computer model, for which a stationary GP is generally inadequate. The computer model is a simulator of the effects of underground nuclear tests on radionuclide diffusion into aquifers at the Yucca Flats in the United States (Fenelon, 2005). We take the same data set as generated by a script in the supplementary material of Pratola and Higdon (2016), which is available online, with $d_1 = 2$, $d_2 = 6$, $n = 10$ and $N = 17600$.

In Pratola and Higdon (2016), the size of the dataset as well as nonstationary modeling is handled with a sum-of-trees regression. We carry out calibration using V-CAL with a two-layer DGP emulator for the computer model to showcase the ability of a more complex emulator to capture the nonstationarity that characterizes this problem. We therefore compare V-CAL with a shallow GP emulator. The implementation details on the initialization can be find in the supplement. Furthermore, we compare against the modularized method with Local Approximate GPs

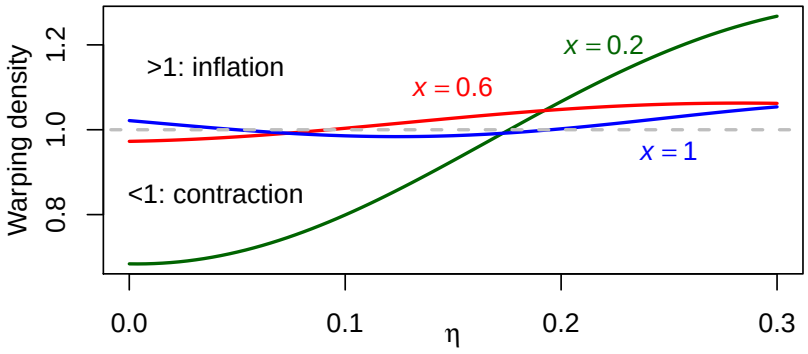

Figure 6: Derivative $\frac{\partial g(\cdot,x)}{\partial \eta}$ for three values of x

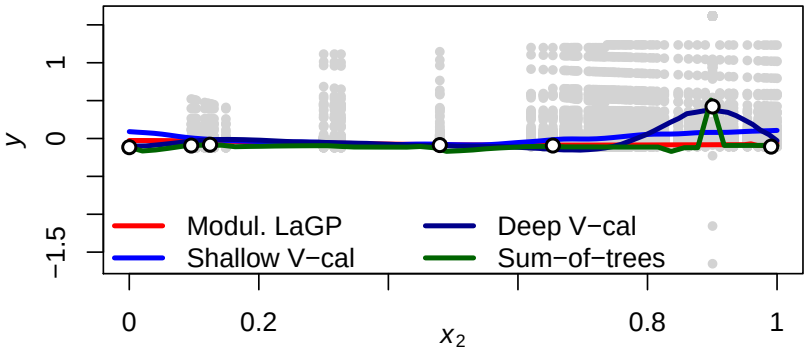

Figure 7: Mean of the posterior over the function $f(\cdot, \boldsymbol{\theta})$ modeling the real observations

(LAGP) of Gramacy et al. (2015).

In Figure 7, we display the posterior over the function $f(\cdot, \boldsymbol{\theta})$ modeling the real observations. We observe that only the deep variational calibration method and the sum-of-trees approach manage to reproduce the nonstationary nature of the data set by capturing the "spike" characterizing one of the observations.

4.4 Data Set Size Scalability

We now showcase the scalability of V-CAL to a large calibration problem in 8 dimensions with one million computer runs, and 100,000 real observations. We use the borehole function $\eta_{\text{bh}}(\boldsymbol{x}, \boldsymbol{t})$, $\boldsymbol{x} \in [0,1]^5$ $\boldsymbol{t} \in [0,1]^3$, which is a widely used function in the literature of computer experiments. The discrepancy function is a rational function (as e.g. in Gramacy et al. (2015)) (see supplement for details). We are interested in retrieving a randomly chosen true value $\boldsymbol{\theta} = [0.089, 0.308, 0.372]^\top$. The locations X, X^* and T are generated with Latin hypercube sampling. To generate $\boldsymbol{Y}$, a white Gaussian noise $\boldsymbol{\varepsilon}$ of standard deviation $\sigma_{\text{bh}} = 5 \times 10^{-3}$ is added: $y_i = \eta(\boldsymbol{x}_i, \boldsymbol{\theta}) + \delta(\boldsymbol{x}_i) + \varepsilon_i$.

We build a shallow V-CAL model with additive discrepancy as described by Equations (1)(3)(4). The covariance functions for the centered GPs η and δ are isotropic Gaussian approximated by 100 random features.

Concerning the sum-of-trees calibration, a sensible computation budget would be to set 2000 posterior samples plus 10000 for burn-in, with 1000 tree cutpoints. However, this corresponds to one month of computation on our computers, so we divided the sampling budget by 5, and set 100 cutpoints, keeping all other default parameters untouched.

We did not compare with the modularized calibration using LAGP, as the current implementation in `R` does not support large amount of real observations. This does not question the relevance of the method, which could be fixed by using a scalable GP for the discrepancy.

We evaluate the performance by comparing the posterior of $\boldsymbol{\theta}$ with the truth (Figure 8) and by evaluating the MSE error between the computer model and observation from the real process $\mathbb{E}_{q(\boldsymbol{\theta})}(||\boldsymbol{Y} - \eta_{\text{bh}}(\boldsymbol{X}, \boldsymbol{\theta})||^2)$ (Table 2). V-CAL provides the best performance both in retrieving $\boldsymbol{\theta}$ and MSE, and it is the fastest by far.

Table 2: Results of calibration on a large data set

CALIBRATION	**TIME** (h)	**MSE**
None (unif. samples on $\mathcal{D}_2$)	0	0.32
V-CAL	1.4	0.03
Sum-of-trees	132.4	0.14

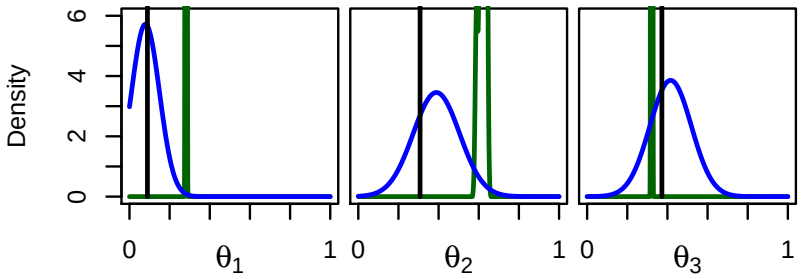

Figure 8: Posterior distribution of $\boldsymbol{\theta}$ on a large data set (black: truth, blue: V-CAL, green: sum-of-trees).

5 CONCLUSIONS

The KOH model and inference in Kennedy and O'Hagan (2001) offers a classical framework to tackle calibration problems where quantification of uncertainty is of primary interest. In this paper we proposed a number of improvements over the KOH calibration model and inference.

From the modeling perspective, we cast the KOH calibration model as a special case of a more general DGP model, where the latent process modeling the real observations is a warped version of the emulator of the computer model. In the experiments, we show that this general calibration model retains the possibility to reason about uncertainty in the mismatch between the computer model and the real process. Furthermore, the proposed approximation of GPs and DGPs with random features and inference through variational techniques gives us a number of advantages, namely the possibility to extend to the use of DGPs to emulate complex computer model outputs, and scalability as demonstrated in the experiments. Crucially, this yields an extremely practical calibration approach that can be implemented using modern development frameworks featuring automatic differentiation.

We are currently investigating the issue of non-identifiability in the context of V-CAL. We are also extending V-CAL to handle cases where the uncertainty in the model can be used to guide the incremental design of the experiment. Finally, we are investigating the application of V-CAL to other large-scale calibration problems in environmental sciences, where the KOH model and related calibration methodologies are usually not the preferred choice due to its limited scalability.

Acknowledgment

Maurizio Filippone gratefully acknowledges support from the AXA Research Fund.

References

P. D. Arendt, D. W. Apley, W. Chen, D. Lamb, and D. Gorsich. Improving identifiability in model calibration using multiple responses. *Journal of Mechanical Design*, 134(10):100909, 2012. Calibration.

G. B. Arhonditsis, S. S. Qian, C. A. Stow, E. C. Lamon, and K. H. Reckhow. Eutrophication risk assessment using bayesian calibration of process-based models: Application to a mesotrophic lake. *Ecological Modelling*, 208(2): 215 – 229, 2007. ISSN 0304-3800.

M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu. A Framework for Validation of Computer Models. *Technometrics*, 49(2):138–154, 2007. Calibration.

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

C. E. Clancy and Y. Rudy. Linking a genetic defect to its cellular phenotype in a cardiac arrhythmia. *Nature*, 400 (6744):566, 1999.

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random feature expansions for deep Gaussian processes. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 884–893, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR.

A. C. Damianou and N. D. Lawrence. Deep Gaussian Processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Proceedings*, pages 207–215. JMLR.org, 2013.

J. M. Fenelon. *Analysis of Ground-Water Levels and Associated Trends in Yucca Flat, Nevada Test Site, Nye County, Nevada, 1951-2003*. US Department ot the Interior, US Geological Survey, 2005.

Y. Gal and Z. Ghahramani. Dropout As a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059. JMLR.org, 2016.

M. Goldstein and J. Rougier. Probabilistic Formulations for Transferring Inferences from Mathematical Models to Physical Systems. *SIAM Journal on Scientific Computing*, 26(2):467–487, 2004. Calibration.

R. B. Gramacy, D. Bingham, J. P. Holloway, M. J. Grosskopf, C. C. Kuranz, E. Rutter, M. Trantham, and R. P. Drake. Calibrating a large computer experiment simulating radiative shock hydrodynamics. *Annals of Applied Statistics 2015, Vol. 9, No. 3, 1141-1168*, Nov. 2015. Calibration.

A. Graves. Practical Variational Inference for Neural Networks. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc., 2011.

M. Gu and L. Wang. Scaled Gaussian Stochastic Process for Computer Model Calibration and Prediction. *arXiv preprint arXiv:1707.08215*, May 2018.

D. A. Henderson, R. J. Boys, K. J. Krishnan, C. Lawless, and D. J. Wilkinson. Bayesian emulation and calibration of a stochastic computer model of mitochondrial dna deletions in substantia nigra neurons. *Journal of the American Statistical Association*, 104(485):76–87, 2009.

D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004. Calibration.

D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer Model Calibration Using High-Dimensional Output. *Journal of the American Statistical Association*, 103(482):570–583, 2008. Calibration.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. ISSN 1467-9868.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, Apr. 2014.

T. Larssen, R. B. Huseby, B. J. Cosby, G. Hst, T. Hgsen, and M. Aldrin. Forecasting acidification effects using a bayesian calibration and uncertainty propagation approach. *Environmental Science & Technology*, 40(24): 7841–7847, 2006. PMID: 17256536.

M. Lázaro-Gredilla, J. Quinonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse Spectrum Gaussian Process Regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, Sept. 1993.

R. M. Neal. *Bayesian Learning for Neural Networks (Lecture Notes in Statistics)*. Springer, 1 edition, Aug. 1996. ISBN 0387947248.

C. J. Paciorek and M. J. Schervish. Nonstationary covariance functions for gaussian process regression. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 273–280, Cambridge, MA, USA, 2003. MIT Press.

M. Plumlee. Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112 (519):1274–1285, 2017.

M. T. Pratola and D. M. Higdon. Bayesian Additive Regression Tree Calibration of Complex High-Dimensional Computer Models. *Technometrics*, 58(2):166–179, 2016.

P. Z. G. Qian and C. F. J. Wu. Bayesian Hierarchical Modeling for Integrating Low-Accuracy and High-Accuracy Experiments. *Technometrics*, 50(2):192–204, 2008. Calibration.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 11 1989.

J. M. Salter, D. B. Williamson, J. Scinocca, and V. Kharin. Uncertainty quantification for spatio-temporal computer models with calibration-optimal bases. *arXiv preprint arXiv:1801.08184*, 2018. Calibration.

B. Sans, C. E. Forest, and D. Zantedeschi. Inferring climate system properties using a computer model. *Bayesian Analysis*, 3(1):1–37, 03 2008.

C. B. Storlie, W. A. Lane, E. M. Ryan, J. R. Gattiker, and D. M. Higdon. Calibration of Computational Models With Categorical Parameters and Correlated Outputs via Bayesian Smoothing Spline ANOVA. *Journal of the American Statistical Association*, 110(509):68–82, 2015. Calibration.

R. Tuo and C. F. J. Wu. A Theoretical Framework for Calibration in Computer Models: Parametrization, Estimation and Convergence Properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795, 2016.

S. Wang, W. Chen, and K.-L. Tsui. Bayesian Validation of Computer Models. *Technometrics*, 51(4):439–451, 2009. Calibration.

B. Williams, D. Higdon, J. Gattiker, L. Moore, M. McKay, and S. Keller-McNulty. Combining experimental data and computer simulations, with an application to flyer plate experiments. *Bayesian Analysis*, 1(4):765–792, 12 2006.

R. K. Wong, C. B. Storlie, and T. C. Lee. A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):635–648, 2017.

F. Xie and Y. Xu. Bayesian Projected Calibration of Computer Models. *arXiv preprint arXiv:1803.01231*, Mar. 2018.

Appendix

A Borehole objective and discrepancy functions

The objective function used in section 4.4 is defined for all $\boldsymbol{x} \in [0,1]^5$ and $\boldsymbol{t} \in [0,1]^3$, with

$$\eta_{\text{bh}}(\boldsymbol{x}, \boldsymbol{t}) = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w)\left(1 + \frac{2LT_u}{\ln(r/r_w) r_w^2 K_w} + \frac{T_u}{T_l}\right)},$$
$$\delta_{\text{bh}}(\boldsymbol{x}) = \frac{2(10x_1^2 + 4x_2^2)}{50x_1x_2 + 10},$$

with $T_u = x_1(115600-63070)+63070$, $H_u = x_2(1110-990)+990$, $H_l = x_3(820-700)+700$, $L = x_4(1680-1120)+1120$, $K_w = x_5(12045-9855)+9855$, $r_w = t_1(0.15-0.05)+0.05$, $r = t_2(50000-100)+100$, $T_l = t_3(116-63.1)+63.1$.

B Initialization of variational parameters in v-cal for the Radionuclide model

We list in the following table the settings of the v-cal models used for the experiments of the section 4.3.

Table 3: Radionuclide Model: Initial Values for the v-cal models

PARAM.	**SHALLOW**	**DEEP**
$\mathbb{E}_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})$	$\frac{1}{2}[1,1,1]^\top$	$\frac{1}{2}[1,1,1]^\top$
$\text{var}_{q(\boldsymbol{\theta})}(\boldsymbol{\theta})$	$\frac{1}{4}[1,1,1]^\top$	$\frac{1}{4}[1,1,1]^\top$
σ_y	10^{-2}	10^{-2}
σ_z	10^{-3}	10^{-3}
A_η, A_δ	$20I$	$20I$
σ_η, σ_δ	1, $\frac{1}{10}$	1, $\frac{1}{10}$
$A_{\text{layer } i}$	$\emptyset$	$2I_{d_1+d_2}$
$\sigma_{\text{layer } i}$	$\emptyset$	1