A review on computer model calibration

Chih-Li Sung*[1] and Rui Tuo[2]

[1]Department of Statistics and Probability, Michigan State University
[2]Department of Industrial & Systems Engineering and Department of Statistics, Texas A&M University

Abstract

Model calibration is crucial for optimizing the performance of complex computer models across various disciplines. In the era of Industry 4.0, symbolizing rapid technological advancement through the integration of advanced digital technologies into industrial processes, model calibration plays a key role in advancing digital twin technology, ensuring alignment between digital representations and real-world systems. This comprehensive review focuses on the Kennedy and O'Hagan (2001) (KOH) framework. In particular, we explore recent advancements addressing the challenges of the unidentifiability issue while accommodating model inadequacy within the KOH framework. In addition, we explore recent advancements in adapting the KOH framework to complex scenarios, including those involving multivariate outputs and functional calibration parameters. We also delve into experimental design strategies tailored to the unique demands of model calibration. By offering a comprehensive analysis of the KOH approach and its diverse applications, this review serves as a valuable resource for researchers and practitioners aiming to enhance the accuracy and reliability of their computer models.

Keywords: Unidentifiability, Gaussian Process, Uncertainty Quantification, Computer Experiments, Experimental Design.

*corresponding author. Email: sungchih@msu.edu

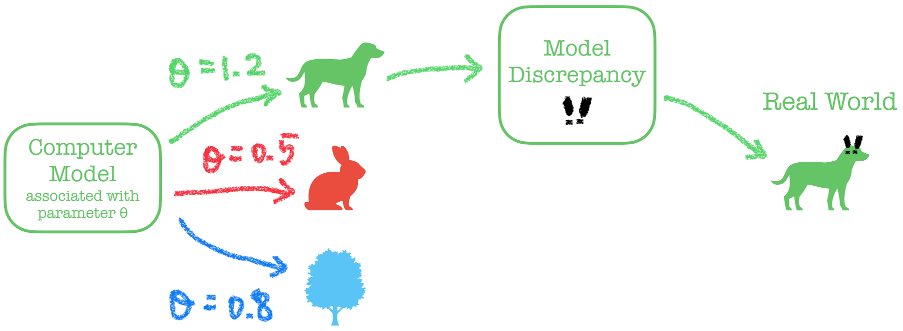

Figure 1: The illustration showcases the model calibration process, particularly within the Kennedy-O'Hagan (KOH) Bayesian calibration framework. The KOH approach serves as a reliable tool for enhancing the accuracy and reliability of computer models, especially in the realm of digital twins for Industry 4.0. Digital twins refer to virtual replicas of physical systems, integrating real-time data for improved performance and decision-making in Industry 4.0 applications.

1 Introduction

With the advent of sophisticated computer modeling, various fields, ranging from engineering to epidemiology, have leveraged these models to simulate complex real-world phenomena. These computer models serve as crucial tools for understanding intricate systems, predicting behaviors, and making informed decisions. However, the effectiveness of these models often hinges on the precise calibration of their underlying parameters, ensuring that their outputs align with observed real-world data.

Model calibration, a fundamental process in computational science and engineering, involves the adjustment of model parameters to optimize the model's performance and enhance its predictive capabilities. These model parameters are often referred to as *calibration parameters*. Through calibration, researchers can fine-tune these models to accurately capture the intricacies of the systems they represent. Notably, the calibration process plays a pivotal role in fields such as nuclear physics (Higdon et al., 2015; Pratola and Higdon, 2016; King et al., 2019; Kejzlar et al., 2020), biology (Henderson et al., 2009; Sung et al., 2020, 2022), environmental sciences (Larssen et al., 2006; Cheng et al., 2021), climatology (Konomi et al., 2017; Forest et al., 2008; Higdon et al., 2013; Salter et al., 2019; Lee et al., 2020), hydrology (Goh et al., 2013; Gramacy et al., 2015; Pratola and Chkrebtii, 2018), manufacturing (Wang et al., 2020), epidemiology (Farah et al., 2014; Wang et al., 2022; Sung and Hung, 2024), health care (Oakley and Youngman, 2017), mechanical engineering (Gattiker et al., 2006), aerospace (Allaire et al., 2012; Huang et al., 2020; Zhou et al., 2023), material science (Generale et al., 2022), transfer learning (Liyanage et al., 2022), robotics (Liu and Negrut, 2021), and digital twins (Kenett and Bortman, 2022; Thelen et al., 2022, 2023), where accurate predictions are essential for informed decision-making.

For instance, in climate science, the calibration of climate models enables scientists to simulate past and present climate conditions accurately, which is crucial for predicting future climate scenarios and assessing the potential impact of climate change. Similarly, in epidemiology, the calibration of disease transmission models helps in understanding the spread of infectious diseases and devising effective strategies for disease control and prevention.

In this paper, we delve into the methodologies of model calibration, with a specific focus on the Bayesian calibration method proposed by KOH (Kennedy and O'Hagan, 2001), which accounts for all sources of uncertainty when using the computer model subsequently for prediction. Figure 1 provides an illustration of the KOH framework.

The concept of model calibration can be traced back to the advent of physics-based models, where statistical methods using nonlinear least squares have long been used to estimate unknown parameters (see, e.g., Box and Hunter (1962)). Over time, more sophisticated approaches emerged, including the generalized likelihood uncertainty estimation method (Romanowicz et al., 1994), the use of Gaussian process models as cost-effective emulators for expensive computer codes, enhancing the efficiency of the parameter search for optimal calibration (Cox et al., 1992; Craig et al., 1996, 2001), and the Bayesian synthesis method (Raftery et al., 1995), and the subsequent Bayesian melding approach (Poole and Raftery, 2000). Despite the notable advancements, the KOH approach (Kennedy and O'Hagan, 2001), the pioneering method to account for all sources of uncertainty, prominently accounted for the critical uncertainty from *model inadequacy*, particularly relevant when models lack the detail necessary to differentiate between distinct conditions leading to varying process values. Its profound impact resonated across diverse fields reliant on computer models, notably influencing the domains of computer model validation (Bayarri et al., 2007a,b; Wang et al., 2009). Despite the comprehensive integration of model inadequacy, the inherent flexibility of the KOH model often renders parameter estimation *unidentifiable*, a dilemma extensively discussed in subsequent literature. Tuo and Wu (2016) notably presented the first theoretical description of this problem and proposed an alternative methodology in Tuo and Wu (2015) aimed at mitigating the issue of unidentifiability. A multitude of subsequent approaches have since been proposed to address this predicament, showcasing the enduring pursuit of enhancing the robustness and applicability of the calibration process.

Distinguished from existing reviews (e.g., Campbell (2006), Xiong et al. (2009), and Baker et al. (2022)) and textbooks (e.g., Santner et al. (2018), Fang et al. (2005), and Gramacy (2020)) within the realm of computer experiments, our discussion delves deeply into recent advances aimed at addressing this unidentifiability issue. We investigate the theoretical underpinnings that support these advancements, considering practical implications and challenges. Moreover, we explore the latest developments in experimental design strategies tailored to the unique demands of model calibration. Furthermore, we highlight the recent applications of the KOH approach, emphasizing its adaptability to complex scenarios such as those involving heteroscedastic outputs, multivariate outputs, and functional calibration parameters.

It is worth noting that, besides KOH model calibration, another important approach in shaping computer/mathematical modeling is the *physics-informed machine learning* (Karniadakis et al., 2021), which has garnered increasing attention and has immediate impacts in engineering and science due to its potential for reducing computational costs and enhancing modeling flexibility. Such physics-informed machine learning integrates data and mathematical models through deep neural networks or other kernel-based regression networks by enforcing fundamental physical laws. For an extensive survey encompassing both model calibration and physics-informed machine learning, along with their intersection, we refer to Viana and Subramaniyan (2021).

The rest of this article is organized as follows. We begin by outlining the problem setting of model calibration in Section 2. In Section 3, we introduce the KOH approach, providing an in-depth understanding of its key components and principles. Section 4 delves into the posterior inference of the KOH model. We then discuss various other calibration approaches in Section 5. Section 6 focuses on the unidentifiability issue, exploring different strategies and solutions proposed to mitigate this challenge. Recent developments on experimental designs in the context of the calibration problem are provided in Section 7. Recent applications of the KOH approach in complex scenarios are discussed in Section 8. Finally, we conclude our discussion in Section 9.

2 Problem Setting

In this section, we elucidate the objectives and components of model calibration. We begin by delineating two distinct sources of data: physical data and computer simulations.

2.1 Physical Data

Consider a real system with d control variables represented as $\mathbf{x}$, where the input space is $\Omega \subseteq \mathbb{R}^d$, i.e., $\mathbf{x} \in \Omega$. We collect a set of n physical data points from a real system, forming input-output pairs $\{(\mathbf{x}_i^p, y_i^p)\}_{i=1}^n$, denoted as $\mathcal{D}_n^p$. The mean process of the output y_i^p, denoted as $\zeta(\mathbf{x}_i^p)$ (also called *true process* in Tuo and Wu (2015, 2016)), is subject to the noise ϵ_i. We formulate this relationship using the following statistical model:

$$y_i^p = \zeta(\mathbf{x}_i^p) + \epsilon_i,$$

where ϵ_i follows an independent and identically distributed (i.i.d.) zero-mean normal distribution with variance σ_ϵ^2, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. For addressing the heteroscedastic assumption concerning σ_ϵ^2, refer to Section 8.1.2. One of the primary objectives here is to estimate the mean process $\zeta(\mathbf{x})$.

2.2 Computer Model

While utilizing the physical data $\mathcal{D}^p$ is one approach, computer models have shown their effectiveness in improving the predictive performance of $\zeta(\mathbf{x})$ when integrated with them.

Consider the scenario where there exists a computer model that can simulate the real system. In the computer model simulations, there are typically $p + q$ control variables. The first p control variables $\mathbf{x}$ coincide with those of the physical system, while the remaining q variables are often referred to as *tuning/calibration parameters*, denoted as $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^q$, where Θ denotes the parameter space. These parameters are typically unobservable in a physical system but are of significant interest in practical applications. For instance, in the implosion simulations of Higdon et al. (2008), the yield stress of steel and the resulting detonation energy are the calibration parameters that are not measurable in physical experiments.

For deterministic computer simulations (i.e., simulations yielding identical outputs for a given input), we express the relationship between the model output, denoted as y^s, and the input using the function f, i.e.,

$$y^s = f(\mathbf{x}, \boldsymbol{\theta}). \tag{1}$$

In the case of stochastic computer simulations, the relationship between the model output and input is expressed as follows:

$$y^s = f(\mathbf{x}, \boldsymbol{\theta}) + v, \quad v \sim \mathcal{N}(0, r(\mathbf{x}, \boldsymbol{\theta})),$$

where $r(\mathbf{x}, \boldsymbol{\theta})$ is the variance of the random error v, which depends on the input $(\mathbf{x}, \boldsymbol{\theta})$ but constant variance is also possible. For a comprehensive review of stochastic computer simulations, refer to Baker et al. (2022). Hereafter, we will primarily focus on deterministic computer simulations and defer the extension to stochastic computer simulations to the interested reader.

Computer models can vary in their computational demands, with some being quick to simulate while others are computationally intensive. For computationally intensive computer simulations, it is often necessary to conduct a computer experiment involving several evaluations of f (i.e., running simulations) to learn a statistical model. The process is called *emulation*, and this statistical model, referred to as an *emulator* or *surrogate model*, provides a cost-effective alternative to the actual simulator f. The Gaussian process (GP) model is a widely employed choice for modeling f in computer experiments (Santner et al., 2018; Rasmussen and Williams, 2006; Gramacy, 2020), with roots dating back to the pioneering work by Sacks et al. (1989). Details regarding GP models are deferred to Section 3.2.

2.3 Imperfect Computer Model

The main objective of model calibration is to identify the optimal set of $\boldsymbol{\theta}$ such that the computer model output f closely approximates the true process ζ, i.e., $f(\mathbf{x}, \boldsymbol{\theta}^*) \approx \zeta(\mathbf{x})$ for any $\mathbf{x} \in \Omega$, with $\boldsymbol{\theta}^*$ being the optimal values. This pursuit ultimately aims to improve the predictive accuracy of ζ by leveraging the computer model f. In essence, calibration represents a statistical *inverse problem*. However, an important uncertainty arises in the form of *model discrepancy*, where the model output may not perfectly match the true process. In other words, finding an optimal $\boldsymbol{\theta}^*$ such that $f(\mathbf{x}, \boldsymbol{\theta}^*) = \zeta(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ may not be feasible. Such a model is termed an *imperfect* or *inexact* computer model. For such imperfect computer models, our main objectives are twofold: first, to accurately estimate the parameter $\boldsymbol{\theta}^*$ that optimally aligns the computer model with the true process, and second, to predict $\zeta(\mathbf{x})$ with precision.

In the next section, we will explore a popular model proposed by KOH that addresses this challenge.

3 Main Approach: Bayesian Calibration

In this section, we delve into a prominent statistical model designed to address model calibration problems. While various popular approaches exist for estimating calibration parameters (e.g., history matching and least squares), our focus in this section centers on the method proposed by KOH. This method strives to achieve two key objectives: parameter estimation and accurate true process prediction, especially when dealing with imperfect computer models. For other methods primarily focusing on estimation, see Section 5.

3.1 KOH Calibration: A Bayesian framework

One of the most widely adopted approaches for model calibration is the Bayesian calibration method introduced by KOH. This statistical model establishes a connection between the physical data and the computer model through the following relationship:

$$\zeta(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x}), \tag{2}$$

or equivalently,

$$y_i^p = f(\mathbf{x}_i^p, \boldsymbol{\theta}) + \delta(\mathbf{x}_i^p) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{3}$$

where $\delta(\mathbf{x})$ is referred to as the *model bias*, *discrepancy*, or *inadequacy*, acknowledging the potential imperfections in the computer model. It is worth noting that while the original KOH paper introduced an unknown regression parameter ρ and assumed $\zeta(\mathbf{x}) = \rho f(\mathbf{x}, \boldsymbol{\theta}) + \delta(\mathbf{x})$, most subsequent research papers, including this one, often omit the parameter ρ and simply assume $\rho = 1$.

KOH employs a Bayesian framework to make inferences about the unknown calibration parameter $\boldsymbol{\theta}$ and the true process $\zeta(\cdot)$. Specifically, this Bayesian framework places a Gaussian process prior on the discrepancy function δ and employs Markov Chain Monte Carlo (MCMC) sampling to infer the posterior distributions of both $\boldsymbol{\theta}$ and ζ. Further details regarding the inference process will be discussed in Section 4.

3.2 Gaussian Process (GP) Prior

Before delving into the inference process, we provide a brief introduction to Gaussian Processes (GPs), which are utilized to model the discrepancy function δ and, when necessary, the computer model f for emulating expensive computer simulations.

A GP prior is a flexible and widely used approach for expressing prior information about functions in a Bayesian framework, with applications ranging from machine learning (Rasmussen and Williams, 2006) to spatial statistics (Stein, 1999). We denote the GP prior as:

$$\delta(\mathbf{x}) \sim \mathcal{GP}(\mu_\delta(\mathbf{x}), \tau_\delta \Phi_\delta(\mathbf{x}, \mathbf{x}')), \tag{4}$$

where $\mu_\delta(\mathbf{x})$ represents the mean function, $\tau_\delta > 0$ is a positive scalar, and $\Phi_\delta(\mathbf{x}, \mathbf{x}')$ is a positive-definite correlation function with $\Phi_\delta(\mathbf{x}, \mathbf{x}) = 1$ for any $\mathbf{x} \in \Omega$. While KOH utilized a linear mean for the mean function $\mu_\delta(\mathbf{x})$, throughout this article, we assume a constant mean function for the sake of simplicity, i.e., $\mu_\delta(\mathbf{x}) \equiv \mu_\delta$.

For correlation functions, squared exponential kernels and Matérn kernels (Stein, 1999) are popular choices. The (anisotropic) squared exponential kernels have the form,

$$\Phi_\delta(\mathbf{x}, \mathbf{x}') = \exp\left(-\sum_{j=1}^{d} \gamma_j \left(x_j - x_j'\right)^2\right),$$

and the Matérn kernels have the form

$$\Phi_\delta(\mathbf{x}, \mathbf{x}') = \frac{1}{\Gamma(\nu) 2^{\nu-1}} (2\sqrt{\nu} \|\boldsymbol{\gamma} \odot (\mathbf{x} - \mathbf{x}')\|_2)^\nu \times B_\nu (2\sqrt{\nu} \|\boldsymbol{\gamma} \odot (\mathbf{x} - \mathbf{x}')\|_2),$$

where $\odot$ is the Hadamard product, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_d)$ is a lengthscale parameter representing a vector of size d, $\|\cdot\|_2$ denotes the Euclidean norm, B_ν is the modified Bessel function of the second kind, and ν represents a smoothness parameter.

The realizations of $\delta(\mathbf{x})$ then follow a multivariate normal distribution, given by:

$$(\delta(\mathbf{x}_1), \ldots, \delta(\mathbf{x}_n))^T \sim \mathcal{N}_n(\mu_\delta \mathbf{1}_n, \tau_\delta \boldsymbol{\Phi}_\delta(\boldsymbol{\gamma})), \tag{5}$$

where $\mathbf{1}_n$ is a vector of ones of size n, and $\boldsymbol{\Phi}_\delta(\boldsymbol{\gamma})$ is an $n \times n$ correlation matrix, with each element $(\boldsymbol{\Phi}_\delta(\boldsymbol{\gamma}))_{i,j} = \Phi_\delta(\mathbf{x}, \mathbf{x}')$ for any $1 \leq i, j \leq n$.

4 Posterior Inference

This section discusses the main inference approaches used to obtain the posterior distributions of $\boldsymbol{\theta}$ and ζ. We begin by discussing scenarios where inexpensive computer simulations are available in Section 4.1, and then delve into cases where expensive computer simulations are involved in Section 4.2.

4.1 Inexpensive Computer Simulation

Consider a scenario where evaluating the computer model is fast, as studied in Higdon et al. (2004), referred to as *unlimited simulation runs*. Denote $\mathbf{y}^p = (y_1^p, \dots, y_n^p)^T$ and $\mathbf{f}(\boldsymbol{\theta}) = (f(\mathbf{x}_1^p, \boldsymbol{\theta}), \dots, f(\mathbf{x}_n^p, \boldsymbol{\theta}))^T$. By (3) and (5), it follows that

$$(\mathbf{y}^p - \mathbf{f}(\boldsymbol{\theta})) \sim \mathcal{N}_n(\mu_\delta \mathbf{1}_n, \tau_\delta \boldsymbol{\Phi}_\delta(\boldsymbol{\gamma}) + \sigma_\epsilon^2 \mathbf{I}_n),$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix. The likelihood of the unknown parameters $\boldsymbol{\theta}$, τ_δ, $\mu_\delta, \sigma_\epsilon^2$, and $\boldsymbol{\gamma}_\delta$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}, \tau_\delta, \mu_\delta, \boldsymbol{\gamma}_\delta, \sigma_\epsilon^2 | \mathcal{D}^p) \propto & \, |\tau_\delta \boldsymbol{\Phi}_\delta(\boldsymbol{\gamma}) + \sigma_\epsilon^2 \mathbf{I}_n|^{-1/2} \\ & \exp\left\{ -\frac{1}{2} (\mathbf{y}^p - \mathbf{f}(\boldsymbol{\theta}) - \mu_\delta \mathbf{1}_n)^T (\tau_\delta \boldsymbol{\Phi}_\delta(\boldsymbol{\gamma}) + \sigma_\epsilon^2 \mathbf{I}_n)^{-1} (\mathbf{y}^p - \mathbf{f}(\boldsymbol{\theta}) - \mu_\delta \mathbf{1}_n) \right\}. \end{aligned} \tag{6}$$

Thus, the parameters, including the calibration parameter $\boldsymbol{\theta}$, can be estimated via maximum likelihood estimation (MLE) or Bayesian inference, such as a Metropolis–Hastings (MH) style MCMC method (Gilks et al., 1995).

The posterior of the mean process $\zeta(\mathbf{x})$, conditional on the parameters $\boldsymbol{\theta}, \tau_\delta, \mu_\delta, \sigma_\epsilon^2, \boldsymbol{\gamma}$, and the data $\mathcal{D}^p$, can be derived using the property of conditional multivariate normal distributions, which follows a normal distribution:

$$\zeta(\mathbf{x}) | \mathcal{D}^p, \boldsymbol{\theta}, \tau_\delta, \mu_\delta, \sigma_\epsilon^2, \boldsymbol{\gamma} \sim \mathcal{N}(m(\mathbf{x}), \sigma^2(\mathbf{x})),$$

where

$$m(\mathbf{x}) = (\mu_\delta \mathbf{1}_n + \mathbf{f}(\boldsymbol{\theta})) + \Phi_\delta(\mathbf{x}, X_n^p; \boldsymbol{\gamma}) \left(\boldsymbol{\Phi}_\delta(\boldsymbol{\gamma}) + \frac{\sigma_\epsilon^2}{\tau_\delta} \mathbf{I}_n \right)^{-1} (\mathbf{y}^p - \mathbf{f}(\boldsymbol{\theta}) - \mu_\delta \mathbf{1}_n)$$

and

$$\sigma^2(\mathbf{x}) = \tau_\delta - \tau_\delta \Phi_\delta(\mathbf{x}, X_n^p; \boldsymbol{\gamma}) \left(\boldsymbol{\Phi}_\delta(\boldsymbol{\gamma}) + \frac{\sigma_\epsilon^2}{\tau_\delta} \mathbf{I}_n \right)^{-1} \Phi_\delta(\mathbf{x}, X_n^p; \boldsymbol{\gamma}), \tag{7}$$

where $X_n^p = (\mathbf{x}_1^p, \dots, \mathbf{x}_n^p)$, and $\Phi_\delta(\mathbf{x}, X_n^p; \boldsymbol{\gamma})$ is a vector of length n with each element $(\Phi_\delta(\mathbf{x}, X_n^p; \boldsymbol{\gamma}))_i = \Phi_\delta(\mathbf{x}, \mathbf{x}_i^p)$. The posterior $\zeta(\mathbf{x})|\mathcal{D}^p$ can be obtained either through MCMC sampling to integrate the parameters $\boldsymbol{\theta}$, τ_δ, μ_δ, σ_ϵ^2, $\boldsymbol{\gamma}$, or by plugging in their MLE estimates.

4.2 Expensive Computer Simulation

In the original KOH approach, the scenario involves evaluating the computer model f being prohibitively expensive, making it necessary to construct an emulator, a common practice in many applications (see, e.g., Mak et al. (2018)). They assume f follows a GP prior and is independent of δ, i.e.,

$$f(\mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{GP}(\mu_f, \tau_f^2 \Phi_f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'))),$$

where μ_f and τ_f^2 are similarly denoted as in (4). Suppose that Φ_f is the squared exponential kernel, then it has the form

$$\Phi_f((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}')) = \exp\left(-\sum_{j=1}^{d} \omega_{j,1} \left(x_j - x_j'\right)^2 - \sum_{j=1}^{q} \omega_{j,2} \left(\theta_j - \theta_j'\right)^2 \right),$$

where $\boldsymbol{\omega} = (\omega_{1,1}, \dots, \omega_{d,1}, \omega_{1,2}, \dots, \omega_{q,2})$. After collecting the evaluations from f at N inputs from the computer simulations, where typically the N input locations are designed to be space-filling in the input space, such as a uniform or Latin hypercube design (Morris and Mitchell, 1995), the input-output pairs are denoted by $\mathcal{D}_N^s = \{((\mathbf{x}_i^s, \boldsymbol{\theta}_i^s), y_i^s)\}_{i=1}^N$ with $y_i^s = f(\mathbf{x}_i^s, \boldsymbol{\theta}_i^s)$. Then, the computer model outputs follow a multivariate normal distribution,

$$(y_1^s, \dots, y_N^s)^T \sim \mathcal{N}_N(\mu_f \mathbf{1}_N, \tau^2 \boldsymbol{\Phi}_f(\boldsymbol{\omega})),$$

where $\boldsymbol{\Phi}_f(\boldsymbol{\omega})$ is an $N \times N$ matrix with each element $(\boldsymbol{\Phi}_f(\boldsymbol{\omega}))_{i,j} = \Phi_f((\mathbf{x}_i, \boldsymbol{\theta}_i), (\mathbf{x}_j, \boldsymbol{\theta}_j))$. The posterior inference for all the parameters $\boldsymbol{\theta}$, τ_δ, μ_δ, $\boldsymbol{\gamma}_\delta$ can be carried out using the full corpus of data from computer simulation data $\mathcal{D}^s$ and physical data $\mathcal{D}^p$, as done by KOH. The posterior of the mean process $\zeta(\mathbf{x})$, conditional on the parameters and the data $\mathcal{D}^s$ and $\mathcal{D}^p$, can be derived similarly to the previous subsection. For further details on the MH sampler for the posteriors in the KOH calibration setting, see Chapter 8.1.1 of Gramacy (2020). Available R packages for the Bayesian approach include CaliCo (Carmassi, 2018) and BACCO (Hankin, 2005).

Fully Bayesian KOH calibration could be computationally intractable, especially when both sample sizes of $\mathcal{D}^p$ and $\mathcal{D}^s$ are large (more approaches to these scenarios will be discussed in Section 8.4). This might lead to parameter/process identification and confounding because coupled δ and f using GP priors could cause MCMC mixing issues. To address this, Bayarri et al. (2007b) and Liu et al. (2009) employed an approach called *modularization*. Instead of coupling the data $\mathcal{D}^s$ and $\mathcal{D}^p$ to infer the posterior, they suggested first training the emulator, $\hat{f}(\mathbf{x}, \boldsymbol{\theta})$, using the predictive mean of the GP posterior based on the simulation data $\mathcal{D}^s$, and then inferring the posterior as done previously by replacing f with $\hat{f}$. On the other hand, Kejzlar et al. (2021) used an empirical Bayes approach, wherein instead of placing a prior distribution on the unknown parameters, including $\boldsymbol{\theta}$, σ_ϵ^2, $\boldsymbol{\gamma}$, $\boldsymbol{\omega}$, μ_δ, and μ_f, the method estimates these parameters directly from the data. To enhance the efficiency of the MCMC methods, Rumsey and Huerta (2021) employed the eigenvalue decomposition to approximate the inverse of the covariance matrix in the likelihood (6), i.e., $(\tau_\delta \boldsymbol{\Phi}_\delta(\boldsymbol{\gamma}) + \sigma^2 \mathbf{I}_n)^{-1}$, which can be computed in nearly quadratic time. Furthermore, Kejzlar and Maiti (2023) used variational Bayes inference (Blei et al., 2017), an alternative Bayesian inference to MCMC, which has been widely used to approximate the posterior distribution through optimization as it tends to be faster and easier to scale to massive datasets.

5 Other Calibration Approaches

Apart from the widely used KOH approach, several other common methods for calibration parameter estimation are discussed in this section.

The ordinary least squares (OLS) estimator, widely applied in calibration problems (see, e.g., Anastassopoulou et al. (2020) and Mahnken (2017)), minimizes the squared difference between physical outputs and simulation outputs, that is,

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \sum_{i=1}^n (y_i^p - f(\mathbf{x}_i^p, \boldsymbol{\theta}))^2. \tag{8}$$

In cases where the computer model f is computationally expensive, an emulator can be constructed using a computer experiment, as discussed in Section 4.2.

History matching (or Bayesian history matching) is an alternative method to KOH calibration, as seen in Craig et al. (1996); Vernon et al. (2014); Williamson et al. (2013); Boukouvalas et al.

(2014); Andrianakis et al. (2015, 2017). The fundamental concept involves utilizing physical outputs to eliminate "implausible" parameter settings. The implausibility measure for $\boldsymbol{\theta} \in \Theta$ is defined (in scenarios with expensive computer simulations where an emulator is necessary) as:

$$I(\boldsymbol{\theta}) = \left(\mathbf{y}^p - \mathbb{E}[\mathbf{f}(\boldsymbol{\theta})]\right)^T \left(\mathbb{V}[\mathbf{y}^p - \mathbb{E}[\mathbf{f}(\boldsymbol{\theta})]]\right)^{-1} \left(\mathbf{y}^p - \mathbb{E}[\mathbf{f}(\boldsymbol{\theta})]\right).$$

If $I(\boldsymbol{\theta})$ exceeds a threshold, the corresponding $\boldsymbol{\theta}$ value is deemed "implausible" and discarded. The threshold setting is problem-specific, as detailed in Pukelsheim (1994); Craig et al. (1996); Williamson et al. (2015); Vernon et al. (2010). History matching's iterative nature involves selecting new computer model simulations to enhance the emulator and calibration, based on preliminary history matching results, often termed "waves" (Salter and Williamson, 2016; Williamson et al., 2017; Salter et al., 2019).

Other methods, adapted from diverse contexts to the calibration setting, include Approximate Bayesian Computation (ABC) methods (Tavaré et al., 1997; Pritchard et al., 1999), which simulate draws from the posterior distribution by approximating likelihood-based algorithms. These methods find applications in calibration problems, as demonstrated in McKinley et al. (2018); Rutter et al. (2019). In addition, Pratola et al. (2013) propose a new criterion, reminiscent of the Kullback–Liebler (KL) divergence (Kullback, 1997), for estimating the calibration parameters. Furthermore, Frenklach et al. (2016) apply the Bound-to-Bound Data Collaboration (BOB) methodology to determine feasible bounds for $\boldsymbol{\theta}$, an optimization-based framework for combining models and experimental data from multiple sources (Feeley et al., 2004, 2006; Frenklach et al., 2002, 2004; Russi et al., 2008, 2010). Another technique, Bayesian melding (Poole and Raftery, 2000), akin to Bayesian calibration, reconciles differences between prior distributions on inputs and outputs of a simulator, with various applications, including Ševčíková et al. (2007); Radtke et al. (2002); Alkema et al. (2007); Fuentes and Raftery (2005). Recent work by Marmin and Filippone (2022) considered a more flexible model by relaxing the additive assumption of KOH in (2) through the use of a wrap function ϕ, which involves assuming the true process $\zeta(\mathbf{x}) = \phi(f(\mathbf{x}, \boldsymbol{\theta}))$ and employing a GP prior over ϕ.

6 Unidentifiability issue for parameter estimation in KOH

Let us redirect our attention to the original KOH approach of Kennedy and O'Hagan (2001). While the KOH method has demonstrated prowess in terms of predictions, the estimation for the calibration parameters $\boldsymbol{\theta}$ is recognized to grapple with the issue of *unidentifiability* between $\boldsymbol{\theta}$ and $\delta(\cdot)$ in the KOH model (3). Specifically, because $\delta(\cdot)$ is arbitrary, for each distinct $\boldsymbol{\theta}$, there exists a $\delta(\cdot)$ that perfectly fits the equation, leading to severe confounding between the two. Consequently, the interpretation of the estimates of $\boldsymbol{\theta}$ becomes obscure. In the original KOH paper, it is cautioned that "It is dangerous to interpret the estimates of $\boldsymbol{\theta}$ that are obtained by calibration as estimates of the true physical values of those parameters."

The issue of non-identifiability was initially highlighted in the discussion of the KOH paper and has since been extensively discussed in subsequent works, including Higdon et al. (2004); Bayarri et al. (2007b); Han et al. (2009); Brynjarsdóttir and O'Hagan (2014). Tuo and Wu (2016) formally delineated this unidentifiability issue mathematically and established the convergence of the calibration parameter $\boldsymbol{\theta}$ of the KOH approach in a large sample setting. Their work demonstrated that the posterior mode of the calibration parameter converges to a value dependent on the prior of the discrepancy in a frequentist setting. We will delve into a detailed description of this issue and introduce the subsequent developments in the following subsections.

6.1 Asymptotic properties of the KOH estimator

In the presence of the identifiability issue of the KOH model, a pertinent question arises: what is the KOH estimator actually estimating? Addressing this requires a deep dive into the intricacies of Bayesian estimation. Of particular interest is the impact of the GP prior on the discrepancy function, denoted as δ. Tuo and Wu (2016) and Tuo et al. (2020) considered a simplified version of the KOH estimator. In this version, the hyperparameters in the GP prior are treated as known quantities, and the focus shifts to the posterior mode rather than the full posterior distribution. Let $\delta(\mathbf{x}, \boldsymbol{\theta}) = \zeta(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\theta})$. Tuo and Wu (2016) showed that, under certain regularity conditions for $\delta(\mathbf{x}, \boldsymbol{\theta})$ in addition to a noiseless condition $\epsilon_i = 0$, the posterior mode converges to the minimizer of $l(\boldsymbol{\theta}) = \|\delta(\cdot, \boldsymbol{\theta})\|_{\mathcal{N}_{\Phi_\delta}}$. Here, $\|\cdot\|_{\mathcal{N}_{\Phi_\delta}}$ denotes the norm of the reproducing kernel Hilbert space (RKHS) generated by the kernel Φ_δ. For a deeper understanding of RKHS and its statistical applications, readers are directed to Wendland (2004) and Gu (2013). Later in Tuo et al. (2020), an analogous convergence result for the KOH method was obtained without necessitating the noiseless condition. These findings suggest that the limit value of the KOH method depends on the choice of the prior, and this dependence is in general hard to interpret. In the wake of these revelations, researchers have introduced methods with limiting values not contingent upon the choice of priors.

6.2 L_2-projection estimator

A practical strategy to combat the identifiability issues is: first define an identifiable parameter as the target parameter, and then propose a method to estimate this parameter. Tuo and Wu (2015, 2016) introduced the L_2-projection as

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \|\zeta(\cdot) - f(\cdot, \boldsymbol{\theta})\|_{L_2}^2, \tag{9}$$

where $\|\cdot\|_{L_2}$ denotes the L_2 norm of a function over its input domain. The L_2-projection can be regarded as a continuous version of the OLS method in (8), and it can be shown that the OLS estimator converges to the L_2-projection if the input points are uniformly distributed in the input domain. However, Tuo and Wu (2015) showed that, the OLS method is not *semi-parametric efficient*, in the sense that there exists estimators whose asymptotic variances are strictly smaller than that of the OLS, unless the computer model is exact.

Tuo and Wu (2015) proposed a two-step approach, called the L_2 calibration. In the first step, estimate ζ using the physical data alone via the kernel ridge regression (KRR). KRR is a frequentist analogy to the GP regression method. From a computational perspective, KRR is equal to the predictive mean of the GP regression. Denote $\hat{\zeta}$ as the KRR estimator. Then the L_2 calibration estimator is given by

$$\hat{\boldsymbol{\theta}}_{L_2} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \|\hat{\zeta}(\cdot) - f(\cdot, \boldsymbol{\theta})\|_{L_2}^2. \tag{10}$$

The asymptotic variance of the L_2 calibration estimator cannot be improved in general.

It is important to note that, despite the issue with estimating the calibration parameter, Tuo and Wu (2018) shows that the KOH method can consistently predict the true process.

A toy example, taken from Tuo and Wu (2015), is presented in Figure 2 to illustrate the L_2-calibration. Suppose that the true process is $\zeta(x) = \exp(x/10)\sin(x)$ for $x \in \Omega = (0, 2\pi)$ (represented as black lines), and the physical data is simulated by $y_i^p = \zeta(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$ and $x_i = 2\pi i/30$ for $i = 1, \ldots, 31$. The computer model is given by

$$f(x, \theta) = \zeta(x) - \sqrt{\theta^2 - \theta + 1}(\sin\theta x + \cos\theta x)$$

for $\theta \in \Theta = (-2, 2)$. This computer model is assumed to be inexpensive, eliminating the need for emulation. Note that there doesn't exist a real number θ satisfying $f(\cdot, \theta) = \zeta(\cdot)$ due to the always-positive quadratic function $\sqrt{\theta^2 - \theta + 1}$. Thus, the computer model is imperfect as described in Section 2.3. The L_2-projection (9) defines the true parameter as $\theta^* \approx -0.1789$.

We employ KOH and L_2-projection estimators, displayed in the left and right panels of Figure 2, respectively. Their estimates are -0.0821 and -0.1844, indicating that the L_2-calibration parameter (10) is closer to the true parameter θ^*. To illustrate the difference between the two estimates, Figure 2 shows that the computer model with the KOH estimate, i.e., $f(x, \hat{\theta}_{\text{KOH}})$ with $\hat{\theta}_{\text{KOH}}$ being the KOH estimate, deviates from the true process (though, as mentioned before, it can still perform well in terms of prediction by compensating for the difference using the discrepancy function). In contrast, the computer model with the L_2 estimate, $f(x, \hat{\theta}_{L_2})$, appears closer to the true process and physical data, making the estimate of θ more interpretable.

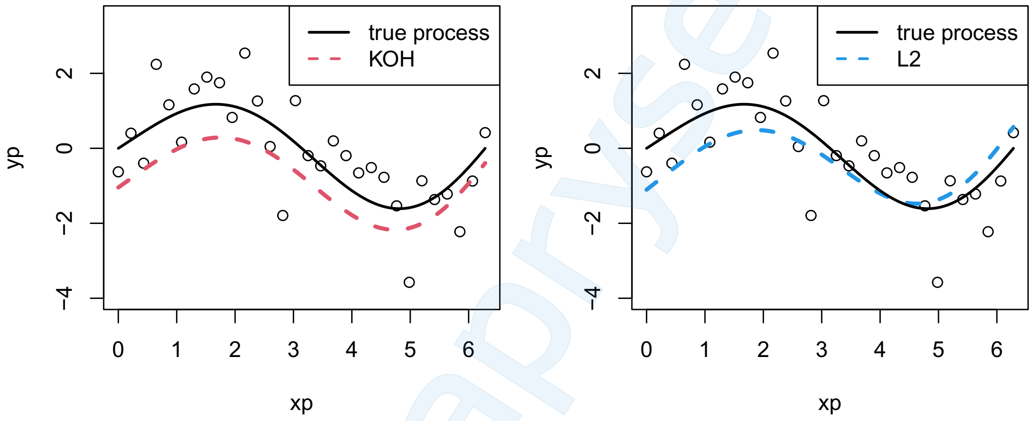

Figure 2: Illustration of L_2-calibration using a toy example, where the black lines denote the true process $\zeta(x)$, and the physical data, $y_1^p, \ldots, y_{31}^p$ are shown as circles. The left panel shows the result using the KOH estimator, where the red dashed line represents the computer model with the KOH estimate, i.e., $f(x, \hat{\theta}_{\text{KOH}})$ with $\hat{\theta}_{\text{KOH}}$ being the KOH estimate. The right panel shows the L_2-projection estimator, with the blue dashed line representing the computer model with the L_2 estimate, $f(x, \hat{\theta}_{L_2})$.

6.3 Bayesian L_2 calibration

A potential drawback of the L_2 calibration is that it does not have an immediate Bayesian version, and this makes the uncertainty quantification and prediction more challenging. Plumlee (2017) proposed a Bayesian approach to rectify the identifiability issue of KOH. The work shows that Bayesian calibration is possible without the presence of idenifiability issue by adopting the assumption that there is some value of the calibration parameter that is optimal under some loss function. Plumlee (2017) considers a broader class of loss functions

$$L_{W_k^2(\mu)}\{\zeta(\cdot) - f(\cdot, \boldsymbol{\theta})\} := \sum_{\|\alpha\|_{l_1} \leq k} \|D^{(\alpha)}\zeta(\cdot) - D^{(\alpha,0)} f(\cdot, \boldsymbol{\theta})\|^2_{L_2(\mu)}, \tag{11}$$

where D denotes the partial derivative operator. Let $\delta(\mathbf{x},\boldsymbol{\theta})=\zeta(\mathbf{x})-f(\mathbf{x},\boldsymbol{\theta})$. Under smoothness assumptions, (11) implies the orthogonality condition

$$\sum_{\|\alpha\|_{l_1}\leq k}\int D^{(\alpha,1)}f(\mathbf{x},\boldsymbol{\theta}^*)D^{(\alpha,0)}\delta(\mathbf{x},\boldsymbol{\theta}^*)d\mu(\mathbf{x})=0. \tag{12}$$

Recall that in KOH, $\delta(\cdot,\boldsymbol{\theta}^*)$ is modeled as a GP. Plumlee (2017) suggested incorporating the orthogonality condition (12) in addition to the GP prior. This leads to a restriction of the prior GP into a linear subset. Plumlee (2017) used his prior work (Plumlee and Joseph, 2018) to show that such a restriction again ends up with a GP, referred to as orthogonal GP, and presented the mean and covariance of the new GP. With this orthogonal GP, calibration can be done in a similar manner of KOH, only by replacing the original GP prior with the orthogonal GP prior.

Tuo (2019) proposed a frequentist version of this method, called the projected kernel calibration, and showed that, under the L_2 loss and other regularity conditions, this method is consistent and can achieve semi-parametric efficiency just as the L_2 calibration.

Xie and Xu (2021) considered another Bayesian version of the L_2 calibration and showed that this Bayesian estimator satisfies similar properties as the L_2 calibration estimator (10). An efficient stochastic approximation algorithm is provided to pursue the Bayesian posterior.

6.4 Confidence set on the parameters

Confidence set (Plumlee, 2019) is a frequentist approach for the uncertainty quantification of the calibration parameter. In analogous to the L_2 calibration, the set of optimal calibration parameter(s), denoted as Θ^*, should minimize the loss function $l(\zeta,f(\cdot,\boldsymbol{\theta}))=\int(\zeta(\mathbf{x})-f(\mathbf{x},\boldsymbol{\theta}))^2d\mu(\mathbf{x})$. An empirical approximation of this loss function is

$$\hat{l}(\text{data},f(\cdot,\boldsymbol{\theta}))=\frac{1}{n}\sum_{i=1}^{n}(y_i^p-f(\mathbf{x}_i^p,\boldsymbol{\theta}))^2.$$

For inexpensive computer models, under an optimal calibration parameter, the empirical loss becomes the sum of squares of the random noise, which can be used directly to build a confidence set. Here a proposed confidence set takes the form $\{\theta\in\Theta:\hat{l}(\text{data},f(\cdot,\boldsymbol{\theta}))\leq q_n(\alpha)/n\}$, which $q_n(\cdot)$ is properly choose such that the coverage rate of this set is no less than $1-\alpha$. For example, the the noise are standard normally distributed, q_n should be the quantile of the χ^2 distribution with degrees of freedom n. In the presence of model discrepancy, however, the empirical loss can be much larger that the previous sum of squares, and therefore, the direct method does not provide sufficient coverage. An additional set of functions, denoted as $I(\boldsymbol{\theta})$ are introduced to offset the model discrepancy, and the *conservative and consistent set* is described by

$$\text{CCS}=\left\{\boldsymbol{\theta}\in\Theta:\min_{d\in I(\boldsymbol{\theta})}\hat{l}(\text{data},f+d)\leq\frac{q_n(\alpha)}{n}\right\}.$$

The definition of $I(\boldsymbol{\theta})$ involves an RKHS, G, with its norm $\|\cdot\|_G$. Consider a ball $D=\{d\in G:\|d\|_G\leq\eta\}$ for a suitably chosen η. Plumlee (2019) assumed that the discrepancy lies in D and shown that to ensure a confidence set with sufficient coverage, the set $I(\boldsymbol{\theta})$ can be as small as

$$I(\boldsymbol{\theta})=\{d\in D:l(f(\cdot,\boldsymbol{\theta})+d(\cdot),f(\cdot,\boldsymbol{\theta}))\leq l(f(\cdot,\boldsymbol{\theta})+d(\cdot),f(\cdot,\mathbf{t}))\text{ for all }\mathbf{t}\in\Theta\}.$$

Plumlee (2019) further proposed two convex optimization problems to give approximated solutions to CCS.

6.5 Other approaches

We review some other methods that address the identifiability issue of calibration in this subsection. Wong et al. (2017) uses a least squares method to estimate $\boldsymbol{\theta}$ and then estimate the discrepancy function with non-parametric regression. A relevant two-stage estimation procedure is presented by Joseph and Melkote (2009). Joseph and Yan (2015) proposed an engineer-driven statistical model to improve the identifiability of the discrepancies. The scaled GP calibration (Gu and Wang, 2018; Gu et al., 2022) intends to reconcile the KOH method with the L_2 calibration. The goal is to build a method that enjoys both the theoretical rigor of the L_2 calibration and the computational convenience of the KOH method. The key idea of this method is to penalize the L_2 norm of the discrepancy while assuming it as a usual GP realization, which can be expressed as a hierarchical Bayesian model. An R package RobustCalibration (Gu, 2023) is available for the scaled GP calibration method. In some other works, it is shown that multiple responses can substantially enhance identifiability (Arendt et al., 2012a,b; Jiang et al., 2016). Sun and Fang (2023b) proposed another estimator by incorporating the L_2 calibration and the smoothing spline ANOVA idea.

7 Experimental designs

Experimental designs are crucial for providing accurate estimation and precise predictions under KOH models by collecting informative samples. Most of the work primarily focuses on selecting samples to ensure prediction accuracy, for which various criteria are proposed to measure the information gain through the collected samples (e.g., minimum IMSPE designs by Leatherman et al. (2017)). On the other hand, due to recent studies on the issue of unidentifiability as discussed in Section 6, more emphasis is placed on experimental designs to mitigate this problem (e.g., robust experimental designs by Krishna et al. (2022)).

It is noteworthy that in some studies, the focus is solely on collecting the physical input $\mathbf{x}_i^p$, whereas others concentrate on collecting simulation input $(\mathbf{x}_i^s, \boldsymbol{\theta}_i^s)$ for building an emulator for expensive simulators, as discussed in Section 4.2, and/or physical input $\mathbf{x}_i^p$.

We discuss two scenarios in two subsections, which are one-shot (or initial) designs, representing designs before collecting any data. Another common scenario is that after obtaining some samples from initial designs, we can estimate model parameters and then sequentially add one new sample (or a batch of samples) to the existing design, known as a sequential design or active learning. These are discussed respectively in Sections 7.1 and 7.2.

7.1 Initial design/One-shot design

One approach to choosing the experimental design is to find the samples that can minimize some degree of uncertainty quantification (either from prediction or estimation). Leatherman et al. (2017) developed minimum integrated mean-squared prediction error (IMSPE) designs for combined physical and simulation data. Specifically, suppose the simulator is inexpensive (and thus there is no need to design experiments for simulation data); in this case, given the physical input X_n^p, the IMSPE is defined as

$$\int_\Omega \sigma^2(\mathbf{x}) \mathrm{d}\mathbf{x},$$

where $\sigma^2(\mathbf{x})$ is denoted as in (7). The IMSPE criterion measures the prediction accuracy, or equivalently, the integrated prediction uncertainty over the domain Ω. Then, the minimum IMSPE design is to find X_n^p that minimizes the criterion.

Contrary to focusing on prediction accuracy, Arendt et al. (2016) and Krishna et al. (2022) aim to mitigate the unidentifiability issue. Arendt et al. (2016) used preposterior analysis to improve identification via space-filling criteria, wherein they calculate the preposterior covariance matrix (Berger, 2013; Carlin and Louis, 2000) of the calibration parameters that can be used to analyze the degree of unidentifiability when sampling the space-filling physical data. Krishna et al. (2022) mitigate the unidentifiability issue by providing a robust design for physical data. The central idea is using the two-stage estimation procedure of Joseph and Melkote (2009), for which they first use an approximate locally D-optimal design and augment it to a space-filling design using the remaining budget.

7.2 Sequential design/Active learning

The settings for sequential designs in the literature are diverse. Some of them focus on simulation data, some on physical data, and some on both.

Ranjan et al. (2011) proposed a batch sequential design for combined physical and simulation data using the IMSPE criterion. Williams et al. (2011) explored entropy and distance-based criteria in a batch sequential design setting for the physical experiment by improving the global prediction of discrepancies inferred from computer model calibration.

Sürer et al. (2023) focuses on sampling the simulation input $\boldsymbol{\theta}_i^s$ by proposing a sequential framework with a criterion for parameter selection that targets learning the posterior density of the parameters. Koermer et al. (2023) focuses on selecting the simulation input $(\mathbf{x}_i^s, \boldsymbol{\theta}_i^s)$ for building the GP surrogate model in the KOH setting using the IMSPE criterion, for which they derive a closed-form expression facilitating the optimization. Results suggest that the selected $\mathbf{x}_i^s$'s are space-filling marginally in their dimension while distinct exploratory behavior is observed in the $\boldsymbol{\theta}$-coordinate when collecting $\boldsymbol{\theta}_i^s$, reflecting the tradeoff between exploration and exploitation.

8 Applications in Diverse Scenarios

The KOH framework has served as an inspiration for a multitude of statistical approaches across diverse calibration problems. In this section, we extend our exploration beyond the problem settings outlined in Section 2 to consider various model calibration scenarios that have been adapted from the KOH model (3).

8.1 Assumptions about outputs

This subsection discusses the methods adapted from the KOH model (3) for addressing problems that deviate from the assumption of normal error with a constant variance.

8.1.1 Multivariate outputs/functional outputs

Multivariate outputs, both in terms of physical observations and simulation outputs, are frequently encountered in calibration problems. As mentioned by Higdon et al. (2008), "Our experience is that high-dimensional simulation output is the rule, rather than the exception."

The notion of multivariate outputs signifies that, given an input, the observed data and simulations yield a vector of outputs. This concept aligns with the broader framework of functional outputs, where the outputs exhibit dependence on time or space. A common approach to these calibration problems is to represent multivariate/functional outputs through a basis expansion, such as the wavelet decomposition method (Bayarri et al., 2007a) and the singular value decomposition (SVD) (Higdon et al., 2008). These basis expansion approaches have enjoyed considerable success and have paved the way for numerous subsequent studies in the field. See, for example, the on-site surrogates developed by Huang and Gramacy (2022) for large-scale, multi-output calibration, the Bayesian additive regression tree (BART) (Chipman et al., 1998, 2010) proposed by Pratola and Higdon (2016) for calibration problems with both high-dimensional outputs and inputs, the Bayesian model calibration with large nonstationary spatial outputs by Chang and Guillas (2019), and the history matching for high-dimensional outputs (Salter and Williamson, 2022).

Alternatively, techniques such as cross-correlation structure or the linear model of co-regionalization (LMC) (Gelfand et al., 2004) have been employed to model the correlations between multiple outputs. These methods, commonly used for emulating multi-output GPs, have found application in the context of model calibration (Arendt et al., 2012b; Paulo et al., 2012). Other techniques primarily focused on emulating multi-output simulations also have the potential for application in model calibration scenarios. Some relevant studies in this area include the works of Rougier (2008), McFarland et al. (2008), Bayarri et al. (2009), Conti and O'Hagan (2010), Fricker et al. (2013), Gu and Berger (2016), and Ma et al. (2022).

8.1.2 Heteroscedastic measurement error

When dealing with replicated physical experiments, it is not uncommon to encounter heteroscedasticity. In response to this challenge, Sung et al. (2022) have introduced a new statistical model that facilitates the estimation of calibration parameters and the generation of predictions in the presence of heteroscedasticity. Specifically, they consider an input-dependent error model:

$$y_i^p = f(\mathbf{x}_i^p, \boldsymbol{\theta}) + \delta(\mathbf{x}_i^p) + \epsilon(\mathbf{x}_i^p),$$

where the measurement error follows an independent normal distribution with heteroscedastic variance, i.e., $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, r(\mathbf{x}))$. The model inadequacy $\delta(\mathbf{x})$ is accommodated using the orthogonal GP outlined in Section 6.3, serving as a remedy for the unidentifiability issue. In the case of replication, they utilize a latent GP prior to model $r(\mathbf{x})$, drawing upon the concepts proposed by Goldberg et al. (1998) and Binois et al. (2018). The R package HetCalibrate (Sung, 2020) is available on an open repository for implementing this approach.

8.1.3 Biased measurement error

In certain scenarios, the error within the KOH model (3) may exhibit bias arising from the data acquisition process, referred to as *measurement bias* or *experimental bias*. Experimentation may introduce specific errors such as setup errors, which, despite occurring randomly during the experiment, can lead to systematic biases in model estimation. An instance of this can be seen in the context of the atmospheric error in satellite interferograms (Gu et al., 2023). To handle such measurement bias, Chang and Joseph (2014) assumes a model of the form

$$y_i^p = f(\mathbf{x}_i^p, \boldsymbol{\theta}) + \delta(\mathbf{x}_i^p) + e_i + \epsilon_i,$$

and considers that the measurement biases e_i are present only at certain locations, enabling a lasso-based estimation to identify and eliminate specific biases entirely. In a similar vein, Gu et al. (2023) proposes a model that considers the dependency of the measurement bias e_i on the input $\mathbf{x}$, expressed as $e(\mathbf{x}_i)$, and assumes that e follows a GP. This approach can be implemented using the R package RobustCalibration (Gu, 2023).

8.1.4 Non-Gaussian outputs

The original KOH model (3) primarily focuses on real-valued outputs y_i^p assumed to follow a normal distribution. Recent advancements in the field of model calibration have extended the scope to encompass scenarios with non-Gaussian outputs across various disciplines. For instance, Grosskopf et al. (2021) calibrated a radiation transport model for neutron detection experiments, dealing with photon counts that represent count data. To accommodate non-Gaussian outputs, Grosskopf et al. (2021) extended the KOH model, incorporating the concept of generalized linear models (McCullagh and Nelder, 2019). The authors employed a suitable link function g, e.g., the log link for count data, i.e., $g(\mathbb{E}[y_i]) = \log \mathbb{E}[y_i]$. Thus, in place of (3), the model assumes

$$y_i^p = g^{-1}(f(\mathbf{x}_i^p, \boldsymbol{\theta}) + \delta(\mathbf{x}_i^p)),$$

and the simulation outputs from (1) are modeled as $y^s = g^{-1}(f(\mathbf{x}, \boldsymbol{\theta}))$, where f and δ are independent GPs. Notably, the original KOH model can be regarded as a special case of this model, where the link function is $g(\mathbb{E}[y_i]) = \mathbb{E}[y_i]$ for Gaussian outputs.

To tackle the issue of unidentifiability, Sung et al. (2020) extended the L_2-projection estimator, as discussed in Section 6.2, to handle binary outputs. Similarly, both Sung and Hung (2024) and Sun and Fang (2023a) adapted this estimator to handle count data, demonstrating their consistency and establishing asymptotic distributions. Applications of these methods included simulations related to cell adhesion and an epidemiological model. The R package calibrateBinary (Sung, 2018) is available on an open repository for implementing the binary calibration proposed by Sung et al. (2020). Alternatively, Wang et al. (2022) proposed the weighted least squares estimation method to compute calibration parameters in an epidemiological model dealing with COVID-19 count data. A promising avenue for future exploration involves integrating the model by Grosskopf et al. (2021) with the approaches outlined in Section 6.3 to address the issue of unidentifiability within a Bayesian framework.

8.1.5 Censored data

Beyond outputs that follow the exponential family, Cao et al. (2018) extended the KOH approach for model calibration with censored data, which was motivated by a liquid stability forecasting application where the precise outcome is unknown and only observed to fall within a specific range. Notably, the work of Chen et al. (2022) introduces experimental designs for GPs under censoring, offering a potential avenue for designing experiments for model calibration with censored data.

8.2 Functional calibration parameters

In various scenarios, the parameter of interest for calibration exists as a function rather than a scalar or a set of scalars. An example of this can be found in the context of the model calibration for ion channel models of cardiac cells (Plumlee et al., 2016). To address such functional parameters, Pourhabib et al. (2015) and Atamturktur et al. (2015) consider parametric approaches where $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{x})$ assumes a parametric functional form. Recent research has leveraged the flexibility of GP models, assuming that the function $\boldsymbol{\theta}(\cdot)$ has a GP prior, including Plumlee et al. (2016) and Brown and Atamturktur (2018), where functional parameters are assumed to follow the distribution

$$g(\theta_j(\mathbf{x})) \overset{\text{indep.}}{\sim} \mathcal{GP}(\mu_j(\mathbf{x}), \tau_j \Phi_j(\mathbf{x}, \mathbf{x}')),$$

where indep. implies that the prior distribution of θ_i is independent of θ_k if $k \neq i$, and μ_j, τ_j and Φ_j are similarly defined as in Section 3.2. While Plumlee et al. (2016) employs the log link function g, Brown and Atamturktur (2018) uses logit, probit, cumulative, or identity functions, as they scale the parameters to lie in the unit hypercube. Sung (2022) relaxes the independence assumption, incorporating the multivariate GP modeling of Fricker et al. (2013) to account for correlated parameters within an epidemiological model.

Notably, Brown and Atamturktur (2018) and Sung (2022) do not consider the model discrepancy $\delta(\mathbf{x})$ within this KOH framework, as the model is already flexible enough. Brown and Atamturktur (2018) also argues that bypassing the need for the discrepancy term can alleviate the identifiability issue, while fostering stronger inferences and enhancing researchers' confidence in using the model for extrapolation.

Alternatively, Tuo et al. (2023) have developed a frequentist approach, providing a theoretical framework for a nonparametric solution to the functional calibration problem. Specifically, Tuo et al. (2023) defines the true functional calibration parameter by extending the definition of (9):

$$\boldsymbol{\theta}^*(\cdot) = \underset{\boldsymbol{\theta}(\cdot)}{\operatorname{argmin}} \int_\Omega (\zeta(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\theta}(\mathbf{x})))^2 \mathrm{d}\mathbf{x}, \quad \text{s.t.} \quad \boldsymbol{\theta}(\mathbf{x}) \in \Theta \quad \text{for all} \quad \mathbf{x} \in \Omega,$$

and proposed the following estimator:

$$\hat{\boldsymbol{\theta}}(\cdot) = \underset{\boldsymbol{\theta}(\cdot)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{p} (y_i^p - f(\mathbf{x}_i^p, \boldsymbol{\theta}(\mathbf{x}_i)))^2 + \lambda \sum_{j=1}^{q} \|\theta_j\|_{\mathcal{N}_{\Phi_j}},$$

where each θ_j lies in an RKHS, $\mathcal{N}_{\Phi_j}$, with the corresponding norm $\|\theta_j\|_{\mathcal{N}_{\Phi_j}}$ serving as a measure of roughness for the jth component of $\boldsymbol{\theta}(\cdot)$, and λ is a smoothing parameter. Ezzat et al. (2018) employed this method and developed a sequential design (of both physical and computer experiments) for functional calibration of computer models.

8.3 Multi-fidelity computer models

In certain applications, the presence of multi-fidelity simulators for the physical process is common. These varying levels of fidelity can emerge due to factors such as the presence of reduced-order physics in lower fidelity models, distinct accuracy levels specified for numerical solvers, or solutions obtained on finer grids. To handle such multi-fidelity simulators, Goh et al. (2013) proposed a Bayesian hierarchical model that incorporates the auto-regressive model introduced by Kennedy and O'Hagan (2000), that is, assuming $f(\mathbf{x}, \boldsymbol{\theta}) = \xi(\mathbf{x}, \boldsymbol{\theta}_l) + \delta_2(\mathbf{x}, \boldsymbol{\theta})$, where $f(\mathbf{x}, \boldsymbol{\theta})$ represents the high-fidelity simulator, $\xi(\mathbf{x}, \boldsymbol{\theta}_l)$ represents the low-fidelity simulator with its associated calibration parameter $\boldsymbol{\theta}_l$, which may not necessarily be the same as $\boldsymbol{\theta}$, and $\delta_2(\mathbf{x}, \boldsymbol{\theta})$ represents the discrepancy between high- and low-fidelity simulators. The functions ξ and δ_2 are assumed to be independent GPs. The KOH model can then be rewritten as

$$y_i^p = \xi(\mathbf{x}_i^p, \boldsymbol{\theta}_l) + \delta_2(\mathbf{x}_i^p, \boldsymbol{\theta}) + \delta(\mathbf{x}_i^p) + \epsilon_i.$$

Further exploration of integrating alternative techniques for combining multi-fidelity simulators, such as those proposed by Qian and Wu (2008) and Tuo et al. (2014), within this calibration framework would be valuable.

8.4 Large-scale dataset

Large-scale data (particularly simulation data) often presents challenges during model calibration, which can render fully Bayesian KOH calibration computationally intractable, as mentioned in Section 4.2, due to the large matrix inverses required for evaluating the likelihood in an MCMC scheme. Various approaches have been developed to address this issue.

Gramacy et al. (2015) utilized local approximate GP modeling (Gramacy and Apley, 2015) to emulate computer simulators, modularizing the KOH hierarchical model as done in Liu et al. (2009), and calibrating parameters by solving a derivative-free maximization of a likelihood term. This method can be implemented using the R package laGP (Gramacy, 2016). Huang et al. (2020) developed an on-site surrogate (OSS) for large-scale calibration that does not require modularization, making a fully Bayesian approach feasible. Instead of building a single large emulator for the simulator $f(\mathbf{x}, \boldsymbol{\theta})$ across the entire $(p+q)$-dimensional input space, they trained separate emulators focused on each input location $\mathbf{x}_i^p$ where field data has been collected.

Alternatively, tree-based models have also been considered to handle large-scale calibration, such as the BART calibration by Pratola and Higdon (2016), Bayesian treed calibration (BTC) by Konomi et al. (2017), and input-dependent Bayesian model calibration (IDBC) by Karagiannis et al. (2019).

Marmin and Filippone (2022) employed deep GPs, an emerging technique in the field of machine learning and uncertainty quantification (Damianou and Lawrence, 2013; Bui et al., 2016; Sauer et al., 2023b,a; Ding et al., 2023), to model both $f(\mathbf{x}, \boldsymbol{\theta})$ and $\delta(\mathbf{x})$, thereby enhancing the flexibility of both models to account for nonstationarity. They adapted techniques based on random feature expansions and stochastic variational inference, building on the work by Cutajar et al. (2017), to facilitate large-scale calibration.

Other approaches include parallel computing through divide-and-conquer methods (Cai and Mahadevan, 2017; Tsai et al., 2021), leveraging stochastic partial differential equations for addressing large nonstationary spatial outputs (Chang and Guillas, 2019), and subsampling techniques proposed by Lv et al. (2023).

Other techniques for large-scale GP, such as sparse approximation (Quiñonero-Candela and Rasmussen, 2005; Sang and Huang, 2012), covariance tapering (Furrer et al., 2006), inducing inputs (Snelson and Ghahramani, 2006), and nearest neighbor GPs (Datta et al., 2016; Finley et al., 2019) (which has been utilized in Cheng et al. (2021) for large-scale calibration), as well as Vecchia-approximated GPs/deep GPs (Katzfuss and Guinness, 2021; Sauer et al., 2023a), are also worth exploring within the framework of KOH calibration. For a comprehensive review of large-scale GP, refer to Liu et al. (2020).

9 Conclusion

In the ever-evolving landscape of modern technology, model calibration remains a critical cornerstone, ensuring the reliability and accuracy of complex computer models. With a focused examination of the Bayesian calibration framework proposed by Kennedy and O'Hagan (2001) (KOH), this review has shed light on the theoretical intricacies and recent advancements addressing the unidentifiability challenge within the KOH framework. Our discussion has emphasized the adaptability of the KOH framework to diverse and complex scenarios, including those involving multivariate outputs and functional calibration parameters. Additionally, we have explored recent developments in experimental design strategies tailored to the unique demands of model calibration. By offering comprehensive insights into the KOH approach and its versatile applications, this review serves as an indispensable guide for researchers and practitioners striving to enhance the precision and robustness of their computer models. Moving forward, continual research and innovation in the field of model calibration will undoubtedly facilitate advancements in digital twin technology and Industry 4.0, fostering a more efficient and interconnected industrial landscape.

Funding Information

The author Chih-Li Sung gratefully acknowledges funding from NSF DMS 2113407. The author Rui Tuo gratefully acknowledges funding from NSF DMS 2312173 and NSF CNS 2328395.

Acknowledgments

The authors gratefully acknowledge helpful advice from the editor and two anonymous referees whose comments greatly strengthen this paper.

References

Alkema, L., Raftery, A. E., and Clark, S. J. (2007). Probabilistic projections of hiv prevalence using Bayesian melding. *The Annals of Applied Statistics*, 1(1):229–248.

Allaire, D., He, Q., Deyst, J., and Willcox, K. (2012). An information-theoretic metric of system complexity with application to engineering system design. *Journal of Mechanical Design*, 134(10):100906.

Anastassopoulou, C., Russo, L., Tsakris, A., and Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS One*, 15(3):e0230405.

Andrianakis, I., McCreesh, N., Vernon, I., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2017). Efficient history matching of a high dimensional individual-based hiv transmission model. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):694–719.

Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2015). 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.

Arendt, P. D., Apley, D. W., and Chen, W. (2012a). Quantification of model uncertainty: Calibration, model discrepancy, and identifiability. *Journal of Mechanical Design*, 134(10):100908.

Arendt, P. D., Apley, D. W., and Chen, W. (2016). A preposterior analysis to predict identifiability in the experimental calibration of computer models. *IIE Transactions*, 48(1):75–88.

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

Atamturktur, S., Hegenderfer, J., Williams, B., Egeberg, M., Lebensohn, R., and Unal, C. (2015). A resource allocation framework for experiment-based validation of numerical models. *Mechanics of Advanced Materials and Structures*, 22(8):641–654.

Baker, E., Barbillon, P., Fadikar, A., Gramacy, R. B., Herbei, R., Higdon, D., Huang, J., Johnson, L. R., Ma, P., Mondal, A., et al. (2022). Analyzing stochastic computer models: A review with opportunities. *Statistical Science*, 37(1):64–89.

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., Walsh, D., et al. (2007a). Computer model validation with functional output. *The Annals of Statistics*, 35(5):1874–1906.

Bayarri, M. J., Berger, J. O., Kennedy, M. C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Lin, C.-H., and Tu, J. (2009). Predicting vehicle crashworthiness: Validation of computer models for functional and hierarchical data. *Journal of the American Statistical Association*, 104(487):929–943.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. (2007b). A framework for validation of computer models. *Technometrics*, 49(2):138–154.

Berger, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Binois, M., Gramacy, R. B., and Ludkovski, M. (2018). Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Boukouvalas, A., Sykes, P., Cornford, D., and Maruri-Aguilar, H. (2014). Bayesian precalibration of a large stochastic microsimulation model. *IEEE Transactions on Intelligent Transportation Systems*, 15(3):1337–1347.

Box, G. E. P. and Hunter, W. G. (1962). A useful method for model-building. *Technometrics*, 4(3):301–318.

Brown, D. A. and Atamturktur, S. (2018). Nonparametric functional calibration of computer models. *Statistica Sinica*, 28(2):721–742.

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

Bui, T., Hernández-Lobato, D., Hernandez-Lobato, J., Li, Y., and Turner, R. (2016). Deep Gaussian processes for regression using approximate expectation propagation. In Balcan, M.-F. and Weinberger, K. Q., editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016*, volume 48, pages 1472–1481. PMLR.

Cai, G. and Mahadevan, S. (2017). Model calibration with big data. In *Model Validation and Uncertainty Quantification, Volume 3: Proceedings of the 35th IMAC, A Conference and Exposition on Structural Dynamics 2017*, pages 315–322. Springer.

Campbell, K. (2006). Statistical calibration of computer simulations. *Reliability Engineering & System Safety*, 91(10-11):1358–1363.

Cao, F., Ba, S., Brenneman, W. A., and Joseph, V. R. (2018). Model calibration with censored data. *Technometrics*, 60(2):255–262.

Carlin, B. P. and Louis, T. A. (2000). Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289.

Carmassi, M. (2018). *CaliCo: Code Calibration in a Bayesian Framework*. R package version 0.1.1.

Chang, C.-J. and Joseph, V. R. (2014). Model calibration through minimal adjustments. *Technometrics*, 56(4):474–482.

Chang, K.-L. and Guillas, S. (2019). Computer model calibration with large non-stationary spatial outputs: application to the calibration of a climate model. *Journal of the Royal Statistical Society: Series C*, 68(1):51–78.

Chen, J., Mak, S., Joseph, V. R., and Zhang, C. (2022). Adaptive design for Gaussian process regression under censoring. *The Annals of Applied Statistics*, 16(2):744–764.

Cheng, S., Konomi, B. A., Matthews, J. L., Karagiannis, G., and Kang, E. L. (2021). Hierarchical Bayesian nearest neighbor co-kriging Gaussian process models; an application to intersatellite calibration. *Spatial Statistics*, 44:100516.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

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

Cox, D. D., Park, J. S., Sacks, J., and Singer, C. E. (1992). Tuning complex computer code to data. In *Proc. 23rd Symp. Interface of Computing Science and Statistics, April 21st-24th, 1991, Seattle*.

Craig, P., Goldstein, M., Seheult, A. H., and Smith, J. A. (1996). Bayes linear strategies for matching hydrocarbon reservoir history. *Bayesian Statistics*, 14:69–95.

Craig, P. S., Goldstein, M., Rougier, J. C., and Seheult, A. H. (2001). Bayesian forecasting using large computer models. *Journal of the American Statistical Association*, 96(454):717–729.

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

Damianou, A. and Lawrence, N. D. (2013). Deep Gaussian processes. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31, pages 207–215. PMLR.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Ding, L., Tuo, R., and Shahrampour, S. (2023). A sparse expansion for deep Gaussian processes. *IISE Transactions*, to appear.

Ezzat, A. A., Pourhabib, A., and Ding, Y. (2018). Sequential design for functional calibration of computer models. *Technometrics*, 60(3):286–296.

Fang, K.-T., Li, R., and Sudjianto, A. (2005). *Design and Modeling for Computer Experiments*. CRC press.

Farah, M., Birrell, P., Conti, S., and Angelis, D. D. (2014). Bayesian emulation and calibration of a dynamic epidemic model for A/H1N1 influenza. *Journal of the American Statistical Association*, 109(508):1398–1411.

Feeley, R., Frenklach, M., Onsum, M., Russi, T., Arkin, A., and Packard, A. (2006). Model discrimination using data collaboration. *The Journal of Physical Chemistry A*, 110(21):6803–6813.

Feeley, R., Seiler, P., Packard, A., and Frenklach, M. (2004). Consistency of a reaction dataset. *The Journal of Physical Chemistry A*, 108(44):9573–9583.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414.

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

Frenklach, M., Packard, A., Garcia-Donato, G., Paulo, R., and Sacks, J. (2016). Comparison of statistical and deterministic frameworks of uncertainty quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):875–901.

Frenklach, M., Packard, A., and Seiler, P. (2002). Prediction uncertainty from models and data. In *Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301)*, volume 5, pages 4135–4140. IEEE.

Frenklach, M., Packard, A., Seiler, P., and Feeley, R. (2004). Collaborative data processing in developing predictive models of complex reaction systems. *International Journal of Chemical Kinetics*, 36(1):57–66.

Fricker, T. E., Oakley, J. E., and Urban, N. M. (2013). Multivariate Gaussian process emulators with nonseparable covariance structures. *Technometrics*, 55(1):47–56.

Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

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

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13:263–312.

Generale, A. P., Hall, R. B., Brockman, R. A., Joseph, V. R., Jefferson, G., Zawada, L., Pierce, J., and Kalidindi, S. R. (2022). Bayesian calibration of continuum damage model parameters for an oxide-oxide ceramic matrix composite using inhomogeneous experimental data. *Mechanics of Materials*, 175:104487.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. CRC press.

Goh, J., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., and Rutter, E. (2013). Prediction and computer model calibration using outputs from multifidelity simulators. *Technometrics*, 55(4):501–512.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. (1998). Regression with input-dependent noise: A Gaussian process treatment. In *Advances in Neural Information Processing Systems*, pages 493–499.

Gramacy, R. B. (2016). laGP: large-scale spatial modeling via local approximate Gaussian processes in R. *Journal of Statistical Software*, 72(1):1–46.

Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. CRC Press.

Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

Gramacy, R. B., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., Rutter, E., Trantham, M., and Drake, R. P. (2015). Calibrating a large computer experiment simulating radiative shock hydrodynamics. *The Annals of Applied Statistics*, 9(3):1141–1168.

Grosskopf, M., Bingham, D., Adams, M. L., Hawkins, W. D., and Perez-Nunez, D. (2021). Generalized computer model calibration for radiation transport simulation. *Technometrics*, 63(1):27–39.

Gu, C. (2013). *Smoothing Spline ANOVA Models*. Springer.

Gu, M. (2023). *RobustCalibration: Robust Calibration of Imperfect Mathematical Models*. R package version 0.5.4.

Gu, M., Anderson, K., and McPhillips, E. (2023). Calibration of imperfect geophysical models by multiple satellite interferograms with measurement bias. *Technometrics*, to appear.

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

Gu, M. and Wang, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1555–1583.

Gu, M., Xie, F., and Wang, L. (2022). A theoretical framework of the scaled Gaussian stochastic process in prediction and calibration. *SIAM/ASA Journal on Uncertainty Quantification*, 10(4):1435–1460.

Han, G., Santner, T. J., and Rawlinson, J. J. (2009). Simultaneous determination of tuning and calibration parameters for computer experiments. *Technometrics*, 51(4):464–474.

Hankin, R. K. S. (2005). Introducing BACCO, an R bundle for Bayesian analysis of computer code output. *Journal of Statistical Software*, 14.

Henderson, D. A., Boys, R. J., Krishnan, K. J., Lawless, C., and Wilkinson, D. J. (2009). 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.

Higdon, D., Gattiker, J., Lawrence, E., Jackson, C., Tobis, M., Pratola, M., Habib, S., Heitmann, K., and Price, S. (2013). Computer model calibration using the ensemble Kalman filter. *Technometrics*, 55(4):488–500.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583.

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

Higdon, D., McDonnell, J. D., Schunck, N., Sarich, J., and Wild, S. M. (2015). A bayesian approach for parameter estimation and prediction using a computationally intensive model. *Journal of Physics G: Nuclear and Particle Physics*, 42(3):034009.

Huang, J. and Gramacy, R. B. (2022). Multi-output calibration of a honeycomb seal via on-site surrogates. *Technometrics*, 64(4):548–563.

Huang, J., Gramacy, R. B., Binois, M., and Libraschi, M. (2020). On-site surrogates for large-scale calibration. *Applied Stochastic Models in Business and Industry*, 36(2):283–304.

Jiang, Z., Arendt, P. D., Apley, D. W., and Chen, W. (2016). Multi-response approach to improving identifiability in model calibration. In Ghanem, R., Higdon, D., and Owhadi, H., editors, *Handbook of Uncertainty Quantification*, pages 69–127. New York: Springer.

Joseph, V. R. and Melkote, S. N. (2009). Statistical adjustments to engineering models. *Journal of Quality Technology*, 41(4):362–375.

Joseph, V. R. and Yan, H. (2015). Engineering-driven statistical adjustment and calibration. *Technometrics*, 57(2):257–267.

Karagiannis, G., Konomi, B. A., and Lin, G. (2019). On the Bayesian calibration of expensive computer models with input dependent parameters. *Spatial Statistics*, 34:100258.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.

Katzfuss, M. and Guinness, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, 36(1):124–141.

Kejzlar, V. and Maiti, T. (2023). Variational inference with vine copulas: an efficient approach for bayesian computer model calibration. *Statistics and Computing*, 33(1):18.

Kejzlar, V., Neufcourt, L., Nazarewicz, W., and Reinhard, P.-G. (2020). Statistical aspects of nuclear mass models. *Journal of Physics G: Nuclear and Particle Physics*, 47(9):094001.

Kejzlar, V., Son, M., Bhattacharya, S., and Maiti, T. (2021). A fast and calibrated computer model emulator: an empirical Bayes approach. *Statistics and Computing*, 31(49):1–26.

Kenett, R. S. and Bortman, J. (2022). The digital twin in Industry 4.0: A wide-angle perspective. *Quality and Reliability Engineering International*, 38(3):1357–1366.

Kennedy, M. C. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13.

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

King, G. B., Lovell, A. E., Neufcourt, L., and Nunes, F. M. (2019). Direct comparison between bayesian and frequentist uncertainty quantification for nuclear reactions. *Physical Review Letters*, 122(23):232502.

Koermer, S., Loda, J., Noble, A., and Gramacy, R. B. (2023). Active learning for simulator calibration. *arXiv preprint arXiv:2301.10228*.

Konomi, B. A., Karagiannis, G., Lai, K., and Lin, G. (2017). Bayesian treed calibration: an application to carbon capture with ax sorbent. *Journal of the American Statistical Association*, 112(517):37–53.

Krishna, A., Joseph, V. R., Ba, S., Brenneman, W. A., and Myers, W. R. (2022). Robust experimental designs for model calibration. *Journal of Quality Technology*, 54(4):441–452.

Kullback, S. (1997). *Information Theory and Statistics*. Courier Corporation.

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

Leatherman, E. R., Dean, A. M., and Santner, T. J. (2017). Designing combined physical and computer experiments to maximize prediction accuracy. *Computational Statistics & Data Analysis*, 113:346–362.

Lee, B. S., Haran, M., Fuller, R. W., Pollard, D., and Keller, K. (2020). A fast particle-based approach for calibrating a 3-d model of the antarctic ice sheet. *The Annals of Applied Statistics*, 14(2):605–634.

Liu, C. K. and Negrut, D. (2021). The role of physics-based simulators in robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:35–58.

Liu, F., Bayarri, M. J., and Berger, J. O. (2009). Modularization in bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020). When Gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423.

Liyanage, D., Ji, Y., Everett, D., Heffernan, M., Heinz, U., Mak, S., and Paquet, J.-F. (2022). Efficient emulation of relativistic heavy ion collisions with transfer learning. *Physical Review C*, 105(3):034910.

Lv, S., Yu, J., Wang, Y., and Du, J. (2023). Fast calibration for computer models with massive physical observations. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):1069–1104.

Ma, P., Mondal, A., Konomi, B. A., Hobbs, J., Song, J. J., and Kang, E. L. (2022). Computer model emulation with high-dimensional functional output in large-scale observing system uncertainty experiments. *Technometrics*, 64(1):65–79.

Mahnken, R. (2017). Identification of material parameters for constitutive equations. *Encyclopedia of Computational Mechanics (second ed.), American Cancer Society*, pages 1–21.

Mak, S., Sung, C.-L., Wang, X., Yeh, S.-T., Chang, Y.-H., Joseph, V. R., Yang, V., and Wu, C. F. J. (2018). An efficient surrogate model for emulation and physics extraction of large eddy simulations. *Journal of the American Statistical Association*, 113(524):1443–1456.

Marmin, S. and Filippone, M. (2022). Deep Gaussian processes for calibration of computer models (with discussion). *Bayesian Analysis*, 17(4):1301–1350.

McCullagh, P. and Nelder, J. A. (2019). *Generalized Linear Models*. New York: Routledge, second edition.

McFarland, J., Mahadevan, S., Romero, V., and Swiler, L. (2008). Calibration and uncertainty analysis for computer simulations with multivariate output. *AIAA journal*, 46(5):1253–1265.

McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2018). Approximate bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statistical Science*, 33(1):4–18.

Morris, M. D. and Mitchell, T. J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43(3):381–402.

Oakley, J. E. and Youngman, B. D. (2017). Calibration of stochastic computer simulators using likelihood emulation. *Technometrics*, 59(1):80–92.

Paulo, R., García-Donato, G., and Palomo, J. (2012). Calibration of computer models with multivariate output. *Computational Statistics & Data Analysis*, 56(12):3959–3974.

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

Plumlee, M. (2019). Computer model calibration with confidence and consistency. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):519–545.

Plumlee, M. and Joseph, V. R. (2018). Orthogonal Gaussian process models. *Statistica Sinica*, 28(2):601–619.

Plumlee, M., Joseph, V. R., and Yang, H. (2016). Calibrating functional parameters in the ion channel models of cardiac cells. *Journal of the American Statistical Association*, 111(514):500–509.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255.

Pourhabib, A., Huang, J. Z., Wang, K., Zhang, C., Wang, B., and Ding, Y. (2015). Modulus prediction of buckypaper based on multi-fidelity analysis involving latent variables. *IIE Transactions*, 47(2):141–152.

Pratola, M. T. and Chkrebtii, O. (2018). Bayesian calibration of multistate stochastic simulators. *Statistica Sinica*, 28(2):693–719.

Pratola, M. T. and Higdon, D. M. (2016). Bayesian additive regression tree calibration of complex high-dimensional computer models. *Technometrics*, 58(2):166–179.

Pratola, M. T., Sain, S. R., Bingham, D., Wiltberger, M., and Rigler, E. J. (2013). Fast sequential computer model calibration of large nonstationary spatial-temporal processes. *Technometrics*, 55(2):232–242.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

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

Qian, P. Z. G. and Wu, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2):192–204.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959.

Radtke, P. J., Burk, T. E., and Bolstad, P. V. (2002). Bayesian melding of a forest ecosystem model with correlated inputs. *Forest Science*, 48(4):701–711.

Raftery, A. E., Givens, G. H., and Zeh, J. E. (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association*, 90(430):402–416.

Ranjan, P., Lu, W., Bingham, D., Reese, S., Williams, B. J., Chou, C.-C., Doss, F., Grosskopf, M., and Holloway, J. P. (2011). Follow-up experimental designs for computer models and physical processes. *Journal of Statistical Theory and Practice*, 5(1):119–136.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Romanowicz, R., Beven, K., and Tawn, J. A. (1994). Evaluation of predictive uncertainty in nonlinear hydrological models using a bayesian approach. In Barnett, V. and Turkman, K. F., editors, *Statistics for the Environment 2: Water Related Issues*, pages 297–317. New York: Wiley.

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, 17(4):827–843.

Rumsey, K. N. and Huerta, G. (2021). Fast matrix algebra for Bayesian model calibration. *Journal of Statistical Computation and Simulation*, 91(7):1331–1341.

Russi, T., Packard, A., Feeley, R., and Frenklach, M. (2008). Sensitivity analysis of uncertainty in model prediction. *The Journal of Physical Chemistry A*, 112(12):2579–2588.

Russi, T., Packard, A., and Frenklach, M. (2010). Uncertainty quantification: Making predictions of complex reaction systems reliable. *Chemical Physics Letters*, 499(1-3):1–8.

Rutter, C. M., Ozik, J., DeYoreo, M., and Collier, N. (2019). Microsimulation model calibration using incremental mixture approximate bayesian computation. *The Annals of Applied Statistics*, 13(4):2189–2212.

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

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

Salter, J. M. and Williamson, D. B. (2022). Efficient calibration for high-dimensional computer model output using basis methods. *International Journal for Uncertainty Quantification*, 12(6):47–69.

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

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 74(1):111–132.

Santner, T. J., Williams, B. J., and Notz, W. I. (2018). *The Design and Analysis of Computer Experiments (Second Edition)*. Springer New York.

Sauer, A., Cooper, A., and Gramacy, R. B. (2023a). Vecchia-approximated deep Gaussian processes for computer experiments. *Journal of Computational and Graphical Statistics*, 32(3):824–837.

Sauer, A., Gramacy, R. B., and Higdon, D. (2023b). Active learning for deep Gaussian process surrogates. *Technometrics*, 65(1):4–18.

Ševčíková, H., Raftery, A. E., and Waddell, P. A. (2007). Assessing uncertainty in urban simulations using Bayesian melding. *Transportation Research Part B: Methodological*, 41(6):652–669.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media.

Sun, Y. and Fang, X. (2023a). A model calibration procedure for count response. *Communications in Statistics-Theory and Methods*, to appear.

Sun, Y. and Fang, X. (2023b). A new L_2 calibration procedure of computer models based on the smoothing spline ANOVA. *Statistical Papers*, to appear.

Sung, C.-L. (2018). *calibrateBinary: Calibration for Computer Experiments with Binary Responses*. R package version 0.1.

Sung, C.-L. (2020). *HetCalibrate: Calibration of Inexact Computer Models with Heteroscedastic Errors*. R package version 0.1.

Sung, C.-L. (2022). Estimating functional parameters for understanding the impact of weather and government interventions on COVID-19 outbreak. *The Annals of Applied Statistics*, 16(4):2505–2522.

Sung, C.-L., Barber, B. D., and Walker, B. J. (2022). Calibration of inexact computer models with heteroscedastic errors. *SIAM/ASA Journal on Uncertainty Quantification*, 10(4):1733–1752.

Sung, C.-L. and Hung, Y. (2024). Efficient calibration for imperfect epidemic models with applications to the analysis of COVID-19. *Journal of the Royal Statistical Society: Series C*, 73(1):47–64.

Sung, C.-L., Hung, Y., Rittase, W., Zhu, C., and Wu, C. F. J. (2020). Calibration for computer experiments with binary responses and application to cell adhesion study. *Journal of the American Statistical Association*, 115(532):1664–1674.

Sürer, Ö., Plumlee, M., and Wild, S. M. (2023). Sequential bayesian experimental design for calibration of expensive simulation models. *Technometrics*, to appear.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518.

Thelen, A., Zhang, X., Fink, O., Lu, Y., Ghosh, S., Youn, B. D., Todd, M. D., Mahadevan, S., Hu, C., and Hu, Z. (2022). A comprehensive review of digital twin—part 1: modeling and twinning enabling technologies. *Structural and Multidisciplinary Optimization*, 65(12):354.

Thelen, A., Zhang, X., Fink, O., Lu, Y., Ghosh, S., Youn, B. D., Todd, M. D., Mahadevan, S., Hu, C., and Hu, Z. (2023). A comprehensive review of digital twin—part 2: roles of uncertainty quantification and optimization, a battery digital twin, and perspectives. *Structural and Multidisciplinary Optimization*, 66(1):1.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature communications*, 12(1):5988.

Tuo, R. (2019). Adjustments to computer models via projected kernel calibration. *SIAM/ASA Journal on Uncertainty Quantification*, 7(2):553–578.

Tuo, R., He, S., Pourhabib, A., Ding, Y., and Huang, J. Z. (2023). A reproducing kernel Hilbert space approach to functional calibration of computer models. *Journal of the American Statistical Association*, 118(542):883–897.

Tuo, R., Wang, Y., and Wu, C. F. J. (2020). On the improved rates of convergence for Matérn-type kernel ridge regression with application to calibration of computer models. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1522–1547.

Tuo, R. and Wu, C. F. J. (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352.

Tuo, R. and Wu, C. F. J. (2016). A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795.

Tuo, R., Wu, C. F. J., and Yu, D. (2014). Surrogate modeling of computer experiments with different mesh densities. *Technometrics*, 56(3):372–380.

Tuo, R. and Wu, C. J. (2018). Prediction based on the Kennedy-O'Hagan calibration model: Asymptotic consistency and other properties. *Statistica Sinica*, 28(2):743–759.

Vernon, I., Goldstein, M., and Bower, R. (2014). Galaxy formation: Bayesian history matching for the observable universe. *Statistical Science*, 29(1):81–90.

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

Viana, F. A. C. and Subramaniyan, A. K. (2021). A survey of Bayesian calibration and physics-informed neural networks in scientific modeling. *Archives of Computational Methods in Engineering*, 28:3801–3830.

Wang, S., Chen, W., and Tsui, K.-L. (2009). Bayesian validation of computer models. *Technometrics*, 51(4):439–451.

Wang, Y., Lu, G., and Du, J. (2022). Calibration and prediction for the inexact SIR model. *Mathematical Biosciences and Engineering*, 19(3):2800–2818.

Wang, Y., Yue, X., Tuo, R., Hunt, J. H., and Shi, J. (2020). Effective model calibration via sensible variable identification and adjustment with application to composite fuselage simulation. *The Annals of Applied Statistics*, 14(4):1759–1776.

Wendland, H. (2004). *Scattered data approximation*, volume 17. Cambridge university press.

Williams, B. J., Loeppky, J. L., Moore, L. M., and Macklem, M. S. (2011). Batch sequential design to achieve predictive maturity with calibrated computer models. *Reliability Engineering & System Safety*, 96(9):1208–1219.

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

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

Williamson, D. B., Blaker, A. T., and Sinha, B. (2017). Tuning without over-tuning: parametric uncertainty quantification for the nemo ocean model. *Geoscientific Model Development*, 10(4):1789–1816.

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

Xie, F. and Xu, Y. (2021). Bayesian projected calibration of computer models. *Journal of the American Statistical Association*, 116(536):1965–1982.

Xiong, Y., Chen, W., Tsui, K.-L., and Apley, D. W. (2009). A better understanding of model updating strategies in validating engineering models. *Computer Methods in Applied Mechanics and Engineering*, 198(15-16):1327–1337.

Zhou, M., Chen, W., Su, X., Sung, C.-L., Wang, X., and Ren, Z. (2023). Data-driven modeling of general fluid density under subcritical and supercritical conditions. *AIAA Journal*, 61(4):1519–1531.