Data-Centric Engineering (2026), 7: e7
doi:10.1017/dce.2025.10035

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Embedded model form uncertainty quantification with measurement noise for Bayesian model calibration

Daniel Andrés Arcones[1,2], Martin Weiser[3], Phaedon-Stelios Koutsourelakis[4] and Jörg F. Unger[2]

[1]School of Engineering and Design, Technical University of Munich, Germany
[2]7.7 Modeling and Simulation, Bundesanstalt für Materialforschung und -prüfung, Germany
[3]Department of Modeling and Simulation of Complex Processes, Zuse-Institut Berlin, Germany
[4]Professorship of Data-driven Materials Modeling, School of Engineering and Design, Technical University of Munich, Germany
Corresponding author: Daniel Andrés Arcones; Email: daniel.andres@tum.de

Received: 23 June 2025; **Revised:** 24 November 2025; **Accepted:** 09 December 2025

Keywords: Bayesian inference; model form uncertainty; model updating; noise

Abstract

A key factor in ensuring the accuracy of computer simulations that model physical systems is the proper calibration of their parameters based on real-world observations or experimental data. Bayesian methods provide a robust framework for quantifying and propagating the uncertainties that inevitably arise. Nevertheless, they produce predictions unable to represent the observed datapoints when paired with inexact models. Additionally, the quantified uncertainties of these overconfident models cannot be propagated to other Quantities of Interest (QoIs) reliably. A promising solution involves embedding a model inadequacy term in the inference parameters, allowing the quantified model form uncertainty to influence non-observed QoIs. In this work, we revisit this embedded formulation and analyze how different likelihood constructions affect the inference of model form uncertainty, particularly under the presence of prescribed measurement noise and unavoidable model discrepancies. Two additional likelihood formulations, the global moment-matching and relative global moment-matching likelihoods, are introduced to explore alternative ways of representing the residual distribution. The behavior of these likelihoods is examined alongside existing formulations to show how different treatments of measurement noise and discrepancies shape the inferred parameter posteriors, and thereby affect the uncertainty ultimately propagated to the QoIs. Particular attention is given to how the uncertainty associated with the model inadequacy term propagates to the QoIs for the posteriors obtained from different likelihood formulations, enabling a more comprehensive statistical analysis of the prediction's reliability. Finally, the proposed approach is applied to estimate the uncertainty in the predicted heat flux from a transient thermal simulation using temperature observations.

Impact Statement

Traditional Bayesian model calibration techniques can underestimate predictive uncertainty when the computational model cannot fully represent the observed data. This paper investigates how different likelihood formulations within an embedded model inadequacy framework influence the handling of measurement noise and model form uncertainties. The results highlight that the likelihood specification influences the posterior distribution, which subsequently governs the uncertainty propagated to predictions and unobserved QoIs. This

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

provides valuable guidance for practitioners employing embedded model inadequacy in simulation-based calibration. By providing a more robust framework for managing uncertainties, this work advances the integration of simulations with measurement data, supporting practical applications such as digital twins.

1. Introduction

In the last few decades, advances in sensor technology have made available large quantities of affordable data that reflect the current state of physical systems and processes. In parallel, simulation models have become ubiquitous in science and engineering for generating predictions on the behavior of a given system based on physical laws. With the ongoing transition to an Industry 4.0 paradigm and the proliferation of Digital Twins, the challenge of calibrating simulation models based on data from physical sensors is becoming increasingly relevant (Vaidya et al., 2018). In the specific case of Digital Twins of physical systems that use simulations to predict the behavior of their real counterpart and make decisions on their behalf, the accuracy and reliability of such predictions are essential (Andrés Arcones et al., 2023). Therefore, their uncertainty must be adequately quantified to achieve such trustworthy predictions.

In most cases, the calibration of the computational models is achieved by estimating a set of parameters that control their behavior based on available observations from the modeled system. Although more uncertainty sources can be identified (Walker et al., 2003), this process introduces two main ones: the error attributed to noise or uncontrolled variations in the data and the discrepancy created by the assumptions used to generate the computational model. They can be related to aleatoric and epistemic uncertainty sources, respectively. While noise errors can be satisfactorily estimated from the observations, the so-called *model discrepancy* or *model inadequacy* presents further challenges. All models are based on assumptions and simplifications that are unavoidable when tackling an infinitely complex reality. Systematic reviews on different approaches for dealing with this model inadequacy exist (Campbell, 2006; Bojke et al., 2009; Gupta et al., 2012; Pernot and Cailliez, 2017; Sung and Tuo, 2024), but they all coincide in the need for further research to achieve reliable methods for quantifying the model form uncertainty.

Bayesian approaches are usually implemented to infer the model parameters (Robert, 2007; Gelman et al., 2013), which allows for obtaining a probability distribution for those parameters based on the observed data. This distribution will narrow down to the optimal parameter value, which may not be able to represent the observations in the presence of model inadequacy (Kaipio and Somersalo, 2006). One classical solution to this problem is the framework proposed by Kennedy and O'Hagan (2001), which extends the model output with a flexible term that corrects the predictions to better reflect the observations. Since its conception, several extensions have been proposed within the framework (Bayarri et al., 2009; Brynjarsdóttir and O'Hagan, 2014; Plumlee, 2017; Barbillon et al., 2024; Leoni et al., 2024). However, one key disadvantage of such implementations is the impossibility of transferring the inferred estimation of the model inadequacy term to other derived quantities of interest (QoI) computed with the same calibrated model (Andrés Arcones et al., 2024). In contrast to these *external* correction approaches that associate the uncertainty with the predictions from the model, *internal* corrections arise as an alternative by attributing the uncertainty to its components instead (Wu et al., 2024). In this way, the inferred uncertainty can be pushed forward to any QoI that uses the same model parameters for its computation.

This family of approaches, also called parameter uncertainty inflation (PUI) methods (Pernot, 2017; Pernot and Cailliez, 2017) have been gaining traction in the last years. The most prominent approach is Sargsyan's parameter embedding (Sargsyan et al., 2015, 2019), which adds a stochastic dimension to the variables to be inferred and fits it together with the other parameters. This methodology has been successfully implemented in the context of ignition reaction (Huan et al., 2017). As alternatives to this methodology, Mortensen et al. (2005) proposes the statistical manipulation of the inferred parameter variance, and Wu et al. (2024) suggests inferring the internal model error structure through Kalman filters.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Alternatively, and related to these internal correction strategies, Stochastic Model Updating (SMU) methods have been developed to jointly address aleatory and epistemic uncertainties by updating the probabilistic descriptions of model parameters based on discrepancies between simulations and experimental data. These methods typically require multiple observations per prediction point to separate the two uncertainty sources and to estimate probability distributions directly. For example, Bi et al. (2019) proposed an SMU framework based on the Bhattacharyya distance within an Approximate Bayesian Computation (ABC) setting, while Lee et al. (2025) introduced a latent-space-based SMU formulation that leverages variational autoencoders to handle high-dimensional data efficiently. In contrast, the embedded model inadequacy framework adopted in this work does not require multiple observations to account for the epistemic uncertainty, as the aleatoric part is prescribed or inferred as a parameter. This distinction makes the embedded approach particularly suited to simulation-based model calibration problems where repeated experimental measurements are rarely feasible.

The aforementioned formulations have the advantage of offering a flexible implementation that can be adapted to the requirements of the problem. Despite being the most promising approach for dealing with model error through an internal correction, Sargsyan's proposal requires further refinement for complex inference cases (Pernot, 2017). This paper aims to address the challenges that arise when using the embedded approach for the calibration of simulation models from sensor data, and in particular a) the impact of misspecified noise models, b) the presence of observations that cannot be explained with modifications of the estimated parameters, and c) how the choice of likelihood affects the uncertainty propagation to other QoIs. A modification in how the prescribed noise is handled is introduced in the two main likelihood formulations proposed in Sargsyan et al. (2015) for the embedded problem—the Independent Normal (IN) and Approximate Bayesian Computation (ABC) likelihoods. The IN likelihood tends to overemphasize low variance locations (Sargsyan et al., 2015) and it will be shown that ABC is specially sensitive to misspecified noise and locations where the model discrepancy cannot be reduced by variations in the parameters. Two new formulations based on the statistical convergence of the residual distribution aimed to alleviate the flaws of IN and ABC are proposed in this paper: the Global Moment-matching (GMM) and Relative Global Moment-matching (RGMM) likelihoods. Further insight is provided for the choice in the control parameters of the likelihoods and their potential shortcomings. Finally, an analysis of the propagation of the uncertainties through QoI is developed. The QoIs obtained from the computational model are observed by *virtual* sensors (Andrés Arcones et al., 2023). In contrast with *real* sensors that are installed in the physical system, virtual sensors observe values provided by the computational model. They are not limited to observable quantities as the real sensors, but can represent any QoI. Quantifying the uncertainty in the output of virtual sensors is key for an assessment of the reliability of the predictions of QoIs.

The advantages and disadvantages of each formulation will first be illustrated on a one-dimensional, linear example with different variations that reflect the possible errors in the observations. Afterwards, a more complex transient thermal model simulating the heat transfer through a reinforced concrete structure is investigated. The objective of the more complex case is to test the propagation of the inferred uncertainty using each different likelihood to a non-related QoI. The structure of the rest of the paper is as follows: Section 2 presents the embedded approach, the likelihood functions and the proposed extensions, Section 3 includes the simple, linear model with its variations (Section 3.1) and the complex transient thermal one (Section 3.2), and conclusions and possible extensions are provided in Section 4. The methods here implemented are included in the Python package `probeye` ([dataset] BAMResearch, 2024).

2. Methodology

2.1. Model form uncertainty framework

Let f be a computable, deterministic function $f : \Theta \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_z}$ for a n_x-dimensional input $x \in \mathbb{R}^{n_x}$, parametrized through a set of n_θ parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{n_\theta}$, which have been restricted to real values for

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

simplicity. This function f models the real response $z:\mathbb{R}^{n_x}\rightarrow\mathbb{R}^{n_z}$, defined as a real-valued map that generates the n_z-dimensional response of a real system. In that case, an additional error term $\varepsilon_{\text{model}}=z-f$ rooted in the inability of f to exactly reproduce z is introduced. This model inadequacy term $\varepsilon_{\text{model}}$ will be referred to as *model form uncertainty.* If the same response is measured through sensor observations y, a noise $\varepsilon_{\text{noise}}$ will introduce an additional model inadequacy with z. The measurement noise $\varepsilon_{\text{noise}}$ is typically prescribed by the sensor manufacturer, often assuming an additive, homoscedastic Gaussian error model with known variance $(\sigma_N)^2$, whereas in other situations, its statistical properties are estimated directly from repeated or reference data observations. In this work, we generally adopt the former case unless explicitly stated otherwise, as it enables a tractable analytical treatment of the noise contribution consistent with the objectives of this study. The relation between observations and computational model predictions can be expressed as

$$y=z(x)+\varepsilon_{\text{noise}}=f(\theta,x)+\varepsilon_{\text{model}}+\varepsilon_{\text{noise}}, \tag{1}$$

modeling the model inadequacy terms additively. This formulation was introduced in the seminal paper of Kennedy and O'Hagan (2001), where they state the need for including the model error in the calibration of computational models. Alternative formulations are possible, such as a multiplicative relation between model inadequacy terms and model predictions.

The objective of the so-called *inverse problem* is to estimate the values for the parameters $\boldsymbol{\theta}^*$ such that f best approximates z as

$$\boldsymbol{\theta}^*=\arg\min_{\boldsymbol{\theta}\in\Theta}\|z-f(\theta,\cdot)\|, \tag{2}$$

where $\|\cdot\|$ is a distance to be defined later. As z is not generally known, it is commonly substituted by the set of n_y sensor observations $\boldsymbol{y}$ for known input parameters $\boldsymbol{x}$, where bold notation will be used to denote vector quantities. The inverse problem then transforms to

$$\boldsymbol{\theta}^*=\arg\min_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{y}-f(\theta,\boldsymbol{x})\|. \tag{3}$$

Bayesian approaches are a popular choice to solve the inverse problem (Kaipio and Somersalo, 2006). They provide a posterior probability density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ on the parameters $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}$ and the computational model $\boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{x})$. The application of Bayes' theorem yields

$$\pi(\boldsymbol{\theta}|\boldsymbol{y})=\frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \tag{4}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is the *posterior* probability distribution, $\pi(\boldsymbol{\theta})$ is the *prior* probability distribution that encompasses the previous knowledge on the unknown parameters, $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is the *likelihood* of the observations $\boldsymbol{y}$ having been generated by $\boldsymbol{\theta}$ and $\pi(\boldsymbol{y})$ is the *marginal* probability distribution of the observations. We assume the prior distribution $\pi(\boldsymbol{\theta})$ to be given, such that the remaining key challenge is the choice and evaluation of the likelihood function $\mathcal{L}(\boldsymbol{\theta})=\pi(\boldsymbol{y}|\boldsymbol{\theta})$ for the given set of observations $\boldsymbol{y}$. Bayesian inference approaches are based on the evaluation of the relationship from Equation 4 to obtain the posterior distribution of the unknown parameters $\boldsymbol{\theta}$.

Once the posterior distribution of the unknown parameters is obtained, it can be pushed forward to generate predictions $\boldsymbol{f}_P(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})$ using the same forward model f or they can be used in a different model $g(\boldsymbol{\theta}):\Theta\rightarrow\mathbb{R}$, for example, that computes a QoI using the same model parameters but evaluating a different quantity. In the first case $\boldsymbol{f}_P(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})$ can usually be compared with $\boldsymbol{y}$ and indirectly $\boldsymbol{z}(\boldsymbol{x})$ through $\varepsilon_{\text{noise}}$.

The calculation of $\mathcal{L}(\boldsymbol{\theta})$ requires the evaluation of the computational model f under the assumption that it generates the real system's response z. However, that is not generally the case and the model error must be considered. Kennedy and O'Hagan (2001) propose to include the model form uncertainty $\varepsilon_{\text{model}}$ in the inference procedure, formulated as a Gaussian Process added to the predicted response as in Equation 1. Despite its potential applications, this model's inadequacy term lacks physical meaning and cannot be employed outside of the use case with which it is inferred. Therefore, the uncertainty

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

quantified by $\varepsilon_{\text{model}}$ cannot be propagated to other QoI through the model g. Notice that this model inadequacy term does not aim to identify the deficiencies in the model and correct them, but to introduce a term that envelops their effects, such that the response is corrected.

More importantly, adding the model inadequacy term to the predicted outputs does not address one of the main challenges of using classical Bayesian inference approaches with imperfect models: the posterior distributions of the unknown parameters collapse to Dirac-delta distributions, which produce model responses that do not correspond with the system that generated the observation. This is commonly known as the problem of misspecification. The Bernstein–von Mises theorem states that under a set of conditions of continuity, differentiability and non-singularity, the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ converges in total variation (TV) distance with the true generating process $\pi_{\boldsymbol{\theta}_0}$ to a multivariate normal distribution centered at the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ and covariance matrix $\left(n_y\right)^{-1}\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$. Here, n_y is the number of observations in $\boldsymbol{y}$ and $\mathcal{I}(\boldsymbol{\theta}_0)$ is Fisher's information matrix at the true values $\boldsymbol{\theta}_0$ of the unknown parameters (van der Vaart, 2000). Formally, this can be formulated as

$$\left\|\pi(\boldsymbol{\theta}|\boldsymbol{y}) - \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \left(n_y\right)^{-1}\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\right)\right\|_{TV} \overset{\pi_{\boldsymbol{\theta}_0}}{\rightarrow} 0. \tag{5}$$

This result is crucial for constructing confidence intervals for the parameters and predictive responses, linking Bayesian and frequentist statistics. For large values of n_y, the posterior distribution becomes concentrated around the MLE $\hat{\boldsymbol{\theta}}$. However, Kleijn and van der Vaart (2012) demonstrated that the confidence intervals derived from Bernstein–von Mises theorem application only reflect the real credibility of the predictions in the case of perfect models that can reproduce the observations exactly.

In the case of imperfect models, where discrepancies exist between the model and the data, the posterior distribution still converges to a multivariate normal distribution that concentrates around the MLE $\hat{\boldsymbol{\theta}}$ for large n_y. However, $\hat{\boldsymbol{\theta}}$ may not generate the observations due to the model inadequacy, even for large n_y, and therefore the associated confidence intervals for the predictive response cannot be interpreted as credible intervals with respect to the true system, as demonstrated by Kleijn and van der Vaart (2012). This implies that the predicted distributions do not adequately capture the variability of the true system, making them unsuitable for quantifying system uncertainty.

When the model inadequacy vector δ_{model}, that models $\varepsilon_{\text{model}}$ at the observation points, is implemented as a correction to the model response, the combined model $\boldsymbol{f}_P + \delta_{\text{model}}$ can reproduce exactly the observations $\mathbf{y}$. As a result, the Bernstein–von Mises theorem holds within the domain of the parameters updated during this process. However, the model $\boldsymbol{f}_P$ alone cannot reliably quantify uncertainty without correcting the $\varepsilon_{\text{model}}$, as it is limited to the observations' domain. Similar to the scenario where no model inadequacy term is used, the model produces overly concentrated posterior distributions for both the parameters and the response, which may fail to represent the observations accurately. As the computation of QoIs through $g(\boldsymbol{\theta})$ usually depends only on the parameters and not on the corrected model response $\boldsymbol{f}_P + \varepsilon_{\text{model}}$, the uncertainty in the QoI will not be representative of the credibility of the model.

A promising solution to these challenges is the use of an embedded formulation where the inadequacy is added to the model through the unknown parameters. The objective is to augment those parameters with an additional stochastic variable that introduces random variations in the pre-existing model parameters. It is the introduction of this variability what prevents the unknown parameters from presenting a concentrated posterior for large n_y. Following the same philosophy as Kennedy and O'Hagan's (KOH) framework, no corrections are introduced in the structure of the physical model which allows the approach to be used non-intrusively with parametrized black-box models. This embedding is presented through Sections 2.2 to 2.4, the likelihood definition and evaluation in Sections 2.5 and 2.6, and the calculation of predicted variables through $\boldsymbol{f}_P$ and QoIs through g are presented in Section 2.7.

2.2. Embedding of model form uncertainty

The objective of the embedding is that, despite the unavoidable collapse of the posterior distributions of the unknown parameters to Dirac-deltas, the distribution of the response should preserve the variance

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

present in the observations. This is achieved by treating the unknown parameters $\boldsymbol{\theta}$ as random variables with known distributions. The objective is now to infer the values of the parameters of the distribution. We will denote these random variables as $\tilde{\boldsymbol{\theta}}$. This idea is similar to the one employed in hierarchical Bayesian approaches, where the prior distribution of $\boldsymbol{\theta}$ is unknown. This turns the originally deterministic model into a stochastic one. With this approach, the extended model's response $f(\tilde{\boldsymbol{\theta}})$ does not necessarily collapse to a Dirac-delta distribution, as a given sample of the inferred parameters still generates a non-concentrated distribution of the random variable $\tilde{\boldsymbol{\theta}}$.

There are different ways to model the random vector $\tilde{\boldsymbol{\theta}}$. One well-known approach, employed in Sargsyan et al. (2015, 2019), uses Polynomial Chaos Expansions (PCE) to represent the stochastic, unknown parameters. Specifically, the parameter vector $\tilde{\boldsymbol{\theta}}$ is expressed as a series expansion as

$$\tilde{\boldsymbol{\theta}} \sim \sum_j \alpha_j \Psi_j(\boldsymbol{\xi}), \tag{6}$$

where $\alpha_j \in \mathbb{R}^{n_\theta}$ are the expansion's coefficients (included in the unknown parameter space Θ), $\Psi_j : \mathbb{R}^{n_\xi} \to \mathbb{R}$ are orthogonal polynomials, and $\boldsymbol{\xi}$ is a set of stochastic variables (also called the stochastic germ) with $\boldsymbol{\xi} \in \mathbb{R}^n_\xi$ where n_ξ is the number of input random variables considered. This PCE-based approach allows for a flexible representation of the distribution of $\tilde{\boldsymbol{\theta}}$.

In this paper, instead of representing $\tilde{\boldsymbol{\theta}}$ as a polynomial expansion, we adopt an explicit representation where the stochasticity is introduced through an additive model inadequacy term. This approach is in line with the proposal of Oliver et al. (2015) for model error characterization and the work of Strong and Oakley (2014) on internal model discrepancies decomposition. Specifically, we model the random vector $\tilde{\boldsymbol{\theta}}$ as

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^m + \delta(\boldsymbol{\theta}^b), \tag{7}$$

where $\boldsymbol{\theta}^m$ represents the deterministic part, and $\delta(\boldsymbol{\theta}^b)$ is a parameterized zero-mean random vector that captures the stochastic deviation from $\boldsymbol{\theta}^m$. The parameters $\boldsymbol{\theta}^b$ control the magnitude of the stochasticity, and the explicit form of $\delta(\boldsymbol{\theta}^b)$ is chosen based on prior assumptions about the model inadequacy structure. In practice, this is equivalent to a first degree PCE using Sargsyan's approach. We advocate, however, for explicitly defining the structure of the additional variables associated with the uncertainty and their relation to the original variables, instead of using the coefficients of the PCE directly.

For simplicity, and because the exact shape of the distribution is often unknown, we assume that the model inadequacy δ is a random variable that depends on $\boldsymbol{\theta}^b$, following a joint normal distribution. Specifically, we assume

$$\delta(\boldsymbol{\theta}^b) \sim \mathcal{N}(0, \mathrm{diag}(\boldsymbol{\theta}^b)), \tag{8}$$

where $\boldsymbol{\theta}^b \in \mathbb{R}^{n_\theta}$ controls the variance of each parameter in $\tilde{\boldsymbol{\theta}}$. This formulation offers a clear interpretation of the model inadequacy associated with each parameter, making it easy to identify the sources of uncertainty and to calibrate the model accordingly. If the parameters are constrained to be positive or known to follow a different distribution, this assumption may need to be revisited. Nevertheless, it is often possible to transform other distributions to a normal through isoprobabilistic transformations, e.g. the Nataf transform. In particular, multi-modality in the parameter distributions must be supported explicitly in the formulation of the stochastic extension if it is expected. The need to explicitly define a joint distribution for the additional uncertainty parameters is one of the main limitations compared with the PCE approximation. However, it allows more control over the structure of the uncertainty, which may be valuable for complex scenarios with interacting variables, such as those described in Strong and Oakley (2014).

Unlike the PCE approach, which implicitly incorporates model inadequacy through a series expansion and introduces many unknown parameters for the PCE coefficients, an explicit model inadequacy representation offers a more direct and interpretable description of uncertainty. This is

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

particularly valuable for model calibration, where identifying the sources of model inadequacy is crucial. Additionally, since the calculation of QoIs may only require some of the calibrated parameters, defining the embedded model inadequacy $\boldsymbol{\theta}^b$ independently for each parameter ensures that only the uncertainty associated with the relevant parameter is transferred. Separating the model inadequacy representation from the inference process also enables a more transparent treatment of uncertainty, reducing the identifiability issues that can arise when simultaneously fitting both the inadequacy and model parameters.

For the remainder of this paper, we will use the formulation in Equation 7, assuming that $\delta(\boldsymbol{\theta}^b) \sim \mathcal{N}(0, \mathrm{diag}(\boldsymbol{\theta}^b))$. Non-additive inadequacy terms or other probability distributions are possible, but they do not affect the generality of the method presented here. Variance structures with correlation can analogously be considered for δ. To generate a stochastic response from the computational model, the probability distribution of $\tilde{\boldsymbol{\theta}}$ must be pushed through the model. This presents computational challenges, as the forward model must be evaluated for each sample of $\tilde{\boldsymbol{\theta}}$. To address this, we implement a PCE approximation of the model's response while keeping the prescribed stochastic structure of the embedding, which is detailed in Section 2.3.

Finally, because the model outputs are now stochastic, traditional inference methods that rely on deterministic outputs are no longer applicable. In Section 2.5, we introduce likelihood formulations inspired by Approximate Bayesian Computation (ABC) methods, which leverage summary statistics to infer the unknown parameters of the model. These methods provide a robust framework for handling stochastic models, enabling a direct comparison between the stochastic model response and the available data. Instead of replicating individual observations exactly, the approach focuses on fitting their statistical moments, resulting in more reliable predictions and avoiding overfitted posterior distributions.

In Section 3, we demonstrate how introducing the stochastic extension significantly improves the probability of the observations being generated by the posterior-predicted model. This improvement is measured through the Mahalanobis distance between the predicted and observed data and is shown to outperform standard Bayesian approaches.

2.3. Forward model evaluation and polynomial chaos expansion (PCE)

The evaluation of the stochastic response of the forward model requires the propagation of the uncertainty introduced by the embedding. A closed-form representation of the response is generally not possible, therefore sampling-based methods are required. A very popular approach for uncertainty propagation in recent years has been the use of generalized polynomial chaos expansions (PCE) (Xiu and Karniadakis, 2002) to approximate the stochastic response once the uncertainty has been propagated. This approach consists of approximating the response f by a linear combination of a basis of orthonormal polynomials $\{\Psi_j\}_{j=0}^{D}$ truncated at degree d.

Let $\Psi_j : \mathbb{R}^{n_\xi} \to \mathbb{R}$ be the number of input random variables, where $\xi \in \mathbb{R}^{n_\xi}$ is the stochastic germ that determines the realization of a given random variable with probability distribution function π_ξ and $n_\xi = n_b$. These stochastic germs must follow an independent distribution inherited from θ. For example, if θ is normal, ξ must also follow a normal distribution. The assumption of independent input random variables, which can be enforced through the use of isoprobabilistic transformations such as Nataf or Rosenblatt transforms (Jakeman et al., 2019). The number of polynomial bases $D+1$ is computed as $D+1 = \binom{d+n_b}{d}$, where n_b is the number of input random variables to be considered.

Then, for each output $f_i(\theta) = f(\theta, x_i)$, which refers to the model response evaluated at the ith point x_i of the domain x (i.e., f_i is f evaluated at x_i with fixed θ), its approximation via PCE $\tilde{f}_i(\theta)$ is

$$f_i(\theta) \approx \tilde{f}_i(\theta) = \sum_{j=0}^{D} \alpha_{ij} \Psi_j(\xi). \tag{9}$$

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

As it can be observed, θ generates a single approximation for each $\tilde{f}_i(\theta)$, which are random variables that can be sampled by evaluating the PCE for different values of ξ. To express this more concisely, the approximation can be written in vector form $\mathbf{f}(\theta) = [f_1(\theta), f_2(\theta), \ldots, f_N(\theta)]$ as

$$\mathbf{f}(\theta) \approx \tilde{\mathbf{f}}(\theta) = \sum_{j=0}^{D} \boldsymbol{\alpha}_j \Psi_j(\xi), \tag{10}$$

where α_{ij} are the PCE coefficients associated with each polynomial Ψ_j that fit the approximation to f_i (Sudret, 2021). The choice of the base of orthonormal polynomials Ψ depends on the distribution of the input random variables to be propagated. Askey's scheme (Xiu and Karniadakis, 2002) assigns popular distributions with their corresponding polynomial basis that ensures orthonormality and numerical stability.

Computing the α_{ij} coefficients requires solving the problem of minimizing the L_2-distance between $\mathbf{f}(\theta, x)$ and $\tilde{\mathbf{f}}(\theta, x)$ as in

$$\min_{\alpha_{ij}} \left\| \mathbf{f}(\theta, x) - \tilde{\mathbf{f}}(\theta, x) \right\|_2. \tag{11}$$

Here, the α coefficients depend on x, and the norm $\| \cdot \|_2$ is defined as the L_2 norm, with the inner product given by

$$\langle f, g \rangle = \int f(\xi) g(\xi) \pi_\xi(\xi) d\xi, \tag{12}$$

where $\pi_\xi(\xi)$ is the probability density function associated with ξ. As $\langle \Psi_i, \Psi_j \rangle = \delta_{ij}$ due to the orthonormality of the polynomial basis, where δ_{ij} is Dirac's delta function, the PCE coefficients that minimize the L_2-distance are given by

$$\alpha_{ij} = \int_{\xi} f_i(\theta(\xi)) \Psi_j(\xi) \pi_\xi(\xi), d\xi, \tag{13}$$

where $\pi_\xi(\xi)$ is the probability density associated with the random variable ξ. This integral can be computed using a Gauss quadrature scheme with weights $w_1, w_2, \ldots, w_P$ and nodes $\xi_1, \xi_2, \ldots, \xi_P$, where $P = p^{n_b}$ is the number of quadrature points. The expansion coefficients are then computed as

$$\alpha_{ij} = \sum_{k=1}^{P} w_k f_i(\theta(\xi_k)) \Psi_j(\xi_k), \tag{14}$$

where the factor $\pi_\xi(\xi_k)$ is already included in the quadrature weights w_k.

The full process to evaluate a sample through the forward model is summarized in Algorithm 1. The Python package `chaospy` (Feinberg and Langtangen, 2015) has been used for the PCE implementation in this paper. The result of such evaluation is the PCE of the forward model's response, which is stochastic and requires post-processing. The polynomial formulation provides direct access to the statistical moments of the response. Let μ^f be the vector of means of the response $\tilde{\mathbf{f}}(\theta, x)$. Then, each of its entries can be obtained as

$$\mu_i^f = \mathbb{E}\left[\tilde{f}_i(\theta)\right] = \alpha_{i0}, \tag{15}$$

which holds under the assumption that $\Psi_0 = 1$. Analogously, the vector of variances σ^f is computed as

$$\left(\sigma_i^f\right)^2 = \mathrm{Var}\left[\tilde{f}_i(\theta)\right] = \sum_{j=1}^{D} \left(\alpha_{ij}\right)^2, \tag{16}$$

because the variance is the sum of squares of the non-constant coefficients.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Algorithm 1 Forward model evaluation with polynomial chaos expansion (PCE) and pseudo-spectral projection

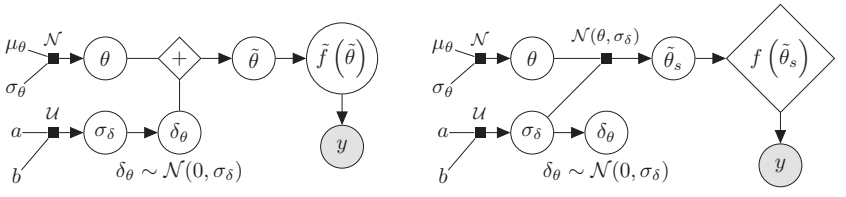
Given a sample θ:
Step 1. Build a joint distribution $\mathcal{J}$ of the stochastic unknown parameters θ^b.
Step 2. If $\mathcal{J}$ is not composed of independent variables, perform an isoprobabilistic transformation to make them independent, e.g., Nataf or Rosenblatt transform.
Step 3. Compute the weights $w_1, w_2, \ldots, w_P$ and nodes $\xi_1, \xi_2, \ldots, \xi_P$ for the Gauss quadrature scheme of degree p given $\mathcal{J}$.
Step 4. Generate the orthonormal polynomial basis Ψ of degree d using Askey's scheme.
Step 5. Evaluate the forward model at the nodes ξ for the sampled θ.
Step 6. For each entry in $\mathbf{f}$, compute the PCE coefficients α_j using Gauss integration with weights w, nodes ξ, and the corresponding model evaluations.

2.4. Embedded model form uncertainty as a hierarchical Bayesian problem

The inverse problem with an embedded model form uncertainty can be interpreted within a hierarchical Bayesian framework. First, a set of *hyperpriors* is prescribed for $\boldsymbol{\theta}^m$ and $\boldsymbol{\theta}^b$, from which they are sampled. Then, a prior distribution for $\boldsymbol{\theta}$ is defined through Equation 7. Finally, the forward model is evaluated based on $\boldsymbol{\theta}$. This structure follows the sequential definition of unknown variables that typically appears in hierarchical Bayesian models (Robert, 2007). This formulation can also be extended to use Approximate Bayesian Computation (ABC) (Turner and Van Zandt, 2013), which has been applied to inferring the model inadequacy in industrial electric motor simulations (John et al., 2021; John, 2021).

We highlight the key difference between embedded and most hierarchical formulations: while hierarchical approaches factorize the parameters, sampling them and inputting these samples into the model to generate a deterministic response, the embedded formulation directly propagates the distributions of the parameters, resulting in a stochastic response. These distinctions are illustrated in Figure 1, where the inadequacy in a single parameter $\tilde{\theta}$ is modeled such that $\tilde{\theta} = \theta^m + \delta(\theta^b)$, where $\delta \sim \mathcal{N}(0, \sigma_\delta)$. Equivalently, $\tilde{\theta} \sim \mathcal{N}(\theta^m, \sigma_\delta)$ for this simple case. This comparison highlights that both the embedded and hierarchical formulations share similar modeling principles but differ fundamentally in how the forward model response is computed. The embedded approach handles the forward model stochastically, leveraging polynomial chaos expansion (PCE) to propagate uncertainty efficiently, while the hierarchical Bayesian approach results in a deterministic forward model after sampling $\tilde{\theta}$.

While both methodologies have practical applications, an exhaustive comparison between them is beyond the scope of this paper. Pernot and Cailliez (2017) compared, among others, an ABC implementation of the embedded approach and a hierarchical one, concluding that the former is a better fit for the



(a) Embedded approach (b) Hierarchical Bayesian approach

Figure 1. *Bayesian graph for the inference of the parameters involved in the embedded formulation of the model form uncertainty. (a) Embedded approach. (b) Classical hierarchical Bayesian approach with Gibbs' sampling. Following usual notation (Dietz, 2010; Obermeyer et al., 2019), circled values with white background represent unknown variables, circled shaded values represent observations, rhomboids represent deterministic operations, and black squares represent drawing a sample from the indicated distribution.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

estimation of parameters under model form uncertainty. For improved results, the hierarchical formulation required a local modification of the parameters, effectively increasing their number. For a more comprehensive discussion on hierarchical Bayesian approaches and their computational challenges, refer to Robert and Casella (2004); Robert (2007); Turner and Van Zandt (2013); Li et al. (2024).

2.5. Likelihood formulation

Due to the uncertainty propagation, the output $\boldsymbol{f}(\boldsymbol{\theta})$ of evaluating the forward model with a given sample $\boldsymbol{\theta}$ or its PCE approximation $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ is a stochastic random variable that follows a probability distribution described by the PCE approximation. While a modified approach that utilizes the statistical moments of the distribution is feasible, directly evaluating a classical Gaussian likelihood function based on the residuals between the predictions, $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$, and the observations, $\mathbf{y}$, is not possible because their domains are not directly comparable, $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ being a random variable and $\mathbf{y}$ being deterministic observations. Approximate Bayesian Computation (ABC) approaches provide a methodology to obtain the likelihood $\pi_{\mathrm{ABC}}(\mathbf{y}|\boldsymbol{\theta})$ from the statistical moments of $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$, which are readily available from the PCE, enabling the formulation of a likelihood function.

Throughout this paper, an ensemble-based Monte Carlo-Markov Chain (MCMC) sampler using a stretch move (Goodman and Weare, 2010) from the package `emcee` (Foreman-Mackey et al., 2013) is used. A threshold on the *effective sample size* (*ESS*) for each of the unknown parameter chains as described in Appendix A is set as stopping criteria for the MCMC algorithm.

2.5.1. ABC-likelihood without noise

Classical ABC approaches usually include two assumptions: (1) measurements are free of noise, where $\varepsilon_{\mathrm{noise}}$ is assumed 0, and (2) the model response can fully represent the system, where $\varepsilon_{\mathrm{model}}$ is assumed 0. Therefore, the model response is expected to fully represent the system's real behavior as $\mathbf{z}=\boldsymbol{f}(\boldsymbol{\theta})$. Those approaches define a residual function $\rho\left(\mathbf{y},\tilde{\boldsymbol{f}}(\boldsymbol{\theta})\right):\mathbb{R}^{n_y}\times\mathbb{R}^{n_z}\rightarrow\mathbb{R}$ between observations $\mathbf{y}$ and predictions $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$, which usually defines a distance, and aim at reducing it below a tolerance value ε by the following steps (Sisson et al., 2018):

1. Sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta})$.
2. Evaluate the forward model $\mathbf{z}=\boldsymbol{f}(\boldsymbol{\theta})\sim p(\mathbf{z}|\boldsymbol{\theta})$.
3. Accept or reject $\boldsymbol{\theta}$ if $\rho(\mathbf{y},\boldsymbol{f}(\boldsymbol{\theta}))<\epsilon$.

If the model is approximated, such as the forward model built in Section 2.3, its evaluation would correspond to $\tilde{\mathbf{z}}=\tilde{\boldsymbol{f}}(\boldsymbol{\theta})\sim p(\tilde{\mathbf{z}}|\boldsymbol{\theta})$, where the model form uncertainty is expected to be fully represented in the stochastic response of the computational model, preserving assumption (2). When this tolerance $\epsilon\rightarrow 0$, the predictions are expected to reproduce exactly the observations, and therefore the samples of θ would belong to the associated posterior distribution. As $\mathbf{y}$ is deterministic and $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ stochastic, the residual function ρ is evaluated on its statistical moments t such that

3a. Accept or reject $\boldsymbol{\theta}$ if $\rho\left(t(\mathbf{y}),t\left(\tilde{\boldsymbol{f}}(\boldsymbol{\theta})\right)\right)<\epsilon$.

A possibility is to use the mean and standard deviation as summary statistics under the assumption of them being statistically *sufficient.* The predictions $\tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ are approximated at each output dimension by $t\left(\tilde{\boldsymbol{f}}(\boldsymbol{\theta})\right)=\left[\boldsymbol{\mu}^f,\boldsymbol{\sigma}^f\right]$. The output observations $\mathbf{y}$ are approximated by its statistical moments as $t(\mathbf{y})=[\boldsymbol{\mu},\boldsymbol{\sigma}]$, where $\boldsymbol{\mu}\approx\mathbf{y}$ and $\boldsymbol{\sigma}\approx\gamma\left|\boldsymbol{\mu}^f-\mathbf{y}\right|$ with γ an additional factor to be specified. The approximation $\boldsymbol{\sigma}\approx\gamma\left|\boldsymbol{\mu}^f-\mathbf{y}\right|$ is necessary due to only one observation being available. A reasonable assumption would be that on, average, the observations $\mathbf{y}$ are located a given distance from the mean $\boldsymbol{\mu}^f$ controlled by their standard deviations $\boldsymbol{\sigma}$, leading to the previous approximation. The choice of summary statistics is not unique and introduces a potential model inadequacy rooted in the approximation that is not present in the formulation (Fearnhead and Prangle, 2012).

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

The acceptance criterion defines the *indicator function* or *kernel* that governs the sampling procedure. Following the criteria from Step 3, the uniform kernel I indicates if a sample is accepted and can be defined as

$$I(\boldsymbol{\theta},\boldsymbol{y}) = \begin{cases} 1 & \text{if} \rho\left(t(\boldsymbol{y}), t\left(\tilde{\boldsymbol{f}}(\boldsymbol{\theta})\right)\right) < \epsilon \\ 0 & \text{else} \end{cases}. \tag{17}$$

Applying Bayes' theorem with the chosen kernel, it is possible to express the approximate posterior distribution of the parameters as

$$\pi_{\text{ABC}}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathbb{R}^{n_z}} I(\boldsymbol{\theta},\boldsymbol{y}) \pi(\boldsymbol{z}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{z} \tag{18}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution for the unknown parameters $\boldsymbol{\theta}$ and $\pi(\boldsymbol{z}|\boldsymbol{\theta})$ the probability of the model output for a given $\boldsymbol{\theta}$. For a deterministic model f, the probability $\pi(\boldsymbol{z}|\boldsymbol{\theta})$ follows a Dirac's delta distribution located at the value for $\boldsymbol{f}(\boldsymbol{\theta})$. Analogously, the summary statistics $t(\tilde{\boldsymbol{z}})$ of the approximated output $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{f}}(\boldsymbol{\theta})$ are deterministic as well, therefore $\pi(t(\tilde{\boldsymbol{z}})|\boldsymbol{\theta})$ follows the same Dirac's delta distribution when they are used as a lower-dimensional representation of the model output. Under this consideration, it can be observed that the kernel acts as the likelihood function that is commonly defined in the classical Bayesian approaches. In particular, the likelihood defined by the aforementioned uniform kernel assigns probability 1 to those samples that fulfil $\rho(t(\boldsymbol{y}), t(\tilde{\boldsymbol{z}})) < \epsilon$ and 0 elsewhere.

Following the same reasoning, it is possible to implement other likelihood functions that reciprocally induce different kernels that govern the acceptance criteria. Such is the case for the *moment-matching* likelihood used in Sargsyan et al. (2015) and Sargsyan et al. (2019), which adopts an exponential structure:

$$\mathcal{L}^{\text{ABC}}(\boldsymbol{\theta}) = \left(2\pi\epsilon^2\right)^{-\frac{1}{2}} \prod_{i=1}^{n_y} \exp\left(- \frac{\left(\mu_i^f - y_i\right)^2 + \left(\sigma_i^f - \gamma\left|\mu_i^f - y_i\right|\right)^2}{2\epsilon^2} \right) \tag{19}$$

hence

$$\pi_{\text{ABC}}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) \mathcal{L}^{\text{ABC}}(\boldsymbol{\theta}). \tag{20}$$

This likelihood assigns higher probabilities to those samples that minimize the squared distance between the summary statistics $\left[\boldsymbol{\mu}^f, \boldsymbol{\sigma}^f\right]$ and the observations, defined as

$$\rho\left(t(\boldsymbol{y}), t\left(\tilde{\boldsymbol{f}}(\boldsymbol{\theta})\right)\right) = \begin{bmatrix} \boldsymbol{\mu}^f - \boldsymbol{y} \\ \boldsymbol{\sigma}^f - \gamma\left|\boldsymbol{\mu}^f - \boldsymbol{y}\right| \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\mu}^f - \boldsymbol{y} \\ \boldsymbol{\sigma}^f - \gamma\left|\boldsymbol{\mu}^f - \boldsymbol{y}\right| \end{bmatrix}. \tag{21}$$

While this distance has proven convergence (Prangle, 2017), it is primarily influenced by the statistical moment with the largest variance. Notably, the value of $\boldsymbol{\mu}^f$ is mainly affected by variations in the original unknown parameters, whereas $\boldsymbol{\sigma}^f$ is contingent upon the likelihood parameters associated with the embedded model inadequacy. The moment-matching likelihood is justified by the assumption of normality in the differences between the predicted summary statistics and those of the observations. The goal is to achieve an exact match between the predicted and observed statistical moments, which occurs as ϵ approaches 0, assuming there is no model inadequacy. When the measurement noise $\varepsilon_{\text{noise}}$ is either zero or can be completely absorbed by the model form uncertainty, the posterior predictive distribution generated using this likelihood converges to the true distribution, provided that the summary statistics are sufficient.

The choice of γ is also not trivial and depends on the assumptions imposed on the residuum's distribution. It can be argued that, supposing a normal independent probability distribution for the residuum, $\gamma = \sqrt{\frac{\pi}{2}}$ should be chosen over $\gamma = 1$ if the desired expected predictive distribution is supposed to have, on average, the magnitude of the residuum as the standard deviation. For a variable u that follows

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

a zero-mean normal distribution $u \sim \mathcal{N}(0,\sigma)$, its absolute value $|u|$ follows a half-normal distribution (Leone et al., 1961; Gelman et al., 2013), and thus

$$\mathbb{E}[|u|] = \sigma\sqrt{\frac{2}{\pi}}. \tag{22}$$

Therefore, if $\boldsymbol{\sigma}^f$ on average should be equal to the residuum $\boldsymbol{\mu}^f - \boldsymbol{y}$ at every point and not to the expectation of $|\boldsymbol{\mu}^f - \boldsymbol{y}|$, a correction of $\gamma = \sqrt{\frac{\pi}{2}}$ is required. In such case, the residuum obtained from the observations is, on average, one standard deviation $\boldsymbol{\sigma}^f$ from the mean. Sargsyan et al. (2019) proposes taking $\gamma = 1$ as a general case instead, specifying then $\boldsymbol{\sigma} \approx |\boldsymbol{\mu}^f - \boldsymbol{y}|$, observing that alternative values of γ should be informed by knowledge on the system (Sargsyan et al., 2019). However, for the likelihood of Equation 19, that would imply underestimating the predictive variance, hence we will employ $\gamma = \sqrt{\frac{\pi}{2}}$ when applying this ABC likelihood. Further discussion on the impact of modifying γ and ϵ for this likelihood can be found in Sargsyan et al. (2015). Equation 20 can be implemented in an MCMC framework to estimate the posterior distribution (Marjoram et al., 2003; Sisson and Fan, 2010).

2.5.2. ABC-likelihood with noise

The likelihood formulated in Equation 19 does not explicitly include the measurement noise that appears in the observations. It supposes that the moments can be matched exactly up to a constant ϵ, which in reality is generally not true. In cases where measurement noise is present, it is not possible to reduce ϵ to zero, which invalidates the moment-matching condition and introduces an error in the posterior. Therefore, the explicit inclusion of noise is required to generate a reliable posterior based on real sensor observations.

Three main alternatives have been proposed by Schälte and Hasenauer (2020) to include noise in ABC approaches: (1) controlling the noise through ϵ, (2) simulating samples from the noisy model $f(\theta) + \varepsilon_{\text{noise}}$, and (3) including the noise model in the sample acceptance criteria, and therefore the likelihood function. Fitting the tolerance to the expected noise is not trivial, as the variation in the summary statistics does not necessarily coincide with the prescribed measurement noise, and it limits the application to uniform noise models (Daly et al., 2017). Alternatively, simulating samples leads to low acceptance rates and largely increases the number of required model evaluations per sample. Finally, the modification of the acceptance criteria has been proven to correctly converge to the posterior of the summary statistics with noise (Wilkinson, 2013). Based on this property, a modification to include the noise in the moment-matching likelihood function is proposed here.

Following Wilkinson (2013), the acceptance step of the general ABC outline is modified to include the noise error:

3b. Accept $\boldsymbol{\theta}$ with probability $\frac{\pi_\varepsilon(\rho(t(\boldsymbol{y}), t(\tilde{\boldsymbol{z}})))}{c}$.

where $\pi_\varepsilon(\rho(t(\boldsymbol{y}), t(\tilde{\boldsymbol{z}})))$ represents the probability error model for the summary statistics, where $\rho(t(\boldsymbol{y}), t(\tilde{\boldsymbol{z}}))$ defines a residual function between them, typically a signed distance, and c is a normalization constant. Applying the MCMC framework and using $\pi_\varepsilon(\rho(t(\boldsymbol{y}), t(\tilde{\boldsymbol{z}})))$ as likelihood, the posterior distribution of the unknown parameters can be approximated as

$$\pi_{\text{ABC}}(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) \approx \pi_{\text{ABC}}(\boldsymbol{\theta}, t(\boldsymbol{z})|t(\boldsymbol{y})) \propto \pi_\varepsilon(\rho(t(\boldsymbol{y}), t(\tilde{\boldsymbol{z}})))\pi(\boldsymbol{\theta}). \tag{23}$$

We refer to Wilkinson (2013) and Schälte and Hasenauer (2020) for a comprehensive proof of the validity and convergence of such a family of likelihoods in an MCMC framework.

The chosen summary statistics are the same as for the original moment-matching likelihood: the mean and standard deviation of the model. The error $\varepsilon_{\text{noise}} = \boldsymbol{y} - \tilde{\boldsymbol{z}}$ follows a prescribed homogeneous Gaussian noise model $\mathcal{N}(0, \sigma_N \boldsymbol{I})$. Therefore, the noise associated with the means has variance $(\sigma_N)^2$. On the other hand, the matching of predicted and estimated standard deviations is considered exactly up to the tolerance ϵ. The predicted standard deviation now not only considers the variation in the output $\boldsymbol{\sigma}^f$ but also includes the noise vector $\boldsymbol{\sigma}_N$. Therefore, the second moment comparison with noise is

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

$\boldsymbol{\sigma} = \sqrt{(\boldsymbol{\sigma}^f)^2 + (\boldsymbol{\sigma}_N)^2} \approx \gamma|\boldsymbol{\mu}^f - \boldsymbol{y}|$. With this new formulation, the variance in the residuals $\boldsymbol{u} = |\boldsymbol{\mu}^f - \boldsymbol{y}|$ will be explained by the prescribed measurement error first, if possible, and the predicted variance would intuitively only be increased to cover for the remaining variance. For the aforementioned homogeneous Gaussian error, the noisy moment-matching likelihood function is

$$\mathcal{L}^{\mathrm{ABC}}(\boldsymbol{\theta}) = \left(2\pi\epsilon^2(\sigma_N)^2\right)^{-\frac{n_y}{2}} \prod_{i=1}^{n_y} \exp\left(-\frac{\left(\mu_i^f - y_i\right)^2}{2(\sigma_N)^2} - \frac{\left(\sqrt{\left(\sigma_i^f\right)^2 + (\sigma_N)^2} - \gamma\left|\mu_i^f - y_i\right|\right)^2}{2\epsilon^2} \right). \tag{24}$$

This modified likelihood presents several advantages in the case of measurements with noise. First and most prominently, the predicted and sampled means are no longer matched exactly, as their fitness is evaluated considering the known measurement noise structure. This avoids over-fitting to the samples and provides more informative posterior distributions that consider a larger part of the available information. Additionally, the variance from measurement noise is no longer absorbed by the stochastic part of the model prediction, which avoids overestimating the model form uncertainty.

Nevertheless, both moment-matching likelihoods are sensitive to badly represented models. As proven in Appendix B, if the prescribed noise is larger than the residuals, the inference procedure will converge to a posterior distribution that does not necessarily represent the real system. This occurs when the sampling algorithm favors samples with a larger residual in order to exactly match the standard deviation of the prediction with the prescribed noise model. However, this situation can be avoided by a careful selection of the prescribed noise model, as is generally the case in practice. Alternatively, the noise parameter can be inferred as an unknown parameter, which is possible only for the noisy moment-matching likelihood function.

The likelihood in Equation 24 assumes additive, homoscedastic Gaussian noise. This assumption is typical in engineering applications, where sensor manufacturers usually specify a constant noise variance. Other noise structures can be accommodated by modifying the likelihood accordingly. In particular, heteroscedasticity can be incorporated by replacing σ_N with the pointwise terms σ_{N_i} and by avoiding factorization of this term outside the summation. Non-Gaussian noise models may also be handled by jointly evaluating the predictive output $f(\tilde{\boldsymbol{\theta}})$ and the noise term $\varepsilon_{\mathrm{noise}}$ to compute their combined statistical moments. While the likelihood assumes normal model errors for the first and second moments, the underlying output distribution itself is not required to be Gaussian. This is a pragmatic simplification, justified by the Central Limit Theorem and commonly adopted in practice. The joint evaluation of prediction and noise can be carried out analytically, through Monte Carlo sampling, or by convolution of their characteristic functions. In cases where the output distribution is markedly non-Gaussian, using only the first and second moments for matching may lead to biased fits. Alternative formulations based on robust summary statistics, such as the median and the median absolute deviation (Huber and Ronchetti, 2009), or on distance metrics such as the Bhattacharyya distance (Bi et al., 2019) can provide more reliable results in such scenarios. The analysis on these modified cases is left for future work.

2.5.3. Independent normal likelihood

An alternative to moment-matching likelihoods is to adapt the classical Gaussian likelihood to take into consideration the predictive variance. Based on the assumption that the residuals follow an independent normal distribution with mean zero and standard deviation of the predicted one, Sargsyan et al. (2015) propose

$$\mathcal{L}^{\mathrm{IN}}(\boldsymbol{\theta}) = (2\pi)^{-\frac{n_y}{2}} \prod_{i=1}^{n_y} \left(\sigma_i^f\right)^{-1} \exp\left(-\frac{\left(y_i - \mu_i^f\right)^2}{2\left(\sigma_i^f\right)^2} \right). \tag{25}$$

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

If noise is accounted for in the variance computation, the likelihood must be modified accordingly, resulting in

$$\mathcal{L}^{\mathrm{IN}}(\boldsymbol{\theta}) = (2\pi)^{-\frac{n_y}{2}} \prod_{i=1}^{n_y} \left(\sqrt{\left(\sigma_i^f\right)^2 + (\sigma_N)^2} \right)^{-1} \exp\left(-\frac{\left(y_i - \mu_i^f\right)^2}{2\left(\left(\sigma_i^f\right)^2 + (\sigma_N)^2\right)} \right). \tag{26}$$

Heteroscedasticity can be included analogously as for Equation 24. However, this formulation would not be suitable for non-Gaussian or non-additive noise models without substantial modifications, as is the case with the classical Gaussian likelihood.

2.5.4. Global moment-matching (GMM) Likelihood with noise

The *global moment-matching likelihood* aims to avoid the fitting of the predictive moments to each y_i individually. This is achieved by considering the predictive distributions of $f^f(\boldsymbol{\theta},\boldsymbol{x}_i)$ centered at zero as the components of an equally weighted mixture model $F(\boldsymbol{\theta})$. Then, a likelihood function is formulated by treating the observed residuals $y_i - \mu_i^f$ as potential samples from $F(\boldsymbol{\theta})$. This approach utilizes the sampling distributions of the moments of a normal distribution to define π_ε in Equation 23.

First, we observe that at each single set of coordinates $\boldsymbol{x}_i$ for all $i = 1, 2, \ldots, n_y$ observation points, we can define a random variable $f^f(\boldsymbol{\theta},\boldsymbol{x}_i)$ with known moments $\mu^f(\boldsymbol{\theta},\boldsymbol{x}_i)$ and $\left(\sigma^f(\boldsymbol{\theta},\boldsymbol{x}_i)\right)^2 + (\sigma_N)^2$ that describes the predictions of the model at $\boldsymbol{x}_i$ for a given sample of the unknown parameters $\boldsymbol{\theta}$. An homoscedastic additive Gaussian noise structure is assumed. As in previous cases, an heteroscedastic noise structure can be introduced by defining σ_{Ni}. We center every observation by the predictive mean as

$$f^f(\boldsymbol{\theta},\boldsymbol{x}_i) - \mu^f(\boldsymbol{\theta},\boldsymbol{x}_i) \sim \mathcal{N}\left(0, \left(\sigma^f(\boldsymbol{\theta},\boldsymbol{x}_i)\right)^2 + (\sigma_N)^2\right). \tag{27}$$

Then, we define $F(\boldsymbol{\theta})$ as the mixture of the centered predictive distributions corresponding to all observation locations as

$$F(\boldsymbol{\theta}) = \frac{1}{n_y} \sum_{i=1}^{n_y} \mathcal{N}\left(0, \left(\sigma^f(\boldsymbol{\theta},\boldsymbol{x}_i)\right)^2 + (\sigma_N)^2\right), \tag{28}$$

where the summation denotes an equally weighted mixture of the predictive distributions. This construction represents the population-level predictive distribution obtained by pooling all local predictive variances, rather than a linear combination of random variables. When the component variances differ, $F(\boldsymbol{\theta})$ is not strictly normal but provides a valid mixture representation of the predictive uncertainty. The moments of $F(\boldsymbol{\theta})$ can be directly computed as

$$t_1(F(\boldsymbol{\theta})) = \mu_F = \mathbb{E}[F(\boldsymbol{\theta})] = 0 \tag{29}$$

$$t_2(F(\boldsymbol{\theta})) = (\sigma_F)^2 = \mathrm{Var}(F(\boldsymbol{\theta})) = \frac{1}{n_y} \sum_{i=1}^{n_y} \left(\left(\sigma^f(\boldsymbol{\theta},\boldsymbol{x}_i)\right)^2 + (\sigma_N)^2 \right) = \overline{\left(\left(\sigma^f(\boldsymbol{\theta},\boldsymbol{x}_i)\right)^2\right)} + (\sigma_N)^2, \tag{30}$$

where, following usual notation, $\mathbb{E}[\cdot]$ denotes expectation of $\cdot$, $\mathrm{Var}(\cdot)$ denotes variance of $\cdot$, and $\bar{\cdot}$ denotes the average of $\cdot$ over all i.

We choose $F(\boldsymbol{\theta})$ to represent the population. We observe that these are estimates of the population, as its true moments are not available only from samples. In this formulation, Cochran's theorem applies approximately because we assume the residuals are sufficiently similar (i.e., approximately homoscedastic) and due to the large value of n_y, allowing the use of the Central Limit Theorem to justify the normality of the sampling distributions of the moment estimators. This is usually not the case and limits in practice the applicability of this likelihood formulation. In particular, the distribution of the sampled mean and variance from $F(\boldsymbol{\theta})$ can be approximated as

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

$$\mu_U - \mu_F \sim \mathcal{N}\left(0, \frac{(\sigma_F)^2}{n_y}\right) \tag{31}$$

$$\frac{n_y(\sigma_U)^2}{(\sigma_F)^2} \sim \chi^2_{n_y-1}, \tag{32}$$

where μ_U and $(\sigma_U)^2$ are the moments of the samples and $\chi^2_{n_y-1}$ is a chi-square distribution of $n_y - 1$ degrees of freedom, which must be obtained from the residuals $u_i = y_i - \mu^f(\boldsymbol{\theta}, \boldsymbol{x}_i)$ with $i = 1, 2, \ldots, n_y$, realizations of the random variable $U \in \mathbb{R}$. The probability distribution of the error based on the samples can then be formulated as

$$\pi_\varepsilon(\rho(t(U(\boldsymbol{\theta})), t(F(\boldsymbol{\theta})))) = \pi_{t_1}(t_1(U(\boldsymbol{\theta})) - t_1(F(\boldsymbol{\theta})))\pi_{t_2}\left(\frac{t_2(U(\boldsymbol{\theta}))}{t_2(F(\boldsymbol{\theta}))}\right). \tag{33}$$

In this case, the first moment t_1 will be the mean and t_2 will be the variance of the set of samples of x. We will use Equations 31 and 32 as error models for π_{t_1} and π_{t_2}, respectively; requiring modeling of the error in π_{t_1} as a difference and in π_{t_2} as a quotient. The mean $t_1(U(\boldsymbol{\theta})) = \mu_U$ and variance $t_2(U(\boldsymbol{\theta})) = (\sigma_U)^2$ are directly estimated from the samples as

$$\hat{t}_1(U(\boldsymbol{\theta})) = \hat{\mu}_U = \overline{\boldsymbol{u}} \tag{34}$$

$$\hat{t}_2(U(\boldsymbol{\theta})) = (\hat{s}_U)^2 = \frac{\sum_{i=1}^{n_y}(u_i - \overline{\boldsymbol{u}})^2}{n_y - 1}, \tag{35}$$

where $\overline{\boldsymbol{u}}$ represents the mean of the residuals vector $\boldsymbol{u}$. The dependency of $\boldsymbol{u}$ and its statistical moments on $\boldsymbol{\theta}$ will not be explicitly stated for notation purposes. Assuming that $(\sigma_F)^2$ is a good approximation for the population variance, the sampled variance $(\sigma_U)^2$ is a sufficient statistic for $(\sigma_F)^2$. Therefore, the requirements from Wilkinson (2013) for a valid ABC probability distribution with noise hold for Equation 33 and the population with variance $(\sigma_F)^2$.

Although the assumption of independence among samples is typically not valid for computational models with correlated outputs, the correlation structure of the residuals is often unknown a priori. While computing and incorporating correlation effects is beyond the scope of this paper, one potential strategy would involve including the covariance structure in the calculation of $(\sigma_F)^2$. Additionally, the concept of pooling the embedded predictions used in the GMM likelihood could be extended to non-Gaussian predictions and non-additive or non-Gaussian error formulations. However, the computation of the global population moments is not direct and would require to perform estimations, for example through sampling, or extensive reformulations.

The likelihood function for the mean would directly be obtained by evaluating the error model of Equation 33 as

$$\mathcal{L}_1^{\text{GMM}}(\boldsymbol{\theta}) = \pi_{t_1}(\hat{t}_1(U(\boldsymbol{\theta})) - t_1(F(\boldsymbol{\theta}))) = \pi_{t_1}(\overline{\boldsymbol{u}}) = \left(\frac{2\pi(\sigma_F)^2}{n_y}\right)^{-\frac{1}{2}} \exp\left(-\frac{n_y\overline{\boldsymbol{u}}^2}{2(\sigma_F)^2}\right), \tag{36}$$

which is equivalent to the independent normal likelihood from Section 2.5.3. This likelihood formulation can be interpreted as the probability of the estimated first moment of the samples $\boldsymbol{u}$ assigned by the error model obtained from the sample distribution of Equation 31.

We proceed analogously to build the likelihood function for the second moment. The density function of a $\chi^2_{n_y-1}$ distribution can be expressed as

$$f\left(x, n_y - 1\right) = \frac{x^{\frac{n_y-1}{2}-1}\exp\left(-\frac{1}{2}x\right)}{2^{\frac{n_y-1}{2}}\Gamma\left(\frac{n_y-1}{2}\right)}. \tag{37}$$

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Hence, substituting $x = \frac{n_y(\hat{s}_U)^2}{(\sigma_F)^2}$ in the $\chi^2_{n_y-1}$ density function, it follows

$$\mathcal{L}_2^{\mathrm{GMM}}(\boldsymbol{\theta}) = \pi_{t_2}\left(\frac{\hat{t}_2(U(\boldsymbol{\theta}))}{t_2(F(\boldsymbol{\theta}))}\right) = \pi_{t_2}\left(\frac{(\hat{s}_U)^2}{(\sigma_F)^2}\right) = \frac{1}{2^{\frac{n_y-1}{2}}\Gamma\left(\frac{n_y-1}{2}\right)}\exp\left(-\frac{(n_y-1)(\hat{s}_U)^2}{2(\sigma_F)^2}\right)\left(\frac{n_y(\hat{s}_U)^2}{(\sigma_F)^2}\right)^{\frac{n_y-1}{2}-1} \tag{38}$$

where $(\hat{s}_U)^2$ and $(\sigma_F)^2$ are obtained from Equations 35 and 30.

The global moment-matching likelihood will then be

$$\pi_\varepsilon(\rho(t(U(\boldsymbol{\theta})), t(F(\boldsymbol{\theta})))) = \pi_\varepsilon^{\mathrm{GMM}}(\boldsymbol{\theta}) = \mathcal{L}_1^{\mathrm{GMM}}(\boldsymbol{\theta})\mathcal{L}_2^{\mathrm{GMM}}(\boldsymbol{\theta}) \tag{39}$$

and in logarithmic form,

$$\begin{aligned}\log \pi_\varepsilon^{\mathrm{GMM}}(\boldsymbol{\theta}) = &-\frac{1}{2}\log\left(\frac{2\pi(\sigma_F)^2}{n_y}\right) - \frac{n_y\overline{u}^2}{2(\sigma_F)^2} - \frac{n_y-1}{2}\log 2 - \log\left(\Gamma\left(\frac{n_y-1}{2}\right)\right) \\ &- \frac{(n_y-1)(\hat{s}_U)^2}{2(\sigma_F)^2} + \left(\frac{n_y-1}{2}-1\right)\log\left(\frac{n_y(\hat{s}_U)^2}{(\sigma_F)^2}\right).\end{aligned} \tag{40}$$

2.5.5. Relative global moment-matching (RGMM) Likelihood with noise

Building the GMM likelihood directly from the residuals of the observations may lead to outliers from points where the predicted variance is high. Therefore, in cases with very heteroscedastic noise or model form uncertainties, the previous likelihood will lead to biased estimations of the embedded parameters, which are mainly controlled by such effects. To compensate for this, we propose a *relative* version of the GMM likelihood that normalizes the centered predictions by the predicted variance. In this formulation, $F_r(\boldsymbol{\theta})$ is interpreted as an equally weighted aggregate of standardized predictive distributions:

$$F_r(\boldsymbol{\theta}) = \frac{1}{n_y}\sum_{i=1}^{n_y}\mathcal{N}(0,1), \quad \text{where each component is} \frac{f^f(\boldsymbol{\theta},\mathbf{x}_i) - \mu^f(\boldsymbol{\theta},\mathbf{x}_i)}{\sqrt{(\sigma^f(\boldsymbol{\theta},\mathbf{x}_i))^2 + (\sigma_N)^2}}. \tag{41}$$

Then, the moments of $F_r(\boldsymbol{\theta})$ are directly $\mu_{F_r} = 0$ and $(\sigma_{F_r})^2 = 1$. As before, the sampled moments are approximated using the Central Limit Theorem and Cochran's theorem. This is easier to achieve here due to the scaling by the predictive variances, which reduces heteroscedasticity and jointly normal residuals are expected.

$$\mu_{U_r} - \mu_{F_r} \sim \mathcal{N}\left(0, \frac{(\sigma_{F_r})^2}{n_y}\right) \tag{42}$$

$$\frac{n_y(\sigma_{U_r})^2}{(\sigma_{F_r})^2} \sim \chi^2_{n_y}, \tag{43}$$

where μ_{U_r} and $\sigma^2_{U_r}$ are the moments normalized by the predictive variances. Defining the relative residuals as

$$u_{ri} = \frac{y_i - \mu^f(\boldsymbol{\theta},\mathbf{x}_i)}{\sqrt{(\sigma^f(\boldsymbol{\theta},\mathbf{x}_i))^2 + (\sigma_N)^2}}, \quad \text{with } i = 1,2,\ldots,n_y, \tag{44}$$

these sample moments are calculated as

$$\hat{t}_1(U_r(\boldsymbol{\theta})) = \hat{\mu}_{U_r} = \overline{\boldsymbol{u}}_r \tag{45}$$

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

$$\hat{t}_2(U_r(\boldsymbol{\theta})) = (\hat{s}_{U_r})^2 = \frac{\sum_{i=1}^{n_y} (u_{ri} - \overline{\mathbf{u}}_r)^2}{n_y}. \tag{46}$$

Following the same reasoning as for the GMM likelihood, we obtain the components of the corresponding RGMM likelihood function $\pi_{\varepsilon}^{\mathrm{RGMM}}(\boldsymbol{\theta}) = \mathcal{L}_1^{\mathrm{RGMM}}(\boldsymbol{\theta})\mathcal{L}_2^{\mathrm{RGMM}}(\boldsymbol{\theta})$, where

$$\mathcal{L}_1^{\mathrm{RGMM}}(\boldsymbol{\theta}) = \left(\frac{2\pi(\sigma_{F_r})^2}{n_y}\right)^{-\frac{1}{2}} \exp\left(-\frac{n_y \overline{\mathbf{u}}_r^2}{2(\sigma_{F_r})^2}\right), \tag{47}$$

and

$$\mathcal{L}_2^{\mathrm{RGMM}}(\boldsymbol{\theta}) = \frac{1}{2^{\frac{n_y}{2}}\Gamma\left(\frac{n_y}{2}\right)} \exp\left(-\frac{n_y(\hat{s}_{U_r})^2}{2(\sigma_{F_r})^2}\right)\left(\frac{n_y(\hat{s}_{U_r})^2}{(\sigma_{F_r})^2}\right)^{\frac{n_y}{2}-1}. \tag{48}$$

The likelihood in logarithmic form is

$$\begin{aligned}\log \pi_{\varepsilon}^{\mathrm{RGMM}}(\boldsymbol{\theta}) = &-\frac{1}{2}\log\left(\frac{2\pi(\sigma_{F_r})^2}{n_y}\right) - \frac{n_y \overline{u}_r^2}{2(\sigma_{F_r})^2} - \frac{n_y}{2}\log 2 - \log\left(\Gamma\left(\frac{n_y}{2}\right)\right) \\ &- \frac{n_y(\hat{s}_{U_r})^2}{2(\sigma_{F_r})^2} + \left(\frac{n_y}{2} - 1\right)\log\left(\frac{n_y(\hat{s}_{U_r})^2}{(\sigma_{F_r})^2}\right).\end{aligned} \tag{49}$$

The same limitations and potential extensions from the GMM likelihood regarding the assumptions of additive, heteroscedastic, and Gaussian noise models, as well the assumed normality of the predictions, also apply for RGMM.

2.6. Summary on likelihood formulations

In this article, four likelihood functions are compared: ABC-likelihood with noise (ABC), global moment-matching (GMM) likelihood, relative global moment-matching (RGMM), and independent normal (IN) likelihood with noise. GMM and RGMM are new proposals developed in this work, while ABC and IN have been adapted from Sargsyan et al. (2019) to explicitly include noise. As a common feature, the four of them aim to match the mean and variance of the predicted output distribution with the data. A summary of the likelihood models is presented in Table 1.

2.7. Pushed-forward prediction of quantities of interest

The solution of the inverse problem provides the joint posterior distribution $\pi(\boldsymbol{\theta}^m, \boldsymbol{\theta}^b|\mathbf{y})$. The predicted response $\boldsymbol{f}_P(\boldsymbol{\theta}^m, \boldsymbol{\theta}^b, \mathbf{x}|\mathbf{y})$ is to be computed. To do so, the posterior distribution of the unknown parameters should be propagated through the forward model. The stochastic extension transforms $\boldsymbol{\theta}$ from a deterministic parameter into a random variable $\tilde{\boldsymbol{\theta}}$ governed by $\boldsymbol{\theta}^m$ and $\boldsymbol{\theta}^b$. Consequently, the output $\boldsymbol{f}(\tilde{\boldsymbol{\theta}})$ is also a random variable, which adds complexity to the propagation process. Each evaluation of samples of $\boldsymbol{\theta}^m$ and $\boldsymbol{\theta}^b$ would require full propagation of the random variable $\tilde{\boldsymbol{\theta}}$. A common approach (Huan et al., 2017; Sargsyan et al., 2019) is to use an estimator $\hat{\boldsymbol{\theta}}$ of the unknown parameters, typically a maximum-a-posteriori (MAP) one. The predicted response is then

$$\boldsymbol{f}_P\left(\hat{\boldsymbol{\theta}}^m, \hat{\boldsymbol{\theta}}^b, \mathbf{x}|\mathbf{y}\right) = \boldsymbol{f}\left(\hat{\boldsymbol{\theta}}^m, \hat{\boldsymbol{\theta}}^b, \mathbf{x}\right). \tag{50}$$

Due to the embedding, $\boldsymbol{f}_P$ generates a stochastic response that usually does not have a tractable form. Therefore, either the polynomial chaos expansion specified in Section 2.3 or Monte-Carlo methods can be used to obtain the moments of the approximated predicted response $\boldsymbol{\mu}_P^f$ and $\boldsymbol{\sigma}_P^f$. One of the main characteristics of $\boldsymbol{f}_P$ as shown by Sargsyan et al. (2019) is that the variance of the predicted response

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Table 1. *Moment-matching likelihoods summary*

Likelihood model	Formulation	Mean matching	Variance matching
Approximate Bayesian computation (ABC)	Equation 24	$-\frac{\left(\mu_i^f - y_i^f\right)^2}{2(\sigma_N)^2}$	$-\frac{\left(\sqrt{\left(\sigma_i^f\right)^2+(\sigma_N)^2}-\gamma\left\|\mu_i^f-y_i^f\right\|\right)^2}{2\epsilon^2}$
Independent normal	Equation 26	$e^{-\frac{\left(y_i-\mu_i^f\right)^2}{2\left(\left(\sigma_i^f\right)^2+(\sigma_N)^2\right)}}$	–
Global moment-matching (GMM)	Equations 36 and 38	$\mu_U-\mu_F\sim\mathcal{N}\left(0,\frac{(\sigma_F)^2}{n_y}\right)$	$\frac{n_y(\hat{s}_U)^2}{\sigma_F^2}\sim\chi^2_{n_y-1}$
Relative global moment-matching (RGMM)	Equations 47 and 48	$\mu_{U_r}-\mu_{F_r}\sim\mathcal{N}\left(0,\frac{(\sigma_{F_r})^2}{n_y}\right)$	$\frac{n_y(\hat{s}_{U_r})^2}{(\sigma_{F_r})^2}\sim\chi^2_{n_y-1}$

does not reduce to zero for large sample sets, preserving the variance due to the dataset and the model inadequacy.

A special case is the propagation of the uncertainty obtained for the inferred parameters through the embedding and solution of the inverse problem to other quantities of interest (QoI) that did not take part in the inference procedure. The function $\boldsymbol{g}$ that represents the QoI based on the inferred parameters is defined analogously to $\boldsymbol{f}$ following Equation 4. Then, the posterior distribution must be pushed through $\boldsymbol{g}$ in the same way as for $\boldsymbol{f}_P$. The predicted QoI will include the uncertainty expressed through the inferred embedded unknown variables.

Nevertheless, using an estimator for $\boldsymbol{\theta}$ presents several drawbacks. First, $\hat{\boldsymbol{\theta}}$ are pointwise estimators despite generating a predictive distribution $\hat{\boldsymbol{f}}_P$. The information they provide is limited, losing the structure and potential correlations that are present in the joint posterior distribution of the unknown parameters. Pointwise estimators fail as well in multimodal distributions to represent other local maxima apart from the optimum, which may be relevant for the QoI. Additionally, the distribution obtained from $\boldsymbol{f}_P\left(\hat{\boldsymbol{\theta}}^m,\hat{\boldsymbol{\theta}}^b,\boldsymbol{x}|\boldsymbol{y}\right)$ only preserves the uncertainty provided by the model form uncertainty parameters $\boldsymbol{\theta}^b$, disregarding the variance obtained through the inference procedure, which leads to potentially overconfident predictions.

As an alternative, in this article, we will opt for a full propagation of the posterior distributions of the unknown parameters as

$$\boldsymbol{f}_P\left(\boldsymbol{\theta}^m,\boldsymbol{\theta}^b,\boldsymbol{x}|\boldsymbol{y}\right)=\boldsymbol{f}\left(\boldsymbol{\theta}^m,\boldsymbol{\theta}^b,\boldsymbol{x}\right), \tag{51}$$

or $\boldsymbol{g}\left(\boldsymbol{\theta}^m,\boldsymbol{\theta}^b,\boldsymbol{x}|\boldsymbol{y}\right)$ for QoIs other than the predictions. The full propagation of the posterior was proposed by Sargsyan et al. (2019), however, the influence of choosing different likelihood formulations has yet to be analyzed. By propagating the joint posterior distribution, the full information obtained during the inference procedure is preserved. This allows not only obtaining a predictive posterior distribution $\boldsymbol{f}_P$ or $\boldsymbol{g}$, but also perform inference on its parameters, providing additional insight into its reliability. For example, it is possible to obtain the pushed posterior distribution of $\boldsymbol{\mu}_P^f$, where propagating the estimators would have only provided a point value. Moving from the propagation of point estimates to that of the full posterior enables a more complete representation of the uncertainty, revealing how variability in the unknown parameters translates into variability in the quantities of interest (QoIs).

In this work, our focus is on showing that this pushed-posterior distribution depends on the likelihood function adopted during inference, and that different likelihood choices lead to distinct propagated uncertainty structures. Two main challenges arise with this approach: the increasing number of evaluations of $\boldsymbol{f}$ or $\boldsymbol{g}$ and the composition of stochastic variables. The additional evaluations come from the need to sample the joint posterior distribution of the unknown parameters, which implies evaluating $\boldsymbol{f}$ or $\boldsymbol{g}$ for each of the samples. To mitigate this computational cost, we employ PCE surrogate models, following

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

the approach introduced by Sargsyan et al. (2019) for posterior propagation. These samples of $\boldsymbol{\theta}$ define random variables as inputs, therefore $\boldsymbol{f}$ and $\boldsymbol{g}$ necessarily generate a stochastic response, which can be approximated by the PCE. If the QoI is the value of a given realization of $\boldsymbol{f}_P$ or $\boldsymbol{g}$, the posterior distribution must account for the variability in the sampled $\boldsymbol{\theta}$ and the variance from the predicted distribution itself. In this paper, this situation will be solved by sampling from the pushed-posterior distributions $\boldsymbol{f}_P(\boldsymbol{\theta}_i^m, \boldsymbol{\theta}_i^b, \boldsymbol{x}|\boldsymbol{y})$ or $\boldsymbol{g}(\boldsymbol{\theta}_i^m, \boldsymbol{\theta}_i^b, \boldsymbol{x}|\boldsymbol{y})$ for every $(\boldsymbol{\theta}_i^m, \boldsymbol{\theta}_i^b)$ in n_P samples of $\boldsymbol{\theta}$ and analyzing the resulting dataset.

3. Applications and discussion

3.1. Application case: simple example

This example aims to test the behavior of the different likelihood formulations under known conditions. The same model is used to generate the dataset and to compute the predictions:

$$y = \theta x + \varepsilon_N \tag{52}$$

where $x \in [0,1]$ is the input variable, y is the output variable, θ is the slope parameter to be inferred and ε_N is a white noise perturbation. The generator aims to provide a sample of observations generated by the computational model given $\theta \sim \mathcal{N}(4.0, 1.0)$. The observations $\boldsymbol{y}$ are then generated by evaluating Equation 52 with a random value of θ from the aforementioned distribution for each entry of the input vector $\boldsymbol{x}$ and adding a random perturbation of $\varepsilon_N \sim \mathcal{N}\left(0, (\sigma_{\text{true}})^2\right)$ with $\sigma_{\text{true}} = 0.01$. If not specified for a particular analysis, $\boldsymbol{x}$ is composed of 120 equidistant samples covering the range $[0.4, 1.0]$. An example of a generated dataset is presented in Figure 2.

The extended variable θ^* with the embedding is defined as

$$\theta^* = t + \delta_b, \quad \delta_b \sim \mathcal{N}\left(0, (\sigma_b)^2\right) \tag{53}$$

where t would correspond to the original unknown variable and σ_b is the new additional unknown variable associated with the model inadequacy. This is equivalent to imposing $\theta^* \sim \mathcal{N}\left(t, (\sigma_b)^2\right)$ and using θ^* instead of θ in Equation 52. As the forward model is a linear function and θ^* follows a normal distribution, a closed-form expression of $\boldsymbol{f}$ is available in this case and can be formulated as

$$\boldsymbol{f}(t, \sigma_b, \boldsymbol{x}) = \theta^* \boldsymbol{x} + \boldsymbol{\varepsilon}_N \Rightarrow f(t, \sigma_b, x_i) \sim \mathcal{N}\left(t x_i, \left((\sigma_b)^2 (x_i)^2 + (\sigma_N)^2\right)\right) \quad \forall i = 1, 2, \ldots, n_y. \tag{54}$$

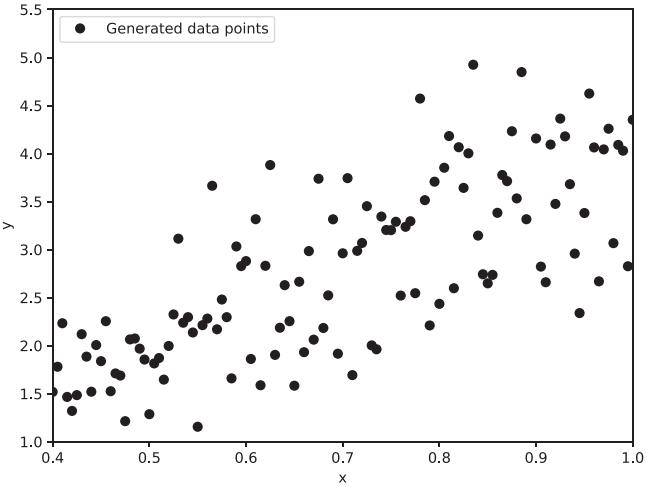


Figure 2. *Generated dataset of the simple example for 120 data points.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

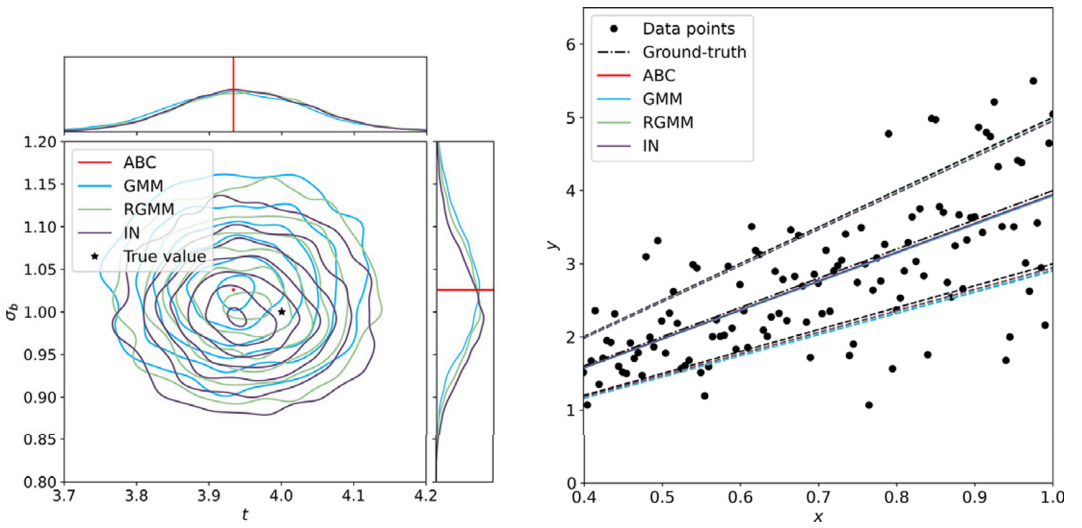
The existence of a closed-form solution provides analytical expressions for the statistical moments of $f(t,\sigma_b,x_i)$ for every x_i in $\mathbf{x}$ and sampled t and σ_b. Under these conditions, $f(t,\sigma_b,x_i)$ follows a Gaussian distribution, ensuring that no additional approximation error arises from the normality assumption employed in the likelihood formulations. This allows a comprehensive analysis of the likelihood formulations and the effect of noise. Nevertheless, a PCE approximation of the forward model response is built in the interest of testing the whole methodology developed in Section 2. As the response is a linear function and the input is normal, a first-degree PCE with Hermite polynomials and two Gauss points is exact and the approximation error tends to zero. The statistical moments of the computed response $\boldsymbol{\mu}^f$ and $\boldsymbol{\sigma}^f$, which are necessary to evaluate the likelihood functions in the inference procedure, are then obtained from Equations 15 and 16.

To obtain the posterior probability $\pi(t,\sigma_b|\mathbf{y})$, an MCMC algorithm with the tested likelihood models is applied to the generated dataset $\mathbf{y}$. The algorithm is run until convergence with a threshold of the estimated sample size $\widehat{ESS}=837$ or a maximum of 10000 MCMC samples. The $\widehat{ESS}$ is calculated following Appendix A. The arbitrarily chosen maximum number of samples was sufficiently large such that it was never reached.

The prior distributions for the parameters are $\pi(t)\sim\mathcal{N}(4.5,0.5^2)$ and $\pi(\sigma_b)\sim\mathcal{LN}(-1,0.5)$. The prior for t is weakly informative and easily covers the true value $t=4.0$, which was chosen primarily for numerical convenience. The prior for σ_b is chosen as a log-normal distribution to ensure positivity. This leads to $\sigma_b\in\left[e^{-1.5},e^{-0.5}\right]$ in the range of one standard deviation. Due to the simplicity of the forward model, the likelihood is sharply peaked around a unique optimum and the effect of the priors on the posterior is therefore negligible and equivalent for all the formulations. Other prior choices would be possible without altering significantly the findings.

The pair plots for the posterior $\pi(t,\sigma_b|\boldsymbol{y})$ from applying the MCMC algorithm with each likelihood formulation and the posterior predictive for its mean are shown in Figure 3. Each likelihood leads to converged values close to the true generating parameters $t=4.0$, $\sigma_b=1.0$. The results are presented in Table 2.

The different likelihood models provide posterior distributions centered around the same values of the unknown parameters. The distribution provided by the ABC likelihood presents significantly less



(a) Pair plot posterior distibution (b) Posterior predictive

Figure 3. *Comparison for the converged solution of linear example. (a) Pair plot of the posterior distribution for t and σ_b. (b) Predictive distribution from the mean of the posterior distributions of t and σ_b. Dashed lines indicate the interval of $\mu^f\pm\sigma^f$.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Table 2. *Posterior results for the linear model and different likelihoods*

	t			σ_b		
Likelihood	Mean	Std	$\widehat{ESS}$	Mean	Std	$\widehat{ESS}$
ABC	3.93	0.00	866	1.03	0.00	845
GMM	3.95	0.13	901	0.89	0.15	868
RGMM	3.97	0.13	896	1.03	0.14	855
IN	3.95	0.09	1117	1.00	0.06	866

variance than the others. This is due to the fitting of σ_b being exact up to ϵ as a difference, while the other likelihoods have larger acceptance probabilities for the values of σ_b. Alternative values for ϵ would have been less restrictive in the fitting of the moments, generating a flatter likelihood distribution that would have led to a wider posterior. There is no significant difference in the convergence speed for this particular dataset. The difference between the means of the parameters and the true generating ones is rooted in the specific realization of the dataset. However, it must be noted that ABC is the only likelihood that does not cover the true generating parameters for any reasonable confidence interval, as it aims to fit exactly the moments of a particular sample dataset.

Four QoI are evaluated for $x = 1.0$: the output y, the mean of the predictive distribution μ_P^f, the standard deviation σ_P^f and the z-value of $y(1.0)$. The QoI generally can have physical meaning or simply be indicators of performance. In this case, due to the simplicity of the application, they represent quantities of statistical importance that can be used for an analysis of the fitness of the predictions. In a classical inference problem, these quantities would be deterministic, which would not allow for an analysis of the reliability of the QoIs themselves. The output y is directly the pushed-forward prediction $\boldsymbol{f}_P = \boldsymbol{f}(\boldsymbol{t}, \boldsymbol{\sigma_b}, \boldsymbol{x})$, as described in Section 2.7. The other QoIs are calculated for each sample of the posterior distribution of the parameters. In particular, the z-value $|Z| = \left|\frac{\mu_P^f - y(1.0)}{\sigma_P^f}\right|$ can be used for testing the hypothesis H_1 : the value $y(1.0)$ has not been generated by a normal distribution $\mathcal{N}\left(\mu_P^f, \sigma_P^f\right)$ against the null hypothesis H_0. The critical threshold of $|Z|$ for a confidence level of 95% is 1.96, allowing to reject H_0 if the value of $|Z|$ is significantly larger. For comparison, the equivalent QoIs obtained with the MAP estimator of t and σ_b are calculated as well. The results are plotted in Figure 4.

The obtained propagated distribution is comparable for all likelihoods, either using an estimator of the posterior or sampling from the distribution. The distributions for μ_P^f and σ_P^f reflect the same conclusions from the pair plot representation. This is due to the linear nature of the system, which propagates directly the posterior distribution of the parameters, as t controls μ_P^f and σ_b, σ_P^f. The z-values are below 1.96 for all likelihoods, failing to reject H_0, meaning it is plausible that the data point $y(1.0) = 5.04$ was generated by the corresponding posterior distribution. In comparison, if no inadequacy term had been introduced and the posterior only considered the prescribed noise as a source of variance, the corresponding z-value would have been 104, which allows us to reject H_0, inferring that the data point could not have been generated by the prediction. It is noted that if only the estimators where propagated, it would not be possible to generate distributions for three of the QoI, and the values indicated in Figure 4 as dashed lines would be taken as the predicted ones.

3.1.1. Dataset dependency analysis

While the obtained results provide insight into the properties of the different likelihoods for this case, it is necessary to compare them with several datasets generated analogously. The described procedure was repeated for 20 datasets with different random seeds for the data generation. The summary of results is presented in Figure 5.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

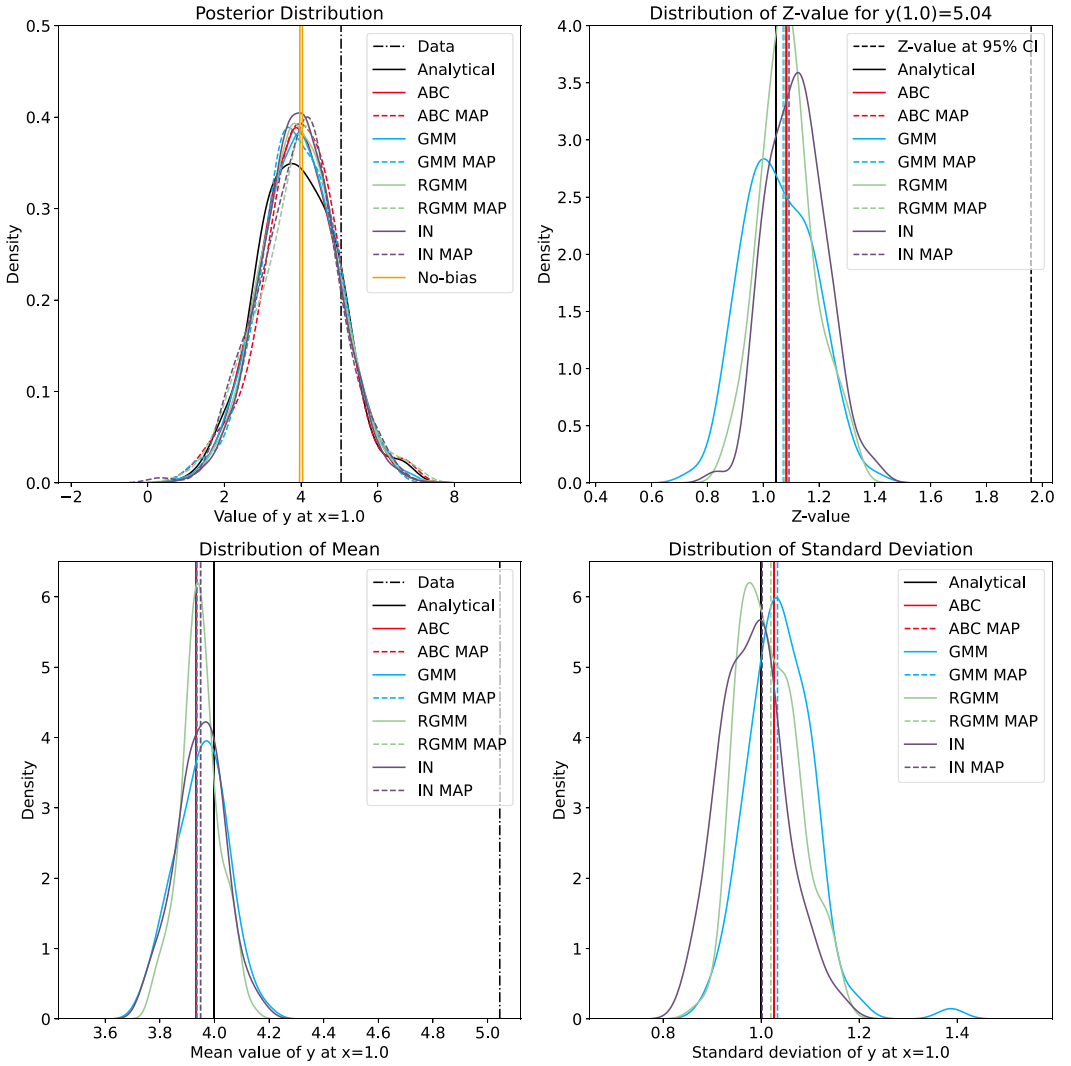


Figure 4. *Analysis of quantities of interest by propagating the posterior distribution of t and σ_b for the simple example.*

As can be observed, the mean value of t using ABC varies the most from dataset to dataset, with the others being mostly equivalent. Coincidentally, ABC presents the smallest variance in the posterior of t, followed by IN, GMM, and RGMM. Due to the limited dataset, ABC tends to overfit to the available points, whose statistics may present a discrepancy from the generating model ones, as ABC aims to fit the moments exactly. When comparing the model form uncertainty results related to σ_b, the means for RGMM and IN are comparable. The only difference comes from GMM, for which the limited dataset does not hold the assumptions of the predictions sharing the same distribution, which is required for this likelihood. An analysis of the speed of convergence is presented in Figure 6. The ABC likelihood presents a larger variance in the number of MCMC samples to converge, with a significant amount of outliers for the $\widehat{ESS}$ of σ_b. Due to the exact matching of the statistical moments, the ABC likelihood may require longer chains in the MCMC algorithm. It must be noted that this can be controlled by the value of the tolerance parameter ϵ. The other likelihood models present comparable convergence behavior with each other.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

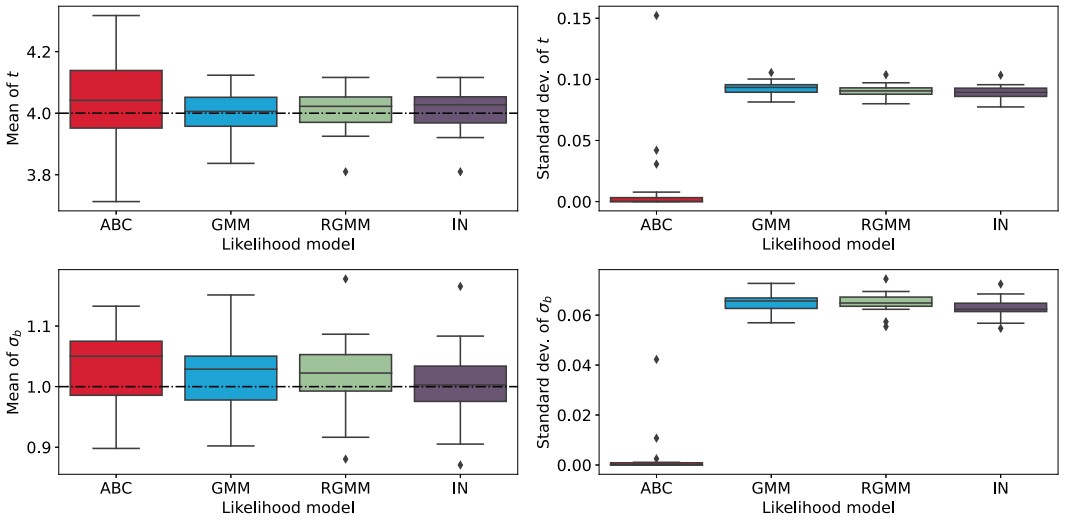


Figure 5. *Box plots for the unknown parameters of the embedded model with different likelihoods. Outliers are indicated as rhomboids. Mean and standard deviation of the posterior distribution after 1000 MCMC evaluations with 10 walkers and 200 burn-in steps for 20 different realizations of the dataset. Mean (top left) and standard deviation (top right) for the slope variable t and mean (bottom left) and standard deviation (bottom right) of the model inadequacy scale parameter σ_b. True generating values are indicated as a dot-dashed line in the means.*

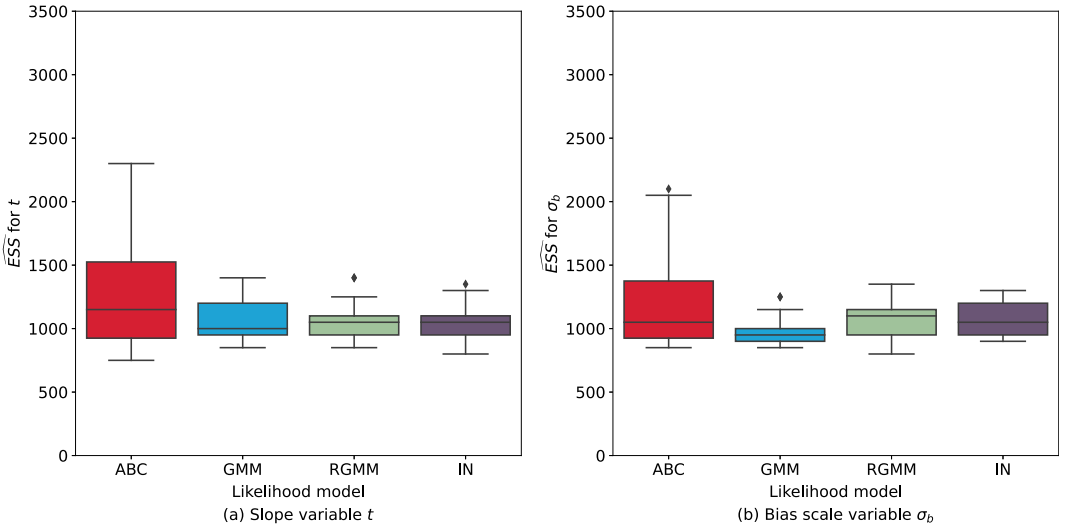


Figure 6. *Boxplot of required number of MCMC samples until reaching the threshold $\widehat{ESS} = 837$ in both variables for the simple example.*

3.1.2. Noise influence analysis

It has been proven in Section 2 and Appendix B, that the ABC likelihood is very sensitive to the prescribed additive noise σ_N, and in particular to its misspecification. We repeat the basic experiment, but with 20 different values for the prescribed σ_N between 0.001 and 10.0 and the same prior distribution. The posterior predictive results are presented in Figure 7 and the compared influence of the noise in Figure 8.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

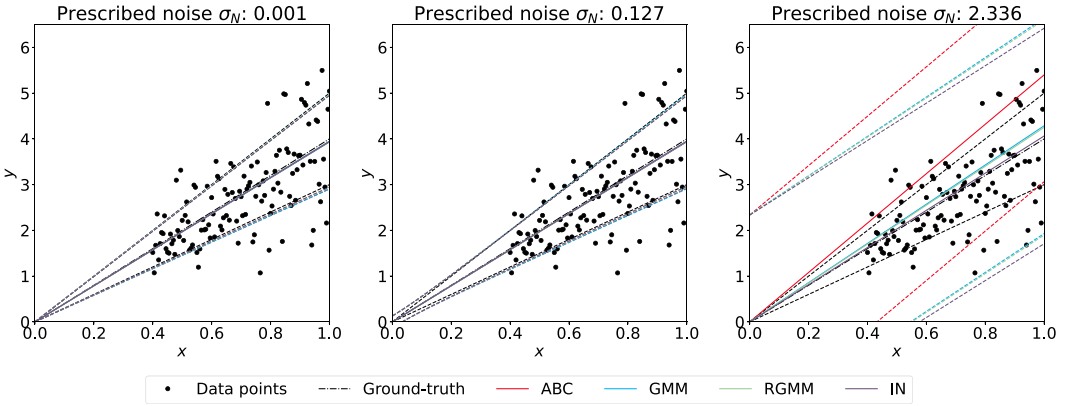


Figure 7. *Posterior prediction comparison for noise value σ_N of 0.001, e^{-2} and $e^{0.85}$, chosen using a logarithmic rule. Dashed lines indicate the interval of $\mu^f \pm \sigma^f$.*

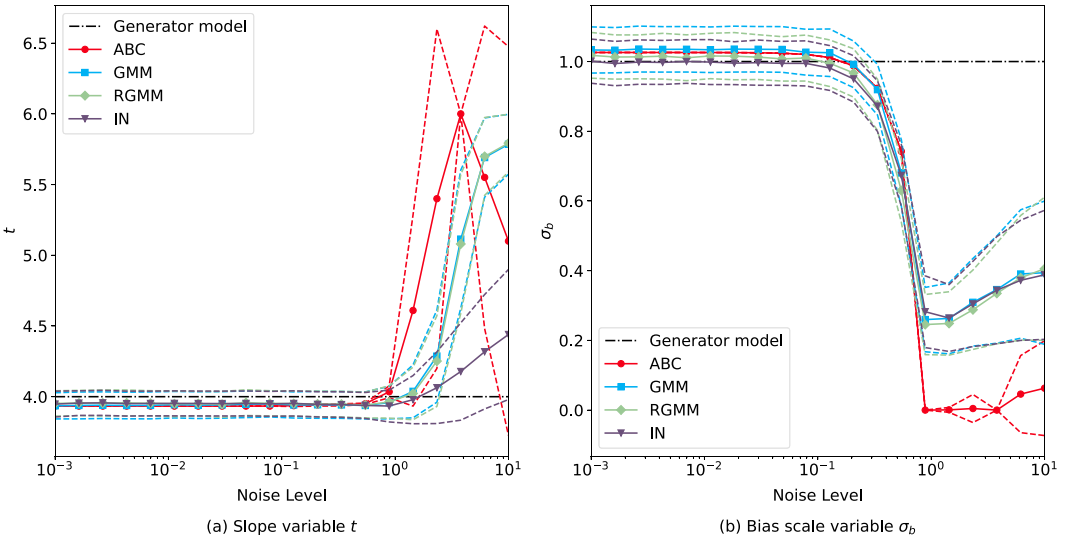


Figure 8. *Influence of prescribed noise value σ_N for the posterior of (a) slope variable t and (b) inadequacy scale variable σ_b. Dashed lines indicate the interval of $\pm\sigma$.*

As expected, the ABC likelihood is the most sensitive to noise misspecification, particularly for values of σ_N exceeding 1.0. When $\sigma_N < 0.01$, the prescribed noise underestimates the true generative noise, and the resulting variability is absorbed by the model form uncertainty. However, due to the scale difference, this effect is not visible in the results. For $0.01 < \sigma_N$, the noise is overestimated for at least some data points, reducing the contribution of σ_i^f to the predictive variance at those locations. Notably, for $\sigma_N > 0.4$, the prescribed noise surpasses the variance of the data generator at certain points, leading to significant shifts in the inferred parameters. Once $\sigma_N > 1.0$, all data samples exhibit overestimated variance. This overestimation strongly affects the unknown parameter t in ABC, which attempts to match both the mean and variance precisely. Although other likelihoods also degrade in performance as noise increases, their sensitivity is less pronounced than that of ABC. Additionally, for large σ_N, ABC exhibits irregular behavior due to the emergence of local minima in the likelihood, causing some chains to converge prematurely and introducing numerical instability. In contrast, GMM, RGMM, and especially IN are

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

considerably more robust to noise misspecification. While GMM and RGMM yield estimates of t that deviate more from the true value ($t = 4.0$) than IN, their confidence intervals remain narrow. For IN, the interval widens with increasing noise. It is also worth noting that due to heteroscedasticity in the discrepancy between predictions and observations, stemming from randomness in θ, σ_b remains nonzero to account for prediction errors not captured by the homoscedastic noise σ_N. These effects are particularly pronounced for unrealistically large values of prescribed noise.

3.1.3. Offset influence analysis

Another influential factor in the choice of the likelihood formulation is the existence of data points that cannot be replicated by the response function. This can be simulated by including an additive offset Δy to the generator as

$$y = \theta x + \Delta y + \varepsilon_N \tag{55}$$

while keeping the original computational model. This case is relevant, for example, for non-stationary problems that depend on initial conditions, which may introduce such an offset. An offset in the range between 0.0 and 1.0 is tested. The posterior predictive results are presented in Figure 9 and the influence comparison in Figure 10. Increasing offset values lead to higher values for t to try to fit the observed points. At the same time, this leads to the need for larger variances to cover the dataset. This effect is more pronounced in RGMM and IN as the offset impacts in particular values with smaller σ_i^f that need larger changes in t and σ_b to cover the new dataset, and coincidentally, are those with the largest weight in those likelihood models. Comparing the values for the mean of t with offsets of 0.0 and 1.0, it can be observed that IN and RGMM evolve less favorably than GMM with the addition of such an offset, as was expected, since the marginalizing effect is stronger in such likelihood models. For smaller datasets where the individual points have a larger weight, we expect this tendency to be more acute, leading to improvements of RGMM over IN as well, driven by the addition of the second moment fitting.

3.1.4. Outlier influence analysis

An alternative situation in which an offset is present is if it only affects a given subset of data points, creating a group of outliers that influence the inference. In this case, Equation 55 is modified to only affect a subset of samples such that

$$y = \begin{cases} \theta x + \varepsilon_N & \text{if } x \in [0.4,\, 0.6) \cup (0.7,\, 1.0] \\ \theta x - \Delta y + \varepsilon_N & \text{if } x \in [0.6,\, 0.7] \end{cases}. \tag{56}$$

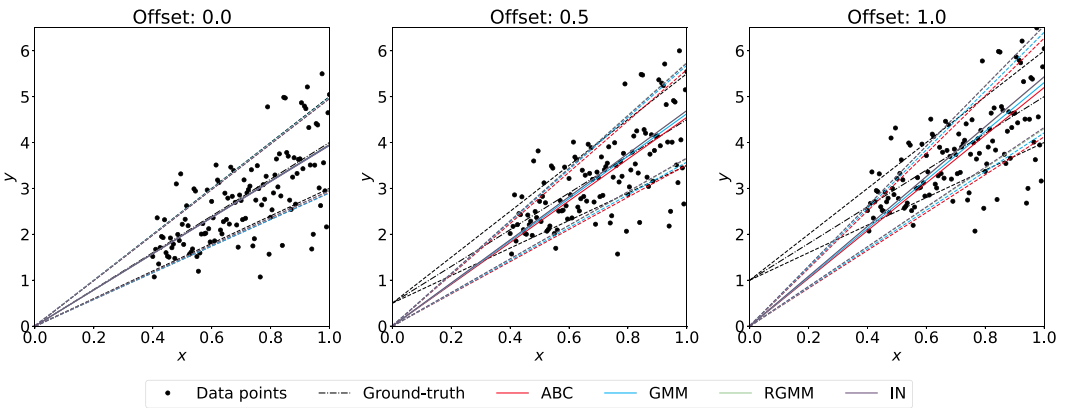


Figure 9. *Posterior prediction comparison for offset values of 0.0, 0.5, and 1.0.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

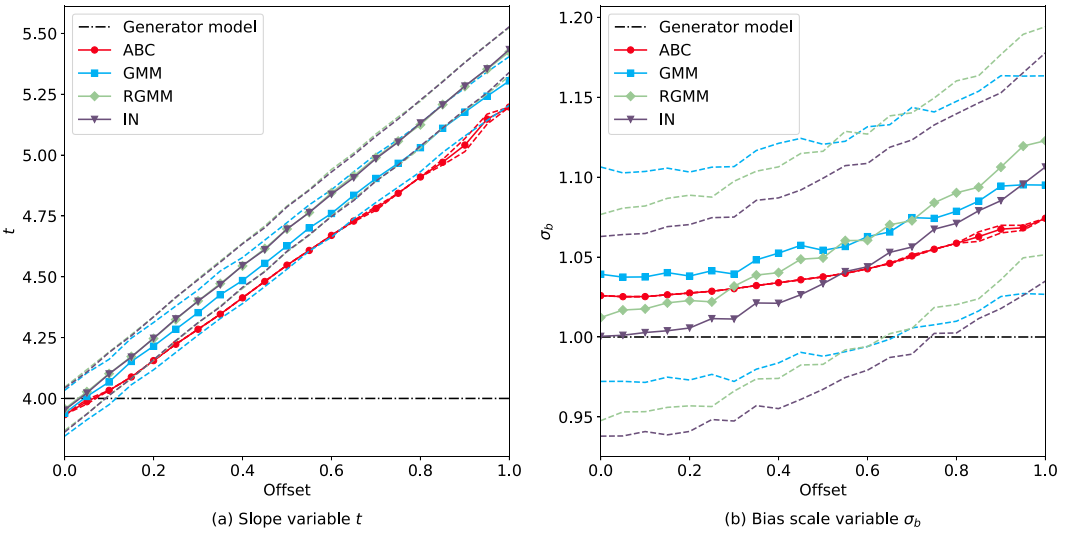


Figure 10. *Influence of prescribed offset Δy for the posterior of (a) slope variable t and (b) model inadequacy scale variable σ_b. Dashed lines indicate the interval of $\pm\sigma$.*

The offset ranges from 0.0 to 2.0 in 21 samples. The posterior predictive results are presented in Figure 11 and the comparison in Figure 12.

In this case, ABC is the most affected in comparison with the other likelihood models. The outliers can be represented by the predicted response, but the fitting of every point with equal weight provokes larger deviations in ABC. Nevertheless, if the outliers are located at data points with low predicted variance, it is expected that RGMM and IN show a larger influence.

3.2. Application case: transient thermal simulation of reinforced concrete

The second application case consists of a transient thermal 2D simulation of the section of a reinforced concrete cylinder with constant external temperature. Due to symmetry and under the assumption of an infinitely long cylinder, such a system can be modeled as a 2D square section with adiabatic boundary conditions at the top, bottom, and left-hand sides and constant temperature at the right-hand side.

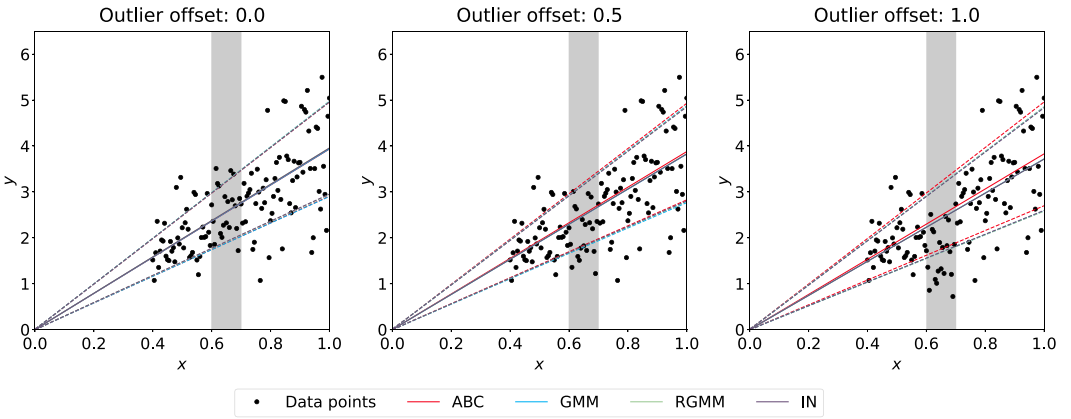


Figure 11. *Posterior prediction comparison for outlier's offset values of 0.0, −0.5, and −1.0. The zones with outliers are shaded in grey. Dashed lines indicate the interval of $\mu^f \pm \sigma^f$.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

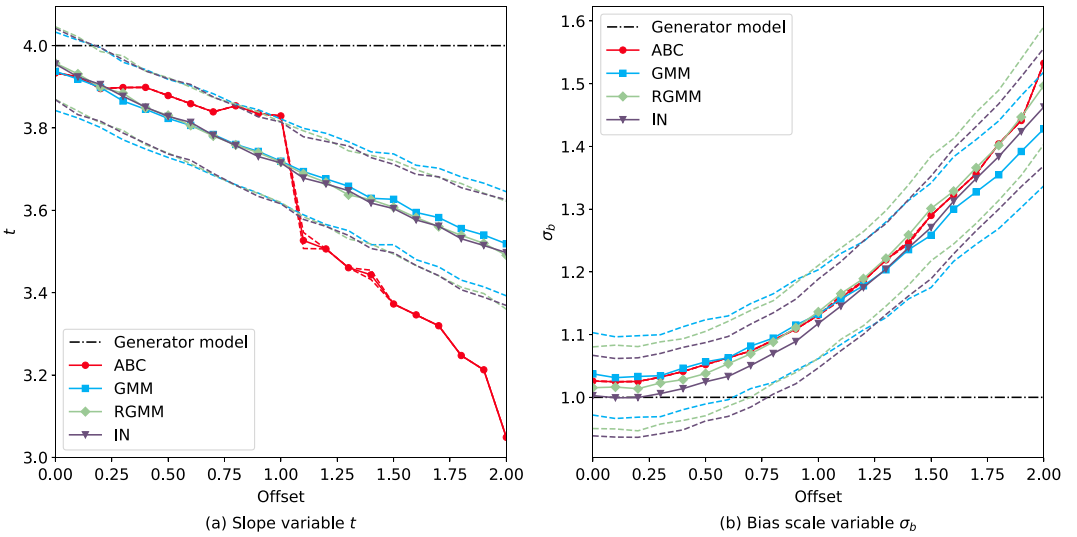


Figure 12. *Influence of outlier data magnitude on the posterior of (a) slope variable t and (b) model inadequacy scale variable σ_b. Dashed lines indicate the interval of $\pm\sigma$.*

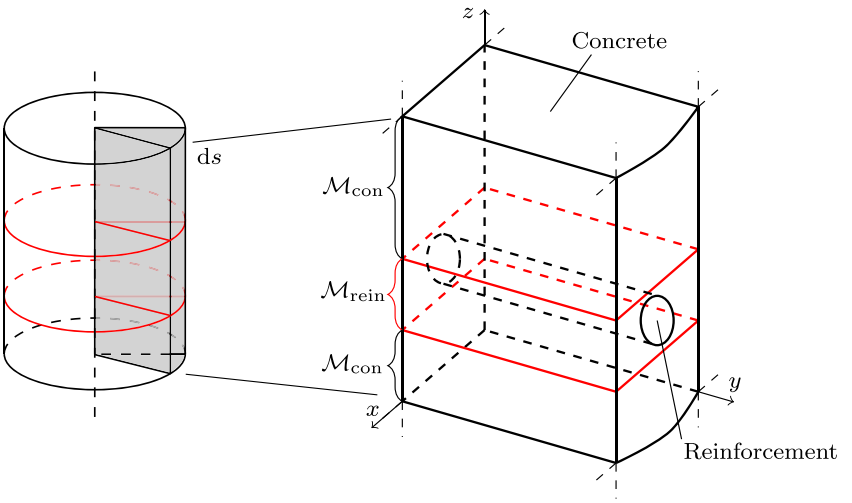


Figure 13. *Schematic diagram of the heat example case for a differential slice d s.*

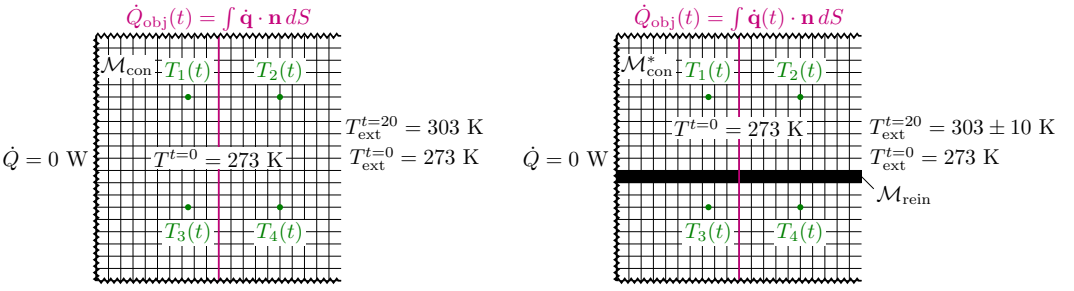
Figure 13 shows a schematic representation of the case. The adiabatic conditions represent the periodic nature of the reinforcement for the sides and an isolation layer for the back. The system is reduced from a 3D cube to a 2D section, assuming constant properties in the X direction. The diagrams in Figure 14 represent the generative and forward models for this application case. The objective is to obtain the total cumulative heat $Q_{\text{obj}}(t)$ that passes through the middle line of the section at a given time.

The transient thermal system follows the heat equation for temperature T

$$\rho c_p \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) = \dot{q}_V \tag{57}$$

where ρ is the density of the material, c_p its heat capacity, k its conductivity, and $\dot{q}_V$ is the volumetric heat flux in the system. There exists a linear relationship between k and c_p for a given ρ which is the diffusivity

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press



(a) Isotropic system (forward model) (b) Reinforced system (generative model)

Figure 14. *Diagrams of the systems modeled in the thermal example. (a) System with isotropic material properties $\mathcal{M}_{\text{con}}$ and (b) system with a band with modified material properties $\mathcal{M}_{\text{rein}}$ that represent a reinforcement bar in that region. The temperature sensors $T_i(t)$ for $i = 1, 2, 3, 4$ act as real sensors and are used for updating the material parameters of $\mathcal{M}_{\text{con}}$. The virtual sensor $\dot{Q}_{\text{obj}}(t)$ predicts the heat through the midline of the system obtained by integrating its normal heat flow obtained from the gradient of the temperature field. The quantity of interest $Q_{\text{obj}}(t)$ is the cumulative heat through the middle line at a given t.*

$\alpha = k/(\rho c_p)$, which requires correcting the heat component as $\dot{q}_V/(\rho c_p)$. The heat flux vector field at a given time t is defined as

$$\dot{\mathbf{q}}(t) = -k\nabla T, \tag{58}$$

where k is calculated for a certain ρ, c_p, and α, and the heat flux through the midline is obtained by integrating the normal heat flux field over the facet S as

$$\dot{Q}_{\text{obj}}(t) = \int \dot{\mathbf{q}}(t) \cdot \mathbf{n}\, dS. \tag{59}$$

The cumulative heat at time t_n is obtained by integrating $\dot{Q}_{\text{obj}}(t)$ over time, which in this case will be done using a trapezoidal rule as

$$Q_{\text{obj}}(t_n) = \int_{t_0}^{t_n} \dot{Q}_{\text{obj}}(t)\, dt \approx \sum_{i=0}^{n-1} \frac{\Delta t_i}{2} \left[\dot{Q}_{\text{obj}}(t_i) + \dot{Q}_{\text{obj}}(t_{i+1}) \right], \tag{60}$$

where t_i are the discretized time points and Δt_i is the interval between two consecutive times t_i and t_{i+1}.

Altogether, the system's geometry, material properties ρ, c_p, and k and its thermal and initial boundary conditions are required to calculate $\dot{Q}_{\text{obj}}(t)$. In this case, the squared slab has a length of 0.4 m in each direction, with adiabatic boundary conditions on three sides ($\dot{Q} = 0$ W) and fixed external temperature $T_{\text{ext}} = 303$ K. The initial temperature at all points of the system is $T(t=0) = 273$ K and we assume a known density for the concrete of 2300 kg/m^3. The diffusivity value α determines the thermal behavior and must be estimated from the observations.

For the parameter estimation, two material models are implemented: an isotropic one $\mathcal{M}_{\text{con}}$ for the forward model and a biphasic one $\mathcal{M}_{\text{con}} + \mathcal{M}_{\text{rein}}$ representing the real reinforced system for the generative model. $\mathcal{M}_{\text{con}}$ is based on the assumption that the whole system shares the same material properties. Therefore, $\mathcal{M}_{\text{con}} = \{\rho = 2300\,\text{kg/m}^3, c_p = 900\,\text{J/(kg}\cdot\text{K)}, \alpha\}$ with α to be inferred. The values of ρ and c_p have been chosen to reflect concrete properties at $T = 293$ K according to EN 1992-1-2:2004 (European Committee for Standardization, 2004). A reference value for the conductivity from these conditions is $k = 2.0$ W/(m·K). This simplified model responds to the situation where the location of the reinforcement bars is not known beforehand or their effect on the simulation is disregarded.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

The biphasic model $\mathcal{M}^*_{\text{con}} + \mathcal{M}_{\text{rein}}$ is used to generate the "real" observations required for the parameter estimation. The system is divided into two materials, $\mathcal{M}^*_{\text{con}}$ for the concrete and $\mathcal{M}_{\text{rein}}$ for the reinforcement. The properties for the concrete are taken from EN 1992-1-2:2004 (European Committee for Standardization, 2004) as $\mathcal{M}^*_{\text{con}} = \{\rho = 2300\,\text{kg/m}^3, c_p = 900\,\text{J/(kg}\cdot\text{K)}, \alpha = 9.66\text{e-}7\,\text{m}^2/\text{s}\}$. To obtain $\mathcal{M}_{\text{rein}}$, a mix of material properties from the steel of the reinforcement and the concrete part must be considered. The region denoted for $\mathcal{M}_{\text{rein}}$ has a height of 0.02 m located between Y=0.16 m and Y=0.18 m, which, assuming the same depth, leads to a total surface perpendicular to the reinforcement of 0.0004 m^2. Using a standard bar diameter of 12 mm, the fractional volumes for concrete and steel are

$$f_{\text{steel}} = \frac{V_{\text{steel}}}{V_{\text{total}}} = \frac{A_{\text{steel}} \cdot L}{A_{\text{total}} \cdot L} = \frac{0.012^2 \cdot 0.25 \cdot \pi}{0.0004} = 0.283 \tag{61}$$

and

$$f_{\text{concrete}} = 1 - f_{\text{steel}} = 0.717. \tag{62}$$

The material parameters M_i of $\mathcal{M}_{\text{rein}}$ are then calculated following a Voigt model as

$$M_i = f_{\text{steel}} M_{\text{steel}} + f_{\text{concrete}} M_{\text{concrete}}, \tag{63}$$

which for $\mathcal{M}_{\text{steel}} = \{\rho = 7850\,\text{kg/m}^3, c_p = 440\,\text{J/(kg}\cdot\text{K)}, \alpha = 1.56\text{e-}5\,\text{m}^2/\text{s}\}$ (obtained from EN 1993-1-2:2005 (European Committee for Standardization, 2005) with $k = 54.0$ W/(m·K)) leads to $\mathcal{M}_{\text{rein}} = \{\rho = 3871\,\text{kg/m}^3, c_p = 770\,\text{J/(kg}\cdot\text{K)}, \alpha = 5.61\text{e-}6\,\text{m}^2/\text{s}\}$ with $k = 16.7$ W/(m·K). The material properties for the generative model are summarized in Table 3. There is a possible argument to use a Reuss model for some of the material properties, in particular the conductivity k. The mixed materials with such a model would be obtained as

$$\frac{1}{M_i} = \frac{f_{\text{steel}}}{M_{\text{steel}}} + \frac{f_{\text{concrete}}}{M_{\text{concrete}}}. \tag{64}$$

However, the use of Voigt's model provides a better approximation than Reuss' model for all the studied properties when the heat flow is parallel to the reinforcement (Ben-Amoz, 1970). More complex formulations are possible, but would only be justified if the direction of the reinforcement were unknown and non-homogeneous.

In a real application, it would not be possible to obtain direct measurements of the heat $\dot{Q}_{\text{objective}}$ to infer the value of α in $\mathcal{M}_{\text{con}}$. It is more common to have available temperature readings from sensors installed inside or at the surface of the concrete specimen. Four temperature sensors T_i for $i = \{1,2,3,4\}$ are placed into the system as in Figure 14 at the coordinates from Table 4. The sensors in the generative model will provide temperature measurements that at a rate of one sample every 5 min for the first 270 min that will be used to infer the unknown parameters. Analogous sensors are defined in the computational model at the same positions. Their observations will be compared during the inference procedure with those collected from the generative model.

Equation 57 is solved over the system to obtain temperature measurements at the sensors. To this end, a finite element (FE) solution is implemented. For simplicity, the geometry is discretized in a regular mesh of 20×20 first-order Lagrange elements. The time integration is performed through a backward Euler

Table 3. *Material properties used for the generative model of the thermal application case*

Material model	Density ρ [kg/m^3]	Heat capacity c_p [J/(kg·K)]	Conductivity k [W/(m·K)]	Diffusivity α [m^2/s]	Volumetric fraction f
$\mathcal{M}^*_{\text{con}}$	2300	900	2.0	9.66e-7	0.717
$\mathcal{M}_{\text{steel}}$	7850	440	54.0	1.56e-5	0.283
$\mathcal{M}_{\text{rein}}$	3871	770	16.7	5.61e-6	-

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Table 4. *Sensor coordinates for the thermal application case*

Sensor	X [m]	Y [m]
T_1	0.15	0.30
T_2	0.30	0.30
T_3	0.15	0.12
T_4	0.30	0.12

scheme with 5 min timesteps, which is adequate for the slow evolution of the temperature profile in concrete. The same implementation is used to derive the temperature field to calculate the heat flux through the system. The first 29 min will be used as a transitory regime where the external temperature linearly ramps up starting from 273 K (the same temperature as the system) to 303 K. Afterwards, the external temperature is composed of a constant component of 303 K and an additional component that simulates variable external conditions. This variable component is composed of a short-term noise sampled at every timestep from a distribution $\mathcal{N}\left(0\text{K}, (1.0\text{K})^2\right)$ and long-term component that is sampled once every five timesteps from a distribution $\mathcal{N}\left(0\text{K}, (10.0\text{K})^2\right)$ and interpolated for the timesteps in between for a smooth transition. Notice that these additional noise components in the temperature are only considered in the data generation, with the isotropic computational model assuming constant external temperature, creating an additional source of model inadequacy. The training and full temperature series are plotted in Figure 15. A comparison of the resolved temperature field at $t = 20$ min is shown in Figure 16.

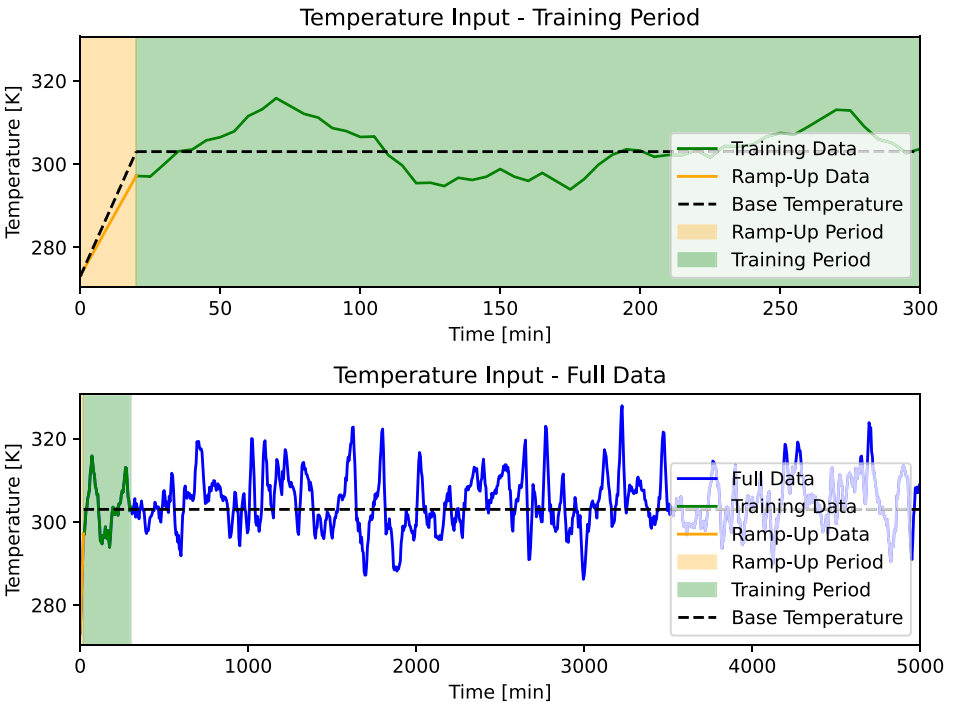
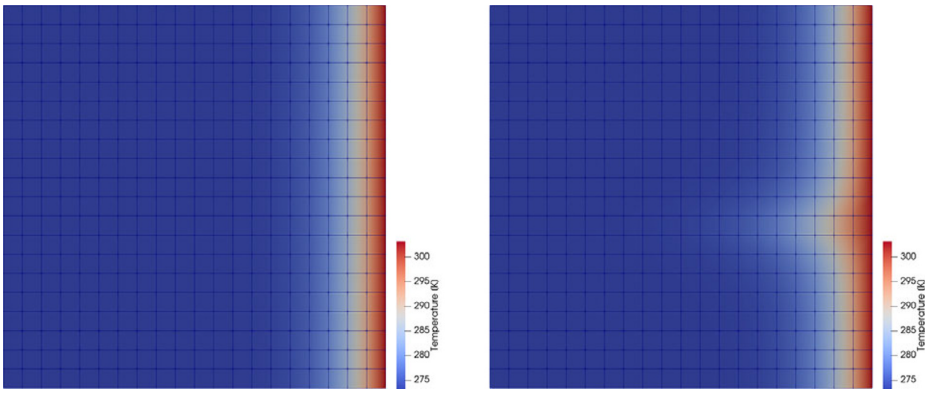


Figure 15. *External temperature series for the reinforced concrete thermal example. Training series (top) and full temperature series (bottom) for QoI evaluation.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press



(a) Isotropic system (forward model) (b) Reinforced system (generative model)

Figure 16. *Resolved temperature field a $t = 20$ min. (a) System with isotropic material properties $\mathcal{M}_{\text{con}}$ and (b) system with a band with modified material properties $\mathcal{M}_{\text{rein}}$ that represent the appearance of a reinforcement bar in that region. The isotropic system presents a uniform temperature gradient from right ($T = 303$ K) to left, while the reinforced system presents a faster development of the temperature front at the position of the reinforcement than in the rest of the system.*

The solution of the application consists of two parts: the inference of the material parameters from the temperature measurements and the prediction of the heat flux through the midline section over time. Only thanks to the embedding it is possible to quantify the uncertainty in the heat due to the model inadequacy that steams from the misspecification of the forward model compared to the generative process.

3.2.1. Temperature inference

To infer the diffusivity α of $\mathcal{M}_{\text{con}}$, an MCMC algorithm is implemented. A training dataset will be used for the parameter estimation and an additional testing dataset will be used for validation. The training dataset $\mathbf{y}$ corresponds to the measurements of $T = [T_1\ T_2\ T_3\ T_4]$ between $t = 20$ min and $t = 220$ min using the generative model with the reinforcement. The first 20 min have been rejected as they belong to the ramp-up phase and the interest is when the external temperature reaches 303 K. A white noise perturbation with $\sigma_N = 0.2$ K is prescribed at the output. For the testing dataset, the simulation is repeated with a new random seed, gathering measurements between $t = 20$ min and $t = 270$ min.The additinola observations in the testing compared to the training are for validation purposes.

The embedding for the diffusivity is defined as

$$\alpha = \alpha_m + \delta_b, \quad \delta_b \sim \mathcal{N}\left(0, (\sigma_b)^2\right), \tag{65}$$

where α_m and σ_b are the unknown parameters that must be estimated by solving the inverse problem. The prior distributions are $\pi(\alpha_m) \sim \mathcal{N}\left(10^{-6}, \left(10^{-7}\right)^2\right)$ and $\pi(\sigma_b) \sim \mathcal{LN}(-16, 0.1)$ (which corresponds to a mean of approximately 1.1e-7 and a standard deviation of approximately 1.1e-8). The PCE for computing the forward model with the embedding consist of a Hermite polynomial expansion of degree 2. The noise σ_N is properly specified with a value of 0.2, which is the same used for data generation. The MCMC algorithm is run until convergence with a threshold of $\widehat{ESS} = 837$ after 250 steps of burn-in. The results for the ABC likelihood are presented in Figure 17 and for the other likelihoods in Appendix C (Figures C1–C3). The summarized inference results are presented in Table 5.

The results do not present considerable variations between likelihood models for this case. As the noise is well prescribed, the data points with low sensitivity to changes in the parameters are not used for the training and the model can cover reasonably well the inadequacy in the model, all likelihood models produce roughly equivalent results. The posterior response is able to cover the test data points as well as

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

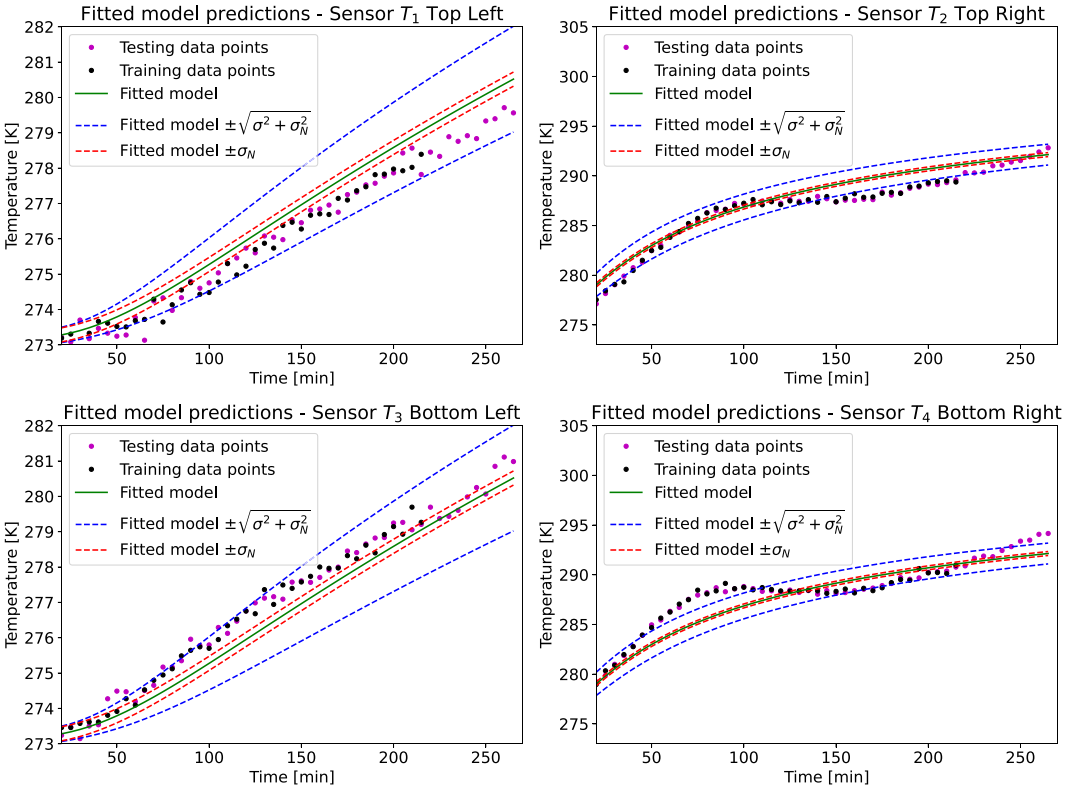


Figure 17. *Temperature sensors predictions for ABC likelihood. Dashed lines indicate the interval of* $\mu^f \pm \sigma^f$.

Table 5. *Posterior results for the thermal model and different likelihoods*

	α_m			σ_b		
Likelihood	Mean	Std	$\widehat{ESS}$	Mean	Std	$\widehat{ESS}$
ABC	1.32e-06	4.42e-13	972	2.40e-07	5.24e-13	868
GMM	1.28e-06	2.31e-08	886	1.27e-07	1.31e-08	864
RGMM	1.27e-06	1.90e-08	1009	1.62e-07	9.66e-09	848
IN	1.27e-06	1.51e-08	852	1.87e-07	8.84e-09	893

the training ones thanks to the embedding being included in the unknown variable itself. However, the values obtained for α_m with mean between 1.27e-6 and 1.32e-6 are not centered at 9.66e-7, concrete's diffusivity value. This is expected, as there is a model inadequacy between the generative system and the computational model due to the reinforcement bars, that increase the diffusivity. The prescribed noise cannot explain the observed temperature measurements, therefore the embedded model inadequacy is justified. Inferring the noise scale σ_N with an additive model would have allowed to explain the variance in the data points as a perturbation in the output, but would not necessarily be able to explain the variance in the testing range. Classical solutions to this case, such as non-additive or heteroscedastic could achieve better results for the prediction of temperature values, but they would not be transferable to other QoI such as the accumulated heat computations. An analogous situation occurs with the model inadequacy formulated following approaches based on Kennedy and O'Hagan's framework (Kennedy and O'Hagan, 2001), as shown in Andrés Arcones et al. (2024).

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

3.2.2. Heat prediction

Once the posterior distribution of the unknown variables α_m and σ_b are estimated by solving the inverse problem, the updated system is used to calculate the cumulative heat through the cross-section. The simulation is run with the full external temperature series of 5000 min, which allows the development of the full thermal profile. Analogously with the simple example, four QoI based on the cumulative heat for $t = 5000$ min are evaluated: the output $Q_{\text{obj}}(t = 5000\,\text{min})$, the mean of the predictive distribution μ_P^f, the standard deviation σ_P^f and z-value with a confidence of 95% for $Q_{\text{obj}}(t = 5000\,\text{min})$. The predicted response taking the mean values of the posterior distribution and $\hat{\alpha}_m$ and $\hat{\sigma}_b$ solving equation with them is presented in Figure 18. The propagated posterior distributions of the QoI are plotted in Figure 19. As observed in Figure 18, the posterior distributions generated by the embeddings are able to reliably envelop the real cumulative heat value. The apparent offset is a product of a larger temperature input at the first 300 min in the especific realization of the dataset, as it can be observed in Figure 15. Such is the case due to the similar sources of model inadequacy affecting the QoI calculation and the model updating, i.e. heterogeneity in the material and variations in the external temperature. If the heat simulation presented different sources of model inadequacy, such as larger temperature variations, the predicted distributions may not be able to reflect the true values. Nevertheless, they always provide a more informative assessment on the reliability of the model predictions. In comparison, the prediction when no model inadequacy term is included differs from the true value by more than 10 kJ despite not providing a significant confidence interval, leading to overconfident results.

Analogous conclusions can be extracted from the analysis of the pushed-forward QoIs. While the model without inadequacy term produces a point distribution due to the lack of a noise model for the cumulative heat prediction, the embedded models produce distributions that could have generated the true value of 91338 J. This is reflected in the z-values, where all the sampled pairs of unknown parameters generate distributions that produce z-values significantly smaller than 1.96. Due to its almost zero variance in the posterior distribution, a z-test would not have been reasonable for the model without inadequacy term unless an error model is prescribed. The choice of likelihood does not impact significantly the mean value of the predictions μ_P^f, laying between 82300 and 82550 J. The main difference resides in the value for σ_P^f and the spread of the distributions for each QoI. In general, ABC produces wider distributions of Q_{obj}, with more concentrated values for μ_P^f and σ_P^f. GMM, RGMM and IN provide comparable results with each other with a more concentrated Q_{obj} but larger dispersion in the other QoIs. Only slight shifts and deformations in the distributions of the QoI can be observed, tending the global likelihoods to wider ones with more variance.

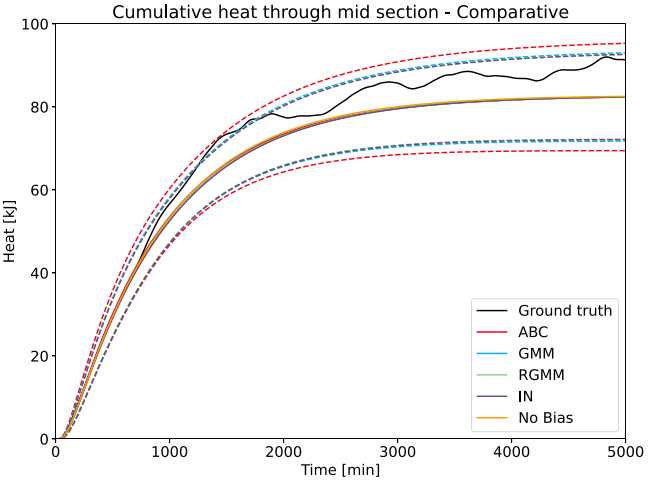


Figure 18. *Temperature heat prediction for each likelihood. Dashed lines indicate the interval of* $\mu^f \pm \sigma^f$.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

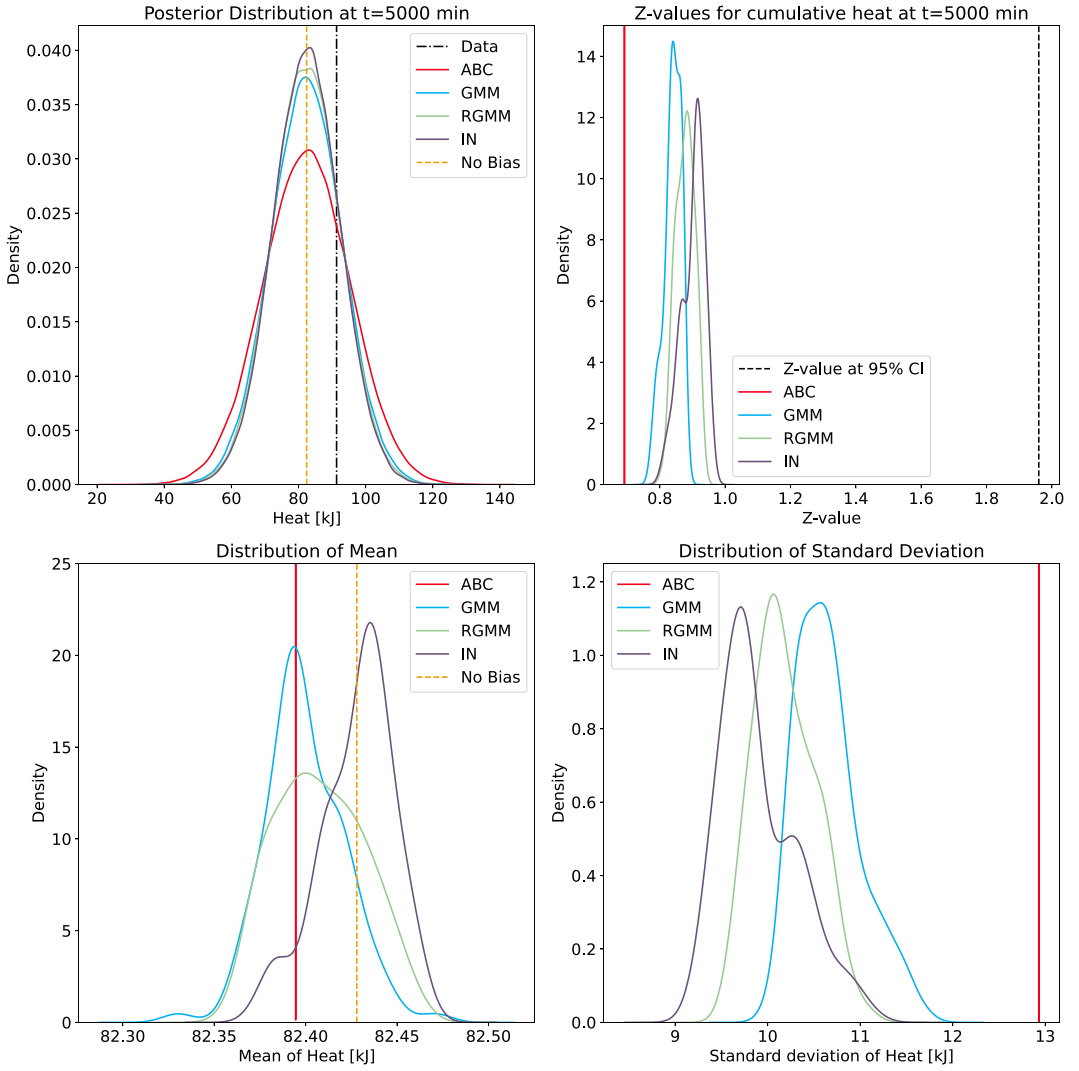


Figure 19. *Analysis of quantities of interest by propagating the posterior distribution of α and σ_b for the thermal example.*

3.3. Discussion on likelihood selection

As it is clear from the formulation and applications, the likelihood models present substantial differences in their behavior that influence their choice over the others depending on the characteristics of the dataset:

- **Number of parameters:** ABC requires prescribing γ and ϵ, while GMM and RGMM only require γ and IN none. As already mentioned, the choice of ϵ is not trivial for datasets with prescribed noise and can largely impact the results. Additionally, the choice of γ corresponds to previous beliefs on the expected predicted distribution. In particular, a value of $\gamma = \sqrt{\frac{\pi}{2}}$ is expected for ABC and $\gamma = 1.0$ for the others. The difference in the value of choice for γ steams from the formulation of the matching of the variance, which induces a half-normal distribution as discussed in Section 2.5.2, while the others do not.
- **Precision specification:** As already mentioned and studied in Sargsyan et al. (2015), the posterior distribution of ABC depends on the specification of the precision parameter ϵ. As the objective is doing a perfect matching of the moments, ϵ is required to be as small as possible. However, it is

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

necessary that $\epsilon > 0$, for numerical stability.In particular, ϵ is usually chosen such that ϵ^2 is not under machine precision, e.g. $\epsilon = 10^{-6}$ assuming a machine precision of 10^{-12}. Larger values of ϵ result in an artificially inflated posterior, despite potentially reducing the sensitivity of ABC to the misspecification of noise by relaxing the matching condition. Therefore, higher arbitrary values of ϵ have no real fundament. Using the noise variance in the residuals mitigates the perfect matching of the mean, but using ϵ is unavoidable unless higher order moments are introduced. Contrarily, the respective variances in GMM, RGMM and IN act naturally as a weight on the precision required in the residuals, penalizing higher values of the variance and avoiding the perfect matching.

- **Noise misspecification:** As shown in Appendix B, ABC is very sensitive to the specification of the noise amplitude, which can mislead the parameter updating procedure to values of $\boldsymbol{\theta}$ further from the expected ones. The misspecification of the prescribed noise generally leads to larger acceptance ratios and wider posterior distribution of $\boldsymbol{\theta}$ for GMM, RGMM and IN.
- **Assumptions of normality:** The four likelihoods considered suppose different distributions of the residuals while still keeping the assumption of normality. ABC fits exactly the moments under the consideration of Gaussian noise. GMM impose the condition that all the residuals are characterized by the same mixture model, which may not be reasonable. RGMM assume that the relative errors between observations and predicted distribution follow such a common standardized mixture model instead. Finally, IN assumes independent normal distributions for each observation. Depending on the application case, some of these assumptions may not hold. An heteroscedasticity in the noise model can be incorporated in all the likelihoods with minimal modifications, while the assumptions of additive Gaussian noise and Gaussian output require extensive reformulations, specially for IN, GMM, and RGMM. ABC can be adapted to those situations with the current formulation as long as the normal error model for the statistical moments still holds.
- **Inferred posterior distribution:** By construction, ABC aims to match the statistical moments exactly up to a tolerance ϵ, which will lead to narrow inferred posterior distributions of the parameters $\boldsymbol{\theta}$ for acceptable values of ϵ compared to GMM, RGMM and IN. In particular, larger predicted variances imply more samples needed for the convergence of the chain (Schälte and Hasenauer, 2020). Additionally, the formulation of $\mathcal{L}_2(\boldsymbol{\theta})$ for GMM and RGMM produces a flatter likelihood function in the direction of the variance-governing parameters, which further hinders the convergence for those parameters.
- **Observations out of the predictive range:** If the observations cannot be replicated by the computational model independently of the choice in the parameters, then embedded approaches lead to suboptimal results. RGMM and IN are expected to be the most impacted if those values are at points with low predictive variance. However, the regularizing effect of considering the whole dataset in RGMM and GMM for the goodness of fit instead of fitting every point, as in IN or ABC, should reduce the impact of these observations if they are not predominant in the dataset.
- **Fit to exact observations:** As pointed out in Sargsyan et al. (2015), marginalizing formulations such as IN would lead to wrong results if σ_i^f is significantly smaller for a given observation i than for the others. In those cases, such observation will be overfitted by reducing the residual to the minimum to the detriment of the other observation points with larger variances. The presence of a prescribed noise reduces this effect in IN, GMM and RGMM likelihoods, as all observations will have at least variance σ_N. For an analysis on this phenomena, refer to Sargsyan et al. (2015).

4. Conclusions

This paper has introduced several significant advances in the implementation of embedded model inadequacy approaches for quantifying model uncertainties. First, a more interpretable embedding representation was developed as an alternative to the one presented by Sargsyan et al. (2019). Although this approach requires specifying a distribution for the model form uncertainty, it reduces the dimensionality of the polynomial chaos expansion (PCE) needed to describe the embedding and improves the separation between fitting the embedding's PCE and inferring the model parameters. Next, it was

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

demonstrated that the Approximate Bayesian Computation (ABC) likelihood formulation, commonly used to update model parameters with an embedded model inadequacy, is highly sensitive to the prescribed measurement noise and requires careful selection of the parameters γ and ϵ. One of the main contributions of this work is the explicit treatment and analysis of measurement noise, which had previously been mentioned in past studies but not thoroughly examined. To address this, the ABC likelihood was extended to properly account for measurement noise and accurately quantify predictive variance by imposing $\gamma = \sqrt{\pi/2}$. Additionally, two alternative likelihood formulations, global moment-matching (GMM) and relative global moment-matching (RGMM), were proposed to mitigate sensitivity issues. These formulations were analyzed and compared against the modified ABC likelihood and the independent normal noise formulation using a simple linear and a complex transient thermal example. Finally, a key contribution of this paper is the analysis of how the choice of likelihood influences the propagation of uncertainties, quantified through the embedding approach, to the quantities of interest (QoIs) computed using the updated model and parameters.

The inclusion of the embedded model inadequacy term enables the quantification of uncertainty arising during parameter updating due to modeling assumptions, potentially capturing the full range of uncertainties in the measured observations. By introducing the embedding model inadequacy term as a stochastic extension of the unknown parameters, the proposed formulation allows for straightforward propagation of these uncertainties to other predicted QoIs that depend on the same model parameters. However, as shown in this work, the prescribed measurement noise critically influences the selection of the appropriate likelihood formulation and affects the convergence of the Bayesian updating process when an embedding is used. In particular, ABC likelihoods have been identified as the most sensitive to poor noise prescriptions within the embedded model form uncertainty framework. Additionally, the embedded model inadequacy formulation limits predictions to the range of responses that the model is capable of producing. As a result, discrepancies that cannot be explained by changes in the model parameters, such as offsets or outlier values, can skew the inferred parameters. The ABC likelihood formulation tends to be more sensitive to outliers, while RGMM and IN are more impacted by offsets or outliers near values with low variance. GMM and RGMM, on the other hand, are reliable only if the assumption of the existence of a common mixture model in the residuals is valid. Identifying which of these effects are present during model updating is not always possible and often depends on data analysis or practitioner experience. However, selecting the right likelihood function is crucial in certain cases. Regardless of the chosen likelihood formulation, the inclusion of embedded model inadequacy improves predictive posterior distributions, allowing for more informed decision-making, albeit at the cost of increased computational effort during the update phase. While hierarchical Bayesian approaches can produce comparable results when the conditional distributions of the hierarchical formulation are known, a thorough comparison between these methodologies is left for future work.

One of the key strengths of this methodology is its ability to propagate the uncertainty captured by the embedded model inadequacy. This paper has demonstrated how propagating the full posterior distribution of the unknown parameters through QoI calculations provides insights that are not possible with point estimates or model inadequacy-free formulations. Generating distributions for the statistical moments of the predicted QoI enables the use of inference techniques for a more robust analysis of the results. Moreover, the interpretable embedding formulation proposed here allows the propagation of uncertainties related only to specific parameters that influence the QoI being calculated. Nevertheless, when multiple parameters are inferred alongside an associated model form uncertainty, it is likely that their inadequacy terms will be correlated. Implementing and propagating such correlation structures to QoIs where only certain parameters are relevant remains as future work.

While the numerical examples presented in this study involve relatively simple probabilistic structures, they were deliberately selected to isolate and clearly illustrate the methodological contributions and the effects of measurement noise, likelihood formulation, and embedding representation. The proposed approach is not inherently limited to unimodal or low-dimensional distributions; however, its performance for highly complex, multimodal, or non-Gaussian priors has not been explicitly demonstrated here. In such cases, the variance introduced by the embedding may encompass several of the modes, effectively merging

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

distinct peaks into a single broad region of uncertainty. This blending effect can obscure the identification of separate modes and lead to an interpretation of multimodality as merely an increase in overall variance.

Careful selection of the embedding structure and potential modifications to the likelihood formulation could help mitigate these issues, though such strategies typically require some degree of prior knowledge about the underlying distribution. In particular, multi-modal summary statistics or histogram-based distances could be integrated in the likelihood formulations in case that multi-modality is expected. This limitation delineates the current scope of applicability and highlights an important direction for future research. Additionally, it is important to note that the examples presented in this work involve a relatively low-dimensional set of parameters, chosen to allow a clear interpretation of the effects of measurement noise and likelihood selection. Extending the proposed formulation to higher-dimensional parameter spaces represents a natural next step. In many engineering applications, the parameters to be inferred are numerous and may be affected by multiple, potentially interacting sources of bias. Accounting for these multiple discrepancy sources within the embedding would require strategies that ensure the separability and identifiability of the variance contributions associated with each parameter group. Such separability is essential to correctly attribute uncertainties and to reliably propagate them to derived QoIs. Addressing these challenges will be key for applying embedded model inadequacy approaches to complex, real-world systems. Introduction of latent-space approaches such as in Lee et al. (2025) present promising alternatives for dealing with such high-dimensional cases.

In conclusion, the advances presented in this paper represent a significant step toward more effective quantification and propagation of uncertainties in model parameters through embedded model inadequacy approaches. One of the main contributions of this work is the explicit consideration and analysis of measurement noise, which had been previously acknowledged in past methods but never thoroughly examined in terms of its impact on the model updating process. Future work should focus on further validating this methodology by applying it to real measurement data from sensors to assess its practical applicability. Additionally, extending the current approach to hierarchical Bayesian formulations and exploring the propagation of correlated inadequacy terms to various QoIs are important avenues for continued research. These developments promise to enhance the robustness and reliability of uncertainty quantification in complex models.

Data availability statement. The data that support the findings of this study are available in https://doi.org/10.5281/zenodo.15720780 ([dataset] Andres Arconés, 2025). The methodology implemented in this article is available in the software package `probeye` ([dataset] BAMResearch, 2024), available in the GitHub repository https://github.com/BAMresearch/probeye.

Acknowledgements. The authors made use of ChatGPT to assist with the drafting of this article for the improvement of text clarity and grammar. GPT-4o was accessed from https://chatgpt.com/ and used without modification on 05/03/2025. GPT-5 was used without modification on 27/10/2025 for the improvement of text clarity and grammar during the review process.

Author contribution. Conceptualization: M.W., P.-S.K., J.F.U.; Methodology: D.A.A.; Software: D.A.A.; Data curation: D.A.A.; Data visualization: D.A.A.; Writing original draft: D.A.A.; Supervision: M.W., P.-S.K., J.F.U.; Review and editing: M.W., P.-S.K., J.F.U. All authors approved the final submitted draft.

Funding statement. This research was supported by the German Research Foundation (DFG) in the special focus program SPP 2388/1 "Hundert plus – Verlängerung der Lebensdauer komplexer Baustrukturen durch intelligente Digitalisierung" (Hundred Plus - Extending the service life of complex building structures through intelligent digitalisation) in the subproject C07: Data driven model adaptation for identification of digital twins of bridges.

Competing interests. The authors declare none.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Andres Arconés D (2025) *Embedded Model Form Uncertainty Quantification with Measurement Noise for Bayesian Model Calibration.* Zenodo [dataset]. https://doi.org/10.5281/ZENODO.15720780.

Andrés Arcones D, **Weiser M**, **Koutsourelakis P-S and Unger JF** (2023) A Bayesian framework for simulation-based digital twins of bridges. *ce/papers 6*(5), 734–740. https://doi.org/10.1002/cepa.2177.

Andrés Arcones D, **Weiser M**, **Koutsourelakis P and Unger JF** (2024) Model bias identification for Bayesian calibration of stochastic digital twins of bridges. *Applied Stochastic Models in Business and Industry 41*, e2897. https://doi.org/10.1002/asmb.2897.

BAMResearch (2024) *Probeye*. Zenodo [dataset]. Retrieved from https://github.com/BAMresearch/probeye and https://doi.org/10.5281/zenodo.15584363

Barbillon P, **Forte A and Paulo R** (2024) Screening the discrepancy function of a computer model. *Technometrics 66*, 394–405. https://doi.org/10.1080/00401706.2024.2319138.

Bayarri MJ, **Berger JO and Liu F** (2009) Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis 4*(1), 119–150. https://doi.org/10.1214/09-ba404.

Ben-Amoz M (1970) The effective thermal properties of two phase solids. *International Journal of Engineering Science 8*(1), 39–47. https://doi.org/10.1016/0020-7225(70)90013-3.

Bi S, **Broggi M and Beer M** (2019) The role of the bhattacharyya distance in stochastic model updating. *Mechanical Systems and Signal Processing 117*, 437–452. https://doi.org/10.1016/j.ymssp.2018.08.017.

Bojke L, **Claxton K**, **Sculpher M and Palmer S** (2009) Characterizing structural uncertainty in decision analytic models: A review and application of methods. *Value in Health 12*(5), 739–749. https://doi.org/10.1111/j.1524-4733.2008.00502.x.

Brynjarsdóttir J and O'Hagan A (2014) Learning about physical parameters: The importance of model discrepancy. *Inverse Problems 30*(11), 114007. https://doi.org/10.1088/0266-5611/30/11/114007.

Campbell K (2006) Statistical calibration of computer simulations. *Reliability Engineering and System Safety 91*(10–11), 1358–1363. https://doi.org/10.1016/j.ress.2005.11.032.

Daly AC, **Cooper J**, **Gavaghan DJ and Holmes C** (2017) Comparing two sequential Monte Carlo samplers for exact and approximate Bayesian inference on biological models. *Journal of the Royal Society Interface 14*(134), 20170340. https://doi.org/10.1098/rsif.2017.0340.

Dietz L (2010) *Directed factor graph notation for generativemodels* (Technical Report, Max Planck Institute for Informatics).

European Committee for Standardization (2004) *EN 1992-1-1 Eurocode 2: Design of Concrete Structures - Part 1-2: Structureal Fire Design*. Brussels: CEN.

European Committee for Standardization (2005) *EN 1993-1-1 Eurocode 3: Design of Steel Structures - Part 1-2: General Rules - Structural Fire Design*. Brussels: CEN.

Fearnhead P and Prangle D (2012) Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society Series B: Statistical Methodology 74*(3), 419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x.

Feinberg J and Langtangen HP (2015) Chaospy: An open source tool for designing methods of uncertainty quantification. *Journal of Computational Science 11*, 46–57. https://doi.org/10.1016/j.jocs.2015.08.008.

Foreman-Mackey D, **Hogg DW**, **Lang D and Goodman J** (2013) Emcee: The MCMC hammer. *Publications of the Astronomical Society of the Pacific 125*(925), 306–312. https://doi.org/10.1086/670067.

Gelman A, **Carlin J**, **Stern H**, **Dunson D**, **Vehtari A and Rubin D** (2013) *Bayesian Data Analysis*, 3rd Edn. London: Chapman and Hall. https://doi.org/10.1201/b16018.

Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136.

Goodman J and Weare J (2010) Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science 5*(1), 65–80. https://doi.org/10.2140/camcos.2010.5.65.

Gupta HV, **Clark MP**, **Vrugt JA**, **Abramowitz G and Ye M** (2012) Towards a comprehensive assessment of model structural adequacy. *Water Resources Research 48*(8). https://doi.org/10.1029/2011wr011044.

Huan X, **Safta C**, **Sargsyan K**, **Geraci G**, **Eldred MS**, **Vane ZP**, **Lacazek G**, **Oefelein JC and Najm HN** (2017) Global sensitivity analysis and quantification of model error for large eddy simulation in scramjet design. In *19th AIAA Non-deterministic Approaches Conference*. Grapevine, Texas: American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2017-1089.

Huber PJ and Ronchetti EM (2009) *Robust Statistics*. New Jersey: Wiley. https://doi.org/10.1002/9780470434697.

Jakeman JD, **Franzelin F**, **Narayan A**, **Eldred M and Plfüger D** (2019) Polynomial chaos expansions for dependent random variables. *Computer Methods in Applied Mechanics and Engineering 351*, 643–666. https://doi.org/10.1016/j.cma.2019.03.049.

John, D. N. (2021). *Uncertainty quantification for an electric motor inverse problem - tackling the model discrepancy challenge* (Doctoral dissertation, Heidelberg University). https://doi.org/10.11588/HEIDOK.00029450

John, D. N., **Stohrer, L.**, **Schillings, C.**, **Schick, M.**, & **Heuveline, V.** (2021). Hierarchical surrogate-based approximate Bayesian computation for an electric motor test bench. ArXiv, https://doi.org/10.48550/arXiv.2106.09597.

Kaipio J and Somersalo E (2006) *Statistical and Computational Inverse Problems*. New York: Springer Science+Business Media.

Kennedy MC and O'Hagan A (2001) Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464. https://doi.org/10.1111/1467-9868.00294.

Kleijn B and van der Vaart A (2012) The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics 6*, 354–381. https://doi.org/10.1214/12-ejs675.

Lee S, **Yaoyama T**, **Kitahara M and Itoi T** (2025) Latent space-based stochastic model updating. *Mechanical Systems and Signal Processing 235*, 112841. https://doi.org/10.1016/j.ymssp.2025.112841.

Leone FC, **Nelson LS and Nottingham RB** (1961) The folded normal distribution. *Technometrics 3*(4), 543–550. https://doi.org/10.1080/00401706.1961.10489974.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Leoni N, **Maître OL**, **Rodio M-G and Congedo PM** (2024) Bayesian calibration with adpative model discrepancy. *International Journal for Uncertainty Quantification 14*(1), 19–41. https://doi.org/10.1615/int.j.uncertaintyquantification.2023046331.

Li Q, **Du X**, **Ni P**, **Han Q**, **Xu K and Bai Y** (2024) Improved hierarchical bayesian modeling framework with arbitrary polynomial chaos for probabilistic model updating. *Mechanical Systems and Signal Processing 215*, 111409. https://doi.org/10.1016/j.ymssp.2024.111409.

Marjoram P, **Molitor J**, **Plagnol V and Tavaré S** (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences 100*(26), 15324–15328. https://doi.org/10.1073/pnas.0306899100.

Mortensen JJ, **Kaasbjerg K**, **Frederiksen SL**, **Nørskov JK**, **Sethna JP and Jacobsen KW** (2005) Bayesian error estimation in density-functional theory. *Physical Review Letters 95*, 216401. https://doi.org/10.1103/PhysRevLett.95.216401.

Obermeyer F, **Bingham E**, **Jankowiak M**, **Pradhan N**, **Chiu J**, **Rush A and Goodman N** (2019) Tensor variable elimination for plated factor graphs. In Chaudhuri K and Salakhutdinov R (eds), *Proceedings of the 36th International Conference on Machine Learning*, Vol. *97*. Long Beach, California: PMLR. pp. 4871–4880.

Oliver TA, **Terejanu G**, **Simmons CS and Moser RD** (2015) Validating predictions of unobserved quantities. *Computer Methods in Applied Mechanics and Engineering 283*, 1310–1335. https://doi.org/10.1016/j.cma.2014.08.023.

Pernot P (2017) The parameter uncertainty inflation fallacy. *The Journal of Chemical Physics 147*(10). https://doi.org/10.1063/1.4994654.

Pernot P and Cailliez F (2017) A critical review of statistical calibration/prediction models handling data inconsistency and model inadequacy. *AICHE Journal 63*(10), 4642–4665. https://doi.org/10.1002/aic.15781.

Plumlee M (2017) Bayesian calibration of inexact computer models. *Journal of the American Statistical Association 112*(519), 1274–1285. https://doi.org/10.1080/01621459.2016.1211016.

Prangle D (2017) Adapting the ABC distance function. *Bayesian Analysis 12*(1). https://doi.org/10.1214/16-ba1002.

Robert CP (2007) *The Bayesian Choice*. New York: Springer Science+Business Media.

Robert CP and Casella G (2004) *Monte Carlo Statistical Methods*. New York: Springer Science+Business Media. https://doi.org/10.1007/978-1-4757-4145-2.

Sargsyan K, **Huan X and Najm HN** (2019) Embedded model error representation for Bayesian model calibration. *International Journal for Uncertainty Quantification 9*(4), 365–394. https://doi.org/10.1615/int.j.uncertaintyquantification.2019027384.

Sargsyan K, **Najm HN and Ghanem R** (2015) On the statistical calibration of physical models. *International Journal of Chemical Kinetics 47*(4), 246–276. https://doi.org/10.1002/kin.20906.

Schälte Y and Hasenauer J (2020) Efficient exact inference for dynamical systems with noisy measurements using sequential approximate Bayesian computation. *Bioinformatics 36*(Supplement 1), i551–i559. https://doi.org/10.1093/bioinformatics/btaa397.

Sisson SA and Fan Y (2010) Likelihood-free MCMC. In Brooks S, Gelman A, Jones G and Meng X-L (eds), *Handbook of Markov Chain Monte Carlo*. Boca Raton, Florida: Chapman and Hall/CRC.

Sisson S, **Fan Y and Beaumont M** (2018) *Handbook of Approximate Bayesian Computation*, first ed. Boca Raton, Florida: Chapman and Hall/CRC. https://doi.org/10.1201/9781315117195.

Sokal A (1997) Monte Carlo methods in statistical mechanics: Foundations and new algorithms. In DeWitt-Morette C, Cartier P and Folacci A (eds), *Functional Integration*. New York: Springer Science+Business Media. pp. 131–192. https://doi.org/10.1007/978-1-4899-0319-8_6.

Strong M and Oakley JE (2014) When is a model good enough? Deriving the expected value of model improvement via specifying internal model discrepancies. *SIAM/ASA Journal on Uncertainty Quantification 2*(1), 106–125. https://doi.org/10.1137/120889563.

Sudret, B. (2021). Polynomial chaos expansions in 90 minutes. https://doi.org/10.3929/ETHZ-B-0005088520

Sung C and Tuo R (2024) A review on computer model calibration. *WIREs Computational Statistics 16*(1), e1645. https://doi.org/10.1002/wics.1645.

Turner BM and Van Zandt T (2013) Hierarchical approximate bayesian computation. *Psychometrika 79*(2), 185–209. https://doi.org/10.1007/s11336-013-9381-x.

Vaidya S, **Ambad P and Bhosle S** (2018) Industry 4.0 – A glimpse. *Procedia Manufacturing 20*, 233–238. https://doi.org/10.1016/j.promfg.2018.02.034.

van der Vaart AW (2000) *Asymptotic Statistics*. New York: Cambridge University Press.

Vats D, **Flegal JM and Jones GL** (2019) Multivariate output analysis for Markov chain Monte Carlo. *Biometrika 106*(2), 321–337. https://doi.org/10.1093/biomet/asz002.

Vats D and Knudson C (2021) Revisiting the Gelman-Rubin diagnostic. *Statistical Science 36*(4). https://doi.org/10.1214/20-sts812.

Walker W, **Harremoës P**, **Rotmans J**, **van der Sluijs J**, **van Asselt M**, **Janssen P and Krayer von Krauss M** (2003) Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment 4*(1), 5–17. https://doi.org/10.1076/iaij.4.1.5.16466.

Wilkinson RD (2013) Approximate bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology 12*(2), 129–141. https://doi.org/10.1515/sagmb-2013-0010.

Wu J-L, **Levine ME**, **Schneider T and Stuart A** (2024) Learning about structural errors in models of complex dynamical systems. *Journal of Computational Physics 513*, 113157. https://doi.org/10.1016/j.jcp.2024.113157.

Xiu D and Karniadakis GE (2002) The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing 24*(2), 619–644. https://doi.org/10.1137/s1064827501387826.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

A. Monte Carlo-Markov Chain stopping criteria

In this paper, the MCMC sampler based on ensembles of chains implemented in `emcee` (Foreman-Mackey et al., 2013) is used for the Bayesian updating framework. In the chosen implementation, a stretch move as defined in Goodman and Weare (2010) proposes the new samples for the next iteration of the MCMC algorithm based on the current values of the ensemble of chains. The chains in a given ensemble are therefore not independent. This correlation invalidates the use of those convergence criteria that are base on the independence of the chains such as Gelman-Rubin's $\hat{R}$ metric (Gelman and Rubin, 1992; Vats and Knudson, 2021). Alternatively, the integrated autocorrelation time τ_∞ (Sokal, 1997) quantifies how long it takes for the values in a chain or ensemble to become effectively uncorrelated with one another and is in practice a metric of the quality of the ensemble that does not require independence of the chains. A stopping criteria based on the effective sample size ESS obtained from τ_∞ is imposed on the MCMC algorithm to determine the number of iterations that must be evaluated.

First, an objective ESS threshold must be define to determine when the chain is converged. Following Vats et al. (2019), an estimated $\widehat{ESS}$ such that

$$\widehat{ESS} \geq W_{p,\alpha,\epsilon} \tag{A1}$$

is calculated a priori. The value $W_{p,\alpha,\epsilon}$ depends only on the dimensionality of the problem p, the level of confidence α of the considered regions and the desired precision ϵ. For large samples, i.e. sufficiently long chains, Vats et al. (2019) prove that a good approximation for $W_{p,\alpha,\epsilon}$ is

$$\widehat{ESS} \geq W_{p,\alpha,\epsilon} = \frac{2^{2/p}\pi}{\left(p\Gamma(p/2)\right)^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\epsilon^2}. \tag{A2}$$

Additionally, a maximum number of iterations steps $n_{\max}$ is established. Therefore, the MCMC algorithm will be continued while, for a given iteration i,

$$ESS_i < \frac{2^{2/p}\pi}{\left(p\Gamma(p/2)\right)^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\epsilon^2} \text{ and } i < n_{\max}. \tag{A3}$$

The effective sample size ESS_i is directly related with τ_∞ at iteration i as

$$ESS_i = \frac{n_i m}{\tau_\infty}, \tag{A4}$$

where n_i is the number of steps at iteration i and m is the number of walkers, i.e. chains in the ensemble. In practice, τ_∞ is approximated by an estimation following the implementation described in Sokal (1997). To avoid excessive computations, ESS_i is not calculated after every sample of the MCMC algorithm but after a batch of iterations with predefined size have been evaluated. The MCMC will be stopped when the thresholds of Equation A3 are exceeded for the chains of each unknown parameter θ that is being sampled. Convergence plots for the simple case of Section 3.1 are presented in Figure A1, where the stopping criteria has been chosen following the procedure previously described.

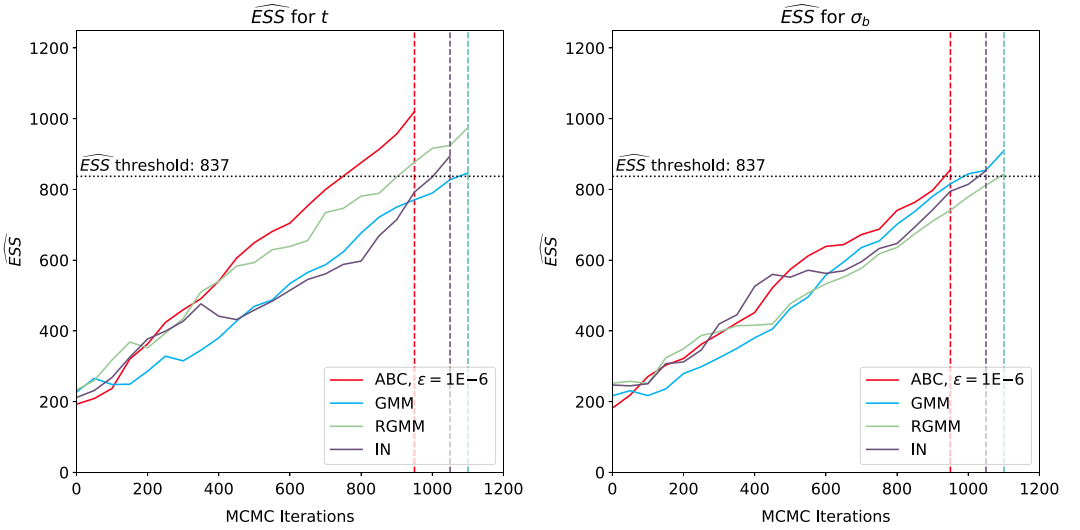


Figure A1. *Estimated sample size (ESS) for the simple example over the MCMC iterations for simple linear application case.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

B. Moment-matching likelihood behavior with measurement noise

Let (A_i) denote the residual at observation i, defined as $A_i = \left|\mu_i^f - y_i\right|$, and let (B_i) denote the total standard deviation of the model, $B_i = \sqrt{(\sigma_N)^2 + \left(\sigma_i^f\right)^2}$. The quantities A_i and B_i are independent, since A_i depends on the unknown model parameters, while B_i depends on the auxiliary parameters introduced through the embedded model inadequacy. Then, the noisy moment-matching likelihood from Equation 24, expressed in logarithmic form, is

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\frac{n_y}{2}\log\left(2\pi\epsilon^2(\sigma_N)^2\right) \sum_{i=1}^{n_y}\left[\frac{(A_i)^2}{2(\sigma_N)^2} + \frac{(B_i - \gamma A_i)^2}{2\epsilon^2}\right]. \tag{B1}$$

The objective is to investigate how changes in A_i (the residuals) and B_i (the total deviations) affect the log-likelihood. By analyzing the signs and magnitudes of its first and second derivatives, we can determine how the optimization process will tend to adjust the model residuals and predicted uncertainties, and where potential optima occur.

Differentiating the log-likelihood with respect to each residual A_i gives

$$c\frac{\partial \log \mathcal{L}}{\partial A_i} = -\frac{A_i}{(\sigma_N)^2} + \frac{\gamma(B_i - \gamma A_i)}{\epsilon^2} = -\left(\frac{1}{(\sigma_N)^2} + \frac{\gamma^2}{\epsilon^2}\right)A_i + \frac{\gamma}{\epsilon^2}B_i, \tag{B2}$$

where $c = \frac{n_y}{2}\log\left(2\pi\epsilon^2(\sigma_N)^2\right)$. Analogously, differentiation with respect to the total deviation B_i yields

$$c\frac{\partial \log \mathcal{L}}{\partial B_i} = -\frac{B_i - \gamma A_i}{\epsilon^2} = \frac{\gamma A_i - B_i}{\epsilon^2}. \tag{B3}$$

These derivatives are defined independently for each observation i, reflecting the separable structure of the log-likelihood. For interpretation, it is convenient to consider the mean derivatives over all observations:

$$\frac{c}{n_y}\sum_{i=1}^{n_y}\frac{\partial \log \mathcal{L}}{\partial A_i} = -\left(\frac{1}{(\sigma_N)^2} + \frac{\gamma^2}{\epsilon^2}\right)\overline{A} + \frac{\gamma}{\epsilon^2}\overline{B}, \quad \frac{c}{n_y}\sum_{i=1}^{n_y}\frac{\partial \log \mathcal{L}}{\partial B_i} = \frac{\gamma}{\epsilon^2}\overline{A} - \frac{1}{\epsilon^2}\overline{B}, \tag{B4}$$

where $\bar{A}$ and $\bar{B}$ are the sample means of A_i and B_i, respectively. A positive mean derivative with respect to A_i implies that increasing the residuals raises the likelihood, suggesting that the optimizer would favor poorer mean-value fits. Conversely, a negative derivative indicates a tendency toward smaller residuals and a better mean-value fit. Similarly, a positive derivative with respect to B_i favors higher predicted model-form uncertainty, while a negative value promotes reducing it.

Because the likelihood is separable across observations, the Hessian matrix is block-diagonal, each block corresponding to one observation pair (A_i, B_i). For each i:

$$c\frac{\partial^2 \log \mathcal{L}}{(\partial A_i)^2} = -\left(\frac{1}{(\sigma_N)^2} + \frac{\gamma^2}{\epsilon^2}\right), \quad c\frac{\partial^2 \log \mathcal{L}}{(\partial B_i)^2} = -\frac{1}{\epsilon^2}, \quad c\frac{\partial^2 \log \mathcal{L}}{\partial A_i \partial B_i} = c\frac{\partial^2 \log \mathcal{L}}{\partial B_i \partial A_i} = \frac{\gamma}{\epsilon^2}. \tag{B5}$$

The corresponding 2×2 Hessian block for observation i is

$$\mathbf{H}_i = c\begin{pmatrix} -\left(\frac{1}{(\sigma_N)^2} + \frac{\gamma^2}{\epsilon^2}\right) & \frac{\gamma}{\epsilon^2} \\ \frac{\gamma}{\epsilon^2} & -\frac{1}{\epsilon^2} \end{pmatrix}, \tag{B6}$$

and its determinant and trace are

$$\det(\mathbf{H}_i) = \frac{c^2}{\epsilon^2(\sigma_N)^2} > 0, \quad \operatorname{tr}(\mathbf{H}_i) = -c\left(\frac{1}{(\sigma_N)^2} + \frac{\gamma^2 + 1}{\epsilon^2}\right) < 0. \tag{B7}$$

Hence, each block is negative definite, and so is the full Hessian, whose determinant is

$$\det(\mathbf{H}) = c^2\left(\frac{1}{\epsilon^2(\sigma_N)^2}\right)^{n_y} > 0. \tag{B8}$$

Therefore, the log-likelihood is globally concave with respect to $\{A_i, B_i\}_{i=1}^{n_y}$, and all stationary points correspond to local maxima. Taking the average form of the gradients,

$$c\frac{\partial \log \mathcal{L}}{\partial \overline{A}} > 0 \Longleftrightarrow -\left(\frac{1}{(\sigma_N)^2} + \frac{\gamma^2}{\epsilon^2}\right)\overline{A} + \frac{\gamma}{\epsilon^2}\overline{B} > 0, \quad c\frac{\partial \log \mathcal{L}}{\partial \overline{B}} > 0 \Longleftrightarrow \frac{\gamma}{\epsilon^2}\overline{A} - \frac{1}{\epsilon^2}\overline{B} > 0. \tag{B9}$$

These expressions describe how the mean residuals $\overline{A}$ and mean predicted deviations $\overline{B}$ evolve during optimization. They form the basis for the behavioral classification summarized in Table B1. For the likelihood form given by Equation 19, equivalent results follow by setting $(\sigma_N)^2 = \epsilon^2$.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

Table B1. *Cases for moment-matching log-likelihood behavior under noise*

	Cases	$\frac{\partial \log \mathcal{L}}{\partial \bar{A}}$	$\frac{\partial \log \mathcal{L}}{\partial \bar{B}}$
(I)	$\bar{B} < \gamma \bar{A}$	Negative	Positive
(II)	$\bar{B} = \gamma \bar{A}$	Negative	Zero
(III)	$\bar{B} \in \left(\gamma \bar{A}, \left(\gamma + \frac{\epsilon^2}{(\sigma_N)^2}\right)\bar{A}\right)$	Negative	Negative
(IV)	$\bar{B} = \left(\gamma + \frac{\epsilon^2}{(\sigma_N)^2}\right)\bar{A}$	Zero	Negative
(V)	$\bar{B} > \left(\gamma + \frac{\epsilon^2}{(\sigma_N)^2}\right)\bar{A}$	Positive	Negative
(VI)	$\bar{A} = 0,\ \bar{B} > 0$[a]	Positive	Negative
(VII)	$\bar{A} = 0,\ \bar{B} = 0$[b]	Zero	Zero

[a] $\bar{A} = 0$ for all cases except (VI) and (VII).

[b] $\bar{B} = 0$ for all cases except (VII).

The seven cases summarized in Table B1 describe the qualitative behavior of the noisy moment-matching log-likelihood depending on the relationship between the mean residual $\bar{A}$ and the mean predicted deviation $\bar{B}$. Each case reflects a characteristic direction of the optimization process in terms of how the model balances residual reduction and predicted variance adjustment. The following interpretations correspond to the derivative signs in each case.

1. **Case (I):** This case arises when the variance predicted by the model is insufficient to cover the observed residuals ($\bar{B} < \gamma \bar{A}$). The optimization tends to increase the predicted variance while reducing the residuals, moving toward better coverage of the data. This situation typically occurs when the model predictions are inaccurate or the observational noise is underestimated.
2. **Case (II):** Here, the model predicts on average the correct level of variance as prescribed by the formulation ($\bar{B} = \gamma \bar{A}$), but the residuals are still larger than ideal. The optimization, therefore, continues to decrease the residuals while maintaining the variance balance.
3. **Case (III):** In this range, the model predicts a variance larger than necessary to explain the observations ($\bar{B} > \gamma \bar{A}$ but smaller than $\left(\gamma + \epsilon^2/(\sigma_N)^2\right)\bar{A}$). Both derivatives are negative, indicating a joint tendency to reduce both the residual and the predicted variance. When the noise level is fixed, further reduction of the total variance may not be feasible.
4. **Case (IV):** This case represents a condition where the residual fitting is optimal under the noisy moment-matching formulation. The equality $\bar{B} = \left(\gamma + \epsilon^2/(\sigma_N)^2\right)\bar{A}$ defines the upper bound of the admissible region. In the limit $\epsilon \to 0$, this condition approaches $\bar{B} = \gamma \bar{A}$, corresponding to the ideal variance balance. For nonzero ϵ, the predicted variation must still decrease slightly to reach the optimum.
5. **Case (V):** This case is the mirror image of Case (I), where the model overestimates the variance relative to the residuals ($\bar{B} > \left(\gamma + \epsilon^2/(\sigma_N)^2\right)\bar{A}$). The optimization response is to increase the residuals while reducing the predicted variance, seeking a better trade-off between the two.
6. **Case (VI):** This is a special instance of Case (V) where the residuals vanish on average ($\bar{A} = 0$) while the predicted variance remains positive. Since $B > 0$ by definition (due to the presence of model noise or predictive uncertainty), the residuals tend to increase until they reach the minimum level compatible with the model's total predicted variance.
7. **Case (VII):** This corresponds to the global maximum of the log-likelihood, achievable only when both the residuals and the predicted variance vanish ($\bar{A} = \bar{B} = 0$). This represents the ideal case in which the model perfectly predicts both the mean and the uncertainty of the observations, a special limiting case of (II) with zero residuals.

In practical scenarios, Cases (VI) and (VII) are rarely attainable due to the inherent presence of model inadequacy and observation noise. The minimum achievable $\bar{A}$ is bounded by the structural limitations of the model, while the minimum $\bar{B}$ is constrained by the noise term σ_N. Upper limits, in contrast, must be prescribed explicitly. If $\min \bar{A} > \min \bar{B}$, the residuals can be fully explained by increasing the predicted model-form uncertainty, and the optimum lies at Case (II), where $\bar{B} = \gamma \min \bar{A}$. Conversely, if $\min \bar{A} < \min \bar{B}$, the optimum is located at Case (IV), where $\bar{A} = \left(\gamma + \frac{\epsilon^2}{(\sigma_N)^2}\right)^{-1} \min \bar{B}$. In this regime, the residuals are forced to increase to match the excessive predicted variance, which is typically undesirable from the standpoint of model fidelity. Such behavior can be avoided by adequately quantifying the observational noise and embedding appropriate parameters to ensure that $\bar{B}$ realistically covers the residuals $\bar{A}$ over their most probable range.

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

C. Results of GMM, RGMM, and IN likelihoods for the thermal experiment

The results for Section 3.2.1 for GMM, RGMM, and IN likelihoods are presented here. It is noted that the fitted model predictions are almost identical for all likelihood models.

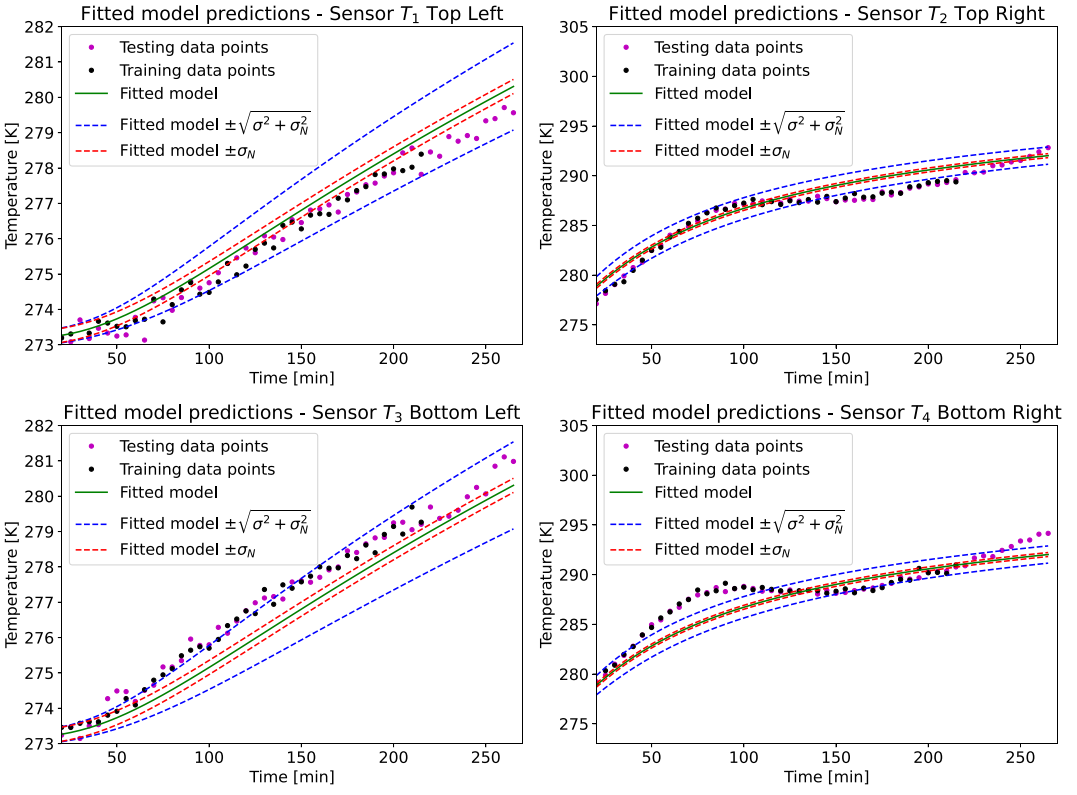


Figure C1. *Temperature sensors' predictions for GMM likelihood.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

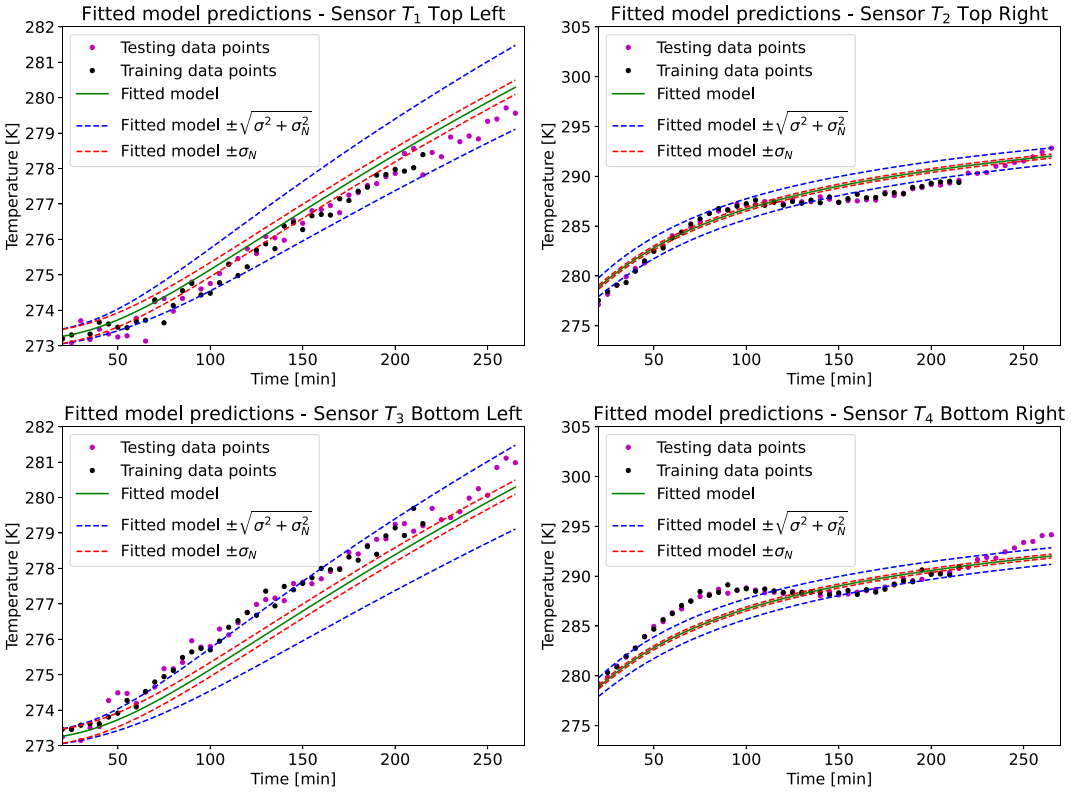


Figure C2. *Temperature sensors' predictions for RGMM likelihood.*

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press

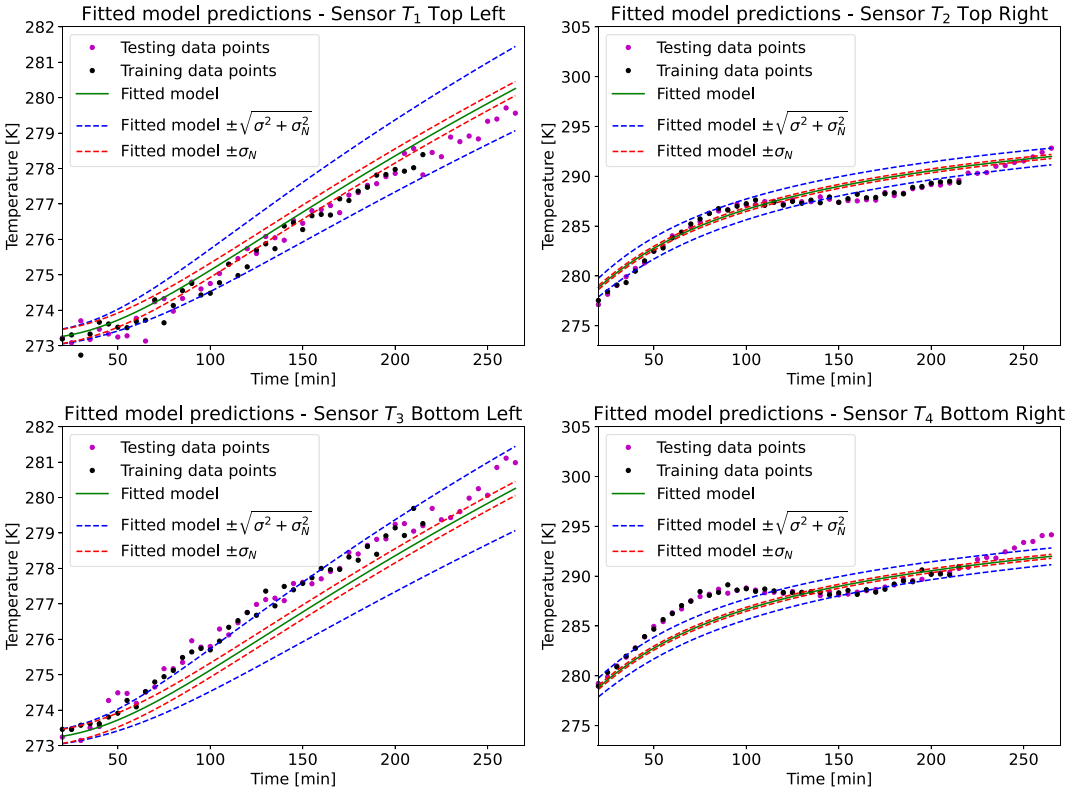


Figure C3. *Temperature sensors predictions for IN likelihood.*

Cite this article: Andrés Arcones D, Weiser M, Koutsourelakis P.-S and Unger JF (2026). Embedded model form uncertainty quantification with measurement noise for Bayesian model calibration. *Data-Centric Engineering*, 7, e7. doi:10.1017/dce.2025.10035

https://doi.org/10.1017/dce.2025.10035 Published online by Cambridge University Press