Structural and Multidisciplinary Optimization (2021) 63:645–660
https://doi.org/10.1007/s00158-020-02715-z

RESEARCH PAPER

A model validation framework based on parameter calibration under aleatory and epistemic uncertainty

Jiexiang Hu[1,2] · Qi Zhou[1] · Austin McKeand[3] · Tingli Xie[3] · Seung-Kyum Choi[3]

Received: 3 January 2020 / Revised: 28 July 2020 / Accepted: 5 August 2020 / Published online: 2 September 2020

Abstract

Model validation methods have been widely used in engineering design to evaluate the accuracy and reliability of simulation models with uncertain inputs. Most of the existing validation methods for aleatory and epistemic uncertainty are based on the Bayesian theorem, which needs a vast number of data to update the posterior distribution of the model parameter. However, when a single simulation is time-consuming, the required simulation cost for the validation of a simulation model may be unaffordable. To overcome this difficulty, a new model validation framework based on parameter calibration under aleatory and epistemic uncertainty is proposed. In the proposed method, a stochastic kriging model is constructed to predict the validity of the candidate simulation model under different uncertainty input parameters. Then, an optimization problem is defined to calibrate the epistemic uncertainty parameters to minimize the discrepancy between the simulation model and the experimental model. K–S test finally decides whether to accept or reject the calibrated simulation model. The performance of the proposed approach is illustrated through a cantilever beam example and a turbine blade validation problem. Results show that the proposed framework can identify the most appropriate parameters to calibrate the simulation model and provide a correct judgment about the validity of the candidate model, which is useful for the validation of simulation models in practical engineering design.

Keywords Model validation · Parameter calibration · Stochastic kriging model · Area metric · Epistemic uncertainty

1 Introduction

Simulation modeling has become an important tool to analyze or predict the behavior of physical systems, especially under some specific scenarios where physical experiments cannot be conducted (often due to limited design cost). However, there are inevitably some differences between the simulation results and experimental observations (Lü et al. 2018), which are generally caused by the uncertainties that exist in the design, manufacture, and experimental process, such as material property, boundary conditions, and machining error (Hu and Mahadevan 2017). To increase the confidence of the simulation model, the validity of the simulation results should be evaluated through the model validation process, which is to determine whether or not a simulation model is an exact representation of the real world within its intended application (Sargent 2010). A simulation model should pass the model validation process before it is further used for design and optimization. Otherwise, it may lead to unreliability system or even the failure of the design.

The uncertainty source in the model validation process can be classified into aleatory uncertainty and epistemic uncertainty. Aleatory uncertainty, also known as stochastic uncertainty, is the inherent variability of an experimental system, such as measurement error and manufacturing errors (Deng et al. 2018). It is often represented by probability theory with random processes or variables (Jiang et al. 2018). When the simulation model only contains aleatory uncertainty parameters,

Responsible Editor: Nam Ho Kim

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00158-020-02715-z) contains supplementary material, which is available to authorized users.

✉ Qi Zhou
qizhouhust@gmail.com

[1] School of Aerospace Engineering, Huazhong University of Science & Technology, Wuhan 430074, People's Republic of China

[2] School of Material Science and Engineering, Huazhong University of Science & Technology, Wuhan 430074, People's Republic of China

[3] George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

the existing validation methods can be generally divided into area metric-based methods (Shen et al. 2015; Xiong et al. 2009; Wang et al. 2018a; Ferson et al. 2008), classical hypothesis testing-based methods (Chen et al. 2004), frequentist's metric-based methods (Moon et al. 2017), Bayes factor-based methods (Rebba and Mahadevan 2006; Sankararaman and Mahadevan 2015; Hu et al. 2018; Lee et al. 2019), and model reliability metric-based methods (Moon et al. 2017; Ao et al. 2017; Rebba and Mahadevan 2008). A comprehensive comparison of the validation methods for aleatory uncertainty can refer to ref. Ling and Mahadevan (2013). Epistemic uncertainty is caused by the lack of knowledge about the physical system, for example, the sparse/imprecise data and model form error. Epistemic uncertainty often exists in the form of sparse data, interval or different type of distributions with uncertain distribution parameters. When validating the simulation model with epistemic uncertainty, Deng et al. (2018) proposed to use the evidence theory to deal with sparse data or interval data while Sankararaman and Mahadevan (2011a) used a likelihood function to describe them. Wang et al. (2018b) developed an interval fitting degree metric to calibrate the distribution of epistemic parameters during the model validation process. Other researches related to the validation for epistemic uncertainty can refer to ref. Deng et al. (2018) and Sankararaman and Mahadevan (2011b).

A more common situation in practical engineering design is that both the aleatory and epistemic uncertainties exist in the input parameters. In this case, the model validation is often accompanied by parameter calibration, namely the epistemic uncertainty parameters are calibrated at first, and then the model validation metric is applied to the calibrated simulation model (Lee et al. 2019). The existing calibration and validation framework can be generally classified into non-Bayesian framework and Bayesian framework. For the first type of validation framework, Youn et al. (2011) proposed a hierarchical model calibration framework, in which the calibration planning (top-down) procedure is utilized to identify and characterize the known and unknown variables, and a calibration execution (bottom-up) procedure is used to calibrate the unknown variables. The goal of the optimization problem used in the calibration execution is maximizing the likelihood function. Jung et al. (2015) further refined the calibration process into three steps, namely model calibration planning, model variable characterization, and model calibration execution. A modified area metric together with the hypothesis testis was also proposed to validate the calibrated model. For the second type of validation framework, Sankararaman and Mahadevan (2015) used the Bayes' theorem to calibrate the model parameter and Bayesian hypothesis testing to validate the updated model. The application of the method was extended to the system with multi-level models. Li and Mahadevan (2016) also proposed a method to quantify the uncertainty from the lower level models to the prediction of the system level. Bayesian inference is used for model calibration and a model reliability metric is utilized to evaluate the validity of the simulation model. Mullins and Mahadevan (2016) described the aleatory uncertainty with a Johnson distribution and used Bayes' theorem in the calibration process. The calibrated simulation model is validated with a modified model reliability metric. Hu and Mahadevan (2017) proposed a Bayesian network to aggregate different sources of uncertainty, an adaptive Bayesian calibration method to reduce the uncertainty, and a Bayesian hypothesis test method in model validation. However, most of the existing validation frameworks generally need to run the simulation multiple times to accurately shape the distribution of the responses, which is further used to check the validity of the simulation model in parameter calibration. If the simulation model is computationally expensive, this process requires a mass of computation cost. What is more, if improperly prior distribution is assigned in the Bayesian validation framework, the parameter calibration process may take a significant amount of iterations to converge (Muehleisen and Bergerson 2016), which also requires a large number of simulations.

To reduce the validation cost, a model validation framework based on parameter calibration (MVBPC) is proposed in this paper. In the proposed method, aleatory uncertainty parameters follow a known form of distributions, and the epistemic uncertainty parameters are expressed as interval parameters, which is a common validation case (Mullins et al. 2016). The epistemic uncertainty parameters are calibrated through an optimization problem with a stochastic kriging model. The calibrated simulation model is finally validated with the K–S test. The performance of the proposed method is illustrated through two engineering examples. Results show that the proposed method can accurately calibrate the candidate model with fewer simulations.

The remainder of the paper is organized as follows. Section 2 gives the background of the stochastic kriging model and model validation metrics used in this study. Section 3 presents the detailed procedures of the proposed method. In Section 4, the proposed method is compared with some validation methods with a cantilever beam validation example and a turbine blade validation problem, followed by a conclusion and future work in Section 5.

2 Technical background

2.1 Area metric

The area metric was proposed by Ferson et al. (2008) to measure the difference of the cumulative distribution functions (CDF) between the simulation responses and experimental observations. The formula of area metric can be described as,

$$d(\boldsymbol{F}_m, \boldsymbol{F}_e) = \int_{-\infty}^{+\infty} |\boldsymbol{F}_m(y) - \boldsymbol{F}_e(y)| dy \quad (1)$$

where $\boldsymbol{F}_m(\cdot)$ is the CDF of the simulation responses, and $\boldsymbol{F}_e(\cdot)$ is the empirical CDF of the experimental observations.

In (1), the integration results reflect the closeness between the two distribution curves. A smaller value indicates less discrepancy between the simulation model and the experimental one. In our method, this metric assesses the discrepancy between the simulation responses under different combinations of uncertainty parameters and the experimental results. The area metric can also validate the simulation model on multiple experimental combinations (or different validation sites) to give an overall assessment of the whole design domain. It incorporates the information from different validation points through the u-pooling procedure, which transforms the responses from physical space to the probability space according to the probability integral transform theorem (Li et al. 2014). For more details of this procedure, readers can refer to ref. Ferson et al. (2008).

2.2 Stochastic kriging

Stochastic kriging is a commonly used interpolation-based model to approximate the relationship between the controllable inputs and the corresponding stochastic simulation responses (Zou and Zhang 2018). Suppose that the design variables are denoted by $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, the prediction of stochastic kriging $\widehat{y}(\boldsymbol{x})$ can be described as,

$$\widehat{y}(\boldsymbol{x}) = f(\boldsymbol{x})^{\boldsymbol{T}}\boldsymbol{\beta} + Z(\boldsymbol{x}) + M(\boldsymbol{x}) \quad (2)$$

where $f(x_i)\boldsymbol{\beta}$ is the regression part, which represents the general trend of the response. $f(\boldsymbol{x})$ is a vector of known basis functions, and $\boldsymbol{\beta}$ is the vector of unknown regression coefficients. $Z(x_i)$ is a realization of zero mean stationary stochastic process. It is often termed as "extrinsic uncertainty" (Ruan et al. 2018). $M(\boldsymbol{x})$ represents the zero mean sampling noise at design point $\boldsymbol{x}$. It is often called "intrinsic uncertainty" (Chen et al. 2013).

The covariance of $Z(\boldsymbol{x})$ between two points can be written as,

$$Cov(Z(x_1), Z(x_2)) = \sigma^2 \boldsymbol{R}(x_1, x_2) \quad (3)$$

where σ^2 is the process variance of $Z(\mathbf{X})$, and $R(\mathrm{x}_1, \mathrm{x}_2)$ is the spatial correlation function, which only depends on the Euclidean distance between two sites x_1 and x_2. In this work, the Gaussian correlation function is adopted,

$$\boldsymbol{R}(x_1, x_2) = \exp\left(-\theta(x_1 - x_2)^2\right) \quad (4)$$

where θ is a roughness parameter to control the variation of the function value with the change of the distance between the two points. The intrinsic variance $M(x_i)$ at a sample point x_i can be calculated by n replications,

$$\boldsymbol{M}(x_i) = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j(x_i) - y(x_i)\right)^2 \quad (5)$$

where $y(x_i) = \frac{1}{n}\sum_{j=1}^{n} y_j(x_i)$, $y_j(x_i)$ is the simulation response of the jth replication.

The prediction of stochastic kriging at an untried point x^* can be written as,

$$\widehat{y}(\boldsymbol{x}^*) = f^T(\boldsymbol{x}^*)\beta + \sum_Z^T(\boldsymbol{x}^*, \cdot)[\sum_Z + \sum_M]^{-1}(\boldsymbol{y} - \boldsymbol{F}\beta) \quad (6)$$

where $\boldsymbol{y} = (y(\boldsymbol{x_1}), \ldots, y(\boldsymbol{x_k}))^T$, $\boldsymbol{F}: (f(\boldsymbol{x_1}), \ldots, f(\boldsymbol{x_k}))^T$. $\sum_Z$ is the covariance matrix of $(Z(\boldsymbol{x_1}), \ldots, Z(\boldsymbol{x_k}))$, $\sum_M$ is the covariance matrix of $(M(\boldsymbol{x_1}), \ldots, M(\boldsymbol{x_k}))$, and $\sum_Z$ is a $k \times 1$ vector whose ith component is $\mathrm{Cov}(Z(\boldsymbol{x^*}), Z(\boldsymbol{x_i}))$.

The estimated mean square error (MSE) is

$$\mathrm{MSE}\left(\widehat{y}(\boldsymbol{x}^*)\right) = \sum_M(\boldsymbol{x}^*, \cdot) + \sum_Z^T(\boldsymbol{x}^*, \cdot)[\sum_Z + \sum_M]^{-1}\sum_M(\boldsymbol{x}^*, \cdot) \quad (7)$$

For more details of stochastic kriging, e.g., the optimization of hyper-parameters, readers can refer to ref. Ankenman et al. (2010), Chen and Kim (2014), and Staum (2009).

2.3 Kolmogorov–Smirnov test

Kolmogorov–Smirnov (K–S) test is a nonparametric hypothesis testing method proposed by Kolmogorov and Smirnov in 1930s. Peacock (1983) proposed a two-dimensional version of it. The two-sample K–S test evaluates the difference between the CDFs of the distributions of the two sample data vectors (Massey Jr 1951); thus, it can be used to check whether two one-dimensional distributions are different from each other. In the two-sample K–S test, the null hypothesis H_0 is that the CDF with a specific distribution can accurately represent the empirical CDF (ECDF) of the given set of statistical data (Gorguluarslan et al. 2017). While the alternative hypothesis is that they are not sampled from the same distribution. The K–S test checks the null hypothesis H_1 by measuring the maximum distance between the CDF curve $F_{Y_m}(y)$ and the ECDF curve $S_{Y_e}(y)$,

$$d_{KS} = \max|F_{Y_m}(y) - S_{Y_e}(y)| \quad (8)$$

Since d_{KS} is a random variable, the CDF of d_{KS} is related to a significance level α as,

$$P\left(d_{KS} \le d_{KS}^{\alpha}\right) = 1 - \alpha \quad (9)$$

For the significance level $1 - \alpha$, the critical value d_{KS}^{α} can be directly obtained from a standard mathematical table for the K–S test. The probability that d_{KS} calculated from the given

sample set smaller than d_{KS}^{α} is defined as the p value, which can be used to test the null hypothesis. The p value can be directly calculated from the Kolmogorov CDF at d_{KS} since d_{KS} follows the Kolmogorov distribution (Marsaglia et al. 2003). If the calculated p value is smaller than α, it means that the null hypothesis H_0 is rejected and the alternative hypothesis is accepted. Otherwise, there is not enough evidence to reject the null hypothesis, and the null hypothesis should be accepted. An advantage of the K–S test is that the size of the two sample sets is explicitly considered when calculating the statistic quantity.

3 Proposed approach

3.1 Description of the parameter calibration problem

To develop a simulation model that can better represent the physical model, the influences of the aleatory and epistemic uncertainty parameters in the model input need to be considered separately (Mullins et al. 2016). Aleatory uncertainty cannot be reduced by including more experimental data. While epistemic uncertainty is caused by a lack of knowledge, it can be calibrated with experimental observations before conducting the model validation. It should be noted that the epistemic uncertainty can be further classified into recognized uncertainty and blind uncertainty according to the ref. Oh et al. (2016). The epistemic uncertainty used in the model calibration process in this paper mainly refers to the first type, such as modeling error caused by the assumptions in the modeling process.

Suppose that the design variables are denoted by $\boldsymbol{x}$, and the aleatory uncertainty input parameters in the experimental model are represented by $\boldsymbol{A}_e$, while the aleatory and epistemic uncertainty input parameters in the candidate simulation mode are denoted by $\boldsymbol{A}_m$ and $\boldsymbol{E}_m$, respectively. The true response $y_t(\boldsymbol{x})$ of an engineering product can be approximated by the experimental results (Oh et al. 2016),

$$y_t(\boldsymbol{x}) = y_e(\boldsymbol{x}) + \varepsilon_e \tag{10}$$

where $y_e(x)$ is the experimental results, and ε_e is the error caused by the aleatory uncertainty parameters $\boldsymbol{A}_e$. The true response can be also approximated by the simulation response by adjusting the epistemic parameters $\boldsymbol{E}_m$,

$$y_t(\boldsymbol{x}) = y_m(\boldsymbol{x}, \boldsymbol{E_m}) + e + \varepsilon_m \tag{11}$$

where $y_m(\boldsymbol{x})$ is the simulation responses, e is the error caused by epistemic uncertainty parameters $\boldsymbol{E}_m$, and ε_m is the error caused by aleatory uncertainty parameters $\boldsymbol{A}_m$. Therefore, the epistemic uncertainty parameters in the simulation model can be calibrated by the experimental results, in which the errors caused by different uncertainty parameters are considered. The parameter calibration problem for epistemic uncertainty in this paper can be formulated as,

$$\begin{aligned} &\min_{\boldsymbol{E}_m} \quad \boldsymbol{C}(y_e(\boldsymbol{x}|\boldsymbol{A}_e), y_m(\boldsymbol{x}|\boldsymbol{A}_m, \boldsymbol{E}_m)) \\ &\text{s.t.} \quad \boldsymbol{A}_e, \boldsymbol{A}_m, \boldsymbol{E}_m \in (\boldsymbol{\Omega}, \boldsymbol{\Psi}, \mathbf{P}) \\ &\qquad \boldsymbol{E}_m = \left\{e_m^1, \ldots, e_m^{ne}\right\}, \quad e_m^i \in U\left(LE_m^i, UE_m^i\right) \end{aligned} \tag{12}$$

where $\mathbf{C}(\cdot)$ is the consistency metric used to reflect the closeness of the simulation responses and experimental results. A smaller metric value indicates a more convincing simulation model. $y_e(\boldsymbol{x}|\boldsymbol{A}_e)$ is the experimental result at the sample point $\boldsymbol{x}$ that is influenced by the aleatory uncertainty, and $y_m(\boldsymbol{x}|\boldsymbol{A}_m, \boldsymbol{E}_m)$ is the simulation response at the sample point $\boldsymbol{x}$ that is affected by both aleatory and epistemic uncertainty. $\boldsymbol{\Omega}$, $\boldsymbol{\Psi}$, and $\mathbf{P}$ are the parameter space of $\boldsymbol{A}_e$, $\boldsymbol{A}_m$, and $\boldsymbol{E}_m$, respectively. $\boldsymbol{LE}_m = \left\{LE_m^1, \ldots, LE_m^{ne}\right\}$ and $\boldsymbol{UE}_m = \left\{UE_m^1, \ldots, UE_m^{ne}\right\}$ are the lower and upper bound of the ne epistemic uncertainty parameters, respectively. Therefore, the value of the objective function in (12) is not deterministic, which leads the optimization problem difficult to be solved. The optimization problem with the uncertainty objective function can be transformed into a deterministic one as suggested in ref. Qian et al. (2016),

$$\begin{aligned} &\min_{\boldsymbol{e}_m} \quad \boldsymbol{C}\left(y_e(\boldsymbol{x}|\boldsymbol{A}_e), y_m\left(\boldsymbol{x}, \boldsymbol{A}_m|\boldsymbol{e}_m\right)\right) \\ &\text{s.t.} \quad \boldsymbol{A}_e, \boldsymbol{A}_m, \boldsymbol{E}_m \in (\boldsymbol{\Omega}, \boldsymbol{\Psi}, \mathbf{P}) \\ &\qquad \boldsymbol{e}_m \in \boldsymbol{E}_m = \left\{e_m^1, \ldots, e_m^{ne}\right\}, \quad e_m^i \in U\left(LE_m^i, UE_m^i\right) \end{aligned} \tag{13}$$

where $\boldsymbol{C}(\cdot)$ is still the transformed objective function used to reflect the consistency of simulation results and experimental results, and $\boldsymbol{e}_m$ is a sample from $\boldsymbol{E}_m$. Thus, the optimization problem is to find the optimal combination of the epistemic uncertainty parameters, which can minimize the discrepancy between the simulation responses and experimental results.

To calculate the value of the objective function in (13), a two-stage nested Monte Carlo simulation (MCS) is utilized to propagate the uncertainty during the optimization. In each iteration, the epistemic uncertainty parameter samples $\boldsymbol{e}_m$ are generated in the outer loop, which is generally implemented by the optimization algorithm. The aleatory uncertainty parameter samples $\boldsymbol{A}_m$ are randomly generated from their distribution intervals in the inner loop. The sample sets $(\boldsymbol{A}_m, \boldsymbol{e}_m)$ are formed to propagate the uncertainty from the inputs to the outputs of the simulation model. The distribution of the response $y_m(\boldsymbol{x}, \boldsymbol{A}_m|\, \boldsymbol{e}_m)$ can be obtained by running the simulation model multiple times. In this way, the consistency between the simulation model and the experimental model under different epistemic uncertainty parameters can be calculated. However, the computation burden of the nested MCS is

obvious because the optimization algorithm generally generates hundreds of or even thousands of epistemic uncertainty parameters for function evaluation before reaching its convergence (An et al. 2018). If a single simulation is computationally expensive, the parameter calibration for the candidate simulation model may be unaffordable.

3.2 Proposed model validation framework

In this section, a new model validation framework based on parameter calibration is proposed to solve the computation-intensive optimization problem with the two-stage nested MCS process. The main idea of the proposed method is to construct a surrogate model to replace the objective function in the optimization, which approximates the relationship between different epistemic uncertainty parameter $\boldsymbol{e}_m$ and the corresponding consistency metric $\boldsymbol{C}(y_e(\boldsymbol{x}, \boldsymbol{A}_e), y_m(\boldsymbol{x}, \boldsymbol{A}_m | \boldsymbol{e}_m))$. There are two advantages of the proposed method over the two-stage nested MCS optimization method, i.e., (1) the nested optimization problem is transformed into a single-stage optimization since the consistency metric under different epistemic uncertainty parameters can be directly predicted. Therefore, the complexity of the optimization problem is reduced and the efficiency is improved. (2) The surrogate model can be constructed by the limited number of sample points, and thus the required simulation cost is greatly reduced.

The proposed parameter calibration process can be formulated as,

$$\begin{aligned} \min_{\boldsymbol{e}_m} \quad & \widehat{\boldsymbol{C}}(\boldsymbol{e}_m) \\ \text{s.t.} \quad & \boldsymbol{e}_m \in \boldsymbol{E}_m = \left\{ e_m^1, \ldots, e_m^{ne} \right\} \\ & e_m^i \in U\left(LE_m^i, UE_m^i\right) \end{aligned} \tag{14}$$

where $\widehat{\boldsymbol{C}}(\boldsymbol{e}_m)$ is the prediction from the stochastic kriging model. To obtain quantitative results about the difference between the distribution functions of the simulation responses and experimental results, the area metric (Ferson et al. 2008) is utilized as the consistency metric in the proposed method, which can also be predicted by the surrogate model.

Even though using the surrogate model can reduce the required simulation cost for solving the optimization problem, it needs to run the simulation model multiple times with the given uncertainty parameters during the uncertainty propagation process. Generally, it requires over 200 samples to obtain a relatively smooth response distribution curve, which is still computationally expensive. To further reduce the simulation cost in the construction of the surrogate model, it is proposed to propagate the uncertainty in the simulation model with a relatively small number of simulations. For example, only 80 simulations are used to obtain the distribution of the simulation responses $y_m(x, \boldsymbol{A}_m | \boldsymbol{e}_m)$ in this study. Then, the bootstrap method (Efron and Tibshirani 1997) is utilized to resample the data set $y_m(x, \boldsymbol{A}_m | \boldsymbol{e}_m)$ to get different sets of simulation responses. The bootstrapped simulation responses are compared with the experimental results to estimate the uncertainty of the consistency metric due to insufficient samples. Stochastic kriging is selected as the surrogate model to achieve higher model accuracy by considering the uncertainty of the consistency metric. The advantage of using stochastic kriging over the deterministic one will be illustrated on a cantilever beam example in Section 4.

After constructing the stochastic kriging model, the minimum value of it can be obtained through solving the optimization problem in (14). The epistemic uncertainty parameters corresponding to the optimal solution can minimize the discrepancy between the simulation model and the experimental model. However, using surrogate model will introduce additional uncertainty (Liu et al. 2016), which may result in a large discrepancy between the calibrated parameters and their actual values. To avoid accepting inaccurate simulation models after validation, the calibrated simulation model runs multiple times to generate the distribution of the responses. The K–S test is applied to finally decide whether to accept or reject the calibrated simulation model.

The flowchart of the proposed method is summarized in Fig. 1. The details of each step are described as follows:

Step 1: Conduct physical experiments and collect the experimental observations $\boldsymbol{y}_e$ under the predefined experiment allocation.

Step 2: Calculated the empirical CDF (ECDF) $\boldsymbol{F}_e$ of the experimental observations $\boldsymbol{y}_e$.

Step 3: Generate samples for the uncertainty parameters. In the computational model, the model parameters with epistemic uncertainty lie in the uncertainty intervals $[\boldsymbol{e}_{\min}, \boldsymbol{e}_{\max}]$, which are generally estimated from empirical data or expert experience. This step generates samples from the uncertainty intervals to construct the stochastic kriging model. To increase the uniformity of the distribution of sample under limited simulation budget, the n initial samples $\boldsymbol{e}_m = \left\{ e_m^1, e_m^2, \ldots, e_m^n \right\}$ can be generated through the uniform sampling (US) (Zhang and Wang 1996) method or Latin hypercube sampling (LHS) method (Helton and Davis 2003).

Step 4: For each sample of $\boldsymbol{E}_m$, run the simulation model multiple times to get the distribution of the responses. When finishing the simulation for all samples of $\boldsymbol{E}_m$, n sets of data $\left\{ \boldsymbol{y}_m^1, \boldsymbol{y}_m^2, \ldots, \boldsymbol{y}_m^n \right\}$ will be obtained.

Step 5: For each data set $\boldsymbol{y}_m^i$ $(i = 1, 2, \ldots, n)$, use the bootstrap method on it for k times. Thus, a total number of $n \times k$ sample sets $\boldsymbol{B}_m^{ij}$ $(i = 1, 2, \ldots, n; j = 1, 2, .., k)$ can be obtained. In this work, $k = 100$ is used for all example points.

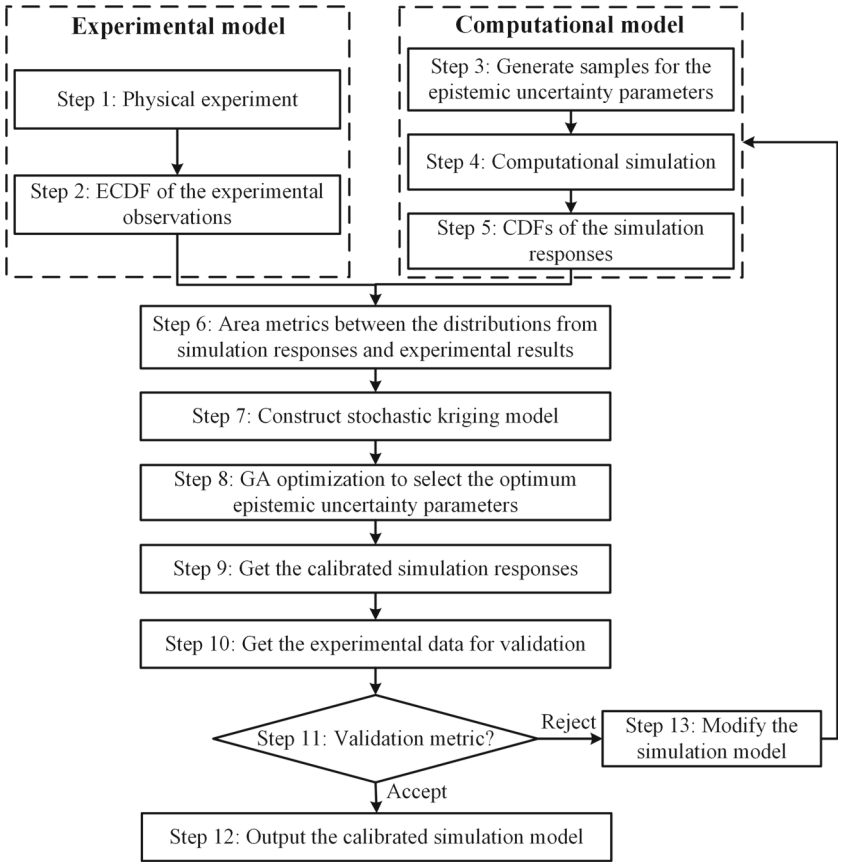


Fig. 1 Flowchart of the proposed parameter calibration and validation framework

Step 5: Calculate the cumulative distribution function for the simulation responses $\boldsymbol{y}_m^i$ and $\boldsymbol{B}_m^{ij}$. The corresponding CDFs are denoted by $\boldsymbol{F}_{my}^i$ and $\boldsymbol{F}_{mb}^{ij}$, respectively.

Step 6: Compare the similarity between the distribution curves of the simulation responses and experimental results. Since the area metric can give a quantified result, it is chosen as the consistency metric in this paper. The area metric can be expressed by the following formula:

$$d = \int_{-\infty}^{+\infty} \left| \boldsymbol{F}_{my} - \boldsymbol{F}_e \right| dy \tag{15}$$

where $\boldsymbol{F}_{my}$ and $\boldsymbol{F}_e$ are the response distribution curves from the simulations and experiments, respectively. The area metric between the simulation responses $\boldsymbol{y}_m^i$ and $\boldsymbol{F}_e$ is denoted by d_{my}^i, and the area metric between the bootstrapped samples $\boldsymbol{y}_m^i$ and $\boldsymbol{B}_m^{ij}$ is represented by d_{mb}^{ij}. The data set d_{mb}^{ij} ($i = 1,2,\ldots,\mathrm{n}$; $j = 1, \ldots, k$) can be reshaped into n data sets $\{d_{mb}^{1j}, \ldots, d_{mb}^{nj}\}$. The variance $\{Var_m^1, \ldots, Var_m^n\}$ of the data sets $\{d_{mb}^{1j}, \ldots, d_{mb}^{nj}\}$ can be calculated.

Step 7: Construct the stochastic kriging model. The stochastic kriging model is used to approximate the relationship between different input parameters and the corresponding area metric values. The stochastic kriging model is constructed with the obtained data set $\left\{ \left(e_m^1, d_{my}^1, Var_m^1\right), \ldots, \left(e_m^n, d_{my}^n, Var_m^n\right) \right\}$. Among them, e_m^i and Var_m^i ($i = 1, 2, \ldots, n$) are the input parameters, while d_{my}^i is the output response.

Step 8: The genetic algorithm (GA) (Davis 1991) is used to select the optimum uncertainty parameter. Specifically, the optimization problem in (14) is solved to find the minimum value of the surrogate model. The optimization problem is a constrained optimization problem, which can be described as,

$$\begin{aligned} &\min_{\boldsymbol{e_m}} \quad \widehat{\boldsymbol{C}}(\boldsymbol{e_m}) \\ &s.t. \quad \boldsymbol{e_m^{\min}} \le \boldsymbol{e_m} \le \boldsymbol{e_m^{\max}} \end{aligned} \tag{16}$$

where $\widehat{\boldsymbol{C}}(\boldsymbol{e_m})$ is the prediction from the stochastic kriging model. The obtained optimum solution e_m^* is utilized as the calibrated parameter to update the simulation model.

Step 9: Run the calibrated simulation model. Update the simulation model with the calibrated parameter e_m^*, and conduct the simulation multiple times to get the responses $\boldsymbol{y}_m^*$. After it, calculate the CDF $\boldsymbol{F}_m^*$ of the responses $\boldsymbol{y}_m^*$.

Step 10: Run the experimental model and get the validation sample set $\boldsymbol{y}_{et}$. The CDF of the validation samples is denoted by $\boldsymbol{F}_{et}$.

Step 11: After obtaining the calibrated simulation model, the validity of it needs to be evaluated through the validation process. The area metric acts as the consistency metric in step 6 and gives a quantified result for constructing the stochastic kriging model. If this metric continues to be used to ultimately decide whether to accept the calibrated simulation model, a threshold value needs to be defined beforehand. However, in practical engineering design, there is often very little information available for the designers to choose this value. An inappropriate threshold may have the risk of accepting inaccurate simulation models. Therefore, the distribution curves $\boldsymbol{F}_{m}^{*}$ and $\boldsymbol{F}_{et}$ are compared through the K–S test in this step. K–S test is a hypothesis-based validation method and can check whether the calibrated simulation responses and the experimental observations are from the same distribution. What is more, the number of experiments is also explicitly considered in the validation of the K–S test.

Step 12: Output the calibrated simulation model. If the null hypothesis cannot be rejected in the K–S test, which means that the distribution of the calibrated simulation responses $\boldsymbol{F}_{m}^{*}$ and the experimental observations $\boldsymbol{F}_{et}$ are very close, the calibrated simulation is accepted and output for future prediction or reliability design.

Step 13: Modify the simulation model. If the simulation model is rejected, which means that adjusting the uncertainty parameter is not enough to match the experiment results, more sample points for constructing the stochastic kriging model is needed. If it still does not work, further modification about the simulation model is required, such as changing the solving equation and adjusting the assumptions in simulation model construction.

Compared with Kennedy and O'Hagan's (KOH) Bayesian calibration method (Kennedy and O'Hagan 2001), no assumption about the distribution of aleatory uncertainty parameters is required in the proposed method. It avoids the problem of selecting an improper prior distribution. It should be noted that the proposed parameter calibration method is only applicable to the case with a single validation site. When applying the proposed method to the case with multiple validation sites, the u-pooling procedure should be added into step 6, which integrates the consistency metric at different validation sites into a single distribution function through the probability integral transform theorem (Li et al. 2014). For more details of this procedure, readers can refer to ref. Ferson et al. (2008). To ensure the extrapolating of the simulation model to application conditions under which experiments may not have been performed, the validation set selected in step 10 should be different from the samples used in the parameter calibration process. The validation data is often chosen from different design sites to given and overall assessment of the calibrated simulation model.

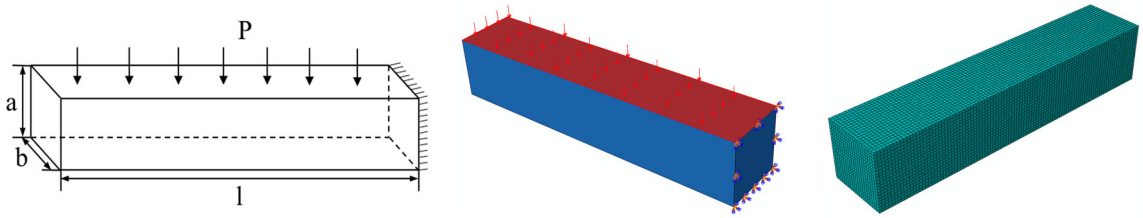
4 Demonstration examples

The performance of the proposed method is illustrated through two examples, including a cantilever beam example and a turbine blade example. The cantilever beam example is utilized to demonstrate the efficiency of the proposed method, while the turbine blade validation problem is used to illustrate the engineering applicability of the proposed method. For comparison, two different parameter calibration methods are also tested. The first one directly solves the optimization problem in (8) with the simulation model and it uses the area metric as the consistency metric, so the first method is denoted by direct optimization with area metric (DOAM) in this work. The second method is proposed by Qian et al. (2016), and it also solves the optimization problem with the simulation model but uses the Mahalanobis distance as the consistency metric, so the second method is denoted by direct optimization with Mahalanobis distance (DOMD).

4.1 Cantilever beam validation example

The proposed method is applied to a cantilever beam validation problem at first. The simulation responses from a high-fidelity simulation model are regarded as the experimental results, which are solved with the Timoshenko beam theory in ABAQUS. The geometry, load and boundary condition, and mesh for the simulation model are plotted in Fig. 2 a, b, and c respectively. One end of the beam is fixed, and the other one is free. The upper surface of the beam is subject to a uniformly distributed pressure $P = 5$ MPa. The height a and width b of the beam are fixed at 10 mm in all simulations. The length of the beam has aleatory uncertainty, which follows normal distribution $l \sim N(40, 0.4^2)$. The Young's modulus and Poisson's ratio are $E_H = 3.5e4$ MPa and $\mu = 0.33$, respectively. The maximum deflection of the beam is the response to be compared during the validation process. In this work, ABAQUS 2018 is chosen as the simulation tool to calculate the maximum deflection. The MATLAB 2017a is used to change the length of the beam in the source file and call the simulations repetitively.

The Euler-Bernoulli beam theory, which is a simplification of the linear theory of elasticity, is the simulation model to be validated in this example. The formula for predicting the maximum deflection of the cantilever can be expressed as (Bauchau and Craig 2009),



(a) Schematic plot of the cantilever beam (b) Load and boundary condition (c) Grid model

Fig. 2 The schematic plot, and simulation details of the cantilever beam. **a** Schematic plot of the cantilever beam. **b** Load and boundary condition. **c** Grid model

$$\delta_{\max} = \frac{PL^4}{8EI} \tag{17}$$

where I is the inertia moment, and $I = ab^3/12$. Compared with the Timoshenko beam theory, the shear deformation effect is ignored in the Euler-Bernoulli beam theory. The geometry parameters a and b and the pressure P in the candidate model are the same as the experimental model. The length of the beam l has the same distribution as in the experimental model. The Young's modulus in the candidate model has epistemic uncertainty, and the uncertainty distribution interval of it is $\boldsymbol{E_L} \sim U$(30000 MPa, 40000 MPa).

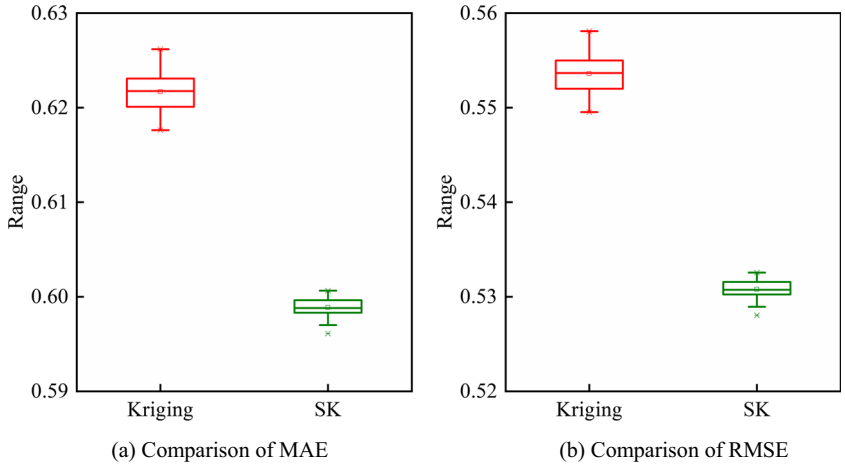
4.1.1 Comparison between stochastic kriging and kriging in model validation

Before validating the simulation model, a comparison is made to illustrate why the stochastic kriging instead of kriging is used in the proposed validation method. Six epistemic uncertainty parameter samples $\boldsymbol{E} = \{e_1, \ldots, e_6\}$ are selected from the uncertainty distribution interval through the Latin hypercube sampling method. The Euler-Bernoulli simulation model is repeated 80 times at each of the six sample points to obtain the distribution of the responses. A total of 30 Timoshenko experimental results are applied to validate the simulation model. Stochastic kriging and kriging are constructed for parameter calibration, respectively. For the stochastic kriging model, the responses of each sample point are bootstrapped 100 times to estimate the uncertainty of the consistency metric due to insufficient samples. Additional 100 test samples are randomly generated from the uncertainty distribution interval $\boldsymbol{E_L}$ to verify the accuracy of the two surrogate models. At each test point, the simulation model is repeated 10,000 times to obtain an accurate distribution of the responses. The CDF curves of the simulation responses and the experimental results are compared with the area metric. The maximum absolute error (MAE) and rooted mean square error (RMSE) are recorded to measure the local and global accuracy of the surrogate model respectively. These two metrics are formulated as,

$$\begin{aligned} \mathrm{MAE} &= \max\left(\left|\widehat{y}_i - y_i\right|\right) \ (i = 1, 2, \ldots, n) \\ \mathrm{RMSE} &= \frac{1}{n}\sqrt{\left(\widehat{y}_i - y_i\right)^2} \end{aligned} \tag{18}$$

where $\widehat{y}_i$ is the prediction from the stochastic kriging model at the test point, y_i is the true response, that is, the area metric value calculated from the 10,000 repetition results and 30 experimental results, and n is the number of test points. This test problem is run 50 times to isolate the influence of the difference between simulation response samples, which is caused by the limited number of simulation repetitions. The box plots of the results are shown in Fig. 3. The top and bottom of the box represent the 25 and 75 percentiles of the results respectively. The line in the middle of the box denotes the median (50 percentile) value of the data. The distances between the top and bottom are the interquartile range. The top and bottom extents of the whiskers are 1.5 times the interquartile range away from the top or bottom of the box. The outliers are caused by different seeds used for the aleatory uncertainty parameters during the uncertainty propagation process.

It is obvious that the median value of the MAE and RMSE from the stochastic kriging model is lower than those from the kriging model, indicating that the stochastic kriging is more accurate than kriging in both local and global performance. What is more, the length of the box of stochastic kriging is smaller than kriging, namely the accuracy of the constructed stochastic kriging is less influence by different sample sets. The difference between kriging and stochastic kriging model can be interpreted that kriging model is an interpolation model, and its performance is easily to be influenced by the noise of the responses. While stochastic kriging model considers not only the mean of the responses but also the variance of them, therefore it can achieve higher model accuracy and provide more robust predictions. To check the convergence of the model



(a) Comparison of MAE (b) Comparison of RMSE

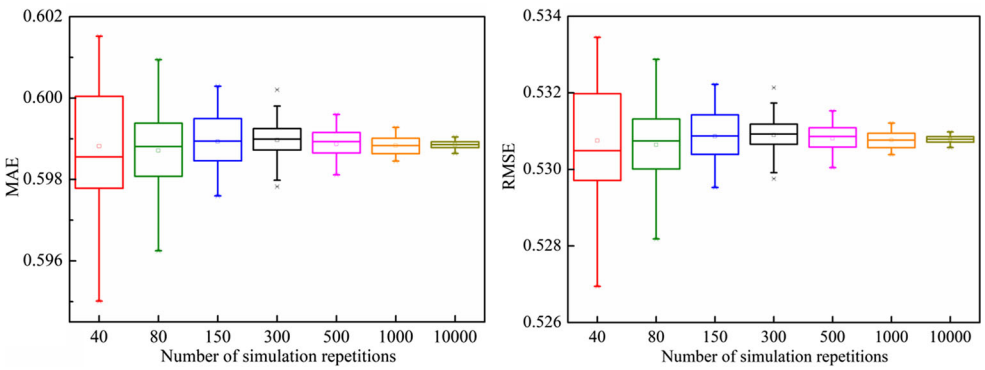
Fig. 3 Performance comparison of kriging and stochastic kriging. **a** Comparison of MAE. **b** Comparison of RMSE

accuracy under the different number of simulation repetitions, seven levels ranging from 40 to 10,000 are set. The model construction process is also repeated 50 times on each level. The comparison results are plotted in Fig. 4. The model construction process is also repeated 50 times on each level. The comparison results are plotted in Fig. 4.

It is obvious that the median values of MAE and RMSE show similar tends, namely they gradually converge to a fixed value as the number of repetitions increases. The variation ranges of MAE and RMSE also reduce with more repetition times. Especially when the repetition time is 1000, the median value of MAE and RMSE can be approximately regarded as the true error of the surrogate model. It can be inferred that the uncertainty of the calculated area metric is reduced with more repetitions, which further enhances the robustness of the predictions from the stochastic kriging model. It should be noted that increasing the number of repetitions cannot directly improve the local or global performance of the surrogate model. To improve the model accuracy, it is required to add more sample points of epistemic uncertainty parameters or adopted some adaptive sampling methods (Qian et al. 2020; Ruan et al. 2020) to allocate the location of sample points more reasonably when building the surrogate model.

4.1.2 Validation with the proposed method

The proposed MVBPC method is applied to check the validity of the simulation model in this section. Firstly, the experimental model is run 30 times to get the experimental observations. Then, 6 samples of the Young's modulus E_L are generated with the LHS from its uncertainty interval, and 80 responses are predicted by (17) for each E_L sample. The bootstrap method with 100 repetitions is applied to the response of every sample point. After comparing each simulation response curve and the experimental results, a stochastic kriging model is constructed, as shown in Fig. 5 a. The variance predicted by the stochastic kriging model (in the 1e-6 order of magnitude) is very small compared with the responses (in the 1e-2 order of magnitude) in this case. Thus, the prediction responses (namely the area metric) and the corresponding prediction variance of 20 test points, which are generated from the design domain, are listed in Table 1. Then, the optimization problem in (16) is utilized to find the minimum value of the constructed surrogate model. The calibrated material property corresponding to the optimal solution is $E^* = 33119.439$ MPa. In the final validation procedure, the calibrated simulation model is run 80 times, while the experimental model is run for extra 20 times to generate the data set for validation. The obtained simulation



(a) Comparison of MAE (b) Comparison of RMSE

Fig. 4 The p-box of area metric under different number of simulation repetitions. **a** Comparison of MAE. **b** Comparison of RMSE

Table 1 The prediction responses and prediction variances of 20 test points

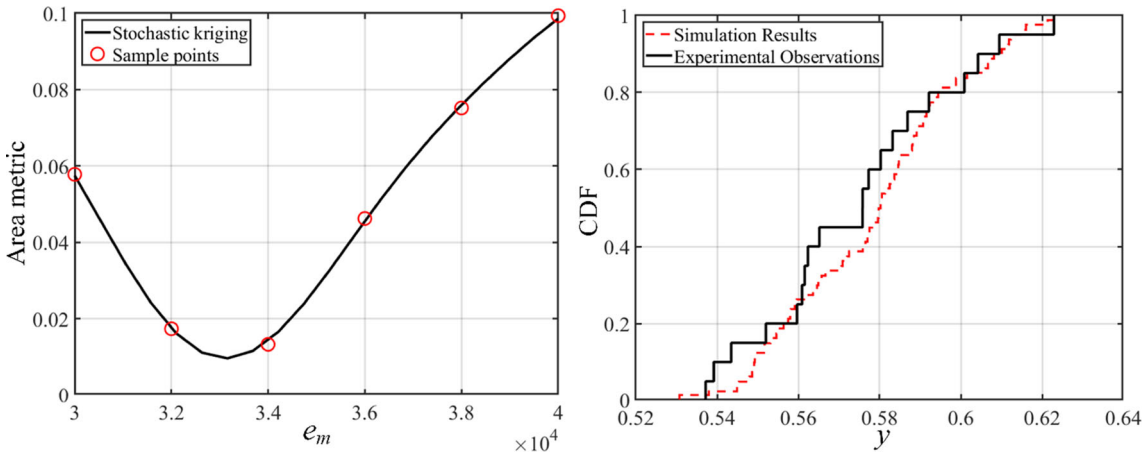
e_m	Response	Variance	e_m	Response	Variance
0.0000	0.0581	7.21E-06	0.5263	0.0338	3.54E-06
0.0526	0.0479	6.58E-06	0.5789	0.0436	3.64E-06
0.1053	0.0368	7.01E-06	0.6316	0.0526	3.90E-06
0.1579	0.0259	5.79E-06	0.6842	0.0606	4.35E-06
0.2105	0.0168	4.57E-06	0.7368	0.0677	4.79E-06
0.2632	0.0107	4.13E-06	0.7895	0.0743	5.47E-06
0.3158	0.0085	3.80E-06	0.8421	0.0810	6.67E-06
0.3684	0.0104	3.32E-06	0.8947	0.0880	7.12E-06
0.4211	0.0159	3.11E-06	0.9474	0.0950	4.96E-06
0.4737	0.0242	3.32E-06	1.0000	0.1017	3.18E-06

responses and the experimental results are plotted in Fig. 5 b. It can be seen that the distribution function of the simulation responses is very close to the experimental observations after parameter calibration. The K–S test is applied to check the consistency of the two distribution functions, and the p value 0.6679 is obtained. Thus, the calibrated simulation model passed the validation, and it can furtherly be used for the design process. Even if the experimental model and the candidate model have the same geometry in this problem, the optimum value for the epistemic uncertainty parameter is not the same as the true value due to different solving theory. The calibration process in the proposed method tries to adjust the epistemic uncertainty parameters to minimize the difference between distribution functions of the simulation responses and the experimental observations, rather than simply approaching the true value in the experimental model.

The two parameter calibration methods, DOMD and DOAM, are also applied to select the optimal value for the epistemic uncertainty parameter. The obtained E^* by DOMD is 32,895.606 MPa, and the obtained p value from the K–S test is 0.4190. Thus, the calibrated simulation model is accepted. While the optimal E^* from DOAM is 33053.813 MPa, the corresponding p value is 0.7540, and the calibrated simulation model is also accepted by the K–S test. The CDF curves of responses from the two calibrated simulation models are plotted in Fig. 6. By adding a parameter calibration process before finally deciding the validity of the simulation model, the response of the simulation model can be transformed into a CDF curve. Since the experimental results are also denoted by a CDF curve, most existing model validation methods, such as area metric or frequentist's metrics, can be directly applied to compare the consistency between the two curves. In addition, the epistemic uncertainty parameters are already calibrated to fixed values minimizing the discrepancy between the simulation model and experimental model, and thus they can be directly used for further design process. No trial and error is required within their uncertainty distribution interval, which also reduces the simulation cost to some extent.

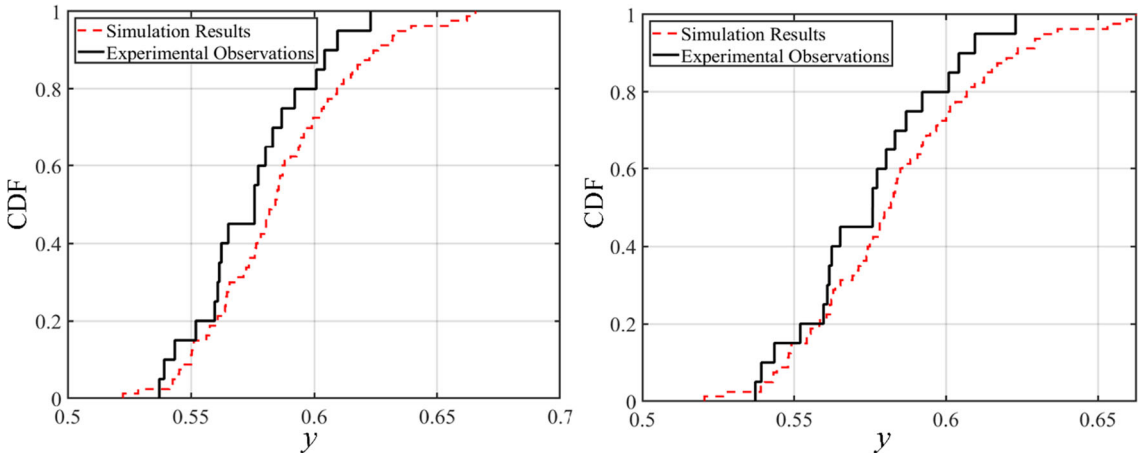
The validation results of the three methods and the corresponding number of simulations used are summarized in Table 2.

From Table 1, it can be seen that (1) the p value of the DOAM method is higher than DOMD methods, and thus the calibrated simulation response with the DOAM method is much closer to the experimental observations. The only difference between these two metrics is the way that is used to check the consistency between the simulation responses and experimental observations: DOMD employs the Mahalanobis distance metric, while DOAM uses the area metric. Therefore, the reason can be inferred that the area metric considers the distribution of the two curves, and calculates the difference between their distribution functions. While



(a) Stochastic kriging model (b) Validation for the calibrated simulation model (MVPBC)

Fig. 5 Validation for a cantilever beam with the proposed method. **a** Stochastic kriging model. **b** Validation for the calibrated simulation model (MVPBC)



(a) Validation for the calibrated simulation model (DOMD)

(b) Validation for the calibrated simulation model (DOAM)

Fig. 6 Validation for a cantilever beam with the DOMD and DOAM method. **a** Validation for the calibrated simulation model (DOMD). **b** Validation for the calibrated simulation model (DOAM)

Mahalanobis distance only measures the covariance distance between the two sample sets, it ignores the distribution information of the response. Thus, the DOAM method is slightly better than the DOMD method. (2) The difference between the proposed MVBPC and DOAM method is whether the surrogate model is used or not. In the proposed MVPBC method, even though using the surrogate model introduces extra uncertainty sources in the validation process, the calibrated simulation model can still pass the K–S test. Furthermore, it saves about 99.96% of the computational cost. In a word, the proposed method can significantly reduce the simulation cost while maintaining high calibration accuracy.

4.2 Turbine blade validation example

Turbine blades are an important component of aircraft engines, and the service life of it directly influences the reliability of the whole engine. However, the uncertainty in the geometry or material properties has a critical influence on the reliability analysis of a turbine blade. It is necessary to consider different sources of uncertainty at the same time when judging the validity of a simulation model. In this section, the engineering applicability of the proposed method is illustrated through a turbine blade model validation problem. Specially, the first natural frequency is selected as the responses to be compared.

A parametric simulation model is selected as the experimental model in this example, and the procedures of generating it are described below. To capture the uncertainty on the geometry that is introduced by the manufacturing process, a turbine blade specimen was manufactured first according to the CAD model in Fig. 7 a. Then, the specimen was scanned through a coordinate measurement machine (CMM) to capture geometry variation. The manufactured specimen and CMM are shown in Fig. 7 b and c. The CAD model of the turbine blade was discretized into 46 line segments, as shown in Fig. 7 d, and about 340 points in each segment were scanned with the CMM. Due to the small errors in the manufacturing process, the coordinate of the measured data differs slightly from the CAD model. The measured points from all segments formed 46 curves, which can be used to describe the geometric variation on the fabricated turbine blade. Based on results from the 46 curves, a Solidworks model of the manufactured turbine blade (denoted as the initial FE model) is generated as in Fig. 7 e.

To determine the material properties of the initial FE model, a vibration test is conducted on the manufactured turbine blade, as shown in Fig. 8 a. The first natural frequency was recorded in the experiment. The previously obtained Solidworks model is imported to ABAQUS, and Young's modulus is adjusted in the simulation to match the result from the experiment result. It was found that when Young's

Table 2 Results comparison for three different parameter calibration methods

	Calibrated parameter E^*	p value	K–S test results	Simulation times (calibration + validation)
MVBPC	33,119.439	0.6679	Accepted	560
DOMD	32,895.606	0.4190	Accepted	2,860,080
DOAM	33,053.813	0.7540	Accepted	1,236,080

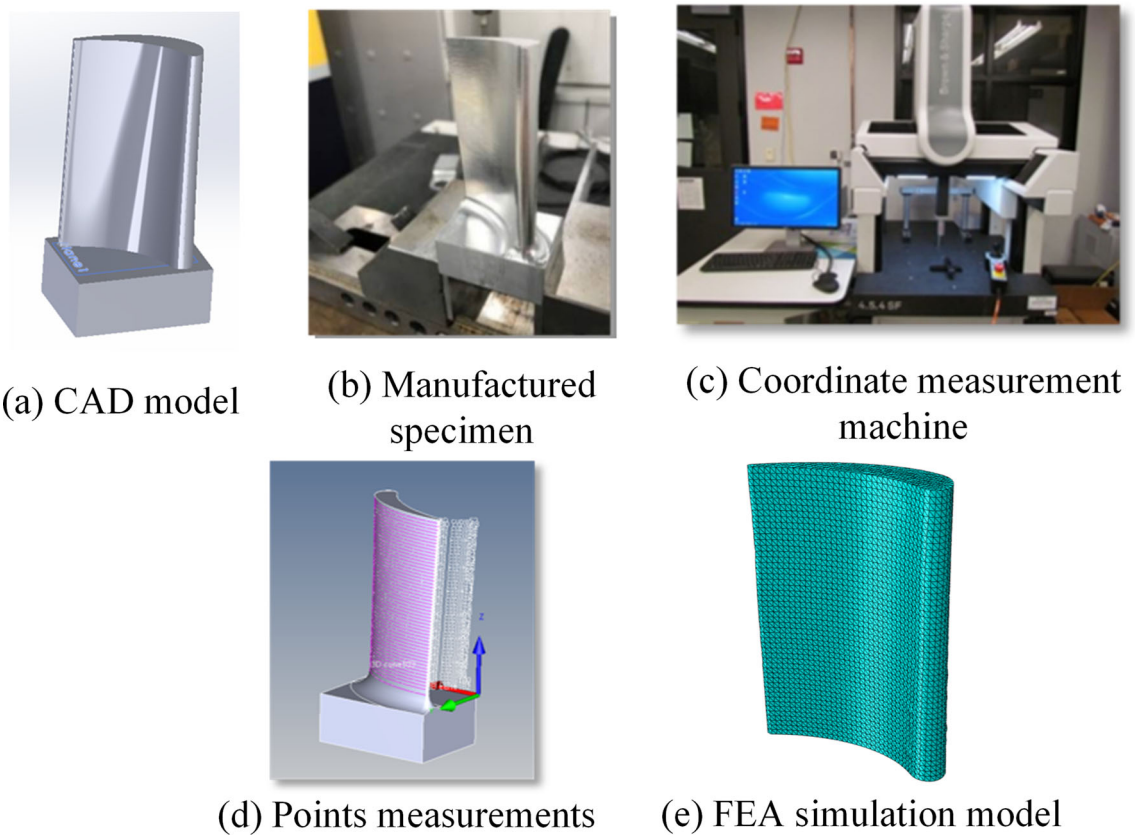
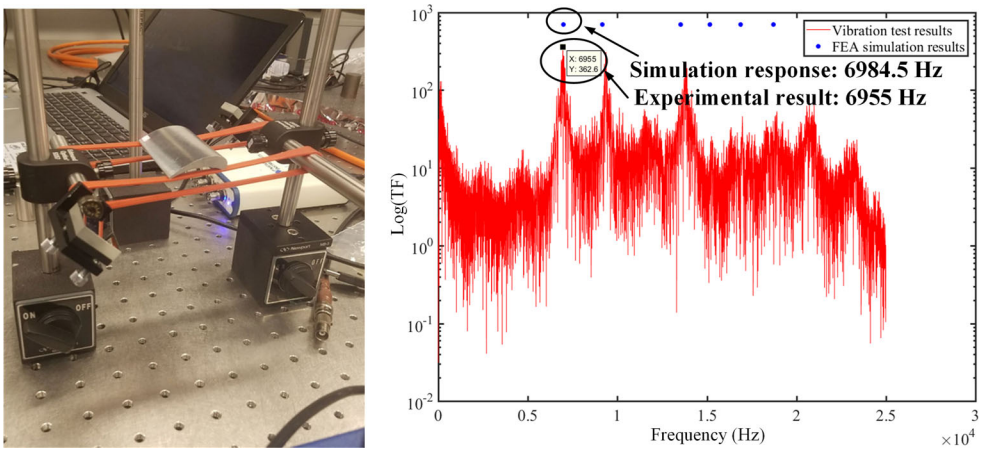


Fig. 7 Procedures for generating the FEA simulation model

modulus of the material is set to $E = 70{,}700$ MPa and Poisson's ratio is set to $\mu = 0.33$, the simulation results are the closest to the experimental results. The comparison result is plotted in Fig. 8 b. The simulation result is 6955 Hz, while the experimental result is 6984.5 Hz. The error is less than 1%. Therefore, the E value in this model is adopted in the following simulations.

To study the spatial dependence of the spatial uncertainties on the surface on the turbine blade, semivariogram analysis is utilized. The coordinates of each point on the original CAD model are the inputs, and the nominal values of the difference between the original CAD data and the measured data are the output. After constructing the semivariogram model, Karhunen-Loeve (KL) expansion is selected to characterize



(a) Vibration test

(b) Comparison between vibration test results and simulation responses

Fig. 8 Vibration test and the results. **a** Vibration test. **b** Comparison between vibration test results and simulation responses

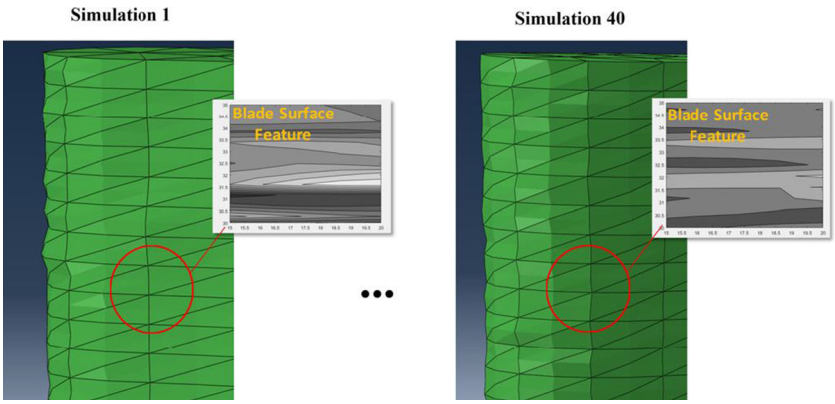


Fig. 9 Example of the geometry variation in the turbine blade FE model

the mathematical model of the spatial variability, which is parameterized by the correlation between different locations. Based on the K-L expansion with a mean of 0 and the correlated matrix that was created from the variogram analysis, a serial of correlated random geometry can be generated. An example of the variations on the surface of the FE model is shown in Fig. 9. The details of the procedures for creating the FE model with random geometry can be found in ref. McKeand et al. (2018). The parametric simulation model with geometry uncertainty (denoted as the fine-scale model) is regarded as the experimental model in this paper. The uncertainty coefficients in K-L expansion, which results in geometric variation on the surface of the model, are the aleatory input parameters of the experimental model. Since the distribution of the aleatory uncertainty of the model cannot simply be described by any standard distributions, the proposed validation method can be used in this example. The previously obtained material properties $E_H = 70{,}700$ MPa and $\mu_H = 0.33$ are applied to this model.

For the candidate simulation model, a coarse-scale model from our previous work (McKeand et al. 2018) is constructed to reduce the computational cost. The candidate model is created with Bezier curves and meshed with the Delaunay Triangulation method. The body of the turbine blade can be divided into 5 sections according to similar variogram functions in the variogram analysis. Thus, it was determined to utilize 6 profiles that are defined by Bezier curves to form the top and bottom profiles of these sections. The coefficients of the Bezier curves have aleatory uncertainty, and the shape of the coarse-scale model can be modified by adjusting the value of these coefficients. To reduce the number of variables in the design, scaling factors on the two directions (except the height of the turbine blade) of the curves are used as the input parameter instead of all of the Bezier curve polynomial coefficients. By changing the scaling factor, the coefficients of the Bezier curves will change automatically. For instance, if the scaling factor Δ is 1, it will generate the same 6 curves as in the initial FE model. When the scaling factor Δ changes to 1.2, the coarse-scale geometry will be 72.8% larger than the volume of the initial FE model. The scaling factor is selected as the aleatory input parameter of the simulation model, and its distribution is obtained from our previous work (McKeand et al. 2018). A comparison of the simulation model with different scaling factors is shown in Fig. 10. The Young's modulus E_L and the the Poisson's ratio μ_L of the coarse-scale model have epistemic uncertainty, and the uncertainty interval for them are E_L~[68,000 *MPa*, 75,000 MPa] and μ_L~[0.3,0.4], respectively. It takes about 5.5 min to run the fine-scale simulation once, while it takes about 22 s to run the coarse-scale simulation once on the computational platform with 3.31 GHz Intel (R) Core(TM) i9-9820X CPU and 64 GB RAM.

In the validation of the candidate simulation model, 40 fine-scale models are generated first based on the K-L expansion. E_H and μ_H are applied to all of the simulation models to get the distribution of the experimental observations. A total of 30 samples (E_L, μ_L) are sampled with the LHS sampling method from their uncertainty interval. The coarse-scale simulation model runs 80 times at each sample point. And the simulation responses at each sample point are bootstrapped 100 times to estimate the

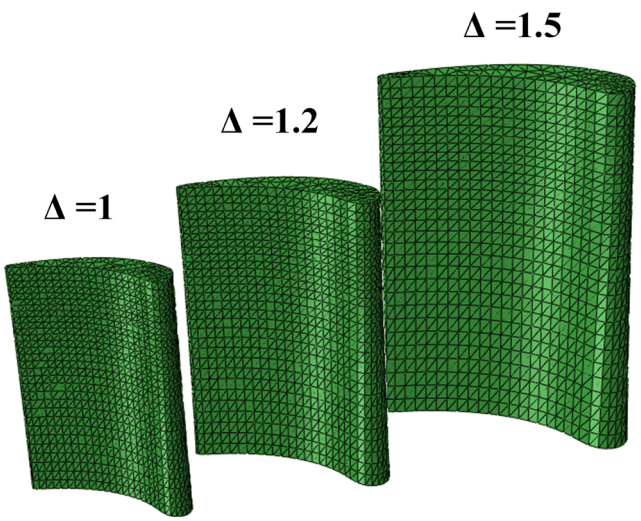
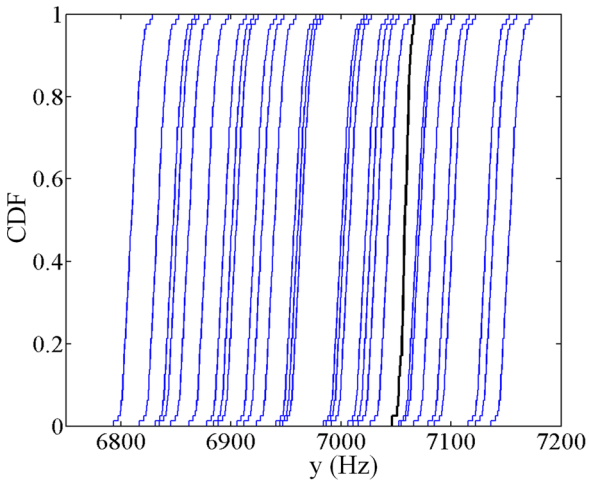


Fig. 10 Example of the candidate simulation model with different scaling factors

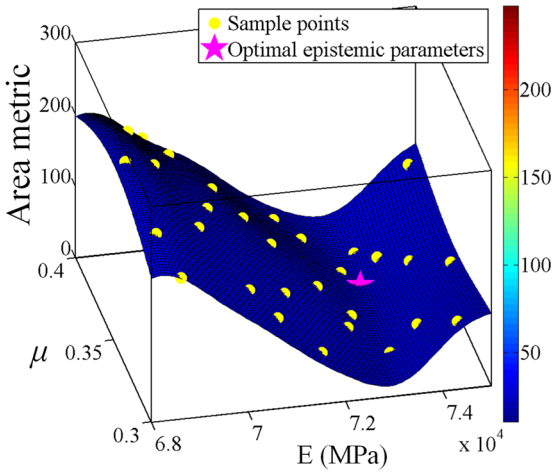
uncertainty of the area metric. The original responses are plotted in Fig. 11 a. It can be seen that the variation range of the experimental observations is much smaller than the range of the simulation responses, and thus the input parameters of the simulation model have large uncertainty. For the proposed validation method, a stochastic kriging model is constructed with the 30 samples, as shown in Fig. 11 b. The optimum value for the coarse-scale model is $E^* = 73272.88$ MPa and $\mu^* = 0.3621$, which are obtained by solving the optimization problem with the constructed surrogate model. Then, the coarse-scale model with the calibrated material property E^* and μ^* is simulated for the additional 200 times. The experimental model is also run for additional 15 times to obtain the experimental observations for validation. The responses from the calibrated simulation model and the experimental model are plotted in Fig. 11 c. The p value in the K–S test is 0.1533, which indicated that the null hypothesis is accepted and the calibrated simulation model finally passes the validation. Therefore, the proposed validation method can accurately calibrate the epistemic uncertainty parameters in the candidate simulation model with relatively less simulation cost.

5 Conclusion

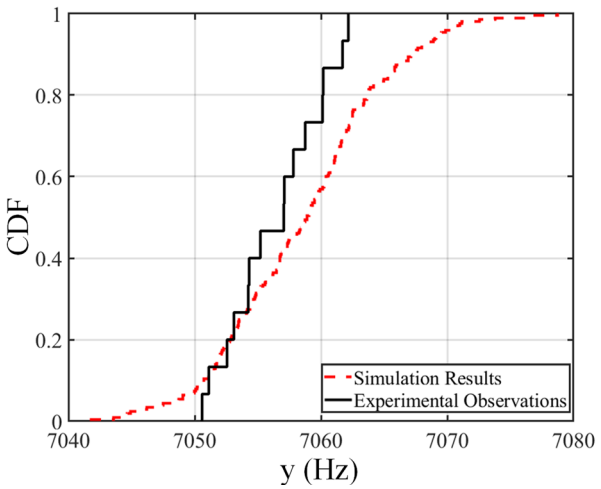
A new model validation method based on parameter calibration under aleatory and epistemic uncertainty is proposed in this paper. In the proposed method, a stochastic kriging model is constructed to reflect the validity of the candidate simulation model under different parameter



(a) Simulation responses and experimental results for validation



(b) Constructed stochastic kriging model



(c) Comparison of the results from the calibrated simulation model and the experimental model

Fig. 11 Parameter calibration and validation results for the turbine blade validation example. **a** Simulation responses and experimental results for validation. **b** Constructed stochastic kriging model. **c** Comparison of the results from the calibrated simulation model and the experimental model

allocations, which greatly reduces the required simulation cost. An optimization problem is utilized to calibrate the value for the parameter with epistemic uncertainty. K–S test is employed to ultimately decide the acceptance of the calibrated simulation model, which considers the number of the samples when calculating the statistic quantity.

The effectiveness and merits of the proposed methods are demonstrated through a cantilever beam example and a turbine blade validation problem. Some desirable merits of the proposed methods over the existing methods are shown from the results. Firstly, the proposed method can accurately calibrate the epistemic uncertainty parameters in the candidate simulation model with relatively less simulation cost. Secondly, increasing the number of sample points improves the accuracy of the surrogate model, while increasing the repetition times at each sample point improves the robustness of the prediction performance. Finally, to prevent the uncertainty of the surrogate model which influences the validation results, the K–S test is applied to test the calibrated simulation model instead of directly verifying the accuracy of the surrogate model. Since using RMSE or MAE metric only indicates the global or maximum error of the surrogate model is within the design limits, it cannot ensure the accuracy at the optimal solution is acceptable.

There are also some limitations of the proposed method. Since constructing an accurate surrogate model in high dimensional cases is difficult, the proposed method is more suitable for the case that the number of the epistemic uncertainty parameter is limited (generally not over four). For the validation problems with a large number of uncertainty parameters, combining the proposed validation framework with dimension reduction methods or hierarchical calibration method as in ref. Youn et al. (2008, 2011) will be investigated as part of our future work. What is more, using the K–S test to validate the calibrated simulation model still has the risk of accepting unreliable simulation models, especially when the design reliability requirements are high, since this metric focuses more on the center of the distribution instead of the tail of the distribution. Therefore, how to choose the validation metrics for the calibrated model according to different design requirements will be also considered for reliability-based design optimization.

Funding information This research has been supported by the National Natural Science Foundation of China (NSFC) under Grant No. 51805179, No. 51775203, and No. 51721092, the National Defense Innovation Program under Grant No. 18-163-00-TS-004-033-01, the research fund under Grant No.61400020401, the Research Funds of the Maritime Defense Technologies Innovation under Grant YT19201901, and the China Scholarship Council with a Scholarship (No. 201706160153).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Replication of results The main step for applying the validation framework has been presented in Section 3. To help readers understand better, the validation code in Section 4.1 has been attached to the supplementary material.

References

An H, Chen S, Huang H (2018) Multi-objective optimization of a composite stiffened panel for hybrid design of stiffener layout and laminate stacking sequence. Struct Multidiscip Optim 57:1411–1426

Ankenman B, Nelson BL, Staum J (2010) Stochastic kriging for simulation metamodeling. Oper Res 58:371–382

Ao D, Hu Z, Mahadevan S (2017) Dynamics model validation using time-domain metrics. Journal of Verification, Validation and Uncertainty Quantification 2:011004

Bauchau O, Craig J (2009) Euler-Bernoulli beam theory. In: Structural analysis. Springer, pp 173–221

Chen X, Kim K-K (2014) Stochastic kriging with biased sample estimates. ACM Transactions on Modeling and Computer Simulation (TOMACS) 24:8

Chen W, Baghdasaryan L, Buranathiti T, Cao J (2004) Model validation via uncertainty propagation and data transformations. AIAA J 42: 1406–1415

Chen X, Ankenman BE, Nelson BL (2013) Enhancing stochastic kriging metamodels with gradient estimators. Oper Res 61:512–528

L. Davis (1991) Handbook of genetic algorithms

Deng W, Lu X, Deng Y (2018) Evidential model validation under epistemic uncertainty. Math Probl Eng 2018:1–11

Efron B, Tibshirani R (1997) Improvements on cross-validation: the 632+ bootstrap method. J Am Stat Assoc 92:548–560

Ferson S, Oberkampf WL, Ginzburg L (2008) Model validation and predictive capability for the thermal challenge problem. Comput Methods Appl Mech Eng 197:2408–2430

Gorguluarslan RM, Choi S-K, Saldana CJ (2017) Uncertainty quantification and validation of 3D lattice scaffolds for computer-aided biomedical applications. J Mech Behav Biomed Mater 71:428–440

Helton JC, Davis FJ (2003) Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliability Engineering & System Safety 81:23–69

Hu Z, Mahadevan S (2017) Uncertainty quantification and management in additive manufacturing: current status, needs, and opportunities. Int J Adv Manuf Technol 93:2855–2874

Hu Z, Mahadevan S, Ao D (2018) Uncertainty aggregation and reduction in structure–material performance prediction. Comput Mech 61: 237–257

Jiang C, Zheng J, Han X (2018) Probability-interval hybrid uncertainty analysis for structures with both aleatory and epistemic uncertainties: a review. Struct Multidiscip Optim 57:2485–2502

Jung BC, Park J, Oh H, Kim J, Youn BD (2015) A framework of model validation and virtual product qualification with limited experimental data based on statistical inference. Struct Multidiscip Optim 51: 573–583

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

Lee G, Kim W, Oh H, Youn BD, Kim NH (2019) Review of statistical model calibration and validation—from the perspective of uncertainty structures. Struct Multidiscip Optim 60:1619–1644

Li C, Mahadevan S (2016) Role of calibration, validation, and relevance in multi-level uncertainty integration. Reliability Engineering & System Safety 148:32–43

Li W, Chen W, Jiang Z, Lu Z, Liu Y (2014) New validation metrics for models with multiple correlated responses. Reliability Eng Syst Saf 127:1–11

Ling Y, Mahadevan S (2013) Quantitative model validation techniques: new insights. Reliability Engineering & System Safety 111:217–231

Liu Y, Shi Y, Zhou Q, Xiu R (2016) A sequential sampling strategy to improve the global fidelity of metamodels in multi-level system design. Struct Multidiscip Optim 53:1295–1313

Lü H, Shangguan W-B, Yu D (2018) A unified method and its application to brake instability analysis involving different types of epistemic uncertainties. Appl Math Model 56:158–171

Marsaglia G, Tsang WW, Wang J (2003) Evaluating Kolmogorov's distribution. J Stat Softw 8:1–4

Massey FJ Jr (1951) The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 46:68–78

McKeand AM, Gorguluarslan RM, Choi S-K (2018) A stochastic approach for performance prediction of aircraft engine components under manufacturing uncertainty. In: ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers Digital Collection, Quebec City, Quebec, Canada

Moon M-Y, Choi K, Cho H, Gaul N, Lamb D, Gorsich D (2017) Reliability-based design optimization using confidence-based model validation for insufficient experimental data. J Mech Des 139: 031404

R.T. Muehleisen, J. Bergerson (2016) Bayesian calibration-what, Why And How

Mullins J, Mahadevan S (2016) Bayesian uncertainty integration for model calibration, validation, and prediction. Journal of Verification, Validation and Uncertainty Quantification 1:011006

Mullins J, Ling Y, Mahadevan S, Sun L, Strachan A (2016) Separation of aleatory and epistemic uncertainty in probabilistic model validation. Reliability Eng Syst Saf 147:49–59

Oh H, Kim J, Son H, Youn BD, Jung BC (2016) A systematic approach for model refinement considering blind and recognized uncertainties in engineered product development. Struct Multidiscip Optim 54: 1527–1541

Peacock JA (1983) Two-dimensional goodness-of-fit testing in astronomy. Mon Not R Astron Soc 202:615–627

Qian X, Li W, Yang M (2016) Two-stage nested optimization-based uncertainty propagation method for model calibration. International Journal of Modeling, Simulation, and Scientific Computing 7:1541003

Qian J, Yi J, Cheng Y, Liu J, Zhou Q (2020) A sequential constraints updating approach for Kriging surrogate model-assisted engineering optimization design problem. Eng Comput 36:993–1009

Rebba R, Mahadevan S (2006) Model predictive capability assessment under uncertainty. AIAA J 44:2376–2384

Rebba R, Mahadevan S (2008) Computational methods for model reliability assessment. Reliability Engineering & System Safety 93: 1197–1207

Ruan X, Zhou Q, Shu L, Hu J, Cao L (2018) Accurate prediction of the weld bead characteristic in laser keyhole welding based on the stochastic Kriging model. Metals 8:486

Ruan X, Jiang P, Zhou Q, Hu J, Shu L (2020) Variable-fidelity probability of improvement method for efficient global optimization of expensive black-box problems. Struct Multidiscip Optim:1–32

Sankararaman S, Mahadevan S (2011a) Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data. Reliability Engineering & System Safety 96:814–824

Sankararaman S, Mahadevan S (2011b) Model validation under epistemic uncertainty. Reliability Engineering & System Safety 96:1232–1241

Sankararaman S, Mahadevan S (2015) Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems. Reliabil Eng Syst Saf 138:194–209

Sargent RG (2010) Verification and validation of simulation models. In: Proceedings of the 2010 Winter Simulation Conference, IEEE, pp 166–183

Shen Z, Chen X, He Q, Zang CP (2015) Study on area metric based upon multiple correlated system response quantities. In: SAE Technical Paper

Staum J (2009) Better simulation metamodeling: the why, what, and how of stochastic kriging. In: Proceedings of the 2009 Winter Simulation Conference (WSC). IEEE, pp 119–133

Wang N, Yao W, Zhao Y, Chen X, Zhang X, Li L (2018a) A new interval area metric for model validation with limited experimental data. J Mech Des 140:061403

Wang C, Matthies HG, Xu M, Li Y (2018b) Epistemic uncertainty-based model validation via interval propagation and parameter calibration. Comput Methods Appl Mech Eng 342:161–176

Xiong Y, Chen W, Tsui K-L, Apley DW (2009) A better understanding of model updating strategies in validating engineering models. Comput Methods Appl Mech Eng 198:1327–1337

Youn BD, Xi Z, Wang P (2008) Eigenvector dimension reduction (EDR) method for sensitivity-free probability analysis. Struct Multidiscip Optim 37:13–28

Youn BD, Jung BC, Xi Z, Kim SB, Lee W (2011) A hierarchical framework for statistical model calibration in engineering product development. Comput Methods Appl Mech Eng 200:1421–1431

Zhang R, Wang Z (1996) Uniform design sampling and its fine properties. Chin J Appl Probab Stat 12:337–347

Zou L, Zhang X (2018) Stochastic kriging for inadequate simulation models. arXiv preprint arXiv:1802.00677