Structural and Multidisciplinary Optimization (2019) 60:1619–1644
https://doi.org/10.1007/s00158-019-02270-2

REVIEW ARTICLE

Review of statistical model calibration and validation—from the perspective of uncertainty structures

Guesuk Lee[1] · Wongon Kim[1] · Hyunseok Oh[2] · Byeng D. Youn[1,3,4] · Nam H. Kim[5]

Received: 8 September 2018 / Revised: 19 March 2019 / Accepted: 22 March 2019 / Published online: 9 May 2019

Abstract
Computer-aided engineering (CAE) is now an essential instrument that aids in engineering decision-making. Statistical model calibration and validation has recently drawn great attention in the engineering community for its applications in practical CAE models. The objective of this paper is to review the state-of-the-art and trends in statistical model calibration and validation, based on the available extensive literature, from the perspective of uncertainty structures. After a brief discussion about uncertainties, this paper examines three problem categories—the forward problem, the inverse problem, and the validation problem—in the context of techniques and applications for statistical model calibration and validation.

Keywords Forward problem · Inverse problem · Validation problem · Uncertainty quantification · Statistical model calibration · Validity check

1 Introduction

With the growth of computing power, computer-aided engineering (CAE) has become an essential instrument for engineering decision-making in various fields of study (e.g., aircrafts, vehicles, electronics, buildings, among others) (Benek et al. 1998; Zhan et al. 2011b; Shi et al. 2012; Fender et al. 2014; Lee and Gard 2014; Silva and Ghisi 2014; Zhu et al. 2016; Jung et al. 2016). However, the credibility of the use of CAE models in real-world applications is a growing concern. To this end, there is an increasing interest in improving and certifying the credibility of computational models (Babuska and Oden 2004; Hills et al. 2008; Oberkampf and Trucano 2008; Kwaśniewski 2009; Roy and Oberkampf 2010; Sargent 2013; Oden et al. 2013; Mousaviraad et al. 2013; Borg et al. 2014; Sankararaman and Mahadevan 2015). For example, the presence of unknown input variables causes credibility concerns in computational model predictions. However, model calibration can improve the credibility of a computational model by estimating the unknown input variables (Kennedy and O'Hagan 2001; Campbell 2006; Higdon et al. 2008; Youn et al. 2011; Zhan et al. 2011b; Arendt et al. 2012b) if sufficient identifiability exists for the unknown input variables (Anderson and Bates 2001). To check the credibility of a computational model, model validation determines the degree to which a model is an accurate representation of the real phenomenon, from the perspective of the model's intended uses (Babuska and Oden 2004; Hills et al. 2008; Oberkampf and Trucano 2008; American Society of Mechanical Engineers 2009). Model validation can be executed at the completion of model calibration to check the predictive robustness of the calibrated model. It should be noted that model validation not only assesses the accuracy of a computational model but also helps the process of improving the model based on the validation results.

Responsible Editor: KK Choi

✉ Hyunseok Oh
hsoh@gist.ac.kr

✉ Byeng D. Youn
bdyoun@snu.ac.kr

[1] Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul 08826, South Korea

[2] School of Mechanical Engineering, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

[3] Department of Mechanical and Aerospace Engineering & Institute of Advanced Machines and Design, Seoul National University, Seoul 08826, South Korea

[4] OnePredict Inc., Seoul 08826, South Korea

[5] Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611, USA

For improving and certifying the credibility of computational models, deterministic approaches have been traditionally adopted, except in the fields of intensive research in statistical model calibration and validation. It may have come from ignorance that does not know the importance of considering various uncertainties or what to do for improving and certifying the credibility of computational models. Although deterministic approaches may be convenient for reducing the disagreement between experimental and computational responses, deterministic approaches not only may incorrectly validate a computational model, but they also may significantly degrade the predictive capability of the computational model (Mousaviraad et al. 2013; Ling and Mahadevan 2013). For this reason, statistical approaches have received significant attention. However, conducting model calibration and validation in a statistical sense is not easy, due to several challenges. Among others, challenges include (1) how to efficiently and accurately conduct uncertainty quantification, (2) how to reduce the degree of uncertainty in the epistemic variables, and (3) how to statistically check the validity of a model. Statistical approaches are beneficial because they attempt to enhance the model's predictive capability by thoroughly addressing uncertainty issues that arise in the experiments and computational models (Oberkampf et al. 2004a; Babuska and Oden 2004; Helton et al. 2004; Oberkampf and Barone 2006; Zhu et al. 2016). Understanding the nature of uncertainties is thus of crucial importance for statistical model calibration and validation.

In general, sources of uncertainties can be defined in three categories: (1) physical, (2) modeling, and (3) statistical (Hills 2006; Urbina et al. 2011; Sankararaman et al. 2011; Zhang et al. 2013; Jung et al. 2014; Zhu et al. 2016). Physical uncertainty arises from the inherent variability in physical quantities. Modeling uncertainty comes from inadequate or erroneous physics models and their numerical implementation and solution. Statistical uncertainty arises from a lack of data associated with uncertainties. In principle, the existence of these uncertainties in engineering systems can be either recognized (accounted for), unrecognized (unaccounted for), or a combination of both. (The terminology "recognized/unrecognized" uncertainty refers to whether the existence of uncertainties is recognized by an engineer, and the terminology "accounted for/unaccounted for" uncertainty refers to whether the uncertainty is addressed by an engineering activity.) Numerous studies have attempted to effectively incorporate various aspects of uncertainty in statistical model calibration and validation (Oberkampf et al. 2002; Chen et al. 2004; Helton and Davis 2003; Ferson et al. 2004; Bayarri et al. 2007). To address the entire spectrum of statistical model calibration and validation problems, this paper presents a preliminary overview of the uncertainty structure. From the perspective of the uncertainty structure, the paper examines three problems in statistical model calibration and validation, specifically, the (1) forward, (2) inverse, and (3) validation problems.

Several review papers related to statistical model calibration and validation can be found in the literature. Kutluay and Winner (2014) presented a literature survey on philosophical model validation concepts and different model validation techniques used in vehicle dynamics simulation models. Earlier, Oberkampf et al. (2004b) presented a state-of-the-art conceptual framework for verification and validation, including code verification, software quality assurance, numerical error estimation, hierarchical experiments for validation, and validation metrics. Later, Trucano et al. (2006) clarified the terminology used in calibration, validation, and sensitivity analysis. Significant new findings in statistical model calibration and validation have been reported since these three prior review papers appeared in the literature. Also, there are a number of new contributions to the literature that have been developed from other societies, such as reliability assessment and reliability-based design optimization, that are worth reviewing in the context of statistical model calibration and validation. Therefore, this paper summarizes the previous literature related to achieving successful statistical model calibration and validation in conjunction with uncertainties. For a systematic review, this paper presents an uncertainty structure for formulating three problems in statistical model calibration and validation. The review includes not only the literature directly related to statistical model calibration and validation but also other related papers deemed necessary for statistical model calibration and validation.

The remainder of the paper is organized as follows. Section 2 gives an overview of uncertainty structure and the framework for statistical model calibration and validation. Section 3 presents the forward problem in the presence of recognized (accounted for) uncertainty. It consists of the uncertainty characterization and the uncertainty propagation process. The inverse problem, in the presence of recognized (accounted for) epistemic uncertainty, is described in Section 4. This review also revisits relevant papers related to model calibration. Section 5 summarizes the validation problem in the presence of unrecognized (unaccounted for) uncertainty and discusses validation metrics for evaluation of model validity. The review paper concludes with suggestions for future work in Section 6.

2 Overview of uncertainty and statistical model calibration and validation

The objective of Section 2 is to provide an overview of uncertainty structure and statistical model calibration and validation. Section 2.1 describes system responses from observation and prediction in a probabilistic sense. The structure of uncertainty is detailed in Section 2.2 from the model calibration and validation point of view. Conceptual definitions and details of three problems—forward, inverse, and validation problem, are presented in Section 2.3, in conjunction with the uncertainty structure.

2.1 Overview of a probabilistic description of system responses

The uncertain nature of randomly varying experimental observation ($\mathbf{Y}_{\mathrm{obs}}$) is caused by various sources of physical uncertainties ($\mathbf{X}$), including material properties, manufacturing tolerances, applied loadings, and boundary conditions, among other factors (Oberkampf et al. 2004a; Buranathiti et al. 2006; Xie et al. 2007; Jung et al. 2011; Jung et al. 2016). In addition to physical uncertainty, errors in measurement (ε) also cause variability in experimental observation (Hills 2006; Harmel et al. 2010; da Silva Hack and Schwengber ten Caten 2012; Ling and Mahadevan 2013; Uusitalo et al. 2015). Measurement errors can be divided into two components: random errors and systematic errors (Ferson and Ginzburg 1996; Liang and Mahadevan 2011; Kim et al. 2018). Random errors are errors in measurement that lead to the measured values being inconsistent when repeated. Often, random errors are neither predictable nor correctable. Systematic errors cause measured values to vary from a true value in a consistent/highly correlated manner from test to test; these errors can often be identified and corrected. An observed response can be expressed in a probabilistic form as

$$\mathbf{Y}_{\mathrm{obs}}(\mathbf{X},\varepsilon)=\mathbf{Y}(\mathbf{X})+\varepsilon \tag{1}$$

where $\mathbf{Y}_{\mathrm{obs}}$ represents all observed system responses from experiments, $\mathbf{Y}$ without any subscript represents the true system responses, $\mathbf{X}$ represents system variables subject to uncertainty, and ε represents a measurement error, which can have both random and systematic components.

It is assumed that simulation is inherently deterministic; when a single set of deterministic inputs is given, a deterministic response is obtained. To compute uncertainty in predicted responses ($\mathbf{Y}_{\mathrm{pre}}$), the uncertainty quantification has been developed by incorporating the physical uncertainties ($\mathbf{X}$) into a computational model (Bae et al. 2003; Pettit 2004; Eldred et al. 2011; Roy and Oberkampf 2011; Mousaviraad et al. 2013). Nonetheless, computational models suffer from the effect of model form uncertainty (e), which embodies the errors from improper assumptions and discretization related numerical solution convergence errors (Sankararaman and Mahadevan 2011b; Roy and Oberkampf 2011; Thacker and Paez 2014) (Voyles and Roy 2015). Generally, depending on the level of expertise of the model developers, model form uncertainty is unrecognized or imprecisely recognized (unaccounted for). Using the model form uncertainty, a predicted response model can be presented in a probabilistic form as

$$\mathbf{Y}_{\mathrm{pre}}(\mathbf{X},\boldsymbol{e})=\mathbf{Y}(\mathbf{X})+\boldsymbol{e} \tag{2}$$

where e represents the model form error and $\mathbf{Y}_{\mathrm{pre}}$ represents all predicted system responses from simulations.

When characterizing physical and modeling uncertainties, statistical uncertainty arises when related data are insufficient. A certain level of data sufficiency can effectively eliminate any *epistemic* uncertainty associated with the characterization of the *aleatory* uncertainty (Oberkampf et al. 2004a; Choi et al. 2010a; Eldred et al. 2011; Urbina et al. 2011; Sankararaman and Mahadevan 2013). Given the sources of uncertainty (i.e., physical, modeling, and statistical) in a computational model, uncertainty structures must be well understood to properly formulate statistical model calibration and validation problems. Thus, the following section discusses the structure of uncertainty for statistical model calibration and validation.

2.2 Uncertainty structure for statistical model calibration and validation

A proper understanding of uncertainty structure is of paramount importance for statistical model calibration and validation. Uncertainties can be classified into recognized and unrecognized uncertainty (Ferson et al. 2004; Oberkampf et al. 2004b; Oliver et al. 2015; Oh et al. 2016), depending on whether the source of uncertainty is recognized or not. Recognized uncertainty, also known as acknowledged uncertainty, comes from conscious investigations by model analysts. Recognized uncertainty is also known as acknowledged uncertainty; it is uncertainty for which a conscious decision has been made to either characterize or deal with it in some way. For instance, when determining the complexity of the physics used to emulate phenomena, modelers often ignore some phenomena and choose a simpler option. Modelers often have good reasons to ignore some phenomena. Unrecognized uncertainty originates from not having the knowledge needed to accurately construct a computational model. In this case, modelers could inadvertently make wrong assumptions in constructing the model. For example, the computational model of a rubber bushing component should be developed with a hyper-elasticity model; however, it could be mistakenly developed using a linear elastic model, leading to unrecognized uncertainty. The predictive capability of the model can be degraded if unrecognized uncertainty is not properly addressed. (The terminology "accounted for/unaccounted for" uncertainty refers to the uncertainties that are recognized but are ignored and not addressed by a proper engineering activity.)

Recognized (accounted for) uncertainties are further classified into aleatory and epistemic uncertainties, depending on whether the uncertainty can be reduced through additional data or information (Parry 1996; Winkler 1996; Youn and Wang 2008; Urbina et al. 2011; Sankararaman and Mahadevan 2011b; Sankararaman and Mahadevan 2011a; Thacker and Paez 2014; Sankararaman et al. 2013; Shah et al. 2015; Liang et al. 2015). Aleatory uncertainty, also referred to as irreducible uncertainty or type A uncertainty, is used to describe the inherent variability associated with a physical system (e.g., material properties) or environment (e.g., operating conditions, manufacturing tolerances) under consideration. On the other hand, epistemic uncertainty, also referred to as reducible uncertainty or type-B uncertainty, is the uncertainty that arises from a lack of knowledge. For example, epistemic uncertainty that obscures the true random or aleatory type A uncertainty can arise when a model input that has considerable degree of uncertainty has a small sample size.

Studies on aleatory uncertainty have focused on how to effectively quantify uncertainty in the form of statistical distributions. Efforts to represent epistemic uncertainty have been developed in two different branches. The first branch starts from the notion that the variable has a true value, but because of the lack of information, the true value is unknown. Since the true value is deterministic in nature, the variable cannot be represented in a statistical distribution. Interval analysis and fuzzy set theory (Moore and Lodwick 2003; Bae et al. 2004; Helton and Oberkampf 2004; Bae et al. 2006; Jiang and Mahadevan 2008; Park et al. 2010; Hanss and Turrin 2010; Liang et al. 2015) belong to this branch. The second branch starts with the idea that the probability distribution in epistemic uncertainty does not represent randomness of a variable, but instead, it represents the shape of knowledge regarding the variable. For example, a uniform distribution can be used if there is no knowledge of the true value. In this approach, well-developed probability theories can easily be adopted to represent the epistemic uncertainty. The Bayesian method (Kennedy and O'Hagan 2001; Jiang and Mahadevan 2007; Babuška et al. 2008; Higdon et al. 2008; Jiang and Mahadevan 2009a; Zhan et al. 2011c; Park and Grandhi 2014) is one of the leading methods of this approach. Both approaches have pros and cons, and this paper considers both approaches as they relate to statistical model calibration and validation.

Epistemic uncertainty is challenging to address by any method. In this paper, system variables ($\mathbf{X}$) of a computational model subject to aleatory and epistemic uncertainties are denoted as $\mathbf{X}_{\mathrm{a}}$ and $\mathbf{X}_{\mathrm{e}}$ respectively.

2.3 Key problems in statistical model calibration and validation

Based on an understanding of uncertainty structures, statistical model calibration and validation can be formulated in three problem categories: (1) the forward problem in the presence of recognized (accounted for) uncertainty (described in Section 2.3.1), (2) the inverse problem in the presence of recognized (accounted for) epistemic uncertainty (see Section 2.3.2), and (3) the validation problem in the presence of unrecognized (unaccounted for) epistemic uncertainty (Section 2.3.3).

2.3.1 Forward problem in the presence of recognized uncertainty

System responses are random due to various sources of uncertainty. However, system responses from computational prediction are inherently deterministic. Uncertainty propagation is thus incorporated into a computational model for uncertainty description of system responses (Helton and Davis 2003; Bae et al. 2003; Pettit 2004; Najm 2009; Eldred et al. 2011; Roy and Oberkampf 2011; Mousaviraad et al. 2013). When all uncertainties of the inputs of a computational model are well recognized, uncertainty propagation can be conducted to estimate the uncertainty of system responses. Uncertainty propagation is defined as a forward problem and is described in the following equation:

$$\mathbf{Y}(\mathbf{X}) \approx \hat{\mathbf{Y}}(\mathbf{X}) = \mathbf{Y}_{\mathrm{pre}}(\mathbf{X}_{\mathbf{a}}) \tag{3}$$

where the caret symbol in $\hat{\mathbf{Y}}(\mathbf{X})$ represents that the predicted system responses may vary from the true system response ($\mathbf{Y}$) in situations where not all variables are accurately characterized with probability distributions and where there exist potential sources of model form errors (e).

Recognized (accounted for) uncertainty should be identified and quantitatively modeled with care, based on the idea of the uncertainty structure discussed in Section 2.2. Probability distribution is a useful way to describe the inherent variability of aleatory uncertainty. When available data are limited, characterizing corresponding uncertainty sources through probability distribution may introduce statistical uncertainty (Moon et al. 2017, 2018). If not ignorable, these sources of statistical uncertainty are thus treated as recognized (accounted for) epistemic uncertainty. A critical review of the forward problem for statistical model calibration and validation is presented in Section 3.

2.3.2 Inverse problem in the presence of recognized epistemic uncertainty

Among recognized (accounted for) uncertainties, often, some are found to be epistemic ($\mathbf{X}_{\mathrm{e}}$) in a computational model because of insufficient information. In situations when epistemic uncertainty is believed to considerably affect the prediction of system responses, uncertainty reduction in epistemic variables is a prerequisite for building a valid computational model. One basic way to achieve uncertainty reduction in epistemic variables is to acquire additional data from experiments. This approach is the most reliable, but it is time consuming and expensive. Another way is the bias correction approach, which defines the inherent difference between experimental observations and computational predictions by a degree of bias (Jiang et al. 2013b; Xi et al. 2013).

A better option is to formulate an inverse problem to identify the epistemic uncertainties. (5), which is the inverse of (4), represents the inverse problem.

$$\mathbf{Y}_{\mathrm{pre}}\left(\hat{\mathbf{X}}_{\mathrm{e}}, \mathbf{X}_{\mathrm{a}}\right) = \mathbf{Y}_{\mathrm{obs}} \tag{4}$$

$$\left[\mathbf{X}_a, \hat{\mathbf{X}}_e\right] = \mathbf{Y}_{\mathrm{pre}}^{-1}(\mathbf{Y}_{\mathrm{obs}}) \tag{5}$$

Formulating the mathematical expression of the inverse problem is based on two assumptions. First, experiments are presumed to be conducted with care, thereby the measurement error ε is negligible. Second, $\mathbf{Y}_{\mathrm{obs}}$ can be considered as a reference, and then the predicted system response ($\mathbf{Y}_{\mathrm{pre}}$) with calibrated epistemic variables $\left(\hat{\mathbf{X}}_e\right)$ should be equivalent to the observed system response ($\mathbf{Y}_{\mathrm{obs}}$). Further review and discussion of the inverse problem are found in Section 4.

2.3.3 Validation problem in the presence of unrecognized uncertainty

If the inverse problem can successfully reduce the uncertainty of recognized (accounted for) epistemic variables in a computational model, it is not necessarily valid for its intended use unless the intended use is at the calibration scenario conditions. Calibration can hide the existence of model-form error if no disagreement exists with the calibration data at the calibration point, so it is not unusual for the model validity check to fail. This usually happens when unrecognized (unaccounted for) uncertainty still exists in the model and the effect of the uncertainty is not ignorable. Examples include misconceptions related to model idealization and assumptions, unrecognized mistakes in mathematical modeling, and ignorance of key physical uncertainties (see (2)) (Trucano et al. 2006; Farajpour and Atamturktur 2012; Atamturktur et al. 2014). The existence of unrecognized (unaccounted for) uncertainty can lead to imprecise or biased prediction of system responses from a computational model. As a consequence of the error due to unrecognized (unaccounted for) uncertainty, the inverse problem in (5) wrongfully calibrates the epistemic variables $\left(\hat{\mathbf{X}}_e\right)$, and thus, the prediction capability of the model deteriorates at untested points. To eliminate the effect of unrecognized (unaccounted for) uncertainty, it must be carefully identified and addressed by refining the computational model. Existence of unrecognized (unaccounted for) uncertainty can be revealed through a validation problem.

The validation problem consists of two processes. First, the system response data set, $\mathbf{Y}_{\mathrm{pre}}$, predicted from computer simulations, is statistically compared to the one from the physical experiments, $\mathbf{Y}_{\mathrm{obs}}$. The comparison is deemed statistically significant if the relationship between the two data sets would be an unlikely realization of the null hypothesis (H_0) according to a threshold probability—the significance level. Second, a decision is made whether or not the prediction of the computational model is acceptable based on the evaluation of a validation metric, which is used as a hypothesis test statistic in the validation problem. The validation metric quantitatively evaluates the statistical difference between the prediction ($\mathbf{Y}_{\mathrm{pre}}$) from computer simulations and the observations ($\mathbf{Y}_{\mathrm{obs}}$) from physical experiments. The hypothesis test is widely used as a decision-making tool for validation problems (Chen et al. 2004; Kokkolaras et al. 2013; Jung et al. 2014).

When the null hypothesis is rejected, sources of unrecognized (unaccounted for) uncertainty still exist and are statistically significant. Potential sources of unrecognized (unaccounted for) uncertainties include model form uncertainty (e), errors in measurement (ε), and erroneous results $\left(\hat{\mathbf{X}}_e\right)$ in the inverse problem. If the null hypothesis is accepted, the computational model is acceptable for its intended use. Section 5 addresses the validation problem.

3 Review of the forward problem

Section 3.1 begins with a discussion of contemporary issues related to the forward problem in conjunction with recognized (accounted for) uncertainty. After that, Sections. 3.2, 3.3, and 3.4 review issues of uncertainty characterization, variable screening, and uncertainty propagation, respectively. (Note that this paper considers the forward problem as a propagation of scalar data, though the propagation of random function and field data are an important issue. The brief discussion of random function and field data are included in Section 6.)

3.1 Contemporary issues of the forward problem

The first step of the forward problem is to identify and characterize all sources of input uncertainties, which can affect the prediction results. In statistical model calibration and validation, the input uncertainties are represented using different types of probability density functions (PDFs). The forward problem can be solved after confirming that the PDFs are sufficient to represent the uncertainties in the populated data. The most frequently addressed issues in uncertainty characterization are (1) how to decide the appropriate types of PDFs to represent the uncertainty of an input, (2) how to handle uncertainty with a limited or nonexistent data to build a PDF for an input, and (3) how to deal with statistical dependency among input uncertainties. The details related to the identification and characterization of uncertainty are covered in Section 3.2.

The second step of the forward problem is a process called variable screening. Variable screening determines which inputs exhibit the most effect on the system responses. Model calibration and validation with a large number of inputs can be conducted more efficiently by focusing on a small number of important inputs. The crucial issue to be considered is that selecting important inputs should be based on uncertainty analysis. Variance-based methods are recommended among various techniques. Several points should be addressed to implement variance-based methods, including (1) how to rank the effectiveness of inputs on system response and (2) how to make good use of uncertainty characterization and propagation techniques. The details related to variable screening are covered in Section 3.3.

The last step of the forward problem is uncertainty propagation, which examines how to effectively understand the effect of input uncertainties on the uncertainty in the system responses. Extensive studies have been conducted and different methods for uncertainty propagation have been discovered. Each category exhibited different computational performances (i.e., efficiency, accuracy, and convergence) depending on the nature of the computational model. One of the major issues is selecting the most suitable method for uncertainty propagation, which requires a study of the tradeoffs between accuracy and computational expense. Other issues in uncertainty propagation include (1) how to effectively fulfill uncertainty quantification, (2) how to account for statistical correlation among input uncertainties, (3) what statistical models are suitable for characterizing system responses, and (4) how to manage uncertainty in highly nonlinear system responses (e.g., multimodal probability density functions).

3.2 Uncertainty characterization

3.2.1 Methods of uncertainty characterization

The most effective way to quantify the uncertainties in an engineered system is to represent the uncertainty in the form of a probability density function (PDF). It is common for a PDF to be parameterized or characterized by statistical parameters, for example, the mean and the variance of a normal distribution (parametric methods). Various types of probability density functions (PDFs) can be applied to express uncertainty. In practice, however, some distributions (e.g., multimodal, mixed distribution) do not follow a parameterized PDF. Unlike parametric methods, nonparametric methods make no assumptions about the PDFs of the inputs. Since nonparametric methods are not involved in population parameters, these can be advantageous when characterizing various shapes of a PDF.

Parametric methods consist of (1) selecting a proper type of PDF to describe the uncertainty in inputs and (2) estimating values for statistical parameters of the selected PDF. First, to decide an appropriate type of PDF, graphical methods and goodness-of-fit (GOF) tests can be used. Graphical methods, such as probability plots, provide strong indications of a proper probability distribution function for the input data. GOF tests verify an assumed PDF by measuring the discrepancy and hypothesis testing (Yates 1934; Smirnov 1948; Massey 1951; Anderson and Darling 1952; Anderson and Darling 1954; Stephens 1974; Plackett 1983). Measures of GOF tests provide the discrepancy between the given data and the values expected under the assumed PDF. Statistical hypothesis tests, such as the chi-squared test, the Anderson–Darling test, and the Kolmogorov–Smirnov test, use such measures to test whether the data follow the assumed PDF. After an appropriate PDF is determined, statistical parameters can be estimated by several methods, including the method of moments, maximum likelihood estimation (MLE), least square methods, Bayesian estimation, and others (Charnes et al. 1976; Scholz 1985; Newey and West 1987; Myung 2003). As long as a parametric PDF for an input can be clearly specified, parametric methods are without doubt powerful tools for characterizing uncertainties.

One important challenge in characterizing uncertainty is that data with nonstandard distributions often appear in unusual shapes, such as a multimodal or mixed shape (Epanechnikov 1969; Adamowski 1985; Bianchi 1997). Extensive studies have been conducted to describe various kinds of PDFs in parametric ways. However, it would be impractical to test all types of PDFs. An alternative way is needed to build a robust framework, regardless of a specific type of PDF. A histogram is the simplest nonparametric estimate of a PDF. An important challenge to using a histogram is its low robustness; different bin sizes can reveal different features of the data. Kernel density estimation (KDE) is widely used to estimate PDF inputs in a nonparametric way (Adamowski 1985; Cao et al. 1994; Pradlwarter and Schuëller 2008; Zambom and Dias 2012). The KDE method uses kernel functions, also called kernel widths (e.g., uniform, triangle, quadratic, tricube, Gaussian, quadratic) to estimate a PDF of inputs (Park and Turlach 1992). In order to estimate PDF by KDE, first, a kernel function is constructed for each sample. Second, those constructed kernel functions are added up and divided by the number of samples. Compared to the discrete histogram method, KDE gets a smooth PDF by substituting each sample into a kernel function. In this way, the smoothness and continuity of the estimated PDF are determined based on suitable usage of the kernel functions. As fewer assumptions (e.g., types of PDF) are made, nonparametric methods have become more widely applicable and more robust than parametric methods. However, in nonparametric methods, a larger number of data are required to achieve the same degree of confidence for estimating a PDF of an input.

3.2.2 Uncertainty characterization under epistemic uncertainty

To accomplish statistical model calibration and validation, uncertainty characterization is used for two purposes (Oberkampf et al. 2004b; Helton et al. 2010; Roy and Oberkampf 2011). First, the physical uncertainties are modeled as a PDF to be incorporated as an input in a computational model. Second, the degrees of uncertainty in the system responses, from both observation and prediction, are modeled as a PDF to be statistically compared. For either purpose, characterizing the aleatory uncertainty in the inputs and the system response is no longer a crucial problem due to the previously introduced methods, provided there is sufficient data. In real-world applications, however, the dearth of data for physical uncertainties and system response is often a concern, resulting in epistemic uncertainty or sampling uncertainty. In the presence of epistemic uncertainty, a special process should be undertaken.

For this process, engineers are requested to identify all uncertainty sources, whether aleatory or epistemic. Although there have been many studies on statistical model calibration and validation, there is at present still no perfect way to split uncertainty sources into aleatory or epistemic. Some studies define inputs without experimental data as epistemic; some studies define inputs with limited experimental data as epistemic (Hoffman and Hammonds 1994; Helton and Burmaster 1996; Winkler 1996; Helton 1997; Oberkampf et al. 2004a; Moon et al. 2017, 2018). Splitting uncertainty sources into aleatory or epistemic remains a controversial topic (Mullins et al. 2016). At present, an efficient way is still needed to classify uncertainty sources into aleatory and epistemic.

Once an uncertainty source is identified as epistemic, several approaches are available. The need to characterize uncertainties even with scarce data has led to several studies (Helton et al. 2004). Using an empirical probability box (p-box) is one possible way, although it can result in large under-estimation or over-estimation (Wu et al. 1990; Helton et al. 2010; Sankararaman and Mahadevan 2011a). An interval is recognized as a simple way to represent epistemic uncertainty; however, it is impossible to be specified with a PDF (Sankararaman and Mahadevan 2011a; Eldred et al. 2011; Urbina et al. 2011). Due to their ease and applicability when only a small amount of data is available, interval-based approaches have garnered attention in many studies (Pradlwarter and Schuëller 2008; Voyles and Roy 2015; Zhu et al. 2016). The evidence theory (Bae et al. 2004; Swiler et al. 2009; Eldred et al. 2011; Salehghaffari and Rais-Rohani 2013; Shah et al. 2015), also referred to as the Dempster-Shafer theory, can be used for interval analysis by aggregating information (e.g., interval data) obtained from different sources and arriving at a degree of belief (i.e., confidence in an interval (Pan et al. 2016)). Similarly, the fuzzy set theory (Moore and Lodwick 2003; Hanss and Turrin 2010; Lima Azevedo et al. 2015) can be used to estimate the interval by combining evidence of different credibility. However, the primary defect of the interval approach is that interval-characterized uncertainties for inputs in computational models need interval arithmetic for statistical model calibration and validation. For instance, the interval-characterized result of uncertainty propagation is produced by the interval-characterized uncertainties. Accordingly, the variations in system response from observation and prediction should be compared by interval. Model calibration and validation based on interval, however, has poor credibility when compared to methods based on expensive PDFs, as the latter methods compare the entire shape of the variation in the system responses (Rebba et al. 2005; Ferson et al. 2008).

In conclusion, representing epistemic uncertainty is still challenging. With existing techniques, constructing an acceptable PDF with scarce data is inherently difficult, unless additional experiments produce more data. In this study, we adopt the statistical model calibration and validation framework (Govers and Link 2010; Youn et al. 2011; Fang et al. 2012; Jung et al. 2014; Bao and Wang 2015). Here, PDFs of inputs that cause epistemic uncertainties are assumed in the forward problem and those epistemic uncertainties are reduced by calibrating the assumed PDFs in the inverse problem. Since next steps of the forward problem can occur after confirming that all PDFs of inputs are acceptable, a PDF for an input causing epistemic uncertainty can be assumed by previously introduced methods, using limited data, knowledge of experts, literature reports, or a mix of these.

3.2.3 Uncertainty characterization with statistical correlation between inputs

Ignoring statistical correlation among inputs of a model may cause unreliable prediction (Li et al. 2014). A variety of statistical methods are available for identifying statistical correlation, such as correlation coefficient (Pearson product moment correlation), Kendall's tau, and Spearman's rho rank correlation. If it turns out to be contributing to the system response, statistical correlation among inputs can be addressed by techniques described in the literature (Jung et al. 2011; Mara and Tarantola 2012; Wei et al. 2015).

A multivariate probability distribution is able to incorporate marginal probability distributions of inputs and statistical correlation between inputs. A copula is a popular way to build a multivariate probability distribution that describes the correlation between inputs (Panchenko 2005; McNeil and Nešlehová 2009). According to Sklar's Theorem (Carley and Taylor 2002; Rüschendorf 2009), any multivariate probability distribution can be written in terms of univariate marginal PDF and a copula function. Two representative families of copulas are (1) Gaussian copulas and (2) Archimedean copulas. A Gaussian copula, also called a Nataf model, is constructed from a multivariate normal distribution and is capable of describing a wider range of correlation coefficients. However, it exhibits undesirable behavior if inputs are highly nonnormal, for example, if marginal PDFs of inputs do not follow a normal PDF. Archimedean copulas are capable of describing multivariate probability distributions; this is not possible with a Gaussian copula. The principal issue in constructing a multivariate distribution model with an Archimedean copula is that only a small range of correlation coefficients can be described. For example, the Morgenstern model, which belongs to the Archimedean family, is only applicable to inputs with low correlation within a range from − 0.3 to 0.3. To exhibit various nonnormal multivariate distributions, as well as a wider range of correlation coefficients, various copula functions, including Clayton, Frank, and Gumbel, have been developed (Genest and Rivest 1993; McNeil 2008; Savu and Trede 2010). The work by Nelson and Joe (Joe 1990; Joe and Hu 1996; Nelsen 2002) provides a clear and detailed introduction to copulas and multivariate probability distribution that considers the relationships between inputs.

3.3 Variable screening

3.3.1 Variable screening in the presence of uncertainty

Variable screening is the study of how the uncertainty in the outputs (system responses) of a model can be apportioned to different sources of uncertainty in its inputs (Helton 1997; Wu and Mohanty 2006; Helton et al. 2006a; Cho et al. 2014; Gan et al. 2014; Wei et al. 2015). The primary purpose of variable screening is to identify inputs that cause significant uncertainty in the system response.

Numerous variable screening methods have been proposed. The easiest way is to prioritize the importance of uncertainty sources based on experts' experience or belief. However, these subjective decisions can lead to adverse effects. Thus, systematic screening studies need to be performed to identify the important inputs that significantly affect the prediction results. A systematic screening analysis should provide evidence to rank the inputs. Typically, a sensitivity analysis, also called a local method, is used by taking the partial derivative of the system response at the design point for the purpose of the design optimization process (Pianosi et al. 2016). For the purpose of variable screening, however, it is important to consider the variation of the system response due to the variations of inputs over a wide range; this is subject to various sources of uncertainty (Chen et al. 2005).

The most applicable approach for considering uncertainty is the variance-based method, a form of global sensitivity analysis (Liang and Mahadevan 2014; Helton 1997; Helton et al. 2006b; Pianosi et al. 2016; Shields and Zhang 2016; Liang et al. 2015; Zhang and Pandey 2014; Gan et al. 2014; Baroni and Tarantola 2014; Sankararaman et al. 2013; Sankararaman and Mahadevan 2013; Plischke et al. 2013). Variance-based methods are based on probabilistic approaches. First, PDFs of inputs and outputs are quantified. To quantify uncertainties in the inputs and outputs, uncertainty characterization (Section 3.2) and uncertainty propagation (Section 3.4) can be used. Next, the importance rank is measured by the amount of variance in the output caused by a single input (main effect index) or by interactions with other inputs (total effect index). The main effect index, also called the first-order index or the Sobol sensitivity index, measures the contribution of the input by itself to the variance in the output (Sobol' 1990; Sobol 2001). On the other hand, the total effect index, also called the global sensitivity index, gives not only the effect on the system response from a single input but also its interaction with other inputs. A more extended explanation of variable screening can be found in the literature (Cacuci and Ionescu-Bujor 2004; Christopher Frey and Patil 2002; Helton et al. 2006b; Pianosi et al. 2016; Gan et al. 2014).

3.3.2 Remarks on variable screening

For variable screening, the variance-based method is recommended to rank the importance of inputs. The variance-based method is carried out using the following procedures. First, the uncertainty in each input is identified and characterized by a PDF. Second, the variation of the system responses is obtained by propagating the input uncertainties through the system model. Last, important inputs are listed in the order of their contribution to system uncertainty, based on the first and second procedure.

Issues related to the first procedure are discussed in Section 3.2. The importance of uncertainty characterization cannot be stressed enough, as it lays the groundwork for statistical model validation. Otherwise, inaccurate characterization of input PDFs can ruin all following procedures. Variable screening commonly assumes independence between inputs; correlation between inputs is often disregarded. However, this assumption can be a serious problem when inputs are strongly correlated (Park et al. 2015).

The second procedure is called uncertainty propagation; this is discussed in Section 3.4. As is always the case in analysis with uncertainty, computational cost is one important problem in variable screening. Various methods have been proposed to reduce the cost of measuring variance-based indices (Chen et al. 2005; Homma and Saltelli 1996; Sobol 2001; Sudret 2008; Zhang and Pandey 2014; Zhang and Pandey 2014). In general, the variance-based method computes sensitivity indices using Monte Carlo simulation. When the number of inputs is large, measuring the index is computationally demanding. A point to notice is that accurate uncertainty propagation is not necessary for variable screening. It is important to recognize that variable screening is a tool to simplify other procedures, not another burden. Sampling-based uncertainty propagation based on surrogate models is an affordable method for variable screening (Helton and Davis 2003). Computationally cheap random designs, such as fractional factorials (Box and Meyer 1986) or Plackett–Burman designs (Tyssedal 2008; Plackett and Burman 1946), can be used, although at the expense of accuracy. Also, dimensional reduction can be used to approximately compute the variance-based indices (Zhang and Pandey 2014).

A complete priority list of inputs provides an order of importance and a quantitative effectiveness on output. Based on this list, the question is how to select inputs for further analysis. The smaller the number of selected inputs, the smaller the computational cost, but with decreased accuracy. After screening out insignificant inputs, they are fixed to the mean value as a deterministic variable, not a random variable. This results in a reduction of the total output variability. If the amount of reduction in the total output variability is large, a loss of accuracy is unavoidable. A recent study proposed hypothesis testing as a way to determine the important variables to address the accuracy problem (Cho et al. 2014).

3.4 Uncertainty propagation

Uncertainty propagation is the study of quantifying uncertainties in the system response. Assuming that model input uncertainties are adequately characterized with a PDF, the uncertainty of the system response is acquired by propagating model input uncertainties through the computational model. Extensive studies have been conducted and different methods developed for uncertainty propagation. These methods are broadly categorized as (1) sampling-based simulation methods and (2) approximate methods. When selecting an uncertainty propagation method, the primary consideration is the tradeoff between accuracy and computational cost.

3.4.1 Sampling-based simulation methods

Sampling-based simulation methods, such as the Monte Carlo method, are not only the most accurate but also the most expensive methods (Bucher 1988; Hurtado and Barbat 1998; Mosegaard and Sambridge 2002; Weathers et al. 2009; Bao and Wang 2015). The basic idea of sampling-based methods is to obtain samples of the system response by repeatedly running the system model with randomly generated input samples. In principle, sampling-based methods can be used to solve any problem having a probabilistic interpretation. For example, sampling-based methods are useful for evaluating (1) multidimensional definite integrals with complicated boundary conditions, (2) nonlinear problems, or (3) simulating systems with many coupled degrees of freedom.

However, sampling-based methods require a large amount of computation time (Fang et al. 2012; Shields et al. 2015). Since the result is deficient if only a few samples are generated, a large number of input samples are needed, which increases the computation time dramatically. Also, the number of function evaluations grows exponentially as the number of dimension increases. For real-world problems, a single simulation can take hours or days to complete. Especially when sampling methods are applied to calibration with optimization techniques in which numerous iteration steps are required, the computational cost burden is too large. To deal with this issue, importance sampling (Melchers 1989; Dubourg et al. 2013; Echard et al. 2013), or adaptive sampling (Bucher 1988; Au and Beck 1999; Gorissen et al. 2010), has been studied to ensure more samples fall within the region of interest, thereby reducing the total number of samples needed for analysis. However, prior knowledge of integrals is needed to do precise importance sampling. Latin hypercube sampling also can help to reduce the total number of samples needed for stable multidimensional distribution (Stein 1987; Huntington and Lyrintzis 1998; Helton and Davis 2003; Helton et al. 2004, 2005; Mousaviraad et al. 2013; Shields et al. 2015; Shields and Zhang 2016). Latin hypercube sampling differs from general random sampling in its consideration of previously generated sample points. Nonetheless, the method still needs a large number of samples.

For sampling-based methods, a computationally expensive model can be replaced with a fast-running surrogate model (Hill and Hunter 1966; Simpson et al. 2001b; Pettit 2004; Giunta et al. 2006; Goel et al. 2007; Viana et al. 2009) (Shan and Wang 2010; Shi et al. 2012; Roussouly et al. 2013; Viana et al. 2014; Tabatabaei et al. 2015; Park et al. 2016b). A surrogate model (also called metamodel, response surface model, or an emulator) is one way to alleviate the burden of expensive computational time by constructing an approximation model. Popular surrogate models include polynomial response surface (Simpson et al. 2001b; Goel et al. 2007; Acar and Rais-Rohani 2009; Goel et al. 2009; Abbas and Morgenthal 2016), the moving least square method (Lancaster and Salkauskas 1981; Levin 1998; Choi et al. 2001; Kang et al. 2010), kriging (Simpson et al. 1998, 2001a; Kaymaz 2005; Goel et al. 2009; Lee et al. 2011; Khodaparast et al. 2011; Zhang et al. 2013; Kleijnen and Mehdad 2014), support vector machines (Hearst et al. 1998; Smola and Schölkopf 2004; Clarke et al. 2005), polynomial chaos expansion (Goel et al. 2009; Hu and Youn 2011; Oladyshkin and Nowak 2012; Kersaudy et al. 2015), artificial neural networks (Gomes and Awruch 2004), and others. For a surrogate model, the challenge is how to build a model that is accurate over the complete domain of interest, while minimizing the simulation cost. For statistical model validation, this challenge is divided into two subissues: (1) design of experiments (DOE) and (2) surrogate model validation.

For the first issue, a surrogate model is constructed based on a computer simulation of intelligently chosen data points. The method used to determine those data points is called design of experiments (DOE; Myers et al. 1995). The goal of DOE is maximizing the amount of information gained from a minimum number of sample points. Two categories of DOE methods are available: classic and modern methods (Hill and Hunter 1966; Steinberg and Hunter 1984; Simpson et al. 2001b). Classic methods, such as full-factorial design, central composite design, Box–Behnken, and D-optimal design, are widely used for designing laboratory experiments. On the other hand, modern methods, such as Latin Hypercube Sampling, orthogonal array design, and uniform design, have been developed for designing computer experiments. Since many software packages provide various kinds of DOE tools today, implementing these methods is no longer difficult. For statistical model validation, the important issue is that the design of experiments should cover the calibration and validation domain. For example, in the calibration procedure, the surrogate model region of interest moves as the input design variables move toward the optimum point. If the DOE does not cover the neighborhood of the optimum point, then the calibration stops before the design variables arrive at the optimum point. Therefore, when designing computational experiments, it is important to clearly understand the calibration and validation domain. A review by C. Viana et al. provides a comprehensive discussion on current research trends of metamodeling (Viana et al. 2014). Two textbooks by Montgomery and cowriters provide comprehensive knowledge on response surface methodology and design of experiments (Myers et al. 1995, 2016; Montgomery 2008).

For the calibration and validation issue, a constructed surrogate model cannot perfectly describe the physical model when considering the true response function is unknown or nonexistent. An overfitting problem arises when the surrogate model has poor predictive performance as it overreacts to minor fluctuations in the training data due to the model's excessive complexity. Thus, validation of the surrogate model is needed. The field of response surface methodology often uses cross-validation (Goel et al. 2007; Viana et al. 2009, 2010; Gorissen et al. 2010). Cross-validation is a model validation technique that assesses how the performance of a constructed surrogate model will generalize to an independent data set. Various ways of cross-validation are available to partition a set of data and to validate a model, including leave-p-out, leave-one-out, and k-fold (Viana et al. 2009; Arlot and Celisse 2010). For cross-validation, first, a set of data is partitioned into complementary subsets. A surrogate model is constructed on one subset and is validated on the other subsets. This procedure is repeatedly performed using different partition ways to separate a set of data. In the end, the validation results from multiple surrogate models are averaged over the repeated processes. Compared with conventional validation, for example, the data set is separated into two sets: 70% for constructing the model and 30% for validating the model. Cross-validation shows better performance with a limited data set. However, cross-validation is computationally more expensive than conventional validation.

3.4.2 Approximate methods

For computationally demanding CAE models, sampling methods quickly become impractical (Fender et al. 2014). This concern can be solved by obtaining an approximate PDF of the system response. To obtain the PDF of the system response, a multidimensional integral is calculated. However, direct integration of multidimensional joint PDF is mathematically infeasible. Numerous studies have thus been conducted to approximately solve the math problem through expansion methods (Lee et al. 2008b; Zhang and Du 2010; Lee et al. 2010), polynomial chaos expansion methods (Wei et al. 2008; Crestaux et al. 2009; Oladyshkin and Nowak 2012), dimension reduction methods (Rahman and Xu 2004; Xu and Rahman 2004; Lee et al. 2008a; Youn et al. 2008; Youn and Wang 2008; Lee et al. 2010), and others.

Approximate methods are much more efficient, but less accurate than sampling-based simulation methods. The expansion methods, including Taylor expansion, the first-order and second-order reliability methods (FORM, SORM), or the perturbation method, approximate statistical moments are calculated (Zhao and Ono 2000; Wojtkiewicz et al. 2001; Youn and Choi 2004; Hua et al. 2008; Khodaparast et al. 2008). In engineering applications, since calculating high-order partial derivatives requires expensive calculations, low-order partial derivatives are predominantly used. Thus, expansion-based methods have difficulties with expressing nonlinearity. Furthermore, when statistical correlations exist between inputs, expansions must be not truncated in low-order partial derivatives. In other words, expansion-based methods cannot consider the statistical correlation that exists in many practical problems. Alternatively, numerical integration-based methods, such as Quadrature formulas (Eldred and Burkardt 2009; Eldred et al. 2011; Mousaviraad et al. 2013), the univariate dimension reduction method (Rahman and Xu 2004; Lee and Chen 2009; Cho et al. 2014), and the eigenvector dimension reduction method (Youn et al. 2008; Jung et al. 2009, 2011), have been proposed to accurately compute multidimensional integrals and simultaneously reduce the computational expense. Specifically, the eigenvector dimension reduction method shows a reasonable accuracy when statistical correlation exists. For situations invloving highly nonlinear function, however, integration-based methods still give misleading results. For example, for the function including a simple product of two inputs, the univariate dimension reduction method and the eigenvector dimension reduction method exhibit a low ability to calculate the statistical moments.

One reason why approximate methods are less accurate than sampling-based simulation methods is that the uncertainty propagation results are statistical moments. Using sampling-based simulation methods, the results of uncertainty propagation are a complete PDF of the system response. For statistical model validation, a complete probability distribution is recommended, as it allows a comparison between experiments and the computational model. On the other hand, approximate methods give the result of uncertainty propagation in the form of statistical moments of the system response (e.g., mean and variance). The basis of the idea is that a sufficient number of statistical moments can provide a good representation of the distribution of the system response. This pertains to the method of moments, which is an uncertainty characterization technique. Devised by Karl Pearson, the Pearson distribution is a family of PDFs that can be specified based upon the first four moments: mean, variance, skewness (normalized third central moment), and kurtosis (normalized fourth central moment; Solomon and Stephens 1978; Lee and Chen 2009; Youn and Wang 2009; Youn and Xi 2009; Youn et al. 2008). The method that selects a parametric PDF for system response based on the Pearson distribution is called the *Pearson system*. The Pearson system, instead of obtaining entire probability distributions of the system response, obtains only the moments of the distributions. Interestingly, the Pearson system was devised in 1895, just a year after the method of moments was proposed by the same researcher. Later, a similar method called the Johnson system was developed in 1949 (Johnson 1949). A limitation of these methods is that only information regarding the central moments is considered. Thus, selection of a PDF based on the central moments of system responses may not properly capture the entire characteristics of the PDF, such as its tails. Additionally, if the PDF of the system response does not belong to the certain family of parametric PDF, for example a multimodal distribution, the selected PDF might have a lower reliability than the actual one.

4 Review of the inverse problem

The inverse problem in statistical model calibration and validation is to estimate the value of epistemic variables in conjunction with experimental observations. In Section 4.1, contemporary issues related to the inverse problem are discussed. Two main approaches are used to solve the inverse problem: (1) calibration using optimization techniques (Section 4.2) and (2) Bayesian updating (Section 4.3).

4.1 Contemporary issues in the inverse problem

The objective of the inverse problem is to statistically calibrate epistemic variables ($\mathbf{X}_e$) of a computational model (Mosegaard and Sambridge 2002; Warner et al. 2015). Issues related to the inverse problem include (1) how to solve the implicit inverse function, (2) how to calibrate epistemic variables given a dearth of data, and (3) how to calibrate multiple epistemic variables.

First, the inverse problem can be solved by taking the inverse of the function ($\mathbf{Y}_{pre}$), as explained in (4) and (5). However, for most computational models that emulate the behavior of engineered systems, it is infeasible to obtain a closed form (explicit form) of the inverse function ($\mathbf{Y}^{-1}{}_{pre}$). To address this difficulty, two approaches have been proposed: (1) optimization-based model calibration and (2) Bayesian-based model calibration. The following subsections (Sections 4.2 and 4.3) provide details on optimization-based model calibration and Bayesian-based model calibration.

Second, epistemic variables ($\mathbf{X}_e$) are calibrated with observed system responses ($\mathbf{Y}_{obs}$) because measuring the system response may be more feasible than directly measuring the quantity of interest, when considering cost and time. However, there is still the problem of performing a sufficient number of experiments on system responses ($\mathbf{Y}_{obs}$) to consider all uncertainty sources. Thus, it is necessary to contrive a way to calibrate epistemic variables ($\mathbf{X}_e$) in a dearth of data.

Third, multiple epistemic variables can arise in the inverse problem. In a large-scale computational model, lack of information on more than two inputs of a computational model is more likely to occur (Zhan et al. 2011b; Fender et al. 2014; Egeberg 2014; Sankararaman and Mahadevan 2015; Jung et al. 2016). When the number of epistemic variables is larger than the number of equations, it leads to an undetermined problem. When an undetermined problem is formulated, various combinations of calibrated epistemic variables may yield comparable fits to observation data ($\mathbf{Y}_{obs}$); this is called the multiple solution problem (Zárate and Caicedo 2008; Manfren et al. 2013). In other words, calibration can produce several models that match experimental data ($\mathbf{Y}_{obs}$), when in fact only one model matches the physical reality.

4.2 Optimization-based model calibration

Optimization techniques provide straightforward calibration methods for solving the inverse problem (Jung et al. 2011, 2016; Youn et al. 2011; Fender et al. 2014; Warner et al. 2015). Section 4.2.1 provides the mathematical formulation of optimization-based model calibration. Section 4.2.2 discusses the calibration metric, which is the objective function for optimization-based model calibration.

4.2.1 Formulation of optimization-based model calibration

The objective of optimization-based model calibration is to inversely estimate epistemic variables so that the prediction is consistent with the observation data. Two ways of achieving this goal are to (1) maximize the agreement or (2) minimize the disagreement between the two probability distributions that were found from the computational prediction and the experimental observation. In this manner, the mathematical statement for the optimization-based calibration can be defined as

$$\underset{\mathbf{X}_e}{\text{Maximize}} \text{ or } \underset{\mathbf{X}_e}{\text{Minimize}} \, f\left(\mathbf{Y}_{obs}, \mathbf{Y}_{pre}(\mathbf{X}_a, \mathbf{X}_e)\right) \tag{6}$$

where *f* denotes the objective function of the optimization problem. Experimental observation ($\mathbf{Y}_{obs}$) and computational prediction ($\mathbf{Y}_{pre}$) of the system response are given as the frequentist PDF under the presence of various uncertainties (i.e., physical, modeling, and statistical uncertainties). If the degree of uncertainty due to measurement error (ε) is negligible, the PDF of the experimental observation ($\mathbf{Y}_{\text{obs}}$) is considered to be a true system response ($\mathbf{Y}$). For computational prediction, all recognized (accounted for) uncertainty sources either from aleatory ($\mathbf{X}_{\text{a}}$) or from epistemic ($\mathbf{X}_{\text{e}}$) variables are incorporated as the inputs ($\mathbf{X}$) to the computational model. In the presence of epistemic variables ($\mathbf{X}_{\text{e}}$), the computational prediction ($\mathbf{Y}_{\text{pre}}$) has an error. An objective function (f) in (6) is then formulated to carry the meaning of the agreement or disagreement between true ($\mathbf{Y}_{\text{obs}}$) and false ($\mathbf{Y}_{\text{pre}}$). By maximizing or minimizing the objective function over the variables, the epistemic variables ($\mathbf{X}_{\text{e}}$) are calibrated.

4.2.2 Calibration metric: objective function for the optimization problem

Establishing a relevant objective function (f) (also simply called a measure or calibration metric) is the key for success of calibration using optimization techniques (Mares et al. 2006; Jung et al. 2011; Fender et al. 2014; Lee et al. 2018). There are a substantial number of similarity or dissimilarity measures encountered in many different fields, such as pattern classification and clustering. Among them, the measures used for optimization-based model calibration are classified into two types: (1) measures that quantify the agreement and (2) measures that quantify the disagreement. For the first type, optimization-based model calibration is performed with the goal of maximizing the value of the measure; for the second type, calibration is performed to minimize the measure. One characteristic that a calibration metric should have is that the function should be globally convex or concave; therefore, it always has an extremum. In addition, it is important to examine how each calibration metric deals with statistical uncertainty due to a lack of data. Basically, a calibrated value that has epistemic uncertainty due to insufficient data cannot be the ultimate answer. However, the important element in solving an inverse problem is that calibrated results with sufficient data should converge to an identical result no matter which approach or method is used.

The likelihood function is the most common measure used to quantify the agreement between two probability distributions (Fonseca et al. 2005; Jung et al. 2011, 2014, 2016; Youn et al. 2011; Xi et al. 2013; Lee et al. 2018). Practically, the log likelihood is commonly incorporated; its probability is estimated in the exponential scale. To evaluate the likelihood function, an assumption needs to be made on the type of PDF (p_{pre}) for computational prediction ($\mathbf{Y}_{\text{pre}}$). Note that the type of statistical parameter ($\theta_{\mathbf{X}}$) is determined based on the selected type of PDF. In other words, uncertainty characterization for computational prediction ($\mathbf{Y}_{\text{pre}}$) is required; this yields another source for statistical uncertainty. However, it is advantageous that uncertainty characterization is not required for experimental observation ($\mathbf{Y}_{\text{obs}}$). Instead, all frequentist information on discrete points are used. Thus, a probability distribution of any shape can be facilitated and directly compared. This results in a low computational cost for evaluation of the likelihood function.

In addition to measures quantifying the agreement, a variety of measures for quantifying the disagreement are available for optimization-based model calibration. Recent studies (Govers and Link 2010; Rui et al. 2013; Bao and Wang 2015) estimated the calibration variables by minimizing the weighted sum of the spatial distance between the statistical moments (e.g., mean, standard deviation, covariance) of observation data and that of prediction data. Because statistical parameters represent a PDF, the distance between the two statistical parameters derived from observation and prediction is acceptable for specifying the difference between prediction and observation. However, with only a few statistical parameters, fully describing the distribution of the system response is difficult, especially when the distribution does not follow a common type of PDF (e.g., normal, lognormal). This method has shown good performance when the PDF of the system response follows a specific PDF, such as a normal, lognormal, and others. In many practical cases, the distributions of system responses are not in a specific type of probability density function (Pradlwarter and Schuëller 2008); for example, they may be a multimodal distribution as a mixture of several distributions. In such a case, there are certain limits to the use of the distance between statistical parameters as an objective function. A subjective assumption that a system response follows a particular PDF could be applied in practical settings; however, an erroneous assumption about the type of PDF is a problem. In this respect, a measure must be developed that can directly compare two PDFs. A comprehensive survey by Cha and Choi (Cha 2007; Choi et al. 2010b) provides various categories of measures that are applicable to compare two PDFs. The survey categorized measures into various families: (1) the L_p Minkowski family, which are developed on the basis of Euclidean distance; (2) the L_1 family, which facilitates the absolute difference; (3) the intersection family, which are in the form of similarity; (4) the inner product family, which incorporates the inner product; (5) the squared-chord family, which are based on the sum of geometric means; (6) the squared L_1 family, which are based upon the squared Euclidean distance; (7) the Shannon's entropy family, which are developed from the relative entropy, also called Kullback–Leibler (Oden et al. 2013); and (8) combinations, which utilize multiple ideas or measures. These measures can be used in future work examining optimization-based model calibration. Detailed explanations for each measure are beyond the scope of this paper. Measures based on Euclidean distance are straightforward. However, a hindrance of this measure is the need to characterize the PDF ($p_{\text{obs}}(y_i)$) of the experiments. The process of characterizing the uncertainty in experimental data may lead to statistical uncertainty, especially with a dearth of data. As a limitation, an accurate calibration using the optimization-based model calibration highly depends on both the quantity and quality of the given experimental data. Also, if the assumption on the distribution type of unknown input variable is wrong, it may lead inaccurate calibrated results.

4.3 Bayesian-based model calibration

The basic principle of Bayesian inference is to derive the posterior probability using a prior probability and likelihood function derived from a statistical model for the observed data (Mahadevan and Rebba 2005; Xiong et al. 2009; Arendt et al. 2012a; Park et al. 2016a). The review of this technique begins by explaining the purpose-built formulation of Bayesian-based model calibration (Section 4.3.1). The following section summarizes research on model calibration using Bayesian inference (Section 4.3.2).

4.3.1 Formulation of Bayesian-based model calibration

Bayesian inference updates the probability of unknown parameters (calibration variables) as more observational data become available. The prior information of the unknown parameters vector ($\boldsymbol{\theta}$) is given in the form of joint PDF ($p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$), and the experimental data ($\mathbf{y}$) are given with variability. The posterior distribution ($p_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y})$) of the parameter ($\boldsymbol{\theta}$) can then be expressed as

$$\boldsymbol{p}_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) \propto \boldsymbol{p}_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})\boldsymbol{p}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \tag{7}$$

where $p_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function that elucidates the probability of observing data ($\mathbf{y}$), given parameters ($\boldsymbol{\theta}$). The Bayesian update in (7) can be applicable when the values of parameters ($\boldsymbol{\theta}$) are observable. In the case of model calibration, the epistemic variables ($\mathbf{X}_e$), which replace parameter vector ($\boldsymbol{\theta}$) in (7), of the model are not directly observable; however, the model responses ($\mathbf{Y}_{\text{obs}}$) are. Therefore, the Bayesian calibration requires the relationship between the model and experimental observations. (In statistical sense, the inverse problem estimates the statistical parameters of the calibration variables.) For Bayesian-based model calibration, the Bayesian formulation can be expressed as

$$\boldsymbol{p}_{\boldsymbol{\theta}_{\mathbf{X}_e}|\mathbf{Y}}(\boldsymbol{\theta}_{\mathbf{X}}|\mathbf{Y} = \mathbf{Y}_{\text{obs}}) \propto \boldsymbol{p}_{\mathbf{Y}|\boldsymbol{\Theta}_{\mathbf{X}}}(\mathbf{Y} = \mathbf{Y}_{\text{obs}}|\boldsymbol{\theta}_{\mathbf{X}})\boldsymbol{p}_{\boldsymbol{\Theta}_{\mathbf{X}}}(\boldsymbol{\theta}_{\mathbf{X}}) \tag{8}$$

where the posterior distribution ($p_{\boldsymbol{\theta}|\mathbf{Y}}(\boldsymbol{\theta}_{\mathbf{X}}|\mathbf{Y}=\mathbf{Y}_{\text{obs}})$) of statistical parameters ($\boldsymbol{\theta}_{\mathbf{X}e}$) of calibration variables ($\mathbf{X}_e$) is proportional to the likelihood ($p_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}=\mathbf{Y}_{\text{obs}}|\boldsymbol{\theta}_{\mathbf{X}})$) times the prior distribution ($p_{\boldsymbol{\theta}}(\boldsymbol{\theta}_{\mathbf{X}})$). In Bayesian statistics, a state of knowledge or belief of an unknown parameter is expressed in terms of PDF ($p(\cdot)$). That is, if the value of an unknown parameter is well known, the distribution will have small uncertainty, and vice versa. In other words, Bayesian-based model calibration uses Bayesian inference to reduce epistemic uncertainty in calibration variables. As experimental data set ($\mathbf{Y}_{\text{obs}}$) becomes larger, the variation in the posterior distribution becomes narrower, which means the degree of uncertainty decreases.

The major benefit of the Bayesian approach is the ability to incorporate prior information. Using prior knowledge of uncertainties, the Bayesian approach has strength in calibration compared to frequentist probability (Liu et al. 2011). One crucial issue in the inverse problem is how to estimate the value of calibration variables with limited observation data (Hemez et al. 2010; Jiang and Mahadevan 2009b). Optimization-based model calibration is based on a frequentist approach, which exclusively relies on the sample data to estimate calibration variables. For optimization-based model calibration, a small number of experimental data can lead to incorrect estimation of the calibration variables. On the other hand, the Bayesian approach utilizes the prior information in conjunction with newly available data to obtain the posterior knowledge for calibration variables. Thus, Bayesian-based model calibration is capable of continuously updating the prior information with evolving experimental data to obtain the posterior information (Higdon et al. 2008; McFarland et al. 2008; Manfren et al. 2013). The major argument encountered against using the Bayesian approach is that it typically requires using expert knowledge or information from previous experiments (Cooke and Goossens 2004; McKay and Meyer 2000; Thorne and Williams 1992; Liu et al. 2011). Improper prior knowledge can lead to improper posteriors (Higdon et al. 2008). When the prior distribution is inconsistent with observed physical data, the process can be either slow to converge or converge to wrong results. Rarely, a prior distribution can be chosen from a conjugated distribution family. When there is no information regarding the prior distribution, uninformative prior or uniform prior can be used (Park and Grandhi 2014). In most cases, the data for both calibration and validation are limited (different data sets should be used for each process); thus, the Bayesian approach seems to be the one that must be studied and developed.

4.3.2 Review of literature on Bayesian-based model calibration

Systematic model calibration using Bayesian inference is mostly based on the seminal publication by Kennedy and O'Hagan (2001), followed by others (Li and Mahadevan 2016; Bayarri et al. 2007; Higdon et al. 2008; McFarland and Mahadevan 2008b; Rebba et al. 2006; Bayarri et al. 2007; Chen et al. 2008; Liu et al. 2008; Qian and Wu 2008; Xi et al. 2013; Lee et al. 2014). The original paper by Kennedy and O'Hagan used the Gaussian process model for Bayesian calibration frameworks, which found the calibration variables that were the most statistically consistent with data from experiments or high-fidelity simulation. An important difference in the approach by Kennedy and O'Hagan is to include not only calibration variables but also a discrepancy function, which represents the effect of model-form error as well as numerical error. In this framework, the relationship between the prediction model and observation can be represented by

$$\mathbf{Y}_{\text{obs}}(\mathbf{X}) = \mathbf{Y}_{\text{pre}}(\mathbf{X}; \theta_{\text{e}}) + \boldsymbol{\delta}(\mathbf{X}) + \varepsilon \tag{9}$$

where $\mathbf{X}$ is the vector of inputs; θ_{e} is a set of epistemic variables (unknown model parameters); $\delta(\mathbf{X})$ is the model error (discrepancy function), which is defined as the difference between model prediction and reality; and ε is the measurement error, which is usually assumed to be a Gaussian distribution ($\varepsilon \sim N(0, \sigma_\varepsilon^2)$). The result from this framework provides a calibrated value of epistemic variables (θ_{e}), corrected hyperparameters (θ_δ) of quantity of model error ($\delta(\mathbf{X})$), and the variance (σ_ε^2) of the measurement error. The prior distributions ($f'(\theta)$) for each parameter (θ_e, θ_δ, σ_ε^2) are predefined. Using the experimental data, the posterior distribution ($f''(\theta)$) is updated as

$$f^{''}(\theta|\mathbf{Y}_{\text{obs}}) = \frac{L(\theta|\mathbf{Y}_{\text{obs}})f^{'}(\theta)}{\int L(\theta|\mathbf{Y}_{\text{obs}})f^{'}(\theta)d\theta} \tag{10}$$

where $L(\theta)$ is the likelihood function, and the prior distribution ($f'(\theta)$) and the posterior distribution ($f''(\theta)$) are the joint PDF. Markov chain Monte Carlo (MCMC) methods are often used for calculating numerical approximations of multidimensional integrals (Li and Mahadevan 2016; Oden et al. 2013). MCMC requires millions of samples to guarantee convergence and is time consuming, although surrogate methods have been attempted (McFarland and Mahadevan 2008a; Yuan et al. 2013; Bao and Wang 2015).

Studies using this approach emphasize that the advantage of introducing the discrepancy function is that it can consider the possible existence of unrecognized (unaccounted for) uncertainty sources. This framework can deal with missing physics and other inaccuracies of the computer model or experimental error, while updating calibration variables (Kennedy and O'Hagan 2001; Bayarri et al. 2007; Chen et al. 2008; Qian and Wu 2008; Xi et al. 2013). The first method by Kennedy and O'Hagan suggests simultaneous estimation of calibration variables and the discrepancy function (Kennedy and O'Hagan 2001). However, a simultaneous process requires high-dimensional integration. Since the roles of calibration variables and the discrepancy function are different, decomposed approaches are often employed, where the calibration variables are identified first using unbiased estimation, and then, the discrepancy function is determined based on the local discrepancy between the observation and the prediction. Arendt et al. developed a modular Bayesian approach, which separates the estimation of the epistemic variables and the discrepancy function into two modules (Arendt et al. 2012a). A Gaussian process model by multiple response was proposed for separating sources of uncertainty (Arendt et al. 2012b). Later, Xi et al. presented a two-step calibration procedure in which first calibration variables were calibrated and then the discrepancy function was identified by the likelihood function (Xi et al. 2013). Park et al. showed that un-biased estimation of calibration variables can lead to a complex discrepancy and suggested that calibration variables be found that lead to the simplest form of the discrepancy function (Park et al. 2017).

In most cases, the data for both calibration and validation are limited (different data sets should be used for each process); thus, the Bayesian approach seems to be the one that must be studied and developed. To this end, many studies follow the work developed by Kennedy and O'Hagan (Li and Mahadevan 2016; Bayarri et al. 2007; Higdon et al. 2008; McFarland and Mahadevan 2008b; Park and Grandhi 2014). A dynamic model discrepancy framework was proposed to correct a discrete time prediction model (Hu et al. 2019). Nevertheless, there are several limitations of using the Bayesian approach accompanied by the discrepancy function, including (1) a corrected model error can be different at different levels of a system, (2) a separate incorporation of the model discrepancy term can lead to inaccurate bias in model predictions, and (3) an increased number of unknown parameters, (including calibration variables, hyper parameters of the discrepancy function, and measurement error) may cause the overfitting problem or the undetermined problem.

5 Review of the validation problem

This section provides a review of the validation problem. Two essential validation problems are the validation metric and the decision problem, as discussed in Section 2.3.3. Section 5.1 begins with a discussion of contemporary issues surrounding the validation problem in statistical model validation. Reviews of the validation metric and the decision problem follow, in Sections 5.2 and 5.3, respectively.

5.1 Contemporary issues related to the validation problem

Previous studies present several issues that surround the validation problems of statistical model calibration and validation, including (1) different types of system responses require different types validation metrics and (2) a criterion must be established to be used to decide whether a computational model is valid or not.

A validation metric is used to quantify the difference in the two system responses that arise from observation and prediction. System responses can be categorized into two types: (1) stationary or steady-state system responses (univariate) and (2) transient or dynamic system responses (multivariate). The first type of system response has a stationary characteristic, for example, a degree of beam deflection by external loading or from the natural frequency of a system. The second type of system response has a periodic character or a complex mixture of many frequencies, such as the velocity of a moving object or vibration frequencies. The uncertainty of the two types of system response appears in different ways. Thus, two distinct validation metrics are required for the two types of system response. Section 5.2 provides reviews of several validation metrics.

A validation metric is engaged to quantify the difference between the observation and prediction system responses. After quantitative comparison by a validation metric, a decision is made about whether to accept the validity of the computational model. The issue disputed in the decision-making process is how to establish a criterion by which the decision is made. Hypothesis testing is a widely used method to make a statistical decision. A review of decision-making in the validation problem is presented in Section 5.3.

5.2 Validation metric

Earlier studies by Oberkampf and Barone (2006), Ferson et al. (2008), and Liu et al. (2011) summarized the desired features of validation metrics. Based on these studies, the desired features of a validation metric can be summarized as "objective" or "stochastic or statistical." It is remarkable that despite research on constructing and utilizing CAE, the model still does not use quantitative methods for validation; instead, qualitative methods are used, such as visual assessment (Ling and Mahadevan 2013). Qualitative validation can give insights into differences between measured and predicted results; however, results from qualitative validation differ by user. In contrast, quantitative validation gives an objective measure (Jiang and Mahadevan 2011; Zhan et al. 2011b, 2012a, b; Murmann et al. 2016; Oberkampf and Barone 2006). Furthermore, in a statistical sense, a system response is presented as a variation, for example, by a probability distribution or a random process, due to existence of various uncertainties. A "statistical or stochastic" validation metric should be able to compare the system responses from experiments and simulations in a variation by considering uncertainties (Jiang and Mahadevan 2011; Sarin et al. 2010; Schwer 2007).

The primary consideration in the selection of an effective metric should be the type of system response. The system response of interest can be (1) stationary, steady-state, or scalars (random variables) or (2) transient, dynamic, or histories (random processes; Murmann et al. 2016; Liu et al. 2011; Teferra et al. 2014). Stationary responses have a determined value regardless of time or space (e.g., natural frequency). On the other hand, dynamic responses denote dynamic histories or curves (e.g., acceleration; Oberkampf and Barone 2006; Dowding et al. 2008; Teferra et al. 2014). For statistical comparison, the stationary response needs a distribution comparison method since the scalar values build a distribution from different conditions of prediction and observation (Halder and Bhattacharya 2011; Mahadevan and Rebba 2005; Chen et al. 2004). However, the validation metrics developed for stationary responses do not have an ability to capture the difference in dynamic responses (Oberkampf and Barone 2006), such as magnitude or phase difference in time-dependent system responses (e.g., fluid velocity in a pipe). In the case of dynamic responses, responses are represented as random processes; accordingly, a validation metric is required that can capture the difference in a dynamic path as well as consider stochastic characteristics in random processes (Oberkampf and Barone 2006; Dowding et al. 2008; Kokkolaras et al. 2013; Sarin et al. 2010; Zhan et al. 2011a, 2012a; Hills 2006; Hasselman et al. 2005). In Sections 5.2.1 and 5.2.2, validation metrics for stationary and dynamic responses are summarized and discussed, respectively.

5.2.1 Validation metrics for stationary system responses

In a probabilistic context, the prediction of stationary responses is represented as a probability distribution due to various uncertainty factors (Kokkolaras et al. 2013). Therefore, a distribution comparison method is used for statistical model validation of stationary responses. Harmel et al. proposed a metric called the degree of overlap for model validation motivated by the methods in the statistics community (Harmel et al. 2010). The basic idea of the degree of overlap is that the closer the measured and predicted results are, the more their probability distributions overlap. In the same manner, the calibration metrics discussed in Section 4.2.2 also can play a role as a validation metric for comparing two probability distributions (Fonseca et al. 2005; Jung et al. 2014; Xiong et al. 2009; Youn et al. 2011). However, significant experimental results are required to use the degree of overlapped area or the calibration metrics. Since minimizing the number of experiments is always an important issue in model validation, the degree of overlapped area and the calibration metrics have fewer practical applications.

As a powerful alternative, a number of works recently in the model validation community show that the area metric possesses most of the desirable features of a validation metric for comparing experiment and simulation (Ferson and Oberkampf 2009; Ferson et al. 2008; Liu et al. 2011; Voyles and Roy 2015; Thacker and Paez 2014; Roy and Oberkampf 2011). To obtain the area metric value, propagated system responses of simulation produce a probability box or p-box, while experimental measurements are used to construct an empirical CDF of system responses (Ferson et al. 2008). Finally, the minimum area between these two structures is referred to as the value of the area validation metric. The comparative studies (Ferson et al. 2008; Oberkampf and Barone 2006; Liu et al. 2011; Roy and Oberkampf 2011; Thacker and Paez 2014) showed the area metric to be promising due to its favorable features compared to other methods. First, the area metric is one of only a few developed distribution comparison metrics. It measures the entire distribution rather than statistical moments, thereby accounting for uncertainties in both the simulation and the experiments. However, the calculated value of the area metric can lead to highly biased validation conclusions. Second, sampling uncertainty due to limited experimental data for the validity check is considered (Jung et al. 2014) in conjunction of the area metric.

U-pooling and T-pooling methods are proposed to assist with usage of the area metric (Ferson et al. 2008; Li et al. 2014; Liu et al. 2011). While the area metric can validate the prediction with a small number of data, conducting a few more experiments could still be a burden to the analysts. The U-pooling method can ease this burden by pooling all experimental observations at different validation sites into a u-value CDF (Ferson et al. 2008; Li et al. 2014; Liu et al. 2011). Through the u-pooling technique, the area metric takes advantage when multiple experiments at various validation sites are available. The original u-pooling method is only applicable for a single-system response at a single validation site. To extend the usage of the area metric for validating correlated multiple responses, Li et al. (2014) proposed the t-pooling method, accompanied by probability integral transformation (PIT). The multivariate PIT method transforms the joint CDF of the system responses into a univariate CDF, and then, the t-pooling method integrates the evidence from all relevant data of multiresponse quantities over an intended validation domain into a single measure to assess the overall disagreement. In general, the surveyed U-pooling and T-pooling validation methods that employ the area metric appear to be useful only when the model turns out to be highly accurate over the set of validation points and the validation uncertainty is small enough to establish this tightly. Then, the model can be justified for prediction without adjusting or correcting for prediction bias and uncertainty before predicting at new conditions beyond the validation database. However, in common where observed and predicted PDFs of system responses (outputs) do not closely overlie each other, then presently, unsolved difficulties exist with the above approach. One very general methodology demonstrated to handle more difficult problem conditions with significant validation bias errors approximately corrected for extrapolative predictions (with extrapolation-scaled uncertainty on the correction) is the following (Romero 2019).

Even though the area metric is the promising method in the current V&V field, there have also been attempts to develop other metrics. A validation metric called a model reliability metric for stationary system responses was proposed by Rebba and Mahadevan (Mullins et al. 2016; Rebba and Mahadevan 2008). The model reliability metric is a direct measure of model prediction quality and has a strength in its computational ease. The metric may be meaningful in that it provides a statistical result of the difference (bound between 0 and 1) between the probability distribution from observation and prediction. However, robustness can be a problem since the expertise or experience of the user is required to determine the tolerance limit. Also, since the number of experiments is limited compared to the number of simulations, the probability distribution of observation cannot be estimated accurately. Therefore, there is a disadvantage in that an accurate value of the model reliability metric cannot be calculated unless a sufficient number of experiments are available. This means that the model reliability metric is still not a better choice than the area metric that takes certain types of epistemic uncertainty into account.

5.2.2 Validation metrics for dynamic system responses

For system responses in a time-dependent or frequency-dependent domain, a validation metric requires an ability to quantify the agreement or the disagreement between two dynamic responses. A validation metric should catch the difference in important features of dynamic response, for example, magnitude or phase. Several validation metrics have been developed and these validation metrics are classified into two categories: (1) single-value metrics and (2) comprehensive metrics (Mongiardini et al. 2010).

Single-value metrics give a single numerical value that represents the agreement between the two dynamic responses. Various types of single-value metrics have been proposed, including relative error (Schwer 2007; Kat and Els 2012), vector norms (Sarin et al. 2010), root mean square error (Mongiardini et al. 2010), Theil's equality (Murray_smith 1998; Whang et al. 1994), Zilliacus Error and Whang's inequality (Murmann et al. 2016; Whang et al. 1994), correlation coefficient (Mongiardini et al. 2010), weight integrated factor (WIFac; Twisk et al. 2007; Twisk et al. 2007), and index of agreement (Willmott et al. 2012). The main limitation of single-value metrics is that they can measure the error in one aspect alone, such as magnitude or phase, but not in multiple aspects. Comprehensive metrics, on the other hand, treat the features of dynamic response separately, such as magnitude, phase, or slope. A single-value comprehensive metric is constructed by combining individual metrics. One important characteristic that a comprehensive metric should have is that each metric should be independent. For example, the phase metric should be insensitive to difference in magnitudes but sensitive to difference in phases. In an earlier study by Geer, a comprehensive metric was proposed (Geers 1984; Kwaśniewski 2009; Mongiardini et al. 2009, 2013; Murmann et al. 2016), followed by various metrics including Russell (Russell 1997), Sprague and Geer (Sprague and Geers 2004; Lee and Gard 2014; Schwer 2007), and Knowles and Gear (Schwer 2007; Lee and Gard 2014; Sarin et al. 2010; Mongiardini et al. 2010). Detailed explanations for each metric are beyond the scope of this paper.

In terms of comprehensive metrics, Sarin et al. (2010) developed an objective rating metric, named Error Assessment of Response Time Histories (EARTH). Similar to other comprehensive metrics, this metric separates the features of dynamic responses into phase, magnitude, and slope. To separate the features into three, the EARTH metric implements dynamic time warping (DTW). In later work by Zhan et al. (Zhan et al. 2012a; Fu et al. 2010), principal component analysis (PCA) was used to quantitatively assess the agreements of important features of multiple dynamic responses simultaneously. The DTW and PCA align multiple points of the other time history that lie in different temporal positions, so as to compensate for temporal shifts. Similarly, for frequency-dependent system response, Jiang and Mahadevan (2011) and McCusker et al. (2010) used a wavelet transform (WT) that decomposes a dynamic system response into a set of time domain basis functions with various frequency resolutions. Feature extraction is an important element for model validation of a dynamic system; however, this paper does not cover that in detail. Further information about feature extraction is found in the literature; Jiang and Mahadevan (2011) provide a good review.

Although each of these metrics has useful characteristics (e.g., magnitude, phase, slope or mixed), the ultimate limitation of them is that they ignore the uncertainties in the dynamic responses from experiments and simulation (Kokkolaras et al. 2013). Sarin and Kokkolaras (Sarin et al. 2010; Zhan et al. 2011a, b) used average residual and its standard deviation to compare time histories for validation of a simulation model. Average residual has a disadvantage in that positive and negative error at various points may cancel each other. To compensate for this, its standard deviation is adopted, which represents how much error is distributed from the average residual. Xi et al. (2015) proposed a validation metric for general dynamic system responses under uncertainty which makes use of the U-pooling approach and extends it for dynamic responses. A future validation metric for dynamic system responses is expected to consider uncertainties in experiments and simulation by adopting the above method.

5.3 Decision-making method

A validation metric is a stand-alone measure that indicates the level of agreement or disagreement between computational and experimental results. However, the acceptance criterion for a yet-to-be-validated model is another important issue. Suppose a desirable validation metric quantified the difference between the experiment and the simulation. It is then necessary to establish a method to determine the validity of the model (Oberkampf and Trucano 2002). Statistical hypothesis testing with a specified level of significance is widely used as a decision-making tool for model validation (Kokkolaras et al. 2013; Chen et al. 2004; Jung et al. 2014). Hypothesis testing aims to determine whether the acceptance or rejection of a model is valid or not using quantitative measurements of the discrepancy between the experiment and the simulation. Two kinds of hypothesis testing have been studied: classical and Bayesian (Oberkampf and Barone 2006). The significant difference between the two testing methods is that classical testing focuses on model *rejection* in the validity check, whereas the Bayesian method focuses on model *acceptance* by using prior information (Liu et al. 2011). The following subsections provide a brief discussion of both classical hypothesis testing and Bayesian hypothesis testing.

5.3.1 Classical hypothesis testing

Classical hypothesis testing is a well-developed statistical method for accepting or rejecting a model based on statistics (Kat and Els 2012; Oberkampf and Barone 2006; Chen et al. 2004; Jung et al. 2014; Ling and Mahadevan 2013). First, the null hypothesis (H_0: $\mu_{pre} - \mu_{exp} = 0$) and the alternative hypothesis (H_1: $\mu_{pre} - \mu_{exp} \neq 0$) are defined. The former means that the difference between the predicted and observed system responses is not statistically significant; the latter means there is a statistically significant difference. Hypothesis testing is based on test statistics. If the observed value of the physical observation test statistic falls outside of the critical region of the test statistic, the null hypothesis is rejected, meaning that the observations from experiments and the simulation prediction are significantly different. The corresponding *P* value is calculated as the probability that the test statistic will fall outside the range defined by the calculated value of the test statistic under the null hypothesis (Ling and Mahadevan 2013).

Classical hypothesis testing is a method to compare two distributions either by using the first two moments or the full statistical distributions. To compare statistical moments, the *t* test statistic and the *F* test statistic can be used to examine the consistency of the mean and variance (Rebba and Mahadevan 2006). To compare the full distribution, commonly, the differences between the empirical and prediction CDFs are measured. For this purpose, the Anderson–Darling test, the Kolmogorov–Smirnov (K–S) test, and the Cramer–von Mises test were introduced and are found in the literature (Rebba and Mahadevan 2008).

Wrong decisions, or statistical errors, can be made in statistical hypothesis testing (i.e., type 1 and 2 errors). Type 1 error denotes an incorrect rejection of a true null hypothesis (H_0) that the system response from a computational model follows the system response from experiments; an alternative hypothesis (H_1) stands for the opposite (Jung et al. 2014). Type 2 error is the failure to reject a false null hypothesis. By adjusting the criteria, the rate of type 1 or 2 error is determined. As type 1 error (or significance level, α) increases, type 2 error decreases. Therefore, a high type 1 error should be used for validity checks.

According to Liu et al. (2011), classical hypothesis testing seldom rejects a better model. A problem arises when a small amount of physical experiment data is available; this is common in many fields. Comparing full distribution data is impossible, but with a small number of data, there is possibility that the result is not trustworthy. In this situation, the confidence interval of prediction distribution can be used by checking if the observed data fall inside the interval (Halder and Bhattacharya 2011; Ghanem et al. 2008; Buranathiti et al. 2006; Chen et al. 2004). However, this strategy can tend not to reject an incorrect model, since a small number of data fall inside the confidence interval of prediction with high possibility (Liu et al. 2011). Furthermore, specifying the confidence level creates another problem, since a small perturbation of the confidence level largely affects the results of acceptance or rejection. On the other hand, a large number of samples can give misleading results, because as the number of samples increases, the null hypothesis tends to be rejected at a given significance level (Ling and Mahadevan 2013). Above all, it should be noted that failing to reject the null hypothesis does not prove that null hypothesis is true.

5.3.2 Bayesian hypothesis testing

Bayesian estimation has been developed for updating statistical models of uncertain parameters in the computational model, especially when there is not enough data for statistical modeling of the input parameter, as discussed in Section 4.3. The goal of the validation problem, however, is to assess the predictive capability of the model, not to optimize the agreement between the model and the measurement. For the purpose of validation, the Bayesian hypothesis is to utilize Bayesian analysis or Bayesian statistical inference in hypothesis testing. Oberkampf describes how Bayesian hypothesis testing has advantages over classical hypothesis testing by incorporating an analyst's prior belief of model validity (Oberkampf and Barone 2006; Li et al. 2014). Mahadevan and coworkers used the Bayesian methodology to quantify the agreement between the computer model prediction and physical test results (Mahadevan and Rebba 2005). Rebba and Mahadevan (2006) and Jiang and Mahadevan (2008) used point Bayesian hypothesis testing to infer how strongly the experimental data supports the null hypothesis (H_0: $\mu_{\text{pre}} - \mu_{\text{exp}} = 0$) to accept the model as opposed to the alternative hypothesis (H_1: $\mu_{\text{pre}} - \mu_{\text{exp}} \neq 0$) to reject the model. Jiang and Mahadevan (2008) later formulated the interval-based hypothesis test (H_0: $|\mu_{\text{pre}} - \mu_{\text{exp}}| < \varepsilon$) and concluded that the interval-based hypothesis test provides a more consistent model validation result because rejection of the point null hypothesis merely proves that the prediction and the observation are not exactly the same. Other work showed that as the amount of data increases, the interval-based method converges to the correct inference (Jiang and Mahadevan 2008).

In some studies, Bayes factor B_0 is used as a validation metric based on the hypothesis testing (Jiang and Mahadevan 2009a, b; Chen et al. 2008; Jiang et al. 2013a). Basically, the Bayes factor looks at the ratio of the posterior distribution of the null and alternative hypothesis to infer whether the experimental data fall inside the statistical populations from the simulation. In addition, it can be interpreted as a ratio of relative likelihood of the null hypothesis that the experimental data support the predictions and the alternative hypothesis that the data does not support the predictions. The Bayes factor (B_0) can be expressed as

$$\begin{aligned} \mathrm{B}_0 &= \frac{Pr\{\text{data}|\mathrm{H}_0 : \mu = \mu_0,\ \ \sigma = \sigma_0\}}{Pr\{\text{data}|\mathrm{H}_1 : \mu \neq \mu_0,\ \ \sigma \neq \sigma_0\}} \\ &= \frac{L(\text{data}|\mu_0, \sigma_0)}{\iint L(\text{data}|\mu, \sigma) f^{\text{prior}}(\mu, \sigma) d\mu d\sigma} \end{aligned} \tag{11}$$

where μ_0 and σ_0 are the mean and standard deviation of the prediction, respectively. In the right-hand side of the (11), $L(\text{data}|\ \mu_0,\ \sigma_0)$ is the likelihood of observing the data under the null hypothesis and the denominator means the integration of any competing distribution that the data can support. The Bayes factor can also be expressed as

$$\mathrm{B}_0 = \left.\frac{f^{\text{post}}(\mu, \sigma|\text{data})}{f^{\text{prior}}(\mu, \sigma)}\right|_{\mu=\mu_0,\ \ \sigma=\sigma_0} \tag{12}$$

where $f^{\text{prior}}(\mu,\ \sigma)$ is the prior PDF of the mean and standard deviation under the alternative hypothesis, and $f^{\text{post}}(\mu,\ \sigma|\ \text{data})$ is the posterior PDF of the mean and standard deviation after being inferred by data. As a large Bayes factor value indicates that the observation from experiments increasingly supports the prediction results, it acts as a quantitative metric as well as a decision method. Jiang and Mahadevan (2007) showed that the threshold of the Bayes factor for model acceptance can be derived based on a risk vs. cost trade-off, thereby aiding in robust, meaningful decision-making.

It is noteworthy that the null hypothesis in classical hypothesis testing is not accepted with confidence even if the test statistic falls into the confidence interval. However, Bayesian hypothesis testing can allow a confident result via the assistance of prior expert knowledge. Even when only one experimental data point is available, the Bayesian method still can identify whether the simulation results and the experiment results belong to the same population. An additional desirable feature is that adding more data increases the confidence. It is also worth noting that the Bayesian method can be used for model updating even when a large amount of data is not available at present, but when additional future experiments are planned. However, it can be misleading if the experts have inconsistent prior knowledge. Furthermore, Bayesian test results are focused on the statistical moments (e.g., mean and standard deviation), not the full distribution.

6 Conclusions and remaining challenges in statistical model calibration and validation

This paper reviews methods and development trends for statistical model calibration and validation. The review begins with a systematic categorization of uncertainty. Based on the categorized uncertainty structure, the entire process of statistical model calibration and validation is divided into three problems: (1) the forward problem, (2) the inverse problem, and (3) the validation problem. For each of the three problems, the review summarizes the published literature and the developed methodologies. This review not only explores research works on statistical model calibration and validation but also examines statistical methodologies that can be used in each of the three problems of statistical model calibration and validation. This study is expected to provide a general guideline for constructing calibration and validation procedures in practical applications.

To perform statistical model calibration and validation successfully, several challenges must be addressed including addressing the forward, inverse, and validation problems. Challenges and future opportunities include:

1) With unlimited resources, statistical model validation based on developed methodologies would not be a problem. However, in real-world settings, there is always a need to minimize the cost of statistical model validation. Most of the costs related to existing methods are paid for designing and conducting validation experiments that include as many sources of uncertainty as possible. Thus, the design of validation experiments is perhaps the most important part of the process. Nevertheless, studies on design of validation experiments are still few in number. The reason for this is because each case study has a different purpose and different resources for the validation experiments. A general study on design of validation experiments with limited resources will likely be proposed in the future. At the same time, proper results of statistical model validation can convey credibility and confidence in statistical analysis in order to make better decisions about whether to believe the capability of a computational model. Furthermore, statistical model validation is not only to assess the degree of model accuracy but can also be very useful for effective design process, decision-making for usage of model, estimating confidence on model prediction, etc. Another challenge on model validation is to deal with epistemic uncertainty due to limited data. Appropriate conservativeness is not easy to achieve given lack of test data. A related matter is the extrapolation of prediction bias (or a correction) from calibration and/or validation points to prediction points in the modeling space. Appropriate packaging of information from the validation assessment to best enable prediction bias correction and its extrapolation is extremely important but not yet well studied or solved by the verification, validation, and uncertainty quantification community. This is an area ripe for intensive future research and will impact the optimal development of validation methods and their use by the engineering analysis community.
2) One function of uncertainty characterization is to construct an intact PDF for fully describing the uncertainty in observed system responses. When observing the system responses through validation experiments, it is difficult to avoid uncertainties during measurement. For example, the same biased sensor or biased experimental environments may cause bias error in observed system responses (systematic error); likewise, aleatoric uncertainty exists in measurement sensors (random error). In general, these two types of error occurring in measurement are often disregarded due to lack of information associated with the two types of error. Future studies on statistical model calibration and validation will enable engineers and designers to avoid bypassing the obvious existence of systematic and random errors.
3) There is a strong demand for acquiring uncertainty of dynamic system responses. As discussed in Section 5, system responses are classified into stationary and dynamic responses. The state-of-the-art technique for uncertainty propagation involves obtaining a PDF of the system response with which only the uncertainty of the stationary system responses can be characterized. A PDF has difficulty in acquiring the uncertainty of dynamic system responses. As a result, model analysts are reluctant to conduct statistical model validation in regard to dynamic system responses. However, considering diverse system responses in model validation helps assessing the accuracy of a computational model in a larger prediction domain. Various statistical methodologies on stochastic or random processes are available. Applying those methodologies, further studies can provide calibration and validation methodologies for various kinds of system responses.
4) In the same context, there is a need for a stochastic validation metric or a calibration metric for dynamic system responses. Numerous deterministic validation metrics are available to compare the predicted and the observed dynamic system responses. These metrics are effective in analyzing the difference in important features of a dynamic response but have limitations in considering uncertainties. Future work is needed to develop a stochastic validation metric based on existing deterministic validation metrics.
5) For the inverse problem, variables without enough information are designated as calibration variables. Because building a large-scale model requires the use of numerous inputs for describing the physical reality, there is a great chance that models for practical applications will have many epistemic variables. From an accuracy perspective, an excessive number of calibration variables may result in an undetermined problem or an indeterminate problem. An indeterminate problem is a problem of multiple solutions. In model calibration, an undetermined problem occurs when the number of calibration variables outnumbers the number of constraints. (A system response or a

physical relationship between variables can act as a constraint that restricts the degrees of freedom produced by the calibration variables.) When an undetermined system is formulated, the calibration result can be unstable and inaccurate, depending on the initial points, because infinitely, many optima are possible for calibration variables. Future studies are needed to find a solution to deal with the undetermined problem.

6) For both optimization-based and Bayesian-based approaches, the results of model calibration depend on the preliminary assumption. In optimization-based model calibration, the initial set value of statistical parameters of the assumed probability distribution is calibrated by the optimization process. In Bayesian-based model calibration, the prior distribution is updated with newly observed data. Thus, a wrongly assumed probability distribution can lead to an inaccurate calibration result. Moreover, for optimization-based model calibration, inappropriate setting of initial values for statistical parameters can increase the number of iterations needed to find the optimum. Generally, an early assumption is made upon experience of model developers. A subjective decision by experts is not negligible; it is sometimes even helpful when the data for calibration is insufficient. However, methods highly dependent on assumptions cannot avoid human error. Therefore, future advanced development of the inverse problem, which is objective and less dependent on prior assumption, is required to build highly predictive computational models.

7) In general, local optimization algorithms or gradient-based optimization algorithms have been used to solve the inverse problem for efficiency for conducting optimization-based model calibration. Alternatively, a global optimization algorithm (nongradient or evolutionary) has an increased chance of finding the global optimum, but at a high computational cost. The problem is when the optimization-based calibration problem is formulated as not globally convex. In this case, local optimization algorithms do not ensure that subsequent iterations converge to the global optimum. The global convexity of the calibration problem can be affected by nonlinearity of a computational model, a calibration metric, or a type of assumed probability distribution. Research is needed for a robust methodology with which the global optimum is always found using gradient-based optimization algorithms starting from the initial calibration variables.

8) What should we do if a constructed computational model is revealed to be invalid? As investigated in this paper, numerous model validation methodologies can consider recognized (accounted for) uncertainties. Validation problems can reveal the presence of unrecognized (unaccounted for) uncertainty, which can cause invalidity of a computational model. The sources of unrecognized (unaccounted for) uncertainty are expected to eventually be uncovered, because the process of model validation is not only a process of assessing the accuracy of a computational model but also a process of improving the model based on the validation results. Therefore, there is a need for future work that enables development of a systematic procedure for uncovering the sources of unrecognized (unaccounted for) uncertainties. There has been little work to date on this topic. One recently proposed method for unrecognized (unaccounted for) uncertainty is called model refinement (Liang and Mahadevan 2014; Jung et al. 2014; Oh et al. 2016). Future work on this issue is expected. Additionally, the feedback information derived from the forward, inverse, and validation problems is an important topic for future research.

Funding information This work was supported by a grant from the Institute of Advanced Machinery and Design at Seoul National University (SNU-IAMD) and the Korea Institute of Machinery and Materials (Project No.: NK213E).

Compliance with ethical standards

Conflict of interest The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abbas T, Morgenthal G (2016) Framework for sensitivity and uncertainty quantification in the flutter assessment of bridges. Probab Eng Mech 43:91–105. https://doi.org/10.1016/j.probengmech.2015.12.007

Acar E, Rais-Rohani M (2009) Ensemble of metamodels with optimized weight factors. Struct Multidiscip Optim 37:279–294

Adamowski K (1985) Nonparametric kernel estimation of flood frequencies. Water Resour Res 21:1585–1590

American Society of Mechanical Engineers (2009) Standard for verification and validation in computational fluid dynamics and heat transfer. American Society of Mechanical Engineers, New York City, NY, USA

Anderson MG, Bates PD (2001) Model validation: perspectives in hydrological science. Chichester, West Sussex, England. John Wiley & Sons, Ltd

Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes Ann Math Stat 23:193–212

Anderson TW, Darling DA (1954) A test of goodness of fit. J Am Stat Assoc 49:765–769

Arendt PD, Apley DW, Chen W (2012a) Quantification of model uncertainty: calibration, model discrepancy, and identifiability. J Mech Des 134:100908

Arendt PD, Apley DW, Chen W, Lamb D, Gorsich D (2012b) Improving identifiability in model calibration using multiple responses. J Mech Des 134:100909

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79

Atamturktur S, Hegenderfer J, Williams B, Egeberg M, Lebensohn R, Unal C (2014) A resource allocation framework for experiment-based validation of numerical models. Mech Adv Mater Struct 22:641–654

Au S, Beck JL (1999) A new adaptive importance sampling scheme for reliability calculations. Struct Saf 21:135–158

Babuska I, Oden JT (2004) Verification and validation in computational engineering and science: basic concepts. Comput Methods Appl Mech Eng 193:4057–4066

Babuška I, Nobile F, Tempone R (2008) A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria. Comput Methods Appl Mech Eng 197:2517–2539. https://doi.org/10.1016/j.cma.2007.08.031

Bae H-R, Grandhi RV, Canfield RA (2003) Uncertainty quantification of structural response using evidence theory. AIAA J 41:2062–2068

Bae H-R, Grandhi RV, Canfield RA (2004) An approximation approach for uncertainty quantification using evidence theory. Reliab Eng Syst Saf 86:215–225

Bae HR, Grandhi RV, Canfield RA (2006) Sensitivity analysis of structural response uncertainty propagation using evidence theory. Struct Multidiscip Optim 31:270–279. https://doi.org/10.1007/s00158-006-0606-9

Bao N, Wang C (2015) A Monte Carlo simulation based inverse propagation method for stochastic model updating. Mech Syst Signal Process 61:928–944

Baroni G, Tarantola S (2014) A general probabilistic framework for uncertainty and global sensitivity analysis of deterministic models: a hydrological case study. Environ Model Softw 51:26–34. https://doi.org/10.1016/j.envsoft.2013.09.022

Bayarri MJ et al (2007) A framework for validation of computer models Technometrics 49:138–154

Benek JA, Kraft EM, Lauer RF (1998) Validation issues for engine-airframe integration. AIAA J 36:759–764

Bianchi M (1997) Testing for convergence: evidence from non-parametric multimodality tests. J Appl Econ 12:393–409

Borg A, Paulsen Husted B, Njå O (2014) The concept of validation of numerical models for consequence analysis. Reliab Eng Syst Saf 125:36–45. https://doi.org/10.1016/j.ress.2013.09.009

Box GE, Meyer RD (1986) An analysis for unreplicated fractional factorials. Technometrics 28:11–18

Bucher CG (1988) Adaptive sampling—an iterative fast Monte Carlo procedure. Struct Saf 5:119–126

Buranathiti T, Cao J, Chen W, Baghdasaryan L, Xia ZC (2006) Approaches for model validation: methodology and illustration on a sheet metal flanging process. J Manuf Sci Eng 128:588–597

Cacuci DG, Ionescu-Bujor M (2004) A comparative review of sensitivity and uncertainty analysis of large-scale systems—II: statistical methods. Nucl Sci Eng 147:204–217

Campbell K (2006) Statistical calibration of computer simulations. Reliab Eng Syst Saf 91:1358–1363

Cao R, Cuevas A, Manteiga WG (1994) A comparative study of several smoothing methods in density estimation. Comput Stat Data Anal 17:153–176

Carley H, Taylor M (2002) A new proof of Sklar's theorem. In: Distributions with given marginals and statistical modelling. Springer, pp 29–34

Cha S-H (2007) Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Models Methods Appl Sci 4:300–307

Charnes A, Frome E, Yu P-L (1976) The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. J Am Stat Assoc 71:169–171

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

Chen W, Jin R, Sudjianto A (2005) Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. J Mech Des 127:875–886

Chen W, Xiong Y, Tsui K-L, Wang S (2008) A design-driven validation approach using Bayesian prediction models. J Mech Des 130:021101

Cho H, Bae S, Choi K, Lamb D, Yang R-J (2014) An efficient variable screening method for effective surrogate models for reliability-based design optimization. Struct Multidiscip Optim 50:717–738

Choi K, Youn BD, Yang R-J (2001) Moving least square method for reliability-based design optimization. In: 4th world congress of structural and multidisciplinary optimization. Liaoning Electronic Press, Shenyang, PRC, pp 4–8

Choi J, An D, Won J (2010a) Bayesian approach for structural reliability analysis and optimization using the Kriging Dimension Reduction Method. J Mech Des 132:051003

Choi S-S, Cha S-H, Tappert CC (2010b) A survey of binary similarity and distance measures. J Syst Cybern Inform 8:43–48

Christopher Frey H, Patil SR (2002) Identification and review of sensitivity analysis methods. Risk Anal 22:553–578

Clarke SM, Griebsch JH, Simpson TW (2005) Analysis of support vector regression for approximation of complex engineering analyses. J Mech Des 127:1077–1087

Cooke RM, Goossens LH (2004) Expert judgement elicitation for risk assessments of critical infrastructures. J Risk Res 7:643–656

Crestaux T, Le Maıˆtre O, Martinez J-M (2009) Polynomial chaos expansion for sensitivity analysis. Reliab Eng Syst Saf 94:1161–1172

da Silva Hack P, Schwengber ten Caten C (2012) Measurement uncertainty: literature review and research trends. IEEE Trans Instrum Meas 61:2116–2124. https://doi.org/10.1109/tim.2012.2193694

Dowding KJ, Pilch M, Hills RG (2008) Formulation of the thermal problem. Comput Methods Appl Mech Eng 197:2385–2389

Dubourg V, Sudret B, Deheeger F (2013) Metamodel-based importance sampling for structural reliability analysis. Probab Eng Mech 33:47–57. https://doi.org/10.1016/j.probengmech.2013.02.002

Echard B, Gayton N, Lemaire M, Relun N (2013) A combined importance sampling and kriging reliability method for small failure probabilities with time-demanding numerical models. Reliab Eng Syst Saf 111:232–240. https://doi.org/10.1016/j.ress.2012.10.008

Egeberg M (2014) Optimal design of validation experiments for calibration and validation of complex numerical models, Master Thesis, Clemson University, Clemson, SC, USA

Eldred M, Burkardt J (2009) Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification. In: Proceedings of the 47th AIAA aerospace sciences meeting and exhibit, vol 1. pp 976, Orlando, FL, January 5-9

Eldred MS, Swiler LP, Tang G (2011) Mixed aleatory-epistemic uncertainty quantification with stochastic expansions and optimization-based interval estimation. Reliab Eng Syst Saf 96:1092–1113

Epanechnikov VA (1969) Non-parametric estimation of a multivariate probability density. Theor Probab Appl 14:153–158

Fang S-E, Ren W-X, Perera R (2012) A stochastic model updating method for parameter variability quantification based on response surface models and Monte Carlo simulation. Mech Syst Signal Process 33:83–96

Farajpour I, Atamturktur S (2012) Error and uncertainty analysis of inexact and imprecise computer models. J Comput Civ Eng 27:407–418

Fender J, Duddeck F, Zimmermann M (2014) On the calibration of simplified vehicle crash models. Struct Multidiscip Optim 49:455–469

Ferson S, Ginzburg LR (1996) Different methods are needed to propagate ignorance and variability. Reliabil Eng Syst Saf 54:133–144

Ferson S, Oberkampf WL (2009) Validation of imprecise probability models. Int J Reliab Saf 3:3–22

Ferson S, Joslyn CA, Helton JC, Oberkampf WL, Sentz K (2004) Summary from the epistemic uncertainty workshop: consensus amid diversity. Reliab Eng Syst Saf 85:355–369

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. https://doi.org/10.1016/j.cma.2007.07.030

Fonseca JR, Friswell MI, Mottershead JE, Lees AW (2005) Uncertainty identification by the maximum likelihood method. J Sound Vib 288:587–599

Fu Y, Zhan Z, Yang R-J (2010) A study of model validation method for dynamic systems. SAE 2010-01-0419

Gan Y et al (2014) A comprehensive evaluation of various sensitivity analysis methods: a case study with a hydrological model. Environ Model Softw 51:269–285. https://doi.org/10.1016/j.envsoft.2013.09.031

Geers TL (1984) An objective error measure for the comparison of calculated and measured transient response histories. Shock and Vibration Information Center The Shock and Vibration Bull 54, Pt 2, p 99–108(SEE N 85-18388 09-39)

Genest C, Rivest L-P (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88:1034–1043

Ghanem RG, Doostan A, Red-Horse J (2008) A probabilistic construction of model validation. Comput Methods Appl Mech Eng 197:2585–2595

Giunta A, McFarland J, Swiler L, Eldred M (2006) The promise and peril of uncertainty quantification using response surface approximations. Struct Infrastruct Eng 2:175–189

Goel T, Haftka RT, Shyy W, Queipo NV (2007) Ensemble of surrogates. Struct Multidiscip Optim 33:199–216

Goel T, Hafkta RT, Shyy W (2009) Comparing error estimation measures for polynomial and kriging approximation of noise-free functions. Struct Multidiscip Optim 38:429–442

Gomes HM, Awruch AM (2004) Comparison of response surface and neural network with other methods for structural reliability analysis. Struct Saf 26:49–67

Gorissen D, Couckuyt I, Demeester P, Dhaene T, Crombecq K (2010) A surrogate modeling and adaptive sampling toolbox for computer based design. J Mach Learn Res 11:2051–2055

Govers Y, Link M (2010) Stochastic model updating—covariance matrix adjustment from uncertain experimental modal data. Mech Syst Signal Process 24:696–706. https://doi.org/10.1016/j.ymssp.2009.10.006

Halder A, Bhattacharya R (2011) Model validation: a probabilistic formulation. In: Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on. IEEE, pp 1692–1697

Hanss M, Turrin S (2010) A fuzzy-based approach to comprehensive modeling and analysis of systems with epistemic uncertainties. Struct Saf 32:433–441

Harmel RD, Smith PK, Migliaccio KW (2010) Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. Trans ASABE 53:55–63

Hasselman T, Yap K, Lin C-H, Cafeo J (2005) A case study in model improvement for vehicle crashworthiness simulation. In: Proceedings of the 23rd International Modal Analysis Conference, Orlando, FL, USA, January 31 - February 3

Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intell Syst Appl 13:18–28

Helton JC (1997) Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. J Stat Comput Simul 57:3–76

Helton JC, Burmaster DE (1996) Guest editorial: treatment of aleatory and epistemic uncertainty in performance assessments for complex systems. Reliab Eng Syst Saf 54:91–94

Helton JC, Davis FJ (2003) Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab Eng Syst Saf 81:23–69

Helton JC, Oberkampf WL (2004) Alternative representations of epistemic uncertainty. Reliab Eng Syst Saf 85:1–10. https://doi.org/10.1016/j.ress.2004.03.001

Helton JC, Johnson JD, Oberkampf WL (2004) An exploration of alternative approaches to the representation of uncertainty in model predictions. Reliab Eng Syst Saf 85:39–71. https://doi.org/10.1016/j.ress.2004.03.025

Helton JC, Davis F, Johnson JD (2005) A comparison of uncertainty and sensitivity analysis results obtained with random and Latin hypercube sampling. Reliab Eng Syst Saf 89:305–330

Helton JC, Johnson JD, Oberkampf W, Sallaberry CJ (2006a) Sensitivity analysis in conjunction with evidence theory representations of epistemic uncertainty. Reliab Eng Syst Saf 91:1414–1434

Helton JC, Johnson JD, Sallaberry CJ, Storlie CB (2006b) Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab Eng Syst Saf 91:1175–1209. https://doi.org/10.1016/j.ress.2005.11.017

Helton JC, Johnson JD, Oberkampf WL, Sallaberry CJ (2010) Representation of analysis results involving aleatory and epistemic uncertainty. Int J Gen Syst 39:605–646

Hemez F, Atamturktur HS, Unal C (2010) Defining predictive maturity for validated numerical simulations. Comput Struct 88:497–505

Higdon D, Nakhleh C, Gattiker J, Williams B (2008) A Bayesian calibration approach to the thermal problem. Comput Methods Appl Mech Eng 197:2431–2441

Hill WJ, Hunter WG (1966) A review of response surface methodology: a literature survey. Technometrics 8:571–590

Hills RG (2006) Model validation: model parameter and measurement uncertainty. J Heat Transf 128:339. https://doi.org/10.1115/1.2164849

Hills RG, Pilch M, Dowding KJ, Red-Horse J, Paez TL, Babuška I, Tempone R (2008) Validation challenge workshop. Comput Methods Appl Mech Eng 197:2375–2380. https://doi.org/10.1016/j.cma.2007.10.016

Hoffman FO, Hammonds JS (1994) Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. Risk Anal 14:707–712

Homma T, Saltelli A (1996) Importance measures in global sensitivity analysis of nonlinear models. Reliab Eng Syst Saf 52:1–17

Hu C, Youn BD (2011) Adaptive-sparse polynomial chaos expansion for reliability analysis and design of complex engineering systems. Struct Multidiscip Optim 43:419–442

Hu Z, Hu C, Mourelatos ZP, Mahadevan S (2019) Model discrepancy quantification in simulation-based design of dynamical systems. J Mech Des 141:011401

Hua X, Ni Y, Chen Z, Ko J (2008) An improved perturbation method for stochastic finite element model updating. Int J Numer Methods Eng 73:1845–1864

Huntington D, Lyrintzis C (1998) Improvements to and limitations of Latin hypercube sampling. Probab Eng Mech 13:245–253

Hurtado J, Barbat AH (1998) Monte Carlo techniques in computational stochastic mechanics. Arch Comput Methods Eng 5:3–29

Jiang X, Mahadevan S (2007) Bayesian risk-based decision method for model validation under uncertainty. Reliab Eng Syst Saf 92:707–718

Jiang X, Mahadevan S (2008) Bayesian validation assessment of multivariate computational models. J Appl Stat 35:49–65

Jiang X, Mahadevan S (2009a) Bayesian inference method for model validation and confidence extrapolation. J Appl Stat 36:659–677

Jiang X, Mahadevan S (2009b) Bayesian structural equation modeling method for hierarchical model validation. Reliab Eng Syst Saf 94:796–809. https://doi.org/10.1016/j.ress.2008.08.008

Jiang X, Mahadevan S (2011) Wavelet spectrum analysis approach to model validation of dynamic systems. Mech Syst Signal Process 25:575–590

Jiang X, Yuan Y, Mahadevan S, Liu X (2013a) An investigation of Bayesian inference approach to model validation with non-normal data. J Stat Comput Simul 83:1829–1851

Jiang Z, Chen W, Fu Y, Yang R-J (2013b) Reliability-based design optimization with model bias and data uncertainty. SAE Int J Mater Manuf 6:502–516

Joe H (1990) Families of min-stable multivariate exponential and multivariate extreme value distributions. Stat Probab lett 9:75–81

Joe H, Hu T (1996) Multivariate distributions from mixtures of max-infinitely divisible distributions. J Multivar Anal 57:240–265

Johnson NL (1949) Systems of frequency curves generated by methods of translation. Biometrika 36:149–176

Jung BC, Lee D-H, Youn BD (2009) Optimal design of constrained-layer damping structures considering material and operational condition variability. AIAA J 47:2985–2995

Jung BC, Lee D, Youn BD, Lee S (2011) A statistical characterization method for damping material properties and its application to structural-acoustic system design. J Mech Sci Technol 25:1893–1904

Jung BC, Park J, Oh H, Kim J, Youn BD (2014) A framework of model validation and virtual product qualification with limited experimental data based on statistical inference. Struct Multidiscip Optim 51:573–583

Jung BC, Yoon H, Oh H, Lee G, Yoo M, Youn BD, Huh YC (2016) Hierarchical model calibration for designing piezoelectric energy harvester in the presence of variability in material properties and geometry. Struct Multidiscip Optim 53:161–173

Kang S-C, Koh H-M, Choo JF (2010) An efficient response surface method using moving least squares approximation for structural reliability analysis. Probab Eng Mech 25:365–371

Kat C-J, Els PS (2012) Validation metric based on relative error. Math Comput Model Dyn Syst 18:487–520

Kaymaz I (2005) Application of kriging method to structural reliability problems. Struct Saf 27:133–151

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc B 63(3):425–464

Kersaudy P, Sudret B, Varsier N, Picon O, Wiart J (2015) A new surrogate modeling technique combining kriging and polynomial chaos expansions – application to uncertainty analysis in computational dosimetry. J Comput Phys 286:103–117. https://doi.org/10.1016/j.jcp.2015.01.034

Khodaparast HH, Mottershead JE, Friswell MI (2008) Perturbation methods for the estimation of parameter variability in stochastic model updating. Mech Syst Signal Process 22:1751–1773. https://doi.org/10.1016/j.ymssp.2008.03.001

Khodaparast HH, Mottershead JE, Badcock KJ (2011) Interval model updating with irreducible uncertainty using the kriging predictor. Mech Syst Signal Process 25:1204–1226. https://doi.org/10.1016/j.ymssp.2010.10.009

Kim T, Lee G, Youn BD (2018) Uncertainty characterization under measurement errors using maximum likelihood estimation: cantilever beam end-to-end UQ test problem Struct Multidiscip Optim 59: 323–333

Kleijnen JPC, Mehdad E (2014) Multivariate versus univariate Kriging metamodels for multi-response simulation models. Eur J Oper Res 236:573–582. https://doi.org/10.1016/j.ejor.2014.02.001

Kokkolaras M, Hulbert G, Papalambros P, Mourelatos Z, Yang R, Brudnak M, Gorsich D (2013) Towards a comprehensive framework for simulation-based design validation of vehicle systems. Int J Veh Des 61:233–248

Kutluay E, Winner H (2014) Validation of vehicle dynamics simulation models – a review. Veh Syst Dyn 52:186–200. https://doi.org/10.1080/00423114.2013.868500

Kwaśniewski L (2009) On practical problems with verification and validation of computational models. Arch Civ Eng 55:323–346

Lancaster P, Salkauskas K (1981) Surfaces generated by moving least squares methods. Math Comput 37:141–158

Lee SH, Chen W (2009) A comparative study of uncertainty propagation methods for black-box-type problems. Struct Multidiscip Optim 37: 239–253

Lee JH, Gard K (2014) Vehicle–soil interaction: testing, modeling, calibration and validation. J Terrramech 52:9–21. https://doi.org/10.1016/j.jterra.2013.12.001

Lee I, Choi KK, Du L, Gorsich D (2008a) Dimension reduction method for reliability-based robust design optimization. Comput Struct 86: 1550–1562. https://doi.org/10.1016/j.compstruc.2007.05.020

Lee I, Choi KK, Du L, Gorsich D (2008b) Inverse analysis method using MPP-based dimension reduction for reliability-based design optimization of nonlinear and multi-dimensional systems. Comput Methods Appl Mech Eng 198:14–27. https://doi.org/10.1016/j.cma.2008.03.004

Lee I, Choi K, Gorsich D (2010) System reliability-based design optimization using the MPP-based dimension reduction method. Struct Multidiscip Optim 41:823–839

Lee I, Choi K, Zhao L (2011) Sampling-based RBDO using the stochastic sensitivity analysis and dynamic kriging method. Struct Multidiscip Optim 44:299–317

Lee T-R, Greene MS, Jiang Z, Kopacz AM, Decuzzi P, Chen W, Liu WK (2014) Quantifying uncertainties in the microvascular transport of nanoparticles. Biomech Model Mechanobiol 13:515–526

Lee G, Yi G, Youn BD (2018) Special issue: a comprehensive study on enhanced optimization-based model calibration using gradient information. Struct Multidiscip Optim 57:2005–2025

Levin D (1998) The approximation power of moving least-squares. Math Comput Am Math Soc 67:1517–1531

Li C, Mahadevan S (2016) Role of calibration, validation, and relevance in multi-level uncertainty integration. Reliab Eng Syst Saf 148:32–43. https://doi.org/10.1016/j.ress.2015.11.013

Li W, Chen W, Jiang Z, Lu Z, Liu Y (2014) New validation metrics for models with multiple correlated responses. Reliab Eng Syst Saf 127: 1–11. https://doi.org/10.1016/j.ress.2014.02.002

Liang B, Mahadevan S (2011) Error and uncertainty quantification and sensitivity analysis in mechanics computational models Int J Uncertain Quantif 11:147–161

Liang C, Mahadevan S (2014) Bayesian framework for multidisciplinary uncertainty quantification and optimization. In: SciTech, 16th AIAA Non-Deterministic Approaches Conference, pp 2014-1347

Liang C, Mahadevan S, Sankararaman S (2015) Stochastic multidisciplinary analysis under epistemic uncertainty. J Mech Des 137:021404

Lima Azevedo C, Ciuffo B, Cardoso JL, Ben-Akiva ME (2015) Dealing with uncertainty in detailed calibration of traffic simulation models for safety assessment. Transp Res Part C: Emerg Technol 58:395–412. https://doi.org/10.1016/j.trc.2015.01.029

Ling Y, Mahadevan S (2013) Quantitative model validation techniques: new insights. Reliab Eng Syst Saf 111:217–231

Liu F, Bayarri M, Berger J, Paulo R, Sacks J (2008) A Bayesian analysis of the thermal challenge problem. Comput Methods Appl Mech Eng 197:2457–2466

Liu Y, Chen W, Arendt P, Huang H-Z (2011) Toward a better understanding of model validation metrics. J Mech Des 133:071005

Mahadevan S, Rebba R (2005) Validation of reliability computational models using Bayes networks. Reliab Eng Syst Saf 87:223–232

Manfren M, Aste N, Moshksar R (2013) Calibration and uncertainty analysis for computer models – a meta-model based approach for integrated building energy simulation. Appl Energy 103:627–641

Mara TA, Tarantola S (2012) Variance-based sensitivity indices for models with dependent inputs. Reliab Eng Syst Saf 107:115–121. https://doi.org/10.1016/j.ress.2011.08.008

Mares C, Mottershead JE, Friswell MI (2006) Stochastic model updating: part 1—theory and simulated example. Mech Syst Signal Process 20:1674–1695. https://doi.org/10.1016/j.ymssp.2005.06.006

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

McCusker JR, Danai K, Kazmer DO (2010) Validation of dynamic models in the time-scale domain. J Dyn Syst Meas Control 132: 061402. https://doi.org/10.1115/1.4002479

McFarland J, Mahadevan S (2008a) Error and variability characterization in structural dynamics modeling. Comput Methods Appl Mech Eng 197:2621–2631. https://doi.org/10.1016/j.cma.2007.07.029

McFarland J, Mahadevan S (2008b) Multivariate significance testing and model calibration under uncertainty. Comput Methods Appl Mech Eng 197:2467–2479

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

McKay M, Meyer M (2000) Critique of and limitations on the use of expert judgements in accident consequence uncertainty analysis. Radiat Prot Dosim 90:325–330

McNeil AJ (2008) Sampling nested Archimedean copulas. J Stat Comput Simul 78:567–581

McNeil AJ, Nešlehová J (2009) Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions Ann Stat 37: 3059–3097

Melchers R (1989) Importance sampling in structural systems. Struct Saf 6:3–10

Mongiardini M, Ray MH, Anghileri M. Development of a software for the comparison of curves during the verification and validation of numerical models. In: Proceedings of the 7th European LS-DYNA Conference, Salzburg, 2009

Mongiardini M, Ray MH, Anghileri M (2010) Acceptance criteria for validation metrics in roadside safety based on repeated full-scale crash tests. Int J Reliab Saf 4:69–88

Mongiardini M, Ray M, Plaxico C (2013) Development of a programme for the quantitative comparison of a pair of curves. Int J Comput Appl Technol 46:128–141

Montgomery DC (2008) Design and analysis of experiments. John Wiley & Sons

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

Moon M-Y, Cho H, Choi K, Gaul N, Lamb D, Gorsich D (2018) Confidence-based reliability assessment considering limited numbers of both input and output test data. Struct Multidiscip Optim 57:2027–2043

Moore R, Lodwick W (2003) Interval analysis and fuzzy set theory. Fuzzy Sets Syst 135:5–9

Mosegaard K, Sambridge M (2002) Monte Carlo analysis of inverse problems. Inverse Probl 18:29–54

Mousaviraad SM, He W, Diez M, Stern F (2013) Framework for convergence and validation of stochastic uncertainty quantification and relationship to deterministic verification and validation. Int J Uncertain Quantif 3:371–395

Mullins J, Ling Y, Mahadevan S, Sun L, Strachan A (2016) Separation of aleatory and epistemic uncertainty in probabilistic model validation. Reliab Eng Syst Saf 147:49–59. https://doi.org/10.1016/j.ress.2015.10.003

Murmann R, Harzheim L, Dominico S, Immel R (2016) CoSi: correlation of signals—a new measure to assess the correlation of history response curves. Mech Syst Signal Process 80:482–502

Murray_smith D (1998) Methods for the external validation of continuous system simulation models: a review. Math Comput Model Dyn Syst 4: 5–31

Myers RH, Montgomery DC, Anderson-Cook CM (1995) Response surface methodology: process and product optimization using designed experiments. Hoboken, NJ, USA. John Wiley & Sons, Inc

Myers RH, Montgomery DC, Anderson-Cook CM (2016) Response surface methodology: process and product optimization using designed experiments. John Wiley & Sons

Myung IJ (2003) Tutorial on maximum likelihood estimation. J Math Psychol 47:90–100

Najm HN (2009) Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics. Annu Rev Fluid Mech 41: 35–52. https://doi.org/10.1146/annurev.fluid.010908.165248

Nelsen RB (2002) Concordance and copulas: a survey. In: Distributions with given marginals and statistical modelling. Springer, pp 169–177

Newey WK, West KD (1987) Hypothesis testing with efficient method of moments estimation. Int Econ Rev 28:777–787

Oberkampf WL, Barone MF (2006) Measures of agreement between computation and experiment: validation metrics. J Comput Phys 217:5–36. https://doi.org/10.1016/j.jcp.2006.03.037

Oberkampf WL, Trucano TG (2002) Verification and validation in computational fluid dynamics. Prog Aerosp Sci 38:209–272

Oberkampf WL, Trucano TG (2008) Verification and validation benchmarks. Nucl Eng Des 238:716–743. https://doi.org/10.1016/j.nucengdes.2007.02.032

Oberkampf WL, DeLand SM, Rutherford BM, Diegert KV, Alvin KF (2002) Error and uncertainty in modeling and simulation. Reliab Eng Syst Saf 75:333–357

Oberkampf WL, Helton JC, Joslyn CA, Wojtkiewicz SF, Ferson S (2004a) Challenge problems: uncertainty in system response given uncertain parameters. Reliab Eng Syst Saf 85:11–19

Oberkampf WL, Trucano TG, Hirsch C (2004b) Verification, validation, and predictive capability in computational engineering and physics. Appl Mech Rev 57:345–384

Oden JT, Prudencio EE, Bauman PT (2013) Virtual model validation of complex multiscale systems: applications to nonlinear elastostatics. Comput Methods Appl Mech Eng 266:162–184. https://doi.org/10.1016/j.cma.2013.07.011

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

Oladyshkin S, Nowak W (2012) Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. Reliab Eng Syst Saf 106:179–190. https://doi.org/10.1016/j.ress.2012.05.002

Oliver TA, Terejanu G, Simmons CS, Moser RD (2015) Validating predictions of unobserved quantities. Comput Methods Appl Mech Eng 283:1310–1335. https://doi.org/10.1016/j.cma.2014.08.023

Pan H, Xi Z, Yang R-J (2016) Model uncertainty approximation using a copula-based approach for reliability based design optimization. Struct Multidiscip Optim 54:1543–1556

Panchenko V (2005) Goodness-of-fit test for copulas. Phys A: Stat Mech Appl 355:176–182

Park I, Grandhi RV (2014) A Bayesian statistical method for quantifying model form uncertainty and two model combination methods. Reliab Eng Syst Saf 129:46–56

Park B, Turlach B (1992) Practical performance of several data driven bandwidth selectors. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE)

Park I, Amarchinta HK, Grandhi RV (2010) A Bayesian approach for quantification of model uncertainty. Reliab Eng Syst Saf 95:777–785. https://doi.org/10.1016/j.ress.2010.02.015

Park C, Kim NH, Haftka RT (2015) The effect of ignoring dependence between failure modes on evaluating system reliability. Struct Multidiscip Optim 52:251–268

Park C, Choi J-H, Haftka RT (2016a) Teaching a verification and validation course using simulations and experiments with paper helicopters. J Verif Valid Uncertain Quantif 1:031002

Park C, Haftka RT, Kim NH (2016b) Remarks on multi-fidelity surrogates Struct Multidiscip Optim 55:1029–1050

Park C, Haftka RT, Kim NH (2017) Simple alternative to Bayesian multi-fidelity surrogate framework. In: 58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, p 0135

Parry GW (1996) The characterization of uncertainty in probabilistic risk assessments of complex systems. Reliab Eng Syst Saf 54:119–126

Pettit CL (2004) Uncertainty quantification in aeroelasticity: recent results and research challenges. J Aircr 41:1217–1229. https://doi.org/10.2514/1.3961

Pianosi F, Beven K, Freer J, Hall JW, Rougier J, Stephenson DB, Wagener T (2016) Sensitivity analysis of environmental models: a systematic review with practical workflow. Environ Model Softw 79:214–232. https://doi.org/10.1016/j.envsoft.2016.02.008

Plackett RL (1983) Karl Pearson and the chi-squared test. Int Stat Rev 51: 59–72

Plackett RL, Burman JP (1946) The design of optimum multifactorial experiments. Biometrika 33:305–325

Plischke E, Borgonovo E, Smith CL (2013) Global sensitivity measures from given data. Eur J Oper Res 226:536–550

Pradlwarter HJ, Schuëller GI (2008) The use of kernel densities and confidence intervals to cope with insufficient data in validation experiments. Comput Methods Appl Mech Eng 197:2550–2560. https://doi.org/10.1016/j.cma.2007.09.028

Qian PZ, Wu CJ (2008) Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. Technometrics 50:192–204

Rahman S, Xu H (2004) A univariate dimension-reduction method for multi-dimensional integration in stochastic mechanics. Probab Eng Mech 19:393–408

Rebba R, Mahadevan S (2006) Validation of models with multivariate output. Reliab Eng Syst Saf 91:861–871

Rebba R, Mahadevan S (2008) Computational methods for model reliability assessment. Reliab Eng Syst Saf 93:1197–1207

Rebba R, Huang S, Liu Y, Mahadevan S (2005) Statistical validation of simulation models. Int J Mater Prod Technol 25:164–181

Rebba R, Mahadevan S, Huang S (2006) Validation and error estimation of computational models. Reliab Eng Syst Saf 91:1390–1397

Romero V (2019) Real-space model validation and predictor-corrector extrapolation applied to the Sandia cantilever beam end-to-end UQ problem, Paper AIAA-2019-1488, 21st AIAA Non-Deterministic Approaches Conference, AIAA SciTech, Jan. 7–11, San Diego, CA

Roussouly N, Petitjean F, Salaun M (2013) A new adaptive response surface method for reliability analysis. Probab Eng Mech 32:103–115

Roy CJ, Oberkampf WL (2010) A complete framework for verification, validation, and uncertainty quantification in scientific computing. In: 48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, 4–7

Roy CJ, Oberkampf WL (2011) A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. Comput Methods Appl Mech Eng 200:2131–2144

Rui Q, Ouyang H, Wang H (2013) An efficient statistically equivalent reduced method on stochastic model updating. Appl Math Model 37:6079–6096

Rüschendorf L (2009) On the distributional transform, Sklar's theorem, and the empirical copula process. J Stat Plann Infer 139:3921–3927

Russell DM (1997) Error measures for comparing transient data: part I: development of a comprehensive error measure. In: Proceedings of the 68th Shock and Vibration Symposium, Hunt Valley. pp 175–184

Salehghaffari S, Rais-Rohani M (2013) Material model uncertainty quantification using evidence theory. Proc Inst Mech Eng C J Mech Eng Sci 227:2165–2181

Sankararaman S, Mahadevan S (2011a) Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data. Reliab Eng Syst Saf 96:814–824

Sankararaman S, Mahadevan S (2011b) Model validation under epistemic uncertainty. Reliab Eng Syst Saf 96:1232–1241

Sankararaman S, Mahadevan S (2013) Separating the contributions of variability and parameter uncertainty in probability distributions. Reliab Eng Syst Saf 112:187–199. https://doi.org/10.1016/j.ress.2012.11.024

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

Sankararaman S, Ling Y, Mahadevan S (2011) Uncertainty quantification and model validation of fatigue crack growth prediction. Eng Fract Mech 78:1487–1504

Sankararaman S, McLemore K, Mahadevan S, Bradford SC, Peterson LD (2013) Test resource allocation in hierarchical systems using Bayesian networks. AIAA J 51:537–550

Sargent RG (2013) Verification and validation of simulation models. J Simul 7:12–24

Sarin H, Kokkolaras M, Hulbert G, Papalambros P, Barbat S, Yang R-J (2010) Comparing time histories for validation of simulation models: error measures and metrics. J Dyn Syst Meas Control 132:061401

Savu C, Trede M (2010) Hierarchies of Archimedean copulas. Quant Finan 10:295–304

Scholz F (1985) Maximum likelihood estimation. Encycl Stat Sci 5:340–351

Schwer LE (2007) Validation metrics for response histories: perspectives and case studies. Eng Comput 23:295–309

Shah H, Hosder S, Winter T (2015) Quantification of margins and mixed uncertainties using evidence theory and stochastic expansions. Reliab Eng Syst Saf 138:59–72. https://doi.org/10.1016/j.ress.2015.01.012

Shan S, Wang GG (2010) Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. Struct Multidiscip Optim 41:219–241

Shi L, Yang R, Zhu P (2012) A method for selecting surrogate models in crashworthiness optimization. Struct Multidiscip Optim 46:159–170

Shields MD, Zhang J (2016) The generalization of Latin hypercube sampling. Reliab Eng Syst Saf 148:96–108. https://doi.org/10.1016/j.ress.2015.12.002

Shields MD, Teferra K, Hapij A, Daddazio RP (2015) Refined stratified sampling for efficient Monte Carlo based uncertainty quantification. Reliab Eng Syst Saf 142:310–325. https://doi.org/10.1016/j.ress.2015.05.023

Silva AS, Ghisi E (2014) Uncertainty analysis of the computer model in building performance simulation. Energy and Buildings 76:258–269. https://doi.org/10.1016/j.enbuild.2014.02.070

Simpson TW, Korte JJ, Mauery TM, Mistree F (1998) Comparison of response surface and kriging models for multidisciplinary design optimization. In: Proceedings of the 7th AIAA/USAF/ NASA/ ISSMO Symp. vol 1, pp 381–391, St. Louis, MO, USA, September 2-4

Simpson TW, Mauery TM, Korte JJ, Mistree F (2001a) Kriging models for global approximation in simulation-based multidisciplinary design optimization. AIAA J 39:2233–2241

Simpson TW, Poplinski J, Koch PN, Allen JK (2001b) Metamodels for computer-based engineering design: survey and recommendations. Eng Comput 17:129–150

Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions. Ann Math Stat 19:279–281

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222

Sobol IM (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 55: 271–280

Sobol' IM (1990) On sensitivity estimation for nonlinear mathematical models. Math Model Comput Simul 2:112–118

Solomon H, Stephens MA (1978) Approximations to density functions using Pearson curves. J Am Stat Assoc 73:153–160

Sprague M, Geers T (2004) A spectral-element method for modelling cavitation in transient fluid-structure interaction. Int J Numer Methods Eng 60:2467–2499

Stein M (1987) Large sample properties of simulations using Latin hypercube sampling. Technometrics 29:143–151

Steinberg DM, Hunter WG (1984) Experimental design: review and comment. Technometrics 26:71–97

Stephens MA (1974) EDF statistics for goodness of fit and some comparisons. J Am Stat Assoc 69:730–737

Sudret B (2008) Global sensitivity analysis using polynomial chaos expansions. Reliab Eng Syst Saf 93:964–979

Swiler LP, Paez TL, Mayes RL (2009) Epistemic uncertainty quantification tutorial. In: Proceedings of the 27th International Modal Analysis Conference, Orlando, FL, USA, February 9-12

Tabatabaei M, Hakanen J, Hartikainen M, Miettinen K, Sindhya K (2015) A survey on handling computationally expensive multiobjective optimization problems using surrogates: non-nature inspired methods. Struct Multidiscip Optim 52:1–25

Teferra K, Shields MD, Hapij A, Daddazio RP (2014) Mapping model validation metrics to subject matter expert scores for model adequacy assessment. Reliab Eng Syst Saf 132:9–19. https://doi.org/10.1016/j.ress.2014.07.010

Thacker BH, Paez TL (2014) A simple probabilistic validation metric for the comparison of uncertain model and test results. In: Proceedings of the 16th AIAA Non-Deterministic Approaches Conference, National Harbor, MD, USA, January 13-17

Thorne M, Williams MMR (1992) A review of expert judgment techniques with reference to nuclear safety. Prog Nucl Energy 27:83–254

Trucano TG, Swiler LP, Igusa T, Oberkampf WL, Pilch M (2006) Calibration, validation, and sensitivity analysis: what's what. Reliab Eng Syst Saf 91:1331–1357. https://doi.org/10.1016/j.ress.2005.11.031

Twisk D, Spit H, Beebe M, Depinet P (2007) Effect of dummy repeatability on numerical model accuracy. SAE 2007-01-1173

Tyssedal J (2008) Plackett–Burman designs. Encycl. Stat in Qual and Reliab, New York, John Wiley & Sons, Inc, 1:1361–1125. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470061572

Urbina A, Mahadevan S, Paez TL (2011) Quantification of margins and uncertainties of complex systems in the presence of aleatoric and epistemic uncertainty. Reliab Eng Syst Saf 96:1114–1125

Uusitalo L, Lehikoinen A, Helle I, Myrberg K (2015) An overview of methods to evaluate uncertainty of deterministic models in decision support. Environ Model Softw 63:24–31. https://doi.org/10.1016/j.envsoft.2014.09.017

Viana FA, Haftka RT, Steffen V Jr (2009) Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. Struct Multidiscip Optim 39:439–457

Viana FAC, Pecheny V, Haftka RT (2010) Using cross validation to design conservative surrogates. AIAA J 48:2286–2298

Viana FA, Simpson TW, Balabanov V, Toropov V (2014) Metamodeling in multidisciplinary design optimization: how far have we really come? AIAA J 52:670–690

Voyles IT, Roy CJ (2015) Evaluation of model validation techniques in the presence of aleatory and epistemic input uncertainties. In: 17th AIAA Non-Deterministic Approaches Conference. ARC, pp 1–16

Warner JE, Aquino W, Grigoriu MD (2015) Stochastic reduced order models for inverse problems under uncertainty. Comput Methods Appl Mech Eng 285:488–514. https://doi.org/10.1016/j.cma.2014.11.021

Weathers JB, Luck R, Weathers JW (2009) An exercise in model validation: comparing univariate statistics and Monte Carlo-based multivariate statistics. Reliab Eng Syst Saf 94:1695–1702. https://doi.org/10.1016/j.ress.2009.04.007

Wei D, Cui Z, Chen J (2008) Uncertainty quantification using polynomial chaos expansion with points of monomial cubature rules. Comput Struct 86:2102–2108

Wei P, Lu Z, Song J (2015) Variable importance analysis: a comprehensive review. Reliab Eng Syst Saf 142:399–432. https://doi.org/10.1016/j.ress.2015.05.018

Whang B, Gilbert WE, Zilliacus S (1994) Two visually meaningful correlation measures for comparing calculated and measured response histories. Shock Vib 1:303–316

Willmott CJ, Robeson SM, Matsuura K (2012) A refined index of model performance. Int J Climatol 32:2088–2094

Winkler RL (1996) Uncertainty in probabilistic risk assessment. Reliab Eng Syst Saf 54:127–132

Wojtkiewicz S, Eldred M, Field R, Urbina A, Red-Horse J (2001) In: Proceedings of the 42rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, pp 1455, Seattle, WA, USA, April 16-19

Wu YT, Mohanty S (2006) Variable screening and ranking using sampling-based sensitivity measures. Reliab Eng Syst Saf 91:634–647. https://doi.org/10.1016/j.ress.2005.05.004

Wu J, Apostolakis G, Okrent D (1990) Uncertainties in system analysis: probabilistic versus nonprobabilistic theories. Reliab Eng Syst Saf 30:163–181

Xi Z, Fu Y, Yang RJ (2013) Model bias characterization in the design space under uncertainty. Int J Perform Eng 9:433

Xi Z, Pan H, Fu Y, Yang R-J (2015) Validation metric for dynamic system responses under uncertainty. SAE Int J Mater Manuf 8:309–314

Xie K, Wells L, Camelio JA, Youn BD (2007) Variation propagation analysis on compliant assemblies considering contact interaction. J Manuf Sci Eng 129:934–942

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. https://doi.org/10.1016/j.cma.2008.11.023

Xu H, Rahman S (2004) A generalized dimension-reduction method for multidimensional integration in stochastic mechanics. Int J Numer Methods Eng 61:1992–2019

Yates F (1934) Contingency tables involving small numbers and the $\chi 2$ test. Suppl J R Stat Soc 1:217–235

Youn BD, Choi KK (2004) An investigation of nonlinearity of reliability-based design optimization approaches. J Mech Des 126:403. https://doi.org/10.1115/1.1701880

Youn BD, Wang P (2008) Bayesian reliability-based design optimization using eigenvector dimension reduction (EDR) method. Struct Multidiscip Optim 36:107–123

Youn BD, Wang P (2009) Complementary intersection method for system reliability analysis. J Mech Des 131:041004

Youn BD, Xi Z (2009) Reliability-based robust design optimization using the eigenvector dimension reduction (EDR) method. Struct Multidiscip Optim 37:475–492

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

Yuan J, Ng SH, Tsui KL (2013) Calibration of stochastic computer models using stochastic approximation methods. Autom Sci Eng IEEE Trans 10:171–186

Zambom AZ, Dias R (2012) A review of kernel density estimation with applications to econometrics arXiv preprint arXiv:12122812

Zárate BA, Caicedo JM (2008) Finite element model updating: multiple alternatives. Eng Struct 30:3724–3730. https://doi.org/10.1016/j.engstruct.2008.06.012

Zhan Z, Fu Y, Yang R-J (2011a) Enhanced error assessment of response time histories (EEARTH) metric and calibration process. SAE 2011-01-0245

Zhan Z, Fu Y, Yang R-J, Peng Y (2011b) An automatic model calibration method for occupant restraint systems. Struct Multidiscip Optim 44:815–822

Zhan Z, Fu Y, Yang R-J, Peng Y (2011c) An enhanced Bayesian based model validation method for dynamic systems. J Mech Des 133:041005

Zhan Z-f, Hu J, Fu Y, Yang R-J, Peng Y-h, Qi J (2012a) Multivariate error assessment of response time histories method for dynamic systems. J Zhejiang Univ Sci A 13:121–131

Zhan Z, Fu Y, Yang R-J, Peng Y (2012b) Bayesian based multivariate model validation method under uncertainty for dynamic systems. J Mech Des 134:034502

Zhang J, Du X (2010) A second-order reliability method with first-order efficiency. J Mech Des 132:101006

Zhang X, Pandey MD (2014) An effective approximation for variance-based global sensitivity analysis. Reliab Eng Syst Saf 121:164–174

Zhang S, Zhu P, Chen W, Arendt P (2013) Concurrent treatment of parametric uncertainty and metamodeling uncertainty in robust design. Struct Multidiscip Optim 47:63–76

Zhao Y-G, Ono T (2000) New point estimates for probability moments. J Eng Mech 126:433–436

Zhu S-P, Huang H-Z, Peng W, Wang H-K, Mahadevan S (2016) Probabilistic physics of failure-based framework for fatigue life prediction of aircraft gas turbine discs under uncertainty. Reliab Eng Syst Saf 146:1–12. https://doi.org/10.1016/j.ress.2015.10.002