Structural and Multidisciplinary Optimization (2019) 60:1171–1187
https://doi.org/10.1007/s00158-019-02288-6

RESEARCH PAPER

A robust and convex metric for unconstrained optimization in statistical model calibration—probability residual (PR)

Hyunseok Oh[1] · Hwanoh Choi[2] · Joon Ha Jung[3] · Byeng D. Youn[3,4,5]

Received: 2 January 2019 / Revised: 20 March 2019 / Accepted: 15 April 2019 / Published online: 6 May 2019

Abstract
Statistical model calibration is a practical tool for computational model development processes. However, in optimization-based model calibration, the quality of the calibrated model is often unsatisfactory due to inefficiency and/or inaccuracy of calibration metrics. This paper proposes a new calibration metric, namely, probability residual (PR). PR quantifies the degree of agreement or disagreement between the computational response and experimental results. The PR metric is defined as the sum of the product of a scale factor and the squared residual. First, the scale factor defines the shape of the squared residual to maintain consistent sensitivity during the optimization process. Thus, the number of function evaluations can be reduced. Second, the mathematical form of the squared residuals is used to make convex optimization feasible. Therefore, the existence of a global minimum is guaranteed. To evaluate the performance of the proposed metric, numerical examples are shown in a case study. Various system functions—including linear, non-linear, and elliptical—are incorporated into the statistical model calibration. A case study that examines journal bearing rotor systems is presented to demonstrate the application of the proposed calibration metric to a real-world engineered system.

Keywords Computational model · Statistical model calibration · Calibration metric · Validity check · Journal bearing rotor system

1 Introduction

Computational models have been widely adopted for virtual testing of engineered systems. Virtual testing can reduce the costs related to physical testing. However, it is often observed that simulation results do not agree with observations from actual tests. This is a serious concern for both modelers and experimenters. To eliminate the disagreement between simulation results and experimental observations, model calibration (or model updating, parameter estimation) techniques have been considered as a practical and useful tool for use in the model development process (AIAA 1998; ASME 2006). Model calibration is defined as the process that adjusts unknown model parameters in the computational model to enhance the model's agreement with experimental observations. Model calibration relies on mathematical means to match simulation results with experimental observations, while model refinement changes physical principles in models or uses other means to improve an invalid model (Oh et al. 2016a). When model calibration is conducted correctly, an accurate computational model can be built efficiently.

Model calibration can be accomplished in a deterministic manner (Trucano et al. 2006). For example, an objective function can be formulated to quantify the disagreement (or agreement) between simulation results and experimental observations. A set of model parameters can be found by minimizing (or maximizing) the objective function. Recently, deterministic model calibration has been employed to develop simulation models, such as anisotropic shear deformable plate models for molecular dynamics. In (Sahmani and Fattahi 2017), individual model parameters were assumed to be a single value, rather than random variables. However, in

Responsible Editor: Felipe A. C. Viana

✉ Byeng D. Youn
bdyoun@snu.ac.kr

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

[2] C4 Advanced Technology Team, Productivity Research Institute, LG Electronics, Pyeongtaek 17709, South Korea

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

[4] Institute of Advanced Machines and Design, Seoul National University, Seoul 08826, South Korea

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

reality, the assumption that model parameters are a fixed value is often violated when the inherent randomness or variability in material properties, product geometries, and/or loading conditions is too large to be neglected.

Uncertainties should be considered in physical systems and computational models. To address this challenge, a Bayesian model calibration approach was developed. In the Bayesian technique by Kennedy and O'Hagan (Kennedy and O'Hagan 2001), the possible sources of uncertainty, such as statistical and physical uncertainties, were incorporated to correct simulations for model adequacy. Recently, the Bayesian approach was adopted for various applications, including dynamic simulations of pyrotechnically actuated devices (Kim et al. 2016), standard k-turbulence models (Guillas et al. 2014), and molecular dynamics simulations (Shin et al. 2016). Nonetheless, the Bayesian approach is limited in that the model parameters are assumed to remain fixed over the physical experiments. Initial lack of information regarding the model parameters is described by prior distributions to them. This assumption is sometimes violated when the model parameters are associated with inherent randomness or variability in material properties and product geometries.

The limitations of the Bayesian calibration approach can be partially overcome with a statistical model calibration approach. The concept of the statistical model calibration approach is identical to that of the deterministic model calibration approach; both approaches formulate an objective function and find a set of model parameters. However, the deterministic approach does not account for uncertainty in model parameters, whereas the statistical approach does (Xiong et al. 2009). The model parameters can be in the form of statistical distributions to incorporate aleatory uncertainty that exists in the real world. In principle, model parameters vary randomly over physical experiments. Several studies attempted to estimate the sample-to-sample variation in physical systems, such as in free-standing thin foils (Ageno et al. 2009) and piezoelectric energy harvesters (Jung et al. 2016). A hierarchical framework for statistical model calibration has also been developed for designing engineered products (Youn et al. 2011).

Optimization-based statistical model calibration consists of two steps, as shown in Fig. 1. The first step is to quantify the degree of disagreement (or agreement) between the two probability distributions. The second step is to find the hyper parameters of the model parameters that minimize disagreement between the probability distributions of simulation results and experimental observations. In optimum design problems, the term objective function is commonly used to evaluate the merits of a given design. In this paper, a term calibration metric is used to explicitly represent the degree of disagreement (or agreement) for the purpose of model calibration. When the value of a calibration metric reaches its minimum, it is assumed that the simulation model is calibrated. Otherwise, another iteration is executed to find a proper set of the hyper parameters by minimizing the calibration metric.

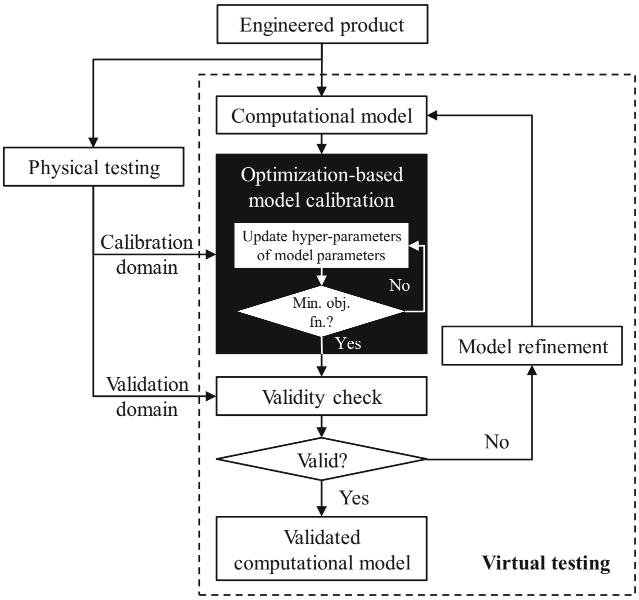


Fig. 1 Schematic of the optimization-based statistical model calibration approach

Several calibration metrics have been used for model calibration. Normalized absolute errors were used in the model updating method that was proposed based on response surface models and Monte Carlo simulation (Fang et al. 2012). The weighted sum of the squared error is the predominant metric used in vibration-based finite element model updating (Bao and Wang 2015; Mares et al. 2006; Simoen et al. 2015). The weights are typically determined by trial-and-error and/or engineering judgment. Likelihood functions were also developed for deterministic model calibration and statistical model calibration (Xiong et al. 2009). To this end, it is obvious that there is no universal calibration metric applicable to all engineering problems (Cha 2007).

This paper presents the formulation of a new calibration metric for the optimization-based calibration approach. In particular, the calibration metric is proposed to overcome two potential problems in existing calibration metrics, including the log-likelihood (log-LK) and the Kullback-Leibler divergence (KLD). The remainder of this paper is organized as follows. Section 2 overviews existing calibration metrics. Strength and limitations of the existing calibration metrics are briefly discussed. Section 3 presents the proposed calibration metric, which we call probability residual (PR). Sections 4 and 5 demonstrate the effectiveness of the proposed calibration metric using numerical examples and the rotor system of a steam turbine in a power plant as case studies. Finally, conclusions and suggestions for future work are provided in Section 6.

2 Existing calibration metrics and limitations

The framework for model calibration can be described as (Campbell 2006):

$$\zeta(x) = z(x) + \varepsilon(x) = S(x, \theta) + \delta(x) \quad (1)$$

where $\zeta(x)$ is the true value at x; x is the controllable input; $z(x)$ and $\varepsilon(x)$ are the experimental observations and measurement errors at x, respectively; $S(x, \theta)$ is the simulation output at x; θ is the simulation parameter that is the object of interest in the context of calibration; and $\delta(x)$ is the discrepancy between the simulation output (i.e., $S(x, \theta)$) and the true value at x (i.e., $\zeta(x)$). In optimization-based model calibration, the goal is to find a proper set of θ that minimizes the discrepancy. Calibration metrics have often been used to quantify the degree of the discrepancy.

2.1 Overview of existing calibration metrics

The choice of calibration metric commonly depends on the type of physical quantity of interest (i.e., $S(x, \theta_0)$) for the model calibration. The physical quantities consist of scalar (e.g., speed, pressure, temperature), vector (e.g., velocity, force, heat flux), and tensor quantities (e.g., stress, strain). The dimension of the physical quantities can be extended in both temporal and spatial fields. For example, the change of hydrodynamic pressure at a particular point of a pipeline can be described as time-series (i.e., vector instead of scalar) with respect to time, although pressures are classified as a scalar quantity. An example of temporal and spatial analysis of an engineered system (Sarin et al. 2010) is the weighted integrated factor (WIFac). For statistical model calibration, the study described in this paper focuses on scalar quantities. Uncertainties in scalar quantities are described via probability distributions. Calibration metrics that quantify the distance or similarity between probability distributions are presented in this section.

Considerable efforts have been put towards finding the relevant distance/similarity measure in different fields. Cha (Cha 2007) conducted a comprehensive survey on distance/similarity measures between probability density functions. As shown in Table 1, the author attempted to group a substantial number of distance/similarity measures in different fields, such as mathematics, physics, statistics, information theory, ecology, and biology. It was observed that each measure was developed to best describe the distance/similarity of the data collected from the corresponding field of research (Gavin et al. 2003; Looman and Campbell 1960). For example, Shannon's entropy (Shannon 1948) was designed to meet the key requirements as a measure of information, including that (1) it is continuous, (2) it exhibits monotonic increase, and (3) the original value of the measure should be the weighted sum of the individual values of the measure. In the same manner, calibration metrics should be designed to meet requirements for statistical model calibration.

In prior work, several existing distance/similarity measures have been employed as calibration metrics in the field of model development and validation. Xiong et al. (Xiong et al. 2009) incorporated a likelihood function to infer uncertain calibration parameters that vary from trial to trial over a physical experiment. The authors demonstrated the effectiveness of the likelihood measure through a thermal challenge problem. Fang et al. (Fang et al. 2012) used normalized absolute errors between structural frequencies of simulations and experiments. They developed a method to quantify parameter variability based on response surface models and Monte Carlo simulation. Bao et al. (Bao and Wang 2015) employed weighted squared errors of statistical moments in model calibration. In Bao's study, a three-degree-of-freedom mass-spring system and an aircraft structure in a laboratory were used to demonstrate the effectiveness of the proposed calibration method. In

Table 1 Distance/similarity measures between probability density functions (Cha 2007)

	Representative example	Some application fields
Minkowsky family L_p	Euclidean L_2; $d_{\mathrm{Euc}} = \sqrt{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2}$	
Normalized L_1	Canberra L_1; $d_{\mathrm{Can}} = \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}$	Ecology
Squared L_2	Pearson χ^2; $d_{\mathrm{P}} = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{Q_i}$	
Intersection family	Intersection; $s_{\mathrm{IS}} = \sum_{i=1}^{d} \min(P_i, Q_i)$	
Inner product family	Inner product; $s_{\mathrm{IP}} = \sum_{i=1}^{d} P_i \cdot Q_i$	Information retrieval Biological taxonomy
Shannon's entropy family	Kullback-Leibler divergence; $d_{\mathrm{KL}} = \sum_{i=1}^{d} P_i \ln \frac{P_i}{Q_i}$	Communication

this section, two representative metrics used for model development and validation, log-LK and KLD, are reviewed. The limitations of the metrics for statistical model calibration are also discussed.

2.2 Log-likelihood

The likelihood function ($L(z;\ \theta)$) is a function of parameters (θ), where z indicates the observed sample values. The function is presented in the form of joint probability density functions (or probability mass functions). In a mathematical form, the likelihood function is:

$$L(z;\theta) = \prod_{i=1}^{N} p(z_i;\theta) \tag{2}$$

where $z = \{z_1, z_2,\ldots, z_N\}$; $\theta = \{\theta_1, \theta_2,\ldots, \theta_M\}$; and N and M are the number of experimental observations and the number of distribution parameters, respectively. Likelihood quantifies the degree of how likely the observed sample is as a function of the possible parameter values.

Calibration with the likelihood metric is illustrated conceptually in Fig. 2. Calibration accuracy can be increased by maximizing the likelihood value. The values of probability density functions for given experimental observations are always between zero and one. If these values are multiplied with a large number of experimental observations with an invalid value of the parameters, the likelihood value will theoretically converge to zero. This can cause a zero convergence problem in the optimization process of statistical model calibration. To avoid the unwanted situation of the likelihood function being negligibly small, the natural logarithm of (2) can be taken (Oh et al. 2016b). Then, the log-LK is:

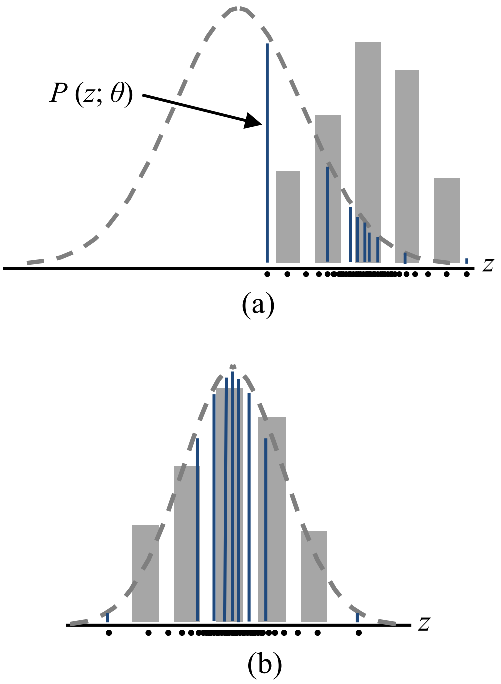


Fig. 2 Concept of log-likelihood in (**a**) initial condition and (**b**) calibrated condition

$$\log L(z;\theta) = \sum_{i=1}^{N} \log p(z_i;\theta) \tag{3}$$

Logarithm measures, such as log-LK and Shannon's entropy, can be effective for a couple of reasons (Shannon 1948). First, taking the natural logarithm of the exponential family distributions makes the logarithm vary in a linear scale. Second, human intuition is more suitable to linear metrics. Last, as stated before, this process can avoid the zero convergence problem.

Xiong et al. (2009) modeled the uncertainty of the parameters (θ) of the likelihood function in statistical model calibration. The M parameters were decomposed into hyper parameters (i.e., $\theta = \{\mu_{\theta 1}, \sigma_{\theta 1}, \mu_{\theta 2}, \sigma_{\theta 2}, \ldots, \mu_{\theta \mathrm{M}}, \sigma_{\theta M}\}$ by assuming a Gaussian distribution. In a same manner, the error (ε) was also decomposed into hyper parameters (i.e., $\varepsilon = \{\mu_\varepsilon, \sigma_\varepsilon\}$).

2.3 Kullback-Leibler divergence

The KLD ($D(P;\ Q)$) is the expectation of the information for discrimination between two probability distributions of P and Q (i.e., $\log p(z_i) / q(z_i)$) (Kullback and Leibler 1951). In mathematical form, the KLD is:

$$D(P;Q) = \sum_{i=1}^{N} p(z_i)\log\frac{p(z_i)}{q(z_i)} \tag{4}$$

where $p(z_i)$ and $q(z_i)$ are the probabilities for observations $i = 1,\ldots, N$ of distributions P and Q, respectively.

The KLD represents a measure of similarity between two probability distributions. As shown in Fig. 3, the KLD values decrease as the overlap between the two probability distributions increases, which is opposite to the case of the likelihood metric. When two probability distributions from simulations and experiments overlap perfectly, the KLD value is zero. The KLD is actively used in research areas of information theory, such as image and speech recognition (Gao et al. 2017; Kim et al. 2017).

The KLD is an asymmetric divergence; $D(P;\ Q)$ is not equal to $D(Q;\ P)$. This property leads to directed divergence when it is used as a calibration metric in model calibration. When the KLD is employed as an objective function in an optimization problem, the solution depends on the direction in which the problem is solved (Abbas et al. 2017). Lee et al. (2017) conducted a detailed analysis of KLD as a calibration metric for statistical model calibration.

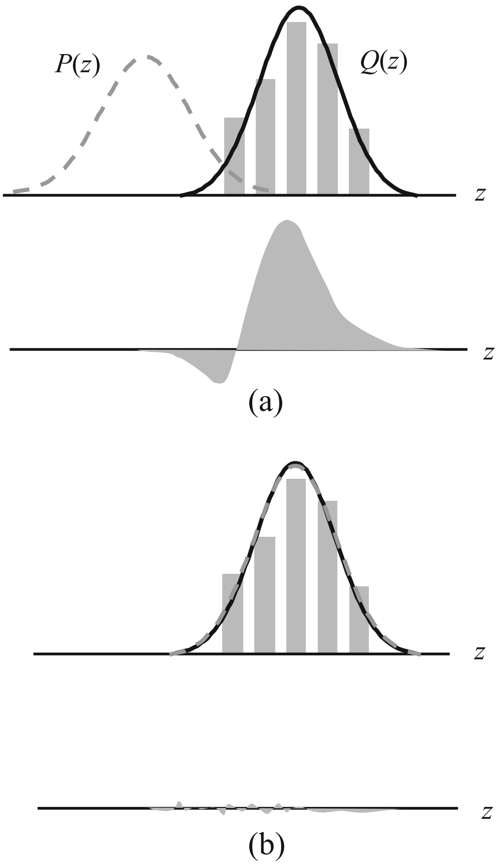


Fig. 3 Concept of KLD in (**a**) initial condition and (**b**) calibrated condition

2.4 Area metric

The area metric $d(F, S_n)$ is defined as the area between two probability distributions (F and S_n). Mathematically, the area metric is expressed (Ferson et al. 2008):

$$d(F, S_n) = \int_{-\infty}^{\infty} |F(x) - S_n(x)| dx \tag{5}$$

where $F(x)$ and $S_n(x)$ are the cumulative distribution function predicted by computational models and the empirical distribution function of experimental data, respectively.

The area metric is a measure of the mismatch between the two distribution functions. As shown in Fig. 4, the area metric value at the initial condition decreases when the calibration is conducted. The gray shaded area illustrates the amount of the mismatch. The area is minimized after calibration. When the two functions overlap perfectly, the area metric value should be zero theoretically. However, in practice, the area metric shows positive values since experimental observations and/or simulation responses are sparse. The area metric was used in the model development and validation such as piezoelectric energy harvester design (Jung et al. 2016).

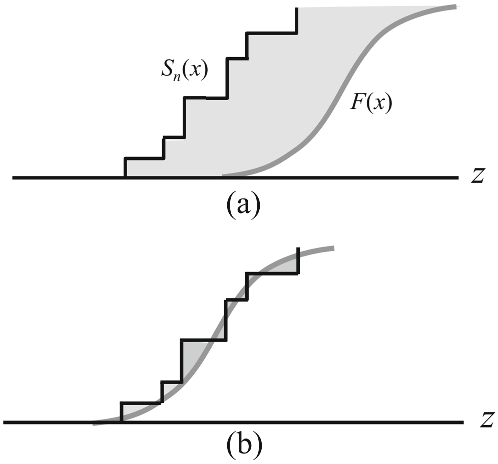


Fig. 4 Concept of area metric in (**a**) initial condition and (**b**) calibrated condition

2.5 Limitations of the existing calibration metrics

The log-LK and KLD that takes the logarithm can suffer from the tail-end effect. The tail-end effect is defined as the phenomenon where the data on the tail of the PDFs can have a larger impact on the logarithmic value than it does on the body of the PDFs. The concept of the tail-end effect is illustrated in Fig. 5. The probability density of a standard normal distribution in Fig. 5 (a) is used as a representative example. When the probability density values converge to zero, as described in Fig. 5 (b), the natural logarithm of the probability density values decreases significantly. For instance, log $p(x)$ is −6.908 when $p(x) = 0.001$; log $p(x)$ is -4.605, when $p(x) = 0.01$; log $p(x)$ is -2.303, when $p(x) = 0.1$. A 10 times difference in $p(x)$ is equivalent to only a two-fold change in log $p(x)$. The impact of data points located at the tail of the PDF has larger impact on the logarithmic value than that of the data points located in the body of the PDF. Consequently, in statistical model calibration, the tail-end effect can lead to better agreement at the tails between simulation and experimental PDFs, while leading to a poor agreement at the body. This is particularly true when an assumed probability distribution of unknown parameters to be calibrated does not exactly match that of an experimentally observed distribution. In principle, the use of the logarithm helps relieve the zero convergence problem by modifying the multiplication operation into the summation operation. However, it will degrade the accuracy of the statistical model calibration results by terminating the optimization process with an inaccurate estimation of design variables with a given convergence criterion.

The gradient of an objective function often determines the rate of convergence in an optimization problem. As the probability density increases, the gradient of the

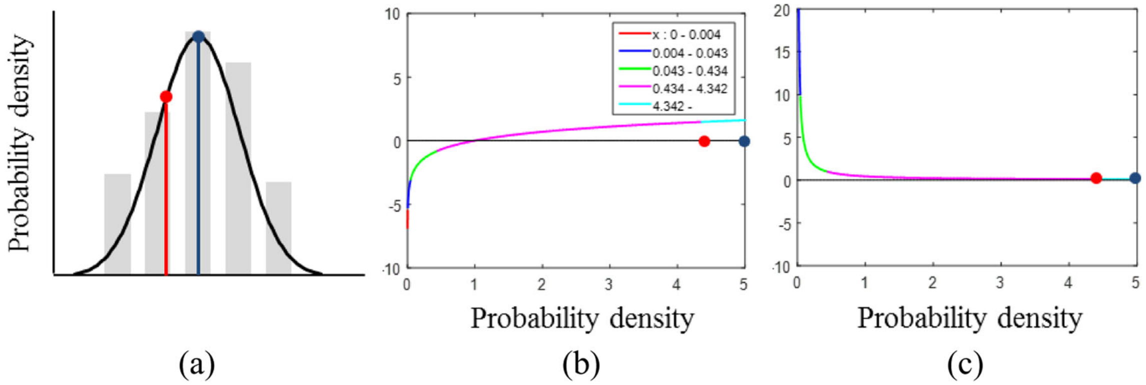


Fig. 5 Low sensitivity problem of calibration metrics with logarithm (**a**) PDF, (**b**) log-operation plot, and (**c**) gradient plot

probabilities decreases. The increment is almost negligible at the end. For example, as shown in Fig. 5 (c), suppose that there are two sets of probability densities. The first set is $p(x_1) = 0.05$ and $p(x_2) = 0.15$, while the second set is $p(x_1) = 0.25$ and $p(x_2) = 0.35$. For the first set, gradient values at the two probability densities are 20.08 and 6.68; their difference is 13.40. For the second set, gradient values at the two probability densities are 3.99 and 2.86; their difference is only 1.13, which is much smaller than that of the first set. This can degrade the efficiency of the optimization process for statistical model calibration. As the log-LK value approaches its maximum (e.g., $p(x_3) = 4.5$ or $p(x_3) = 5.0$ in Fig. 5 (c)), it requires more computational resources to find an optimal solution, lowering the efficiency. The same logic applies to KLD, since both are based on the logarithmic operator.

The area metric has limitations from the perspective of optimization-based model calibration. First, the existence of the global minimum is not guaranteed. Thus, the solution can be local minima. Second, predominant optimization algorithms, such as conjugate gradient and/or quasi-Newton methods, may not be useful to solve the local minima problem. Evolutionary algorithms such as genetic algorithms can be effective. However, the computational cost will be much higher.

3 Proposed calibration metric

This section proposes a new calibration metric to overcome the limitations of the existing calibration metrics. The key idea of the proposed calibration metric comes from distance measures of the squared L_2 family and the χ^2 family (Cha 2007). The proposed calibration metric, which we call probability residual (PR), is explained in Section 3.1. The PR consists of two components: squared residual (or squared Euclidean distance) and scale factor; these components are explained in Sections 3.2 and 3.3, respectively.

3.1 Probability residual

The probability residual (PR) metric is defined as the sum of the product of the scale factor (*S*) and the squared residual (SR):

$$PR(P, Q) = S \times SR(P, Q) \tag{6}$$

The SR is a metric that quantifies how much two probability distributions do not overlap. When the probability distribution of the computational response (*P*) perfectly overlaps with that of the experimental results (*Q*), the PR becomes zero, like KLD, as illustrated in Fig. 6.

3.2 Squared residual

Inherent randomness, or variability of the performance of interest (PoI), in engineered systems is commonly

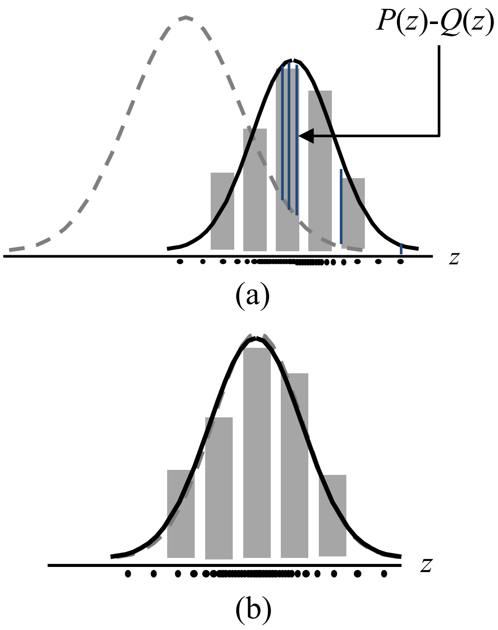
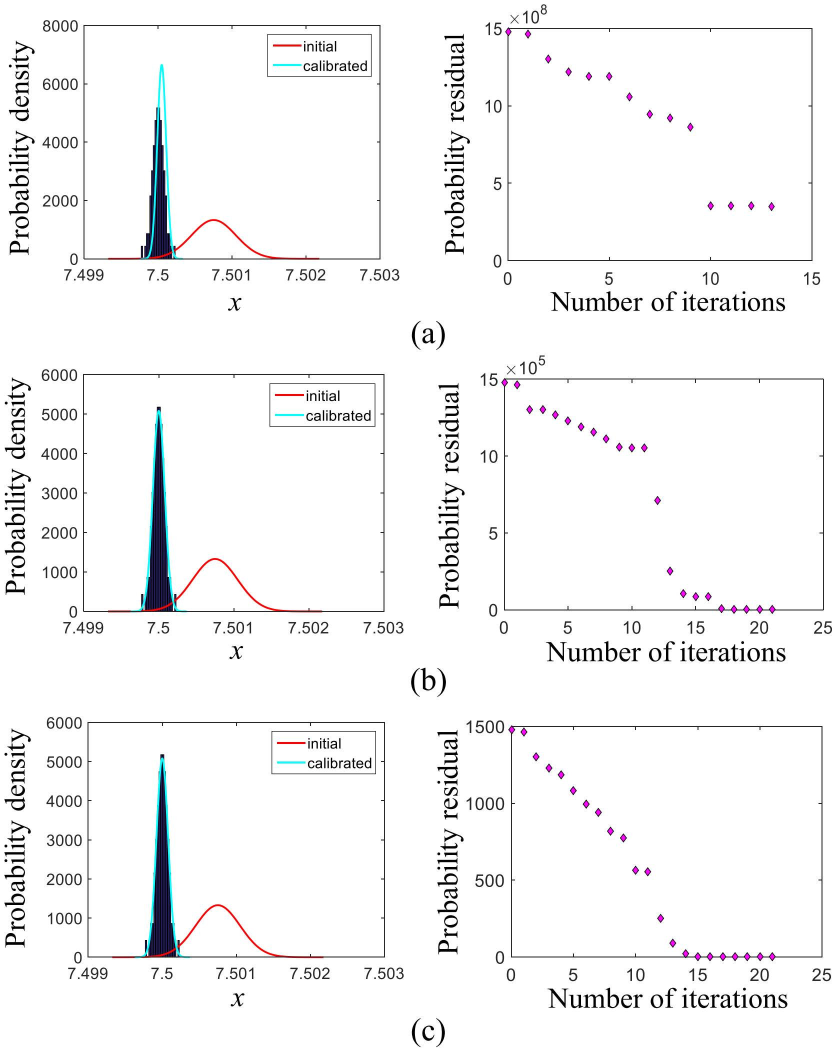


Fig. 6 Concept of PR in (**a**) initial condition and (**b**) calibrated condition

Fig. 7 Calibration result for a narrow probability distribution with different scale factors: (**a**) $C = 1$, (**b**) $C = 10$ and (**c**) $C = 100$



described with the use of empirical probability distributions. Suppose Z is a set of N elements whose possible values are discrete and finite. A histogram $q(Z)$ of a set Z presents the normalized frequency of the individual values. The normalized frequency for the i^{th} bin from experiments is denoted as $q(z_i)$. The probability of the corresponding bin $p(z_i)$ from simulations can be defined in the same manner. The SR for i^{th} bin can be described as:

$$SR_i(P,Q) = (p(z_i)-q(z_i))^2 \tag{7}$$

The SR between simulation results and experimental observations is the summation of the R_i for possible samples in set Z. Specifically,

$$SR(P,Q) = \sum_{i=1}^{n} (p(z_i)-q(z_i))^2 \tag{8}$$

where n is the number of discrete bins for set Z. As the quantity in (8) depends on the number of bins, the proper number of bins should be chosen. If the number of bins is not selected properly, the features of the experimental observations cannot be captured sufficiently. Previous studies (Indira et al. 2011; Wand 1997) discussed how to select an optimal number of data size and bin size.

In statistical model calibration, simulation results are represented by parametric or non-parametric probability distributions. The probability distribution from simulation results can be estimated by using kernel density estimation methods. When the parameters of the estimated probability distributions are θ, (8) becomes:

$$SR(P,Q) = \int_{-\infty}^{\infty} (p(z;\theta)-q(z))^2 dz \tag{9}$$

The SR can be described as $SR(P, Q)$ in either (8) or (9). When the two probability distributions overlap perfectly, the

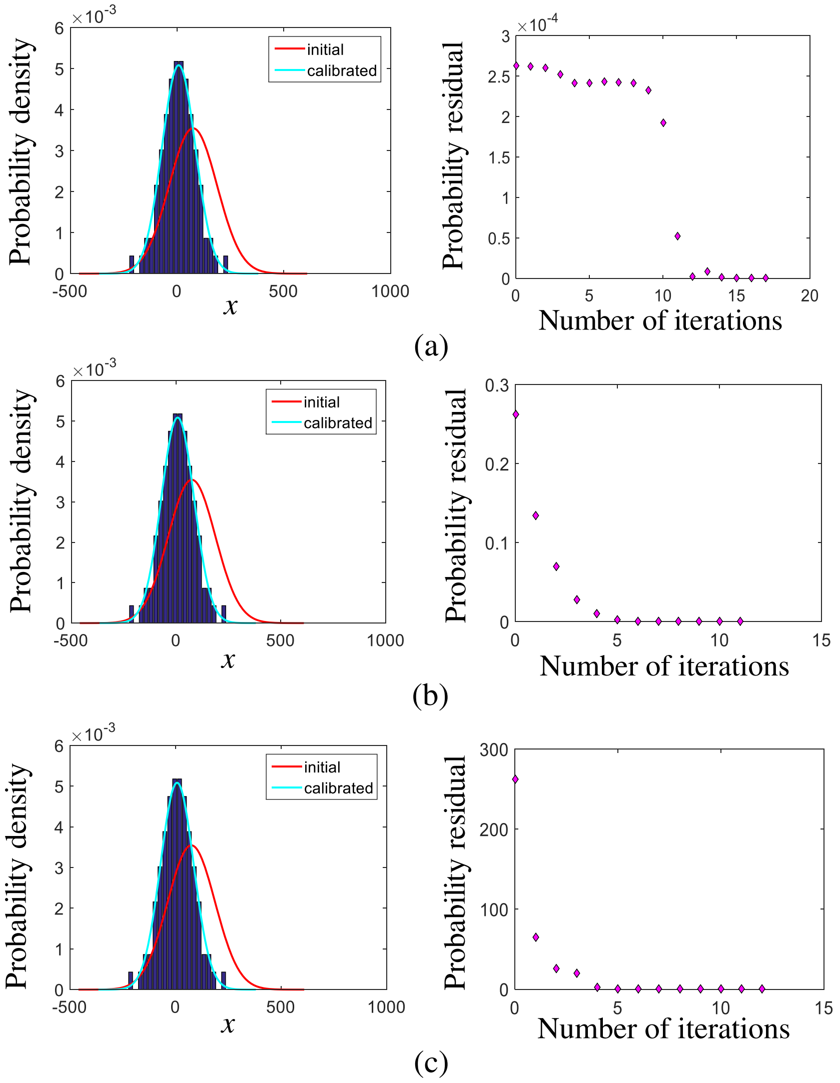


Fig. 8 Calibration result for a wide probability distribution with different scale factors: (**a**) $C = 1$, (**b**) $C = 10$ and (**c**) $C = 100$

SR equals to zero. From the perspective of optimization, calibration metrics based on the PR have two advantages. First, a global minimum (or maximum) exists in the optimization problem, as the probability residual has a quadratic form, i.e., convex. Thus, it can avoid local minima. Second, predominant optimization algorithms, such as conjugate gradient and/or quasi-Newton methods, can be used. The characteristics of "convexity" make the proposed calibration metric robust.

3.3 Scale factor

For statistical model calibration, the convergence rate and the function evaluation number are critical issues during the optimization process. It is desirable to have a high convergence rate, while minimizing the number of function evaluations. To achieve this goal, a scale factor is devised in conjunction with the SR, as described in Section 3.1.

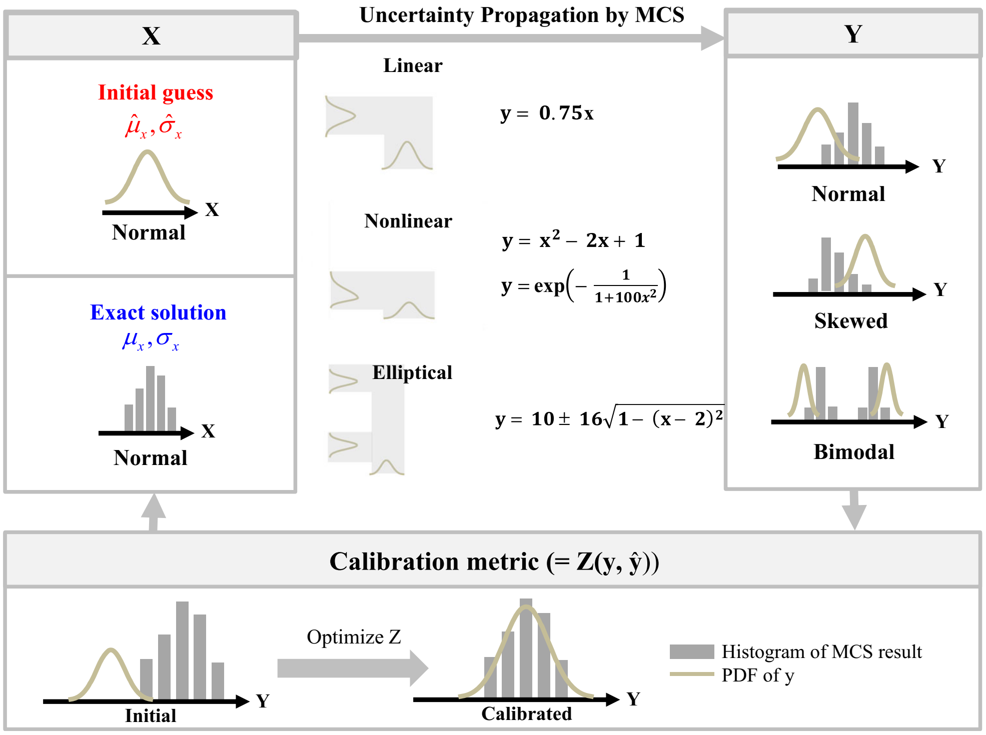
The scale factor is defined as C to the power of the logarithm of the largest magnitude of the empirical probability distribution ($\max(Q)$):

$$S(P) = \frac{1}{C^{\lfloor \log_{10}(\max(Q)) \rfloor}} \tag{10}$$

where C is the constant that is defined by the user. The bracket is the floor function (or Gauss notation) that gives the greatest integer less than or equal to the number in the bracket. When the value of the "$\max(Q)$" is between one and 10, the scale factor becomes one with any choice of constant value. This indicates that there is no need to employ the scale factor since the SR provides a high convergence rate in the optimization process. The scale factor is fixed during the optimization process once the experimental data are given.

The constant C should be selected considering the maximum value of Q. If the C is too small, the calibration accuracy

Fig. 9 Process for performance evaluation of the PR approach



will be degraded, sometimes leading to divergence of the optimization. On the other hand, if the C is too large, the calibration speed will be degraded, leading to slow convergence. To understand the effect of the constant C on the convergence rate, a parametric study was conducted with mathematical examples. The parameters of probability distributions from simulations (e.g., mean and standard deviation of normal distributions) were calibrated by maximizing the PR. It was assumed that the probability distributions from simulations and experiments follow normal distributions. The calibration process used unconstrained optimization problems. Next, the effect of the scale factor on the convergence rate was evaluated. For a narrow probability distribution with max(Q) of 5000, the use of a small value of $C = 1$ resulted in divergence, as shown in Fig. 7 (a). With the use of larger values of $C = 10$ and $C =$ 100, the calibration converged with 22 and 22 iterations of function evaluations, respectively, as shown in Fig. 7 (b) and (c). For a wide probability distribution with max(Q) of 0.005, a similar result was observed, as shown in Fig. 8. From the results of the mathematical example, the convergence rate appears to be the best with $C = [10\sim1000]$. Therefore, this paper uses 100 for C.

4 Case study 1: mathematical examples

The performance of the proposed calibration metric (PR) was compared to that of existing calibration metrics (i.e., log-LK and KLD). The input variable of the systems was assumed to follow a normal distribution. When the uncertainty of the input variable was propagated through different system functions, responses could be calculated by Monte Carlo simulation with a sample size of one million. In this case study, four types of system functions, including linear and nonlinear A, nonlinear B and elliptical, were used as representative numerical examples. These system functions are shown in (11), (12), (13), and (14), respectively.

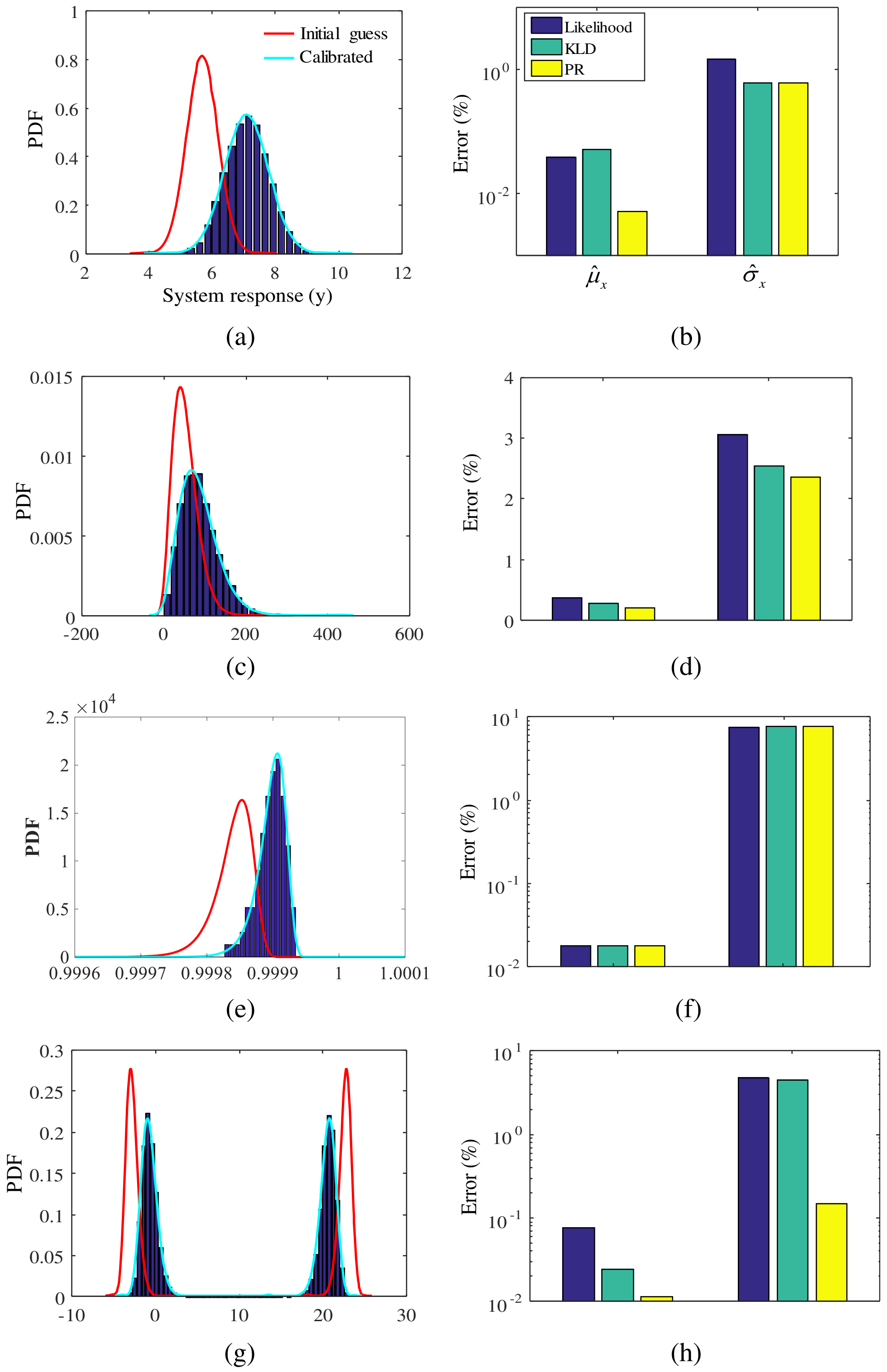
$$y = 0.75x \tag{11}$$

$$y = x^2 – 2x + 1 \tag{12}$$

$$y = \exp\left(-\frac{1}{1 + 100x^2}\right) \tag{13}$$

$$y = 10 \pm 16\sqrt{1–(x–2)^2} \tag{14}$$

where x is the input variable and y is the system response. The system functions generated three types of responses: normal, skewed, and bimodal distributions. As shown in Fig. 9, the initial guess of the input variable was calibrated to converge to



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 10 System response before and after calibration using PR for (**a**) linear, (**c**) nonlinear A, (**e**) nonlinear B, and (**g**) elliptical system functions. Errors between distribution parameters of the calibrated input variable and exact solution when different calibration metrics are used for (**b**) linear, (**d**) nonlinear A, (**f**) nonlinear B, and (**h**) elliptical system functions

the exact solution by minimizing (or maximizing) the calibration metric.

Figure 10 (a), (c), (e), and (g) show system responses before and after calibration using the PR metric. From visual inspection, the PDFs after calibration matched the histograms when the four system functions were evaluated. Figure 10 (b), (d), (f), and (h) compare errors between the PDF of the calibrated input variable and the PDF of

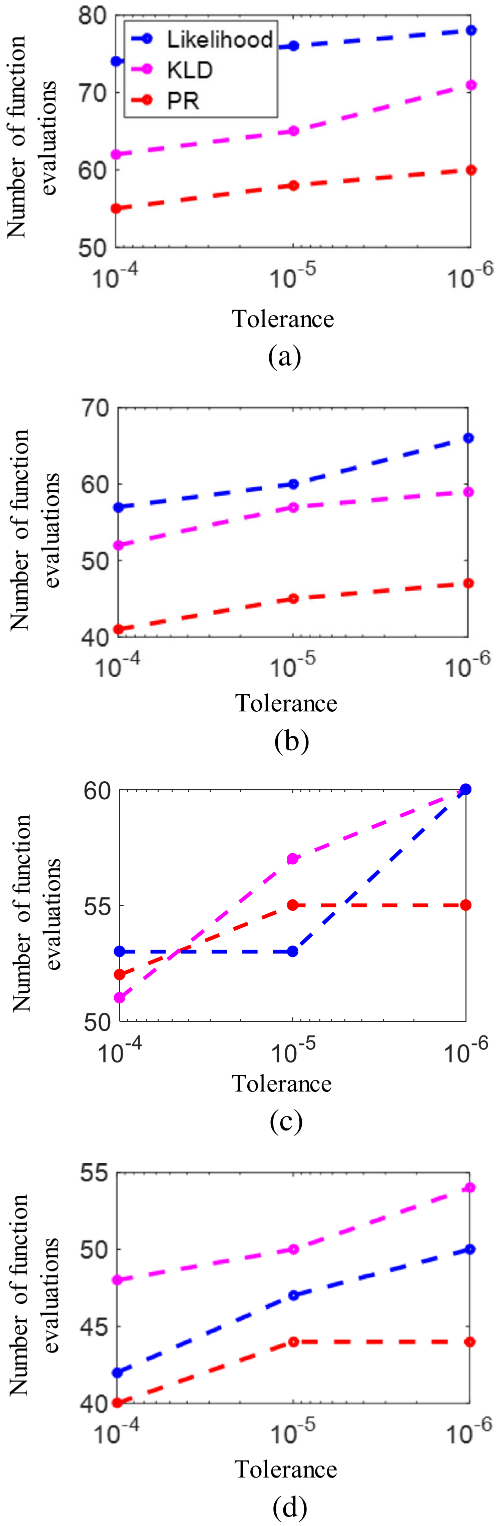


Fig. 11 Comparison of the number of function evaluations for likelihood, KLD, and PR when (**a**) linear, (**b**) nonlinear A, (**c**) nonlinear B, and (**d**) elliptical system functions are used

the exact solution. The absolute percentage errors of the means for the PR metric were less than one percentage point, regardless of the type of system function. For each of the four system responses, the proposed PR metric outperformed or was equivalent to the existing metrics.

Figure 11 (a) compares the number of function evaluations for log-LK, KLD, and PR. With three different levels of x tolerances in the optimization process, the number of function evaluations using the PR metric was smaller than that using the existing metrics. A similar result was observed when different system functions were employed, as shown in Fig. 11 (b) and (d). For a highly-nonlinear system function in Fig. 11 (c), the performance of the three metrics was comparable each other. Consequently, the effectiveness of the proposed calibration metric was demonstrated.

5 Case study 2: journal bearing rotor systems

Journal bearing rotor systems are one of the key systems in steam turbines of power plants. Journal bearings are used to support rotating shafts that are subjected to heavy loading and high-speed conditions. The safety and reliability of journal bearings must be ensured during their design life, which may be 25 to 40 years. Unexpected failure of a rotor system should be avoided by proper maintenance. To address this challenge, rotor diagnostic techniques have been studied extensively in research communities.

Rotor diagnostics often employs model-based and data-driven approaches. The data-driven diagnostic approach requires a sufficient amount of data. Data should be collected from both healthy and faulty rotor systems under various environmental and operational conditions. The failure modes of the rotor systems should also be known. However, while healthy data are relatively easy to acquire, fault data seldom exists for real-world power plants. The model-based diagnostic approach has the potential to overcome the problem of insufficient data. If simulation models can be built to emulate the dynamic behaviors of the journal bearing rotor systems in real power plants, theoretically, an unlimited amount of data can be collected from the simulation models.

Rotor systems have been analyzed using the Timoshenko beam theory, the transfer matrix method, and finite element methods. Industrial rotating machines, such as turbo compressors, vacuum pumps, and induction motors, were modeled for various reasons, including resonance avoidance, stability evaluation, and vibration suppression. Nevertheless, building an accurate simulation model remains extremely challenging. One of the major challenges is epistemic uncertainty that arises due to a lack of available information. In this case study, a simulation model of a journal bearing rotor system with uncertain input variables is calibrated using the proposed statistical model calibration technique.

Table 2 Statistical model calibration with two-step approach

	1st step	2nd step
System condition	Normal	Rubbing fault
Employed test	Impact hammer test	Rotor operational test
Simulation	Modal analysis	Rotor steady state analysis
Unknown variable to be calibrated	Journal bearing stiffness, journal bearing damping coefficient	Penalty stiffness

5.1 Overview of statistical model calibration

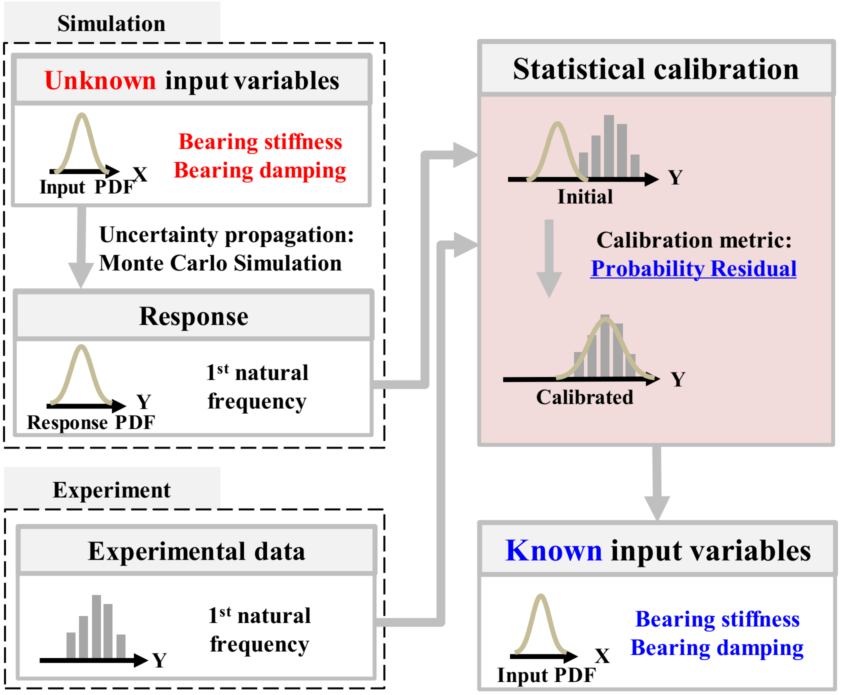
Through expert knowledge and sensitivity analysis, three variables (among many) were found to be unknown. To determine the unknown variables, a two-step approach was used, as shown in Table 2. In the first step, the two unknown variables (i.e., stiffness and damping coefficient) for the journal bearings were set to be unknown, as shown in Fig. 12. The initial PDFs of the unknown variables were propagated by meta modeling and Monte Carlo simulation to calculate the PDF of the simulation response (1st natural frequency). Then, the disagreement between the two PDFs (one derived from simulation and the other from experiments) were quantified using the proposed calibration metric. The unknown variables were fine-tuned until the calibration metric met a threshold. Throughout this iterative process, the two unknown variables were calibrated.

In the second step, the calibration shown in Fig. 12 was repeated for another unknown variable (i.e., penalty stiffness). It should be noted that the two variables calibrated in the 1st step were incorporated as known variables in the 2nd step. The crest factor was used for comparison, instead of the 1st natural frequency. The three unknown variables were determined by the two-step approach. When a calibrated simulation model was obtained, the validity of the simulation model was evaluated with another set of experimental data.

5.2 Experiments and finite element analysis

Figure 13 (a) presents the General Electric (GE) Bently-Nevada RK4 rotor kit that consists of short and long shafts, a flexible coupling, and three journal bearings. The long, 10 mm diameter shaft supports a disk of 0.8 kg. A small amount of unbalance typically exists at the normal state even though balancing of the rotor systems is conducted periodically. As shown in Fig. 13 (b), a mass of 0.45 g was pinned to the disk to emulate the normal state. Rubbing is one of the major failure modes in journal bearing rotor systems. A specially designed jig was used to emulate the rubbing. As depicted in Fig. 13 (c), the severity of rubbing can be precisely

Fig. 12 Statistical model calibration process: 1st step



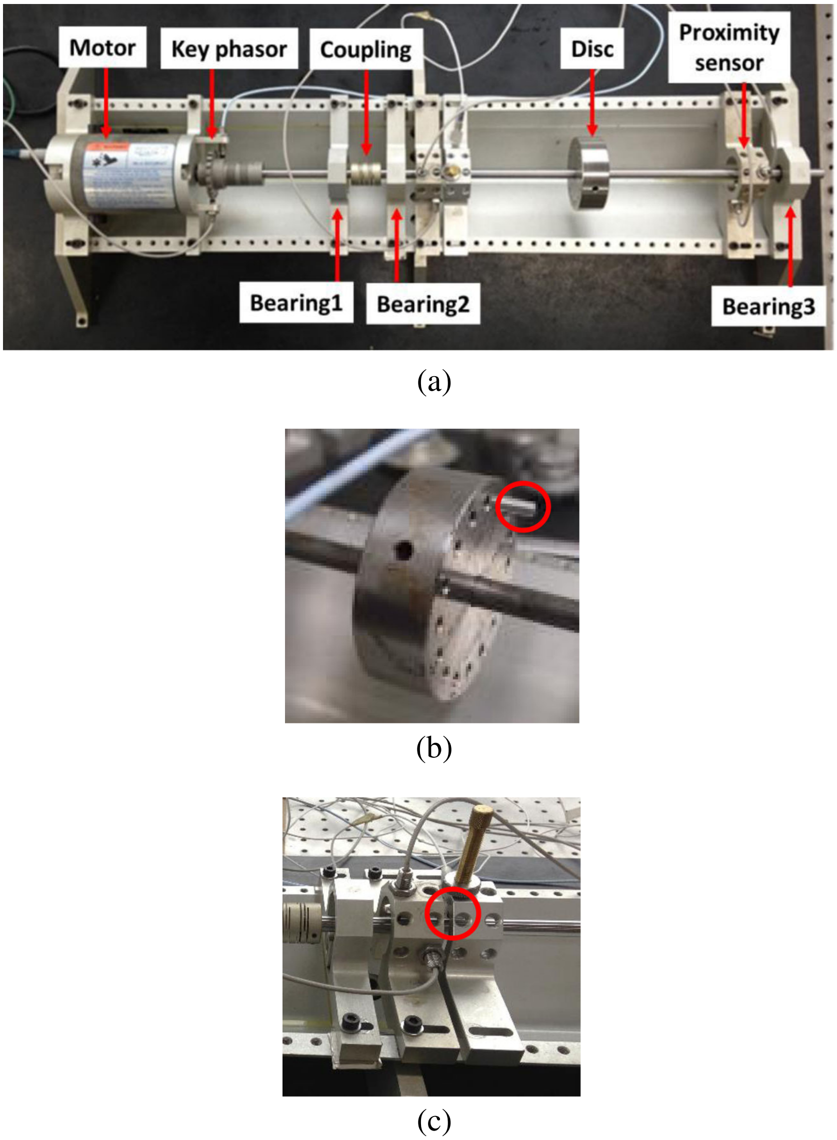


Fig. 13 GE Bently-Nevada RK4: (**a**) rotor kit, (**b**) normal state and (**c**) rubbing state

controlled by regulating the rubbing screw with which the shaft was rubbed. The rotor kit operated at a rotating speed of 3600 rpm, the normal speed of steam turbines in thermal power plants. A pair of proximity sensors (Bently-Nevada 3300) were mounted between the 2nd and 3rd journal bearings to measure vibration. The sampling rate was 8500 Hz. For more details, refer to Jung et al. (2017).

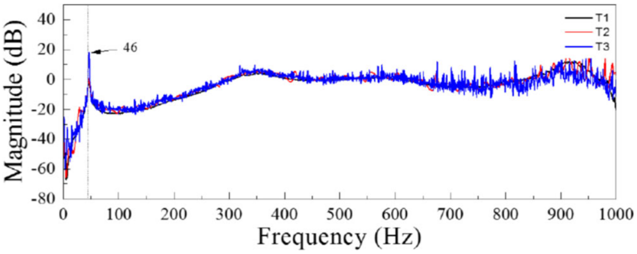


Fig. 14 Impact hammer test result of RK4 rotor kit (Kwon et al. 2018)

Figure 14 describes a representative plot for impact hammer testing of the RK4 rotor kit. The test was repeated 15 times. The first natural frequency could be described as a normal distribution, $N(45.69, 0.19)$. This variable was used to calibrate the stiffness and damping coefficient of the journal bearings.

When the rotor kit was operated in the rubbing state, it was expected that multiple harmonics, as well as 1x frequencies would exist in the frequency domain analysis. As expected, the FFT results confirmed this, as shown in Fig. 15. The severity of the rubbing can be quantified using several features, such as root mean square (RMS), crest factor, and kurtosis. The crest factor is known to be appropriate for capturing the random vibration shock produced by the rubbing. The crest factor in the rubbing state could be described as a normal distribution, $N(1.73, 0.14)$. This variable was used to calibrate the penalty stiffness. As a validity check, another set of

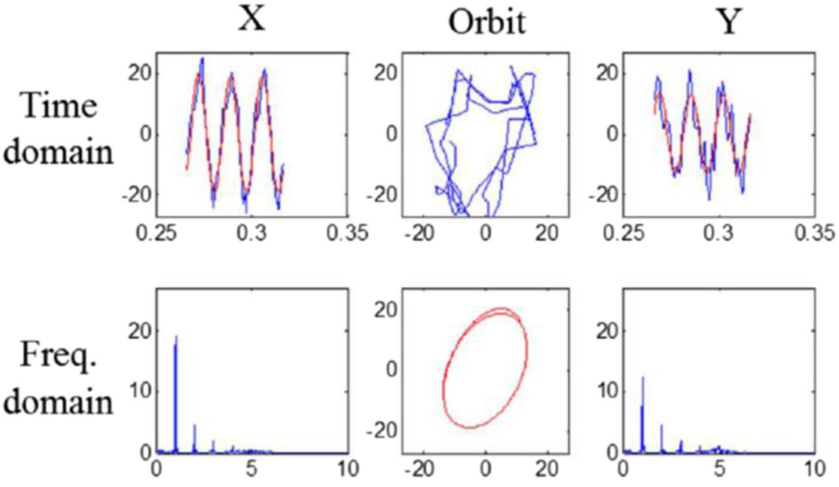


Fig. 15 x- and y-axes proximity sensor signals at the rubbing state

experimental data was collected at the rubbing state with an additional unbalance mass of 0.5 g (0.95 g in total).

Rotor dynamics can be expressed by a 2nd order ordinary differential equation in a matrix form. Here,

$$M\ddot{x}(t) + (C + G)\dot{x}(t) + (K + B)x(t) = f(t) \tag{15}$$

where $x(t)$ and $f(t)$ are the displacement and force vector, respectively; M, C, and K are the mass, damping coefficient, and spring constant, respectively. G and B are the gyroscopic moment matrix and rotating damping matrix, respectively. Figure 16 (Kwon et al. 2018) shows a simulation model that was built using a commercial software package, ANSYS APDL. The simulation model contains mechanical components, as explained in Fig. 13 (a). Short and long shafts, three journal bearings, and a disk were modeled with three-dimensional beam elements (BEAM188), two-dimensional spring elements (COMBI214), and concentrated mass elements (MASS21), respectively. The flexible coupling was modeled by combining one-dimensional spring elements (COMBIN14) with concentrated mass elements (MASS21). The rubbing was emulated by employing penalty stiffness (CONTA178) that prevents a particular node from penetrating neighboring nodes.

5.3 Results

Figure 17 presents the first natural frequency before and after calibration. With the use of the proposed calibration metric, PR, the disagreement of the first natural

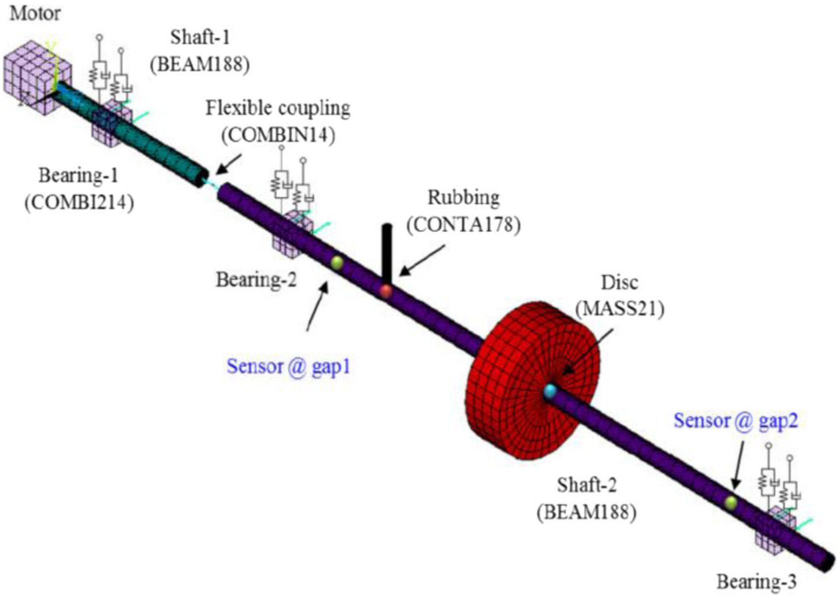


Fig. 16 GE Bently-Nevada RK4 simulation model (Kwon et al. 2018)

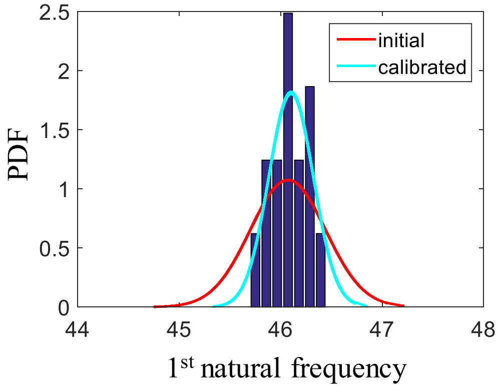


Fig. 17 Calibration result: 1st natural frequency in the normal state

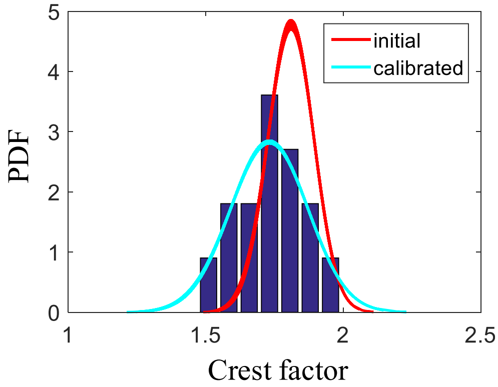


Fig. 18 Calibration result: crest factor in the rubbing state

frequencies between the simulation result and the experimental result was minimized, as shown in Table 3. The errors of the mean and standard deviation were only 0.02% and 0.88%, respectively, which are negligible. The initial values of the stiffness and damping coefficient were 266 kN/m^2 and 10.55 Ns/m, respectively. The coefficient of variation of 10% was used, and normal distributions were assumed. After calibration, they were determined to be N(266.50, 15.99) kN/m^2 and N(15.89, 0.64) Ns/m, respectively.

Figure 18 shows the crest factor before and after calibration. As expected, the disagreement of the crest factor was minimized. The errors of the mean and standard deviation were almost zero (0.0001%); this amount of error can be ignored. The initial value of the penalty stiffness was 400 GN/m^2. The coefficient of variation of 10% was used, and a normal distribution was assumed. After calibration, it was determined to be N(362.3, 67.3) GN/m^2.

Figure 19 describes the validity check of the calibrated simulation model. The histogram for the crest factor in the rubbing state is shown in Fig. 19 (a). The empirical CDF plot shown in Fig. 19 (b) was used to calculate the u metric. The blue and red lines indicate the ideal and experimental CDFs, respectively. The discrepancy between the two CDFs was the magnitude of the area metric. For a validity check, hypothesis testing was employed. The null hypothesis was that the simulation responses were not statistically different from the experimental results. Since the area metric (i.e., 0.0364) was smaller than the threshold (i.e., 0.1051)—with a sample size of 20 and a significance level of 5%, as shown in Fig. 19 (c)—the null hypothesis could not be rejected. Therefore, it was confirmed that the calibrated model is valid for future use.

Table 3 Comparison of experimental result and simulation response after calibration: 1st natural frequency

	1st natural frequency	
	Mean	Standard deviation
Experiment (Hz)	45.6935	0.1925
Simulation (Hz)	45.6855	0.1908
Error (%)	0.017	0.883

6 Conclusions

Statistical model calibration can be made practical for computational model development processes by employing relevant calibration metrics since the calibration metric serves as the objective function in optimization-based calibration. This study presented the limitations of existing calibration metrics, including log-likelihood (log-LK) and Kullback-Leibler divergence (KLD), in terms of calibration accuracy and efficiency. The log-LK and KLD that takes the logarithm metrics suffer from the tail-end effect. In principle, the use of the logarithm helps relieve the zero convergence problem. However, it was shown that the accuracy of the statistical model calibration results was degraded by terminating the optimization process with an inaccurate estimation of design variables with a given convergence criterion.

To address these limitations, a new calibration metric, called probability residual (PR), was proposed in this paper. There are three primary merits of the proposed metric. First, the characteristic of "convexity" makes the PR metric robust. A global minimum (or maximum) exists in the optimization problem, as the probability residual has a quadratic form. Predominant optimization algorithms, such as conjugate gradient and/or quasi-Newton methods, can be used. Second, the PR metric is efficient due to the scale factor that controls shape of the probability density function (PDF). The PR metric was designed to have a high convergence rate, while minimizing the number of function evaluations, by adjusting the proper value of the

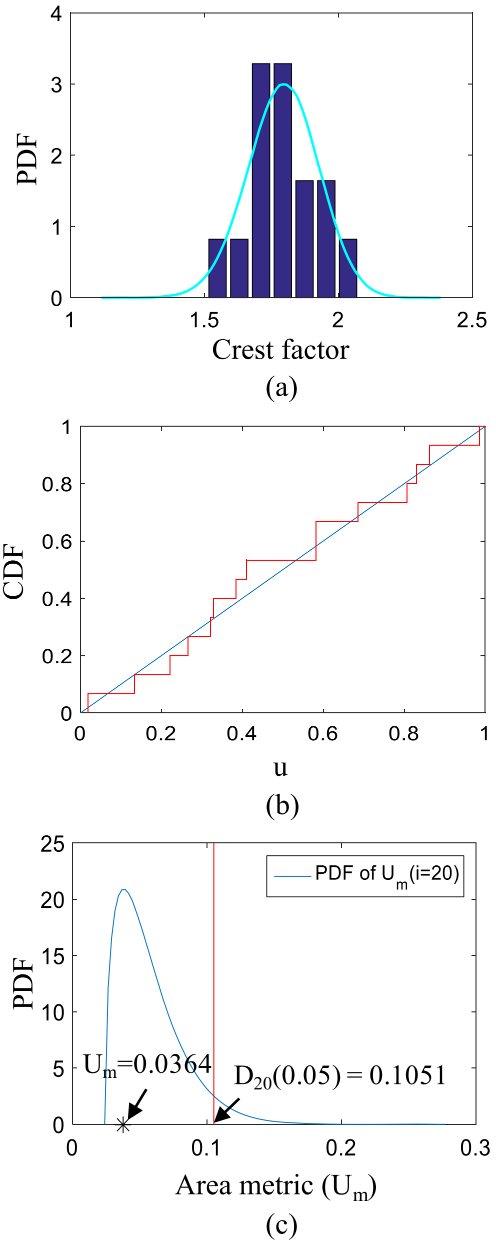


Fig. 19 Validity check of the calibrated simulation model (**a**) crest factor in rubbing state with additional unbalance mass of 0.5 g, (**b**) area metric calculation, and (**c**) hypothesis test

scale factor. Finally, the PR metric outperformed the existing calibration metrics for the given system functions, including linear, nonlinear, and elliptical.

A case study of a journal bearing rotor system was used to demonstrate the applicability of the PR metric to statistical model calibration of real systems. Three unknown variables (stiffness and damping coefficient of journal bearings, and penalty stiffness of node contacts) in the simulation model were calibrated using the two-step calibration approach. The validity of the calibrated model was confirmed by hypothesis testing. The performance predicted by the calibrated model was not significantly different from that found from the additional set of experimental data. With a significance level of 5%, the null hypothesis could not be rejected. The case study showed that the proposed PR metric is a promising method for building an accurate computational model.

In future work, the root causes of inaccurate calibrated results that arise from using existing calibration metrics should be examined. As discussed earlier in this paper, in optimization-based model calibration, it was empirically observed that the existing calibration metrics often suffer from inaccurate calibrated results or divergence of the results. This needs to be investigated in detail in future studies.

7 Replication of results

MATLAB codes are disclosed for the proposed calibration metric (i.e., Probability Residual; PR) as well as widely-accepted ones (likelihood and Kullbalk-Leiblier divergence).

Acknowledgments This work was supported by the R&D project (R17GA08) of Korea Electric Power Corporation (KEPCO) and the Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant no. 20172010105590 funded by the Korean Ministry of Trade, Industry, & Energy (MOTIE).

Compliance with ethical standards

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

References

Abbas A, H. Cadenbach A, Salimi E (2017) A Kullback–Leibler view of maximum entropy and maximum log-probability methods. Entropy 19:232

Ageno M, Bolzon G, Maier G (2009) An inverse analysis procedure for the material parameter identification of elastic–plastic free-standing foils. Struct Multidiscip Optim 38:229–243. https://doi.org/10.1007/s00158-008-0294-8

AIAA (1998) Guide for the verification and validation of computational fluid dynamic simulations, American Institute of Aeronautics and Astronautics

ASME (2006) Guide for verification and validation in computational solid mechanics. American Society of Mechanical Engineers, New York

Bao N, Wang C (2015) A Monte Carlo simulation based inverse propagation method for stochastic model updating. Mech Syst Signal Process 60-61:928–944. https://doi.org/10.1016/j.ymssp.2015.01.011

Campbell K (2006) Statistical calibration of computer simulations. Reliab Eng Syst Saf 91:1358–1363. https://doi.org/10.1016/j.ress.2005.11.032

Cha S-H (2007) Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Models Methods Appl Sci 1:300–307

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. https://doi.org/10.1016/j.ymssp.2012.06.028

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

Gao BB, Xing C, Xie CW, Wu J, Geng X (2017) Deep label distribution learning with label ambiguity. IEEE Trans Image Process 26:2825–2838. https://doi.org/10.1109/TIP.2017.2689998

Gavin DG, Oswald WW, Wahl ER, Williams JW (2003) A statistical approach to evaluating distance metrics and analog assignments for pollen records. Quat Res 60:356–367. https://doi.org/10.1016/S0033-5894(03)00088-7

Guillas S, Glover N, Malki-Epshtein L (2014) Bayesian calibration of the constants of the k–ε turbulence model for a CFD model of street canyon flow. Comput Methods Appl Mech Eng 279:536–553. https://doi.org/10.1016/j.cma.2014.06.008

Indira V, Vasanthakumari R, Sakthivel NR, Sugumaran V (2011) A method for calculation of optimum data size and bin size of histogram features in fault diagnosis of mono-block centrifugal pump. Expert Syst Appl 38:7708–7717. https://doi.org/10.1016/j.eswa.2010.12.140

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. https://doi.org/10.1007/s00158-015-1310-4

Jung JH, Jeon BC, Youn BD, Kim M, Kim D, Kim Y (2017) Omnidirectional regeneration (ODR) of proximity sensor signals for robust diagnosis of journal bearing systems. Mech Syst Signal Process 90:189–207. https://doi.org/10.1016/j.ymssp.2016.12.030

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc Ser B Stat Methodol 63:425–464. https://doi.org/10.1111/1467-9868.00294

Kim H-S, Jang S-G, Kim N-H, Choi J-H (2016) Statistical calibration and validation of elasto-plastic insertion analysis in pyrotechnically actuated devices. Struct Multidiscip Optim 54:1573–1585. https://doi.org/10.1007/s00158-016-1545-8

Kim M, Kim Y, Yoo J, Wang J, Kim H (2017) Regularized speaker adaptation of KL-HMM for dysarthric speech recognition. IEEE Trans Neural Syst Rehabil Eng 25:1581–1591. https://doi.org/10.1109/TNSRE.2017.2681691

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Kwon KB, Jung J, Han JS (2018) Abnormal condition analysis and validation of RK4 multi axis rotor systems using finite element analysis. Trans Korean Soc Noise Vib Eng 28:204–213

Lee G, Yi G, Youn BD (2017) A comprehensive study of calibration metric for optimization-based model calibration. Paper presented at the 12th World Congress of Structural and Multidisciplinary Optimization, Braunschweig, Germany, June 5–9

Looman J, Campbell JB (1960) Adaptation of Sorensen's K (1948) for estimating unit affinities in prairie vegetation. Ecology 41:409–416. https://doi.org/10.2307/1933315

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

Oh H, Kim J, Son H, Youn BD, Jung BC (2016a) A systematic approach for model refinement considering blind and recognized uncertainties in engineered product development. Struct Multidiscip Optim 54: 1527–1541. https://doi.org/10.1007/s00158-016-1493-3

Oh H, Wei HP, Han B, Youn BD (2016b) Probabilistic lifetime prediction of electronic packages using advanced uncertainty propagation analysis and model calibration. IEEE Trans Compon Packag Manuf Technol 6:238–248. https://doi.org/10.1109/TCPMT.2015.2510398

Sahmani S, Fattahi AM (2017) Calibration of developed nonlocal anisotropic shear deformable plate model for uniaxial instability of 3D metallic carbon nanosheets using MD simulations. Comput Methods Appl Mech Eng 322:187–207. https://doi.org/10.1016/j.cma.2017.04.015

Sarin H, Kokkolaras M, Hulbert G, Papalambros P, Barbat S, Yang RJ (2010) Comparing time histories for validation of simulation models: error measures and metrics. J Dyn Syst Meas Control 132:061401–061401-061410. https://doi.org/10.1115/1.4002478

Shannon CE (1948) A mathematical theory of communication SIGMOBILE. Mob Comput Commun Rev 5:3–55. https://doi.org/10.1145/584091.584093

Shin H, Chang S, Yang S, Youn BD, Cho M (2016) Statistical multiscale homogenization approach for analyzing polymer nanocomposites that include model inherent uncertainties of molecular dynamics simulations. Compos Part B 87:120–131. https://doi.org/10.1016/j.compositesb.2015.09.043

Simoen E, De Roeck G, Lombaert G (2015) Dealing with uncertainty in model updating for damage assessment: a review. Mech Syst Signal Process 56:123–149. https://doi.org/10.1016/j.ymssp.2014.11.001

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

Wand MP (1997) Data-based choice of histogram bin width. Am Stat 51: 59–64. https://doi.org/10.2307/2684697

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

Youn BD, Jung BC, Xi Z, Kim SB, Lee WR (2011) A hierarchical framework for statistical model calibration in engineering product development. Comput Methods Appl Mech Eng 200:1421–1431. https://doi.org/10.1016/j.cma.2010.12.012

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.