Reliability Engineering and System Safety 148 (2016) 32–43

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Role of calibration, validation, and relevance in multi-level uncertainty integration

Chenzhao Li, Sankaran Mahadevan *

Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 10 February 2015
Received in revised form
26 October 2015
Accepted 7 November 2015
Available online 3 December 2015

Keywords:
Calibration
Validation
Uncertainty
Bayesian
Model reliability metric
Sobol indices

ABSTRACT

Calibration of model parameters is an essential step in predicting the response of a complicated system, but the lack of data at the system level makes it impossible to conduct this quantification directly. In such a situation, system model parameters are estimated using tests at lower levels of complexity which share the same model parameters with the system. For such a multi-level problem, this paper proposes a methodology to quantify the uncertainty in the system level prediction by integrating calibration, validation and sensitivity analysis at different levels. The proposed approach considers the validity of the models used for parameter estimation at lower levels, as well as the relevance at the lower level to the prediction at the system level. The model validity is evaluated using a model reliability metric, and models with multivariate output are considered. The relevance is quantified by comparing Sobol indices at the lower level and system level, thus measuring the extent to which a lower level test represents the characteristics of the system so that the calibration results can be reliably used in the system level. Finally the results of calibration, validation and relevance analysis are integrated in a roll-up method to predict the system output.

1. Introduction

Parameters of computational models are often calibrated using experimental data. For a complicated system it may be difficult to conduct full-scale experiments, but it may be possible to obtain data at lower levels of complexity (e.g., isolated physics or simpler configurations). Fig. 1 shows such a multi-level problem with two lower levels (G_1, G_2) and a system level (H). The lower levels and the system level constitute a hierarchy, and different levels have the same set of model parameters ($\boldsymbol{\theta}_m$) that need to be calibrated.

In order to predict the system level output when data are only available at lower levels, a reasonable route is to quantify the model parameters using lower level data, and propagate the results through the computational model at the system level. Several issues need to be addressed in realizing such a multi-level parameter estimation problem. First, even if model input and output are measured in the lower level tests, thereby forming pairwise input–output data, the calibration result can still be uncertain due to several sources, including: 1) model errors in the lower level computational models; 2) measurement errors in the experiments; and 3) sparse experimental data. Second, the existence of *multiple* lower levels provides multiple possibilities to conduct model calibration and leads to multiple calibration results. In a multi-level problem, model calibration can be conducted using the data from a single level or multiple levels. For the problem in Fig. 1 with two lower levels, 3 calibration options are possible: 1) calibration using the data and model from Level 1 alone; 2) calibration using the data and model from Level 2 alone; and 3) calibration using the data and models from both Level 1 and Level 2. Generally, if data are available at n different levels, $2^n - 1$ model calibration options are possible to quantify the uncertainty of model parameters [1].

This paper uses Bayesian inference for model calibration, thus the result of model calibration is a joint posterior distribution of model parameters. As Kennedy and O'Hagan [2] pointed out, the posterior distribution is the "best-fitting" results in the sense of representing the calibration data faithfully, not necessarily representing the true physical values. *The main objective of this paper is to determine the appropriate distribution for model parameters $\boldsymbol{\theta}_m$ to be used in system level prediction.* One possibility is to use all the lower level data in model calibration and propagate the resultant posterior distribution to predict the system level output. However, this result is conditioned on the event that both the models at Level 1 and Level 2 are valid, which may or may not be true [3]. This paper answers this question by assigning a "confidence" measure to each posterior distribution. Note that this paper is not using the term "confidence" in the same sense as is used in statistics (as in confidence interval). This "confidence" measure

* Corresponding author. Tel.: +1 615 322 3040.
E-mail address: sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

http://dx.doi.org/10.1016/j.ress.2015.11.013
0951-8320/

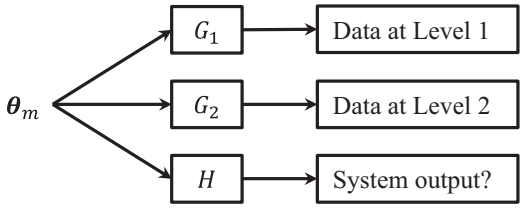

Fig. 1. Multi-level parameter estimation problem.

constitutes of two components: 1) the model validity at the corresponding lower level (one can think of this as local confidence regarding each lower level); 2) the relationship between the lower level and the system level, i.e., the relevance of the posterior distribution obtained at the lower level to the system level prediction problem (one can think of this as inter-level confidence). The relationship between two lower levels can be also important. However, this relationship is not considered here since in this paper the obtained information in a lower level is extrapolated to the system level, but not to another lower level.

Before quantifying the local confidence, the relationship between model calibration and model validation should be clarified. The purpose of model calibration is to adjust a set of parameters associated with a computational model so that the agreement between model prediction and experimental observation is maximized [4]. The term "model validation" has had different interpretations in different studies, and this paper follows the AIAA definition [5], i.e., model validation is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. Generally model validation is realized by comparing the model prediction against experimental data. Both model calibration and model validation are conducted in this paper, but they use different sets of experimental data (no calibration data is used in model validation). Comprehensive reviews on model validation can be found in [5–8]. A methodology for integrating model validation results from multiple experiments, each of which tests one part of the physics in the target application, can be found in [9].

Model calibration and model validation are distinct activities. Theoretically, for a computation model $F(\boldsymbol{\theta}_m;\boldsymbol{x})$ where $\boldsymbol{x}$ is a set of model inputs and $\boldsymbol{\theta}_m$ is a set of model parameters, model validation can be conducted exclusive of any model calibration [5] if the model parameters are assumed to be known. However, the model parameters $\boldsymbol{\theta}_m$ are often unknown. Therefore, prior to model validation, model calibration can be conducted to quantify the values of $\boldsymbol{\theta}_m$ or reduce the uncertainty about their values. The KOH framework [2] of model calibration used in this paper not only reduces the analyst's uncertainty about $\boldsymbol{\theta}_m$ by Bayesian inference, but also quantifies the model error $\delta(\boldsymbol{x})$ which is defined as the difference between model prediction and reality. The corrected prediction model under the KOH framework is $F(\boldsymbol{\theta}_m;\boldsymbol{x})+\delta(\boldsymbol{x})$. Compared to the original computational model, the new model is different in two aspects: 1) reduced uncertainty in $\boldsymbol{\theta}_m$; and 2) introduction of model error $\delta(\boldsymbol{x})$. In this paper, the model to be assessed in model validation is this "corrected" model. Thus validation is a subsequent and distinct activity after calibration in this paper. In other words, we consider model calibration and model validation as two distinct activities, and use two different sets of experimental data for these two activities, as suggested in [10,11]. Thus the calibration results of $\delta(\boldsymbol{x})$ and $\boldsymbol{\theta}_m$ within a single level do not change as a result of model validation in our approach.

With the calibration and validation perspectives to be used in this paper defined as above, the reason to use model validation to quantify the local confidence is explained next. In model validation, the assessed model validity of the corrected prediction model $F(\boldsymbol{\theta}_m;\boldsymbol{x})+\delta(\boldsymbol{x})$ at a lower level is a combined effect of three components: 1) $F(\boldsymbol{\theta}_m;\boldsymbol{x})$; 2) $\delta(\boldsymbol{x})$; and 3) the posterior distribution of $\boldsymbol{\theta}_m$. The third aspect corresponds to the "local confidence" (not to be confused with confidence intervals used in statistics), thus this paper takes the model validity as one factor affecting our confidence in extrapolating the posterior distribution of the model parameter from the lower level to the system level. This is reasonable since the model parameter has been calibrated with a model corresponding to the lower level experiment, and it is important to know whether the model was calibrated accurately; the calibration result is obviously affected by how accurately the lower level model represents the physics in the lower level experiment.

Model validation is about comparing the model prediction against experimental data, and a model validation metric is needed to quantify this comparison. Among the validation metrics in the literature, classical hypothesis testing gives an acceptance/rejection decision. Confidence intervals have also been calculated for the difference between model prediction and observed data [5]. Although the confidence intervals may provide a quantitative measure of the model validity at a single level, it is not possible to apply them in uncertainty propagation and integration across multiple levels, since the concept of propagation of confidence interval does not exist in classical statistics. Validation metrics resulting in a single quantitative value indicating the degree of model validity have also been developed. In Bayesian hypothesis testing [10,12], the posterior distribution obtained by model calibration is used as the null hypothesis and an alternative distribution is selected for the alternative hypothesis. The result of Bayesian hypothesis testing is a Bayes factor (the likelihood ratio between the null and alternate hypotheses), measuring the support from validation data to the null and alternate hypotheses. This is a relative measure significantly depending on the choice of distribution of the alternate hypothesis. In contrast, Ferson et al. [13,14] proposed an area metric, which is the difference between CDFs and has the same unit as the prediction/data. For the case that the model output is stochastic at fixed model input, this metric measures the area between the CDF of model output and the EDF (empirical distribution function) of experimental data at a fixed model input. (Note that in this paper model inputs $\boldsymbol{x}$ and model parameters $\boldsymbol{\theta}_m$ are different quantities, thus the model output can be stochastic at fixed model inputs $\boldsymbol{x}=\boldsymbol{x}^*$ if the model parameters $\boldsymbol{\theta}_m$ are still uncertainty. In addition, uncertain model errors, surrogate model uncertainty are other reasons that the model output can be stochastic at fixed model inputs) If data are from experiments with different inputs, this metric is still applicable by building a single EDF for all the data with u-pooling method [13].

The model validation metric used in this paper is the model reliability metric proposed by Rebba and Mahadevan [15] and further developed by Sankararaman and Mahadevan [16]. This metric measures the model validity by "model reliability", which is defined as the probability that the difference between model prediction and observed data is less than a pre-defined tolerance. Here the model prediction is stochastic, whose uncertainty is caused by the uncertainty in the posterior distribution of model parameters as well as the uncertainty regarding the model error. In other words, the model reliability metric considers the combined effect of these two sources of uncertainty. The value of model reliability is between 0 and 1, thus it can be conveniently used as a weighting term in subsequent uncertainty integration across multiple levels.

For a given validation data point, the model reliability is a deterministic value. However, its value is different for different data points. To capture this variability in model reliability, this paper proposes a stochastic model reliability metric where the model reliability is treated as a random variable instead of a

deterministic value. In addition, this paper extends the model reliability metric to handle multivariate output.

1) As mentioned earlier, the inter-level confidence to extrapolate a lower level posterior distribution to the system level is about the relationship between the lower level and the system level. In this paper, the relationship between the lower level and the system level is quantified by a proposed relevance analysis. The necessity of relevance analysis is explained here. An inherent assumption in the proposed relevance analysis is that if the physical configuration and inputs of a lower level experiment (say Level 2 in Fig. 1) is more similar to the system level than another lower level experiment (say Level 1 in Fig. 1), it is reasonable to assign higher confidence to the calibration result at this level (i.e., Level 2). Thus the relevance of the lower level to the system level is the degree to which the experimental configuration and inputs at a lower level reflects the physical characteristics of the system so that the calibration results can be reliably used in the system level prediction. The relevance decides the inter-level confidence on the calibration at lower levels, and influences the uncertainty integration. This paper proposes a method to quantify the relevance using Sobol indices and the cosine similarity of sensitivity vectors.

With the local confidence and inter-level confidence quantified, uncertainty integration is needed to aggregate all the available information from model calibration, model validation (for local confidence) and relevance analysis (for inter-level confidence). A roll-up methodology for uncertainty integration was proposed in [3], which results in the integrated distribution of model parameters as a weighted average of the posterior distributions, and the weight terms are the model reliability at lower levels. A brief introduction to this methodology is given in Section 5, and this paper extends it to incorporate more information from the lower levels, including: 1) the stochastic model reliability; and 2) the relevance between any lower level and the system level.

In summary, the motivation of this paper is to quantify the distributions of model parameters to be used in system level prediction, by using the available information at multiple levels from model calibration, model validation, relevance analysis, and uncertainty integration. The posterior distributions of model parameters are computed by Bayesian inference. The integration of multiple posterior distributions for each model parameter is assisted by model validation and relevance analysis, and realized in a proposed new roll-up method. This paper develops a methodology to compute the relevance using Sobol indices and cosine similarity of vectors. In model validation, the model reliability metric is extended to capture the variability in model reliability among different validation points and to consider multivariate output. Finally, the integrated distributions of model parameters are propagated through the computational model at the system level to predict the system output and quantify its uncertainty.

2. Model calibration

This section provides a brief summary of Bayesian inference for model calibration, as conducted in this paper. No new developments are reported; however, since calibration is the first step of the proposed methodology, a brief summary is given here for the sake of completeness.

Suppose the physical input–output relationship at a single level is described by a computational model $y_c = F(\boldsymbol{\theta}_m;\boldsymbol{x})$, where y_c is the computational model output, and $\boldsymbol{\theta}_m$ is a set of unknown model parameters, and $\boldsymbol{x}$ is the model input. Kennedy and O'Hagan (KOH) [2] expressed the relationship between the experimental observation z and the computational model as:

$$z = F(\boldsymbol{\theta}_m;\boldsymbol{x}) + \delta(\boldsymbol{x}) + \varepsilon_m \tag{1}$$

where $\delta(\boldsymbol{x})$ is the model error (input-dependent); ε_m is the measurement error which is usually assumed to be Gaussian distribution $N(0,\sigma_m^2)$. The model error $\delta(\boldsymbol{x})$ can be modeled using different formulations [17], which introduces more parameters. In addition, to reduce the computational effort, the computational model $F(\boldsymbol{\theta}_m;\boldsymbol{x})$ may be replaced by a surrogate model; several options such as polynomial response surface [18], polynomial chaos expansion [19], Gaussian process (GP) model [20,21] etc. are available. This paper uses the GP model. The parameters of $\delta(\boldsymbol{x})$ and the surrogate model for $F(\boldsymbol{\theta}_m;\boldsymbol{x})$ are also uncertain and need to be estimated. These parameters are also called hyper-parameters to distinguish them from model parameters $\boldsymbol{\theta}_m$. In sum, all the parameters to calibrate include: 1) model parameters $\boldsymbol{\theta}_m$; 2) hyper-parameters in the surrogate model for $F(\boldsymbol{\theta}_m;\boldsymbol{x})$; 3) hyper-parameters $\boldsymbol{\theta}_\delta$ of the model error $\delta(\boldsymbol{x})$; and 4) standard deviation σ_m of ε_m. The presence of so many calibration parameters is challenging if calibration data are sparse.

This paper ignores the hyper-parameter uncertainty in the GP model of $F(\boldsymbol{\theta}_m;\boldsymbol{x})$ for three reasons: 1) enough training points are used to build an accurate GP model with small variance in the GP prediction, thus the hyper-parameter uncertainty is expected to be small; 2) considering this hyper-parameter uncertainty will bring enormous computational effort [22] in model calibration and validation, whereas this hyper-parameter uncertainty is not the focus of this paper; and 3) the uncertainty in the hyper-parameters is typically negligible compared to actual model parameters [23]. Thus we first estimate the hyper-parameters of the GP model and then fix them as deterministic values in the subsequent calibration of model parameters $\boldsymbol{\theta}_m$. In addition, if the model input is fixed, then the input dependent model discrepancy $\delta(\boldsymbol{x})$ will become a single parameter δ. In the numerical example of this paper, for each lower level calibration test, the model/experimental input is fixed and so the vector of calibration parameters $\boldsymbol{\theta}$ includes: 1) model parameters $\boldsymbol{\theta}_m$; 2) model error δ; and 3) the standard deviation σ_m of measurement error ε_m.

In a multi-level problem, each lower level may provide data for multivariate output quantities, and each output quantity at any level has a corresponding model error $\delta(\boldsymbol{x})$ and measurement error standard deviation σ_m to be calibrated. In Fig. 1, if calibration data consist of two output quantities at Level 1, model calibration includes two model error terms and two measurement error terms; and if two output quantities at Level 2 are also included for calibration, model calibration includes four model error terms and four measurement error terms.

For the model error $\delta(\boldsymbol{x})$, we need to select the prior distribution for each hyper-parameter in the above formulation. But if model input $\boldsymbol{x}$ is fixed and the hyper-parameters are fixed, we only need to select a prior distribution for δ. In the numerical example in Section 6, since there is no information available on δ, a uniform prior distribution is assumed as $\delta \sim U(a,b)$ where a and b are the lower and upper bounds of the uniform distribution. The prior distribution of σ_m is chosen as the non-informative Jeffrey's prior $f'(\sigma_m) \propto 1/\sigma_m$, which is invariant under re-parameterization [24]. In addition, the prior distributions for $\boldsymbol{\theta}_m$ are constructed based on expert opinion.

With prior distributions for $\boldsymbol{\theta} = \boldsymbol{\theta}_m, \boldsymbol{\theta}_\delta, \sigma_m$ defined and experimental data at lower levels obtained, the Bayesian inference expresses the posterior distribution of $\boldsymbol{\theta}$ as:

$$f''(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}) f'(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}) f'(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}} \tag{2}$$

where $L(\boldsymbol{\theta})$ is the likelihood function of $\boldsymbol{\theta}$ and $f'(\boldsymbol{\theta})$ is the joint prior PDF of $\boldsymbol{\theta}$. The samples of $f''(\boldsymbol{\theta})$ are often generated numerically by

Markov Chain Monte Carlo (MCMC) methods [25]. Note that if the computational model $F(\boldsymbol{\theta}_m;\boldsymbol{x})$ is replaced by a GP model $GP(\boldsymbol{\theta}_m;\boldsymbol{x}) \sim N(\mu_s(\boldsymbol{\theta}_m;\boldsymbol{x}), \sigma_s^2(\boldsymbol{\theta}_m;\boldsymbol{x}))$, this paper not only considers its mean prediction $\mu_s(\boldsymbol{\theta}_m;\boldsymbol{x})$ but also its variance $\sigma_s^2(\boldsymbol{\theta}_m;\boldsymbol{x})$. Therefore Eq. (1) will change to $z = N(\mu_s(\boldsymbol{\theta}_m;\boldsymbol{x}), \sigma_s^2(\boldsymbol{\theta}_m;\boldsymbol{x})) + \delta(\boldsymbol{x}) + N(0, \sigma_m^2)$, and the likelihood function $L(\boldsymbol{\theta})$ is established based on this modified equation so that the surrogate model uncertainty is also incorporated in model calibration.

3. Model validation

As mentioned in Section 1, a multi-level problem with n lower levels can provide $2^n - 1$ alternative model calibration results, but model calibration cannot answer the question regarding how to integrate them. Thus model validation is necessary to assess the validity of the model calibration before using the calibrated model parameters for system output prediction.

In this paper, the basic concept in uncertainty integration is to combine all the information from lower levels and results in an integrated distribution of model parameter θ as the weighted average of multiple posterior distributions. To make the integrated distribution as a valid PDF, the sum of the weight terms computed in model validation should be unity. The model reliability metric directly satisfies this requirement and is selected in this paper.

Section 3.1 introduces the model reliability metric; Section 3.2 extends it to consider the model reliability as a stochastic variable to aggregate the validation results at different validation points; and Section 3.3 extends the model reliability metric to deal with multivariate output.

3.1. Model reliability metric

In model reliability metric, for a specific application, the model is defined to be valid if the difference between the model prediction y and the corresponding validation measurement is less than a predefined tolerance λ. Due to the measurement error ($\varepsilon_m \sim N(0, \sigma_m^2)$), the measurement is actually a random variable. For a single observed value D, this random variable is denoted by d with mean value D and standard deviation σ_m, i.e. $d \sim N(D, \sigma_m^2)$. Let G denote the event that the model is valid, then the model reliability is defined as the probability of event G:

$$P(G|D) = P(|y - d| < \lambda) \tag{3}$$

The probability in Eq. (3) is used as a metric to measure model validity, thus this metric is named as "model reliability metric". If y and σ_m are deterministic, Eq. (4) computes the model reliability where ϵ is a dummy variable for integration:

$$P(G|D) = \int_{-\lambda}^{\lambda} \frac{1}{\sigma_m\sqrt{2\pi}} \exp\left[-\frac{(\epsilon - (y - D))^2}{2\sigma_m^2}\right] d\epsilon \tag{4}$$

In this paper, the model prediction y refers to the computational model output corrected by the model error, i.e., $y = F(\boldsymbol{\theta}_m;\boldsymbol{x}) + \delta(\boldsymbol{x})$. Although model input $\boldsymbol{x}$ is known, the model prediction y is still stochastic due to the uncertainty of $\delta(\boldsymbol{x})$ and $\boldsymbol{\theta}_m$. Furthermore, another calibration parameter σ_m can be also uncertain. In this case, the model reliability is:

$$P(G|D) = \int P(G|\boldsymbol{\theta}, D) f''(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{5}$$

where $P(G|\boldsymbol{\theta}, D)$ is given by the right side of Eq. (4), and $f''(\boldsymbol{\theta})$ is the joint posterior distribution of $\boldsymbol{\theta} = \boldsymbol{\theta}_m, \boldsymbol{\theta}_\delta, \sigma_m$. Note that if the computational model $F(\boldsymbol{\theta}_m;\boldsymbol{x})$ is replaced by a GP model $GP(\boldsymbol{\theta}_m;\boldsymbol{x}) \sim N(\mu_s(\boldsymbol{\theta}_m;\boldsymbol{x}), \sigma_s^2(\boldsymbol{\theta}_m;\boldsymbol{x}))$, this paper not only considers its mean prediction $\mu_s(\boldsymbol{\theta}_m;\boldsymbol{x})$ but also its variance $\sigma_s^2(\boldsymbol{\theta}_m;\boldsymbol{x})$, thus the model prediction will be $y = N(\mu_s(\boldsymbol{\theta}_m;\boldsymbol{x}), \sigma_s^2(\boldsymbol{\theta}_m;\boldsymbol{x})) + \delta(\boldsymbol{x})$. Then the model reliability in Eq. (5) is computed based on this formula so that the surrogate model uncertainty is also incorporated in model validation.

Eqs. (4) and (5) are only suitable for a single observed value D from an output quantity. If multiple data points are observed for an output quantity (i.e., multiple validation experiments), then Eqs. (4) and (5) are not correct. Model validation is further complicated if experimental data are observed for a multivariate output and multiple validation data points are available. Therefore the concept of the model reliability metric needs to be extended to deal with multiple data points and multivariate output. The first issue will be addressed in Section 3.2 by proposing a stochastic model reliability metric, while the second issue will be addressed in Section 3.3.

3.2. Stochastic model reliability metric

As shown in Eqs. (4) and (5), the value of model reliability $P(G)$ is deterministic at a single data point D, but changes over different data points. If model inputs $\boldsymbol{x}$ of these data points are known, a mathematical function $P(G|\boldsymbol{x}) = S(\boldsymbol{x})$ can be established where $P(G|\boldsymbol{x})$ is the model reliability at model input $\boldsymbol{x}$. However, this function may be not accurate due to validation data sparseness (only five validation points are available in the numerical example in Section 6). Thus constructing a mathematical function for model reliability (as a function of $\boldsymbol{x}$) is not considered in this paper. Instead, this paper uses a probability distribution to represent the variability in $P(G)$, and this distribution is constructed using the model reliability values at different validation data points. (The first option could be considered if a large number of validation experiments are conducted).

In this paper, model reliability $P(G)$ is assumed to have a beta distribution since $P(G) \in [0, 1]$ and the sample space of beta distribution is also the interval $[0, 1]$. If a data set $\boldsymbol{D} = D_1, D_2, \cdots, D_n$ of one output quantity is observed for model validation from n experiments with different inputs, the corresponding model reliability values computed by Eq. (5) at each experiment are $\boldsymbol{D}_R = D_{R1}, D_{R2}, \cdots, D_{Rn}$. Using $\boldsymbol{D}_R$, several methods can be used to construct the PDF of model reliability, such as the method of maximum likelihood, method of moments, or Bayesian inference. This paper uses the method of moments to construct the PDF of $P(G)$. In summary, this approach gives a stochastic representation of model reliability, i.e., $P(G)$ is not a single value but represented by a probabilistic distribution. The next section extends the model reliability metric to deal with multivariate output.

3.3. Model reliability metric with multivariate output

If K output quantities are observed in a validation experiment, we have a set of K models sharing the same model input and model parameters:

$$\boldsymbol{y} = \boldsymbol{F}(\boldsymbol{\theta}_m;\boldsymbol{x}) + \boldsymbol{\delta}(\boldsymbol{x}) \leftrightarrow \begin{cases} y_1 = F_1(\boldsymbol{\theta}_m;\boldsymbol{x}) + \delta_1(\boldsymbol{x}) \\ y_2 = F_2(\boldsymbol{\theta}_m;\boldsymbol{x}) + \delta_2(\boldsymbol{x}) \\ \cdots \\ y_K = F_K(\boldsymbol{\theta}_m;\boldsymbol{x}) + \delta_K(\boldsymbol{x}) \end{cases} \tag{6}$$

where $F_j(\boldsymbol{\theta}_m;\boldsymbol{x})$ and $\delta_j(\boldsymbol{x})$ ($j = 1 \text{to} K$) are the computational model and model error of the $j$thquantity. Each quantity also has a measurement error $\varepsilon_{m_j} \sim N(0, \sigma_{m_j}^2)$ and the corresponding variable $z_j = y_j + N(0, \sigma_{m_j}^2)$ representing the measurement. We denote $\boldsymbol{z} = \{z_1, \ldots, z_j, \ldots z_K\}^T$. Assume that n experiments are conducted. In the ith experiment ($i = 1$ to n), data points for K quantities form a data set $\boldsymbol{D}_i = \{D_{i1}, \ldots, D_{ij}, \ldots, D_{iK}\}^T$. In addition, the pre-defined tolerance for each quantity is included in a vector $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_j, \ldots, \lambda_K\}^T$.

The distance between $\boldsymbol{z}$ and $\boldsymbol{D}_i$ can be measured by multiple distance functions such as the Euclidean distance, Chebyshev distance, Manhattan distance, and Minkowski distance [26]. This paper uses the Mahalanobis distance [27]. The Mahalanobis distance between $\boldsymbol{z}$ and $\boldsymbol{D}_i$ is defined as $M=\sqrt{(\boldsymbol{z}-\boldsymbol{D}_i)^T\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}(\boldsymbol{z}-\boldsymbol{D}_i)}$ where $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ is the covariance matrix of $\boldsymbol{z}$. The Mahalanobis distance transfers $\boldsymbol{z}$ and $\boldsymbol{D}_i$ into the normalized principal component (PC) space [27] by using $\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}$. Compared to other distance functions, the Mahalanobis distance brings two advantages: 1) the correlations between output quantities are considered; and 2) the output quantities are normalized to the same scale to prevent any quantity from dominating the metric simply due to large numerical values. Using the Mahalanobis distance, the model reliability for multivariate output is defined as:

$$P(G|\boldsymbol{D}_i)=P(M<\lambda_M)=P\left(\sqrt{(\boldsymbol{z}-\boldsymbol{D}_i)^T\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}(\boldsymbol{z}-\boldsymbol{D}_i)}<\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}\boldsymbol{\lambda}}\right) \quad (7)$$

where $\lambda_M=\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\boldsymbol{z}}^{-1}\boldsymbol{\lambda}}$ is the normalized tolerance.

Generally the posterior distributions obtained in model calibration are numerical samples generated by MCMC, so the subsequent model reliability in Eqs. (4) and (5) is also computed numerically. Numerical computation also facilitates the realization of the extended model reliability in Eq. (7). Here the model reliability is expressed as:

$$\begin{aligned}P(G|\boldsymbol{D}_i)=P(M<\lambda_M|\boldsymbol{D}_i)&=\int_0^{\lambda_M}f(M|\boldsymbol{D}_i)\mathrm{d}M\\&=\int_0^{\lambda_M}\left(\int f(M|\boldsymbol{D}_i,\boldsymbol{\theta})f''(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right)\mathrm{d}M\end{aligned} \quad (8)$$

Eq. (8) indicates a numerical algorithm to compute the model reliability:

1. Generate a random sample of $\boldsymbol{\theta}$ from its posterior distribution $f''(\boldsymbol{\theta})$;
2. Generate a sample of M conditioned on $\boldsymbol{\theta}$ by generating a sample of $\boldsymbol{z}$ and computing its Mahalanobis distance from $\boldsymbol{D}_i$;
3. Repeat steps 1 and 2 to obtain N samples of M; these samples can be used to construct the distribution $f(M\boldsymbol{D}_i)$, which is not conditioned on $\boldsymbol{\theta}$;
4. If N' out of N samples in step 3 satisfy $M<\lambda_M$, the model reliability is $P(G\boldsymbol{D}_i)=N'/N$.

The model reliability $P(G\boldsymbol{D}_i)$ by Eq. (8) is regarding a single experiment and $P(G\boldsymbol{D}_i)$ is a deterministic value. Thus n experiments will give n different model reliability values $\{P(G\boldsymbol{D}_1),\ldots,P(G\boldsymbol{D}_n)\}$. As proposed in Section 3.2, these values can be used to build a probability distribution for the model reliability $P(G)$, by treating $P(G)$ as a random variable instead of a deterministic value.

4. Relevance analysis

Section 1 explains the necessity to assign larger weight to the level physically "closer" or more relevant to the system level than the other. For instance, to predict the battery temperature of a spacecraft on the way to Mars, the data of the same quantity collected from its journey to the Moon will be more valuable than the data collected in any laboratory experiment on earth, since the former ones come from a physical environment more similar to the system of interest. Hence this section develops a method for relevance analysis, which measures the degree to which the experimental configuration and inputs at a lower level reflect the physics captured in the system-level model. Currently such measure is only intuitive and qualitative; an objective quantitative measure of relevance is needed for uncertainty integration.

The methodology to measure relevance should have two desired features. First, the defined methodology needs no mathematical details of the model in each level, since the model in each level could be a black box. Second, the resultant relevance measure can be used conveniently as a weighting term in uncertainty integration. To fulfill these two criteria, a relevance analysis using Sobol indices is proposed in this section.

Consider a model $Y=F(\boldsymbol{X})$ where $\boldsymbol{X}=X^1,\ldots,X^N$ is a vector containing all the inputs. Sensitivity analysis measures the contribution of each input to the uncertainty of $\boldsymbol{Y}$ [28]. Compared to local sensitivity analysis, global sensitivity analysis (GSA) considers the entire probability distribution of the input, not just the contribution at a local point. The Sobol indices for GSA have been developed in the literature based on the variance decomposition theorem [29], including first-order index and total effects index. For a particular input X^i, its first-order index is $S_1^i=V(E(Y|X^i))/V(Y)$; and its total effects index is $S_T^i=1-V(E(Y|X^{-i}))/V(Y)$ where X^{-i} means all the inputs other than X^i. The first-order index S_1^i measures the contribution of X^i by itself, and the sum of first-order indices of all inputs is always less than or equal to unity. The difference between this sum and unity is the contribution of the interaction among inputs. In contrast, the total effects index S_T^i contains not only the contribution of X^i, but also the interaction effect of X^i with other inputs. The interaction between variables will be ignored if the first-order index is used, thus this paper uses the total effects index to develop a method to quantify the relevance. In the following discussion the term sensitivity index indicates the total effects index.

Without loss of generality, this paper takes the multi-level problem in Fig. 1 for the illustration of relevance analysis. To predict the system output y_s (such as the maximum acceleration at the top mass in the numerical example in Section 6), the same quantity is also measured at lower levels (in the numerical example the maximum acceleration at the top mass is also measured at Level 1 and Level 2). The three prediction models for this quantity at different levels are $y_{L_1}=GP_{L_1}(\boldsymbol{\theta}_m,\boldsymbol{x}_{L_1})+\delta_{L_1}(\boldsymbol{x}_{L_1}), y_{L_2}=GP_{L_2}(\boldsymbol{\theta}_m,\boldsymbol{x}_{L_2})+\delta_{L_2}(\boldsymbol{x}_{L_2}), y_{L_1}=GP_{L_s}(\boldsymbol{\theta}_m,\boldsymbol{x}_{L_s})$ where $\boldsymbol{\theta}_m$ are model parameters and $\boldsymbol{x}_{L_1},\boldsymbol{x}_{L_2},\boldsymbol{x}_s$ are the model inputs at each level. Note that 1) the computational models are replaced by the GP models to improve computational efficiency; 2) model errors are considered in Level 1 and Level 2; and 3) model error at the system level is not considered since no information on it is available. These prediction models are stochastic, i.e., the output is stochastic even at fixed values of model inputs and model parameters. However, the Sobol indices computation requires a deterministic model, i.e., deterministic output at given values of model inputs and model parameters. This paper applies the auxiliary variable methodology based on the probability integral transform, as developed in Refs. [30,31], to obtain a deterministic value of the output for a given realization of inputs and model parameters; thus the Sobol indices can be computed.

Assume model parameters, model inputs, auxiliary variables constitute N_{L_1} elements in total at Level 1; since each element has a corresponding sensitivity index, a N_{L_1}-dimensional vector V_{L_1} of sensitivity indices will be obtained at Level 1. Similarly, a N_{L_2}-dimensional sensitivity vector V_{L_2} will be obtained at Level 2 and a N_s-dimensional sensitivity vector V_s will be obtained at the system level.

Rigorously, measuring the relevance requires comparing the mathematical model of the lower level and the mathematical model of the system level. However, this comparison is not easy if the models at different levels have distinct formats and are addressing different physical configurations (3-mass-spring vs. 3-mass-spring-on-beam in the numerical example) and are under different inputs (sinusoidal inputs vs. random process inputs in the

numerical example). Further, the model sometimes may be a black box; thus we cannot access its mathematical details and a direct comparison would be difficult. The obtained sensitivity vectors quantify the contribution of each model input/parameter towards the uncertainty in the model output. In other words, the sensitivity vector indicates which model input/parameter is more important in affecting the model output uncertainty. Actually whether the model input/parameter is important is determined by the physics of the model, thus the sensitivity vector is a representative of the physics, to the extent that the model represents the physics accurately. Therefore, this paper considers the sensitivity vector as an indicator of the physics captured in the model. (Of course, how well the physics is captured in the model is already indicated by the model reliability metric); thus the comparison of the vectors from two different levels is used to quantify the relevance between these two levels.

One issue in the comparison of $V_{L_i}(i=1,2)$ and V_s is that they may have different sizes (N_{L_1}, N_{L_2}, N_s may not be equal to each other) and some elements in one vector may not be present in the other vector. The shared dimensions of V_{L_i} and V_s are model parameters $\boldsymbol{\theta}_m$; and the unshared dimensions are the different model inputs and auxiliary variables at each level. To solve this problem we add the unshared dimension in V_{L_i} or V_s to the other vectors but set the corresponding sensitivity indices as zero since the added dimensions have no effect in the computation of the original sensitivity vector. Thus all the vectors V_{L_i} or V_s are brought to the same size.

Several methods are available to compare two vectors, such as Euclidean distance [26], Manhattan distance [26], Chebyshev distance [26], and cosine similarity [26,32]. To include the relevance in the subsequent uncertainty integration conveniently, we define the relevance index R as the square of cosine similarity of the sensitivity vectors, where the cosine similarity is the normalized dot product of two vectors:

$$R=\left(\frac{V_{L_i}\bullet V_s}{\|V_{L_i}\|\,\|V_s\|}\right)^2 \tag{9}$$

In other words, the above relevance index is the square of the cosine value of the angle between two sensitivity vectors, the elements in which are all positive. If the angle is zero, the relevance among the two levels is 1; if the two vectors are perpendicular, the relevance is 0.

In addition, this definition of relevance generates a value on the interval [0, 1]; and its complement, the square of the sine value, indicates physical non-relevance; hence the sum of "relevance" and "non-relevance" is the unity. Here the relevance index is a plausibility model for the proposition "The lower level model reflects the physical characteristics of the system level model", and the plausibility of this proposition is the relevance index. Based on Cox's theorem [33], this plausibility model is isomorphic to probability, since 1) the relevance index is a real value depending on the information of sensitivity vectors we obtained; and 2) the relevance index changes sensibly as the sensitivity vectors change. Thus the relevance index can be converted to probability by scaling, which has been done since the relevance index defined in Eq. (9) is already on the interval [0, 1]. Therefore in the roll-up methodology proposed in Section 5, we treat the relevance index as a probability and conveniently include it as a weighting term in the uncertainty integration.

However, the relevance index is only calculated based on the prediction models at each level, and data at lower levels; but no system-level observation data is assumed to be available. Therefore, if the system-level model does not capture the system-level physics very well, the relevance index cannot capture the effect of this discrepancy. Thus the proposed relevance index approach is not a fully physics-based approach, and does not provide a comprehensive comparison of the actual physics at different levels. However, the sensitivity vector does provide an indication of the *physics captured in the models* through variance decomposition, and we seek to include this information in the distributions of those system level model parameters that are inferred using lower level tests and models.

When the system-level model has additional physics, there may be additional parameters in the system level model to reflect this. The sensitivity vector of the system level model will quantify the contribution of these additional parameters, as well as the contribution of the parameters shared with the lower level models. The relevance index is based on the dot product of sensitivity vectors for the models at two different levels. Therefore, if the additional physics parameters in the system-level model have a significant contribution, then the physics in the Level 1 model may not be closely related to the physics in the system-level model. In that case, the two corresponding sensitivity vectors will diverge, and the relevance index of Level 1 will be small. Similarly, if the physics in the Level 2 model is not closely related to the physics in the system-level model, the relevance index of Level 2 will be small.

A further question arises in the computation of relevance index. Sobol indices consider the entire distribution of the influencing variable, but the posterior distribution of $\boldsymbol{\theta}_m$ (to be used in system level prediction) is unknown before the uncertainty integration. In order to solve this problem, a straightforward iterative algorithm to compute the relevance index R is proposed below:

1. Set an initial value of R.
2. Obtain the integrated distribution of each model parameter using the current relevance and the proposed roll-up method in Section 5 below.
3. Use the integrated distributions from step 2 to compute the sensitivity indices, and re-compute the updated relevance index R.
4. Repeat steps 2 and 3 until the relevance index R converges.

Thus, the results of calibration and validation at each lower level, and relevance indices between the lower levels and the system level have been obtained. The next task is to construct the integrated distribution of the system level model parameters and predict the system output.

5. Uncertainty integration

For a multi-level problem, the purpose of uncertainty integration is to combine all the available information (from calibration, validation and relevance analysis) from the lower levels and predict the response at the system level. In this paper the information from the lower level includes: 1) the posterior distributions from model calibration by considering data at each individual lower level, as well as data from multiple lower levels; 2) the model reliability distributions from model validation at each lower level; and 3) the relevance indices between each lower level and the system level. A roll-up methodology has been proposed in [3] for uncertainty integration. For the multi-level problem in Fig. 1, this methodology results in an integrated distribution [34] for a model parameter $\theta \in \boldsymbol{\theta}_m$:

$$\begin{aligned} f\left(\theta\middle|D_1^{C,V},D_2^{C,V}\right) &= P(G_1)P(G_2)f\left(\theta\middle|D_1^{C},D_2^{C}\right)+P(G_1')P(G_2)f\left(\theta\middle|D_2^{C}\right) \\ &\quad +P(G_1)P(G_2')f\left(\theta\middle|D_1^{C}\right)+P(G_1')P(G_2')f(\theta) \end{aligned} \tag{10}$$

In Eq. (10) the integrated distribution $f\left(\theta\middle|D_1^{C,V},D_2^{C,V}\right)$ is a weighted average of multiple posterior distributions and contains four terms: in the first term the posterior distribution $f(\theta|D_1^C,D_2^C)$

uses the calibration data of both Level 1 and Level 2 and its weight $P(G_1)P(G_2)$ is the probability that both of the models are valid; in the second and third terms the posterior distribution $f(\theta|D_i^C)$ uses the calibration data at Level i alone and its weight is the probability that the model at Level i is valid but the model at another level is invalid; in the last term the weight $P(G_1')P(G_2')$ of the prior distribution $f(\theta)$ is the probability that both of the models are invalid. After obtaining the integrated distributions for all the parameters in $\boldsymbol{\theta}_m$, the system response can be predicted by propagating all these integrated distributions through the computational model of the system level.

Obviously, the weight of each PDF on the right hand side of Eq. (10) is purely decided by model validation. This paper proposes an extension of Eq. (10) to include two additional concepts:

1. *Stochastic model reliability*: the model reliability $P(G_i)$ in Eq. (10) is a deterministic value, where G_i is the event that the model at Level i is valid; and this paper proposes the stochastic model reliability metric, where $P(G_i)$ is a random variable with PDF $f(P(G_i))$ as explained in Section 3.2;
2. *Relevance index*: this has been defined in Section 4 as the square of the cosine value of the angle between the sensitivity vectors at lower level and system level. We treat the relevance index similar to probability in the roll-up methodology, based on Cox's theorem. If S_i denotes the event that Level i is relevant to the system level, then the probability $P(S_i|G_i)$ is equal to the value of the relevance index R; this probability is conditioned on G_i since the computation of the relevance index uses the model at Level i; in contrast $P(S_i'|G_i)$ denotes the probability of non-relevance, and is equal to $1-R$.

The roll-up formula in Eq. (10) can be extended to consider stochastic model reliability by rewriting the left hand side as $f\left(\theta D_1^{C,V}, D_2^{C,V}, P(G_1), P(G_2)\right)$ and averaging it over $f(P(G_1))$ and $f(P(G_2))$. But a new formula is required to include the relevance index. Take the multi-level problem in Fig. 1 as an example. The integrated distribution of a model parameter θ conditioned on the calibration and validation data and model reliability $P(G_i)(i=1,2)$ is redefined as:

$$\begin{aligned} f\left(\theta|D_1^{C,V}, D_2^{C,V}, P(G_1), P(G_2)\right) = {} & P(G_1G_2S_1S_2)f\left(\theta|D_1^C, D_2^C\right) \\ & + P(G_1S_1 \cap (G_2' \cup S_2'))f\left(\theta|D_1^C\right) \\ & + P(G_2S_2 \cap (G_1' \cup S_1'))f\left(\theta|D_2^C\right) \\ & + P((G_1' \cup S_1') \cap (G_2' \cup S_2'))f(\theta) \end{aligned} \tag{11}$$

From the view of generating samples, Eq. (11) indicates two criteria: 1) whether a level is relevant to the system level; 2) whether a level has a valid model. A sample of θ is generated from $f\left(\theta D_1^C, D_2^C\right)$ only when both levels satisfy both criteria; a sample of θ is generated from $f\left(\theta D_i^C\right)$ if level i satisfies both criteria but the other level does not; and a sample of θ is generated from the prior distribution $f(\theta)$ if neither level satisfies both criteria. By assuming independence of model validity and relevance between different lower levels, the weight terms in Eq. (11) are computed by using the values of $P(G_i), P(S_i|G_i)$ and two fundamental probability relationships: $P(G_iS_i) = P(G_i)P(S_iG_i), P(G_i' \cup S_i') = 1 - P(G_iS_i)$. Eq. (11) also implies the option of "using only data from one level". If both the model validity and relevance are 1 for Level 1, and either model validity or relevance is 0 for Level 2, Eq. (11) reduces to $f\left(\theta D_1^{C,V}, D_2^{C,V}\right) = f\left(\theta D_1^C\right)$, i.e., only Level 1 data is used.

The integrated distribution of θ, which is conditioned on both calibration and validation data, can now be computed as:

$$f\left(\theta|D_1^{C,V}, D_2^{C,V}\right) = \iint f\left(\theta|D_1^{C,V}, D_2^{C,V}, P(G_1), P(G_2)\right) f(P(G_1))f(P(G_2))\mathrm{d}P(G_1)\mathrm{d}P(G_2) \tag{12}$$

Eqs. (11) and (12) express the proposed approach to integrate calibration, validation and relevance results at lower levels. Note that Eq. (12) accounts for stochastic model reliability. The analytical expression of $f\left(\theta|D_1^{C,V}, D_2^{C,V}\right)$ is difficult to derive since the results we collect in model calibration and validation are all numerical. A single loop sampling approach is proposed to construct $f\left(\theta|D_1^{C,V}, D_2^{C,V}\right)$ numerically, as follows:

1. Generate a sample of $P(G_1)$ and $P(G_2)$ from their distributions.
2. Compute the weight terms in Eq. (11). Divide the interval [0, 1] into four ranges; the length of the kth range is equal to the value of the kth weight in Eq. (11).
3. Generate a random number from the uniform distribution $U(0,1)$.
4. Generate a sample of θ using stratified sampling, i.e., from $f(\theta|D_1^C, D_2^C)$ if the random number in step 3 is located in the first range; from $f(\theta|D_1^C)$ if located in the second range; from $f(\theta|D_2^C)$ if located in the third domain; from $f(\theta)$ if located in the fourth domain.
5. Repeat steps 1 to 4 to obtain multiple samples of θ; then construct the PDF $f\left(\theta|D_1^{C,V}, D_2^{C,V}\right)$ by any method such as kernel density estimation [35].

After obtaining the integrated distributions of all the model parameters, the final step is to propagate the integrated distributions through the computational model of the system of interest to predict the system level output. This can be done by Monte Carlo sampling or other preferred stochastic analysis methods. Due to the uncertainty in the model parameters, the predicted system output will also be stochastic, and its distribution can be constructed by kernel density estimation. The distribution of the system output now systematically includes the contributions form calibration and validation activities at lower levels, and also accounts for the relevance of the lower levels to the actual system.

6. Numerical example

6.1. Problem description

A multi-level structural dynamics challenge problem provided by Sandia National Laboratories [36] is used to illustrate the methodology developed in Sections 2–5. As shown in Fig. 2, Level 1 contains three mass-spring-damper dynamic components in series, and a sinusoidal force input $P = 300\ \sin(500t)$ is applied to m_1. At Level 2, the dynamic system is mounted on a beam supported by a hinge at one end and a spring at the other end; a sinusoidal force input $P = 3000\ \sin(350t)$ is applied on the beam. The configuration of the system level is the same as Level 2, but the input is a random process loading (indicating difference in usage condition). Here Level 1 and Level 2 are defined as lower levels, and experimental data are assumed to be available only at the lower levels. All levels share six model parameters: three spring stiffnesses $k_i(i=1,2,3)$ and three damping ratios $\zeta_i(i=1,2,3)$; and they are assumed to be deterministic but unknown parameters, which are to be calibrated. The units of all quantities are non-dimensional.

Suppose 10 experiments are conducted at each of Level 1 and Level 2; and the displacement, velocity and acceleration history at each degree of freedom are recorded. Six quantities at each lower

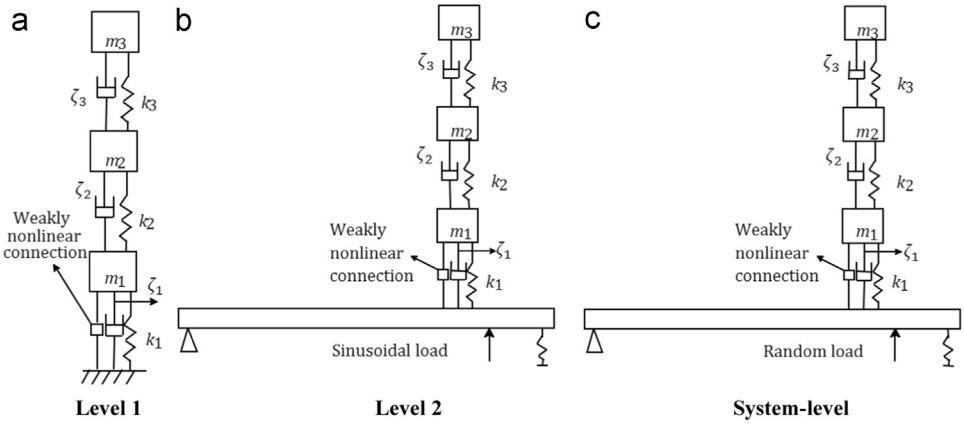

Fig. 2. Structural dynamics challenge problem.

Table 1
Synthetic experimental data at Level 1.

	Calibration data				Validation data					
A_1	10,749	8146	9195	9500	10,185	9940	10,233	9887	9837	10,409
A_2	6362	6827	6780	5759	6319	6579	6346	6730	6160	6126
A_3	1509	1465	1431	1556	1512	1416	1288	1293	1548	1360
D_1	93,230	93,059	84,033	86,102	92,717	84,258	89,758	95,249	85,275	90,709
D_2	8110	7283	8377	8590	8736	7490	8407	8127	8710	8477
D_3	33,948	30,740	30,693	34,290	24,536	34,579	31,193	29,959	33,172	33,723

level are extracted from these records as the synthetic experimental data in model calibration and validation:

1) $A_i(i=1,2,3)$: the maximum acceleration in the ith mass;

2) $D_i(i=1,2,3)$: the energy dissipated by the ith damper in 1000 time units.

The synthetic experimental data are listed in Tables 1 and 2. The data points for each quantity from the first five tests are selected as calibration data and the rest as validation data.

Computational models for the three levels have been established. The method to solve the dynamic problem at Level 1 can be found in structural dynamics text books [37]; and the computational models using the finite element method for Level 2 and the system level are provided by Sandia National Laboratories [32].

Since the model input at each level is fixed, the input-dependent model error is an unknown deterministic value. Thus the parameters to be calibrated in this example are: the spring stiffnesses $k_i(i=1,2,3)$, the damping ratios $\zeta_i(i=1,2,3)$, model error δ and the output measurement error standard deviation σ_m if the data of the corresponding quantity are used in model calibration. Based on expert opinion, suppose the prior distribution of each k_i and ζ_i is assumed to be lognormal with a coefficient of variation of 10% and mean values of $\mu_{k_1}=5000$, $\mu_{k_2}=9000$, $\mu_{k_3}=8000$, $\mu_{\zeta_i}=0.025$ $(i=1,2,3)$. The prior distribution of model error is assumed to be uniform, i.e., $\delta \sim U(a,b)$ and the prior of σ_m is Jeffrey's prior $f'(\sigma_m) \propto 1/\sigma_m$.

The objective in this numerical example is to quantify the uncertainty in the prediction of maximum acceleration at m_3 in the system level, by using available models and experimental data.

Since as many as six quantities are measured, we can choose any combination of these six quantities in the analysis. Measurement data on more output quantities reduce the uncertainty in the system output prediction, but the computational effort will also increase and each quantity will bring two more related terms (δ and σ_m) for calibration. For the sake of brevity, only the calibration and validation results using the test data for all six quantities are provided below. But a plot showing the reduction in the uncertainty of system output prediction with the increase of output quantity measurements is also provided at the end.

Table 2
Synthetic experimental data at Level 2.

	Calibration data				Validation data					
A_1	3876	4110	4372	4187	4443	4486	3912	4237	4394	4807
A_2	4316	4051	4488	3947	4596	4347	5008	4930	4455	4809
A_3	3648	4133	4311	4558	4126	4410	4037	4380	4523	4277
D_1	8593	9009	8966	8910	9746	8606	8644	8757	9050	8458
D_2	1566	1563	1749	1616	1602	1718	1577	1597	1614	1451
D_3	2490	2975	2679	2891	3017	2654	2834	3021	2983	3121

6.2. Results and analysis

In order to reduce the computational effort, Gaussian process (GP) surrogate models are established to replace the computational models for all the output quantities. The surrogate model uncertainty introduced by the GP models is incorporated in model calibration and validation, as explained in Sections 2 and 3.1. The calibration results of k_i and ζ_i using the calibration data of the six output quantities at different levels are shown in Fig. 3, including all the PDFs needed in Eq. (11). As more data are used in the calibration, the uncertainty of the model parameters will decline. Thus Fig. 3 shows that the posterior distributions using the data at both levels always have less uncertainty than those using data at a single level. The difference between the posterior distributions within each sub-figure also indicates that the posterior distribution is a best-fitting result in the sense of representing that particular data-set, but we do not yet know how to combine these alternatives in the subsequent prediction. This is answered by model validation and relevance analysis.

Next model validation is performed using the stochastic model reliability metric with multivariate output. The tolerance for each

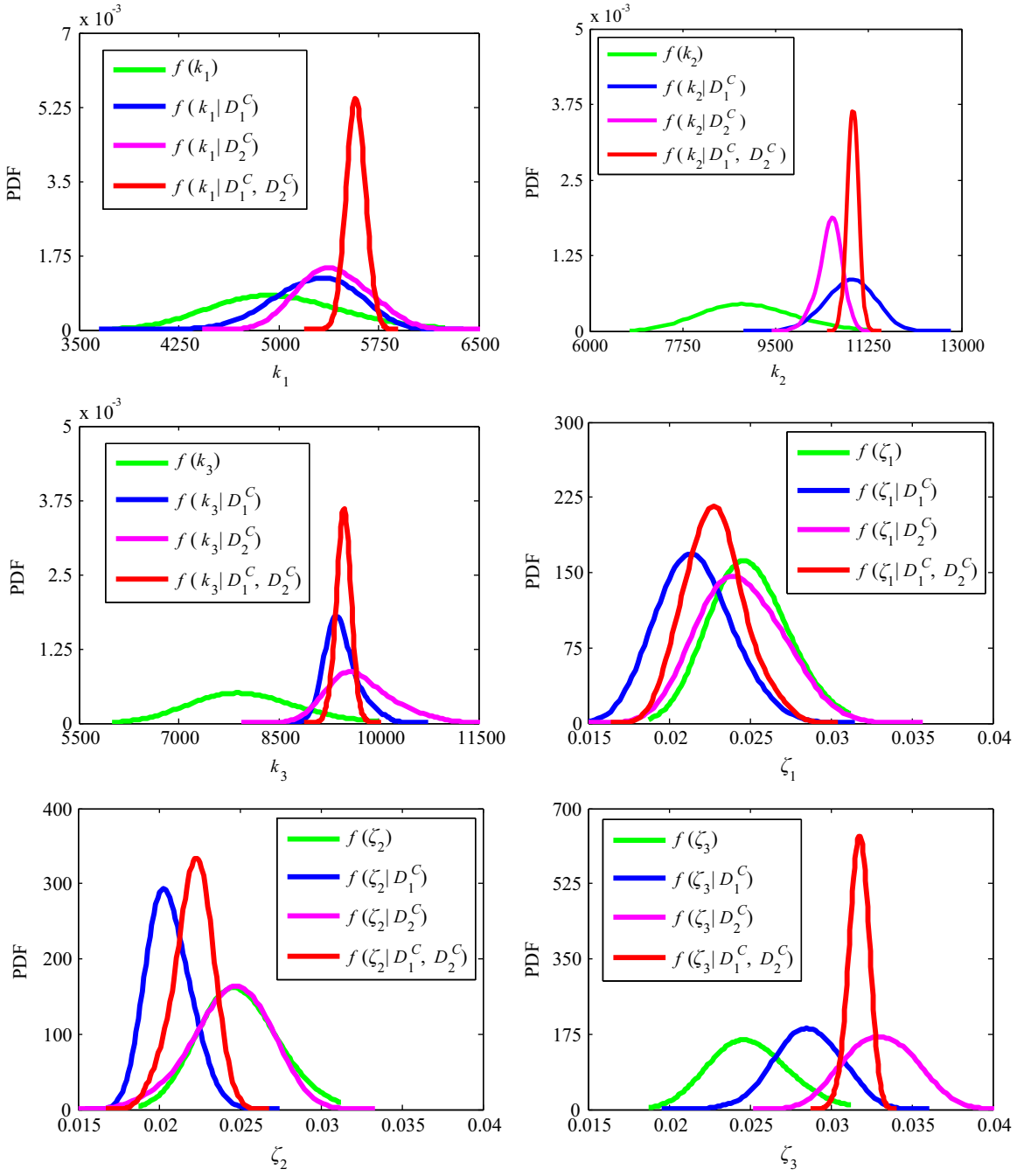

Fig. 3. Posterior distributions of model parameters.

quantity is chosen to be 15% of the validation data. Level 2 is expected to have lower model reliability value for two main factors:

1. The discretization error at Level 2 due to a limited number of finite elements for the beam (41 in this example). But this factor is not effective here since the data at Level 2 are synthetic data generated using the computational model, meaning that the difference between the computational model and the physics model is ignored. This factor will come into play if experimental data instead of synthetic data are used.
2. The coupling between the beam and the damped mass-spring system brings stronger nonlinearity at Level 2. Under the same number of training points, the GP surrogate model at Level 2 has more surrogate uncertainty (larger GP model prediction variance) than the GP surrogate model at Level 1. This factor is included in the numerical example.

Table 3
Model reliability values.

Validation test	1	2	3	4	5
Model reliability at Level 1	0.9702	0.9580	0.9398	0.9828	0.9800
Model reliability at Level 2	0.9616	0.8564	0.9208	0.9796	0.7904

The model reliability values given by the validation data from each validation test are listed in Table 3, which indicate lower model reliability at Level 2. In Fig. 4, these values are used to construct the distributions of model reliability at Level 1 and Level 2 using the method of moments.

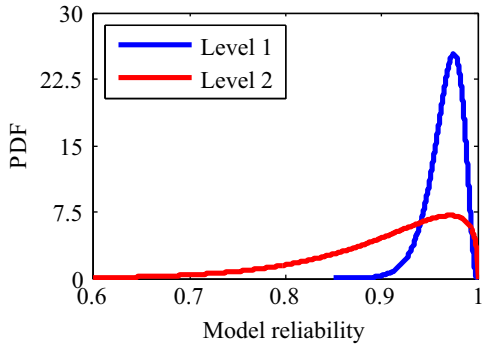

Fig. 4. Distribution of model reliability.

However, even though the model at Level 1 has higher model reliability than the model at Level 2, Level 2 is closer to the system level of interest since they have the same configuration. Therefore relevance analysis also needs to be considered.

The relevance index of each lower level to the system level is computed using the iterative algorithm in Section 4. The initial values of relevance indices for both lower levels are set as 1. The algorithm converges after three iterations for Level 1; and after five iterations for Level 2. The results are: $P(S_1) = 0.5785, P(S_2) = 0.8971$. This result means that Level 2 is more relevant to the system level, which is consistent with our intuition since Level 2 has the same structural configuration as the system and differs only in the load input (sinusoidal vs. random process). Compared with the result of model

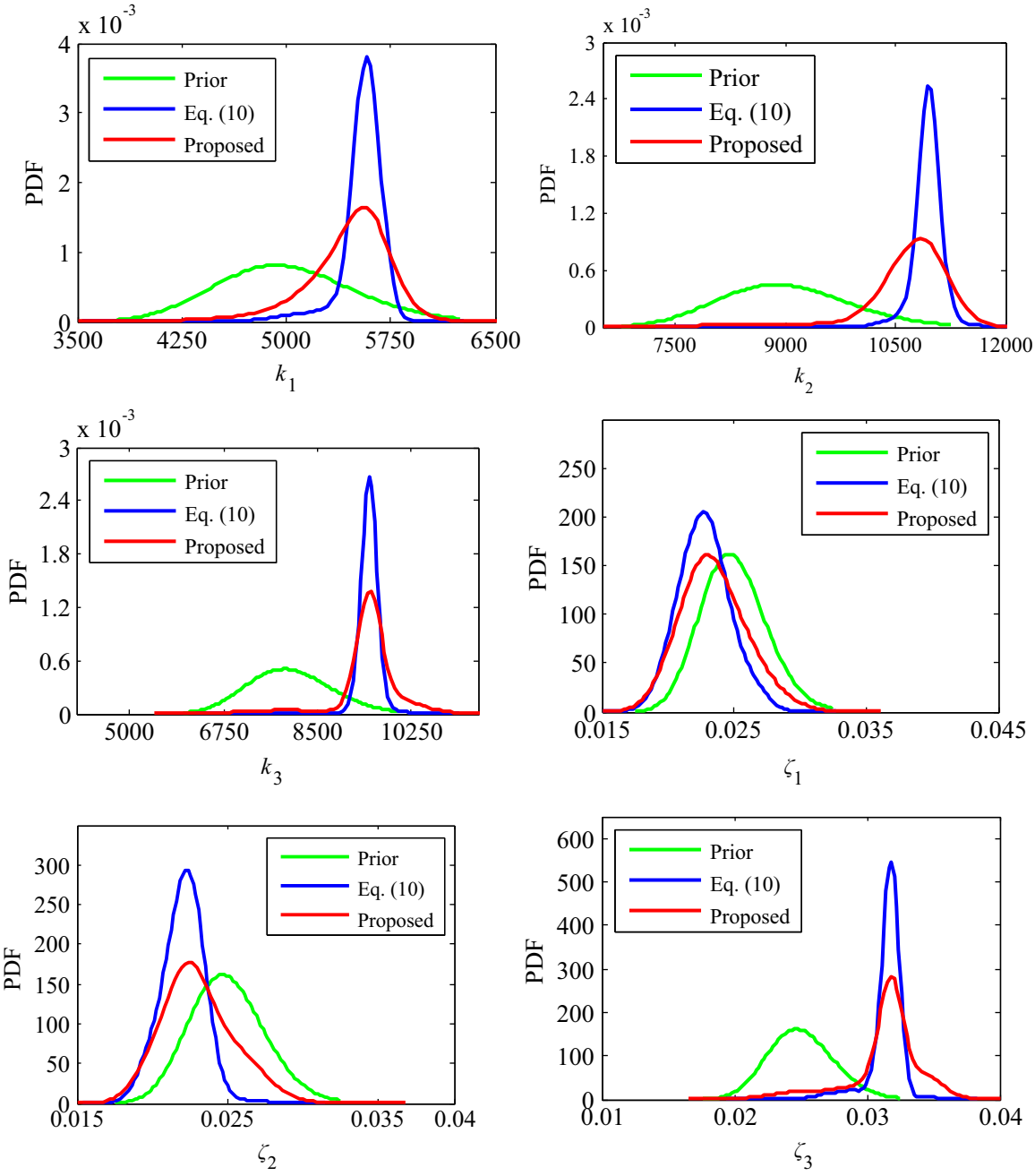

Fig. 5. Integrated distributions of model parameters.

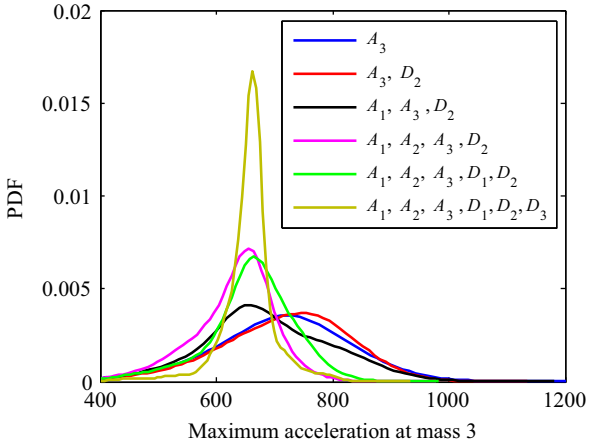

Fig. 6. System output prediction.

Table 4
Mean values and variances of predictions.

Number of quantities	1	2	3	4	5	6
Mean values	710	713	690	632	655	656
Variance	12,202	10,499	10,959	4868	5432	2301

validation, Level 2 has a lower value of model reliability but higher relevance index.

Based on all the information from calibration, validation and relevance analyses, the integrated distributions of all six model parameters are constructed in Fig. 5 using Eqs. (11) and (12). Fig. 5 also shows the result by considering validation only (no relevance) using the previous roll-method in Eq. (10) but extended for stochastic model reliability metric. It is shown that the proposed roll-up method is more conservative than the previous one, since we add one more criterion of relevance during the generation of samples from the posterior distribution.

The system output is predicted by propagating the integrated distribution of model parameters through the computational model at the system level. Fig. 6 gives not only the prediction using the data of all six quantities but also the prediction by other combinations of quantities whose names are shown in the legend. The mean values and variances of the predictions are shown in Table 4. As more quantities are employed, the mean value of prediction decreases from 712 to 656; and the variance shows an overall decreasing tendency, but not monotonic (the variance increases slightly when the number of outputs considered rises from 2 to 3, and from 4 to 5).

7. Summary

This paper developed a methodology to quantify the uncertainty in the system level output in a multi-level problem if experimental data are available only at lower levels and no data is available at the system level. The particular focus of this paper was to determine the appropriate distribution for model parameters $\boldsymbol{\theta}_m$ to be used in system level prediction, using calibration, validation, and sensitivity analyses at lower levels.

Note that the focus is not on improving the precision of calibration, but on including as much information as possible. The lower level models have different physical configurations and/or excitation compared to the system level prediction model (e.g., 3-mass-spring vs. 3-mass-spring-on-beam and sinusoidal inputs vs. random process inputs), and no calibration data is available corresponding to the system level configuration. Thus the proposed approach results in increasing the uncertainty of the posteriors because the lower-level models do not have 100% reliability, or 100% relevance to the system level.

The quantification of relevance is an important contribution to uncertainty integration. The relevance index quantifies the extent to which the lower level model reflects the physics captured in the system level model, and contributes to the weight of each posterior distribution in the uncertainty integration. In the proposed method, the relevance index is computed using the Sobol indices, and defined as the square of the cosine of the angle between two sensitivity vectors. As mentioned in Section 4, this approach does not provide a comprehensive comparison of the actual physics at different levels, but seeks to include the indication of physics given by variance-based sensitivity analysis, based on the prediction models at different levels.

For model validation, the proposed stochastic model reliability metric solves the problem of properly integrating results from multiple validation experiments. This paper also extends the model reliability metric to deal with multivariate data, i.e., measurements of multiple output quantities.

The third contribution of this paper is the development of the roll-up formula (Eqs. (11) and (12)) to integrate the information from three sources: 1) posterior distribution of model parameters by model calibration; 2) stochastic model reliability in model validation; 3) and relevance index of each lower level to the system level. The steps to realize this integration numerically are also developed.

In conclusion, model calibration obtains posterior distributions of each parameter within and across different lower levels; model validation evaluates the model reliability at each lower level separately; and the relevance analysis reveals the relationship between each lower level and the system level. All the above activities provide information to obtain the integrated distribution of model parameters. Using all this information, the system level output is predicted by propagating the integrated distributions of model parameters through the computational model at the system level.

Acknowledgments

The research in this paper is partially supported by funds from Sandia National Laboratories through Contract no. BG-7732 (Technical Monitor: Dr. Angel Urbina). This support is gratefully acknowledged. The authors also thank Dr. Josh Mullins, Dr. Shankar Sankararaman and Dr. You Ling for valuable discussions.

References

[1] Mullins J, Li C, Sankararaman S, Mahadevan S. Probabilistic integration of validation and calibration results for prediction level uncertainty quantification: application to structural dynamics. In: Proceedings of the 54th AIAA/ASME/ASCE/AHS/ ASC structures, structural dynamics, and materials conference; 2013.
[2] Kennedy MC, O'Hagan A. Bayesian calibration of computer models. J R Stat Soc 2001;63(3):425–64.
[3] Sankararaman S, Mahadevan S. Comprehensive framework for integration of calibration, verification and validation. In: Proceedings of the 53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference, 2012, no. April, p. 1–12.
[4] Trucano TG, Swiler LP, Igusa T, Oberkampf WL, Pilch M. Calibration, validation, and sensitivity analysis: What's what. Reliab Eng Syst Saf 2006;91(10–11):1331–57.
[5] Oberkampf WL, Barone MF. Measures of agreement between computation and experiment: validation metrics. J Comput Phys 2006;217(1):5–36.
[6] Oberkampf WL, Trucano TGG. Verification and validation in computational fluid dynamics. Prog Aerosp Sci 2002;38(3):209–72.
[7] Roache PJ. Fundamentals of verification and validation. Socorro, NM, US: Hermosa Press; 2009.

[8] Oberkampf WL, Roy CCJ. Verification and validation in scientific computing. Cambridge, UK: Cambridge University Press; 2010.
[9] Hills RG. Roll-up of validation results to a target application. Albuquerque, NM, US: Sandia Natl. Lab.; 2013 Rep. SAND2013-7424.
[10] O'Hagan A. Fractional Bayes factors for model comparison. J R Stat Soc 1995;57(1):99–138.
[11] Mullins J, Li C, Mahadevan S, Urbina A. Optimal selection of calibration and validation test samples under uncertainty. In: Proceedings of the IMAC XXXII; 2014. p. 391–01.
[12] Mahadevan S, Rebba R. Validation of reliability computational models using Bayes networks. Reliab Eng Syst Saf 2005;87(2):223–32.
[13] Ferson S, Oberkampf WL, Ginzburg L. Model validation and predictive capability for the thermal challenge problem. Comput Methods Appl Mech Eng 2008;197(29–32):2408–30.
[14] Ferson S, Oberkampf WL, Ginzburg L. Validation of imprecise probability models. Int J Reliab Saf 2009;3(1–3):3–22.
[15] Rebba R, Mahadevan S. Computational methods for model reliability assessment. Reliab Eng Syst Saf 2008;93(8):1197–207.
[16] Sankararaman S, Mahadevan S. Assessing the reliability of computational models under uncertainty. In: Proceedings of the 54th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference; 2013. p. 1–8.
[17] Ling Y. Uncertainty quantification in time-dependent reliability analysis. Nashville, TN, US: Vanderbilt University; 2013.
[18] Rajashekhar MR, Ellingwood BR. A new look at the response surface approach. Struct Saf 1993;12:205–20.
[19] Ghanem R, Spanos PD. Polynomial chaos in stochastic finite elements. J Appl Mech 1990;57(1):197–202.
[20] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge, MA, US: MIT Press; 2006.
[21] Xu P, Su X, Mahadevan S, Li C, Deng Y. A non-parametric method to determine basic probability assignment for classification problems. Appl Intell 2014;41(3):681–93.
[22] Arendt PD, Apley DW, Chen W. Quantification of model uncertainty: calibration, model discrepancy, and identifiability. J Mech Des 2012;134(10):100908.
[23] Liu F, Bayarri MJ, Berger JO. Modularization in Bayesian analysis, with emphasis on analysis of computer models. Bayesian Anal 2009;4(1):119–50.
[24] Jeffreys H. An invariant form for the prior probability in estimation problems. Proc R Soc Lond A Math Phys Sci 1946;186(1007):453–61.
[25] Tierney L. Markov chains for exploring posterior distributions. Ann Stat 1994;22(4):1701–28.
[26] Cha S. Comprehensive survey on distance/similarity measures between probability density functions. Int J Math Model Methods Appl Sci 2007;1(4).
[27] De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemom Intell Lab Syst 2000;50(1):1–18.
[28] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. Global sensitivity analysis: the primer. Chichester, UK: John Wiley & Sons; 2008.
[29] Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 2001;55(1–3):271–80.
[30] Li C, Mahadevan S. Global sensitivity analysis for system response prediction using auxiliary variable method. In: Proceedings of the 17th AIAA non-deterministic approaches conference; 2015.
[31] Li C, Mahadevan S. Relative contributions of aleatory and epistemic uncertainty sources in time series prediction. Int J Fatigue 2015.
[32] Singhal A. Modern information retrieval: a brief overview. IEEE Data Eng Bull 2001;24(4):35–43.
[33] Van Horn KS. Constructing a logic of plausible inference: a guide to Cox's theorem. Int J Approx Reason 2003;34(1):3–24.
[34] Li C, Mahadevan S. Sensitivity analysis for test resource allocation. In: Proceedings of the IMAC XXXIII; 2015.
[35] Rosenblatt M. Remarks on some nonparametric estimates of a density function. Ann Math Stat 1956;27(3):832–7.
[36] Red-Horse JR, Paez TL. Sandia National Laboratories Validation Workshop: Structural dynamics application. Comput Methods Appl Mech Eng 2008;197(29–32):2578–84.
[37] Chopra AK. Dynamics of structures: theory and applications to earthquake engineering. 4th ed. . Upper Saddle River, NJ, US: Prentice Hall; 2011.