Dan Ao
Department of Civil and Environmental Engineering,
Vanderbilt University,
279 Jacobs Hall,
VU Mailbox: PMB 351831,
Nashville, TN 37235
e-mail: dan.ao@vanderbilt.edu

Zhen Hu
Department of Civil and Environmental Engineering,
Vanderbilt University,
279 Jacobs Hall,
VU Mailbox: PMB 351831,
Nashville, TN 37235
e-mail: zhen.hu@vanderbilt.edu

Sankaran Mahadevan[1]
Professor
Department of Civil and Environmental Engineering,
Vanderbilt University,
272 Jacobs Hall,
VU Mailbox: PMB 351831,
Nashville, TN 37235
e-mail: sankaran.mahadevan@vanderbilt.edu

Dynamics Model Validation Using Time-Domain Metrics

Validation of dynamics model prediction is challenging due to the involvement of various sources of uncertainty and variations among validation experiments and over time. This paper investigates quantitative approaches for the validation of dynamics models using fully characterized experiments, in which both inputs and outputs of the models and experiments are measured and reported. Existing validation methods for dynamics models use feature-based metrics to give an overall measure of agreement over the entire time history, but do not capture the model's performance at specific time instants or durations; this is important for systems that operate in different regimes in different stages of the time history. Therefore, three new validation metrics are proposed by extending the model reliability metric (a distance-based probabilistic metric) to dynamics problems. The proposed three time-domain model reliability metrics consider instantaneous reliability, first-passage reliability, and accumulated reliability. These three reliability metrics that perform time-domain comparison overcome the limitations of current feature-based validation metrics and provide quantitative assessment regarding the agreement between the simulation model and experiment over time from three different perspectives. The selection of validation metrics from a decision-making point of view is also discussed. Two engineering examples, including a simply supported beam under stochastic loading and the Sandia National Laboratories structural dynamics challenge problem, are used to illustrate the proposed time-domain validation metrics.
[DOI: 10.1115/1.4036182]

Keywords: model validation, uncertainty, dynamics model, model reliability metric, time-series prediction

1 Introduction

Computational models are increasingly being used to emulate the physics of engineering systems and predict their behavior under untested conditions. However, the simulation model needs to be guaranteed to well represent the actual physical system before it can be applied in practice, and this need is addressed by model validation. Model validation is the process of determining the agreement to which a model is an accurate representation of the real world from the perspective of the intended use of the model [1,2]. Qualitative validation methods (i.e., graphical comparison) describe the agreement visually, while quantitative methods (using a validation metric) numerically characterize the degree of agreement [3]. The quantitative assessment of agreement is complicated by the presence of uncertainty sources that affect both prediction and observation. Therefore, significant research efforts have been pursued during the past two decades to develop effective model validation metrics in the presence of uncertainty. Commonly studied quantitative validation metrics include mean-based methods [4,5], hypothesis testing-based methods [3], area metric [6], and distance or reliability metric [2,7]. Considering the advantages and disadvantages of each method, analysts may select different validation metrics according to their specific problem, quantity of interest (QoI), and available information.

The aforementioned validation metrics have been intensively investigated for time-independent problems. These metrics, however, cannot be directly applied to the validation of time-dependent model predictions (i.e., response is a function of time), which are very common in practical engineering problems. Time-dependent responses are observed in cases with random process input (e.g., loading is a random process) [8–10] or in dynamics problems (e.g., acceleration and displacement under dynamic loading) [11]. Validation of time-dependent models is more challenging than that of time-independent models since the models need to be validated not only over the input space but also over the time domain. Validation of time-dependent models has not been studied until recent years. For instance, McFarland and Mahadevan [12] assessed a dynamics model by comparing the maximum absolute acceleration from prediction and observation. Sundermeyer et al. [13] performed model validation of time history using classical multivariate hypothesis testing based on the Fourier coefficients. Sprague–Geers (S&G) metric is commonly used to measure the similarity of the amplitude and phase between the test and simulation curves [14]; similar to S&G metric, Russell's error measure has been used to provide a measurement of the difference between time histories [15]; and dynamic time warping (DTW) has also been applied to quantify the discrepancy between two time histories [15]. Xi et al. [16] extended the U-pooling approach to dynamic response by using principal component analysis (PCA) [17] to transfer many correlated random variables into a few uncorrelated principal components. The Karhunen–Loève (KL) expansion [18] is also used to extract coefficients from time series and use the area metric for model validation based on the coefficients [19]. Jiang and Mahadevan [20] proposed a Bayesian wavelet method for model assessment by decomposing the time series and conducting Bayesian hypothesis testing based on the main energy components. Jiang and Mahadevan [21] also proposed a wavelet spectrum analysis approach to validate a dynamics model considering both time-domain and frequency-domain features, using the wavelet cross-spectrum and the wavelet time-frequency coherence, respectively.

These validation techniques for dynamics models may be referred to as feature-based model assessment techniques [20], i.e., they compute and compare several features of the time series prediction and observation. Feature-based methods have also been used in structural health monitoring (also involving comparison between prediction and observation) for damage detection. For example, statistical classification [22], autoregressive models [23,24], neural networks [25,26], and many other data-driven techniques are used to recognize the pattern of the time series in

[1]Corresponding author.

Manuscript received November 30, 2016; final manuscript received February 26, 2017; published online March 24, 2017. Assoc. Editor: David Moorcroft.

system identification problems. Modal properties (i.e., modal frequencies, modal damping ratio, and mode shapes) have often been applied in damage detection [27]. Other researchers have extracted features from the power spectral density (PSD) based on the Fourier transform, such as dominant frequencies or the first few central moments [28]. Some other transforms such as Fisher criterion minimization and divergence analysis can also be used to determine the appropriate features for specific problems [28–30]. Yang and Chang [31], and Jiang and Mahadevan [20,32] also developed wavelet-based method to extract signal features.

These feature-based techniques measure the agreement of predictions and observations to an overall degree. But they risk losing important information regarding the finer details of model performance in different regimes of application, especially if we are interested in the time varying agreement of the model prediction with experimental observation.

In order to overcome the drawback of feature-based comparison, Sankararaman and Mahadevan [26] proposed a Bayesian method to quantify and continuously update the uncertainty in damage detection. Basically they conducted Bayesian hypothesis testing on the mean value of the residuals, and continuously updated the distribution of the mean value and the likelihood ratio (between the null and alternative hypothesis) as more measurements are collected. Similar to feature-based model validation metrics, they give an overall evaluation on the agreement between model prediction and system measurement based on the accumulated evidence. The difference is that they directly operate in the time domain. Even though this approach can evaluate the time varying quality of the model to some degree, it cannot fully capture the performance of the time-dependent model.

This paper proposes new model validation metrics by directly working in the time domain. When both model output and experimental data involve randomness, the difference between the responses for the same input is also random. For given distributions of model output and experimental data, the difference can also be described as a distribution. In this situation, the model reliability metric proposed by Mahadevan and coworkers [2,7], which expresses the validation result as a simple probability, has several advantages. It can clearly separate the contribution of aleatory and epistemic uncertainty sources to the validation result [33,34], and therefore facilitate the allocation of resources to reduce the epistemic uncertainty. Further, the reliability metric has a clear probabilistic interpretation regarding the validation result; thus, it can be used in the aggregation of uncertainty quantification results from calibration, verification, and validation activities as developed in Refs. [35] and [36].

In this paper, the model reliability metric is extended to time-domain comparison in order to overcome the disadvantages of feature-based comparison. (Note that the model reliability metric can already accommodate feature-based comparison.) Three new time-domain model reliability metrics are proposed, namely, instantaneous reliability, first-passage reliability, and accumulated reliability. The instantaneous reliability metric evaluates the agreement between model and experiment at each time instant without considering the correlations between responses at different time instants. The first-passage reliability assesses the agreement between model and experiment over time duration of interest. The accumulated reliability tracks the accumulated evidence over time regarding model performance. Since the three validation metrics are defined from three different perspectives, they provide different types of information regarding the accuracy of the simulation model. The decision maker can choose the specific validation metric suitable to the particular application of interest. For example, if the decision maker is interested in the quality of the model at each time instant, the instantaneous reliability metric can be employed; in random vibration, shock and crash simulation, the first-passage reliability metric could be used if the decision maker wants to know that how the model quality changes over time; if the overall quality of the model is of interest, the accumulated reliability metric should be employed.

The remainder of the paper is organized as follows: Section 2 discusses the challenges of model validation in dynamics problems when uncertainty is considered, and reviews existing feature-based validation metrics. Section 3 develops three new model validation methods for time-domain comparison based on the model reliability metric, and discusses the advantages and disadvantages of each proposed metric. Section 4 presents two numerical examples and demonstrates the effectiveness of the proposed metrics. Section 5 provides concluding remarks.

2 Validation Metrics for Dynamics Models

2.1 Validation of Time-Independent Model Prediction. Concept and techniques of model validation have been intensively investigated with various representations of uncertainty in literature. Current model validation methods can be roughly grouped as hypothesis testing-based methods and nonhypothesis testing-based methods [2].

In hypothesis testing, we compare the evidence of two hypotheses; the null hypothesis (H_0) is that the prediction agrees with the observation, and the alternative hypothesis (H_1) is that it does not. The hypothesis testing can be based on classical or Bayesian statistics [2]. In classical hypothesis testing, a statistical test is used to determine whether there is enough evidence in a sample of data to reject the null hypothesis. For example, p-value in t-test is used to examine mean at a certain significance level, and chi-square is a test to check variance [37]. In Bayesian hypothesis testing, the likelihood ratio of the two hypotheses (known as Bayes factor) is used to assess data support for the two hypotheses [38]. Bayesian hypothesis testing has also been investigated for equality hypotheses [3], interval hypotheses [3], and for validation data from fully or partially characterized experiments [39]. Numerous nonhypothesis testing methods have also been developed. Commonly used metrics include the Mahalanobis distance [40], K–L divergence [41], area metric [6,42], and distance, or reliability metric [7]. The area metric [6,42] uses the area between the cumulative distributions (CDF) of model prediction and experimental data to quantify the disagreement between prediction and observation data. By transforming the model prediction and experimental data into standard uniform space, a U-pooling metric is further developed based on the area metric [6]. The reliability metric proposed by Rebba and Mahadevan [7] represents the model reliability in terms of the probability that the difference between model prediction and observation is within a predefined threshold.

Though various validation metrics have been developed in the literature, they can only be directly applied to time-independent models. For time-dependent models, where validation is needed over both input space and time domain, these metrics require further research.

2.2 Validation of Time-Dependent Model Prediction. In dynamics problems, the response is a function of both experiment inputs and time. At each time instant, uncertainty in the predicted response needs to be considered and the correlation between responses over time cannot be ignored. (In a validation experiment, when we measure input and output, and the model output corresponding to the measured input is compared to the measured output, the measurement error in the input which affects the model prediction is treated as aleatory uncertainty in the prediction.) The randomness over both input domain and time makes it difficult to directly apply traditional time-independent validation metrics to time-dependent problems. A common way to address this challenge is to convert the responses of time-dependent models into time-independent features and then perform validation in the feature space using traditional time-independent validation metrics.

Current techniques use features either in time domain or in frequency domain. The features in time domain are straightforward. For example, McFarland and Mahadevan [12] used the maximum absolute acceleration over the time series as the feature. Some

researchers recognized the pattern of time series using statistical classification [22], autoregressive models [23,24], neural networks [25,26], etc. Numerous types of transformation have been explored for feature extraction in the frequency domain. Principal components analysis (PCA) [17], Karhunen–Loève (KL) expansion [18], and wavelet packet decomposition [20] have been applied to compute features for model validation. Other transformation methods (e.g., Fourier transform, Fisher criterion minimization, and divergence analysis) [28,32], which are commonly used for damage detection in structural health monitoring, can also be extended to model validation. Three representative examples of feature-based comparison in the literature include PCA-based methods, the Bayesian wavelet method (extracting wavelet packet component energy), and wavelet spectrum analysis approach. These three methods are briefly reviewed.

Xi et al. [16] quantify the statistical disagreement between model prediction and experimental data by combining PCA with the area metric. After transferring the high-dimensional response into low-dimensional principal components, the joint cumulative distribution function (CDF) of coefficients of these components can be obtained for both model prediction and observation based on independence assumption. Then, the area metric can be easily implemented to assess the difference between the two CDFs. The Bayesian wavelet method proposed by Jiang and Mahadevan [20] assesses the time-dependent model by applying Bayesian hypothesis testing to the main energy components obtained from wavelet packet decomposition. In this method, main energy components are extracted for both model prediction and observation, and the discrepancies between these two sets of components are assumed to follow a multivariate normal distribution $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, leading to the comparison of Bayes factor as the validation metric (i.e., likelihood ratio of null and alternate hypotheses). Jiang and Mahadevan [21] also proposed a wavelet spectrum analysis approach to validate a dynamics model based on features in both time and frequency domains simultaneously. The wavelet transform (WT) decomposes a time series into a set of time-domain basis functions (i.e., wavelets) with various frequency resolutions, and the wavelets consist of a family of mathematical scaling functions used to represent a signal in both time and frequency domains. The wavelet cross-spectrum is then calculated to construct a time-frequency phase difference map, and wavelet time-frequency coherence is computed and assessed using statistical significance tests.

2.3 Analysis of Current Feature-Based Validation Metrics. The above feature-based validation methods are found to have four main disadvantages as summarized below.

(1) The feature-based validation metrics cannot explicitly represent the agreement between model and experiment over the time-domain since the features are obtained from the decomposition of the entire time series of the response. Thus, they only provide an overall assessment of the agreement between the model prediction and experiment data.
(2) Even if PCA, K–L expansion, Fourier transform or wavelet transform, etc., can decompose the time-variant response into a low dimensional set of time-independent responses for some problems, it cannot guarantee that the time-independent features are always low-dimensional. For responses with low correlation over time, the first several terms (e.g., principal components, dominant frequencies, or main energy components) may not be enough to fully capture the characteristics of the response over time. If a larger number of terms are used, model validation using validation metrics such as area metric, involves high-dimensional integration, which is computationally intensive.
(3) Even if the feature comparison is feasible for some cases, the way of dealing with the feature terms is not straightforward for the PCA, K–L expansion, Fourier transform or the wavelet transform-based methods. Take PCA as an example. Since model prediction and experimental observation belong to different stochastic processes, different principal components need to be used to decompose the responses, i.e., the two sets of principal components are in two different spaces. In that case, the validation needs to consider both the variation in the coefficients and the deviation of the principal components. If the same principal components are used for both simulation experimental output, the problem can be simplified. However, using the same principal components for both simulation and experimental output is not theoretically correct since the principal component for the simulation output may not be the optimal one for the experimental output, and vice versa. This problem also exists in other feature-based methods.
(4) Another issue is that conclusions about model validity based on different types of features (e.g., principal components, dominant frequencies, or main energy components) may differ from one another, and sometimes conflict with each other.

In order to overcome the above drawbacks, we propose three new validation metrics based on time-domain comparison, for the validation of dynamics models from three perspectives. In the proposed new validation metrics, the validation is performed directly in the time domain and has clear physical and probability interpretations.

3 Proposed Time-Domain Validation Metrics

In this section, we first briefly review the model reliability metric originally proposed for the validation of time-independent models. Based on that, we propose the new validation metrics for time-dependent problems.

3.1 Model Reliability Metric. The model reliability metric proposed by Rebba and Mahadevan [7] quantifies the probability of the difference (d) between model output and experimental data being within a desired tolerance interval as

$$r(\mathbf{x}) = \Pr\{|d| < \varepsilon\},\ d = Y_D - Y_M \tag{1}$$

where $r(\mathbf{x})$ is the reliability metric for a given input point $\mathbf{x}$ within the validation domain, Y_D indicates experimental data corresponding to $\mathbf{x}$, Y_M is model prediction at $\mathbf{x}$, ε is the tolerance representing the acceptable prediction error threshold from the decision maker's perspective, i.e., larger values of ε corresponds to a lower accuracy requirement [43,44]. Thus, the reliability metric is basically a probabilistic distance metric.

Subject domain knowledge is needed to choose the appropriate ε for the implementation of Eq. (1). To make the threshold more general, we can normalize Eq. (1) as

$$\begin{aligned} r(\mathbf{x}) &= \Pr\{|\tilde{d}| < \lambda\},\ \tilde{d} = (Y_D - Y_M)/Y_D \\ &= \Pr\{|d| < [\varepsilon = \lambda|Y_D|]\} \end{aligned} \tag{2}$$

where λ is a value between 0 and 1.

From the above discussion, we can see that the computation of $r(\mathbf{x})$ requires the probability distribution of absolute value of d, which can be obtained by comparing the model prediction with each replicated experimental measurement for each input setting. Based on this distribution, we can compute the reliability metric either based on a normal distribution assumption or directly using Monte Carlo simulation (MCS) to estimate the probability of the difference falling within a tolerance interval [7]. When validation experiments are available at multiple input settings, the joint distribution of the absolute value of d across all the input settings need to be considered.

3.2 Reliability Metric for Time-Dependent Model Prediction. Let $\mathbf{X}_{in}$ be the experiment input and t be time; the mean values of a subset of $\mathbf{X}_{in}$ are controllable during experiments and are

denoted as $\boldsymbol{\mu}_{\mathbf{X}}$, i.e., some inputs are controlled by the experimenter and some are not. We can perform simulations and experiments by changing the mean values of the controllable variables. Experiments can be performed with different mean values of the controllable inputs; we call them different *validation sites*. (We use mean values in this discussion, in order to acknowledge the uncertainty in actually realizing the specified value of the input.) The controllable input can be considered as a vector of random variables represented as $\mathbf{X}$. Based on the above definitions, we define the time-dependent responses of the experiment data (D) and simulation models (M) as $Y_D(\mathbf{X}, \mathbf{X}_a, t)$ and $Y_M(\mathbf{X}, \mathbf{X}_a, t)$, respectively. $\mathbf{X}_a$ is a vector of uncontrollable random variables during the experiments, which include both the uncontrollable experimental inputs and other random factors, such as observation errors.

Due to the uncertainty in the experimental environment and uncertainty in realizing specified experiment input $\mathbf{X}$, we may get different responses from replicated experiments even if the experiment input settings $\boldsymbol{\mu}_{\mathbf{X}}$ are the same. Figure 1 shows the realizations of three experiments with the same input settings $\boldsymbol{\mu}_{\mathbf{X}_1}$ and the corresponding model predictions with the same input settings by considering the uncertainty in the input $\mathbf{X}_1$. Since we are interested in analyzing the difference between the model and experiment, we also plot the differences between the model prediction and experimental observation for the three experiments in Fig. 1. The research issue that needs to be addressed now is how to quantitatively validate the model based on the data obtained similar to Fig. 1.

Aiming to validate $Y_M(\mathbf{X}, \mathbf{X}_a, t)$ using time-domain comparison, we propose three new reliability metrics, namely instantaneous reliability metric, first-passage reliability metric, and accumulated reliability metric. Details of the proposed three new reliability metrics are given in Secs. 3.2.1–3.2.3. In addition, phase lag differences between the two time series (prediction and observation) are very common. One way is to iteratively adjust one time series to maximize its cross-correlation or cross-coherence with the other time series, which is usually based on Fourier transform or wavelet transform [45,46]. In this paper, the developed method is applicable after eliminating the phase lag in the data. For example, a dynamic time warping technique has been studied for phase alignment [47]. Note that the focus of this paper, however, is not about how to eliminate the phase lag.

3.2.1 Instantaneous Reliability Metric. The most straightforward way of comparing time-dependent model and experimental outputs is to compute the model reliability at each instant. The validation therefore becomes a time-independent validation problem at each time instant. Assume that n_i replicate experiments are performed at nominal input settings $\boldsymbol{\mu}_{\mathbf{X}_i}$, $i = 1, 2, \cdots, N$, where N is the number of validation sites; and M simulations are performed by considering the uncertainty in the input $\mathbf{X}_i$.

For a given nominal input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$ and the jth experiment with this input setting, the reliability metric at time instant t_q is written as follows based on the definition given in Eq. (1):

$$\begin{aligned}&\left\{r(t_q)\left|\left[\boldsymbol{\mu}_{\mathbf{X}_i}, Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right)\right]\right.\right\}\\ &\quad = \Pr\left\{\left|d = \left[Y_M\left(\mathbf{X}_i\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right)\right]\right.\right.\\ &\qquad \left.\left.- \left[Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right)\right]\right| < \varepsilon(t_q)\right\} = \frac{1}{M}\sum_{k=1}^{M} I_{ijk}(t_q)\end{aligned} \tag{3}$$

where $r(t_q)\left|\left[\boldsymbol{\mu}_{\mathbf{X}_i}, Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right)\right]\right.$ is the reliability metric for given input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$ and observation data $Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right)$ collected from the jth experiment, $\varepsilon(t_q)$ is the accuracy requirement for the model at time instant t_q, $\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}$ is a vector of random variables $\mathbf{X}_i$ conditioned on input settings $\boldsymbol{\mu}_{\mathbf{X}_i}$, and $I_{ijk}(t_q)$ is given by

$$I_{ijk}(t_q) = \begin{cases} 1, \text{ if } |d_{ijk}(t_q)| < \varepsilon(t_q) \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

in which $d_{ijk}(t_q) = Y_M\left(\mathbf{x}_i^{(k)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(k)}, t_q\right) - Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right)$, where $Y_M\left(\mathbf{x}_i^{(k)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(k)}, t_q\right)$ is the kth realization of the prediction from the simulation model with input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$ by considering the uncertainty in the random variables, and $Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right)$

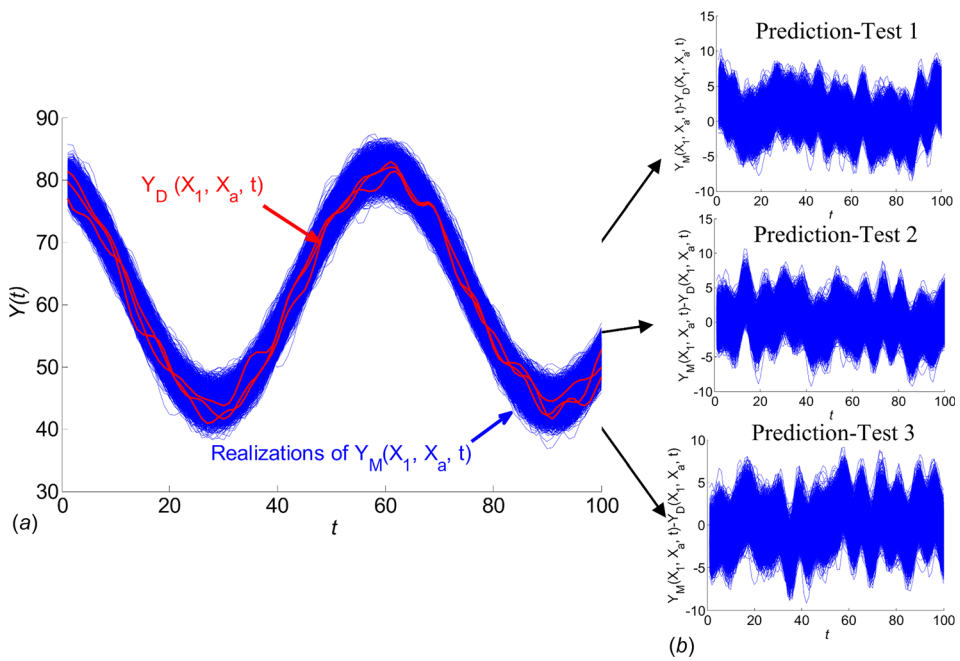

Fig. 1 Illustration of difference between experiment and model at one validation sites: (*a*) experimental observation and model prediction and (*b*) difference

is the response measurement from the jth experiment with input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$. Note that M is usually much larger than N since simulation models are often cheaper than experiments.

After considering the variation over replicated experiments, based on the total probability theorem, the reliability metric given in Eq. (3) can be rewritten as

$$\begin{aligned} r(t_q)\left|\boldsymbol{\mu}_{\mathbf{X}_i}\right. &= \sum_{j=1}^{n_i}\Pr\left\{\left|d = Y_M\left(\mathbf{X}_i\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right.\right)\right.\right. \\ &\quad \left.\left. - Y_D\left(\mathbf{x}_i^{(j)}\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right.\right)\right| < \varepsilon(t_q)\right\} \\ &\quad \times \Pr\left\{Y_D\left(\mathbf{x}_i^{(j)}\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right.\right)\right\} \\ &= \frac{1}{n_i}\sum_{j=1}^{n_i}\Pr\left\{\left|Y_M\left(\mathbf{X}_i\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right.\right)\right.\right. \\ &\quad \left.\left. - Y_D\left(\mathbf{x}_i^{(j)}\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right.\right)\right| < \varepsilon(t_q)\right\} \end{aligned} \tag{5}$$

After substituting Eq. (3) into Eq. (5), we have

$$r(t_q)\left|\boldsymbol{\mu}_{\mathbf{X}_i} = \frac{1}{n_i M}\sum_{j=1}^{n_i}\sum_{k=1}^{M} I_{ijk}(t_q)\right. \tag{6}$$

If there is no uncertainty in the experiment input settings, Eq. (3) will reduce to the original reliability metric proposed in Ref. [7].

When the validation is performed at multiple validation sites with input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$, $i = 1, 2, \cdots, N$, the validation metric given in Eq. (5) becomes

$$\begin{aligned} r(t_q) = \Pr\left\{\left|Y_M\left(\mathbf{X}_1\left|\boldsymbol{\mu}_{\mathbf{X}_1}, \mathbf{X}_a, t_q\right.\right) - Y_D\left(\mathbf{X}_1\left|\boldsymbol{\mu}_{\mathbf{X}_1}, \mathbf{X}_a, t_q\right.\right)\right| < \varepsilon(t_q) \cap \right. \\ \left. \cdots \cap \left|Y_M\left(\mathbf{X}_N\left|\boldsymbol{\mu}_{\mathbf{X}_N}, \mathbf{X}_a, t_q\right.\right) - Y_D\left(\mathbf{X}_N\left|\boldsymbol{\mu}_{\mathbf{X}_N}, \mathbf{X}_a, t_q\right.\right)\right| < \varepsilon(t_q)\right\} \end{aligned} \tag{7}$$

In the above equation, $Y_M\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right)$, $i = 1, 2, \cdots, N$ may be correlated random variables due to the shared random variables $\mathbf{X}_a$ in the simulation model. When we have a large number of experiments for each input setting, $Y_D\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right)$, $i = 1, 2, \cdots, N$ are also random variables modeled based on the experiment data. Based on the correlation analysis of $Y_M\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right)$, $i = 1, 2, \ldots, N$ using simulation models and random variable modeling of $Y_D\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right)$, $i = 1, 2, \cdots, N$, Eq. (7) can be estimated. When we have a limited number of experiments, Eq. (7) can be written in a discrete form based on the total probability theorem as

$$\begin{aligned} r(t_q) = \frac{1}{n_1 n_2 \cdots n_N}\sum_{j_1=1}^{n_1}\cdots\sum_{j_N}^{n_N}\Pr\left\{\left|Y_M\left(\mathbf{X}_1\left|\boldsymbol{\mu}_{\mathbf{X}_1}, \mathbf{X}_a, t_q\right.\right)\right.\right. \\ \left. - Y_D\left(\mathbf{x}_1^{(j_1)}\left|\boldsymbol{\mu}_{\mathbf{X}_1}, \mathbf{x}_a^{(j_1)}, t_q\right.\right)\right| < \varepsilon(t_q) \cap \\ \cdots \cap \left|Y_M\left(\mathbf{X}_N\left|\boldsymbol{\mu}_{\mathbf{X}_N}, \mathbf{X}_a, t_q\right.\right)\right. \\ \left.\left. - Y_D\left(\mathbf{x}_N^{(j_N)}\left|\boldsymbol{\mu}_{\mathbf{X}_N}, \mathbf{x}_a^{(j_N)}, t_q\right.\right)\right| < \varepsilon(t_q)\right\} \end{aligned} \tag{8}$$

where $Y_D\left(\mathbf{x}_i^{(j_i)}\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j_i)}, t_q\right.\right)$ is the j_i − th experiment under input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$.

If the correlation between $Y_M\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right)$, $i = 1, 2, \ldots, N$ is ignored, the time instantaneous reliability is given by

$$\begin{aligned} r(t_q) &= \frac{1}{n_1 n_2 \cdots n_N}\sum_{j_1=1}^{n_1}\cdots\sum_{j_N}^{n_N}\left[\prod_{i=1}^{N}\Pr\left\{\left|Y_M\left(\mathbf{X}_i\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, t_q\right.\right)\right.\right.\right. \\ &\quad \left.\left.\left. - Y_D^{(j_i)}(\boldsymbol{\mu}_{\mathbf{X}_i}, t_q)\right| < \varepsilon(t_q)\right\}\right] \\ &= \frac{1}{n_1 n_2 \cdots n_N}\sum_{j_1=1}^{n_1}\cdots\sum_{j_N}^{n_N}\left[\prod_{i=1}^{N}\frac{1}{M}\sum_{k=1}^{M} I_{ijk}(t_q)\right] \end{aligned} \tag{9}$$

If the instantaneous reliability metric is used to compute the model reliability versus time, the reliability curve fluctuates frequently with time. The instantaneous method provides the degree of agreement at every instant, and is easy to implement. However, the resulting noisy reliability curve actually gives very little useful information regarding model validity. The instantaneous reliability metric is useful if we are interested in the validity of the model at specific time instants.

3.2.2 First-Passage Reliability Metric. As shown in Fig. 2, first-passage failure is defined as the first time instant when the system response crosses a certain threshold [48]. In first-passage reliability problems, the system is defined to fail once the quantity of interest (i.e., $Y(t)$) crosses a threshold.

The first-passage reliability has been extensively studied in structural reliability research during the past several decades [48,49]. Computation of the reliability metric for model validation can employ the same methods that are well established in time-dependent reliability analysis. Thus, the concept of first-passage reliability can be extended to model validation to develop the first-passage reliability metric. The first-passage reliability metric for model validation can be defined as

$$r(t) = \Pr\left\{\left|d(\tau) = Y_D(\mathbf{X}, \mathbf{X}_a, \tau) - Y_M(\mathbf{X}, \mathbf{X}_a, \tau)\right| \le \varepsilon(\tau), \forall\tau \in [0, t]\right\} \tag{10}$$

The above equation gives the probability that the difference between the simulation model and experiment is less than a certain threshold over a duration $[0, t]$. For a given nominal input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$ and the jth experiment with this input setting, the first-passage reliability metric $r(t)\left|\boldsymbol{\mu}_{\mathbf{X}_i}, Y_D\left(\mathbf{x}_i^{(j)}\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, \tau\right.\right), \tau \in [0, t]\right.$ over a duration $[0, t]$ is given by

$$\begin{aligned} r(t)\left|\left[\boldsymbol{\mu}_{\mathbf{X}_i}, Y_D\left(\mathbf{x}_i^{(j)}\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, \tau\right.\right), \tau \in [0, t]\right.\right] \\ = \Pr\left\{\left|d(\tau) = \left[Y_M\left(\mathbf{X}_i\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{X}_a, \tau\right.\right)\right]\right.\right. \\ \left.\left. - \left[Y_D\left(\mathbf{x}_i^{(j)}\left|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, \tau\right.\right)\right]\right| \le \varepsilon(\tau), \forall\tau \in [0, t]\right\} \end{aligned} \tag{11}$$

where " ∀ " means "for all."

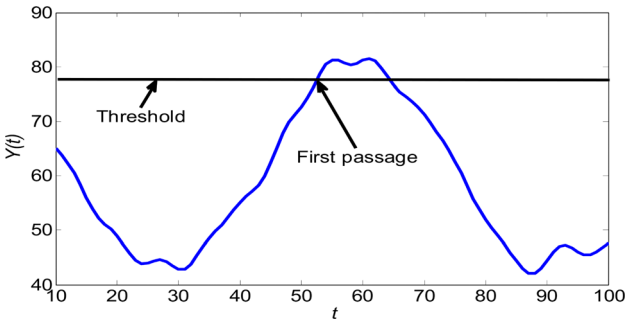

Fig. 2 Illustration of first-passage failure

After considering the uncertainty in the simulation model, based on a sampling-based approach, Eq. (11) is rewritten as

$$r(t)\left|\left[\boldsymbol{\mu}_{\mathbf{X}_i},\, Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\mathbf{x}_a^{(j)},\,\tau\right),\,\tau\in[0,\,t]\right]=\frac{1}{M}\sum_{k=1}^{M}I_{ijk}(t)\right. \quad (12)$$

where $I_{ijk}(t)$ is given by

$$I_{ijk}(t)=\begin{cases}1,\ \text{if}\ |d_{ijk}(\tau)|<\varepsilon(\tau),\ \forall\tau\in[0,\,t]\\ 0,\ \text{otherwise}\end{cases} \quad (13)$$

in which $d_{ijk}(\tau)=Y_M\left(\mathbf{x}_i^{(k)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{x}_a^{(k)},\,\tau\right)-Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{x}_a^{(j)},\,\tau\right)$ is the distance between the kth realization of the prediction from the simulation model with input setting $\mathbf{X}_i$ and the jth experiment with input setting $\mathbf{X}_i$ at time instant τ.

Since $|d_{ijk}(\tau)|-\varepsilon(\tau),\ \forall\tau\in[0,\,t]<\max_{\tau\in[0,\,t]}\left\{|d_{ijk}(\tau)|-\varepsilon(\tau)\right\}$, we have $|d_{ijk}(\tau)|<\varepsilon(\tau),\forall\tau\in[0,t]$ if $\max_{\tau\in[0,t]}\left\{|d_{ijk}(\tau)|-\varepsilon(\tau)\right\}<0$ and Eq. (13) reduces to the following equation:

$$I_{ijk}(t)=\begin{cases}1,\ \text{if}\ \max_{\tau\in[0,\,t]}\left\{|d_{ijk}(\tau)|-\varepsilon(\tau)\right\}<0\\ 0,\ \text{otherwise}\end{cases} \quad (14)$$

Equation (11) can be further written as

$$\begin{aligned}&r(t)\left|\left[\boldsymbol{\mu}_{\mathbf{X}_i},\, Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\mathbf{x}_a^{(j)},\,\tau\right),\,\tau\in[0,\,t]\right]\right.\\ &\quad=\Pr\left\{d_{\max}^{(j)}\left(\mathbf{X}_i\middle|\boldsymbol{\mu}_{\mathbf{X}_i},t\right)=\max_{\tau\in[0,\,t]}\left(\left|\left[Y_M\left(\mathbf{X}_i\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{X}_a,\,\tau\right)\right]\right.\right.\right.\\ &\qquad\left.\left.\left.-\left[Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{x}_a^{(j)},\tau\right)\right]\right|-\varepsilon(\tau)\right)\le 0\right\}\end{aligned} \quad (15)$$

where $d_{\max}^{(j)}(t)$ is a random variable affected by the variability in both model prediction and experimental observation over the time duration of interest.

Similar to the instantaneous reliability metric, Eq. (15) can be further written as follows by considering the variability in the experiments:

$$\begin{aligned}r(t)\left|\boldsymbol{\mu}_{\mathbf{X}_i}\right.=&\frac{1}{n_i}\sum_{j=1}^{n_i}\Pr\left\{d_{\max}^{(j)}\left(\mathbf{X}_i\middle|\boldsymbol{\mu}_{\mathbf{X}_i},t\right)=\max_{\tau\in[0,\,t]}\left(\left|\left[Y_M\left(\mathbf{X}_i\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\mathbf{X}_a,\tau\right)\right]\right.\right.\right.\\ &\left.\left.\left.-\left[Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\mathbf{x}_a^{(j)},\tau\right)\right]\right|-\varepsilon(\tau)\right)\le 0\right\}\end{aligned} \quad (16)$$

where $d_{\max}^{(j)}(t)$ is the extreme value distribution of the difference between the simulation model and the jth experiment.

Analytically estimating the extreme value distribution $d_{\max}^{(j)}\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i},\,t\right)$ is difficult. In this paper, a simulation-based method is employed. Since realizations of the simulation model and the data of experiment are collected at each time instant τ, Eq. (16) can be estimated based on the samples of the simulation model and experiment data. There are basically three steps as shown in Fig. 3 and explained in the text following the figure. The three steps are difference analysis, extreme value analysis, and first-passage reliability computation.

Assume that n_i experiments are performed at input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$, $i=1,\,2,\,\cdots,\,N$, where N is the number of validation sites; and M simulations are performed by considering the uncertainty in the input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$. In *step* 1 (*difference step in Fig. 3*), the difference between the model outputs and experiment outputs is computed at each time instant $\tau\in[0,\,t]$ as follows:

$$d_{ijk}(\tau)=Y_M\left(\mathbf{x}_i^{(k)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{x}_a^{(k)},\,\tau\right)-Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{x}_a^{(j)},\,\tau\right) \quad (17)$$

where $Y_M\left(\mathbf{x}_i^{(k)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{x}_a^{(k)},\,\tau\right)$ is the kth ($k=1,\,2,\,\cdots,\,M$) realization of the simulation model, $Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\,\mathbf{x}_a^{(j)},\,\tau\right)$ is the jth ($j=1,\,2,\,\cdots,\,n_i$) experiment at input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$, and τ is a time instant belonging to time interval $[0,\,t]$.

In **step 2** (*extreme value step in Fig. 3*), the kth sample of the extreme value distribution of difference corresponding to the jth ($j=1,\,2,\,\cdots,\,n_i$) experiment at input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$ is obtained based on the maximum absolute difference as below

$$d_{ijk}^{\max}=\max_{\tau\in[0,\,t]}\left\{\left|d_{ijk}(\tau)\right|-\varepsilon(\tau)\right\},\ k=1,\,2,\,\cdots,\,M \quad (18)$$

After the samples of the extreme values of the differences are obtained, the first-passage reliability metric is evaluated in **step 3** (*validation step in Fig. 3*) as

$$r(t)\left|\boldsymbol{\mu}_{\mathbf{X}_i}\right.=\frac{1}{n_iM}\sum_{j=1}^{n_i}\sum_{k=1}^{M}I_{ijk}^{\max}(t) \quad (19)$$

where $I_{ijk}^{\max}(t)=\begin{cases}1,\ \text{if}\ d_{ijk}^{\max}<0\\ 0,\ \text{otherwise}\end{cases}$.

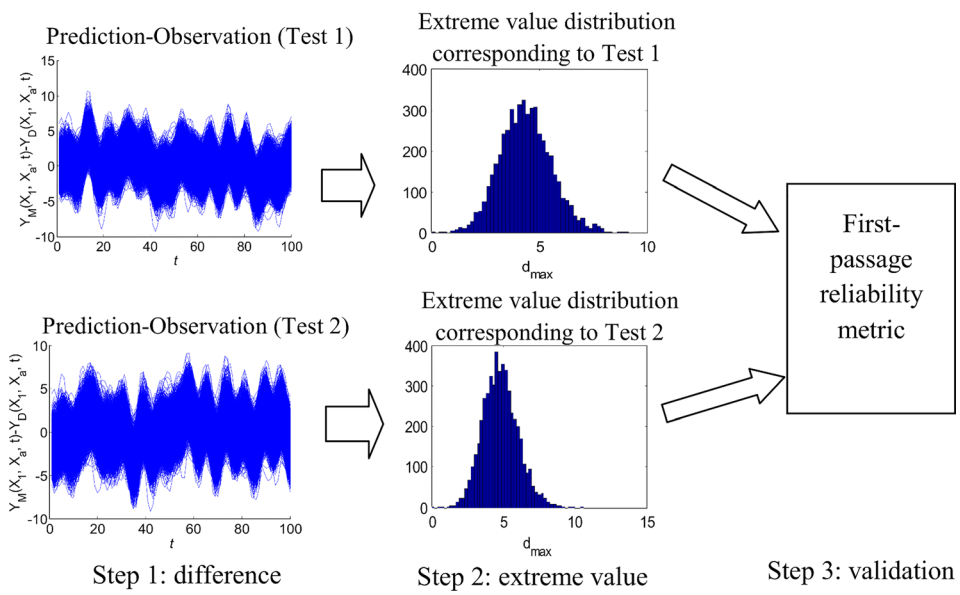

Fig. 3 Three steps of the first-passage reliability metric computation at one validation site

The above discussion is for only one validation site. When there are multiple validation sites (i.e., input settings), the first-passage reliability metric is given by

$$
\begin{aligned}
r(t) = & \frac{1}{n_1 n_2 \cdots n_N} \sum_{j_1=1}^{n_1} \cdots \sum_{j_N}^{n_N} \Pr\left\{ \max_{\tau \in [0,\, t]} \left(\left| Y_M\left(\mathbf{X}_1 \middle| \boldsymbol{\mu}_{\mathbf{X}_1},\, \mathbf{X}_a,\, \tau \right) \right.\right.\right. \\
& \left. - Y_D\left(\mathbf{x}_1^{(j)} \middle| \boldsymbol{\mu}_{\mathbf{X}_1},\, \mathbf{x}_a^{(j)},\, \tau \right) \right| - \varepsilon(\tau) \Big) \le 0 \cap \\
& \cdots \cap \max_{\tau \in [0,\, t]} \left(\left| Y_M\left(\mathbf{X}_N \middle| \boldsymbol{\mu}_{\mathbf{X}_N},\, \mathbf{X}_a,\, \tau \right) \right.\right. \\
& \left. \left. - Y_D\left(\mathbf{x}_N^{(j)} \middle| \boldsymbol{\mu}_{\mathbf{X}_N},\, \mathbf{x}_a^{(j)},\, \tau \right) \right| - \varepsilon(\tau) \right) \le 0 \Big\} \\
= & \frac{1}{n_1 n_2 \cdots n_N} \sum_{j_1=1}^{n_1} \cdots \sum_{j_N}^{n_N} \Pr\left\{ d_{\max}^{(j_1)}\left(\mathbf{X}_1 \middle| \boldsymbol{\mu}_{\mathbf{X}_1},\, t \right) \right. \\
& \left. \le 0 \cap \cdots \cap d_{\max}^{(j_N)}\left(\mathbf{X}_N \middle| \boldsymbol{\mu}_{\mathbf{X}_N},\, t \right) \le 0 \right\}
\end{aligned} \tag{20}
$$

The correlations between $d_{\max}^{(j_i)}\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i},\, t\right)$, $i = 1,\, 2,\, \cdots,\, N$ can be analyzed using either Monte Carlo simulation or analytical approximations of the response function using the first-order reliability method (FORM) [8] or second-order reliability method (SORM) [50]. If the correlation between $d_{\max}^{(j_i)}\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i},\, t\right)$, $i = 1,\, 2,\, \cdots,\, N$ is ignored, the first-passage reliability metric is computed by

$$
\begin{aligned}
r(t_q) &= \frac{1}{n_1 n_2 \cdots n_N} \sum_{j_1=1}^{n_1} \cdots \sum_{j_N}^{n_N} \left[\prod_{i=1}^{N} \Pr\left\{ d_{\max}^{(j_i)}\left(\mathbf{X}_i \middle| \boldsymbol{\mu}_{\mathbf{X}_i},\, t \right) < 0 \right\} \right] \\
&= \frac{1}{n_1 n_2 \cdots n_N} \sum_{j_1=1}^{n_1} \cdots \sum_{j_N}^{n_N} \left[\prod_{i=1}^{N} \frac{1}{M} \sum_{k=1}^{M} I_{ij_ik}^{\max}(t) \right]
\end{aligned} \tag{21}
$$

where $I_{ij_ik}^{\max}(t)$ is given in Eq. (19).

The above computation of the first-passage reliability metric is affordable if the physics simulation model is inexpensive to evaluate. When the simulation model is expensive, several simplified options reported in the structural reliability literature can be used to reduce the computational effort. Examples include surrogate modeling-based time-dependent reliability analysis [49], the upcrossing rate method [8], and importance sampling [9,51]. In the first-passage reliability metric, the model is recognized as valid over a duration t only when the absolute difference $|d|$ is less than the threshold ε at every time instant within this duration. From the model validation point of view, the first-passage reliability metric can thus be applied when the requirement of the agreement between model prediction and observation is stringent. This metric can be used to evaluate the performance of the model over different time intervals. The first-passage reliability metric, however, may reject a good model when the time interval is long since the longer the time interval, the lower the first passage model reliability.

3.2.3 Accumulated Reliability Metric. From the above definitions and discussions, it can be seen that the instantaneous reliability metric provides an evaluation at each time instant and the first-passage reliability metric gives the evaluation over a time interval. The instantaneous reliability metric always fluctuates, whereas the first-passage reliability metric is nonincreasing; in other words, the first-passage reliability decreases as the duration of the time interval increases (since there is increased probability of the difference exceeding the threshold). However, in the validation of time-dependent models, sometimes we desire to get increasing confidence in our validation result (i.e., the model reliability becomes stable) with time. In other words, we wish to conclude whether the simulation model is accurate or not as more and more experimental data are collected over time. Based on this motivation, we propose another metric called the accumulated reliability metric. In the accumulated reliability metric, the evidence about the validity of the model is accumulated over time and the validation result will converge to a value when enough data are collected.

The accumulated reliability metric is proposed by extending the reliability metric proposed by Rebba and Mahadevan in Ref. [7] to time-dependent models. The basic idea is to compute the accumulated probability that the difference between the model and experiment is within a specific accuracy requirement. In Ref. [7], a distribution is assumed for the difference between the model and experiment. Then the distribution parameters are updated when new experiment data are collected over time. In this paper, we directly use the samples from the simulation model and experiment to compute the accumulated reliability without assuming any distribution since the samples are available. Assume that n_i experiments are performed at input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$, $i = 1,\, 2,\, \cdots,\, N$; and M simulations are performed by considering the uncertainty in the input $\mathbf{X}_i$, and the time interval of interest $[0,\, t]$ is divided into n_t time instants. For given input setting $\boldsymbol{\mu}_{\mathbf{X}_i}$ and given experiment data $Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\, \mathbf{x}_a^{(j)},\, \tau\right)$, $\tau \in [0,\, t]$, the accumulated reliability metric is computed as

$$
r(t)\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\, Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\, \mathbf{x}_a^{(j)},\, \tau\right),\, \tau \in [0,\, t] = \Pr\left\{ d_a^{(j)}\left(\mathbf{X}_i \middle| \boldsymbol{\mu}_{\mathbf{X}_i} \right) \le 0 \right\} \tag{22}
$$

where $d_a^{(j)}\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}\right)$ is a random variable representing the accumulated samples of difference of the distance between simulation model and the jth experiment and the error threshold ($\varepsilon(t_q)$). The probability density function (PDF) of $d_a^{(j)}\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}\right)$ is given by

$$
f_{d_a}(d) = \frac{1}{n_t} \sum_{q=1}^{n_t} f_{d_q}(d) \tag{23}
$$

in which $f_{d_q}(d)$ is the PDF of random variable $d_q\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}\right) = \left| [Y_M(\mathbf{X}_i|\, \boldsymbol{\mu}_{\mathbf{X}_i},\, \mathbf{X}_a,\, t_q)] - \left[Y_D\left(\mathbf{x}_i^{(j)}|\boldsymbol{\mu}_{\mathbf{X}_i},\, \mathbf{x}_a^{(j)},\, t_q\right) \right] \right| - \varepsilon(t_q)$.

The above equation indicates that we assign equal weights to the difference d_q at each time instant q in the accumulated reliability metric. When the sampling-based method is employed to evaluate Eq. (22), the accumulated reliability metric $r(t)\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\, Y_D\left(\mathbf{x}_i^{(j)}|\boldsymbol{\mu}_{\mathbf{X}_i},\, \mathbf{x}_a^{(j)},\, \tau\right)$, $\tau \in [0,\, t]$ can be computed as

$$
\begin{aligned}
r(t)\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\, Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i},\, \mathbf{x}_a^{(j)},\, \tau\right),\, \tau \in [0,\, t] &= \frac{1}{n_t} \sum_{q=1}^{n_t} \frac{\sum_{k=1}^{M} I_{ijk}(t_q)}{M} \\
&= \frac{\sum_{q=1}^{n_t} \sum_{k=1}^{M} I_{ijk}(t_q)}{n_t M}
\end{aligned} \tag{24}
$$

where

$$
I_{ijk}(t_q) = \begin{cases} 1, \text{if} \left| Y_M\left(\mathbf{x}_i^{(k)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(k)}, t_q\right) - Y_D\left(\mathbf{x}_i^{(j)}\middle|\boldsymbol{\mu}_{\mathbf{X}_i}, \mathbf{x}_a^{(j)}, t_q\right) \right| - \varepsilon(t_q) < 0 \\ 0, \text{otherwise} \end{cases} \tag{25}
$$

After considering the variability in repeated experiments, the accumulated reliability metric is given by

$$r(t)|\boldsymbol{\mu}_{\mathbf{X}_i} = \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\sum_{q=1}^{n_t}\sum_{k=1}^{M} I_{ijk}(t_q)}{n_t M} \tag{26}$$

When multiple validation sites are considered, the accumulated reliability metric becomes

$$r(t) = \frac{1}{n_1 n_2 \cdots n_N}\sum_{j_1=1}^{n_1}\cdots\sum_{j_N}^{n_N} \Pr\left\{d_a^{(j_1)}\left(\mathbf{X}_1|\boldsymbol{\mu}_{\mathbf{X}_1}\right) \le 0 \cap \cdots \cap d_a^{(j_N)}\left(\mathbf{X}_N|\boldsymbol{\mu}_{\mathbf{X}_N}\right) \le 0\right\} \tag{27}$$

Analytically solving the above equation is complicated and sampling-based method is recommended to estimate the accumulated reliability. If the correlation between $d_a^{(j_i)}\left(\mathbf{X}_i|\boldsymbol{\mu}_{\mathbf{X}_i}\right)$, $i = 1, \cdots, N$ can be ignored, $r(t)$ can be estimated using sampling-based method as

$$r(t) = \frac{1}{n_1 n_2 \cdots n_N}\sum_{j_1=1}^{n_1}\cdots\sum_{j_N}^{n_N}\left[\prod_{i=1}^{N}\left(\frac{1}{n_t M}\sum_{q=1}^{n_t}\sum_{k=1}^{M} I_{ij_ik}(t_q)\right)\right] \tag{28}$$

As the duration of the time interval increases, and more samples of the difference are accumulated, the value of $r(t)$ is observed to converge in these specific numerical examples; however, in some other problems, it is possible that it may not converge. Note that the accumulated reliability metric provides an overall evaluation regarding the validity of the simulation model.

The above accumulated reliability metric computation uses Monte Carlo simulation (MCS), which is straightforward and easy to implement. The Bayesian tracking method proposed by Sankararaman and Mahadevan [26] is actually another way to update the model reliability with accumulating evidence. In Ref. [26], Bayesian hypothesis testing is conducted on the mean value of the residuals (i.e., difference between model prediction and observation). They assumed that the mean value follows a prior distribution, and then continuously updated the distribution of the mean value and Bayes factor as more measurements are collected. The Bayesian method can be used in time-dependent problems, and is especially useful when residuals are only available at some of the time instants (i.e., sparse data).

Until now, we have defined the proposed three reliability metrics for time-dependent models. Next, we will discuss the properties (advantages and disadvantages) of the three reliability metrics.

3.2.4 Comparison of the Three Time-Domain Reliability Metrics. From the above descriptions of the proposed reliability metrics, it can be found that no feature analysis (e.g., principal components analysis, Fourier transform, wavelet transform) needs to be performed in the process of model validation. The validation is directly performed over the time domain and the implementation procedure is straightforward. Each of the proposed reliability metrics, however, has its own advantages and disadvantages. The properties of these reliability metrics are summarized as follows.

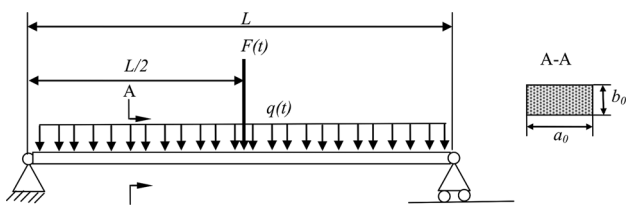

Fig. 4 Simply supported beam under stochastic loads

- *Instantaneous Reliability Metric*: The implementation of this reliability metric is the most straightforward. It directly gives a quantitative measurement on how the validity of the model changes over time. The disadvantage of this metric is that it is only an instantaneous measurement and has no implication on the quality of the model over the time interval of interest.
- *First-Passage Reliability Metric*: This reliability metric is extended from the concept of structural reliability analysis to the reliability of simulation models. The advantage of this metric is that it can provide an evaluation on the quality of the simulation model over different time intervals. When there is a decrease in the quality of the model, this metric can effectively reflect the change in the quality of the model. An important feature of the first-passage reliability metric is that the metric is nonincreasing. The longer the time interval, the lower is the reliability metric. The first-passage reliability metric requires that the difference between model prediction and experimental data be within the accuracy threshold at every instant in the time interval where the model reliability is evaluated. Therefore, the accuracy requirement is stringent.
- *Accumulated Reliability Metric*: This metric can give an overall evaluation on the quality of the model, and will converge to a single value as the duration of the time interval increases. The impact of model failure at a particular instant on the overall accumulated reliability is less significant compared to that of the first-passage reliability metric. Only large and sustained changes in the model quality will be reflected in this reliability metric since it is defined from an overall perspective. (For illustration, suppose the model has been accurate for a period of time and its accuracy suffers at some instant. This lapse in accuracy will be overshadowed by the accumulated evidence of past performance. The lapse in accuracy will need to be significant and sustained for a while to be reflected in the change of the metric.) This reliability metric is similar to feature-based model validation metrics, but operates in the time domain, without any transformations used in feature analysis. The disadvantage of the accumulated reliability metric can be made up by the advantage of the first-passage reliability metric.

In general, the time-dependent characteristics of model accuracy can be fully captured by the proposed three reliability metrics. The decision maker can choose the metric based on the desired application.

4 Numerical Examples

In this section, we use two numerical examples to illustrate the application of the proposed reliability metrics in the validation of time-dependent problems.

4.1 Beam Under Stochastic Time-Varying Loads

4.1.1 Problem Statement. A simply supported beam as shown in Fig. 4 is used as our first example [48]. The beam is subjected

Table 1 Uncertainty parameters of beam variables

Variable	Mean	Standard deviation	Distribution	Autocorrelation
a_0	0.2 m	0.005 m	Lognormal	N/A
b_0	0.04 m	0.0008 m	Lognormal	N/A
$F(t)$	1000 $(1+\sin(2+t))$ N	100 N	Gaussian	Eq. (30)
$q(t)$	450 N/m	20 N/m	Gaussian	Eq. (31)
L	5 m	0	Deterministic	N/A
ρ_{st}	78.5 kN/m^3	0	Deterministic	N/A
E	30 Gpa	0	Deterministic	N/A

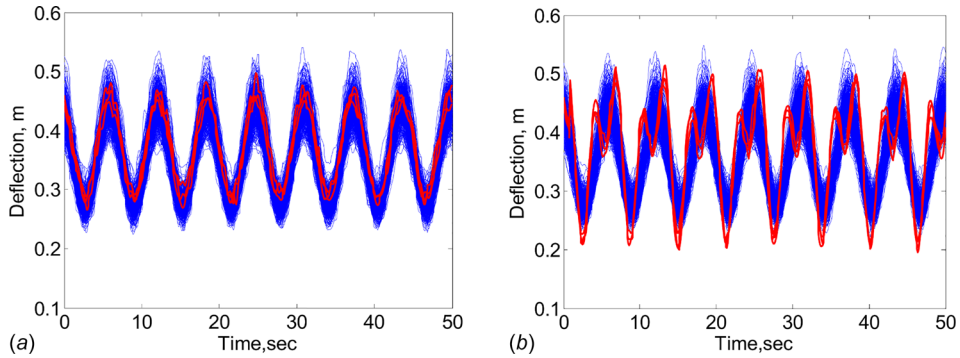

Fig. 5 Simulation and experimental output: (*a*) good model and (*b*) bad model

to two stochastic external loads, namely a concentrated load $F(t)$ at the midpoint and a uniformly distributed load $q(t)$. The beam is also subjected to gravity load due to its own weight. The system response of interest is the beam deflection at the midpoint. The performance function (deflection at the midpoint) is predicted by the mathematical model as

$$g(\mathbf{X}, t) = \frac{F(t)L^3}{48\text{EI}} + \frac{5\left(q(t) + \rho_{st} a_0 b_0\right)L^4}{384\text{EI}} \tag{29}$$

where $I = \left(a_0 b_0^3/12\right)$ and ρ_{st} is the density of the material. Here, a_0 and b_0 are the dimensions of the beam cross section.

The controllable experiment inputs are the mean values of the stochastic loads. Table 1 gives the random variables of the prediction model.

The autocorrelation functions of $F(t)$ and $q(t)$ are given by

$$\rho_F(t_1, t_2) = \exp\left(-(8(t_1 - t_2))^2\right) \tag{30}$$

and

$$\rho_q(t_1, t_2) = \exp\left(-(2(t_1 - t_2))^2\right) \tag{31}$$

In the above two correlation functions, the time unit is second. In general, the load $q(t)$ as well as the variables a_0, b_0, and E will also have variability along the length of the beam (random field). This spatial variability is ignored in this example.

To demonstrate the effectiveness of the proposed three reliability metrics, two cases, namely good model and bad model, are investigated. The prediction model is the same in both cases; the model for generating synthetic experimental data is different. In the good model case, we simply add a noise term to the prediction model to generate the experimental data; whereas in the bad model case, we add a noise term *plus* a bias term to the prediction model to generate the experimental data. The white noise term is given by $\varepsilon_{\text{obs}} \sim N\left(0, 0.1^2\right)$ and the bias term in the bad model is given by

$$\delta_g = 0.08B * \sin(2t) \tag{32}$$

where, $B* = 0$ if $t < 1$ second, otherwise $B* = 1$. It implies that the model becomes bad after 1 s.

4.1.2 Results and Discussion. Based on the above information, we generate synthetic experimental data for the good and bad models, respectively. In each case, three realizations of synthetic experimental data are generated and 1×10^5 realizations are generated in the prediction model. Figure 5 gives the time-dependent experimental output (bolded lines) and several realizations of the simulation output at one validation site (i.e., experiments performed with a set of values of the controllable inputs). It shows that the experimental and simulation outputs are very close for the good model and the difference between experiment and simulation is large for the bad model. We then perform quantitative validation in both cases in order to quantify the level of agreement between the model and experimental output.

Note that the three sets of experimental data are different for both good and bad models as shown in Fig. 5. This is because we add a bias to the experimental data in the bad mode case. Since an accuracy threshold needs to be specified in computing the reliability metric, the threshold employed in this example is given by

$$\varepsilon(t) = \lambda |Y_D(\mathbf{X}, t)| \tag{33}$$

where λ is a percentage value. In this paper, we report the results for $\lambda = 8\%$ and $\lambda = 16\%$ for both good and bad models. Note that the percentage values used here are just illustrative values. The decision makers can choose the appropriate threshold according to their requirement.

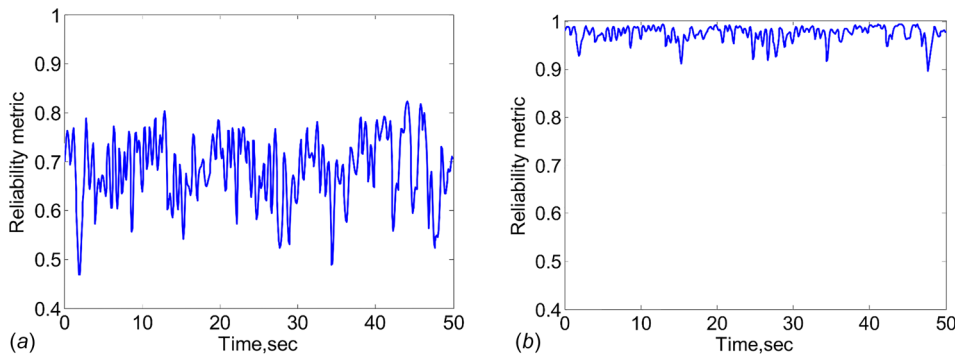

Fig. 6 Validation results using time-instantaneous reliability metric for good model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

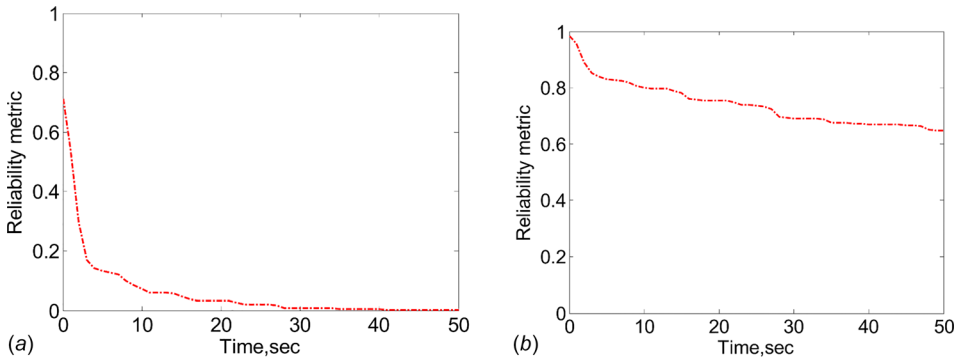

Fig. 7 Validation results using the first-passage reliability metric for good model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

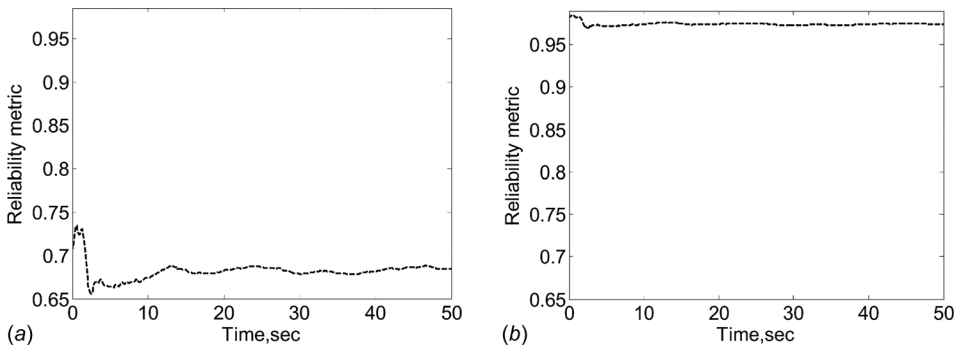

Fig. 8 Validation results using the accumulated reliability metric for good model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

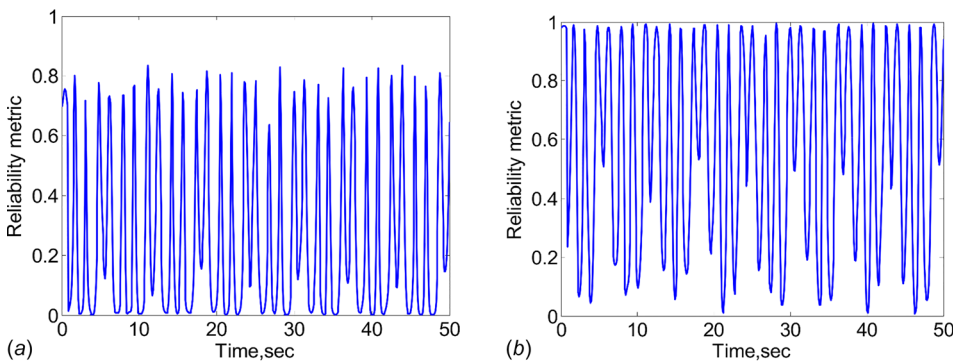

Fig. 9 Validation results using time-instantaneous reliability metric for bad model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

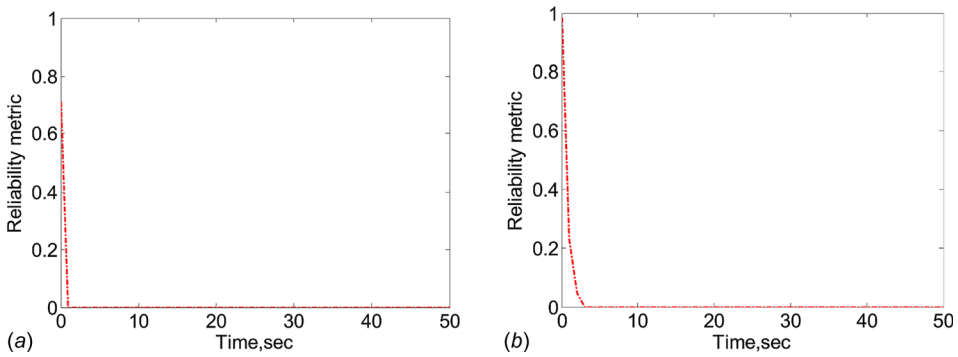

Fig. 10 Validation results using the first-passage reliability metric for bad model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

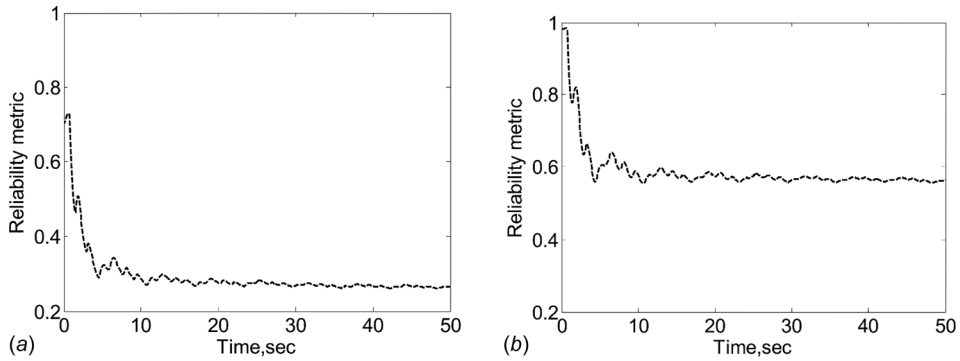

Fig. 11 Validation results using the accumulated reliability metric for bad model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

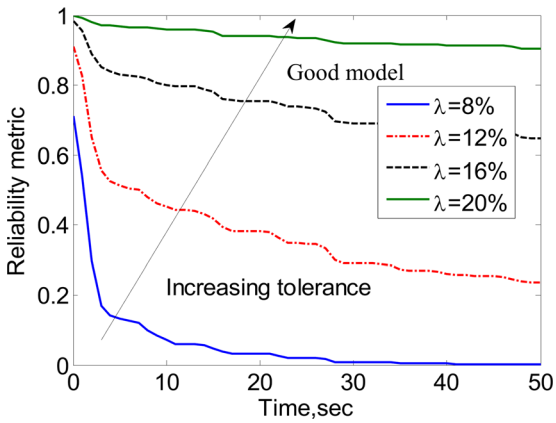

Fig. 12 Validation results using the first-passage reliability metric under different thresholds

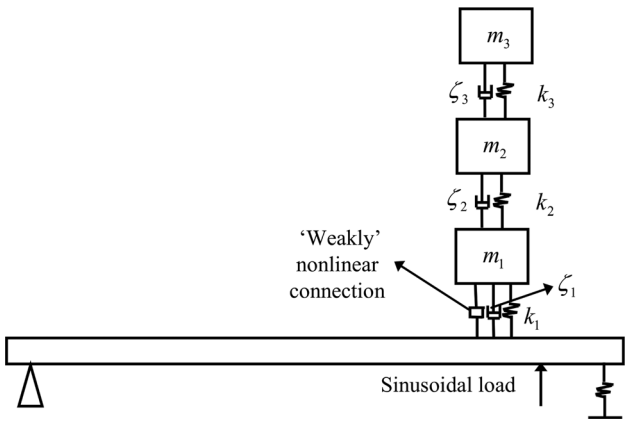

Fig. 14 Mass–spring–dampers on a beam

Figures 6–8 give the validation results of the good model using the instantaneous reliability metric, the first-passage reliability metric, and the accumulated reliability metric, respectively. From the validation results, it is seen that the instantaneous reliability metric fluctuates randomly over time, the first-passage reliability metric decreases over time, and the accumulate reliability receives a stable conclusion regarding the model validity as the time increases. (Note that the values in the vertical axes are different for $\lambda = 8\%$ and $\lambda = 16\%$.)

In addition to the above findings, we can also see that the first-passage reliability metric is sensitive to the threshold of the reliability metric. When the threshold is large (i.e., low requirement on the model accuracy), the first-passage reliability metric and the accumulated reliability metric get a similar conclusion regarding the validity of the simulation model. As shown in Figs. 7 and 8, both the first-passage reliability metric and the accumulated reliability metric tend to accept the model when $\lambda = 16\%$. Since the first-passage reliability metric is decreasing over time, we may reject a good model when the time interval is long since the longer the time interval, the higher the probability that the difference between the simulation model and experiment is larger than the accuracy threshold. This phenomenon, however, does not happen in the accumulated reliability metric.

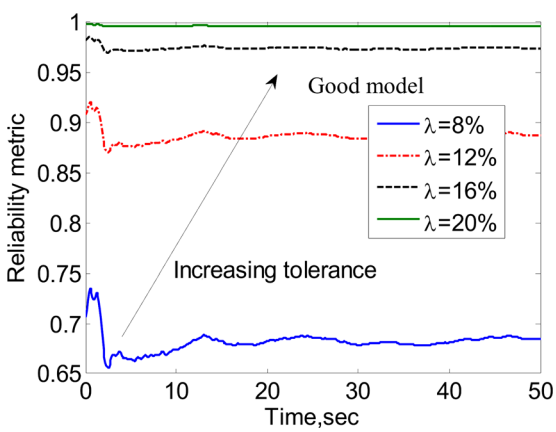

Fig. 13 Validation results using the accumulated reliability metric under different thresholds

We also performed model validation for the bad model case using the data shown in Fig. 5(*b*). Figures 9–11 give the validation results of the bad model obtained using the instantaneous reliability metric, the first-passage reliability metric, and the accumulated reliability metric, respectively. Similar conclusions can be obtained as those with the good model case.

The first-passage reliability metric and the accumulated reliability metric can effectively reject the bad model. The instantaneous reliability can hardly provide any useful information for the validation of time-dependent models. The accumulated reliability converges to a stable conclusion when more and more experimental data are collected.

Table 2 Random variables of the Sandia challenge problem

Variable	k_1	k_2	k_3	ζ_1	ζ_2	ζ_3
Mean	5000	9000	8000	0.025	0.025	0.025
Standard deviation	100	180	160	0.0005	0.0005	0.0005

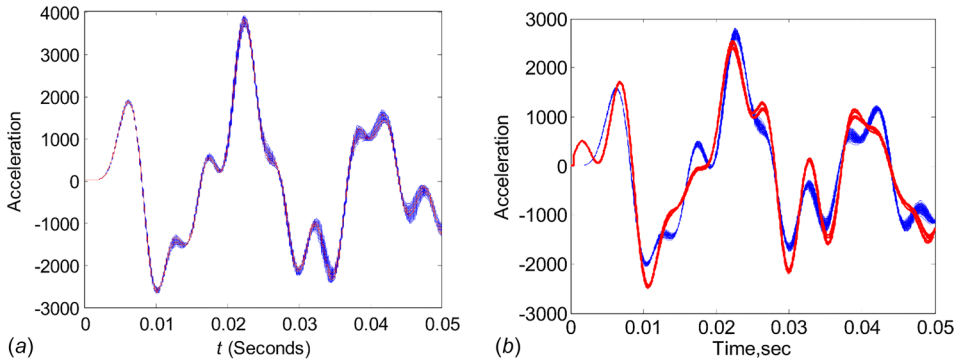

Fig. 15 Simulation and experiment data of the good model and bad model: (*a*) good model and (*b*) bad model

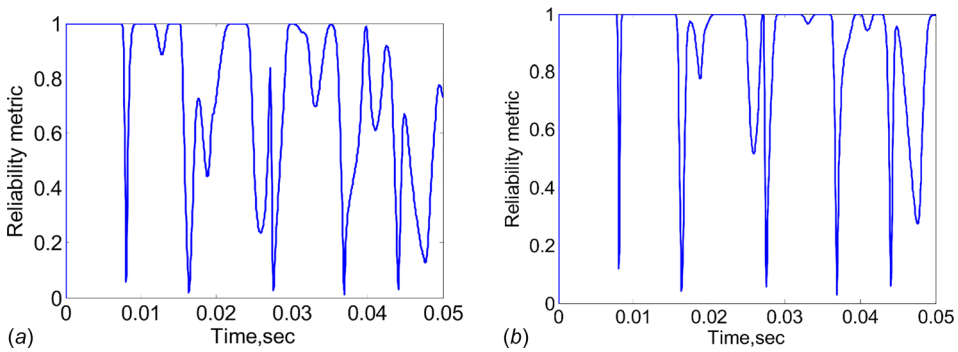

Fig. 16 Validation results using time-instantaneous reliability metric for good model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

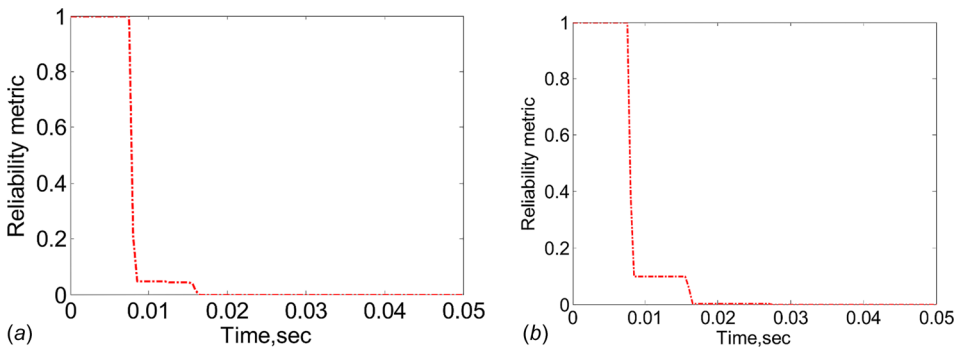

Fig. 17 Validation results using the first-passage reliability metric for good model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

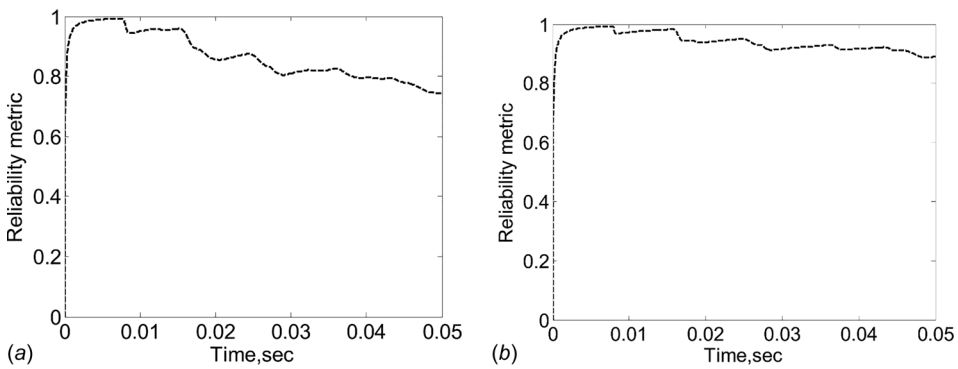

Fig. 18 Validation results using the accumulated reliability metric for good model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

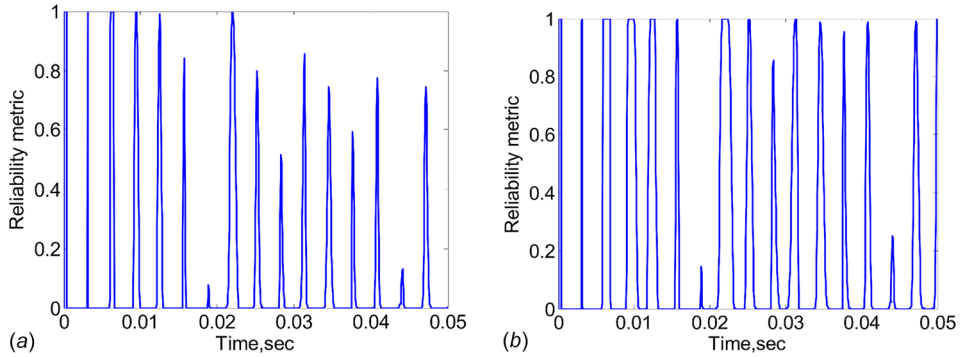

Fig. 19 Validation results using time-instantaneous reliability metric for bad model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

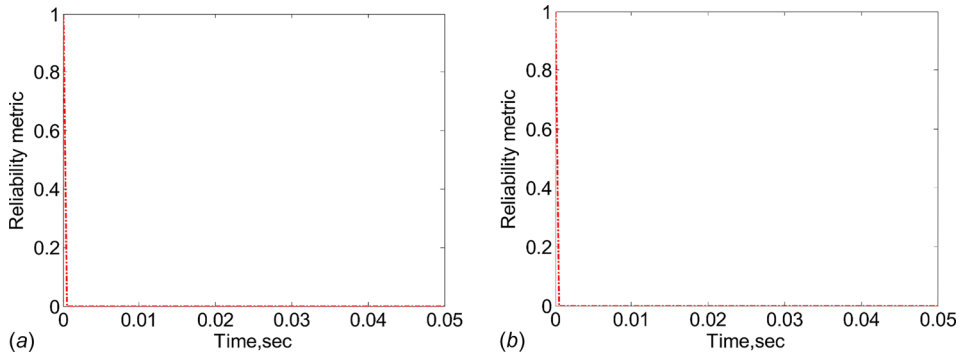

Fig. 20 Validation results using the first-passage reliability metric for bad model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

4.1.3 Parametric Study of the Threshold. Since the validation results of the three reliability metrics are dependent on the specified accuracy threshold, we also performed a parametric study for the threshold given in Eq. (2). In the good model case, validation is performed by gradually increasing the value of λ. Figures 12 and 13 show the validation results of the first-passage reliability metric and the accumulated reliability metric with different values of the threshold.

The results show that it is easier to accept the model as the tolerance (threshold) increases. As shown in Fig. 12, the first-passage metric rejects a good model when the threshold is low. This demonstrates that the first-passage reliability metric is more sensitive to the threshold than the accumulated reliability metric. Note that the parametric study is not performed for the instantaneous reliability metric because it has been shown in Sec. 4.1.2 that the instantaneous reliability metric can hardly give useful information.

4.2 Sandia Dynamics Challenge Problem

4.2.1 Problem Description. A structural dynamics problem developed at Sandia National Laboratories is used as the second example to illustrate the proposed three reliability metrics [35,52,53]. Figure 14 shows the configuration of the system. The structure consists of a beam with a substructure connected to it at a point via a weakly nonlinear connection. In the subsystem, three mass–spring–damper components are connected in series. A sinusoidal force input $P = 3000\sin(350t)$ is applied on the beam. Spring stiffness and damping ratio are assumed to follow lognormal distributions with mean values and standard deviations shown in Table 2.

The objective of this problem is to predict the acceleration of mass m_3, and use the experimental data to validate the model. Monte Carlo simulation (MCS) is used to evaluate the response (i.e., time series) for each input setting by considering randomness in spring stiffness and damping ratio. Multirealizations of stiffness and damping ratio result in a collection of time series for model prediction at each input setting. The proposed metrics evaluate the model reliability by considering the uncertainty in model response prediction.

Similar to the beam example in Sec. 4.1, five sets of experimental data are synthetically generated. For the "good model" case, the experimental data are generated using the prediction model itself by only adding noise. For the "bad model" case, the experimental data are generated by also adding a bias to the mass of m_1, and the following model is used to generate the experimental data:

$$g_b^i = g_g^i + 500\sin(100t) + \varepsilon \tag{34}$$

where $g_b^i(t)$ and $g_g^i(t)$ are the ith set (three sets in total) of experimental data in the case of bad model and good model, respectively; the second term on the right of the Eq. (34) is the bias, and it is added from the tenth step (i.e., 0.0005 s); ε is white noise and $\varepsilon \sim N(0, 1^2)$.

4.2.2 Results and Analysis. Figure 15 shows the predictions (blue curves) and experimental observations (red curves) of the acceleration at m_3 obtained from the good model and bad model, respectively. Based on the prediction and experimental data, validation of the dynamics model is performed using the proposed three reliability metrics. Note that the experimental data (red curves) are different for repeated experiments. This is due to the aleatory uncertainty sources in the experiments. Figures 16–18 give the validation results for the good model case by using two thresholds. It shows that the instantaneous reliability fluctuates significantly over time while the first-passage reliability metric decreases very quickly. The accumulated reliability metric gives the most stable evaluation of the model validity. It implies that the

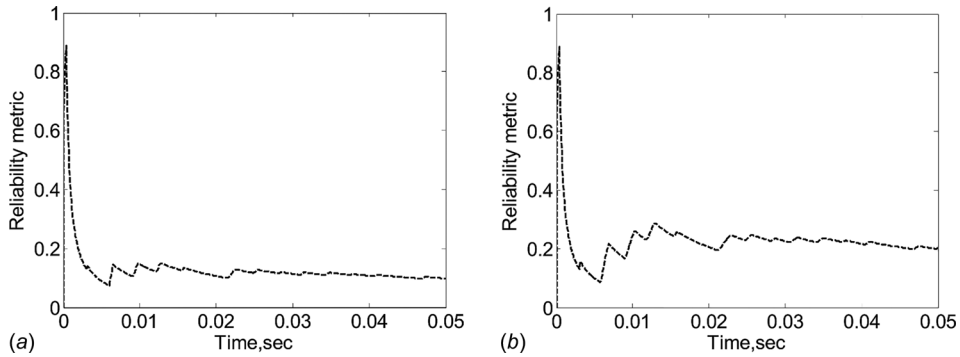

Fig. 21 Validation results using the accumulated reliability metric for bad model: (*a*) $\lambda = 8\%$ and (*b*) $\lambda = 16\%$

accumulated reliability metric is the most robust among the proposed three reliability metrics. Figures 19–21 give the validation results for the bad model case. Similar conclusions can be obtained as that of the good model case. In this particular problem, we conclude that the accumulated reliability metric is most useful for validation of time-dependent models because the instantaneous reliability is too trivial and the first-passage reliability drops too fast. If we have a very strict requirement on the accuracy of the model, the first-passage reliability metric is recommended. In Fig. 17, there are step changes in the reliability metric curve due to significant change in model quality.

5 Conclusion

In some validation applications, agreement between model prediction and observation data needs to be assessed both in the input space and over the time domain when we validate a time-dependent model. Feature-based validation comparisons for time-dependent models give overall measures and miss peculiarities over the time history. Three new time-domain validation metrics are proposed in this paper to overcome the limitations of feature-based validation metrics, in particular to avoid the loss of critical information.

The three validation metrics, namely instantaneous, first-passage, and accumulated reliability metrics, assess a time-dependent model from three different perspectives in the time domain. These three metrics overcome the disadvantages of feature-base validation metrics (as summarized in Sec. 2.3) by assessing the quality of the model over time instead of providing just one overall assessment. The first one shows model validity at each instant, while the other two focus on the agreement over a time interval. It is worth noting that the first-passage reliability metric and accumulated reliability metric reveal different types of agreement and different sensitivities to the validation threshold. Two examples demonstrate that the proposed validation metrics are capable of assessing the validity of a time-dependent model from different perspectives. The proposed reliability metrics are straightforward to implement, and successfully address the disadvantages of feature-based validation metrics. As indicated in the numerical examples, the proposed validation metrics are applicable to problems where only a limited number of experiments are possible. In addition, since there is no assumption made regarding the properties of the time history, the proposed metrics are also applicable to problems with nonstationary/ergodic responses, such as a shock environment. In this paper, the validation metrics are developed based on the assumption that the experiment is fully characterized. For dynamics problems with partially characterized experiments (i.e., when some of the inputs are not measured), the validation is more difficult. Extending the proposed reliability metrics to time-dependent models with partially characterized experiments may be studied in future.

Note that both feature-based validation metrics and the proposed time-domain metrics have their own benefits and drawbacks. For problems where it is difficult to create simulation time histories with the same duration as experimental time histories, feature-based validation metrics are more useful; and when the simulated time duration is the same as the experimental time duration, such that phase alignment is possible, the proposed time-domain metrics are useful. The choice of the metric and the acceptance criteria may depend on the usage requirement of the model, and the level of risk the decision maker is willing to take, which may be different in different situations. In the case where the model is required to be reliable over a short duration (e.g., shock), the first-passage reliability metric might be more useful, whereas in the case where the quantity of interest is response after a long duration (e.g., crack growth due to fatigue), the accumulated reliability metric might be more useful. The methodology developed in this paper helps risk-informed decision making by providing quantitative information about the model reliability.

Acknowledgment

This study was supported by funds from the National Science Foundation (Grant No. 1404823, CDSE Program). The support is gratefully acknowledged.

References

[1] AIAA, 1998, "Guide for the Verification and Validation of Computational Fluid Dynamics Simulations," American Institute of Aeronautics and Astronautics, Reston, VA, Standard No. AIAA G-077-1998(2002).
[2] Ling, Y., and Mahadevan, S., 2013, "Quantitative Model Validation Techniques: New Insights," Reliab. Eng. Syst. Saf., **111**, pp. 217–231.
[3] Rebba, R., Mahadevan, S., and Huang, S., 2006, "Validation and Error Estimation of Computational Models," Reliab. Eng. Syst. Saf., **91**(10), pp. 1390–1397.
[4] Kleijnen, J. P., 1995, "Verification and Validation of Simulation Models," Eur. J. Oper. Res., **82**(1), pp. 145–162.
[5] Drignei, D., Mourelatos, Z. P., Kokkolaras, M., and Pandey, V., 2014, "Reallocation of Testing Resources in Validating Optimal Designs Using Local Domains," Struct. Multidiscip. Optim., **50**(5), pp. 825–838.
[6] Ferson, S., Oberkampf, W. L., and Ginzburg, L., 2008, "Model Validation and Predictive Capability for the Thermal Challenge Problem," Comput. Methods Appl. Mech. Eng., **197**(29), pp. 2408–2430.
[7] Rebba, R., and Mahadevan, S., 2008, "Computational Methods for Model Reliability Assessment," Reliab. Eng. Syst. Saf., **93**(8), pp. 1197–1207.
[8] Hu, Z., and Du, X., 2012, "Reliability Analysis for Hydrokinetic Turbine Blades," Renewable Energy, **48**, pp. 251–262.
[9] Singh, A., Mourelatos, Z., and Nikolaidis, E., 2011, "Time-Dependent Reliability of Random Dynamic Systems Using Time-Series Modeling and Importance Sampling," SAE Int. J. Mater. Manuf., **4**(1), pp. 929–946.
[10] Zhang, J., and Du, X., 2011, "Time-Dependent Reliability Analysis for Function Generator Mechanisms," ASME J. Mech. Des., **133**(3), p. 031005.
[11] Wang, Z., Mourelatos, Z. P., Li, J., Singh, A., and Baseski, I., 2013, "Time-Dependent Reliability of Dynamic Systems Using Subset Simulation With Splitting Over a Series of Correlated Time Intervals," ASME Paper No. DETC2013-12257.
[12] McFarland, J., and Mahadevan, S., 2008, "Error and Variability Characterization in Structural Dynamics Modeling," Comput. Methods Appl. Mech. Eng., **197**(29), pp. 2621–2631.
[13] Sundermeyer, J., Betts, J. F., and Walker, M. A., 2016, "Simulation Time History Validation Via Multivariate Hypothesis Testing of Fourier Coefficients," ASME Paper No. VVS2016-8844.

[14] Marzougui, D., Samaha, R. R., Cui, C., Kan, C., and Opiela, K. S., 2012, "Extended Validation of the Finite Element Model for the 2010 Toyota Yaris Passenger Sedan," National Crash Analysis Center, George Washington University, Washington, DC, Report No. NCAC 2012-W-005.
[15] Sarin, H., Kokkolaras, M., Hulbert, G., Papalambros, P., Barbat, S., and Yang, R.-J., 2008, "A Comprehensive Metric for Comparing Time Histories in Validation of Simulation Models With Emphasis on Vehicle Safety Applications," ASME Paper No. DETC2008-49669.
[16] Xi, Z., Pan, H., Fu, Y., and Yang, R.-J., 2015, "Validation Metric for Dynamic System Responses Under Uncertainty," SAE Int. J. Mater. Manuf., **8**(2), pp. 309–314.
[17] Wold, S., Esbensen, K., and Geladi, P., 1987, "Principal Component Analysis," Chemom. Intell. Lab. Syst., **2**(1–3), pp. 37–52.
[18] Hu, Z., and Mahadevan, S., 2015, "Time-Dependent System Reliability Analysis Using Random Field Discretization," ASME J. Mech. Des., **137**(10), p. 101404.
[19] Wang, Z., Fu, Y., Yang, R.-J., Barbat, S., and Chen, W., 2016, "Validating Dynamic Engineering Models Under Uncertainty," ASME J. Mech. Des., **138**(11), p. 111402.
[20] Jiang, X., and Mahadevan, S., 2008, "Bayesian Wavelet Method for Multivariate Model Assessment of Dynamic Systems," J. Sound Vib., **312**(4), pp. 694–712.
[21] Jiang, X., and Mahadevan, S., 2011, "Wavelet Spectrum Analysis Approach to Model Validation of Dynamic Systems," Mech. Syst. Signal Process., **25**(2), pp. 575–590.
[22] DeSimio, M., Miller, I., Derriso, M., Brown, K., and Baker, M., 2003, "Structural Health Monitoring Experiments With a Canonical Element of an Aerospace Vehicle," IEEE Aerospace Conference, Big Sky, MT, Mar. 8–15, pp. 3105–3111.
[23] Sohn, H., Farrar, C. R., Hunter, N. F., and Worden, K., 2001, "Structural Health Monitoring Using Statistical Pattern Recognition Techniques," ASME J. Dyn. Syst. Meas. Control, **123**(4), pp. 706–711.
[24] Nichols, J., Nichols, C., Todd, M., Seaver, M., Trickey, S., and Virgin, L., 2004, "Use of Data-Driven Phase Space Models in Assessing the Strength of a Bolted Connection in a Composite Beam," Smart Mater. Struct., **13**(2), p. 241.
[25] Qian, Y., and Mita, A., 2008, "Acceleration-Based Damage Indicators for Building Structures Using Neural Network Emulators," Struct. Control Health Monit., **15**(6), pp. 901–920.
[26] Sankararaman, S., and Mahadevan, S., 2013, "Bayesian Methodology for Diagnosis Uncertainty Quantification and Health Monitoring," Struct. Control Health Monit., **20**(1), pp. 88–106.
[27] Tobe, R. J., 2010, "Structural Health Monitoring of a Thermal Protection System for Fastener Failure With a Validated Model," Ph.D. thesis, Wright State University, Dayton, OH.
[28] Guratzsch, R. F., 2007, "Sensor Placement Optimization Under Uncertainty for Structural Health Monitoring Systems of Hot Aerospace Structures," Ph.D. dissertation, Vanderbilt University, Nashville, TN.
[29] Therrien, C. W., 1989, *Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics*, Wiley, New York.
[30] Farrar, C. R., and Sohn, H., 2000, "Pattern Recognition for Structural Health Monitoring," Workshop on Mitigation of Earthquake Disaster by Advanced Technologies, Las Vegas, NV, Nov. 30–Dec. 1, Paper No. LA-UR-00-5565.
[31] Yang, J., and Chang, F.-K., 2006, "Detection of Bolt Loosening in C–C Composite Thermal Protection Panels—I: Diagnostic Principle," Smart Mater. Struct., **15**(2), p. 581.
[32] Jiang, X., Mahadevan, S., and Guratzsch, R., 2009, "Bayesian Wavelet Methodology for Damage Detection of Thermal Protection System Panels," AIAA J., **47**(4), pp. 942–952.
[33] Sankararaman, S., and Mahadevan, S., 2011, "Model Validation Under Epistemic Uncertainty," Reliab. Eng. Syst. Saf., **96**(9), pp. 1232–1241.
[34] Mullins, J., Ling, Y., Mahadevan, S., Sun, L., and Strachan, A., 2016, "Separation of Aleatory and Epistemic Uncertainty in Probabilistic Model Validation," Reliab. Eng. Syst. Saf., **147**, pp. 49–59.
[35] Sankararaman, S., and Mahadevan, S., 2015, "Integration of Model Verification, Validation, and Calibration for Uncertainty Quantification in Engineering Systems," Reliab. Eng. Syst. Saf., **138**, pp. 194–209.
[36] Mullins, J., and Mahadevan, S., 2016, "Bayesian Uncertainty Integration for Model Calibration, Validation, and Prediction," J. Verif. Valid. Uncertainty Quantif., **1**(1), p. 011006.
[37] Haldar, A., and Mahadevan, S., 2000, *Probability, Reliability, and Statistical Methods in Engineering Design*, Wiley, New York.
[38] Mahadevan, S., and Rebba, R., 2005, "Validation of Reliability Computational Models Using Bayes Networks," Reliab. Eng. Syst. Saf., **87**(2), pp. 223–232.
[39] Rebba, R., and Mahadevan, S., 2006, "Validation of Models With Multivariate Output," Reliab. Eng. Syst. Saf., **91**(8), pp. 861–871.
[40] De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L., 2000, "The Mahalanobis Distance," Chemom. Intell. Lab. Syst., **50**(1), pp. 1–18.
[41] Halder, A., and Bhattacharya, R., 2011, "Model Validation: A Probabilistic Formulation," 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC), Orlando, FL, Dec. 12–15, pp. 1692–1697.
[42] Ferson, S., and Oberkampf, W. L., 2009, "Validation of Imprecise Probability Models," Int. J. Reliab. Saf., **3**(1–3), pp. 3–22.
[43] Mullins, J., and Mahadevan, S., 2014, "Variable-Fidelity Model Selection for Stochastic Simulation," Reliab. Eng. Syst. Saf., **131**, pp. 40–52.
[44] Ling, Y., Mullins, J., and Mahadevan, S., 2014, "Selection of Model Discrepancy Priors in Bayesian Calibration," J. Comput. Phys., **276**, pp. 665–680.
[45] Bullock, T., McClune, M., Achimowicz, J., Iragui-Madoz, V., Duckrow, R., and Spencer, S., 1995, "EEG Coherence Has Structure in the Millimeter Domain: Subdural and Hippocampal Recordings From Epileptic Patients," Electroencephalogr. Clin. Neurophysiol., **95**(3), pp. 161–177.
[46] Grinsted, A., Moore, J. C., and Jevrejeva, S., 2004, "Application of the Cross Wavelet Transform and Wavelet Coherence to Geophysical Time Series," Nonlinear Processes Geophys., **11**(5/6), pp. 561–566.
[47] Aach, J., and Church, G. M., 2001, "Aligning Gene Expression Time Series With Time Warping Algorithms," Bioinformatics, **17**(6), pp. 495–508.
[48] Hu, Z., and Du, X., 2013, "Time-Dependent Reliability Analysis With Joint Upcrossing Rates," Struct. Multidiscip. Optim., **48**(5), pp. 893–907.
[49] Hu, Z., and Mahadevan, S., 2016, "A Single-Loop Kriging Surrogate Modeling for Time-Dependent Reliability Analysis," ASME J. Mech. Des., **138**(6), p. 061406.
[50] Du, X., and Chen, W., 2004, "Sequential Optimization and Reliability Assessment Method for Efficient Probabilistic Design," ASME J. Mech. Des., **126**(2), pp. 225–233.
[51] Dey, A., and Mahadevan, S., 1998, "Ductile Structural System Reliability Analysis Using Adaptive Importance Sampling," Struct. Saf., **20**(2), pp. 137–154.
[52] Red-Horse, J., and Paez, T., 2008, "Sandia National Laboratories Validation Workshop: Structural Dynamics Application," Comput. Methods Appl. Mech. Eng., **197**(29), pp. 2578–2584.
[53] Li, C., and Mahadevan, S., 2016, "Role of Calibration, Validation, and Relevance in Multi-Level Uncertainty Integration," Reliab. Eng. Syst. Saf., **148**, pp. 32–43.