Reliability Engineering and System Safety 111 (2013) 217–231

Contents lists available at SciVerse ScienceDirect

Reliability Engineering and System Safety

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

Quantitative model validation techniques: New insights

You Ling, Sankaran Mahadevan *

Department of Civil and Environmental Engineering, Vanderbilt University, TN 37235, United States

ARTICLE INFO

Article history:
Received 27 May 2011
Received in revised form
13 November 2012
Accepted 14 November 2012
Available online 28 November 2012

Keywords:
Model validation
Hypothesis testing
Bayesian statistics
Reliability
MEMS

ABSTRACT

This paper develops new insights into quantitative methods for the validation of computational model prediction. Four types of methods are investigated, namely classical and Bayesian hypothesis testing, a reliability-based method, and an area metric-based method. Traditional Bayesian hypothesis testing is extended based on interval hypotheses on distribution parameters and equality hypotheses on probability distributions, in order to validate models with deterministic/stochastic output for given inputs. Formulations and implementation details are outlined for both equality and interval hypotheses. Two types of validation experiments are considered—fully characterized (all the model/experimental inputs are measured and reported as point values) and partially characterized (some of the model/experimental inputs are not measured or are reported as intervals). Bayesian hypothesis testing can minimize the risk in model selection by properly choosing the model acceptance threshold, and its results can be used in model averaging to avoid Type I /II errors. It is shown that Bayesian interval hypothesis testing, the reliability-based method, and the area metric-based method can account for the existence of directional bias, where the mean predictions of a numerical model may be consistently below or above the corresponding experimental observations. It is also found that under some specific conditions, the Bayes factor metric in Bayesian equality hypothesis testing and the reliability-based metric can both be mathematically related to the p-value metric in classical hypothesis testing. Numerical studies are conducted to apply the above validation methods to gas damping prediction for radio frequency (RF) micro-electro-mechanical-system (MEMS) switches. The model of interest is a general polynomial chaos (gPC) surrogate model constructed based on expensive runs of a physics-based simulation model, and validation data are collected from fully characterized experiments.

1. Introduction

Model validation is defined as the process of determining the degree 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 such as graphical comparison between model predictions and experimental data are widely used in engineering. However, statistics-based quantitative methods are needed to supplement subjective judgments and to systematically account for errors and uncertainty in both model prediction and experimental observation [3].

Previous research efforts include the application of statistical hypothesis testing methods in the context of model validation [4–9], and development of validation metrics as measures of agreement between model prediction and experimental observation [9–13]. Discussions on the pros and cons of these validation methods can be found in [9,14]. Based on these existing methods and related studies, this paper is motivated by several issues that remain unclear in the practice of model validation: (1) validation with fully characterized vs. partially characterized experimental data; (2) validation of deterministic vs. stochastic model predictions; (3) accounting for the existence of directional bias; and (4) choice of thresholds in different validation metrics.

The four issues cannot be discussed without an introduction to the terminology concerning model and validation experiments. When a model is developed, the physical quantity of interest Y is postulated to be a function of a set of variables $\{\boldsymbol{x},\boldsymbol{\theta}\}$, i.e., $Y=f(\boldsymbol{x},\boldsymbol{\theta})$. This function is not exactly known and hence is approximated using a model with output Y_m, i.e., $Y_m=g(\boldsymbol{x},\boldsymbol{\theta})$. Y is observable through some experiments and $\boldsymbol{x}$ is the set of input variables to the experiments. The term "input" is referred to as the variables in a model that can be measured in experiments. We assume that the same set of variables goes into the model and validation experiments as inputs (i.e., the terms "model input" and "experimental input" represent the same set of variables), and we are comparing the outputs of the model and experiments during validation. Although there is an infinite set of variables in nature that can affect validation experiments, "model input" and "experimental input" only contain a subset of these variables due to the assumptions and approximations that are made in

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

0951-8320/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ress.2012.11.011

modeling. Model parameter set ($\boldsymbol{\theta}$) contains variables that may be measurable in concept but are usually difficult or impractical to directly measure in validation experiments. Therefore, the values/distributions of these variables are usually obtained from model calibration (performed prior to validation with a different set of input–output experimental data), and assumed invariant during multiple replicates of the validation experiment. For example, the maximum deflection (w) of a cantilever beam subjected to a point load at the free end is the quantity of interest, and it is known to be a function of the load value P, beam geometry, and Young's modulus (E). The solution $w = PL^3/3EI$ based on Euler–Bernoulli beam theory is a model that approximates this unknown function. The model input set $\boldsymbol{x}$ includes the load P, the moment of inertia I, and beam length L; the model parameter set $\boldsymbol{\theta}$ contains only one variable E.

The division of variables into two sets (model input $\boldsymbol{x}$ and model parameter $\boldsymbol{\theta}$), which is usually decided by modelers, helps classify the sources of uncertainty involved in validation. If a variable is measured and reported as point value, the uncertainty in this value is due to measurement error; if a variable is not measured but a range/distribution is assigned, the uncertainty is due to imprecise data; if a variable is inferred from model calibration, the sources of uncertainty include natural variability (the variability between different experiments), data uncertainty (uncertainty due to measurement error and insufficient data) and model uncertainty. It should be noted that the diagnostic quality and the bias in experiments are not considered as "input". Instead, they are classified as components of the measurement error, which is represented by ε_D in this paper.

With the terminology introduced above, we can now continue with the discussion of the four issues. First, there are two possible types of validation data, resulting from (1) fully characterized experiments (i.e., all the inputs of the model/experiment are measured and reported as point values), or (2) partially characterized experiments (i.e., some inputs of the model/experiment are not measured or are reported as intervals). For instance, some input variables of the model/experiment may not be measured, but we may have expert opinion about the possible ranges or probability distributions of these input variables, and thus this experiment is "partially" characterized. In other words, there will be more uncertainty in the data from partially characterized experiments than from fully characterized experiments, due to the uncertainty in the input variables. Some partially characterized experiments with limited uncertainty may be considered for validation by practitioners. While most of the previous studies only focus on validation with fully characterized experimental data, this paper explores the use of both types of data in various validation methods.

Second, due to the existence of aleatory and epistemic uncertainty, both the model prediction (denoted as Y_m) and the physical quantity to be predicted (denoted as Y) can be uncertain, and this has been the dominant case studied in the literature [6–9,12,13,15]. However, in practice it is possible that either Y_m or Y can be considered as deterministic. Deterministic Y_m implies that for given values of the model input variables, the output prediction of the model is deterministic. The application of various validation methods to these different cases is examined in this paper.

Third, in this study, we define two terms to characterize the difference between model prediction and validation data—bias and directional bias. Bias is defined as simply the difference between the mean value of model prediction and the statistical mean value of experiment data, and the term "directional bias" implies the persistence of bias in one direction as one varies the inputs of model and experiment. This paper explores whether various validation metrics are able to account for the existence of directional bias.

Fourth, although different validation metrics are developed to measure the agreement between model prediction and validation data from different perspectives, this paper shows that under certain conditions some of the validation metrics can be mathematically related. These relationships may help decision makers to select appropriate validation metrics and the corresponding model acceptance/rejection thresholds.

Various quantitative validation metrics, including the p-value in classical hypothesis testing [16], the Bayes factor in Bayesian hypothesis testing [17], a reliability-based metric [9], and an area metric [12,13], are investigated in this paper. Based on the original definition of Bayes factor, we formulate two types of Bayesian hypothesis testing, one on the accuracy of the predicted mean and standard deviation of model prediction, and the other one on the entire predicted probability distribution of model prediction. These two formulations of Bayesian hypothesis testing can be applied to both fully characterized and partially characterized experiments. The use of these two types of experimental data in the other validation methods is also investigated. The first formulation of Bayesian hypothesis testing, along with the modified reliability-based method and the area metric-based method, takes into account the existence of directional bias. The mathematical relationships among the metrics used in classical hypothesis testing, Bayesian hypothesis testing, and the reliability-based method are investigated.

Section 2 discusses the general procedure of quantitative model validation in the presence of uncertainty. Sections 3 and 4 investigate the aforementioned model validation methods for (1) fully characterized and partially characterized experimental data, (2) application to the case when model prediction and the quantity to be predicted may or may not be uncertain, (3) sensitivity to the existence of directional bias, and (4) the mathematical relationships among some of these validation methods. A numerical example is presented in Section 5 to illustrate the validation of a MEMS switch damping model, which is a generalized polynomial chaos (gPC) surrogate model [18] that has been constructed to predict the squeeze-film damping coefficient. The gPC model is used to replace the expensive micro-scale fluid simulation model and thus expedite the probabilistic analysis of the MEMS device.

2. Quantitative validation of model prediction

Suppose a computational model is constructed to predict a physical quantity. Quantitative model validation methods involve the comparison between model prediction and experimental observation. In this paper, we use the following notations

- Y represents the "true value" of the system response.
- Y_m is the model prediction of this true response Y.
- Y_D is the experimental observation of Y.

The development of quantitative validation metrics is usually based on assumptions regarding Y, Y_m, and Y_D, and these assumptions relate to the various sources of uncertainty and the types of available validation data. In order to select appropriate validation methods, the first step is to identify the sources of uncertainty and the type of validation data.

As mentioned earlier, the available validation data can be from fully characterized or partially characterized experiments. In the case of fully characterized experiments, the model/experimental inputs $\boldsymbol{x}$ are measured and reported as point values. Note that Y and Y_m can be stochastic because of other uncertainty sources other than the input uncertainty. For example, Young's modulus of a certain material can be stochastic due to variation in the

material micro-structure, and the output of a regression model for given inputs is stochastic because of the random residual term. If the experiment is partially characterized, some of the inputs $\boldsymbol{x}$ are not measured or are reported as intervals, and thus the uncertainty in $\boldsymbol{x}$ should be considered. In the Bayesian approach, the lack of knowledge (epistemic uncertainty) about $\boldsymbol{x}$ is represented through a probability distribution (subjective probability). Then, since both Y and Y_m are considered as functions of $\boldsymbol{x}$, they also get treated through probability distributions. Non-probabilistic approaches have also been proposed to handle the epistemic uncertainty; in this paper, we only focus on probabilistic methods.

Note that Y_D results from the addition of measurement error to the true value of the physical quantity Y, i.e., $Y_D = Y + \varepsilon_D$, where ε_D represents measurement error. Hence, the uncertainty in the experimental observation (Y_D) can be split into two parts, the variability in the physical system response (Y) and the measurement error in experiments (ε_D). It should be noted that experimental data with poor quality can hardly provide any useful information on the validity of a model. The discussions in this paper are restricted to the cases when uncertainty in data (due to the error in measuring experimental input and output variables) is limited.

Table 1 summarizes the applicability of the various validation methods investigated in this paper to the different scenarios discussed above, and more details will be presented in Sections 3 and 4.

After selecting a validation method and computing the corresponding metric, another important aspect of model validation is to decide if one should accept or reject the model prediction based on the computed metric and the selected threshold. Sections 3 and 4 will provide some discussions on the decision threshold. The flowchart in Fig. 1 describes a systematic procedure for quantitative model validation.

Table 1
Scenarios of validation and the corresponding methods.

Experimental data	Quantity of interest Y (to be predicted)	Prediction Y_m (from model)	Applicable methods
Fully characterized	Stochastic	Deterministic	1, 2, 4, 5
	Deterministic	Stochastic	1, 2, 4, 5
	Stochastic	Stochastic	1, 2, 3, 4, 5
Partially characterized	Stochastic	Stochastic	1, 2, 3, 4, 5

Methods considered:
1. Classical hypothesis testing.
2. Bayesian interval hypothesis testing.
3. Bayesian equality hypothesis testing.
4. Reliability-based method.
5. Area metric-based method.

Note: Y_D is always treated as a random variable due to measurement error.

3. Hypothesis testing-based methods

Statistical binary hypothesis testing involves deciding between the plausibility of two hypotheses—the null hypothesis (denoted as H_0) and the alternative hypothesis (denoted as H_1). H_0 may be something that one believes could be true, whereas H_1 is a hypothesis opposite to H_0 [19]. For example, given a model for damping coefficient prediction, H_0 can be the hypothesis that the model prediction is equal to the actual damping coefficient, and correspondingly H_1 states that the model prediction is not equal to the actual damping coefficient. The null hypothesis H_0 will be rejected if it fails the test, and will not be rejected if it passes the test. Two types of error can possibly occur from this exercise: rejecting the correct hypothesis (type I error), or failing to reject the incorrect hypothesis (type II error). In the context of model validation, it should be noted that the underlying subject matter domain knowledge is also necessary for the implementation of the hypothesis testing-based methods, especially in the formulation of test hypotheses (H_0 and H_1) and the selection of model acceptance threshold. To formulate appropriate H_0 and H_1 for the validation of a model with stochastic output prediction Y_m, we need to be clear about the physical interpretation of "model being correct". In other words, we need to decide whether or not accurate prediction of statistical moments or the entire PDF of Y_m suggests that the model is correct.

3.1. Classical hypothesis testing

Classical hypothesis testing is well established and has been explained in detail in many statistics textbooks. A brief overview is given here, only to facilitate the development of mathematical relationships between various validation methods in later sections.

In classical hypothesis testing, a test statistic is first formulated and the probability distributions of this statistic under the null and alternative hypothesis are derived theoretically or by approximations. Thereafter, one can compute the value of the test statistic based on validation data and thus calculate the corresponding p-value, which is the probability that the test statistic falls outside a range defined by the computed value of the test statistic under the null hypothesis. The p-value can be considered as an indicator of how good the null hypothesis is, since a better H_0 corresponds to a narrower range defined by the computed value of the test statistic and thus a higher probability of the test statistic falling outside the range.

The practical outcome of model validation should be to provide useful information for decision making in terms of model use. The decisions whether or not to reject the null hypothesis can be made based on the acceptable probabilities of making type I and type II errors (specified by the decision maker). The concept of significance level α defines the maximum probability of making type I error, and the probability of making type II error β can be estimated based on α and the probability distribution of the test statistic under H_1. The null hypothesis will be rejected if the

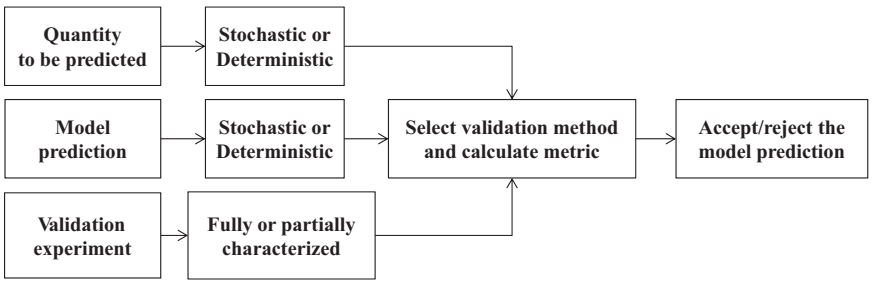

Fig. 1. Decision process in quantitative model validation (Note: The last two steps involve decision making.).

computed p-value is less than α, or the computed β exceeds the acceptable value. Correspondingly, we will reject the model if H_0 is rejected, and accept the model if H_0 is not rejected. An alternative approach to comparing p-value and α is to use confidence intervals. A confidence interval can be constructed based on the confidence level $\gamma = 1-\alpha$, and the null hypothesis will be rejected if the confidence interval does not include the predicted value from the model.

It should be noted that failing to reject H_0 indicates that the accuracy of the model is acceptable, but it does not prove that H_0 is true. Also note that the comparison between p-value and significance level α becomes meaningless when the sample size of experimental data is large. Since almost no null hypothesis H_0 is true, the p-value will decrease as the sample size increases, and thus H_0 will tend to be rejected at a given significance level α as the sample size grows large [19]. In addition, the over-interpretation of p-value and the corresponding significance testing result can be misleading and dangerous for model validation. Criticisms on over-stressing p-value and significance level can be found in [20,21].

Various test statistics have been developed corresponding to different types of hypotheses. The t-test or z-test can be used to test the hypothesis that the mean of a normal random variable is equal to the model prediction; the chi-square test can be used to test the hypothesis that the variance of a normal random variable is equal to the model prediction; and the hypothesis that the observed data come from a specific probability distribution can be tested using methods such as the chi-square test, the Kolmogorov–Smirnov (K–S) test, the Anderson–Darling test and the Cramer test [22]. The tests on variance or probability distribution require relatively large amounts of validation data and thus only the tests on distribution mean are discussed in this subsection, namely the t-test and the z-test.

The t-test is based on Student's t-distribution. Suppose the quantity of interest Y is a normal random variable with mean μ and standard deviation σ, and the measurement error ε_D is a normal random variable with zero mean and standard deviation σ_D. Thus, the experimental observation $Y_D = Y+\varepsilon_D \sim N(\mu,\sigma^2+\sigma_D^2)$, i.e., Y_D also follows a normal distribution with mean μ and variance $\sigma^2+\sigma_D^2$. For the sake of simplicity, let $\sigma_{Y_D} = \sqrt{\sigma^2+\sigma_D^2}$. The validation data are a set of random samples of Y_D with size n (i.e., y_{D1}, y_{D2}, …, y_{Dn}) and the corresponding sample mean is $\overline{Y}_D$ and sample standard deviation is S_D. The variable $(\overline{Y}_D-\mu)/(S_D/\sqrt{n})$ is modeled with a t-distribution with $(n-1)$ degrees of freedom. Therefore, if one assumes that the model mean prediction μ_m (if model prediction is deterministic, μ_m equals to the prediction value) is the mean of Y, i.e., the null hypothesis is $H_0 : \mu = \mu_m$, then the corresponding test statistic t (follows the same t-distribution) is

$$t = \frac{\overline{Y}_D-\mu_m}{S_D/\sqrt{n}} \tag{1}$$

The p-value for the two-tailed t-test (considering both ends of the distribution) can be obtained as

$$p = 2F_{T,n-1}(-|t|) \tag{2}$$

where $F_{T,n-1}$ is the cumulative distribution function (CDF) of a t-distribution with $(n-1)$ degrees of freedom. If the chosen significance level is α, then one will reject the null hypothesis H_0 if $p < a$, or fail to reject H_0 if $p > a$.

The t-test requires a sample size $n \geq 2$ in order to estimate the sample standard deviation S_D. If $n=1$, the z-test can be used instead by assuming that the standard deviation of Y is equal to the standard deviation of model prediction Y_m, i.e., $\sigma = \sigma_m$ and $\sigma_{Y_D} = \sqrt{\sigma_m^2+\sigma_D^2}$. Thus, the variable $(\overline{Y}_D-\mu)/(\sigma_{Y_D}/\sqrt{n})$ follows the standard normal distribution. Under the null hypothesis $H_0 : \mu = \mu_m$, the test statistic z is

$$z = \frac{\overline{Y}_D-\mu_m}{\sigma_{Y_D}/\sqrt{n}} \tag{3}$$

The corresponding p-value for the two-tailed z-test can be computed as

$$p = 2\Phi(-|z|) \tag{4}$$

where Φ is the CDF of the standard normal variable. Similar to the t-test, one will reject H_0 if $p < a$, or fail to reject H_0 if $p > a$.

To compute the probability of making type II error β, an alternative hypothesis H_1 is needed and a commonly used formulation is $H_1 : \mu = \mu_m + \epsilon_\mu$. In t-test, under the alternative hypothesis $H_1 : \mu = \mu_m + \epsilon_\mu$, the t statistic follows a non-central t-distribution with noncentrality parameter $\delta = \sqrt{n}\epsilon_\mu/\sigma_{Y_D}$ [23,24], the probability of making type II error β can then be estimated as

$$\beta = 1-\Pr(|t| > t_{1-\alpha/2,n-1}|\delta) \tag{5}$$

where the term $\Pr(|t| > t_{\alpha/2,n-1}|\delta)$ is called the power of the test in rejecting H_0. In this paper, we use "Pr" in mathematical expressions and equations to represent the probability of a certain event. Similarly, β in the z-test can be estimated as

$$\beta = 1-\Pr(|z-\delta| > \Phi^{-1}(1-\alpha/2)) \tag{6}$$

Note that the above discussion is for the case when both Y and Y_m are stochastic. If Y is deterministic, the standard deviation σ becomes zero; if Y_m is deterministic, σ_m becomes zero. However, the computation procedure of p-value remains the same.

Applying classical hypothesis testing to fully characterized experiments is straightforward as one can directly compare the data against the model predictions for given inputs. For partially characterized experiments, some of the inputs of the model/experiments are available in the form of intervals or probability distributions based on measurements or expert opinions. Let data that have inputs with the same intervals or probability distributions form a sample set. The aforementioned t-test and z-test can then be conducted by comparing the mean of the sample set against the mean of the unconditional probability distribution of model output ("unconditional" means that the probability distribution is not dependent on the point values of inputs). The unconditional probability distribution of model output can be obtained by propagating uncertainty from the input variables to the output variable [25].

3.2. Bayesian hypothesis testing

In probability theory, Bayes' theorem reveals the relationship between two conditional probabilities, e.g., the probability of occurrence of an event A given the occurrence of an event E (denoted as $\Pr(A|E)$), and the probability of occurrence of the event E given the occurrence of the event A (denoted as $\Pr(E|A)$). This relationship can be written as [26]

$$\Pr(E|A) = \frac{\Pr(A|E)\Pr(E)}{\Pr(A)} \tag{7}$$

Suppose event A is the observation of a single validation data point y_D, and event E is hypothesis H_0 being true (or hypothesis H_1 being true). Using Bayes' theorem, we can calculate the ratio between the posterior probabilities of the two hypotheses given validation data y_D as

$$\frac{\Pr(H_0|y_D)}{\Pr(H_1|y_D)} = \frac{\Pr(y_D|H_0)}{\Pr(y_D|H_1)} * \frac{\Pr(H_0)}{\Pr(H_1)} \tag{8}$$

where $\Pr(H_0)$ and $\Pr(H_1)$ are the prior probabilities of H_0 and H_1 respectively, representing the prior knowledge one has on the

validity of these two hypotheses before collecting experimental data; and $\Pr(H_0|y_D)$ and $\Pr(H_1|y_D)$ are the posterior probabilities of H_0 and H_1 respectively, representing the updated knowledge one has after analyzing the collected experimental data. The likelihood function $\Pr(y_D|H_i)$ in Eq. (8) is the conditional probability of observing the data y_D given the hypothesis H_i ($i=0$ or 1), and the ratio $\Pr(y_D|H_0)/\Pr(y_D|H_1)$ is known as the Bayes factor [17,27] and is used as the validation metric.

The original intent of the Bayes factor was to compare the data support for two models [28], and thus the two hypotheses become H_0: model M_i is correct and H_1: model M_j is correct. If $\boldsymbol{\theta_i}$ and $\boldsymbol{\theta_j}$ are the parameters of the two competing models respectively, the Bayes factor is calculated as

$$B=\frac{\Pr(y_D|H_0)}{\Pr(y_D|H_1)}=\frac{\int \Pr(y_D|\boldsymbol{\theta_i})\pi(\boldsymbol{\theta_i})\,d\boldsymbol{\theta_i}}{\int \Pr(y_D|\boldsymbol{\theta_j})\pi(\boldsymbol{\theta_j})\,d\boldsymbol{\theta_j}} \tag{9}$$

where $\pi(\boldsymbol{\theta_i})$ and $\pi(\boldsymbol{\theta_j})$ are the probability density distributions of $\boldsymbol{\theta_i}$ and $\boldsymbol{\theta_j}$ respectively.

In the context of validating a single model, H_0 and H_1 need to be formulated differently. Rebba and Mahadevan [9,7] proposed the equality-based formulation ($H_0: y_m=y_D, H_1: y_m\neq y_D$) and the interval-based formulation ($H_0: |y_m-y_D|<\epsilon, H_1: |y_m-y_D|>\epsilon$) for Bayesian hypothesis testing, where y_m is the model prediction for a particular input $\boldsymbol{x}$.

Consider the case when both the model prediction Y_m and the quantity to be predicted Y are random variables. Two null hypotheses can be formulated: (1) the hypothesis that the difference between the means of Y_m and Y, and the difference between the standard deviations of Y_m and Y, are within desired intervals respectively; (2) the hypothesis that the PDF of Y_m is equal to the PDF of Y. With the first formulation, it is straightforward to derive the likelihood functions under the null and alternative hypothesis, and the existence of directional bias can be reflected in the test, as will be shown below. The advantages of the second formulation are that it avoids the setting of interval width in the first formulation, and leads to a direct test on probability distributions instead of distribution parameters. For the case that either Y or Y_m is deterministic, the first formulation can still be applicable by setting the standard deviation of the deterministic quantity to be zero; however, the second formulation only applies to the case when both Y and Y_m are stochastic. These two formulations are applicable to both fully characterized and partially characterized experiments. Note that in the case where the model output follows a tail-heavy distribution, formulating hypotheses on higher order moments (instead of the mean and standard deviation) may be necessary in order to assess the validity of the model. In this paper, the prediction distribution of the damping model in the application example (Section 5) is close to a normal distribution. Therefore, we only consider hypotheses on the first two moments (mean and standard deviation) and the entire PDF for the purpose of illustration.

Interval hypothesis on distribution parameters: The interval hypothesis can be formulated as $H_0: \epsilon_{\mu 1}\le \mu_m-\mu\le \epsilon_{\mu 2}$, $\epsilon_{\sigma 1}\le \sigma_m-\sigma\le \epsilon_{\sigma 2}$, and $H_1: \mu_m-\mu>\epsilon_{\mu 2}$ or $\mu_m-\mu<\epsilon_{\mu 1}$, $\sigma_m-\sigma>\epsilon_{\sigma 2}$ or $\sigma_m-\sigma<\epsilon_{\sigma 1}$. μ_m and μ are the means of Y_m and Y respectively, and σ_m and σ are the standard deviations of Y_m and Y respectively. $\epsilon_{\mu 1}$, $\epsilon_{\mu 2}$, $\epsilon_{\sigma 1}$ and $\epsilon_{\sigma 2}$ are constants which define the width of interval. Note that $\epsilon_{\mu 1}<\epsilon_{\mu 2}$, $\epsilon_{\sigma 1}<\epsilon_{\sigma 2}$.

Under the interval hypothesis H_0, μ can be any value between $[\mu_m-\epsilon_{\mu 2},\ \mu_m-\epsilon_{\mu 1}]$. So $\mu\sim \mathrm{Unif}(\mu_m-\epsilon_{\mu 2},\mu_m-\epsilon_{\mu 1})$, and the PDF $\pi_0(\mu|\mu_m)=1/(\epsilon_{\mu 2}-\epsilon_{\mu 1})$. Similarly, $\sigma\sim \mathrm{Unif}(\sigma_m-\epsilon_{\sigma 2},\sigma_m-\epsilon_{\sigma 1})$, and the PDF $\pi_0(\sigma|\sigma_m)=1/(\epsilon_{\sigma 2}-\epsilon_{\sigma 1})$. Thus

$$\pi_0(y|\mu_m,\sigma_m)=\iint \pi(y|\mu,\sigma)\pi_0(\mu|\mu_m)\pi_0(\sigma|\sigma_m)\,d\mu\,d\sigma$$

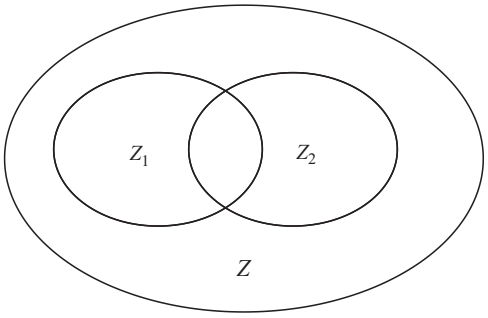

Fig. 2. Graphical illustration of the combined test.

$$=\frac{1}{(\epsilon_{\mu 2}-\epsilon_{\mu 1})(\epsilon_{\sigma 2}-\epsilon_{\sigma 1})}\int_{\sigma_m-\epsilon_{\sigma 2}}^{\sigma_m-\epsilon_{\sigma 1}}\left\{\int_{\mu_m-\epsilon_{\mu 2}}^{\mu_m-\epsilon_{\mu 1}}\pi(y|\mu,\sigma)\,d\mu\right\}d\sigma \tag{10}$$

In the presence of measurement error, the experimental observation is a random variable with conditional probability $\Pr(y_D|y)$. Hence, the likelihood function under the null hypothesis H_0 can be derived as

$$\Pr(y_D|H_0)=\int \Pr(y_D|y)\pi_0(y|\mu_m,\sigma_m)\,dy \tag{11}$$

Under the alternative hypothesis H_1, μ can be any value outside $[\mu_m-\epsilon_{\mu 2},\ \mu_m-\epsilon_{\mu 1}]$, but the uniform distribution is not applicable to infinite space in practical cases. To avoid this issue, we can assume that the possible values of μ are within a finite interval $[\mu_l,\ \mu_u]$ based on the underlying physics. Therefore $\mu\sim \mathrm{Unif}(\mu_l,\mu_m-\epsilon_{\mu 2})\cup(\mu_m-\epsilon_{\mu 1},\mu_u)$, and the PDF $\pi_1(\mu|\mu_m)=1/(\mu_u-\mu_l+\epsilon_{\mu 1}-\epsilon_{\mu 2})$. Similarly, $\sigma\sim \mathrm{Unif}(\sigma_l,\sigma_m-\epsilon_{\sigma 2})\cup(\sigma_m-\epsilon_{\sigma 1},\sigma_u)$, and the PDF $\pi_1(\sigma|\sigma_m)=1/(\sigma_u-\sigma_l+\epsilon_{\sigma 1}-\epsilon_{\sigma 2})$. Thus

$$\pi_1(y|\mu_m,\sigma_m)=\iint \pi(y|\mu,\sigma)\pi_1(\mu|\mu_m)\pi_1(\sigma|\sigma_m)\,d\mu\,d\sigma$$
$$=\frac{A}{(\mu_u-\mu_l+\epsilon_{\mu 1}-\epsilon_{\mu 2})(\sigma_u-\sigma_l+\epsilon_{\sigma 1}-\epsilon_{\sigma 2})} \tag{12}$$

where A is calculated as

$$A=\int_{\sigma_l}^{\sigma_m-\epsilon_{\sigma 2}}\left\{\int_{\mu_l}^{\mu_m-\epsilon_{\mu 2}}\pi(y|\mu,\sigma)\,d\mu+\int_{\mu_m-\epsilon_{\mu 1}}^{\mu_u}\pi(y|\mu,\sigma)\,d\mu\right\}d\sigma$$
$$+\int_{\sigma_m-\epsilon_{\sigma 1}}^{\sigma_u}\left\{\int_{\mu_l}^{\mu_m-\epsilon_{\mu 2}}\pi(y|\mu,\sigma)\,d\mu+\int_{\mu_m-\epsilon_{\mu 1}}^{\mu_u}\pi(y|\mu,\sigma)\,d\mu\right\}d\sigma \tag{13}$$

The likelihood function under H_1 can then be derived as

$$\Pr(y_D|H_1)=\int \Pr(y_D|y)\pi_1(y|\mu_m,\sigma_m)\,dy \tag{14}$$

The Bayes factor for the Bayesian interval hypothesis testing can be calculated by dividing $\Pr(y_D|H_0)$ in Eq. (11) by $\Pr(y_D|H_1)$ in Eq. (14).

It is straightforward to apply this method to the case that Y_m is deterministic and the case that Y is deterministic. For the first case, let σ_m be zero and the rest of the computation remains the same. For the second case, the interval assumption will only be made on μ and μ_m, since we know σ is zero. The other steps will be the same as above.

The directional bias defined in Section 1 can be captured by conducting two separate Bayesian interval hypothesis tests. In the first test, we set $\epsilon_{\mu 1}=-\epsilon_\mu$ and $\epsilon_{\mu 2}=0$, and thus under the null hypothesis $-\epsilon_\mu\le \mu_m-\mu\le 0$. In the second test, we set $\epsilon_{\mu 1}=0$ and $\epsilon_{\mu 2}=\epsilon_\mu$, and thus under the null hypothesis $0\le \mu_m-\mu\le \epsilon_\mu$. The model will fail if any of these two null hypotheses fails the corresponding test. Therefore, the existence of directional bias will increase the chance of a model to fail the combined test. Fig. 2 illustrates this combined test using the concept of data space. Suppose Z is the overall validation data space, Z_1 is the set of data

which does not support the model in the first Bayesian interval hypothesis test, and Z_2 is the set of data which does not support the model in the second test. Then, the union of Z_1 and Z_2 is the set of data that does not support the model combining these two tests.

Equality hypothesis on probability density functions: To further validate the entire distribution of Y_m predicted by a probabilistic model, H_0 or H_1 can be formulated correspondingly as the predicted distribution $\pi(y_m)$ being or not being the true distribution of the quantity to be predicted Y, i.e., $H_0 : \pi(y_m) = \pi(y)$, and $H_1 : \pi(y_m) \neq \pi(y)$. The Bayes factor in this case becomes

$$B = \frac{\Pr(y_D|H_0)}{\Pr(y_D|H_1)} = \frac{\int \Pr(y_D|y)\pi_0(y)\,dy}{\int \Pr(y_D|y)\pi_1(y)\,dy} \tag{15}$$

where $\Pr(y_D|y)$ is the conditional probability of observing noisy data y_D given that the actual value of Y is y; $\pi_0(y)$ is the PDF of Y under the null hypothesis H_0 and hence $\pi_0(y) = \pi(y_m)$; $\pi_1(y)$ is the PDF of Y under the alternative hypothesis H_1. If no extra information about $\pi_1(y)$ is available, it can be assumed as a non-informative uniform PDF. Note that the bounds of this uniform distribution will affect the value of the estimated Bayes factor, and thus it should be carefully selected based on available information.

Note that $\Pr(y_D|y)$ is proportional to the value of the PDF of Y_D conditioned on y which is evaluated at $Y_D = y$, i.e., $\Pr(y_D|y) \propto \pi(y_D|y)$. Therefore, Eq. (15) can be rewritten as

$$B = \frac{\int \pi(y_D|y)\pi_0(y)\,dy}{\int \pi(y_D|y)\pi_1(y)\,dy} \tag{16}$$

Validation data from fully/partially characterized experiments. If the validation data point is from a fully characterized experiment, i.e., all the input variables $\boldsymbol{x}$ of the experiment are measured and the point values of $\boldsymbol{x}$ are available, μ_m and σ_m used in the Bayesian interval hypothesis testing are the mean and standard deviation of the model prediction given the measured $\boldsymbol{x}$, and the PDF of Y_m ($\pi(y_m)$) used in the Bayesian equality hypothesis testing is also conditioned on the measured $\boldsymbol{x}$. If the experiment is partially characterized, i.e., some of the input variables $\boldsymbol{x}$ corresponding to observation y_D are not measured or are reported as intervals, we can assume that $\boldsymbol{x}$ have the PDF $\pi(\boldsymbol{x})$ based on the reported intervals or expert opinions [29]. One can first calculate the unconditional PDF of model prediction $\pi(y_m)$ via propagating uncertainty from $\boldsymbol{x}$ to model output Y_m

$$\pi(y_m) = \int \pi(y_m|\boldsymbol{x})\pi(\boldsymbol{x})\,d\boldsymbol{x} \tag{17}$$

and then calculate μ_m and σ_m from $\pi(y_m)$. If data from both fully characterized and partially characterized experiments are available, we can first calculate Bayes factors corresponding to these two types of data points separately using different μ_m and σ_m (in the Bayesian interval hypothesis testing), or $\pi(y_m)$ (in the Bayesian equality hypothesis testing) as shown above, and then multiply these Bayes factors to obtain an overall Bayes factor, as discussed below.

Bayesian hypothesis testing with multiple data points: If multiple validation experiments are conducted for various test input combinations (including replicas), we can first compute individual Bayes factors B_i's ($i = 1,2,\ldots,n$) corresponding to each data point based on the methods describe above, and then multiply these individual Bayes factors to obtain an overall Bayes factor $B = \prod_{i=1}^{n} B_i$, assuming model predictions (as well as experimental measurements) corresponding to different data points are independent. If this assumption of independence does not hold, a more rigorous approach would be to account for the statistical correlation between different data points in calculating the Bayes factor [30]. If there is a large difference between Bayes factors corresponding to different experiments, we can express these Bayes factors on a logarithmic scale in order to achieve better visualization (e.g., Figs. 6 and 7).

Interpretation of Bayesian hypothesis testing results: If the Bayes factor calculated is greater than 1, it is indicated that the data favor the null hypothesis; if the Bayes factor is less than 1, it is indicated that the data favor the alternative hypothesis. In addition, Jeffreys [31] gave a heuristic interpretation of Bayes factor in terms of the level of support that the hypotheses obtain from data. The threshold value of Bayes factor B_{th} can be related to the so-called Bayes risk in detection theory [32,33], which is the sum of costs due to different decision scenarios—failing to reject the true/wrong hypothesis and rejecting the true/wrong hypothesis. It has been shown that appropriate selection of B_{th} can help reduce the Bayes risk [32]. If one assumes that the cost of making correct decisions (failing to reject the true hypothesis or rejecting the wrong hypothesis) is zero, the costs of type I and type II errors are the same, and the prior probabilities of the null and alternative hypothesis being true are equal, then the resulting $B_{th} = 1$ [33]. However, it should be noted that as a part of the decision making process, the choice of thresholds for Bayes factor inevitably contains subjective elements.

Before collecting validation data, there may be no evidence to support or reject the model. In such cases, it may be reasonable to assume that the prior probabilities of the null hypothesis and alternative hypothesis are equal ($=0.5$), and thus a simple expression of the posterior probability of the null hypothesis can be derived in terms of the Bayes factor [7], which is a convenient metric to assess the confidence in model prediction:

$$\begin{aligned}\Pr(H_0|y_D) &= \frac{\Pr(y_D|H_0)\Pr(H_0)}{\Pr(y_D|H_0)\Pr(H_0)+\Pr(y_D|H_1)\Pr(H_1)} \\ &= \frac{\Pr(y_D|H_0)}{\Pr(y_D|H_0)+\Pr(y_D|H_1)} \\ &= \frac{B}{1+B}\end{aligned} \tag{18}$$

An advantage of Bayesian hypothesis testing is that the posterior probabilities of H_0 and H_1 obtained from the validation exercise can both be used through a Bayesian model-averaging approach [15,34,35] to reflect the effect of the model validation result on the uncertainty in model output as

$$\overline{\pi}(y) = \pi_0(y)\Pr(H_0|y_D) + \pi_1(y)\Pr(H_1|y_D) \tag{19}$$

where $\overline{\pi}(y)$ is the predicted PDF of Y combining the PDFs of Y under the null and alternative hypotheses. Therefore, instead of completely accepting a single model, one can include the risk of using this model in further calculations. This helps to avoid both type I and type II errors, i.e., accepting an incorrect model or rejecting a correct model.

3.3. Relationship between p-value and Bayes factor

Although the p-value in classical hypothesis testing and the Bayes factor B are based on different philosophical assumptions and formulated differently, it has been shown that these two metrics can be mathematically related for some special cases [36]. In the discussion below, the Bayes factor based on the hypothesis of probability density functions for a fully characterized experiment is found related to the p-value in t-test and z-test, if the model prediction Y_m is a normal random variable with mean μ_m and standard deviation σ_m.

Starting from the formula of Bayes factor in Eq. (16), since we assume that the PDF of the quantity to be predicted Y under the alternative hypothesis H_1 is uniform, the integration term in the denominator is not affected by the target model and thus can be

treated as a constant $1/C$. Based on the null hypothesis H_0, the quantity of interest $Y \sim N(\mu_m, \sigma_m^2)$. Recall the relationship $Y_D = Y + \varepsilon_D$, and $\varepsilon_D \sim N(0, \sigma_D^2)$, we know that $Y_D \sim N(\mu_m, \sigma_m^2 + \sigma_D^2)$. Thus the numerator of Eq. (16) can be calculated as

$$\int \pi(y_D|y)\pi_0(y|\boldsymbol{x})\, dy = \frac{1}{\sqrt{\sigma_m^2+\sigma_D^2}}\phi\left(\frac{y_D-\mu_m}{\sqrt{\sigma_m^2+\sigma_D^2}}\right) \tag{20}$$

where $\phi(*)$ is the PDF of the standard norm random variable.

If the variance of measurement noise is negligible compared to the variance of Y_m, i.e., $\sigma_D^2 \ll \sigma_m^2$, we have $\sigma_m^2 + \sigma_D^2 \approx \sigma_m^2$. Also note that for a single data point $\overline{Y}_D = y_D$. Therefore Eq. (16) becomes

$$B = \frac{C}{\sigma_m} * \phi\left(\frac{\overline{Y}_D-\mu_m}{\sigma_m}\right) \tag{21}$$

Based on Eqs. (1) and (3), we have

$$\overline{Y}_D-\mu_m = \begin{cases} t * S_D/\sqrt{n} & \text{for } t\text{-test} \\ z * \sigma_{Y_D}/\sqrt{n} & \text{for } z\text{-test} \end{cases} \tag{22}$$

Substituting Eq. (22) into Eq. (21), we obtain

$$B = \begin{cases} C/\sigma_m * \phi[(t * S_D)/(\sigma_m\sqrt{n})] & \text{for } t\text{-test} \\ C/\sigma_m * \phi[(z * \sigma_{Y_D})/(\sigma_m\sqrt{n})] & \text{for } z\text{-test} \end{cases} \tag{23}$$

where ϕ is the probability density function of a standard normal variable.

From Eq. (23), we can see that the Bayes factor can be related to either the z statistic or the t statistic, and hence it can be related to the p-value in both z-test and t-test. Combining Eqs. (4) and (23), we obtain the relation between Bayes factor and the p-value in the z-test as

$$B = \frac{C}{\sigma_m} * \phi\left[\Phi^{-1}\left(\frac{p}{2}\right)\frac{\sigma_{Y_D}}{\sigma_m\sqrt{n}}\right] \tag{24}$$

where Φ^{-1} is the inverse standard normal CDF. Similarly, the relation between Bayes factor and the p-value in the t-test can be obtained by combining Eqs. (2) and (23) as

$$B = \frac{C}{\sigma_m} * \phi\left\{\left[S_D * F_{T,n-1}^{-1}\left(\frac{p}{2}\right)\right]/(\sigma_m\sqrt{n})\right\} \tag{25}$$

where $F_{T,n-1}^{-1}$ is the inverse CDF of a t-distribution with $(n-1)$ degrees of freedom.

If the chosen significance level in z-test or t-test is α, the corresponding threshold Bayes factor B_{th} can be calculated using Eq. (24) or (25) by letting $p = \alpha$. In that case, the z-test/t-test with significance level α and Bayesian hypothesis testing with the corresponding threshold value B_{th} will give the same model validation result.

4. Non-hypothesis testing-based methods

Besides the binary hypothesis testing methods discussed above, various other validation metrics have been developed to quantify the agreement between models and experimental data from other perspectives, such as the Mahalanobis distance [8,23], Kullback–Leibler divergence [37,38], probability bounds [10], confidence intervals [11], reliability-based metric [9], and area metric [12,13]. Our discussion is restricted to the reliability-based metric and the area metric, since these two metrics have clear probabilistic or physical interpretations regarding the degree of model validity, and both can be applied to validation of a model with multiple input variables using data from discrete test combinations.

4.1. Reliability-based metric

The reliability metric r proposed by Rebba and Mahadevan [9] is a direct measure of model prediction quality and is relatively easy to compute. It is defined as the probability of the difference (d) between observed data (Y_D) and model prediction (Y_m) being less than a given tolerance limit ϵ

$$r = \Pr(-\epsilon < d < \epsilon), \quad d = Y_D - Y_m \tag{26}$$

As mentioned in Section 2, experimental observation is random due to measurement error. In the Bayesian framework, epistemic uncertainty is also represented through probability distributions, and therefore model output is treated as stochastic as well under the combined effect of epistemic and aleatory uncertainty. As the difference between two random variables, d is treated as a random variable, and the probability distribution of d can be obtained from the probability distributions of Y_D and Y_m. For instance, if the model prediction $Y_m \sim N(\mu_m, \sigma_m^2)$, and the corresponding observation $Y_D \sim N(\mu, \sigma_{Y_D}^2)$ (see discussion in Section 3.1), the difference $d \sim N(\mu-\mu_m, \sigma^2+\sigma_D^2+\sigma_m^2)$. For the sake of simplicity, let $\sigma_d = \sqrt{\sigma^2+\sigma_D^2+\sigma_m^2}$. In this case, the reliability-based metric r can be derived as

$$r = \Phi\left[\frac{\epsilon-(\mu-\mu_m)}{\sigma_d}\right] - \Phi\left[\frac{-\epsilon-(\mu-\mu_m)}{\sigma_d}\right] \tag{27}$$

In this paper, experimental data are considered as samples of the random variable Y_D. Therefore, if the size of experimental data set (n) is relatively large, e.g., $n > 30$, the sample variance S_D^2 can be assumed to be a good estimator of $\sigma_{Y_D}^2$ (the true variance of Y_D), which is needed to compute the reliability metric. If n is small and no prior information on σ is available, we can assume that $\sigma = \sigma_m$, which is the same assumption used in z-test. By assuming further that the mean of validation data $\overline{Y}_D$ is equal to μ, Eq. (27) can be re-written as

$$r = \Phi\left[\frac{\epsilon-(\overline{Y}_D-\mu_m)}{\sigma_d}\right] - \Phi\left[\frac{-\epsilon-(\overline{Y}_D-\mu_m)}{\sigma_d}\right] \tag{28}$$

By substituting Eq. (22) into Eq. (28), the relation between the reliability-based metric r and the test statistic in the t-test or z-test is obtained as

$$r = \begin{cases} \Phi[(\epsilon - t * S_D/\sqrt{n})/\sigma_d] + \Phi[(\epsilon + t * S_D/\sqrt{n})/\sigma_d] - 1 & \text{for } t\text{-test} \\ \Phi[(\epsilon - z * \sigma_{Y_D}/\sqrt{n})/\sigma_d] + \Phi[(\epsilon + z * \sigma_{Y_D}/\sqrt{n})/\sigma_d] - 1 & \text{for } z\text{-test} \end{cases} \tag{29}$$

By combining Eqs. (2), (4) and (29), the reliability-based metric can be further related to the p-value in the t-test or z-test as

$$r = \begin{cases} \Phi[(\epsilon - F_{T,n-1}^{-1}(p/2) * S_D/\sqrt{n})/\sigma_d] \\ \quad + \Phi[(\epsilon + F_{T,n-1}^{-1}(p/2) * S_D/\sqrt{n})/\sigma_d] - 1 & \text{for } t\text{-test} \\ \Phi[(\epsilon - \Phi^{-1}(p/2) * \sigma_{Y_D}/\sqrt{n})/\sigma_d] \\ \quad + \Phi[(\epsilon + \Phi^{-1}(p/2) * \sigma_{Y_D}/\sqrt{n})/\sigma_d] - 1 & \text{for } z\text{-test} \end{cases} \tag{30}$$

If one chooses to test models based on a threshold reliability value r_{th} calculated by letting $p = \alpha$ in Eq. (30) above, the result of model validation will be the same as that in the t-test or z-test with significance level α.

Note that the threshold r_{th} used in the reliability-based method represents the minimum probability of the difference d falling within an interval $[-\epsilon, \epsilon]$, and the decision of accepting/rejecting a model can be made based on the decision maker's acceptable level of model reliability.

Since the reliability-based metric is the probability of d being within a given interval, it can also reflect the existence of

directional bias by modifying the intervals. Similar to the Bayesian interval hypothesis testing, we can take two different intervals $[0,\epsilon]$ and $[-\epsilon,0]$, and calculate the corresponding values of metric r^1 and r^2 as

$$r^1 = \Phi\left[\frac{\epsilon-(\mu-\mu_m)}{\sigma_d}\right]-\Phi\left[\frac{-(\mu-\mu_m)}{\sigma_d}\right]$$
$$r^2 = \Phi\left[\frac{-(\mu-\mu_m)}{\sigma_d}\right]-\Phi\left[\frac{-\epsilon-(\mu-\mu_m)}{\sigma_d}\right] \quad (31)$$

By comparing the values of r^1 and r^2 against the threshold $r_{th}/2$ (half of the original threshold value because the width of intervals considered is half of the original one), the model may be judged to have failed the validation test if either r^1 or r^2 is less than $r_{th}/2$.

Note that for the case that the quantity of interest Y is deterministic, σ becomes zero; for the case that the model prediction Y_m is deterministic, σ_m becomes zero.

4.2. Area metric-based method

The area metric proposed by Ferson et al. [13,12] measures the difference between the cumulative distribution functions (CDF) of model output and experimental data, and is defined as

$$d(F_{Y_m},S_{Y_D}) = \int_{-\infty}^{+\infty} |F_{Y_m}(y)-S_{Y_D}(y)|\, dy$$

where $F_{Y_m}(y)$ is the cumulative distribution function (CDF) of model output, and $S_{Y_D}(y)$ is the empirical CDF of experimental data. When the model prediction Y_m is deterministic, the area metric-based method is still applicable by considering the model output to follow a degenerate distribution, i.e., $F_{Y_m}(y)=0$ for $y<y_m$, $F_{Y_m}(y)=1$ for $y\geq y_m$.

Different from the validation metrics in hypothesis testing methods and the reliability-based method, the area metric has no probability interpretation; it is the difference between two CDFs; its physical unit is the same as for the quantity of interest (Y), and thus the area metric can be viewed as a direct measure of prediction error.

The area metric can incorporate fully characterized experiments using the so-called "u-pooling" procedure (transformation from physical space to probability space), and thus to validate models with sparse data on multiple experimental combinations [14]. For a single experimental combination with input $\boldsymbol{x_i}$, suppose $F^m_{\boldsymbol{x_i}}$ is the corresponding CDF of model output Y_m and y_{Di} is the observation, one can compute $u_i = F^m_{\boldsymbol{x_i}}(y_{Di})$ for this experimental combination . Based on the probability integral transform theorem [39], if the observation y_{Di} is a random sample from the probability distribution of model output, the computed u_i will be a random sample from the standard uniform distribution, and thus the empirical CDF of all the u_i's $(i=1,2,\ldots,N)$ should match the CDF of the standard uniform random variable. The difference between these two CDF curves is a measure of the disparity between model predictions and experimental observations. Hence, an area metric in the transformed probability space can be developed as [12]

$$d(F_u,S_u) = \int_0^1 |F_u-S_u|\, du \quad (32)$$

where F_u is the empirical CDF of all the u_i's and S_u is the CDF of the standard uniform random variable. If the value of $d(F_u,S_u)$ is small/large, the model predictions are correspondingly close/not close to experimental observations.

The area metric can reflect the existence of directional bias, i.e., when the experimental observations are consistently below or above the corresponding mean predictions of numerical model. For example, if the model outputs at different test combinations are normal random variables, the values of $F^m_{\boldsymbol{x_i}}(y_{Di})$ will all be less than 0.5 if y_{Di}'s are smaller than the mean of the corresponding normal variables. Therefore, instead of being uniformly distributed between [0,1], u_i's are distributed between [0,0.5], causing a large area between the empirical CDF of u_i and the standard uniform CDF.

The area metric defined in u-space based on Eq. (32) can be transformed back to physical space to retrieve its physical interpretation. As suggested in [12], one can use the CDF of model output (G_y) at a certain point to perform a back-transformation: $y_i = G_y^{-1}(u_i)$, and then compute the area metric in the physical space

$$d(F_y,G_y) = \int |F_y-G_y|\, dy \quad (33)$$

where y_i is the transformed variable with the physical unit of the quantity of interest, and F_y is the empirical CDF of y_i.

Since the area metric has the physical unit of the quantity of interest and represents the prediction error of a model, the threshold of model rejection/acceptance can be set up based on the error tolerance limit in the prediction domain.

5. Numerical example

In this section, the aforementioned model validation methods are illustrated via an application example on damping prediction for MEMS switches. The quantity of interest, the damping coefficient, is treated as a random variable due to the lack of understanding in physical modeling, in other words, the epistemic uncertainty of damping coefficient is represented by a subjective probability distribution following the Bayesian way of thinking; the corresponding computational model is also stochastic as will be shown in Section 5.1.1. The validation data are obtained from fully characterized experiments, and it is found that the directional bias defined in Section 1 exists between model prediction and validation data.

5.1. Damping model and experimental data

Despite the superior performance provided in terms of signal loss and isolation compared with silicon devices [40], the use of RF MEMS switches in applications requiring high reliability is hindered by significant variations in device lifetime [41]. Rigorous quantification of the uncertainty sources contributing to the observed life variations is necessary in order to achieve the design of reliable devices. Within the framework of uncertainty quantification in the modeling of RF MEMS switches, the validation of squeeze-film damping model emerges as a crucial issue due to two factors: (1) damping strongly affects the dynamic behavior of the MEMS switch and therefore its lifetime [42]; (2) it is difficult to accurately model micro-scale fluid damping and available models are applicable to limited regimes [43].

5.1.1. Uncertainty quantification in micro-scale squeeze-film damping prediction

For the purpose of illustration, this study considers damping prediction using the Navier–Stokes slip jump model [44]. Two major sources of uncertainty have been shown to affect the prediction of gas damping [41]. The first one is epistemic uncertainty related to the lack of understanding of fundamental failure modes and related physical models. The second one is aleatory uncertainty in model parameters and inputs due to variability in either the fabrication process or in the operating environment. Uncertainty quantification approaches usually require large numbers of deterministic numerical simulations. In order to reduce the computational cost, a generalized polynomial

chaos (gPC) surrogate model [18] is constructed and trained using solutions of the Navier–Stokes (N–S) equation for a few input combinations, thus avoiding repetitively solving the N–S equation. Note that several other surrogate modeling techniques are also available, including Kriging or Gaussian Process (GP) interpolation [45], support vector machine (SVM) [46], relevance vector machine [47], etc. The gPC model is used for the purpose of illustration. This model approximates the target stochastic function using orthogonal polynomials in terms of the random inputs [41]. A Pth order gPC model $y_m(\boldsymbol{x})$ that approximates a random function $y(\boldsymbol{x})$ can be written as

$$y(\boldsymbol{x}) \approx y_m(\boldsymbol{x}) = \sum_{i=1}^{M} a_i \phi_i(\boldsymbol{x}) + \varepsilon_m, \quad M = \begin{pmatrix} n_x + P \\ n_x \end{pmatrix} \tag{34}$$

where ϕ_i's are the orthonormal polynomial bases such as Legendre polynomials, Hermite polynomials, and Wiener–Askey polynomials; n_x is the dimension of input $\boldsymbol{x}$ and P is the order of the polynomial; ε_m is the error of the gPC model. The coefficients a_i's can be obtained as

$$a_i = \frac{\int y_m(\boldsymbol{x})\phi_i(\boldsymbol{x})\, d\boldsymbol{x}}{\int \phi_i^2(\boldsymbol{x})\, d\boldsymbol{x}} = \frac{1}{h_i}\sum_{j=1}^{N} w_j y(\boldsymbol{x_j})\phi_i(\boldsymbol{x_j}) \tag{35}$$

where $h_i = \int \phi_i^2(\boldsymbol{x})\, d\boldsymbol{x}$ is constant for a given polynomial basis $\phi_i(\boldsymbol{x})$, and $\{\boldsymbol{x_j}, w_j\}_{j=1}^{N}$ is a set of nodes and weights of the quadrature rule for numerical integration.

Based on the calculated damping coefficient values $y(\boldsymbol{x_j})$ at the quadrature nodes $\boldsymbol{x_j}$ by solving the Navier–Stokes Slip Jump model, the gPC model $y_m(\boldsymbol{x})$ can be constructed using Eqs. (34) and (35). For any given input $\boldsymbol{x_k}$, $\mu_m(\boldsymbol{x_k}) = \sum_{i=1}^{M} a_i \phi_i(\boldsymbol{x_k})$ is deterministic, while the residual term ε_m is random. Under the Gauss–Markov assumption, ε_m asymptotically follows a normal distribution with zero mean, and the variance can be estimated as [48,49]

$$\sigma_m^2 = \sigma^2[1 + \boldsymbol{\phi}^T(\boldsymbol{x_k})(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\phi}(\boldsymbol{x_k})] \tag{36}$$

where σ_m^2 is a function of model input $\boldsymbol{x_k}$; the vector $\boldsymbol{\phi}(\boldsymbol{x_k}) = [\phi_1(\boldsymbol{x_k}), \phi_2(\boldsymbol{x_k}), \ldots, \phi_M(\boldsymbol{x_k})]^T$; the matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\boldsymbol{x_1}), \boldsymbol{\phi}(\boldsymbol{x_2}), \ldots, \boldsymbol{\phi}(\boldsymbol{x_N})]^T$; and $\sigma^2 = 1/(N-M)\sum_{j=1}^{N}[\mu_m(\boldsymbol{x_j}) - y(\boldsymbol{x_j})]^2$.

Therefore, for a given input combination $\boldsymbol{x_k}$, the prediction of damping coefficient based on the gPC model is a random variable with normal distribution $N(\mu_m(\boldsymbol{x_k}), \sigma_m(\boldsymbol{x_k}))$. The methods presented in Sections 3 and 4 will be applied to the validation of this predicted distribution.

The example RF MEMS switch modeled as a membrane is shown in Fig. 3. To construct a gPC model for the damping coefficient, the variables $\boldsymbol{x}$ need to be specified first. A probabilistic sensitivity analysis shows that the membrane thickness t, the gap height g, and the frequency ω are the major sources of variability in the damping coefficient. Hence, these three variables are included in the gPC model and they are all measured in the validation experiment, i.e., $\boldsymbol{x} = [t, g, \omega]$. The coefficients a_i's in Eq. (34) are the parameters ($\boldsymbol{\theta}$) of the gPC model. Note that the parameters of this gPC surrogate model are estimated using the simulation data $\{\boldsymbol{x_j}, y(\boldsymbol{x_j})\}$ from the Navier–Stokes Slip Jump model as shown in Eq. (35), instead of using experimental data. Four different gas pressures – 18798.45 Pa, 28664.31 Pa, 43596.41 Pa, and 66661.19 Pa – are considered and correspondingly four gPC models are constructed. This example uses a third order gPC model with Legendre polynomial bases [41]. The representation accuracy of the surrogate model can be quantified by the standard deviation (σ_m) of the surrogate model error term (ε_m) in Eq. (36). In this example, the magnitude of σ_m is limited to less than 5% of the model prediction over the training (sampling) domain, which we consider acceptable. However, it should be noted that the validity of the surrogate model does not guarantee the validity of the original model. We only have access to the surrogate model and validation experimental data; therefore in this example we are only assessing the validity of the surrogate model.

If the original model is to be validated, the number of model evaluations needed to compute a validation metric may be of interest in practice as the original model could be computationally demanding in some problems. In general, z-test, t-test, and Bayesian interval hypothesis testing require less number of model evaluations, since only the mean and variance of the model output are used to compute the validation metric. More model evaluations are needed in Bayesian equality hypothesis testing, the reliability-based method, and the area metric-based method, since the entire probability distribution of model output is needed. In this example, the output of each gPC model follows a normal distribution as shown in Eq. (34), which is fully described by the mean and variance. Therefore, the number of surrogate model evaluations needed in each validation approach is the same.

5.1.2. Experimental data for validation

In the experiment, seven devices with different geometric dimensions are considered. For each of the four pressures mentioned above, five tests are conducted on each of the seven devices with slightly different frequencies, and hence 140 data points are collected. Since the input set $[t, g, \omega]$ are recorded for each of the data points, these experiments are fully characterized and the 140 data points correspond to 140 different test input combinations. That is to say, there are 140 sample sets and each set contains only one sample. We assume that the variability of samples in each sample set is due to measurement error, and measurement errors for different test combinations are treated as statistically independent. Therefore, the sample sets are also statistically independent from one another.

Fig. 4(a) shows a graphical comparison between the mean gPC model prediction and experimental data under the four different pressures by aggregating predictions and data with respect to the

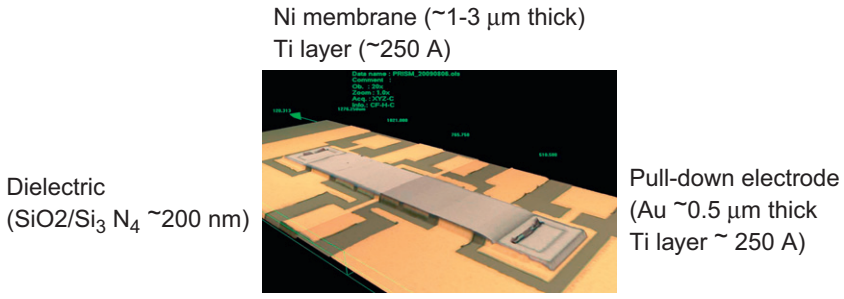

Fig. 3. Example RF MEMS switch (Courtesy: Purdue PRISM center).

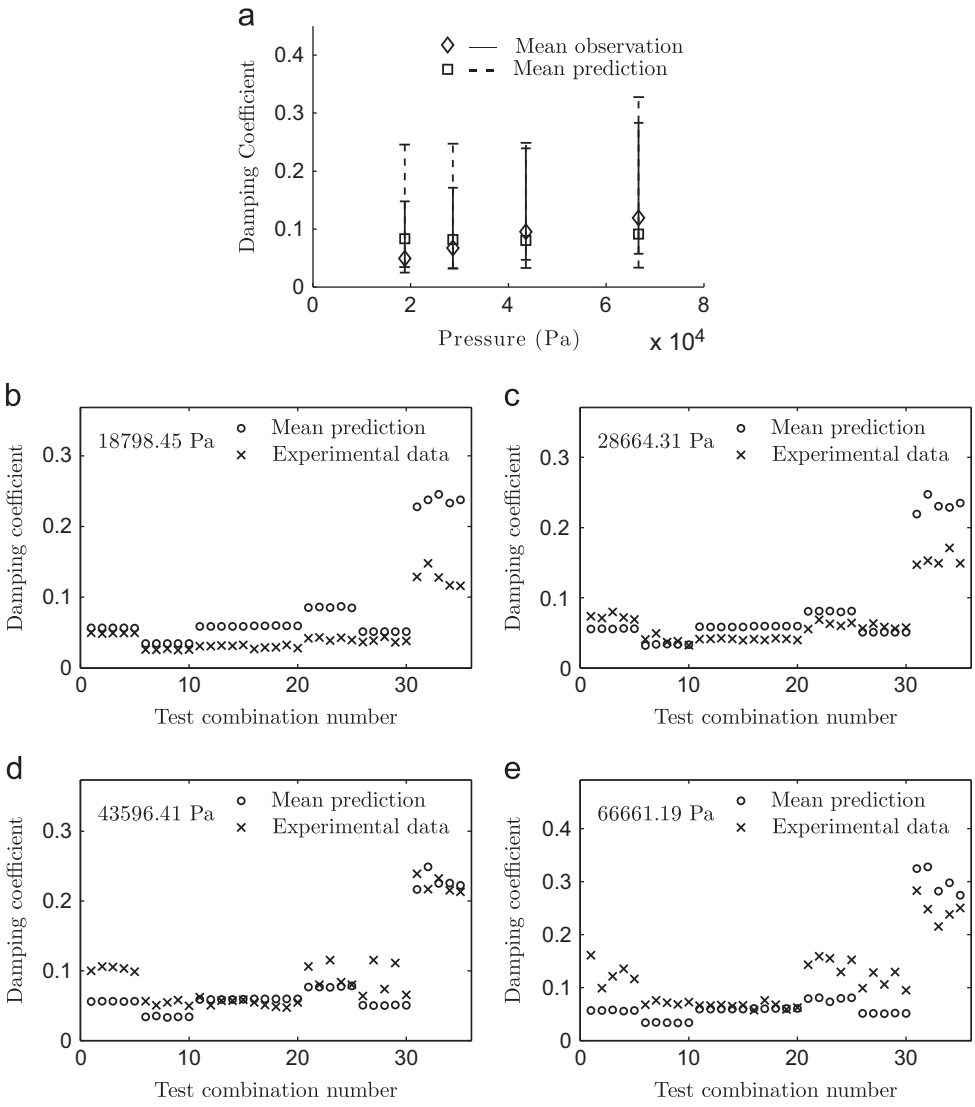

Fig. 4. Graphical comparisons between gPC predictions and experimental data.

35 test combinations for each pressure value. The top/bottom points are correspondingly the maximum/minimum value of model mean predictions and experimental data, and the square/diamond markers are the average values of predictions/data on the 35 test combinations. Note that Fig. 4(a) ignores the difference between the seven devices, and thus should not be considered as a rigorous comparison. A more detailed graphical comparison showing mean prediction of the gPC model vs. experimental data on each of the individual test combinations is provided in Fig. 4(b)–(e).

From the graphical comparison, we can see that the gPC model performs better under the middle two values of pressure. Also note that there is a systematic bias between the gPC model and experimental observations at the low pressure value (18798.45 Pa), i.e., the mean predictions of the gPC model are always larger than the experimental data.

5.2. Validation based on binary hypothesis testing

5.2.1. Classical hypothesis testing

Because the sample size for each experimental combination is only 1, the *t*-test is not applicable and instead *z*-test is used. The *p*-values calculated using Eq. (4) are shown in Fig. 5. The dashed lines in Fig. 5 represent the significance level $\alpha = 0.05$. The model is considered to have failed at the experimental combinations where the corresponding *p*-values fall below the dashed line. Note that a more rigorous test will need to include the probability of making type II error (β). The individual numbers of failures of the four gPC models are shown in Table 2.

5.2.2. Bayesian hypothesis testing

Interval hypothesis on distribution parameters: As discussed in Section 3.2, combination of two Bayesian hypothesis tests based on the interval null hypotheses H_0^1 and H_0^2 respectively can reflect the existence of directional bias. In practical, the parameters ϵ_μ, $\epsilon_{\sigma 1}$, and $\epsilon_{\sigma 2}$ that define the intervals can be determined based on the strictness requirement of validation. For the purpose of illustration, we set $\epsilon_\mu = 0.025$, $\epsilon_{\sigma 1} = -0.005$, and $\epsilon_{\sigma 2} = 0.005$. μ_l and μ_u that define the possible range of μ are set as 0 and 1 respectively since the MEMS device considered is under-damped. σ_l and σ_u are set to be 0.001 and 0.05 respectively. The results of Bayesian interval hypothesis testings are calculated using Eq. (10)–(14), and are shown in Fig. 6 and Table 3.

Equality hypothesis on probability density functions: In this study, the possible values of damping coefficient range from 0 to 1 since the system is under-damped. Hence the limit of

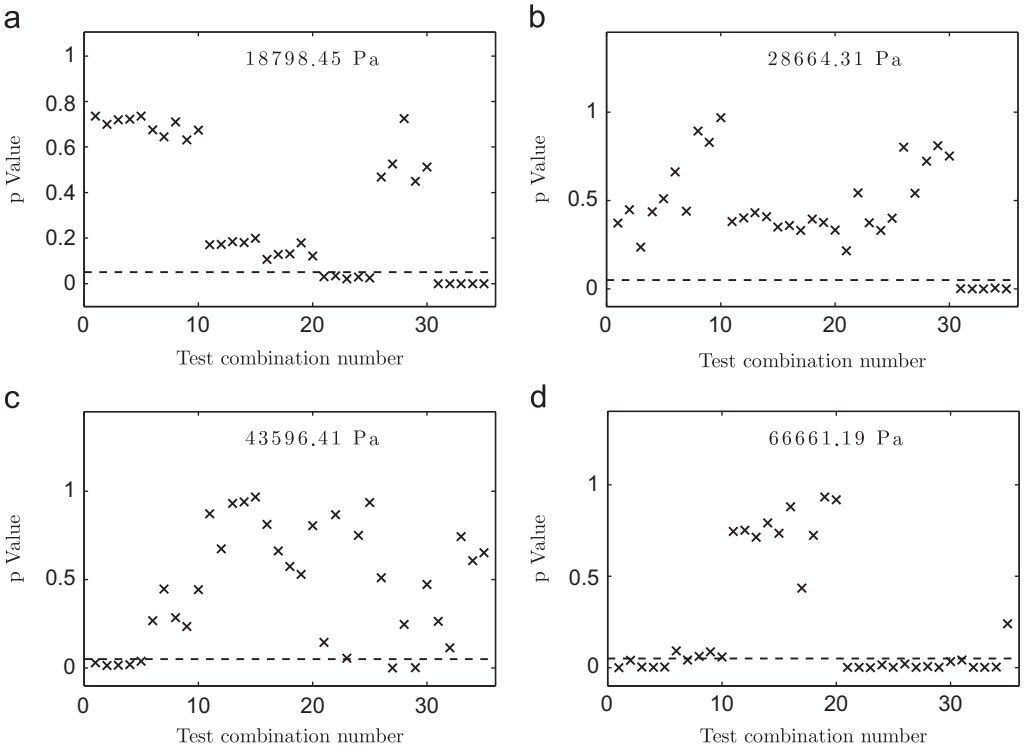

Fig. 5. p-value of z-test.

Table 2
Performance of gPC models in z-test with $\alpha = 0.05$.

Pressure (Pa)	18 798.45	28 664.31	43 596.41	66 661.19
Number of failures	10	5	7	20
Failure percentage	28.6	14.3	20.0	57.1

integration in the denominator of Eq. (16) is [0, 1], while the limit of integration in the numerator is $[-\infty, +\infty]$.

The performance of the gPC models in Bayesian hypothesis testing are shown in Fig. 7 and Table 4. The values of Bayes factor are calculated using Eq. (16), and the threshold Bayes factor $B_{th} = 1$ (this threshold value is chosen based on the discussion in Section 3.2). Although the performance of the gPC model in terms of failure percentage is different for the two hypothesis tests as shown in Tables 2 and 4, if one increases the threshold Bayes factor B_{th} to 2.88, which is calculated using Eq. (24) with $p = 0.05$ in Section 3.3, the result of Bayesian hypothesis testing in terms of the number of failures becomes the same as in the z-test in Section 5.2.1. The reason for this coincidence has been explained in Section 3.3. Note that the performance of the second gPC model (for pressure = 28664.31 Pa) remains the same when B_{th} is raised from 1 to 2.88, and this can be easily observed from Fig. 7(b).

By comparing the results based on interval hypothesis on distribution parameters and equality hypothesis on probability density functions (Tables 3 and 4), it can be observed that the performance of the gPC model for pressure 18798.45 Pa in the first test is significantly worse than in the second test, while the models for the other three pressures have similar failure percentages in these two tests. As shown in Fig. 4(b), the data are all located below the mean predictions of this gPC model, which indicates the existence of directional bias, and thus the gPC model for pressure 18798.45 Pa performs worse in the Bayesian interval hypothesis testing.

5.3. Validation using non-hypothesis testing-based methods

5.3.1. Reliability-based metric

Fig. 8 and Table 5 show the calculated values of the reliability-based metric r, r^1 and r^2 (Eqs. (27) and (31)), the failure percentage of models with $\epsilon = 0.025$ and the decision criterion $r_{th} = 0.2325$. This decision criterion is obtained using Eq. (30) with the significance level $\alpha = 0.05$, and thus the results of validation (comparing r with r_{th}) in terms of failure percentage are the same as in the z-test in Section 5.2.1. It can also be observed that the failure percentage of the gPC model for pressure 18798.45 Pa increases significantly in the test that comparing r^1 and r^2 with $r_{th}/2$ due to the existence of directional bias.

5.3.2. Area metric-based method

The area metrics for the four gPC models in both u-space and physical space are computed using Eqs. (32) and (33), and the results are shown in Fig. 9 and Table 6. Note that the gPC model for pressure 18798.45 Pa has the highest area value in u-space. This is due to the directional bias between mean predictions and experimental data, and it is reflected in the area metric as discussed in Section 4.2. Since the area metric in physical space ($d(F_y, G_y)$) can be interpreted as prediction error, the decision of rejecting/accepting the models can be made by comparing the values of $d(F_y, G_y)$ against a specified tolerance limit. If we use the same tolerance limit $\epsilon = 0.025$ as in the reliability-based method, the gPC model for pressure 66661.19 Pa will be rejected as the corresponding area metric ($= 0.033$) is larger than 0.025, whereas the other three gPC models will be accepted.

5.4. Discussion

This section demonstrated a numerical example of validating the gPC surrogate model for the RF switch damping coefficient using the validation methods presented in Sections 3 and 4, and

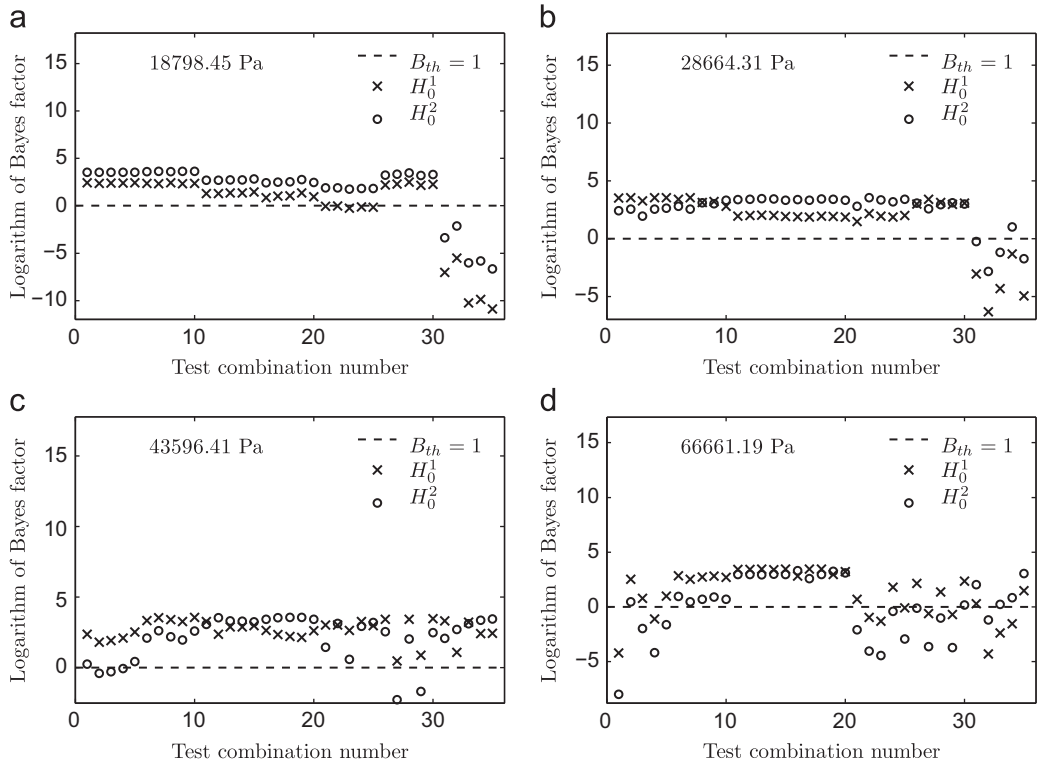

Fig. 6. Bayes factor in interval-based hypothesis testing (on logarithmic scale).

Table 3
Performance of gPC models in interval-based Bayesian hypothesis testing with $\log B_{th} = 0$.

		Pressure (Pa)	18798.45	28 664.31	43 596.41	66 661.19
H_0^1:	$-\epsilon_\mu \le \mu_m - \mu \le 0$	Number of failures	10	5	0	10
	$\epsilon_{\sigma 1} \le \lvert\sigma_m - \sigma\rvert \le \epsilon_{\sigma 2}$	Overall Bayes factor	3.1	58.3	92.9	44.1
H_0^2:	$0 \le \mu_m - \mu \le \epsilon_\mu$	Number of failures	5	4	5	14
	$\epsilon_{\sigma 1} \le \lvert\sigma_m - \sigma\rvert \le \epsilon_{\sigma 2}$	Overall Bayes factor	63.9	87.1	74.1	1.4
	Combined test	Number of failure	10	5	5	16
		Failure percentage	28.6	14.3	14.3	45.7

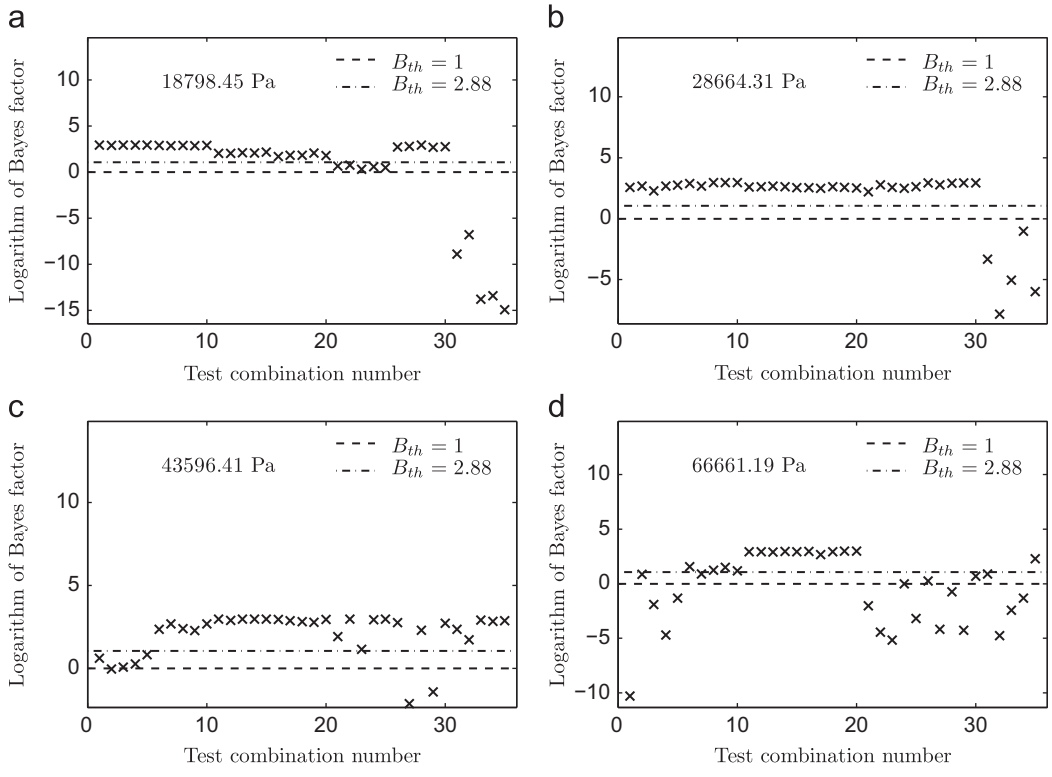

Fig. 7. Bayes factor in equality-based hypothesis testing (on logarithmic scale).

Table 4
Performance of gPC models in equality-based hypothesis testing with $\log B_{th} = 0$.

Pressure (Pa)	18 798.45	28 664.31	43 596.41	66 661.19
Number of failures	5	5	3	15
Failure percentage	14.3	14.3	8.6	42.9
Overall Bayes factor (log-scale)	7.4	57.2	72.3	−10.2

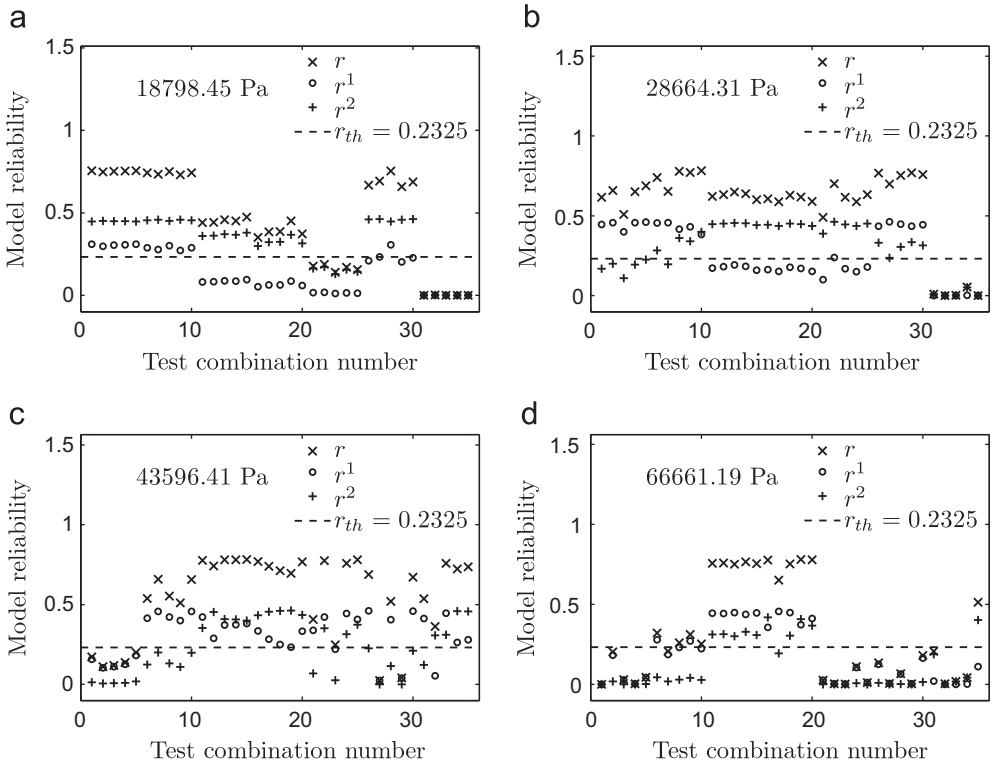

Fig. 8. Reliability-based metric.

Table 5
Performance of gPC models in reliability-based method with $r_{th} = 0.69$.

	Pressure (Pa)	18 798.45	28 664.31	43 596.41	66 661.19
r vs. r_{th}	Number of failures	10	5	7	20
	Failure percentage	28.6	14.3	20.0	57.1
r^1 and r^2 vs. $r_{th}/2$	Number of failures	20	7	12	25
	Failure percentage	57.1	20.0	34.3	71.4

140 fully characterized experimental data points. Based on the performance of the gPC model in these validation tests, it can be concluded that the prediction of the gPC model has better agreement with observation under the middle two values of pressure (28664.31 Pa and 43596.41 Pa), whereas less agreement can be found under the lowest and highest pressure values (18798.45 Pa and 66661.19 Pa). The decision on model acceptance can be formed based on the failure percentages with the hypothesis testing methods and the reliability-based method, and the values of area metric, given the desired prediction error tolerance. It is shown that the z-test and the reliability-based metric give the same results in terms of failure percentage when r_{th} is selected corresponding to the significance level α used in z-test. Similarly, classical and Bayesian hypothesis testing give the same result by choosing a specific threshold Bayes factor as shown in Section 3.3. It is also shown that the existence of directional bias can be reflected in the Bayesian interval hypothesis testing, reliability-based method with modified intervals, and the area metric-based method. Models that have directional bias will perform worse in these three validation methods than in classical hypothesis testing and in Bayesian hypothesis testing with equality hypothesis on probability density functions.

6. Conclusion

This paper explored various quantitative validation methods, including classical hypothesis testing, Bayesian hypothesis testing, a reliability-based method, and an area metric-based method, in order to validate computational model prediction. The numerical example featured a generalized polynomial chaos (gPC) surrogate model which predicts the micro-scale squeeze-film damping coefficient for RF MEMS switches.

A Bayesian interval hypothesis testing-based method is formulated, which validates the accuracy of the predicted mean and standard deviation from a model, taking into account the existence

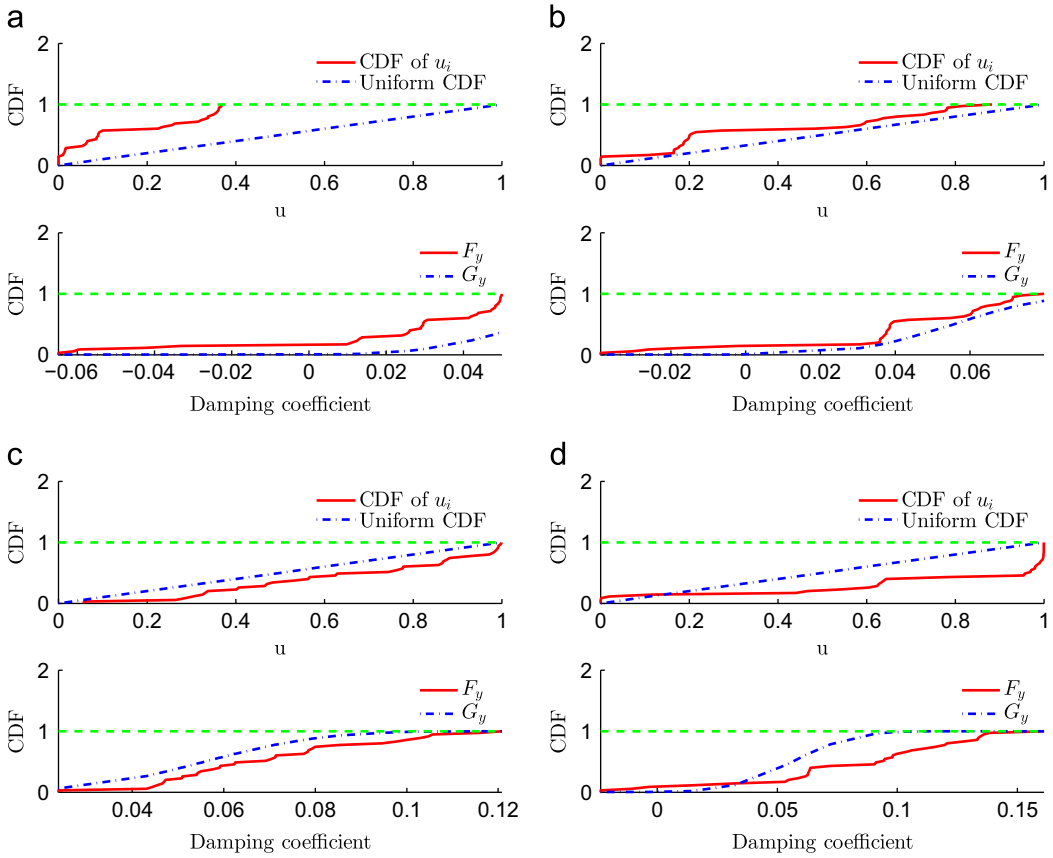

Fig. 9. Comparison of CDFs in the u-space and the physical space.

Table 6
Area metric for gPC models.

Pressure (Pa)	18 798.45	28 664.31	43 596.41	66 661.19
$d(F_u,S_u)$	0.343	0.139	0.151	0.249
$d(F_y,G_y)$	0.024	0.014	0.013	0.033

of directional bias. Further, Bayesian hypothesis testing to validate the entire PDF of model prediction is formulated. These two formulations of Bayesian hypothesis testing can be applied to both fully characterized and partially characterized experiments, and the case when multiple validation points are available. It is shown that while the classical hypothesis testing is subject to type I and type II errors, the Bayesian hypothesis testing can minimize such risk by (1) selecting a risk-based threshold and (2) subsequent model averaging using posterior probabilities. It is observed that under some conditions, the p-value in the z-test or t-test can be mathematically related to the Bayes factor and the reliability-based metric.

The area metric in the transformed probability space (u-space) is shown to be sensitive to the direction of bias between model predictions and experimental data, and so are the Bayesian interval hypothesis testing-based method and the reliability-based method. The Bayesian model validation result and reliability-based metric can be directly incorporated in long-term failure and reliability analysis of the device, thus explicitly accounting for model uncertainty [30]. In addition, due to the use of likelihood function in the Bayesian hypothesis testing, the Bayesian model validation method can be extended to the case that the validation data is in the form of interval, as shown in [29,50].

Acknowledgments

This paper is based upon research partly supported by the Department of Energy [National Nuclear Security Administration] under Award Number DE-FC52-08NA28617 to Purdue University (Principal Investigator: Prof. Jayathi Murthy), and Subaward to Vanderbilt University. The support is gratefully acknowledged. The authors also thank the U. S. DOE (NNSA) PSAAP Center for Prediction of Reliability, Integrity and Survivability of Microsystems (PRISM) at Purdue University for providing the models and validation data for the numerical example.

References

[1] AIAA, AIAA guide for the verification and validation of computational fluid dynamics simulations. American Institute of Aeronautics and Astronautics, AIAA-G-077-1998, Reston, VA, 1998.
[2] ASME, Guide for verification and validation in computational solid mechanics. American Society of Mechanical Engineers, ASME Standard V&V 10-2006, New York, NY; 2006.
[3] Oberkampf W, Trucano T. Verification and validation in computational fluid dynamics. Progress in Aerospace Sciences 2002;38(3):209–72, http://dx.doi.org/10.1016/S0376-0421(02)00005-2.
[4] Hartmann C, Smeyers-Verbeke J, Penninckx W, Vander Heyden Y, Vankeerberghen P, Massart D. Reappraisal of hypothesis testing for method validation: detection of systematic error by comparing the means of two methods or of two laboratories. Analytical Chemistry 1995;67(24):4491–9, http://dx.doi.org/10.1021/ac00120a011.
[5] Hills RG, Trucano TG. Statistical validation of engineering and scientific models : a maximum likelihood based metric. Sandia Technical Report (SAND2001-1783); 2001.
[6] Hills RG, Leslie IH. Statistical validation of engineering and scientific models: validation experiments to application. Sandia Technical Report (SAND2003-0706); 2003.
[7] Rebba R, Mahadevan S, Huang S. Validation and error estimation of computational models. Reliability Engineering & System Safety 2006;91(10–11):1390–7, http://dx.doi.org/10.1016/j.ress.2005.11.035.

[8] Rebba R, Mahadevan S. Validation of models with multivariate output. Reliability Engineering & System Safety 2006;91(8):861–71, http://dx.doi.org/10.1016/j.ress.2005.09.004.
[9] Rebba R, Mahadevan S. Computational methods for model reliability assessment. Reliability Engineering & System Safety 2008;93(8):1197–207, http://dx.doi.org/10.1016/j.ress.2007.08.001.
[10] Urbina A, Paez TL, Hasselman T, Wathugala W, Yap K. Assessment of model accuracy relative to stochastic system behavior. In: 44 th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference, 2003.
[11] Oberkampf W, Barone M. Measures of agreement between computation and experiment: validation metrics. Journal of Computational Physics 2006;217(1):5–36, http://dx.doi.org/10.1016/j.jcp.2006.03.037.
[12] Ferson S, Oberkampf W, Ginzburg L. Model validation and predictive capability for the thermal challenge problem. Computer Methods in Applied Mechanics and Engineering 2008;197(29–32):2408–30, http://dx.doi.org/10.1016/j.cma.2007.07.030.
[13] Ferson S, Oberkampf W. Validation of imprecise probability models. International Journal of Reliability and Safety 2009;3(1):3–22, http://dx.doi.org/10.1504/IJRS.2009.026832.
[14] Liu Y, Chen W, Arendt P. Toward a better understanding of model validation metrics. Journal of Mechanical Design 2011;133(7):071005, http://dx.doi.org/10.1115/1.4004223.
[15] Zhang R, Mahadevan S. Bayesian methodology for reliability model acceptance. Reliability Engineering & System Safety 2003;80(1):95–103, http://dx.doi.org/10.1016/S0951-8320(02)00269-7.
[16] Schervish MJ. P values: what they are and what they are not. American Statistician 1996;50(3):203–6, http://dx.doi.org/10.2307/2684655.
[17] O'Hagan A. Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society. Series B (Methodological) 1995;57(1):99–138.
[18] Xiu D, Karniadakis G. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM Journal on Scientific Computing 2002;24(2):619–44, http://dx.doi.org/10.1137/S1064827501387826.
[19] Marden J. Hypothesis testing: from p values to Bayes factors. Journal of the American Statistical Association 2000;95(452):1316–20, http://dx.doi.org/10.2307/2669779.
[20] Ziliak S, McCloskey D. The cult of statistical significance. University of Michigan Press; 2008.
[21] Ambaum M. Significance tests in climate science. Journal of Climate 2010;23(22):5927–32, http://dx.doi.org/10.1175/2010JCLI3746.1.
[22] Lehmann E, Romano JP. Testing statistical hypotheses. 3rd ed.Springer; 2005.
[23] Srivastava MS. Methods of multivariate statistics. 1st ed.Wiley-Interscience; 2002.
[24] McFarland JM. Uncertainty analysis for computer simulations through validation and calibration. PhD thesis. Vanderbilt University; 2008.
[25] Hills RG, Trucano TG. Statistical validation of engineering and scientific models: background. Sandia Technical Report (SAND99-1256); 1999.
[26] Haldar A, Mahadevan S. Probability, reliability, and statistical methods in engineering design. New York: Wiley; 2000.
[27] Kass R, Raftery A. Bayes factors. Journal of the American Statistical Association 1995;90(430):773–95.
[28] Pericchi LR. Handbook of statistics, vol. 25: Bayesian thinking, modeling and computation, 1st ed. North Holland; 2005. p. 115–49 [Chapter 6].
[29] Sankararaman S, Mahadevan S. Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data. Reliability Engineering & System Safety 2011;96(7):814–24, http://dx.doi.org/10.1016/j.ress.2011.02.003.
[30] Sankararaman S. Uncertainty quantification and integration in engineering systems. PhD thesis. Vanderbilt University; 2012.
[31] Jeffreys H. Theory of probability. 3rd ed. USA, London: Oxford University Press; 1983.
[32] Kay SM. Fundamentals of statistical signal processing, vol. 2: detection theory, 1st ed. Prentice Hall; 1998.
[33] Jiang X, Mahadevan S. Bayesian risk-based decision method for model validation under uncertainty. Reliability Engineering & System Safety 2007;92(6):707–18, http://dx.doi.org/10.1016/j.ress.2006.03.006.
[34] Hoeting J, Madigan D, Raftery A, Volinsky C. Bayesian model averaging: a tutorial. Statistical Science 1999;14(4):382–401.
[35] Zhang R, Mahadevan S. Model uncertainty and Bayesian updating in reliability-based inspection. Structural Safety 2000;22(2):145–60, http://dx.doi.org/10.1016/S0167-4730(00)00005-9.
[36] Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review 2009;16(2):225–37, http://dx.doi.org/10.3758/PBR.16.2.225.
[37] Seghouane A, Bekara M, Fleury G. A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence. Signal Processing 2005;85(7):1405–17, http://dx.doi.org/10.1016/j.sigpro.2005.02.004.
[38] Jiang X, Mahadevan S. Bayesian cross-entropy methodology for optimal design of validation experiments. Measurement Science and Technology 2006;17(7):1895, http://dx.doi.org/10.1088/0957-0233/17/7/031.
[39] Angus J. The probability integral transform and related results. SIAM Review 1994;36(4):652–4, http://dx.doi.org/10.1137/1036146.
[40] Rebeiz GM. RF MEMS: theory, design, and technology. 1st ed. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2003.
[41] Guo X, Li J, Xiu D, Alexeenko A. Uncertainty quantification models for microscale squeeze-film damping. International Journal for Numerical Methods in Engineering 2010;84(10):1257–72, http://dx.doi.org/10.1002/nme.
[42] Snow M, Bajaj A. Uncertainty quantification study for a comprehensive electrostatic mems switch model. In: Third international conference on uncertainty in structural dynamics - USD2010. Leuven, Belgium: Department of Mechanical Engineering, Katholieke Universiteit; 2010.
[43] Bidkar R, Tung R, Alexeenko A, Sumali H, Raman A. Unified theory of gas damping of flexible microcantilevers at low ambient pressures. Applied Physics Letters 2009;94(16):163117, http://dx.doi.org/10.1063/1.3122933.
[44] Gad-el Hak M, editor. MEMS: introduction and fundamentals, 2nd ed. CRC Press; 2005.
[45] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. The MIT Press; 2006.
[46] Vapnik VN. An overview of statistical learning theory. IEEE Transactions on Neural Networks/A Publication of the IEEE Neural Networks Council 1999;10(5):988–99, http://dx.doi.org/10.1109/72.788640.
[47] Tipping M. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 2001;1:211–44.
[48] Liang B, Mahadevan S. Error and uncertainty quantification and sensitivity analysis in mechanics computational models. International Journal for Uncertainty Quantification 2011;1(2):147–61.
[49] Seber G, Wild C. Nonlinear regression. Wiley series in probability and statistics. Wiley-Interscience; 2003.
[50] Sankararaman S, Mahadevan S. Model validation under epistemic uncertainty. Reliability Engineering & System Safety 2011;96(9):1232–41, http://dx.doi.org/10.1016/j.ress.2010.07.014.