Yu Liu
Visiting Predoctoral Student at Northwestern University
School of Mechatronics Engineering,
University of Electronic Science and Technology of China,
Chengdu 611731, China

Wei Chen[1]
Department of Mechanical Engineering,
Northwestern University,
2145 Sheridan Road, Tech B224,
Evanston, IL 60208
e-mail: weichen@northwestern.edu

Paul Arendt
Department of Mechanical Engineering,
Northwestern University,
Evanston, IL 60208

Hong-Zhong Huang
School of Mechatronics Engineering,
University of Electronic Science and Technology of China,
Chengdu 611731, China

Toward a Better Understanding of Model Validation Metrics

Model validation metrics have been developed to provide a quantitative measure that characterizes the agreement between predictions and observations. In engineering design, the metrics become useful for model selection when alternative models are being considered. Additionally, the predictive capability of a computational model needs to be assessed before it is used in engineering analysis and design. Due to the various sources of uncertainties in both computer simulations and physical experiments, model validation must be conducted based on stochastic characteristics. Currently there is no unified validation metric that is widely accepted. In this paper, we present a classification of validation metrics based on their key characteristics along with a discussion of the desired features. Focusing on stochastic validation with the consideration of uncertainty in both predictions and physical experiments, four main types of metrics, namely classical hypothesis testing, Bayes factor, frequentist's metric, and area metric, are examined to provide a better understanding of the pros and cons of each. Using mathematical examples, a set of numerical studies are designed to answer various research questions and study how sensitive these metrics are with respect to the experimental data size, the uncertainty from measurement error, and the uncertainty in unknown model parameters. The insight gained from this work provides useful guidelines for choosing the appropriate validation metric in engineering applications. [DOI: 10.1115/1.4004223]

1 Introduction

With the increase of computing capacity, computational models play an increasing role as predictive models for complex engineering systems. *Model validation*, 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 uses of the model* [1,2], is often required to either choose among alternative models or decide whether a model is acceptable or not before it is used for engineering analysis and design. The fundamental concept and terminology of model validation have been intensively investigated by professional societies and standard committees [3–6]; however, there still exists no unified approach. Our interest in this work is to examine the existing metrics for validation with the consideration of uncertainty in both predictions and physical experiments and to achieve a better understanding of the advantages and disadvantages of each method. Even though design-driven validation metrics have been proposed in previous research [7], our focus here is on assessing the agreement between model predictions and physical observations as opposed to utilizing the design objective to guide a validation process.

By definition, a *validation metric* provides a quantitative measure of agreement between a predictive model and physical observations. In engineering design, the metrics become useful for model selection, when alternative models are being considered and the improvement of an updated model needs to be assessed [8,9]. Additionally, the predictive capability of a computational model needs to be evaluated before it is used in engineering analysis and design [7,10]; such information is crucial in model-based design applications [11–13]. Traditional validation activities are frequently based on deterministic frameworks where no uncertainty is acknowledged in both predictions and physical observations, and the discrepancy between these two sets of data is qualitatively measured through visual inspection of graphic plots [14–16]. However, the measure of agreement from "graphical validation" is not rigorous and varies from person to person [17]. Further, graphical validation does not consider uncertainties which are inevitable in model validation. Based on the work of Kennedy and O'Hagan [18], several different sources of uncertainty can be identified in engineering computer models and experiments, including the *lack of knowledge uncertainty* resulting from model *parameter uncertainty* and *model inadequacy*; *numerical* or *algorithmic uncertainty* introduced from numerical implementations of the computer model such as numerical integration; *experimental uncertainty* in the form of measurement error, systematic error, and random errors; and *interpolation uncertainty* due to lack of samples. Depending on whether the uncertainty can be reduced by collecting more data, the above sources of uncertainty have been broadly classified into two categories in the literature, i.e., *aleatory* versus *epistemic* uncertainty, for which various representations have been proposed correspondingly [19,20]. In this work, it is assumed that stochastic characterizations are used to quantify both types of uncertainty. As a result, model validation needs to quantitatively compare statistical distributions resulting from both simulation predictions and experimental observations. Ideally, a validation metric should also consider the predictive capability of a computational model in both the experimentally tested and untested design regions [21–23].

In this paper, existing validation metrics are classified into different categories based on a few key characteristics. As shown in Table 1, validation metrics belong to either *deterministic* or *stochastic* category based on whether uncertainty is considered in comparing computational and experimental data [24–33]. Moreover, with the consideration of uncertainty, the responses of the

Table 1 Classification of validation metrics

Uncertainty type	Number of response quantities	Number of controllable input settings
Deterministic	Univariate	Single
Stochastic	Multivariate	Multiple

[1]Corresponding author.

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received October 22, 2010; final manuscript received April 3, 2011; published online July 7, 2011. Assoc. Editor: David Gorsich.

computer model prediction and experimental observation could be *univariate* or *multivariate*. A single response of interest is considered in univariate validation [28,30–34], whereas multivariate validation refers to the case where there is more than one response of interest from the same experiment. There are mainly two situations resulting in multivariate responses: (1) A physical experiment and the computational model generate multiple responses or measurements [15,35,36]. These responses usually have distinct magnitudes and scales, e.g., acceleration versus displacement; (2) The responses of interest measured from the same experiment is a function of spatial [37] and temporal [22,38–40] variables. In both cases, there is a strong correlation between any pair of response quantities from the same experiment [21,41,42]. Furthermore, validation metrics may be employed at either a single setting (validation) of controllable inputs or multiple settings of controllable inputs over an intended prediction region. In the latter case, the global predictive capability of a model should be assessed [30,31].

In this paper, taking into account the uncertainty in both predictions and experimental observations, four main types of validation metrics, namely classical hypothesis testing, Bayes factor, frequentist's metric, and area metric, are examined to achieve a better understanding of existing methods under the category of stochastic validation. Our main focus is revealing their differences in fundamental principles and the advantages and disadvantages of each metric. Using mathematical examples, a set of numerical studies are designed to answer various research questions and study how sensitive these metrics are with respect to the experimental data size, the uncertainty from measurement error, and the uncertainty in unknown model parameters.

The remainder of the paper is organized as follows: the desired features of a validation metric and an overview of stochastic validation metrics is provided in Sec. 2. A detailed examination is carried out by numerical examples in Sec. 3 to illustrate the effectiveness of each validation metric. This is followed by a summary and remarks in Sec. 4. Section 5 is the closure of the paper.

2 Overview of Validation Metrics

2.1 Desired Features of Validation Metrics. By combining several different sources [22,43], including our own views, a list of desired features (properties) of model validation metrics is provided as follows with the emphasis on assessing the accuracy of a predictive model. In Sec. 3, the validation metrics will be examined against these desired features.

- A metric should be a *quantitative* measure of the agreement between predictions and physical observations [22]. A metric should also be *objective*, which means that given the predictive and experimental data sets, the metric will produce the same assessment for every analyst independent of their individual preferences [43].
- Criteria used by an analyst/designer to accept a model for prediction should be separate from the validation metric itself [22], i.e., the criteria used for determining whether a model is acceptable or not should not be a part of the metric which is expected to provide a quantitative measurement only.
- The uncertainties resulting from both computer models and experiments need to be considered, together with the *correlation* among multivariate responses [41]. Ideally, the value of a stochastic validation metric should *degenerate to the value from a deterministic comparison* between scalar values when uncertainty is absent [43].
- Additionally, the validation metric is desired to provide a statistical confidence level associated with the amount of available experimental data.
- A validation metric should differentiate between models containing greater and lesser amounts of uncertainty, for example, its value should not be improved if the analyst introduces additional sources of uncertainty into modeling, e.g., widening the probability distribution of a model parameter to gain a greater chance of encompassing physical observations.
- A metric should have the flexibility of measuring the agreement of prediction and physical observations either at *a single setting* or *multiple settings* of controllable inputs over an intended prediction region to assess the *global predictive capability*. This last feature is critical from the viewpoint of engineering design.

2.2 Validation Metrics Considering Uncertainty in Prediction. While various representations of uncertainty, such as probabilistic [15,34,43–45] and interval-based methods [23,28,29,33], are considered in the model validation literature, the comparative study in this work is focused on the metrics that utilize the probabilistic representation and treat the outputs from both model and experiments as stochastic quantities. Validation with stochastic uncertainty involves quantification of the statistical distribution of model predictions and comparing the result with physical observations, which also follows a statistical distribution [21,46]. When multivariate responses are considered, the associated uncertainty needs to be characterized by a multivariate joint probability distribution. In such situations, validation metrics should measure the agreement not only in the marginal distribution of each response but also in the dependency among multiple responses [41]. Other desired properties, shown in Sec. 2.1, should also be taken into account. In the remaining part of Sec. 2.2, four validation metrics commonly used in engineering applications with the consideration of stochastic uncertainty are introduced.

2.2.1 Classical Hypothesis Testing. The primary idea of classical hypothesis testing is to construct an unbiased test statistic with the underlying hypothesis that the physical observations come from the prediction populations. According to the estimated test statistic from the available physical observations, one can decide whether there is enough evidence to reject or not reject a null hypothesis. Two hypotheses, called the null hypothesis H_0 and the alternative hypothesis H_1, should be defined before calculating a test statistic. Under the assumption that the null hypothesis is true, a test statistic S which follows a certain distribution (e.g., t distribution, F distribution, etc.) can be constructed. If the observed value of the test statistic $\hat{S}$, based on the physical observations, falls outside of the critical region $[-S_{\text{crit}}, S_{\text{crit}}]$ of the test statistic S, the null hypothesis will be rejected. The critical region is constructed using a confidence level of $100(1-\alpha)\%$ which indicates $(100\times\alpha)\%$ (e.g., 5%, 10%) of making a type I error, i.e., rejecting a null hypothesis when it is actually true, $\Pr\{|S| > S_{\text{crit}}|H_0\}$ [2,44]. On the other hand, classical hypothesis testing also faces a type II error, i.e., accepting a null hypothesis when it is actually false, $\Pr\{|\hat{S}| < S_{\text{crit}}|H_1\}$. Therefore, prespecifying a higher confidence level $100(1-\alpha)\%$, i.e., lower α, would widen our acceptance region, hence reducing the chance of rejecting a valid null hypothesis but increasing the probability of accepting an invalid null hypothesis [15,30,31,45].

The test statistic S may vary from case to case depending on the hypothesis being tested, the underlying assumption, and the available sample size. Two common hypothesis scenarios include (1) comparison of the means of the predictions and physical observations and (2) comparing the full statistical distributions of the predictions and physical observations. With the assumption that the populations of the predictions and physical observations are normally distributed, the t-test statistic and F-test statistic could be used for examining the consistency of mean and variance, respectively [21,44]. Instead of comparing only the first two moments, hypothesis testing can be extended to measure the differences between the empirical and prediction cumulative density functions (CDFs), e.g., Anderson–Darling test [47], Kolmogorov–Smirnov (K-S) test [48], Cramér–von Mises test [48], etc. Further extensions to multivariate scenarios are discussed in Ref. [41].

For the situation where only one physical experiment data point is available at each validation site, the aforementioned test statistics become inapplicable, since the estimated mean and variance

of physical observations cannot be computed with only one sample. Hills and Trucano [15] proposed a hypothesis testing procedure to handle such scenario, and the method has been applied in several different fields [30,31,35,45,49]. Hypothesis testing states that if the physical observation falls within the performance range obtained from the $100(1-\alpha)\%$ confidence bound of the prediction distribution (could be a joint distribution for multivariate responses [15,35]), the predictive model is consistent with the experimental results at this validation site (i.e., one does not have enough evidence to reject the model); otherwise, one can reject the model with $100(1-\alpha)\%$ confidence level. The downside of this method, as to be illustrated in the numerical study in Sec. 3, is that the method tends not to reject an incorrect model, because the single physical observation happens to fall inside the distribution of a model prediction which has a large amount of uncertainty.

2.2.2 Bayes Factor. The Bayes factor approach to model validation is rooted in Bayesian hypothesis testing. The statistical parameters (e.g., mean and/or standard deviation) of the prediction distribution are treated as random variables and can be updated via the observed physical data. The validation metric is based on the ratio of posterior distributions of the null and alterative hypothesis to infer whether the experimental data comes from one of the statistical populations derived from the predictive model.

With the assumption of normality for predictions and physical observations, the Bayes factor for a general case is defined as

$$B_0 = \frac{\Pr\{\text{data} \mid H_0 : \mu = \mu_0, \sigma = \sigma_0\}}{\Pr\{\text{data} \mid H_1 : \mu \neq \mu_0, \sigma \neq \sigma_0\}} = \frac{L(\text{data} \mid \mu_0, \sigma_0)}{\int\int L(\text{data} \mid \mu, \sigma) f^{\text{prr}}(\mu, \sigma) d\mu d\sigma} \tag{1}$$

In the above formulation, μ_0 and σ_0 are the mean and standard deviation of the model prediction, respectively. The above is an extended formulation compared to the original Bayes factor approach [21,44,47], where only the mean is examined. The Bayes factor is interpreted as a ratio of relative likelihood of the null hypothesis that the experimental data supports the predictions and the alternative hypothesis that the data does not support the predictions, which could follow any competing distribution. The Bayes factor ratio can also be expressed as [32]

$$B_0 = \left. \frac{f^{\text{pst}}(\mu, \sigma \mid \text{data})}{f^{\text{prr}}(\mu, \sigma)} \right|_{\mu=\mu_0, \sigma=\sigma_0} \tag{2}$$

In Eqs. (1) and (2), $L(\text{data} \mid \mu_0, \sigma_0)$ is the likelihood of observing the data under the null hypothesis; $f^{\text{prr}}(\mu, \sigma)$ is the prior density function of mean and standard deviation under the alternative hypothesis and $f^{\text{pst}}(\mu, \sigma \mid \text{data})$ is the posterior density function of mean and standard deviation given the physical observations. If both the prior probabilities of the null and alternative hypotheses are 0.5, the corresponding model acceptance confidence is directly related to the Bayes factor and is written as

$$\Pr(H_0 : \mu = \mu_0, \sigma = \sigma_0 \mid \text{data}) = \frac{B_0}{1 + B_0} \tag{3}$$

Here, the Bayes factor acts as a metric in model validation, and the predictive model is accepted at a test site if $B_0 > 1$. A larger Bayes factor indicates that the physical observations increasingly favor the predictive model and vice versa. This validation metric can be further extended to the multivariate case, where a joint likelihood should be used in Eq. (1). For assessing the global predictive capability of a model, the current approach is to multiply the values of Bayes factor at multiple validation sites [21]. Since the values of Bayes factors could switch between larger than 1.0 and less than 1.0 at different validation sites, the product of all the Bayes factors is sensitive to the locations of validation sites and the end result does not necessarily provide a direct association between its value and the global accuracy for a model.

Instead of focusing on rejecting the null hypothesis, Bayesian model validation emphasizes accepting a null hypothesis with certain posterior confidence $\Pr(H_0 \mid \text{data})$. One cannot accept the null hypothesis in classical hypothesis testing even if the test statistic falls into the critical region, but Bayesian model validation can claim the acceptance with a certain confidence under the given prior knowledge [50]. A better way to understand the Bayes factor approach is that the prior distribution (knowledge) provides the probability of all possible statistical populations that analysts believe the physical observations may come from, and the posterior distribution reflects the updated probabilities of a statistical population that favor these physical observations. If the posterior probability is larger than the prior probability, it means the physical observations support the population from the predictive model and vice versa. Based on the posterior probability, the type I and type II errors can be quantified when the analysts make a decision to accept or reject the model.

In the most recent work of Rebba and Mahadevan [47], an interval Bayesian hypothesis testing is proposed. The core idea is to introduce an interval null hypothesis which includes the desired model accuracy (or allowable error) instead of the point null hypothesis in Eq. (1). The resulting Bayes factor will reflect the model adequacy under the allowable error. Since the specification of the interval (allowable error) is problem dependent, this paper focuses on the point null hypothesis situation presented above that is not problem dependent.

2.2.3 Frequentist's Metric. Instead of making a "yes" or "no" statement about the agreement between the predictive model and physical observations from the classical hypothesis testing and/or the Bayes factor approach, the frequentist's metric proposed by Oberkampf et al. [1,22,51] quantifies the agreement from a different perspective by measuring the distance between the mean of the predictions and the estimated mean of the physical observations. Due to the lack of sufficient physical observations, the uncertainty of the distance is quantified by a confidence bound. The validation metric, at validation sample site x_i, is interpreted as the estimated error in the predictive model $\hat{e}$ with a confidence level of $100(1-\alpha)\%$ that the true error is in the interval

$$\left(\hat{e} - t_{\alpha/2}(N-1) \cdot \frac{s}{\sqrt{N}}, \hat{e} + t_{\alpha/2}(N-1) \cdot \frac{s}{\sqrt{N}} \right) \tag{4}$$

where $t_{\alpha/2}(N-1)$ is the $1-\alpha/2$ quantile of the t distribution for $v = N-1$ degrees of freedom. $\hat{e}$ is the estimated prediction error computed as $\hat{e} = \sum_{l=1}^{N} (y_l^e(x_i) - \mu_{x_i}^m)/N$. $y_l^e(x_i)$ is assumed to be independently, identically, and normally distributed, and s is the estimated standard deviation of the N repetitive physical observations. One can see from Eq. (4) that, as the amount of experimental data increases, the uncertainty of experimental observations reduces and the confidence bound narrows correspondingly.

To assess the global predictive capability of the computational model, a single global validation metric can be computed by integrating the estimated errors over the entire test region. For example, the *average absolute error metric* is defined as

$$|\hat{e}|_{\text{abs}} = \frac{1}{(x_U - x_L)} \int_{x_L}^{x_U} \left| \mu_{x_i}^m - \bar{y}^e(x_i) \right| dx_i \tag{5}$$

with the *average absolute confidence indicator* as

$$|\text{CI}|_{\text{abs}} = \frac{t_{\alpha/2}(N-1)}{(x_U - x_L)\sqrt{N}} \int_{x_L}^{x_U} |s(x_i)| dx_i \tag{6}$$

In the frequentist's metric, the decision criterion of accepting or rejecting a model is separate from the metric itself. Since the frequentist's metric directly measures the distance between the

predictions and experimental observations, the analysts/designers can accept the predictive model with allowable accuracy based on either the metric at a single validation site or the global metric over a region, depending on the practical needs.

2.2.4 Area Metric. With the aim of measuring the agreement of the entire distributions of predictions and observations, Ferson et al. [29,43] proposed a validation metric which uses the area between the prediction distribution $F^m_{x_i}(\cdot)$ and the observation distribution $F^e_{x_i}(\cdot)$. The quantitative measure of the mismatch between the two distributions is formulated as

$$d\left(F^e_{x_i}, F^m_{x_i}\right) = \int_{-\infty}^{+\infty} \left|F^e_{x_i}(x) - F^m_{x_i}(x)\right| dx \tag{7}$$

Compared to the aforementioned validation metrics, the area metric has two important merits. First, the area metric measures the differences between the entire distributions from the observations and predictions. Second, the metric can be used when only a few data points from predictions or experiments are available.

Built upon the idea of the area metric, the *u-pooling* method was proposed by Ferson et al. [43] to measure the agreement between full distributions of predictions and physical observations when the data is sparse at multiple validation sites. A favorable feature of the u-pooling method is that it allows for pooling all physical experiments over the intended prediction domain at multiple validation sample sites into a single aggregated metric. The method begins with calculating a *u-value*, u_i, for each experimental data point by calculating the CDF at $y^e_l(x_i)$ as

$$u_i = F^m_{x_i}(y^e_l(x_i)) \tag{8}$$

where $F^m_{x_i}(\cdot)$ is the corresponding CDF generated by the predictive model at the validation site x_i where the experiment was conducted. Figure 1(*a*) provides an illustration of calculating the values u_i for three observations $y^e(x_1)$, $y^e(x_2)$, and $y^e(x_3)$ at three different validation sites. After pooling all values of u_i for all physical observations, a distribution of u_i is characterized. According to Ferson et al. [43], if each experimental observation $y^e(x_i)$ hypothetically comes from the same "mother" distribution $F^m_{x_i}(\cdot)$, all u_i are expected to constitute a standard uniform distribution on the range of [0, 1]. By comparing the area difference of the empirical distribution of u_i to that of the standard uniform distribution (depicted as the shaded region in Fig. 1(*b*) with a range of $[0, 0.5]$), henceforth termed "u-pooling" metric, can be used to quantify the mismatch or dispersion of the distributions of outputs from both experiments and predictions in a global sense. A larger area difference indicates less agreement and, therefore, a less accurate computer model.

Similar to the frequentist's metric, the area metric does not include any criterion of accepting the model, but only quantifies the discrepancy between predictions and observations. Model acceptance will be determined according to the accuracy requirement of individual problems.

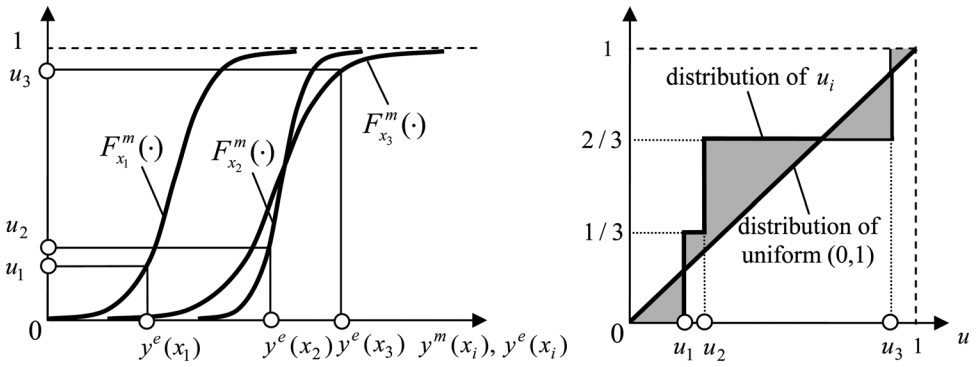
3 Numerical Case Study and Comparisons

3.1 Test Settings. In this work, a set of numerical studies are designed to compare the performance of the four validation metrics against the desired features listed in Sec. 2.1. Two sets of test problems are created. Test set 1 focuses on examining whether the metrics can provide a correct judgment of model validity and how sensitive the validation metrics are with respect to the experimental measurement error. Test set 2 is used to study whether the metric can differentiate between models of greater and lesser uncertainty. The purpose of test set 2 is to explore the desired feature of a validation metric to not increase the chance of accepting a model after more uncertainty is introduced. In all cases, the physical observation data is artificially generated using the following model:

$$y^e(x, \theta) = \sin(x - 0.5\pi\theta) + \cos(\theta + 0.25\pi) + 0.2x + \varepsilon_e \tag{9}$$

where x $(0 \le x \le 8)$ is the deterministic input that defines the settings of controllable variables. θ is a model parameter fixed at $\theta = 1.5$. Measurement error, ε_e, follows a Gaussian distribution $N(0, \sigma^2_\varepsilon)$ and the variance is specified based on the test set. The tested predictive models in test sets 1 and 2 are summarized in Table 2. To study how sensitive the metric is with respect to the number of physical observations, it is assumed that the measurement error is pre-estimated and included as a part of the predictive model.

3.1.1 Test Set 1. In test set 1, two predictive models are validated against the hypothetical physical observations. Model I is considered to be a correct predictive model with model parameter θ exactly equal to 1.5; model II is set to be an incorrect predictive model with θ equal to 1.2. To examine the influence of the amount of uncertainty from measurement errors, two cases, with the standard deviation of measurement error σ_ε equal to 0.08 and 0.2, respectively, are considered. The mean and 95% confidence interval of model predictions and physical observations are plotted in Fig. 2. Because model I is exactly the same as the physical observations, the curves of the computer model and the physical observations completely overlap with each other as shown in Fig. 2(*a*). The same figure shows that model II is incorrect along x, but the discrepancy varies from site to site. For example, there is no overlap between the confidence bounds of the predictions and



(*a*) deriving u values (*b*) computing area metric

Fig. 1 Illustration of the u-pooling method. (*a*) *u*-values at multiple validation sites; (*b*) area metric of mismatch between the empirical distribution of *u*-values and the standard uniform distribution.

Table 2 Formulae of the predictive models

Test set	Model ID	Formula	Notes
Set 1	I	$y_{\mathrm{I}}^{m}(x) = y^{e}(x, \theta = 1.5)$	Model parameter is exact
	II	$y_{\mathrm{II}}^{m}(x) = y^{e}(x, \theta = 1.2)$	Model parameter is incorrect, predictions have discrepancy
Set 2	III*a*	$y_{\mathrm{IIIa}}^{m}(x) = y^{e}(x, \theta \sim N(1.5, 0.1^2))$	Mean of the model parameter is exact, uncertainty is smaller
	III*b*	$y_{\mathrm{IIIb}}^{m}(x) = y^{e}(x, \theta \sim N(1.5, 0.2^2))$	Mean of the model parameter is exact, uncertainty is larger

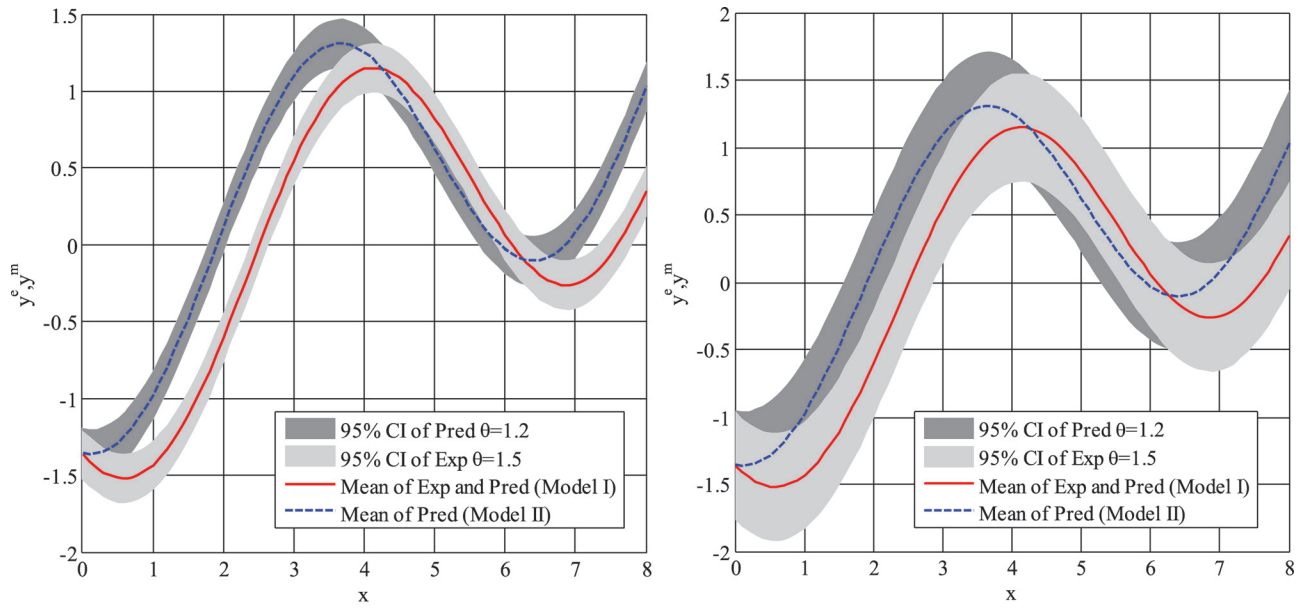
observations in the range of $x = 1.0 \sim 3.5$, but the two curves overlap in the range of $x = 4.0 \sim 6.0$. It is noted from Fig. 2(*b*) that as the uncertainty of measurement error increases, the predictions from the incorrect model II overlap more with the physical observations from Eq. (9). Test set 1 represents the scenario where the distributions of predictions and physical observations partially overlap, e.g., at validation site $x = 6.0$. Ideally, the metrics should indicate that model I is better than model II.

3.1.2 Test Set 2. The predictive models in this test set have an uncertain model parameter θ due to lack of knowledge. In model III*a*, θ follows a Gaussian distribution $N(1.5, 0.1^2)$, whereas the parameter in model III*b* follows $N(1.5, 0.2^2)$. The standard deviation of measurement error σ_ε is set to 0.08 in all cases of this test set. As stated in Sec. 2.1, one desired feature of a validation metric is that one cannot manipulate the improvement of a metric by simply introducing additional uncertainty into the model, e.g., widening the distribution of a model parameter. The mean and 95% confidence bounds of model predictions versus the true physical observations are plotted in Fig. 3 for models III*a* and III*b*. Since model III*a* and model III*b* both have a perfect match of θ in mean but model III*b* has a wider distribution of uncertainty of θ (θ is a constant in Eq. (9) for generating the experimental data), ideally, it is expected that the validation metric should favor model III*a* rather than model III*b*.

3.2 Comparison of Results From Different Validation Metrics. Due to the uncertain nature of physical observations, statistical performance of the validation metrics is assessed in the following comparative studies. For a given number of physical experiments (e.g., $N = 1, 5, 15$), 1000 data sets of physical observations are randomly generated for the replicated experiments at each validation site to evaluate the statistics of the values of the validation metrics. Physical data is sampled from Eq. (9) with $\theta = 1.5$ and the standard deviation of measurement error σ_ε specified as 0.08 or 0.2. Validation site $x = 6.0$ is used as a representative site for point validation. At this site, the predictions and experimental observations partially overlap in test set 1 and the predictions completely encompass the experimental observations in test set 2.

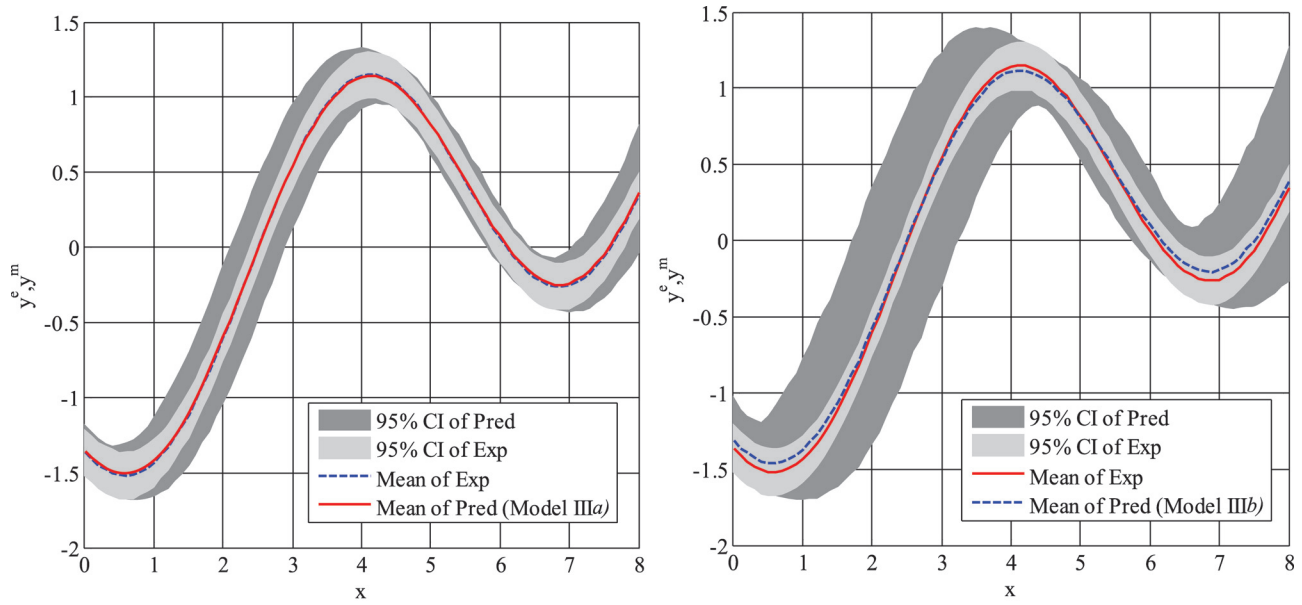
3.2.1 Classical Hypothesis Testing. In our study, the confidence level for each hypothesis test is set to 95%, indicating 5% type I error. As the classical hypothesis testing focuses on whether a model is rejected, the statistical performance of the validation metric is captured by the percentage of sets of physical observations where the predictive model is rejected.

- *Observations from test set 1.* Figure 4 plots the trend of the percentage of model rejection along with the number of physical observations at validation site $x = 6.0$. As shown in Fig. 4, adding more physical observations has a minimal impact on the rejection rate of model I, the correct model. On the other hand, as the number of physical observations increases, the percentage of rejecting model II, the incorrect model, at validation site $x = 6.0$ rises. This observation indicates that the method has a higher chance to reject the incorrect model. Hence, it is able to identify model I as a better model. Comparing the two curves for model II in Fig. 4, it is noted that a larger measurement error ($\sigma_\varepsilon = 0.2$) results in a lower rejection rate because the distributions of predictions and observations overlap more (see Fig. 2).
- *Observations from test set 2.* The general trend of the percentage of model rejection versus the amount of physical observations at site $x = 6.0$ is illustrated in Fig. 5 for models III*a* and III*b*. As more physical observations are collected, the percentage of rejecting model III*a* and model III*b* both rise as shown in Fig. 5. Nevertheless, the increasing rate is lower for model III*a*, which



(*a*) Measurement error $\sigma_\varepsilon = 0.08$ (*b*) Measurement error $\sigma_\varepsilon = 0.2$

Fig. 2 Physical observations (Exp) versus model predictions (Pred) for test set 1



(*a*) Model IIIa vs. Exp (*b*) Model IIIb vs. Exp

Fig. 3 Physical observations (Exp) versus model predictions (Pred) for test set 2

contains a smaller uncertainty of θ. When the number of physical observations is more than one, model III*b*, which has a wider distribution of uncertainty than model III*a*, always has a greater chance to be rejected, which is a desirable feature of validation metrics. However, if only one experimental data point is available, the probability of both model III*a* and model III*b* to not be rejected is zero, as shown in Fig. 5, indicating that the classical hypothesis testing is unable to identify the better model in this situation.

- *Discussion*. According to the desired features stated in Sec. 2.1 and the observations from the above studies, the *advantages* of the classical hypothesis testing method are summarized as follows:

(1) It provides a quantitative test statistic as a validation metric, and the observed value of the test statistic is objectively determined by the available physical observations.
(2) As observed in test set 1, the better model always has a lower chance to be rejected. By adding physical observations, the chance of rejecting an incorrect model always rises. The above analysis indicates that the hypothesis testing metric is capable of identifying the model which has closer predictions with the experimental population.
(3) By considering the correlation among multiple responses in the null hypothesis, the classical hypothesis testing can be further extended to multivariate cases [21,41].

On the other hand, there are several *disadvantages* of this metric:

(1) Since the classical hypothesis method focuses more on model rejection rather than acceptance [2,32], not having enough evidence to reject the null hypothesis (predictive model) does not necessarily indicate that the predictive model is valid. Even though the equivalent testing method has been proposed to stress model acceptance rather than

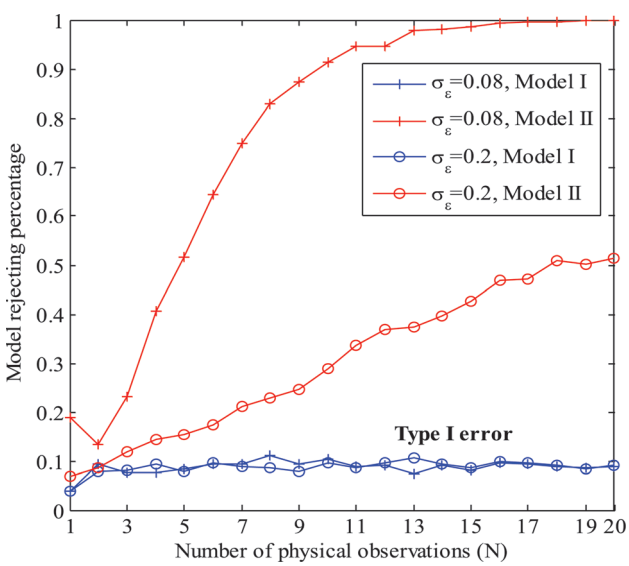


Fig. 4 Percentage of rejecting model versus the number of observations at validation site $x = 6.0$ in test set 1

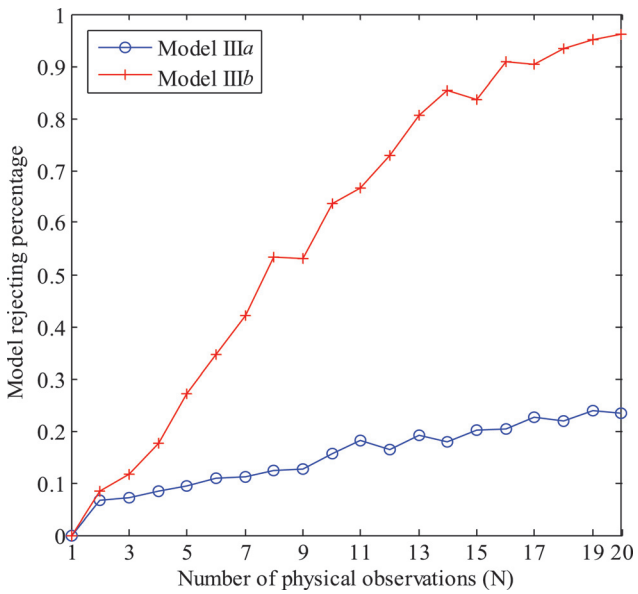


Fig. 5 Percentage of rejecting model versus the number of observations at validation site $x = 6.0$ in test set 2

Table 3 Percentage of $B_0 \geq 1.0$ at $x = 6.0$ among 1000 sets of experimental data in test set 1

	$\sigma_\varepsilon = 0.08$		$\sigma_\varepsilon = 0.2$	
	Model I	Model II	Model I	Model II
$N = 1$	82.2%	53.6%	81.0%	75.8%
$N = 5$	81.1%	14.2%	82.1%	62.1%
$N = 15$	90.2%	0.7%	90.1%	55.7%

rejection [52], there are difficulties in defining the appropriate test statistic [53,54].

(2) Small perturbations in the prespecified confidence level can have a significant impact on rejecting or not rejecting a predictive model [23]. The type II error is usually difficult to quantify for the classical hypothesis testing [47,55,56].
(3) When only one physical observation is available, the classical hypothesis method is more likely to not reject a model when the distribution of predictions encompasses the true distribution of physical observations as observed in test set 2 (see Fig. 5 for $N = 1$). This dramatically increases the risk of type II error.
(4) The Boolean result, either rejecting or not rejecting, does not quantitatively measure the discrepancy between predictions and observations and is not applicable for the case where the asymptotic limit of uncertainty goes to zero.
(5) The classical hypothesis testing method lacks the ability to integrate the validation results at multiple sites, because the conclusions (rejecting or not rejecting) at difference validation sites might conflict with each other.

3.2.2 Bayes Factor. To alleviate the impact of the chosen prior in using the Bayes factor approach, the prior distributions for the mean and standard deviation under the alternative hypothesis are assumed noninformative. Using Eq. (1), the prior distribution of the mean follows a uniform distribution with the range of $[(\mu^m_{x_i} - \sigma^m_{x_i}), (\mu^m_{x_i} + \sigma^m_{x_i})]$, while the standard deviation is also uniform in the range of $[0.5\sigma^m_{x_i}, 2\sigma^m_{x_i}]$; $\mu^m_{x_i}$ and $\sigma^m_{x_i}$ represent the mean and standard deviation of the outputs from predictive models at validation site x_i, respectively. In this scenario, the population of the predictions from models I and II is compared to all competitive populations within the prior mean and standard deviation to infer which population the experimental data could be from. The statistical performance of the validation metric is accounted for by examining the variation of the Bayes factor B_0.

- *Observations from test set 1.* The percentages of B_0 larger than 1.0 (i.e., the predictive model is acceptable) is listed in Table 3. It is found that as the experimental data increases from $N = 1$ to 5, and 15, the chance to accept the correct predictive model (model I) rises, whereas it decreases for the incorrect predictive model (model II). When the magnitude of uncertainty of measurement error is increased from $\sigma_\varepsilon = 0.08$ to $\sigma_\varepsilon = 0.2$, the error does not significantly impact the acceptance percentage for model I. However, due to the overlaps between observations and predictions, which results from the larger measurement error, a higher percentage of accepting model II is observed as seen in Table 3. To examine the variation of B_0, the distributions of $\log(B_0)$ at validation site $x = 6.0$ are plotted in Fig. 6. Having $\log(B_0) > 0$ indicates that physical observations favor the population of the outputs from the predictive model but not all competitive populations within the prior distribution and vice versa. Also, the larger $\log(B_0)$, the greater the confidence level of accepting the predictive model. It can be seen from Fig. 6(*a*), as the number of physical observations increase, the distribution of $\log(B_0)$ of model I shifts toward the right, and the distribution moves toward the left for model II. This indicates that when increasing the amount of physical observations, the confidence level of accepting a model increases for the correct predictive model and decreases for the incorrect predictive model. By comparing Fig. 6(*a*) with Fig. 6(*b*), it can be found that for the incorrect predictive model with a larger measurement error, there is a shift of the distribution of $\log(B_0)$ toward the right. This indicates the confidence level of accepting the incorrect model increases when more experimental uncertainty is introduced.

The global predictive capability of a model is examined by multiplying the Bayes factor from multiple validation sites ($x = 0.0, 1.0, \ldots, 8.0$). The products of the Bayes factor with respect to different amounts of physical observation are shown in Table 4.

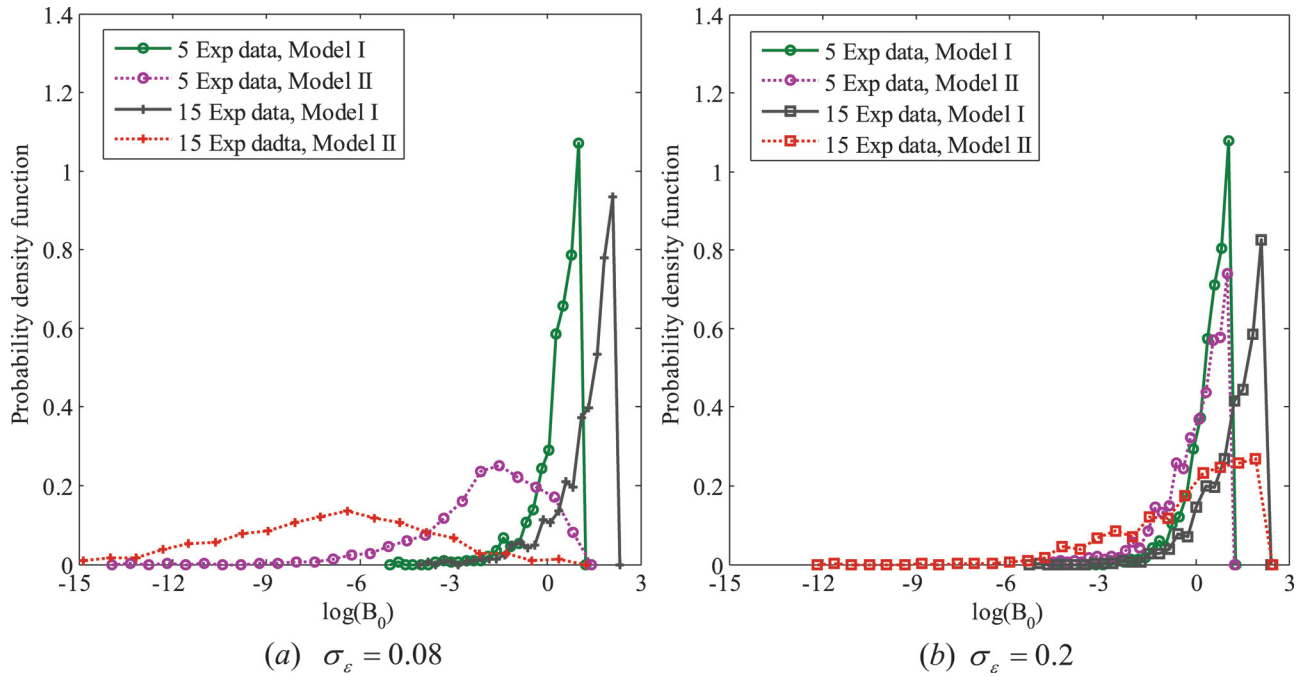


Fig. 6 Distributions of $\log(B_0)$ at validation site $x = 6.0$

Table 4 Mean value of the product of Bayes factor at multiple locations among 1000 sets of experimental data in test set 1

	$\sigma_\varepsilon = 0.08$		$\sigma_\varepsilon = 0.2$	
	Model I	Model II	Model I	Model II
$N=1$	1.46	2.0×10^{-19}	1.48	0.01
$N=5$	5.34	5.0×10^{-131}	5.50	1.2×10^{-14}
$N=15$	116.46	0.0	138.1	9.7×10^{-59}

Results from both $\sigma_\varepsilon = 0.08$ and $\sigma_\varepsilon = 0.2$ overwhelmingly favor model I compared to model II because the magnitudes of the products of Bayes factors are significantly different for these two models.

• *Observations from test set 2*. To avoid the impact from the prior distributions, in test set 2, the priors of mean and standard deviation of predictions under alternative hypothesis are set to be identical for model III*a* and model III*b* and encompasses a large possible range. The mean is set to be a uniform distribution within a range of [−0.1, 0.25]; while the standard deviation is uniformly distributed within [0.05, 0.3]. The percentage of accepting the predictive models and the associated mean of Bayes factors versus the number of available physical observations is plotted in Fig. 7. It is observed that both percentages of accepting model III*a* and model III*b* decrease rapidly as the number of physical observations increases. Regardless of the number of physical observations, model III*b* always has a lower percentage of acceptance, but also a smaller mean of Bayes factors, indicating the confidence of accepting model III*b* is lower than accepting model III*a*.

• *Discussion*. Summarized from our study, the Bayes factor approach has several *advantages*:

(1) The Bayes factor, as a quantitative validation metric, focuses on model acceptance rather than model rejection. As illustrated in the two test sets, even when only one experimental data point is available, the metric still has the capability to identify the predictive model whose outputs are closer to the experimental population.
(2) Widening the uncertainty bounds of predictions does not improve the validation metric as shown in test set 2.
(3) By incorporating the analyst's belief (prior) on the alternative hypothesis, the Bayes factor can be used to quantify the adequacy (confidence level) of accepting the null hypothesis (predictive model is correct) and assess the associated type I and type II errors in making a model accepting or rejecting decision.
(4) To deal with the correlation among multiple responses, the Bayes factor approach can be further extended to multivariate cases by replacing the marginal prior distribution in Eq. (1) to a joint distribution for multiple responses [21,41,42].
(5) As observed from the test sets, adding more physical data drives the distribution of $\log(B_0)$ (confidence level) toward the right direction depending on whether the model is correct or not, which is a desired feature of validation metrics.

Despite its advantages, the Bayes factor approach has several *disadvantages* as listed here:

(1) The Bayes factor is sensitive to the prior knowledge of the alternative hypothesis [32,47] when there is a lack of physical observations.
(2) Since the metric is also rooted in hypothesis testing, it is not applicable for comparing deterministic quantities, i.e., when the asymptotic limit of uncertainty goes to zero in the system.
(3) Since the Bayes factors at different validation sites represent the different degrees (confidence levels) of accepting a predictive model in a statistical sense [47], there is no direct physical meaning when multiplying the values of the Bayes factors at multiple locations for assessing the global accuracy of a predictive model.

3.2.3 Frequentist's Metric. The error (distance) between the means of predictions and observations is measured in the frequentist's metric as opposed to answering a yes or no question about accepting or rejecting a model. To study the impact of the amount of available experimental data, the variation of the estimated errors between means are investigated and compared to their true values.

• *Observations from test set 1*. In test set 1, the true prediction errors at $x = 6.0$ are 0 and 0.093 for model I and model II, respectively. The 95% confidence bounds of estimated errors obtained from the 1000 randomly generated experiments versus the amount

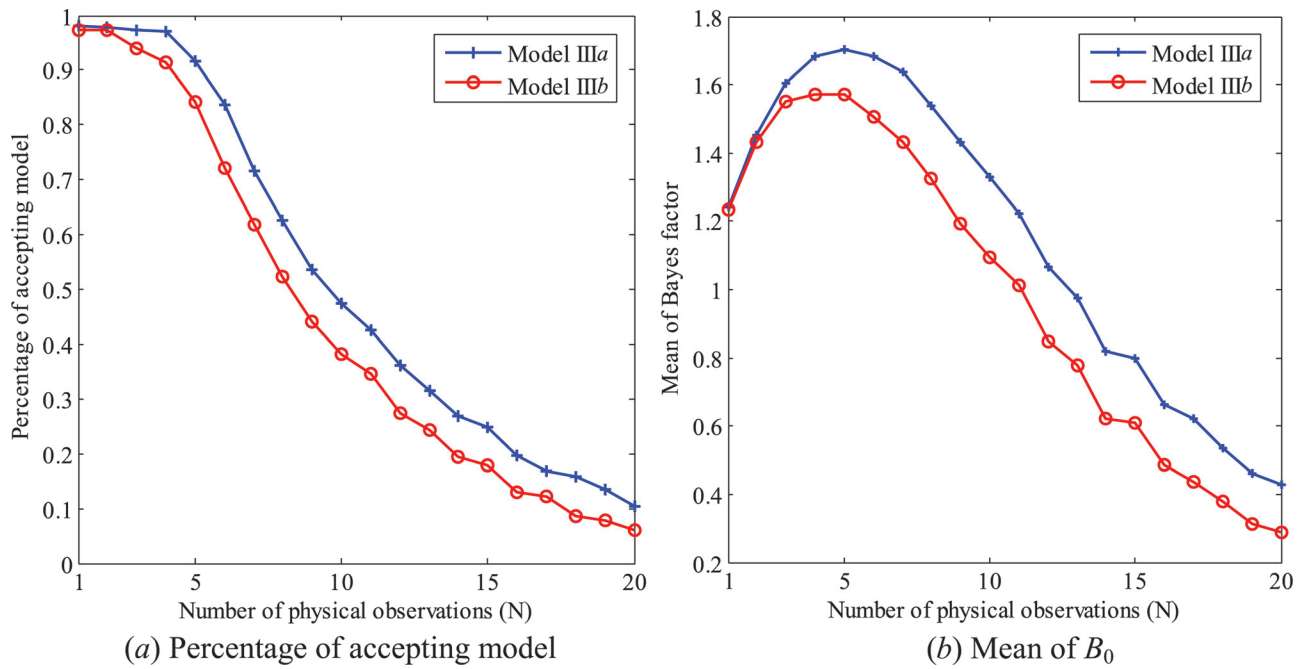


Fig. 7 Impact from the amount of physical observations at $x = 6.0$

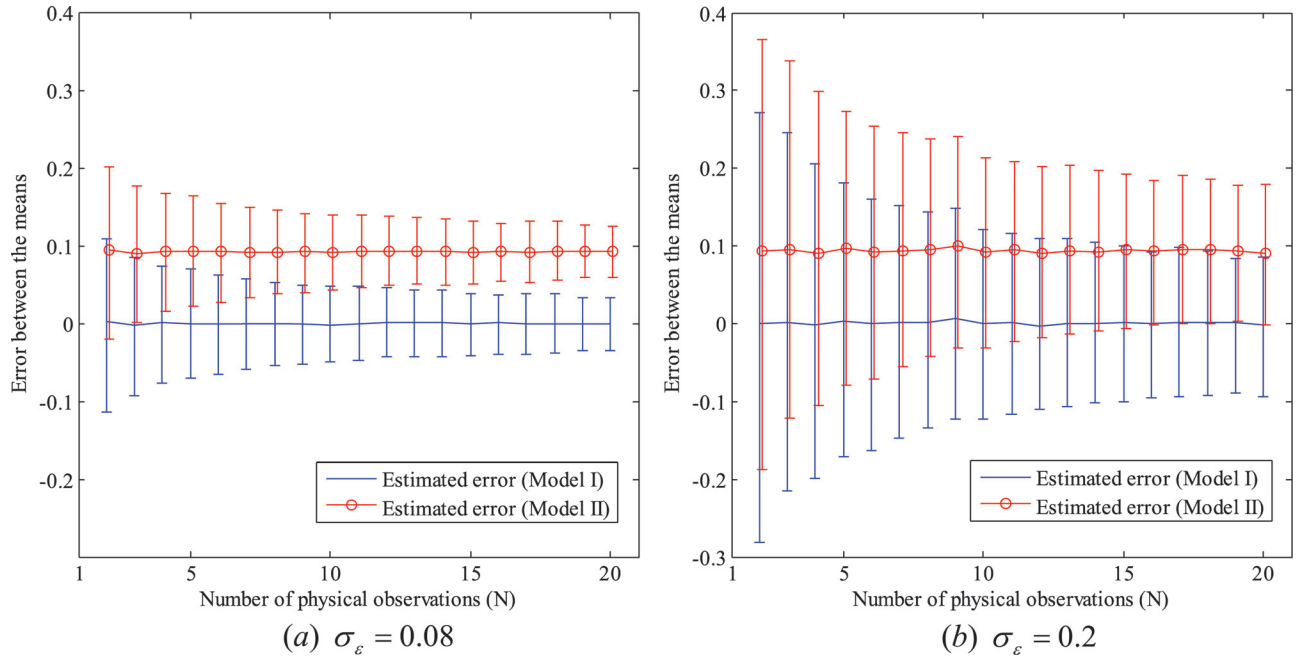


Fig. 8 95% confidence bounds of estimated error versus the number of physical observations at $x = 6.0$ in test set 1

of physical observations are plotted in Fig. 8 for two different magnitudes of experimental measurement error. It is observed that adding more experimental data can reduce the variation of the estimated errors. Also noted from Fig. 8, model I has a smaller error magnitude compared to model II because model I's error magnitudes are closer to zero. Comparing Fig. 8(*a*) to Fig. 8(*b*), it is noted that a larger measurement uncertainty results in a larger variation of the estimated error. By introducing the associated confidence bounds of the estimated error (see Eq. (4)), the frequentist's metric is capable of quantifying the variation of the estimated error due to the lack of sufficient experimental data.

To measure the global predictive capability, the average absolute error metric (see Eq. (5)) is derived as a global metric by integrating the absolute error metric at multiple validation sites ($x = 0.0, 1.0, \ldots, 8.0$). The mean values of the global metric for the 1000 sets of physical observations are listed in Table 5 with $N = 5$ and 15 observations at each site. Results show adding more uncertainty into the measurement error, i.e., increasing σ_ε, increases the average error for both models. Results also indicate that model I, the accurate predictive model, always has a smaller average absolute error metric than model II.

- *Observations from test set 2*. Since the means of predictions from model III*a* and model III*b* are extremely close to the mean of the physical observations (see Fig. 3), the estimated errors of these two predictive model are almost identical at $x = 6.0$. Because the frequentist's metric only measures the distance between the means of computational and physical data, the magnitude of the model parameter uncertainty does not have any impact on the metric value.

Discussion. Several *advantages* of the frequentist's metric are summarized as follows:

(1) The uniqueness of the frequentist's metric is that the agreement between the physical observations and model predictions is objectively quantified by the distance between the means of these two sets of data.
(2) The frequentist's metric allows for integrating the metric (distance) at multiple sites into a global metric to provide a global assessment of model accuracy (nevertheless, this is only done in a limited sense based on the discrepancy of means).
(3) Without specifying a null hypothesis, the frequentist's metric completely separates out the criterion of accepting/rejecting a model from the metric itself.
(4) The adequacy of measurement is quantified by the confidence level associated with estimating the mean of the physical observations as shown in Eq. (4).
(5) It is noted from Eq. (4), by increasing the amount of experimental data (increasing N), the uncertainty bound of the estimated mean of physical observations will shrink.
(6) When uncertainty is absent, the metric is reduced to a deterministic metric, measuring the distance between two scalar values.

Three *disadvantages* of the metrics are noteworthy:

(1) Since the metric can only measure the discrepancy of means, comparing the means will be insufficient to validate a predictive model as in test set 2. The metric also cannot discern model predictability when the tested system has non-negligible random variability.
(2) With multiple responses, the correlations among the responses cannot be accounted for by only examining the central tendency of data. Therefore, this metric is inapplicable to the multivariate scenario.
(3) Equation (4) requires at least two physical observations to compute the estimated error and the associated confidence bounds. When only one experimental data point is available, the metric cannot be used.

Table 5 Mean of average absolute error metric in test set 1

	$\sigma_\varepsilon = 0.08$		$\sigma_\varepsilon = 0.20$	
	$N = 5$	$N = 15$	$N = 5$	$N = 15$
Model I	0.0283	0.0166	0.0714	0.0415
Model II	0.3520	0.3504	0.3586	0.3533

3.2.4 Area Metric. In our study, the statistical performance of the metric is examined by comparing the distributions of the metric obtained from running 1000 sets of physical observations following the given measurement error in the test problems.

- *Observations from test set 1*. Since model I is an exact and correct predictive model, the true area difference between the distributions of predictions and observations at the validation site

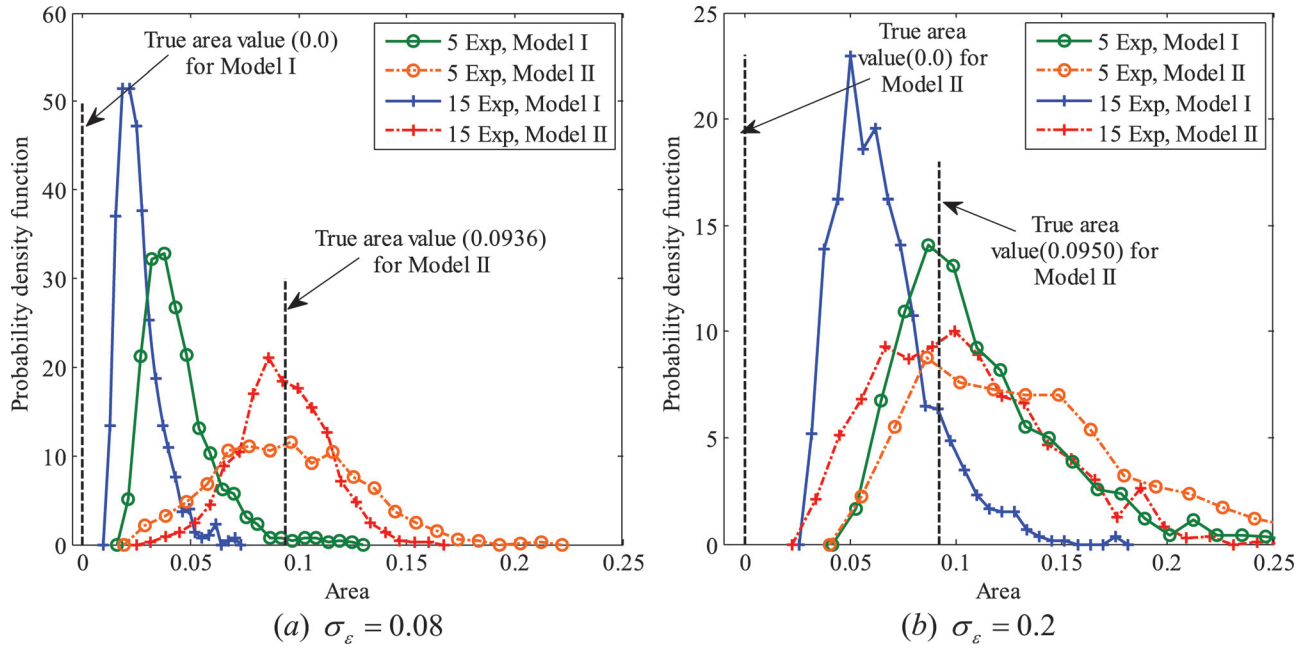


Fig. 9 Distributions of area metrics at validation site $x = 6.0$ in test set 1

$x = 6.0$ is expected to be zero. Alternatively, for model II (an incorrect predictive model), the true area difference at $x = 6.0$ should be 0.0936 when the standard deviation of measurement error is $\sigma_\varepsilon = 0.08$. The distribution of the resulting area metric from running 1000 data sets of physical observations with several different sizes of physical experiments are plotted in Fig. 9(*a*). It is observed that even after increasing the number of physical observations from $N = 5$ to $N = 15$, the area difference between model I and physical observations is greater than zero. The area metric will never approach 0 (overestimate the true discrepancy), even for model I, because the empirical CDF will never be exactly the same as the true CDF. Therefore, there will always be a positive difference between the two CDF curves. The area metrics for model II are distributed around the true value but with large deviations.

If the amount of physical observations is not sufficient, the use of empirical distributions for the physical observations may result in underestimation or overestimation of the area metric. The potential risk resulting from underestimating or overestimating the area metric is examined by evaluating the chance (percentage) of the area metric for model I to be greater than the area metric for model II with the same experimental data set. In the case of smaller measurement error of $\sigma_\varepsilon = 0.08$, the resulting percentages are 30.8%, 9.9%, and 1.1% corresponding to $N = 1$, 5, and 15, respectively. If the magnitude of σ_ε is 0.2, the resulting probabilities increase to 42.0%, 30.8%, and 19.6%, respectively. Comparing Fig. 9(*a*) to Fig. 9(*b*), it is also noted that the larger measurement error, the more the area metric deviates from the true value. In conclusion, when there is a lack of sufficient physical observations or large measurement uncertainty, there exists a

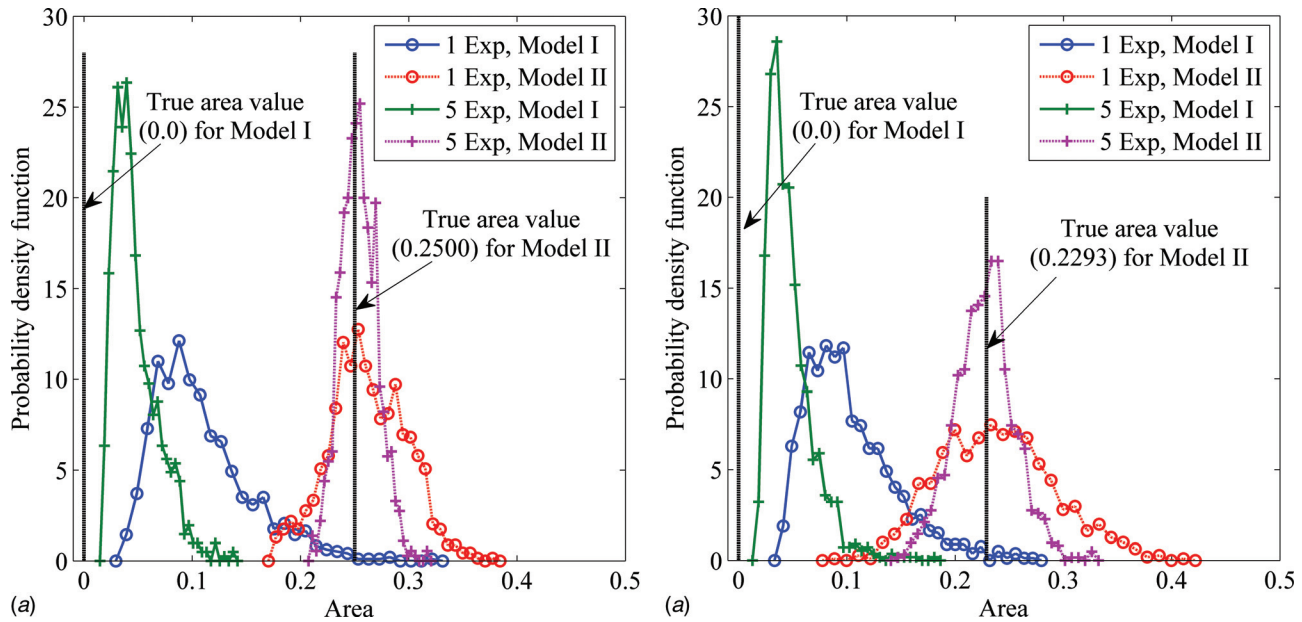


Fig. 10 Distributions of global metrics (u-pooling metrics) in test set 1

high risk of underestimation or overestimation of the true area metric.

Despite the above problem, an advantageous characteristic of the area metric is that it can pool all the sparse physical observations at multiple validation sites into a global area metric, the so-called u-pooling metric [8,43]. Figure 10 plots the distributions of the global metric for the predictive models. The physical observation data was collected at multiple validation sites ($x = 0.0, 1.0, \ldots, 8.0$) with $N = 1$ and 5 observations at each site. As with the previously discussed area metric, the u-pooling metric also contains a risk of underestimation or overestimation. The distributions can be seen in Fig. 10 to be moving toward the true values as the number of physical observations increases, while the range of variation becomes smaller.

The potential risk of identifying model I as a worse model than model II is examined by computing the percentages when the u-pooling metric of model I is larger than that of model II. In the case of smaller measurement error of $\sigma_\varepsilon = 0.08$, the risk is only 1.0% and 0.0% corresponding to the physical sample sizes $N = 1$ and 5, respectively, while the risk is 3.6% and 0.0% for the case $\sigma_\varepsilon = 0.20$. It is interesting to note that the results from using the u-pooling metric over a prediction domain are less influenced by the sample size compared to using the area metric at given settings of controllable variables. This is because the u-pooling method pools all the physical observations at multiple validation sites into one metric. The method provides more evidence to evaluate the model accuracy and therefore lowers the risk due to the lack of sufficient data uncertainty.

- *Observations from test set 2*. A similar study is performed on test set 2. The results at validation site $x = 6.0$ are provided in Fig. 11. It is noted that the distribution of the area metric of the predictive model III*a*, which has a smaller uncertainty of the model parameter, is closer to zero compared to model III*b*. This indicates that the area metric can differentiate between models containing greater and lesser amounts of uncertainty.
- *Discussion*. The area metric possesses several desirable features for validation metrics as described in Sec. 2.1. These *advantages* are listed as follows:

(1) The method directly uses the area between the distributions of predictions and physical observations as an objective metric to quantify the agreement. Therefore, widening the distribution of a predictive model will not increase the chance of accepting an incorrect model, because a wider distribution will most likely lead to a greater discrepancy.

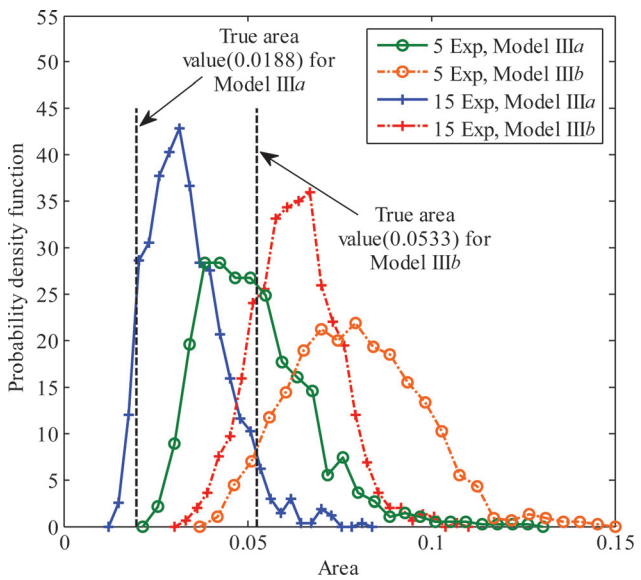


Fig. 11 Distributions of area metrics at validation site $x = 6.0$ for test set 2

(2) The area metric generalizes to a deterministic comparison when uncertainty goes to zero and degenerates the comparison of two stochastic quantities to a comparison between two scalar values from predictions and physical observations.
(3) One distinctive advantage of the method is its capability of integrating sparse physical observations at multiple validation sites to assess the global predictive capability over a specified domain of interest.

Nevertheless, a few *disadvantages* of the area metric are noted as follows:

(1) When there is a lack of sufficient experimental data, using the empirical distribution derived from physical observations may lead to underestimation or overestimation. Thus, the existing area metric does not provide a confidence level of the metric due to the lack of sufficient data.
(2) The existing area metric only compares marginal distributions; therefore, it is better suited for single response or uncorrelated multiple responses. One possible extension for multiple correlated responses is to compare the joint CDFs from physical observations and model predictions, a topic worth further investigation.

4 Summary of Observations and Remarks

Based on the desired features of model validation metrics presented in Sec. 2.1 and the observations from the tests in Sec. 3, the main characteristics of each validation metric tested are summarized and compared in Table 6. These findings are expected to provide a general guideline for selecting a proper validation metric or a combination of several metrics for practical applications when measuring the accuracy of a predictive model is the end goal.

In summary, both classical hypothesis testing and Bayes factor belong to the category of hypothesis testing methods. The null hypothesis is usually defined as the predictive model is accurate, but the alternative hypothesis is often difficult to determine. In classical hypothesis testing, since the alternative hypothesis is defined as any model different from the predictive model, type II error cannot be assessed. Using Bayes factor, a prior validity probability for all possible alternative populations where the experimental data might be from is provided by analysts. The method allows us to compute the type I and type II errors in a Bayesian framework. Additionally, if the analysts/designers have some knowledge about the possible alternatives, the Bayes factor method is an effective method to incorporate such information. If the analysts do not have prior knowledge, then Bayes factor can be misleading because it is sensitive to the prior. Overall, hypothesis testing will only answer a yes or no question when assessing whether the predictive model is accurate or not at certain validation sites. The method cannot provide insight into how accurate a predictive model is compared to the true physical system. Although Bayes factor can provide an assessment of the global predictive capability by multiplying Bayes factor values at multiple locations, the product does not have a clear physical meaning.

Instead of making a yes or no conclusion, the frequentist's metric and the area metric quantify the agreement of predictions and observations by measuring the distance between the means or the entire distributions of the two sets of populations. We refer to these validation metrics as distance-based methods. Distance-based metrics have an intuitive physical meaning that allows analysts to integrate the results at multiple validation sites into a global metric to assess the global predictive capability of a model over a region of interest, which is critical for using predictive models in engineering design and optimization [10–13]. These metrics can be used in model selection by comparing the accuracy of different models based on the distance measurement. The distance measurement can also be useful for quantifying the model error to update the model prediction when ignoring the systematic

Table 6 Summary of the main characteristics of the validation metrics

	Classical hypothesis	Bayes factor	Frequentist's metric	Area metric
Quantitative measure	Yes	Yes	Yes	Yes
Objective measure	Yes	No	Yes	Yes
Excludes any belief and criterion of accepting model	No	No	Yes	Yes
Includes all uncertainty sources	Yes	Yes	No	Yes
Feasible for multivariate case	Yes	Yes	No	No, but can be extended
Generalizes deterministic comparisons	No	No	Yes	Yes
Considers confidence level associated with amount of experimental data	Yes	Yes	Yes	No
Model not improved by widening the distribution of model parameter	Yes, if experimental data is more than one; otherwise No	Yes	No	Yes
Assesses global predictive capability	No	Yes, but needs improvement	Yes	Yes

uncertainty that exists in the experimental input conditions and/or response measurements. Further, the criterion of accepting a model or not is completely separated out from the metric itself and should be determined by the allowable error based on practical needs.

In practice, when only the central trend is being evaluated, the frequentist's method is sufficient. When there is a lack of sufficient experimental data, the associated adequacy (confidence level) of the frequentist's metric is quantified by the confidence bound, and the confidence bound will narrow when more experimental data is added. Since the area metric measures the discrepancy of the distributions from the predictive model and experiments, the metric is extremely useful when the dispersion or variance of a response is of importance in validation. On the other hand, the area metric is unable to quantify the uncertainty due to a lack of sufficient data, for which the analyst needs to consider the potential risk of overestimation and underestimation as observed in the test problems.

Additionally, as observed from test set 1, a larger measurement error would result in either a lower chance to reject the inaccurate model when using hypothesis testing-based metrics or a greater variation when using the distance-based metrics. Therefore, the measurement error should be eliminated as much as possible before conducting model validation.

5 Closure

Model validation metrics are important to select the best model from several different candidates and to facilitate the model updating process in a practical validation application. In this paper, by combining the selected views from the literature and our own views, we highlight a set of desired features that model validation metrics should possess. Four popular stochastic validation metrics have been reviewed and examined against the desired features using carefully designed numerical examples. A summary of comparison is provided in Sec. 4 and the findings can be used as a general guideline for choosing the appropriate validation metric or metrics given by the application.

We believe the hypothesis testing-based validation metric is only suitable for reaching a Boolean conclusion (yes or no) in model validation, whereas the distance-based method is not only feasible for model selection but can also be used to quantify the model error to update the model prediction in practical validation activities. Global predictive capability of a model can be easily obtained by integrating the values of distance-based metrics at multiple validation sites. However, due to a lack of sufficient data, the distance-based metrics may underestimate or overestimate the real discrepancy. To avoid this, the confidence bound of the distance-based metrics needs to be quantified.

It should be noted that the focus of this paper is on examining the metrics for assessing the *agreement* (*accuracy*) between a predictive model and physical observations, but not on the best metric for assessing *model validity*. In other words, our emphasis is on validating the prediction for a particular quantity of interest rather than on validating the model itself. Beyond accuracy, other important aspects, such as whether a model captures the general trend of performance, may be more important to consider in certain applications. The subject of "model validity" is broader and deeper than what is currently covered in this paper. Additionally, this work is focused on the probabilistic (stochastic) representation of uncertainty and does not intend to compare the existing metrics that introduce other types of uncertainty representations, e.g., a mixture of probabilistic and interval representations for modeling both aleatory and epistemic uncertainties. Besides, this work is focused on using a metric as a measure of the accuracy of a predictive model compared to physical observations, and not on how to set the bound of an accuracy requirement in model acceptance. In this later task, it is important to take into account the affordable and achievable experimental resolution uncertainty as it sets a lower bound. In addition, one should keep in mind that a validation effort should not be constrained by the available data, i.e., better observational data may need to be sought to achieve better validation in some applications. Finally, it should be noted that the systematic uncertainty that exists in the experimental input conditions and/or response measurements is not considered in the current study.

Acknowledgment

The grant supports from the U.S. National Science Foundation (CMMI-0928320) and the China Scholarship Council are greatly acknowledged. The views expressed are those of the authors and do not necessarily reflect the views of the sponsors.

Nomenclature

- CDF = cumulative density function
- PDF = probability density function
- $y^m(x)$ = predictive model
- $y_l^e(x_i)$ = the lth repetitive physical observation at validation sample site x_i
- H_0 = null hypothesis
- H_1 = alterative hypothesis
- $f_{x_i}^m(\cdot)$ = PDF of predictions at validation sample site x_i
- $F_{x_i}^m(\cdot)$ = CDF of predictions at validation sample site x_i
- $\mu_{x_i}^m$ = mean of predictions at validation sample site x_i
- $\sigma_{x_i}^m$ = standard deviation of predictions at validation sample site x_i
- $f_{x_i}^e(\cdot)$ = PDF of physical observations at validation sample site x_i
- $F_{x_i}^e(\cdot)$ = CDF of physical observations at validation sample site x_i
- $\mu_{x_i}^e$ = mean of physical observations at validation sample site x_i
- $\sigma_{x_i}^e$ = standard deviation of physical observations at validation sample site x_i

$\bar{y}^e(x_i)$ = estimated mean of repetitive physical observations at validation sample site x_i

ε_e = measurement error

σ_ε = standard deviation of measurement error

References

[1] Oberkampf, W. L., Trucano, T. G., and Hirsch, C., 2004, "Verification, Validation, and Predictive Capability in Computational Engineering and Physics," Appl. Mech. Rev., **57**(3), pp. 345–384.

[2] Sornette, D., Davis, A. B., Ide, K., Vixie, K. R., Pisarenko, V., and Kamm, J. R., 2007, "Algorithm for Model Validation: Theory and Applications," Proc. Natl. Acad. Sci. U.S.A., **104**(16), pp. 6562–6567.

[3] Ang, J. A., Trucano, T. G., and Luginbuhl, D. R., 1998, "Confidence in ASCI Scientific Simulations," Sandia National Laboratories, Report No. SAND98–1525C.

[4] Defense Modeling and Simulation Office, 1996, *DoD Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide*, Office of the Director of Defense Research and Engineering.

[5] AIAA, 1998, "Guide for the Verification and Validation of Computational Fluid Dynamics Simulations," AIAA Paper No. G–077–1998.

[6] ASME, 2006, "Guide for Verification and Validation in Computational Solid Mechanics" ASME, PTC 60/V&V 10.

[7] Chen, W., Xiong, Y., Tsui, K. L., and Wang, S. A., 2008, "Design-Driven Validation Approach Using Bayesian Prediction Models," ASME J. Mech. Des., **130**(2), p. 021101.

[8] Xiong, Y., Chen, W., Tsui, K. L., and Apley, D. W., 2009, "A Better Understanding of Model Updating Strategies in Validating Engineering Models," Comput. Methods Appl. Mech. Eng., **198**(15–16), pp. 1327–1337.

[9] Messer, M., Panchal, J. H., Krishnamurthy, V., Klein, B., Yoder, P. D., Allen, J. K., and Mistree, F., 2010, "Model Selection Under Limited Information Using a Value-of-Information-Based Indicator," ASME J. Mech. Des., **132**(12), p. 121008.

[10] Malak, R. J., and Paredis, C. J. J., 2010, "Using Support Vector Machines to Formalize the Valid Input Domain of Predictive Models in Systems Design Problems," ASME J. Mech. Des., **132**(10), p. 101001.

[11] Shan, S., and Wang, G. G., 2010, "Metamodeling for High Dimensional Simulation-Based Design Problem," ASME J. Mech. Des., **132**(5), p. 051009.

[12] Apley, D. W., Liu, J., and Chen, W., 2006, "Understanding the Effect of Model Uncertainty in Robust Design with Computer Experiments," ASME J. Mech. Des., **128**(4), pp. 945–958.

[13] Shao, T., and Krishnamurty, S., 2008, "A Clustering-Based Surrogate Model Updating Approach to Simulation-Based Engineering Design," ASME J. Mech. Des., **130**(4), p. 041101.

[14] Mayer, D. G., and Butler, D. G., 1993 "Statistical Validation," Ecol. Modell., **68**(1–2), pp. 21–32.

[15] Hills, R. G., and Trucano, T. G., 1999, "Statistical Validation of Engineering and Scientific Models: Background," Sandia National Laboratories, Report No. SAND99–1256.

[16] Sugawara, Y., Shinohara, K., and Kobayashi, N., 2009, "Quantitative Validation of Dynamic Stiffening Represented by Absolute Nodal Coordinate Formulation," *Proceedings of the ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, DETC2009–86955.

[17] Oberkampf, W. L., and Trucano, T. G., 2008, "Verification and Validation Benchmarks," Nucl. Eng. Des., **238**(3), pp. 716–743.

[18] Kennedy, M. C., and O'Hagan, A., 2001, "Bayesian Calibration of Computer Models," J. R. Stat. Soc. Ser.B, **63**(3), pp. 425–464.

[19] Helton, J. C., 1994, "Treatment of Uncertainty in Performance Assessments for Complex Systems," Risk Anal., **14**(4), pp. 483–511.

[20] Helton, J. C., Johnson, J. D., and Oberkampf, W. L., 2004, "An Exploration of Alternative Approaches to the Representation of Uncertainty in Model Predictions," Reliab. Eng. Syst. Saf., **85**(1–3), pp. 39–71.

[21] Rebba, R., and Mahadevan, S., 2006, "Model Predictive Capability Assessment Under Uncertainty," AIAA J., **44**(10), pp. 2376–2384.

[22] Oberkampf, W. L., and Barone, M. F., 2006, "Measures of Agreement Between Computation and Experiment: Validation Metrics," J. Comput. Phys., **217**(1), pp. 5–36.

[23] Romero, V. J., Luketa, A., and Sherman, M., 2010, "Application of a Versatile 'Real Space' Validation Methodology to a Fire Model," AIAA J. Thermophys. Heat Transfer, **24**(4), pp. 730–744.

[24] Sprague, M. A., and Geers, T. L., 2003, "Spectral Elements and Field Separation for an Acoustic Fluid Subject to Cavitation," J. Comput. Phys., **184**(1), pp. 149–162.

[25] Russell, D. M., 1997, "Error Measures for Comparing Transient Data: Part I, Development of a Comprehensive Error Measure," *Proceedings of the 68th Shock and Vibration Symposium*, pp. 175–184.

[26] Russell, D. M., 1997, "Error Measures for Comparing Transient Data: Part II, Error Measures Case Study," *Proceedings of the 68th Shock and Vibration Symposium*, pp. 185–198.

[27] Schwer, L. E., 2007, "Validation Metrics for Response Histories: Perspectives and Case Studies," Eng. Comput., **23**(4), pp. 295–309.

[28] Romero, V. J., Luketa, A., and Sherman, M., 2009, "Application of a Pragmatic Interval Based 'Real Space' Approach to Fire CFD Model Validation Involving Aleatory and Epistemic Uncertainty," *Proceedings of the 50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, AIAA Paper No. 2009–2279.

[29] Ferson, S., and Oberkampf, W. L., 2009, "Validation of Imprecise Probability Models," Int. J. Reliab. Saf., **3**(1–3), pp. 3–22.

[30] Chen, W., Baghdasaryan, L., Buranathiti, T., and Cao, J., 2004, "Model Validation via Uncertainty Propagation and Data Transformations," AIAA J., **42**(7), pp. 1406–1415.

[31] Buranathiti, T., Cao, J., Chen, W., Baghdasaryan, L., and Xia, Z. C., 2006, "Approaches for Model Validation: Methodology and Illustration on a Sheet Metal Flanging Process," ASME J. Manuf. Sci. Eng., **128**(2), pp. 588–597.

[32] Mahadevan, S., and Rebba, R., 2005, "Validation of Reliability Computational Models Using Bayes Networks," Reliab. Eng. Syst. Saf., **87**(1), pp. 223–232.

[33] Lew, J. S., 2008, "Model Validation Using Coordinate Distance with Performance Sensitivity," Mathematical Problems in Engineering, vol. 2008, 298146, p. 8

[34] Zhang, R., and Mahadevan, S., 2003 "Bayesian Methodology for Reliability Model Acceptance," Reliab. Eng. Syst. Saf., **80**(1), pp. 95–103.

[35] Hills, R. G., 2006, "Model Validation: Model Parameter and Measurement Uncertainty," ASME J. Heat Transfer, **128**(4), pp. 339–351.

[36] Yang, R. J., Barbat, S., and Weerappuli, P., 2009, "Bayesian Probabilistic PCA Approach for Model Validation of Dynamic Systems," *Proceedings of the SAE World Congress*, 2009–01–1404.

[37] Dowding, K. J., Pilch, M., and Hills, R. G., 2008, "Formulation of the Thermal Problem," Comput. Methods Appl. Mech. Eng., **197**(29–32), pp. 2385–2389.

[38] Sarin, H., Kokkolaras, M., Hulbert, G., Papalambros, P., Barbat, S., and Yang, R. J., 2009, "A Comprehensive Metric for Comparing Time Histories in Validation of Simulation Models with Emphasis on Vehicle Safety Applications," *Proceedings of the ASME 2008 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, DETC2008–49669.

[39] Red-Horse, J. R., and Paez, T. L., 2008, "Sandia National Laboratories Validation Workshop: Structural Dynamics Application," Comput. Methods Appl. Mech. Eng., **197**(29–32), pp. 2578–2584.

[40] Yang, R. J., Li, G., and Fu, Y., 2007, "Development of Validation Metrics for Vehicle Frontal Impact Simulation," *Proceedings of ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, DETC2007–34124.

[41] Rebba, R., and Mahadevan, S., 2006, "Validation of Models with Multivariate Output," Reliab. Eng. Syst. Saf., **91**(8), pp. 861–871.

[42] Jiang, X., and Mahadevan, S., 2008, "Bayesian Validation Assessment of Multivariate Computational Models," J. Appl. Stat., **35**(1), pp. 49–65.

[43] 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**(29–32), pp. 2408–2430.

[44] Rebba, R., Huang, S., Liu, Y., and Mahadevan, S., 2006, "Statistical Validation of Simulation Models," Int. J. Mater. Prod. Technol., **25**(1–3), pp. 164–181.

[45] Hills, R. G., and Trucano, T. G., 2002, "Statistical Validation of Engineering and Scientific Models: A Maximum Likelihood Based Metric," Sandia National Laboratories, Report No. SAND2001–1783.

[46] Rebba, R., Mahadevan, S., and Huang, S., 2006, "Validation and Error Estimation of Computational Models," Reliab. Eng. Syst. Saf., **91**(10–11), pp. 1390–1397.

[47] Rebba, R., and Mahadevan, S., 2008, "Computational Methods for Model Reliability Assessment," Reliab. Eng. Syst. Saf., **93**(8), pp. 1197–1207.

[48] Haldar, A., and Mahadevan, S., 2000, *Probability, Reliability, and Statistical Methods in Engineering Design*, Wiley, New York.

[49] Ghanem, R. G., Doostan, A., and Red-Horse, J., 2008, "A Probabilistic Construction of Model Validation," Comput. Methods Appl. Mech. Eng., **197**(29–32), pp. 2585–2595.

[50] Kass, R. E., and Raftery, A., 1995, "Bayesian Factors," J. Am. Stat. Assoc., **90**(430), pp. 773–795.

[51] Oberkampf, W. L., and Trucano, T. G., 2002, "Verification and Validation in Computational Fluid Dynamics," Prog. Aerosp. Sci., **38**(2), pp. 209–272.

[52] Loehle, C., 1997, "A Hypothesis Testing Framework for Evaluating Ecosystem Model Performance," Ecol. Modell., **97**(3), pp. 153–165.

[53] Robinson, A. P., and Froese, R. E., 2004, "Model Validation Using Equivalence Tests," Ecol. Modell., **176**(3–4),pp. 349–358.

[54] Wellek, S., 2003, *Testing Statistical Hypotheses of Equivalence*, Chapman and Hall, London.

[55] Balci, O., and Sargent, R. G., 1981, "A Methodology for Cost-Risk Analysis in the Statistical Validation of Simulation Models," Commun. ACM, **24**(4), pp. 190–197.

[56] Trucano, T. G., Swiler, L. P., Igusa, T., Oberkampf, W. L., and Pilch, M., 2006, "Calibration, Validation, and Sensitivity Analysis: What's What," Reliab. Eng. Syst. Saf., **91**(10–11), pp. 1331–1357.