Reliability Engineering and System Safety 147 (2016) 49–59

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Separation of aleatory and epistemic uncertainty in probabilistic model validation

Joshua Mullins [a], You Ling [a], Sankaran Mahadevan [a,*], Lin Sun [b], Alejandro Strachan [b]

[a] Department of Civil and Environmental Engineering, Vanderbilt University, TN 37235, United States
[b] School of Materials Engineering, Purdue University, IN 47907, United States

ARTICLE INFO

Article history:
Received 23 August 2014
Received in revised form
11 July 2015
Accepted 5 October 2015
Available online 19 October 2015

Keywords:
Model validation
Validation metrics
Imprecise data
Reliability
Epistemic uncertainty

ABSTRACT

This paper investigates model validation under a variety of different data scenarios and clarifies how different validation metrics may be appropriate for different scenarios. In the presence of multiple uncertainty sources, model validation metrics that compare the distributions of model prediction and observation are considered. Both ensemble validation and point-by-point approaches are discussed, and it is shown how applying the model reliability metric point-by-point enables the separation of contributions from aleatory and epistemic uncertainty sources. After individual validation assessments are made at different input conditions, it may be desirable to obtain an overall measure of model validity across the entire domain. This paper proposes an integration approach that assigns weights to the validation results according to the relevance of each validation test condition to the overall intended use of the model in prediction. Since uncertainty propagation for probabilistic validation is often unaffordable for complex computational models, surrogate models are often used; this paper proposes an approach to account for the additional uncertainty introduced in validation by the uncertain fit of the surrogate model. The proposed methods are demonstrated with a microelectromechanical system (MEMS) example.

1. Introduction

Engineering systems are described by computational models in order to predict the behavior of the system under scenarios that cannot be tested with an experiment (usually due to economic constraints). Most models have parameters that cannot be directly measured, and these parameters must be inferred based on experimental data of relevant inputs and outputs, in a process known as model calibration [1–4] before the model has any predictive capability. However, the calibrated model should not be trusted for prediction without evidence that it is a good representation of reality in other input scenarios, both in terms of the inferred parameters and the underlying form of the model. This evidence should come from experiments conducted within a regime of testable input conditions that is different from those used in the calibration phase. The new experimental data is used for model validation, which can be defined as the process of assessing the adequacy of a computational model for an intended prediction application. The prediction of the calibrated model is stochastic if parameter uncertainty and/or model errors are considered, and the experimental observation is affected by measurement uncertainty. As a result, several recent quantitative validation methods are based on the comparison of probability distributions for prediction and observation. The construction of the model output distribution under parameter uncertainty requires multiple Monte Carlo runs of the model, and this may be computationally unaffordable if the model is expensive. Therefore, surrogate models are often used to construct the model output distribution, which leads to additional uncertainty about the validation result.

Validation methods that have been developed in the literature include classical hypothesis testing [5–7], Bayesian hypothesis testing [8–11], the area metric [12–14], and the model reliability metric [15,16]. The connections between these various metrics as well as their strengths and weaknesses have also been explored [17,18]. Each of these existing approaches assesses the agreement between model prediction and validation observation, but they differ in how they are applied. One view, as is usually taken with the area metric, is to look at the set of validation observations collectively and compare the distribution of the prediction over the entire input domain against the distribution of the observation data. When the input and corresponding output are measured for each validation experiment (with corresponding stochastic predictions for each input), a synthesis across the domain is accomplished via the "u-pooling" approach [12]. An alternate view, as

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

http://dx.doi.org/10.1016/j.ress.2015.10.003
0951-8320/

taken with the model reliability metric, is to perform a series of point comparisons, one for each validation input condition, and assess the predictive capability of the model as a function of the location in the input domain. Both classical and Bayesian hypothesis testing may be cast in a way that is consistent with either of these views by choosing different hypotheses. The proper interpretation depends largely on the type of data that is available to the analyst. This paper investigates different validation scenarios where one of these two views (ensemble validation vs. point-by-point validation) is more suitable.

A further distinction between these methods is in the interpretation of the results. Conventionally, model validation has resulted in a single positive or negative result that indicates whether or not the model should be used in prediction. By choosing thresholds for the quantitative results, any of the previously mentioned methods could be interpreted in this manner. Alternatively, Bayesian hypothesis testing and the model reliability metric enable the result to be interpreted as a probability of agreement between prediction and observation. Thus, the result is not a single pass/fail decision, but a degree of validity. This paper focuses primarily on these probabilistic approaches because they enable other ongoing research efforts that are aimed at including the validation result in the subsequent prediction of a quantity of interest in the usage condition [19–21].

An important aspect of this discussion is the distinction between aleatory and epistemic uncertainty sources. Aleatory uncertainty is the natural variation of inputs that impact outputs of interest. This uncertainty is irreducible, and it is commonly treated with probability theory. Epistemic uncertainty results from the lack of knowledge about the system of interest, and it can be further separated into model uncertainty (e.g. parameter uncertainty, solution approximation errors, and model form uncertainty) and data uncertainty (e.g. measurement uncertainty and sparse or imprecise data). These two types of epistemic uncertainty are inherently coupled since data is used to improve computational models (by inferring the model parameters or altering the model form). Since it stems from the lack of knowledge, epistemic uncertainty can be reduced by obtaining additional information. In the literature, this uncertainty has been modeled in a number of different ways, including Bayesian probability [22], interval analysis [23], evidence theory [24], possibility theory [25], fuzzy logic [26], and generalized information theory [27]. Regardless of the approach to epistemic uncertainty characterization, researchers have become increasingly aware of the importance of separating aleatory and epistemic uncertainty sources [28–30]. Aleatory uncertainty is unavoidable and must be accounted for in prediction models; however, it is not directly pertinent to decisions about risk and uncertainty reduction because its contribution cannot be eliminated. On the other hand, information about epistemic uncertainty directly supports decisions about data collection and model improvement. Therefore, the focus of this paper is the impact of epistemic uncertainty on model validation and how to separate the contributions of aleatory and epistemic uncertainty when the available data permits.

Within this context, this paper aims to address three issues that impact the validation assessment: (1) the type of input–output measurements that are made in validation experiments, (2) the "proximity" of the validation tests to the prediction regime of interest, and (3) the use of surrogate models for uncertainty propagation. The first issue is addressed in Section 3 where three different types of validation data scenarios are explored, and appropriate validation approaches are identified. The second issue is addressed in Section 4 where a method for weighting validation results by the relevance to the prediction is proposed. The third issue is addressed in Section 5 where Gaussian process surrogate modeling is discussed and the effect of the added uncertainty on the validation result is quantified. The proposed methods are demonstrated with numerical examples relating to a microelectromechanical system (MEMS) device in Section 6, and the paper is concluded in Section 7.

2. Comparing uncertain predictions and observations

In the presence of multiple uncertainty sources, both the prediction of the quantity of interest and the observation of the same quantity are uncertain. In a probabilistic framework, they are treated as stochastic variables that are described by probability distributions. These distributions may be compared by comparing the moments, the shapes, or the samples of the distributions. In the area metric [12–14] and Kullback–Leibler (K–L) divergence [31] approaches, the shapes of the distributions themselves are compared directly. In the model reliability metric [15,16] approach, the distance between sampled realizations of prediction and observation is evaluated. Hypothesis testing methods e.–>(i.e., classical hypothesis testing [5–7] and Bayesian hypothesis testing [8–11]) may be cast in different ways by choosing different hypotheses (e.g. equality of moments or distribution parameters, equality of prediction and observation samples, or allowable distance between prediction and observation samples). A key factor in the choice of comparison is the stochastic dependence between the prediction and observation. As noted in [13], samples cannot be uniquely generated without some knowledge of the dependence, so it is only possible to compare samples if the dependence information e.–>(i.e., the correlation structure) is known. In such a scenario, a comparison of sampled differences can make a stronger statement about the agreement between prediction and observation. For example, positive correlation between prediction and observation may suggest better predictive capability than negative correlation. In Section 3, we discuss how the separation of uncertainty sources in point-by-point validation enables dependence information to be isolated, such that independent samples can be drawn. However, this separation may not always be possible, and when no such dependence information is known, a shape-based comparison can be performed in order to bypass this requirement. The result can then be bounded for different possible dependence structures [13].

In this paper we focus on the area metric and the model reliability metric comparison approaches; these two methods are described in detail (the area metric in Section 2.1 and the model reliability metric in Section 2.2). The applicability of these approaches depends on the type of information that is available to the analyst.

2.1. Area validation metric

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

$$d(F_{Y_m}, S_{Y_D}) = \int_{-\infty}^{+\infty} |F_{Y_m}(y) - S_{Y_D}(y)|\, dy \tag{1}$$

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 model reliability metric, 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 data from different input conditions by using the so-called "u-pooling" procedure (transformation from physical space to probability space) to validate models with sparse data on multiple experimental combinations [17]. 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 [32], 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]

$$d(F_u, S_u) = \int_0^1 |F_u - S_u| \, du \tag{2}$$

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 defined in u-space based on Eq. (2) 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

$$d(F_y, G_y) = \int |F_y - G_y| \, dy \tag{3}$$

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.

2.2. Model reliability metric

The model reliability metric r proposed by Rebba and Mahadevan [15] is a direct measure of model prediction quality and is relatively easy to compute. It is defined as the probability of the difference (Δ) between observed data (Y_D) and model prediction (Y_m) being less than a given tolerance limit ϵ:

$$r = \Pr(-\epsilon < \Delta < \epsilon), \quad \Delta = Y_D - Y_m. \tag{4}$$

In Eq. (4), experimental observation is treated as a random variable due to measurement error, and under the Bayesian perspective, model output is assigned a probability distribution accounting for the combined effect of epistemic and aleatory uncertainty. Note that upper case variables (e.g. Y_m and Y_D) denote random variables whereas lower case variables (e.g. y_D and y_m) denote particular samples of random variables as in Section 2.1. As the difference between two random variables, Δ is treated as a random variable, and the probability distribution of Δ can be obtained from the probability distributions of Y_D and Y_m. Then, the model reliability metric, as defined in Eq. (4), is computed by the integration of the distribution of Δ.

$$r = \int_{-\epsilon}^{\epsilon} f_\Delta(\omega) \, d\omega = F_\Delta(\epsilon) - F_\Delta(-\epsilon) \tag{5}$$

For instance, if the model prediction, $Y_m \sim N(\mu_m, \sigma_m^2)$, and the corresponding observation, $Y_D \sim N(\mu, \sigma^2_{Y_D})$, are independent, the distribution of the difference can be computed analytically, $\Delta \sim N(\mu - \mu_m, \sigma^2_{Y_D} + \sigma_m^2)$. For the sake of simplicity, let $\sigma_\Delta = \sqrt{\sigma^2_{Y_D} + \sigma_m^2}$. In this case, the reliability-based metric r can be computed by evaluating the standard normal CDF Φ as

$$r = \Phi\left[\frac{\epsilon - (\mu - \mu_m)}{\sigma_\Delta}\right] - \Phi\left[\frac{-\epsilon - (\mu - \mu_m)}{\sigma_\Delta}\right] \tag{6}$$

Note that observations are typically limited, and the distribution of Y_D is not precisely known. Therefore, the integration in Eq. (5) is rarely performed analytically. Instead, the metric is estimated via sampling. Samples are drawn from the distribution of Y_m, and each sample is compared with a discrete sample from the set of available observations from Y_D (there may be any number). The value of r is the percentage of the sample pairs that yield differences smaller than ϵ. The precise value of ϵ is problem specific, and its magnitude is closely tied to the magnitude of Y. Whenever possible, it should be determined in communication with decision makers and domain experts, and the value of ϵ that is used should always be included along with r when reporting results.

Since the result of this computation is a probability, we refer to the model reliability approach as a probabilistic validation metric. Note that Bayesian hypothesis testing generally leads to a single scalar result known as the Bayes factor [8,33,34], but the Bayes factor may also be converted to a probability measure. Thus, the methods that are developed in this paper for probabilistic validation metrics are also applicable to Bayesian hypothesis testing although they are only illustrated for the model reliability metric.

3. Aleatory and epistemic uncertainty in model validation

3.1. Validation with fully, partially, or uncharacterized experimental data

Validation experimental data of course report observations on output quantities of interest, but three possible scenarios exist with respect to input measurements: (1) fully characterized (i.e., all the input variables of individual experiments and corresponding outputs are measured and reported as point values), (2) partially characterized (i.e., some inputs and/or outputs of individual experiments are not measured or are reported as intervals), or (3) uncharacterized (i.e., experiments are performed on multiple input combinations, but these input combinations are not measured or are reported as a single interval). In the cases of partially characterized or uncharacterized validation data, the input $\boldsymbol{X}$ is treated as a random vector due to the lack of measurements or the imprecision of the measurements. The reported intervals and expert opinion (if available) are needed to construct a probability distribution of $\boldsymbol{X}$. Note that in the Bayesian approach, the lack of knowledge (epistemic uncertainty) is represented through a probability distribution (subjective probability). This point is critical to the discussion that follows later in this section; the implication is that the "true" output of a single experiment is not a probability distribution, but a single value that cannot be precisely observed. Likewise, the corresponding model prediction would also be a single deterministic value for each experiment if all inputs and parameters were precisely known. Non-probabilistic approaches have also been proposed to handle the epistemic uncertainty; in this paper, we only focus on probabilistic methods.

For partially characterized validation data, input distributions are assigned to different experiments separately, and these distributions $f_{\boldsymbol{X_i}}(\boldsymbol{x})(i = 1, \ldots, n$ for n validation input conditions) represent input data uncertainty in each individual experiment.

Table 1
Three types of validation experiments and the corresponding input–output data.

Fully characterized	Input	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	…	$\boldsymbol{x}_n$
	Output	y_{D_1}	y_{D_2}	…	y_{D_n}
Partially characterized	Input	$f_{X_1}(\boldsymbol{x})$	$f_{X_2}(\boldsymbol{x})$	…	$f_{X_n}(\boldsymbol{x})$
	Output	y_{D_1}	y_{D_2}	…	y_{D_n}
Uncharacterized	Input		$f_X(\boldsymbol{x})$		
	Output	y_{D_1}	y_{D_2}	…	y_{D_n}

For example, suppose experiments were conducted at n different nominal load values, but each of the load values is only known up to an interval $[x_i-\epsilon, x_i+\epsilon]$. For uncharacterized validation data, a single distribution is assigned to the variable over multiple experiments, and this distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$ represents the uncertainty due to both natural variability and input data uncertainty. For example, suppose the same n experimental outputs are available; however, there is not a nominal load value for each individual experiment, but rather a single interval that encompasses the load values for all experiments $[x_L, x_U]$. Table 1 shows a typical format of input–output data collected from the three types of experiments. Although uncertainty exists in the inputs of partially characterized or uncharacterized experiments, the resulting data may still be considered for validation by practitioners if the input uncertainty is limited.

3.2. Ensemble vs. point-by-point validation

As mentioned in Section 1, there are two possible views of validation. The data can be viewed collectively and compared against the overall distribution of the model prediction across the input domain, or the data can be viewed individually and compared against a separate stochastic prediction at each input condition. If the validation assessment is performed only once over the collection of data (i.e., ensemble validation), it is difficult to separate the contributions of aleatory and epistemic uncertainty sources to the validation result. Once the model prediction has been corrected for solution approximation errors and/or calibrated for bias (often referred to as model discrepancy [1]), the distributions of both the prediction and observation are a result of aleatory uncertainty (input variations) and epistemic uncertainty (parameter uncertainty in the prediction and measurement uncertainty in the observation). There is no reasonable expectation that the epistemic uncertainty contributions to the total uncertainty in the prediction and observation should be similar to each other because the two sources are independent. In particular, parameter uncertainty is related to the quantity and quality of available calibration data. As more calibration data is collected, parameter uncertainty can be reduced via Bayesian updating. Since the validation data set should be separate from the calibration data in order to make a proper assessment of the model's predictive capability, the measurement uncertainty in the validation data is generally different from the calibration measurement uncertainty. Furthermore, even if the distributions of the measurement errors in the calibration data and the validation data are similar, there is still no reason to expect correlation between particular samples of measurement error. Therefore, the only uncertainty contribution that is common to both the prediction and observation is the aleatory uncertainty in the input.

In the collective view of validation, one option for separating the aleatory and epistemic contributions is the p-box approach [14]. In this treatment, epistemic uncertainty is expressed as an interval while aleatory uncertainty is expressed with probability distributions. Such a treatment is particularly suitable for uncharacterized data because the data quality does not enable point-by-point separation. However, when the dominant effect is epistemic uncertainty rather than aleatory uncertainty, comparing observations to a p-box may not be very informative since the epistemic uncertainty gives a wide window of acceptance for the model [13,14].

In many problems, the epistemic contributions are, in fact, large since economic constraints in realistic applications often lead to very sparse/imprecise data. For this reason, the model reliability approach is aimed at epistemic uncertainty in both the observation and the prediction. Note again that parameter uncertainty in this paper refers to the subjective probability description of a deterministic parameter value, not aleatory uncertainty. It is possible that parameters may also be affected by aleatory variability across experiments, but we address this issue by localizing calibration to particular experimental configurations. The parameter uncertainty is expressed by a subjective probability distribution separately for each test, and it is then reduced via Bayesian updating with replicate testing as seen in the example in Section 6. Aiming the assessment at epistemic uncertainty leads directly to decisions about what improvements are most necessary (either in the data or the model) in order to improve the predictive quality of the model.

Therefore, when information is available about the particular input condition associated with each data point (either fully or partially characterized data), we propose the use of individual comparisons at each location with the model reliability metric. The metric is computed for a stochastic prediction and an uncertain observation, but the metric is not maximized when the spreads in the two distribution are the same. This behavior occurs because the metric is not a shape-based comparison; it comes from sampling the distributions to compute the distribution of the difference $\boldsymbol{\Delta}$ (see Section 2). As mentioned in that discussion, the distribution of $\boldsymbol{\Delta}$ can only be obtained if the stochastic dependence between prediction and observation is known. However, as described in the opening of this section, the correlation between these two variables only occurs through aleatory uncertainty that is common to both, and the epistemic uncertainty sources are independent. Therefore, at a particular input condition, since the stochastic prediction and observation are only sampled over epistemic uncertainty sources, the samples are *conditionally independent*. Since $\boldsymbol{\Delta}$ is simply the distribution of bias between deterministic samples of prediction and observation, the maximum reliability metric occurs when the distributions of prediction and observation are unbiased from each other, and each has minimum uncertainty (see Fig. 1). This behavior agrees with our intuition about how to improve the result if the validation agreement is poor. By reducing measurement uncertainty *or* reducing parameter uncertainty, the validation result at each input can be improved. For the shapes of the distributions to agree, both the measurement uncertainty *and* the parameter uncertainty must be reduced in order to improve the agreement. It is an unnecessary requirement that the shapes agree since they are representing only independent epistemic uncertainty sources. Both collecting more calibration data (to reduce parameter uncertainty) and collecting more precise data (to reduce data uncertainty) should individually improve confidence in the model if the model is actually predicting well.

For these reasons, shape-based comparisons are not intended for purely epistemic uncertainty-based comparisons. They should not be used for this purpose because it is possible for the contributions of one or both of the uncertainty sources to increase and improve the comparison. For example, the point comparison shown in Fig. 1 poses two scenarios, both with the same stochastic prediction. Fig. 1(b) gives an idealized scenario where the observation data is “perfect” (i.e., no measurement uncertainty). In this scenario, clearly the shapes of the two distributions are not the same, and the distributions will actually match more closely

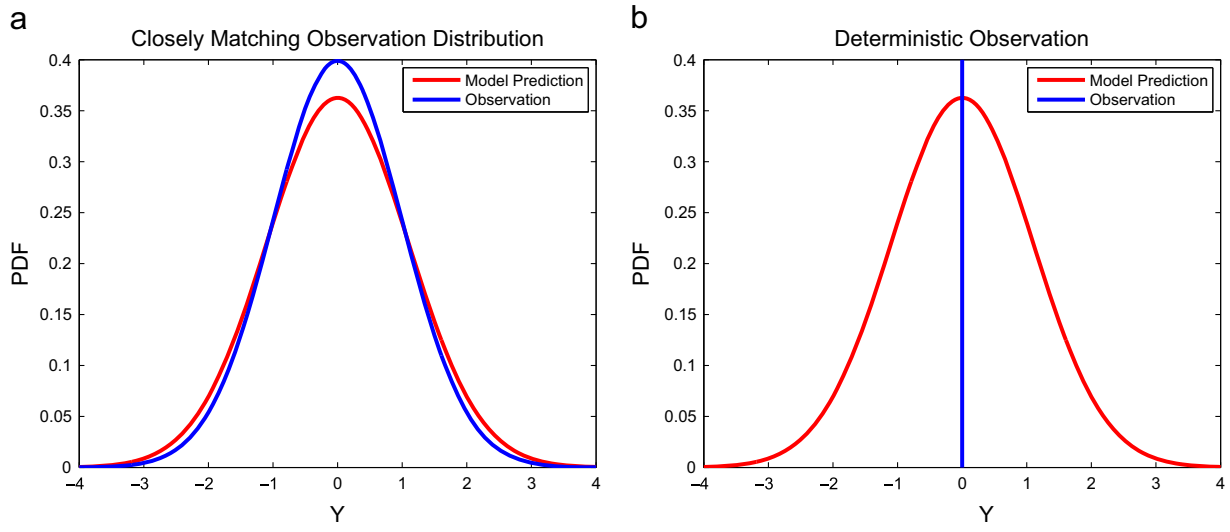

Fig. 1. Decreasing measurment uncertainty for the same stochastic prediction improves the confidence in the model if the observation is unbiased. (a) Measurement uncertainty and parameter uncertainty are similar (Model reliability =0.86). (b) Zero measurement uncertainty with the same parameter uncertainty as in 1(a) (Model reliability =0.95).

(improving a shape-based measure) by injecting more uncertainty into the observation as in Fig. 1(a). This result does not occur with the model reliability approach because the metric is lower for larger uncertainty in the observation (i.e., we have less confidence in the assessment because the observation data is not adequate). Since, at a single known input point (fully or partially characterized), the uncertainty sources are completely epistemic, both the prediction and observation would be deterministic values if no epistemic uncertainty existed. Thus, the scenario shown in Fig. 1 (b) (where ideal quality observation data is available) is actually preferable because there is a higher probability that the deterministic prediction and observation would agree if both were known precisely.

An additional advantage of point-by-point comparison is that it demonstrates the quality of the model as a function of input condition. This information may be very useful in determining whether the model will be appropriate in its intended use, and it may also help to isolate potential systematic errors arising from model form inadequacy. For example, if the model is consistently performing poorly for large values of some input (e.g. loading), this may be evidence that the model does not capture some higher order physical behavior (e.g. non-linearity) that is activated by extreme conditions. Additionally, if different values of model reliability are computed at different inputs, the weighting approach that is presented in Section 4 becomes possible, and preferences for particularly important regions of the input domain based on the intended application can be incorporated.

In summary, we conclude that ensemble validation is best suited for uncharacterized data scenarios, and point-by-point validation is preferable when information is known about the corresponding input conditions (partially or fully characterized validation data scenarios) for the following reasons: (1) distributions of prediction and observation can only be expected to agree when the dominant uncertainty source is aleatory variability that is common to both distributions; (2) point-by-point comparisons with the model reliability metric separate aleatory and epistemic uncertainty and penalize large epistemic uncertainty (from any source) by returning a lower validation result; (3) point-by-point comparisons allow systematic error trends to be isolated in the model; and (4) a set of point-by-point comparisons can be weighted based on relevance to the intended use of the model.

4. Integration of model validation results from multiple input conditions

By utilizing the model reliability approach, a value for the validation metric can be computed for each validation input condition. This information is itself useful for decision making about the model adequacy since developers can look at regions of the input domain that perform poorly in validation and investigate potential model improvements. However, the ultimate goal of the validation activity is to assess the current model's prediction capability, and recent research [19–21] has the additional goal of performing this assessment quantitatively so that it may be included in the prediction. In some applications, including the validation result in a prediction framework may require a single overall measure of the model quality across the entire domain of interest. This measure should be representative of the quality of the model in its intended application condition where the prediction will be made. Thus, the method given by Eq. (7) is proposed:

$$v_{overall} = \int v(\boldsymbol{x})\pi(\boldsymbol{x})\, d\boldsymbol{x}. \tag{7}$$

Here, $v(\boldsymbol{x})$ is the value of the validation metric at a particular point in the validation domain, represented by the n-dimensional input vector $\boldsymbol{x}$, and $\pi(\boldsymbol{x})$ is the n-dimensional joint probability density of the point $\boldsymbol{x}$ in the prediction domain. This distribution comes from the best available knowledge of the input conditions that will be encountered in the intended application of the model; the distribution may describe both aleatory and epistemic uncertainty. Effectively, the joint density becomes a weighting function for the importance of each validation result according to how likely that input condition is in the prediction scenario. In evaluating the integral in Eq. (7), note that $v(\boldsymbol{x})$ is only available at some discrete values of $\boldsymbol{x}$. Therefore, the integral may be approximated by a weighted sum taken over a set of m validation tests as

$$v_{overall} = \frac{\sum_{i=1}^{m} v(\boldsymbol{x_i})w_i}{\sum_{i=1}^{m} w_i}. \tag{8}$$

The computation of the weight w_i is straightforward for fully characterized validation data; it is obtained by computing $\pi(\boldsymbol{x_i})$, which is a single value for each validation experiment. When input measurement uncertainty exists, the validation data is considered to be partially characterized, and $\boldsymbol{X_i}$ is not a point value, but rather

a random variable. In this scenario, the weighting for the intended application can be obtained for each validation test by taking the expected value over the distribution of the corresponding input measurement uncertainty $f_{\boldsymbol{X}_i}(\boldsymbol{x})$:

$$w_i = \int \pi(\boldsymbol{x}) f_{\boldsymbol{X}_i}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{9}$$

Once all the weights are computed, Eq. (8) results in a single deterministic measure of the probabilistic performance of the model over the expected prediction domain. It is an approximation since the set of validation input conditions generally does not cover the full prediction domain of interest; therefore, the summation must be normalized in order to obtain a valid probability. In fact, in some cases the prediction scenario may be for values of $\boldsymbol{x}$ that are not close to the validation domain. In this situation, the validation input conditions fall in the tail of the distribution for the intended application, and $\pi(\boldsymbol{x_i})$ is small for all the validation points. This would imply that none of the validation experiments are in the regime that is most relevant to the intended application, and the prediction represents a significant extrapolation of the model. Such extrapolation scenarios can be dangerous applications of the model, but they are often unavoidable in practice. Additional conservatism is needed for this situation, and the analyst should be especially aware of any trends in the point-by-point validation results that suggest that model inadequacy will be magnified in the prediction regime. Ongoing research efforts are exploring quantitative methods of setting boundaries for the extrapolation of the model and applying additional conservatism to the extrapolation scenarios when they are practically necessary [19–21,35,14].

The proposed integration approach has been described for situations where a single probabilistic value can be obtained from the model reliability metric at each input condition. When additional epistemic uncertainties exist, the validation metric uncertainty can be described by a probability distribution at each validation input, and the overall metric will also be a probability distribution accounting for these additional uncertainties. One example is a stochastic model discrepancy term as in the Kennedy–O'Hagan approach to model calibration [1]. If the discrepancy term is used as a correction for the model prediction, different realizations of the stochastic discrepancy yield different validation results. As another example, when surrogate models are used to generate the distribution of the model output that is used in the validation assessment, different realizations of the surrogate model prediction also lead to different validation results. These additional uncertainty sources should also be accounted for; thus, the surrogate model scenario is explored in Section 5. Note that the mathematics of treating stochastic model discrepancy would follow similarly.

5. Validation with surrogate models

Probabilistic approaches to model validation, as described in Section 2, require the propagation of parameter uncertainty through the model at each validation input condition. This propagation is typically performed via Monte Carlo sampling, which requires a large number of model evaluations. When the computational model is expensive, it is often replaced by a surrogate model to improve the efficiency of the propagation. Ultimately, the goal of the validation assessment is to make a statement about the adequacy of the physics-based, original computational model and not the surrogate, since the former will be used for prediction. Since the surrogate model is not a perfect representation of the original computational model, additional uncertainty is added to the validation result. In this paper, we use Gaussian process (GP) surrogate models as described in Section 5.1 because they provide a natural way of quantifying the uncertainty due to the discrepancy between the surrogate and the original computational model. This uncertainty then propagates to uncertainty in the validation assessment as described in Section 5.2.

5.1. Gaussian process interpolation

Let $y = G(\boldsymbol{x})$ be the target function which we intend to replace with a surrogate model. GP interpolation (or kriging) assumes that the output y over the domain of input $\boldsymbol{x}$ is a Gaussian random process with a mean function $m(\boldsymbol{x})$ and a covariance function $k(\boldsymbol{x},\boldsymbol{x}')$, i.e.,

$$\begin{aligned} &\mathrm{E}[y\,|\,\boldsymbol{x}] = m(\boldsymbol{x}) \\ &\mathrm{Cov}(y,y') = k(\boldsymbol{x},\boldsymbol{x}'). \end{aligned} \tag{10}$$

Given a set of training data $\{\boldsymbol{X}_T,\boldsymbol{y}_T\}$ and the input $\boldsymbol{X}_p$ where prediction is desired, the conditional probability distribution of the output $\boldsymbol{y}_P$ follows a multivariate Gaussian distribution [36] as

$$\begin{aligned} &\boldsymbol{y}_P\,|\,\boldsymbol{X}_P,\boldsymbol{X}_T,\boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &\boldsymbol{\mu} = m(\boldsymbol{X}_P) + \boldsymbol{\Sigma}_{PT}\boldsymbol{\Sigma}_{TT}^{-1}\left(\boldsymbol{y}_T - m(\boldsymbol{X}_T)\right) \\ &\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{PP} - \boldsymbol{\Sigma}_{PT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{PT}^{T} \end{aligned} \tag{11}$$

where $\boldsymbol{\mu}$ is the mean vector of the prediction $\boldsymbol{y}_P$ conditioned on the training data, and $\boldsymbol{\Sigma}$ is the conditional covariance matrix of $\boldsymbol{y}_P$; $\boldsymbol{\Sigma_{PT}}$ is the covariance matrix between the prediction and the training data; $\boldsymbol{\Sigma_{TT}}$ is the covariance matrix of the training data; $\boldsymbol{\Sigma}_{PP}$ is the unconditional covariance matrix of $\boldsymbol{y}_P$.

In order to determine the multivariate Gaussian distribution in Eq. (11), we need to formulate the mean function $m(\boldsymbol{x})$ and the covariance function $k(\boldsymbol{x},\boldsymbol{x}')$. $m(\boldsymbol{x})$ is often formulated as a polynomial function of $\boldsymbol{x}$. The form of $k(\boldsymbol{x},\boldsymbol{x}')$ may be selected from a number of commonly used covariance functions based on the desired properties (order of continuity, stationary/non-stationary, isotropic/anisotropic) [36]. Note that the selected formulations of $m(\boldsymbol{x})$ and $k(\boldsymbol{x},\boldsymbol{x}')$ contain unknown coefficients. Let $\boldsymbol{\phi}$ denote the coefficients of $m(\boldsymbol{x})$, and $\boldsymbol{\varphi}$ denote the coefficients of $k(\boldsymbol{x},\boldsymbol{x}')$. These coefficients can be estimated after constructing their likelihood function, which is proportional to the probability density of training data conditioned on $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$, i.e.,

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\varphi}) \propto \pi(\boldsymbol{y}_T\,|\,\boldsymbol{X}_T,\boldsymbol{\phi},\boldsymbol{\varphi}) \tag{12}$$

Either Bayesian inference or maximum likelihood estimation (MLE) may be used to compute the parameters; here, we use MLE. Since $\boldsymbol{y}_T\,|\,\boldsymbol{X}_T,\boldsymbol{\phi},\boldsymbol{\varphi} \sim \mathcal{N}\left(m(\boldsymbol{X}_T),\boldsymbol{\Sigma}_{TT}\right)$, the conditional probability density function $\pi(\boldsymbol{y}_T\,|\,\boldsymbol{X}_T,\boldsymbol{\phi},\boldsymbol{\varphi})$ is a multivariate Gaussian PDF as

$$\begin{aligned} \pi(\boldsymbol{y}_T\,|\,\boldsymbol{X}_T,\boldsymbol{\phi},\boldsymbol{\varphi}) = (2\pi)^{-n/2}\,|\,\boldsymbol{\Sigma}_{TT}\,|^{-1/2} * \\ \exp\left(-\tfrac{1}{2}[\boldsymbol{y}_T - m(\boldsymbol{X}_T)]^T\,\boldsymbol{\Sigma}_{TT}^{-1}\,[\boldsymbol{y}_T - m(\boldsymbol{X}_T)]\right) \end{aligned} \tag{13}$$

where n is the size of the training data set.

If MLE is used to estimate the coefficients, one needs to be cautious about the choice of optimization algorithms. In a high-dimensional coefficient space, the likelihood function is rarely convex, and thus gradient-based local optimization algorithms may not be effective. When global optimization methods such as the DIRECT algorithm [37] and simulated annealing [38] are used, careful selection of algorithm parameters is suggested. In order to ensure achieving the global maximum, we may need to manually divide the coefficient space into multiple smaller regions, and then search for maximums in these regions separately.

Note that the inverse of the covariance matrix $\boldsymbol{\Sigma}_{TT}$ is needed in the computation of the likelihood function in Eq. (12), as well as the construction of model prediction in Eq. (11). Some discussions on the efficient numerical strategies for inverting $\boldsymbol{\Sigma}_{TT}$ can be found in [39,40]. Also note that the size of $\boldsymbol{\Sigma}_{TT}$ increases with the size of

training data set, which may lead to ill-conditioned matrices and high computational cost. In such cases, sparse Gaussian process approximations may be used, which estimate the inverse of $\boldsymbol{\Sigma}_{TT}$ via projections from the high-dimension training data space to a low-dimension latent space [41].

5.2. Inclusion of surrogate model uncertainty

When GP surrogate models are available, they can be used for affordable uncertainty propagation that is necessary in probabilistic model validation. The issue with this approach is that it creates an additional source of uncertainty in validation. The validation result must be applied to the physics-based computational model (not the surrogate model) since it will be used in the prediction domain. To make this assessment, the additional uncertainty stemming from the uncertain fit of the surrogate to the computational model must be accounted for. Using a GP model $\hat{f}$ as previously described to approximate the model as a function of input $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$ provides a multivariate jointly Gaussian distribution at a set of prediction points arising from surrogate uncertainty as $\boldsymbol{Y_m} = \hat{f}(\boldsymbol{x}, \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_{Y_m}, \boldsymbol{\Sigma}_{Y_m})$. This represents a family of distributions for different values of $\boldsymbol{\theta}$. This family of distributions may be collapsed by employing the auxiliary variable approach [42] in which the dependence on the mean vector $\boldsymbol{\mu}_{Y_m}$ and prediction covariance matrix $\boldsymbol{\Sigma}_{Y_m}$ can be mapped to a dependence on a vector of CDF values $\boldsymbol{u}$ at the set of prediction points as in Eq. (14).

$$\boldsymbol{u} = F_{\boldsymbol{Y_m}}(\boldsymbol{y_m} | \boldsymbol{\mu}_{Y_m}, \boldsymbol{\Sigma}_{Y_m}) = \int_{-\infty}^{\boldsymbol{y_m}} f_{\boldsymbol{Y_m}}(\boldsymbol{\omega} | \boldsymbol{\mu}_{Y_m}, \boldsymbol{\Sigma}_{Y_m})\, d\boldsymbol{\omega} \tag{14}$$

Since the set of auxiliary variables $\boldsymbol{u}$ represent CDF values (each ranges from 0 to 1), the model reliability R becomes a random variable itself and can be written as a function of the random variables $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{\Theta}$. As shown in Eq. (15), the model reliability metric at input $\boldsymbol{x}$ can be computed for any realization of $\boldsymbol{u}$ by integrating over the distribution of $\boldsymbol{\Theta}$ as in Eq. (5). Then, the model reliability at any $\boldsymbol{x}$ is weighted by the pdf of $\boldsymbol{x}$ in the prediction domain (as in Eq. (7)) to obtain the overall distribution of the metric as a function of the surrogate model uncertainty as shown in Eq. (16).

$$r(\boldsymbol{x}, \boldsymbol{u}) = \int_{|y_m - y_d| < \epsilon} [y_m(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{u}) - y_D(\boldsymbol{x})] f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{15}$$

$$r_{overall}(\boldsymbol{u}) = \int r(\boldsymbol{x}, u) \pi(\boldsymbol{x})\, d\boldsymbol{x} \tag{16}$$

The resulting distribution of the validation metric can be computed by sampling realizations of the set of auxiliary variables to demonstrate the contribution of the GP uncertainty to the validation result. The process of sampling realizations of the auxiliary variables is equivalent to sampling realizations of the Gaussian random process. The spread in the distribution of the validation metric is the cost of using the surrogate model for propagation. This uncertainty may be reducible by improving the surrogate model by adding additional training points.

The proposed approach formalizes the validation assessment when using surrogate models for uncertainty propagation. When possible, it is preferable to use the original computational model directly, but constraints on computational effort often make such an approach unaffordable. When surrogates are necessary, the additional uncertainty can be included via the method described above. An alternative approach is to use sparse sampling of the original computational model for propagation, but sparse sampling also adds uncertainty to the output distribution [43]. The choice between sparse sampling and a surrogate model is problem specific and depends on the expense of the computational model as well as the smoothness of the underlying function. Note that each of the uncertainty integration steps that are discussed in this section and in Section 4 is performed numerically via sampling. Thus, incorporating each of these sources of uncertainty can be computationally expensive; this issue further motivates the need for surrogate models. Other efficient sampling techniques [43] can also be employed for propagation to further manage the computational expense.

6. Numerical example

6.1. Validation of MEMS device simulation

To demonstrate the proposed validation methodology, a microelectromechanical system (MEMS) example is introduced. The radio frequency (RF) MEMS switch, shown in the conceptual diagram in Fig. 2, is subjected to electrostatic loading that causes the membrane to deform. The mechanical properties of the membrane resist the deformation, but at some voltage, known as the pull-in voltage, the electrostatic force pulls the membrane into contact with the substrate. At a voltage level known as the pull-out voltage, the membrane can then be released from contact with the substrate. The pull-in and pull-out voltages are predicted by device simulation, and they are also measured in validation experiments (20 replicate tests on each of six devices).

Five variables, membrane thickness h, gap between one end of the membrane and the substrate g_1, gap between the other end of the membrane and the substrate g_2, Young's modulus E, and contact height d_c are identified as inputs to the model and experiments. Due to the imprecision of the measurement techniques, the geometry parameters g_1 and g_2 are described by distributions that represent input measurement uncertainty for each of the six devices. Direct measurements of E and d_c are not available, but the ranges of these two variables are obtained via multiscale simulation [44,45]. The thickness of the membrane h cannot be measured accurately, so it is treated as a calibration parameter.

This parameter is important to both the pull-in voltage and pull-out voltage predictions, so it could be calibrated from either of these sets of measurement data. However, it is often preferable to reserve an independent set of information for validation in order to avoid over-fitting during calibration and then use the validation data to explore the predictive capability of the estimated parameters for a separate prediction. With this approach, poor validation results provide evidence that the parameter estimates should be further investigated before being used for subsequent predictions. Therefore, in this example demonstration, the pull-in voltage measurements are used for calibration, and the pull-out voltage measurements are used for validation. From the pull-in voltage measurements, the membrane thickness is estimated separately for each device via Bayesian inference. Then, the calibrated distribution of thickness for each device is propagated through the simulation to predict pull-out voltage and compare against the pull-out voltage measurements. The measurement for

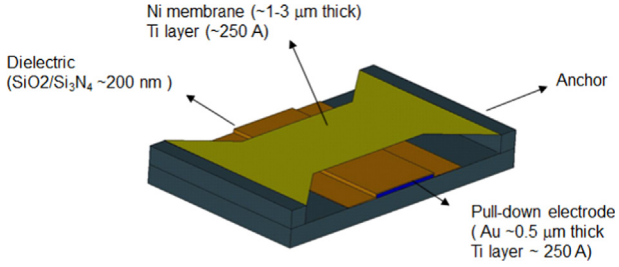

Fig. 2. RF MEMS switch.

each device corresponds to a combination of the input set [h, g_1, g_2, E, d_c], each with associated uncertainty. Thus, the validation measurements are partially characterized.

In a partially characterized data scenario, input measurement uncertainty can be treated in the same manner as parameter uncertainty when performing the validation assessment. For a single device, each of these inputs has a single value in reality, but it cannot be measured precisely. Aleatory uncertainty is only present in the form of device-to-device variation. Therefore, the source of the uncertainty in the prediction for a particular device (i.e., a particular input condition) is completely epistemic. The uncertainty in the observation is attributed to output measurement uncertainty, which is also epistemic. Therefore, a point-by-point comparison for each device using the model reliability metric can be performed.

For example, Fig. 3 demonstrates the model reliability metric computation for one of the six devices. The tolerance ϵ is set to 5 V, and the distribution of the difference between prediction and observation Δ is integrated over the interval $(-5, 5)$ to obtain a model reliability of 0.74. The prediction distribution shown in Fig. 3(a) is generated by propagating input measurement uncertainty through the prediction model. Since the computational model that predicts the pull-out voltage is expensive (approximately 6 h per evaluation) and a large number of Monte Carlo samples of the input measurement uncertainty are needed in order to converge the output distribution (10,000 were used in this illustration), using the computational model for propagation is unaffordable. Therefore, GP surrogate models are constructed to improve the efficiency of the computation. Each of the GP models that is constructed for this example uses a constant mean function with the squared exponential covariance function. The MLE of the parameters is used, computed according to the approach described in Section 5.1. For illustration, the surrogate uncertainty is not included in the result shown in Fig. 3; only the mean prediction from the GP model is used. If the computational model were not expensive, the uncertainty propagation could be performed without constructing a surrogate model, and the computation of the model reliability would proceed exactly as shown, resulting in a single value of the model reliability for each device. However, as mentioned, a surrogate model is needed for this example, and this uncertainty must also be included in the assessment. As a result, the model reliability is instead described by a distribution for each device. This consideration is demonstrated in Section 6.2.

6.2. Inclusion of surrogate uncertainty

The framework in Section 5 is applied to the validation assessment for each of the six devices. For each device, the model prediction is made for a set of samples of the input uncertainty. By sampling the auxiliary variable according to the approach described in Section 5.2, many realizations of the GP model are taken; each of these is a candidate prediction of the underlying computational model. The set of realizations produces a family of predictions that represents the possible outcomes for the validation assessment that could be obtained if the computational model were used directly. Note that these realizations are obtained by sampling the auxiliary variable and using the covariance function of the GP model, so the outputs at different samples of the input uncertainty are highly correlated. This correlation may result in a family of predictions with greater uncertainty than the standard deviation at a single prediction point would indicate. For each candidate model prediction, Eq. (15) is applied to obtain a value for the model reliability metric. This set of values for the model reliability is used to construct a histogram for the validation result for each device. The histograms are normalized to obtain the frequency diagrams shown in Fig. 4.

For several of the devices, the mean model reliability is very low because the mean prediction and mean observation were substantially biased from each other. This result may occur due to inadequacies in the model and/or inconsistencies in the observed data. As described in Section 3, both input and output measurement uncertainty may also contribute to the poor performance of the model (input measurement uncertainty increases the spread in the prediction while output measurement uncertainty increases the spread in the observation). Additionally, the spread in the potential outcomes of the model reliability indicates that the GP uncertainty is significant. By obtaining more training data, this particular source of epistemic uncertainty can be reduced, and the model reliability would be expected to converge toward the single value that would be obtained by performing the propagation with the computational model directly.

For most applications, the validation results shown in Fig. 4 would not provide sufficient confidence to use the model going forward in prediction. Either the model form should be improved, or the quality of the observation data should be thoroughly evaluated, and if necessary additional validation data should be collected. However, for illustration, the approach for integrating these

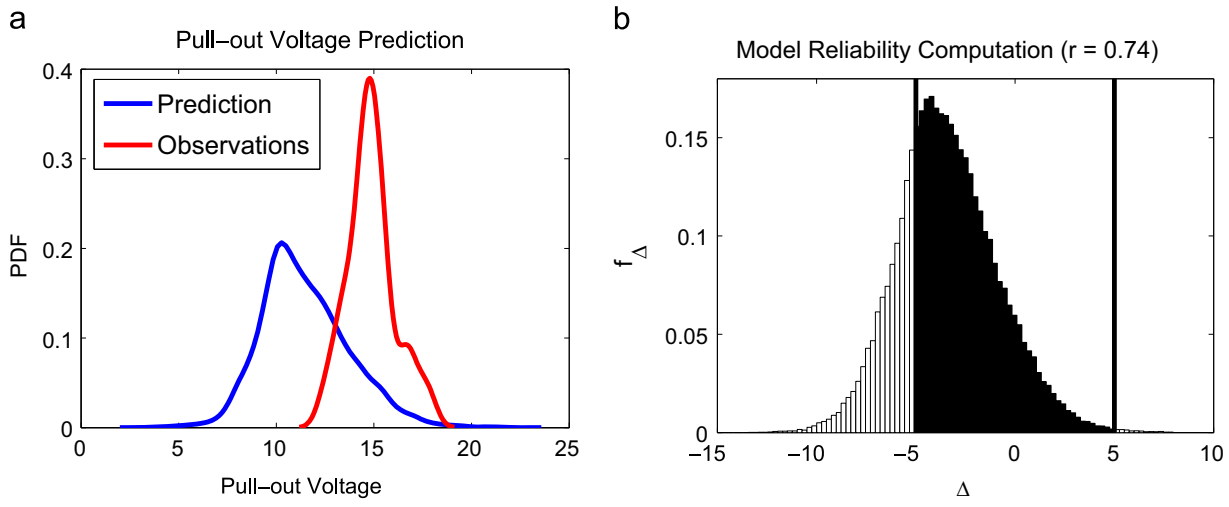

Fig. 3. Computation of model reliability for partially characterized validation data. (a) Comparison of model prediction and observation with associated epistemic uncertainty. (b) Computation of model reliability using the difference between prediction and observation. For $\epsilon = 5$, the difference Δ is integrated over the interval $(-5, 5)$ as shown.

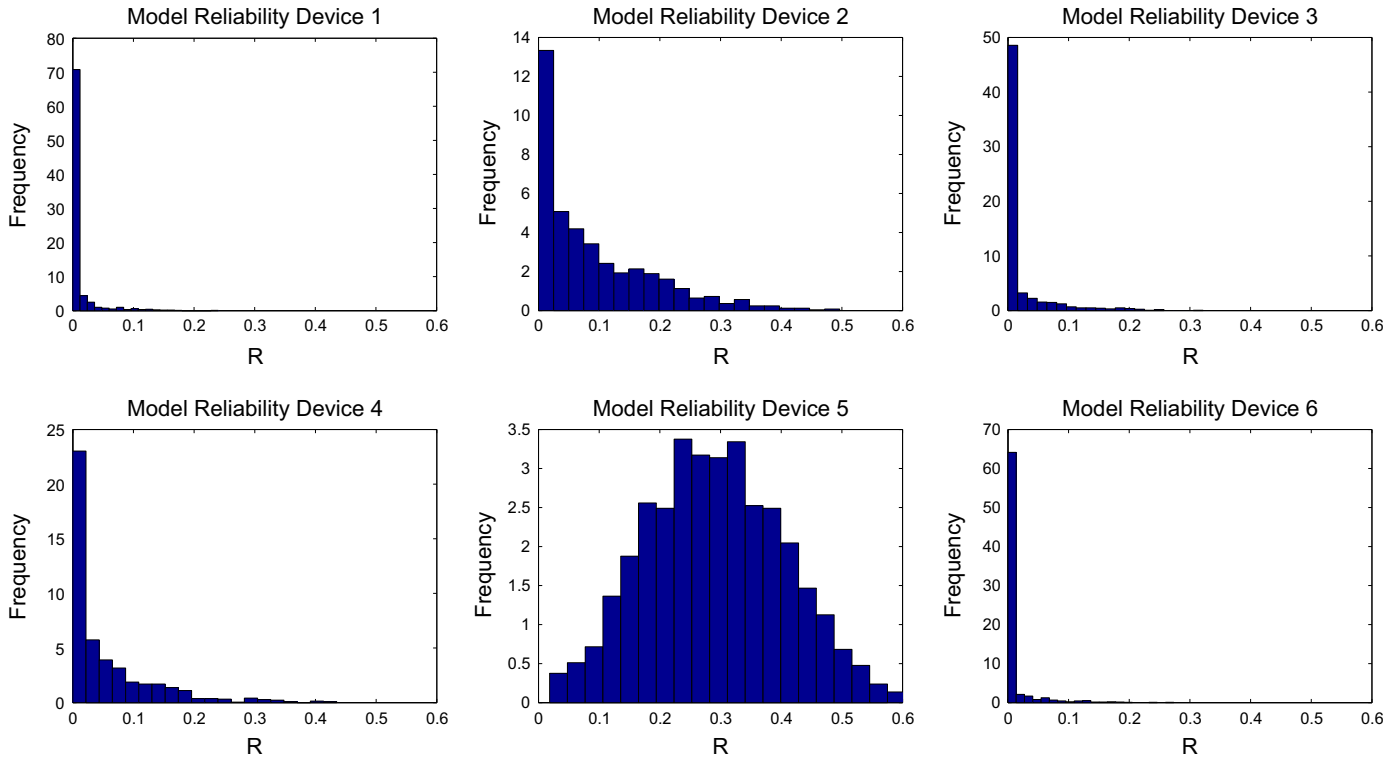

Fig. 4. Frequency diagrams of model reliability for each of 6 devices.

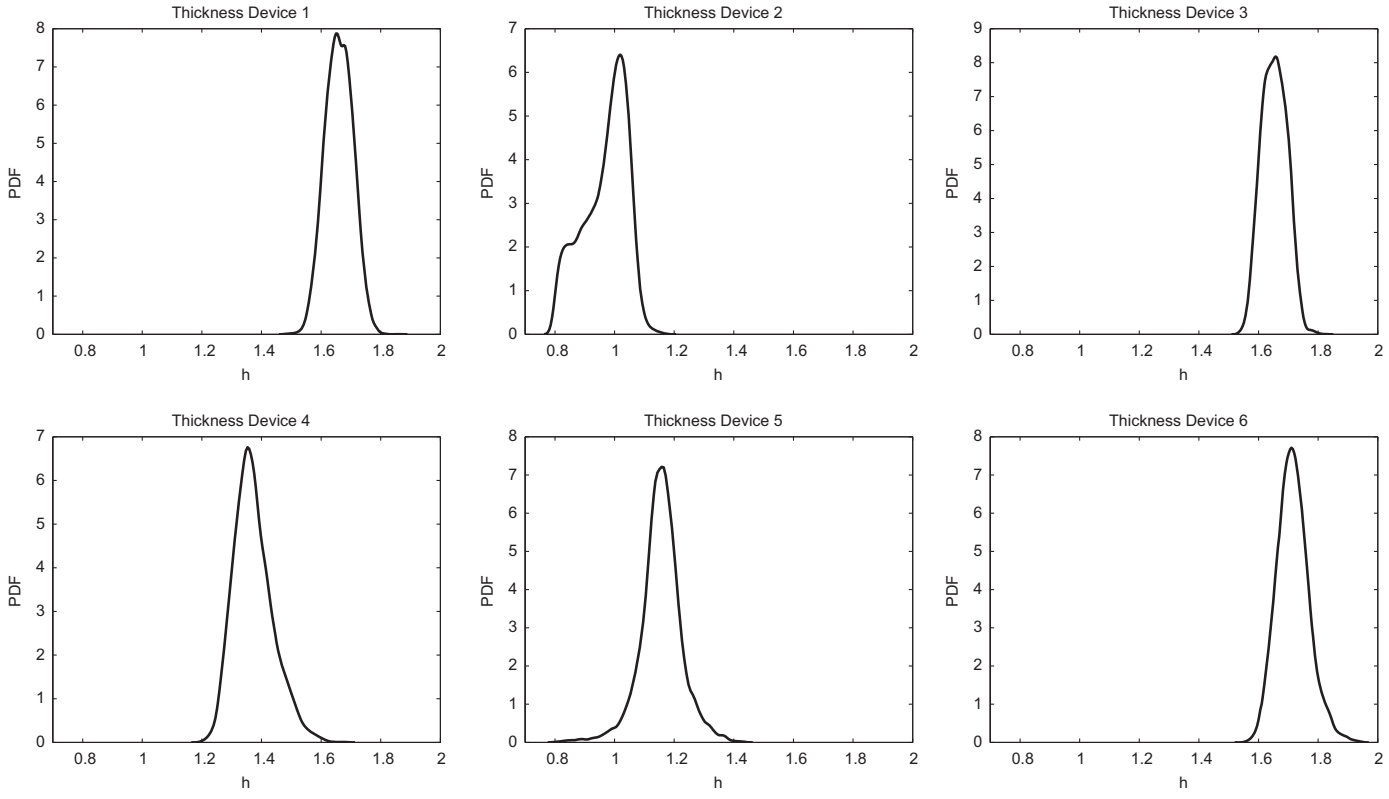

Fig. 5. Input uncertainty for the thickness of the 6 devices.

Table 2
Weights for two different prediction scenarios.

	Device 1	Device 2	Device 3	Device 4	Device 5	Device 6
$\pi(\boldsymbol{x}) \sim \mathcal{N}(1.2, 0.1)$	1.18e−4	0.114	1.11e−4	0.244	0.642	1.57e−5
$\pi(\boldsymbol{x}) \sim \mathcal{N}(1.7, 0.1)$	0.328	4.08e−10	0.323	9.56e−3	2.71e−5	0.339

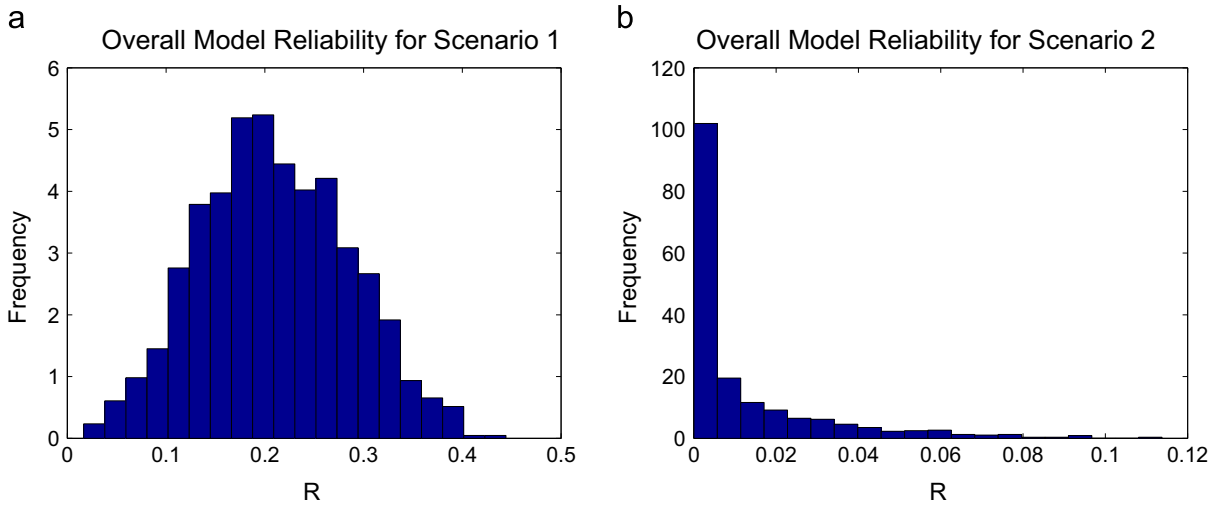

Fig. 6. The model is expected to perform much better for the first scenario since device 5 is the most relevant and also the best performer in the validation assessment. (a) Distribution of $r_{overall}$ for a prediction scenario with $\pi(\boldsymbol{x}) \sim \mathcal{N}(1.2, 0.1)$. (b) Distribution of $r_{overall}$ for a prediction scenario with $\pi(\boldsymbol{x}) \sim \mathcal{N}(1.7, 0.1)$.

results from different devices into a single result is demonstrated in Section 6.3 below.

6.3. Integration of validation results from multiple devices

Once individual validation results have been obtained for several different devices, it is useful to determine which of the results is most relevant to the prediction of interest. For example, if the beam thickness h is an input of particular interest, it is helpful to assess to the predictive capability of the model as a function of what thickness will be encountered. The validation tests that were conducted for thicknesses similar to those in the prediction scenario are most relevant. The calibrated thickness distributions for each of the six devices are shown in Fig. 5.

Suppose the model will be used for two different prediction scenarios in which the thicknesses will be $\mathcal{N}(1.2, 0.1)$ and $\mathcal{N}(1.7, 0.1)$, respectively. By applying Eq. (9) and normalizing the weights, the weights for the 6 devices are shown for the two scenarios in Table 2. It is clear from this table that device 5 is most relevant to the first prediction scenario while device 4 and device 2 are also somewhat relevant, and the other three devices are not. The second scenario has three device tests that are of nearly equal relevance (devices 1, 3, and 5), and the other three have negligible weight. By using these weights in Eq. (8), the integration in Eq. (16) can be approximated to produce the distributions for $R_{overall}$ shown in Fig. 6.

This validation assessment has shown that there is low confidence in the model in general, but this comparison shows that the model is much more adequate for the first prediction scenario than the second. Since only device 5 gave reasonable prediction quality in the validation assessment, it is reasonable to conclude that if the model is used in prediction at all, it should only be for input scenarios that are similar to the measured inputs of device 5. Therefore, the predictive capability of the model is very limited, which again emphasizes the need to improve both the model and the observation data.

7. Conclusion

This paper presents a model validation methodology for handling different data scenarios. When validation data is uncharacterized (corresponding inputs are not measured for each experiment), an ensemble validation approach is suitable. However, when inputs are also measured in validation tests (either fully or partially characterized data), it is preferable to perform validation individually for each input scenario. This enables aleatory and epistemic uncertainty sources to be separated from one another, which aids in decision making for uncertainty reduction when the model performance is inadequate. Additionally, understanding the reliability of the model as a function of the input may help to identify systematic inadequacies in model form. The individual metric values can be integrated into a single metric by weighting each value with the probability of observing the corresponding input in the prediction domain (i.e., relevance to the intended application of the model). When the computational expense of the model causes uncertainty propagation to be intractable, surrogate models are needed to obtain the distribution of the model prediction. This approach adds additional uncertainty into the assessment that should also be included in the analysis. With a GP surrogate model, the surrogate uncertainty can be readily obtained from the covariance structure of the model, and this uncertainty results in the model reliability metric itself being treated as a random variable with epistemic uncertainty. Once the model reliability metric is obtained (either a single value or a distribution), the metric can be interpreted probabilistically; this allows the validation result to be incorporated into the prediction.

Acknowledgements

This paper is based upon research partly supported by the U.S. 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] Kennedy MC, O'Hagan A. Bayesian calibration of computer models. J R Stat Soc: Ser B (Stat Methodol) 2001;63(3):425–64. http://dx.doi.org/10.1111/1467-9868.00294.
[2] Higdon D, Kennedy M, Cavendish J, Cafeo J, Ryne R. Combining field data and computer simulations for calibration and prediction. SIAM J Sci Comput 2005;26(2):448–66. http://dx.doi.org/10.1137/S1064827503426693.
[3] Trucano T, Swiler L, Igusa T, Oberkampf W, Pilch M. Calibration, validation, and sensitivity analysis: what's what. Reliab Eng Syst Saf 2006;91(10–11):1331–57. http://dx.doi.org/10.1016/j.ress.2005.11.031.
[4] Arendt PD, Apley DW, Chen W. Quantification of model uncertainty: calibration, model discrepancy, and identifiability. J Mech Des 2012;134(10):100908. http://dx.doi.org/10.1115/1.4007390 http://link.aip.org/link/JMDEDB/v134/i10/p100908/s1&Agg=doi.
[5] 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. Anal Chem 1995;67(24):4491–9. http://dx.doi.org/10.1021/ac00120a011.
[6] Hills RG, Trucano TG. Statistical validation of engineering and scientific models: background, Sandia Technical report, SAND99-1256.
[7] Ghanem R, Doostan A, Red-Horse J. A probabilistic construction of model validation. Comput Methods Appl Mech Eng 2008;197(29–32):2585–95. http://dx.doi.org/10.1016/j.cma.2007.08.029.
[8] O'Hagan A. Fractional Bayes factors for model comparison. J R Stat Soc, Ser B (Methodol) 1995;57(1):99–138.
[9] Rebba R, Mahadevan S, Huang S. Validation and error estimation of computational models. Reliab Eng Syst Saf 2006;91(10–11):1390–7. http://dx.doi.org/10.1016/j.ress.2005.11.035.
[10] Rebba R, Mahadevan S. Validation of models with multivariate output. Reliab Eng Syst Saf 2006;91(8):861–71. http://dx.doi.org/10.1016/j.ress.2005.09.004.
[11] Wang S, Chen W, Tsui K-L. Bayesian validation of computer models. Technometrics 2009;51(4):439–51. http://dx.doi.org/10.1198/TECH.2009.07011.
[12] Ferson S, Oberkampf W, Ginzburg L. Model validation and predictive capability for the thermal challenge problem. Comput Methods Appl Mech Eng 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. Int J Reliab Saf 2009;3(1):3–22. http://dx.doi.org/10.1504/IJRS.2009.026832.
[14] Roy C, Oberkampf W. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. Comput Methods Appl Mech Eng 2011;200(25–28):2131–44. http://dx.doi.org/10.1016/j.cma.2011.03.016.
[15] Rebba R, Mahadevan S. Computational methods for model reliability assessment. Reliab Eng Syst Saf 2008;93(8):1197–207. http://dx.doi.org/10.1016/j.ress.2007.08.001.
[16] Sankararaman S, Mahadevan S. Assessing the reliability of computational models under uncertainty. In: 54th AIAA/ASME/ASCE/AHS/ASC structures, atructural dynamics and materials conference, Boston, MA; 2013. ⟨http://dx.doi.org/10.2514/6.2013-1873⟩.
[17] Liu Y, Chen W, Arendt P. Toward a better understanding of model validation metrics. J Mech Des 2011;133(7):071005. http://dx.doi.org/10.1115/1.4004223.
[18] Ling Y, Mahadevan S. Quantitative model validation techniques: new insights. Reliab Eng Syst Saf 2013;111:217–31. http://dx.doi.org/10.1016/j.ress.2012.11.011.
[19] Hills RG, Leslie IH. Statistical validation of engineering and scientific models: validation experiments to application, Sandia Technical report (SAND2003-0706).
[20] Romero V, Luketa A, Sherman M. Application of a versatile "Real-Space" validation methodology to a fire model. J Thermophys Heat Transf 2010;24(4):730–44. http://dx.doi.org/10.2514/1.46358.
[21] Sankararaman S, Mahadevan S. Comprehensive framework for integration of calibration, verification and validation. In: 53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference, Honolulu, HI; 2012. ⟨http://dx.doi.org/10.2514/6.2012-1367⟩.
[22] O'Hagan A, Oakley JE. Probability is perfect, but we can't elicit it perfectly. Reliab Eng Syst Saf 2004;85(1–3):239–48. http://dx.doi.org/10.1016/j.ress.2004.03.014.
[23] Jaulin L, Kieffer M, Didrit O, Walter E. Applied interval analysis. New York, NY: Springer-Verlag; 2001.
[24] Shafer G. A mathematical theory of evidence. Princeton, NJ: Princeton University Press; 1976.
[25] Dubois D, Prade H. Possibility theory: an approach to computerized processing of uncertainty. New York, NY: Plenum Press; 1986.
[26] Ross TJ. Fuzzy logic with engineering applications. New York, NY: McGraw-Hill; 1995.
[27] Klir GJ, Wierman MJ. Uncertainty-based information: elements of generalized information theory, 2nd ed., vol. 15, Heidelberg, DE: Physica-Verlag; 1998.
[28] Helton J, Sallaberry C. Uncertainty and sensitivity analysis: from regulatory requirements to conceptual structure and computational implementation. In: Dienstfrey A, Boisvert R, editors. Uncertainty quantification in scientific computing, IFIP Advances in information and communication technology, vol. 377, Berlin Heidelberg: Springer; 2012. pp. 60–77. ⟨http://dx.doi.org/10.1007/978-3-642-32677-6_5⟩.
[29] Oberkampf WL, Helton JC, Joslyn Ca, Wojtkiewicz SF, Ferson S. Challenge problems: uncertainty in system response given uncertain parameters. Reliab Eng Syst Saf 2004;85(1–3):11–9. http://dx.doi.org/10.1016/j.ress.2004.03.002.
[30] Kiureghian A. Aleatory or epistemic? Does it matter? Struct Saf 2009;31(2):105–12. http://dx.doi.org/10.1016/j.strusafe.2008.06.020.
[31] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22(1):79–86. http://dx.doi.org/10.1214/aoms/1177729694.
[32] Angus J. The probability integral transform and related results. SIAM Rev 1994;36(4):652–4. http://dx.doi.org/10.1137/1036146.
[33] Kass R, Raftery A. Bayes factors. J Am Stat Assoc 1995;90(430):773–95.
[34] Pericchi LR. Handbook of statistics, vol. 25: Bayesian thinking, modeling and computation, 1st ed., North Holland; 2005. pp. 115–149 (Chapter 6).
[35] Hombal VK, Mullins J, Mahadevan S. Extrapolation confidence assessment for predictions of computational engineering models, Comput Methods Appl Mech Eng, Submitted for Publication.
[36] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. The MIT Press; 2006.
[37] Finkel D, Kelley C. Convergence analysis of the DIRECT algorithm. Optim Online 2004:1–10.
[38] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983;220(4598):671–80. http://dx.doi.org/10.1126/science.220.4598.671.
[39] McFarland JM. Uncertainty analysis for computer simulations through validation and calibration [Ph.D. thesis]. Vanderbilt University; 2008.
[40] Haarhoff LJ, Kok S, Wilke DN. Numerical strategies to reduce the effect of ill-conditioned correlation matrices and underflow errors in Kriging. J Mech Des 2013;135(4):044502. http://dx.doi.org/10.1115/1.4023631.
[41] Quinonero-Candela J, Rasmussen CE. A unifying view of sparse approximate Gaussian process regression. J Mach Learn Res 2005;6:1939–59.
[42] Sankararaman S, Mahadevan S. Separating the contributions of variability and parameter uncertainty in probability distributions. Reliab Eng Syst Saf 2013;112:187–99. http://dx.doi.org/10.1016/j.ress.2012.11.024.
[43] Haldar A, Mahadevan S. Probability, reliability, and statistical methods in engineering design. New York: Wiley; 2000.
[44] Koslowski M, Strachan A. Uncertainty propagation in a multiscale model of nanocrystalline plasticity. Reliab Eng Syst Saf 2011;96(9):1161–70. http://dx.doi.org/10.1016/j.ress.2010.11.011.
[45] Kim H, Venturini G, Strachan A. Molecular dynamics study of dynamical contact between a nanoscale tip and substrate for atomic force microscopy experiments. J Appl Phys 2012;112(9):094325. http://dx.doi.org/10.1063/1.4762016 http://link.aip.org/link/JAPIAU/v112/i9/p094325/s1&Agg=doi.