Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 283 (2015) 1310–1335
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Validating predictions of unobserved quantities

Todd A. Oliver[a,*], Gabriel Terejanu[b], Christopher S. Simmons[a], Robert D. Moser[a,c]

[a] *Center for Predictive Engineering and Computational Sciences, Institute for Computational Engineering and Sciences, The University of Texas at Austin, USA*
[b] *Department of Computer Science and Engineering, University of South Carolina, USA*
[c] *Department of Mechanical Engineering, The University of Texas at Austin, USA*

Received 21 April 2014; received in revised form 15 August 2014; accepted 20 August 2014
Available online 29 August 2014

Abstract

The ultimate purpose of most computational models is to make predictions, commonly in support of some decision-making process (e.g., for design or operation of some system). The quantities that need to be predicted (the quantities of interest or QoIs) are generally not experimentally observable before the prediction, since otherwise no prediction would be needed. Assessing the validity of such extrapolative predictions, which is critical to informed decision-making, is challenging. In classical approaches to validation, model outputs for observed quantities are compared to observations to determine if they are consistent. By itself, this consistency only ensures that the model can predict the observed quantities under the conditions of the observations. This limitation dramatically reduces the utility of the validation effort for decision making because it implies nothing about predictions of unobserved QoIs or for scenarios outside of the range of observations. However, there is no agreement in the scientific community today regarding best practices for validation of extrapolative predictions made using computational models. The purpose of this paper is to propose and explore a validation and predictive assessment process that supports extrapolative predictions for models with known sources of error. The process includes stochastic modeling, calibration, validation, and predictive assessment phases where representations of known sources of uncertainty and error are built, informed, and tested. The proposed methodology is applied to an illustrative extrapolation problem involving a misspecified nonlinear oscillator.

Keywords: Extrapolative predictions; Validation; Model discrepancy; Bayesian inference

1. Introduction and motivation

Advances in computing hardware and algorithms in recent decades, along with accompanying advances in the fidelity of computational models, enable simulation of physical phenomena and systems of unprecedented complexity. This capability is revolutionizing engineering and science. For example, results of computational simulations are used heavily in the design of nearly all complex engineering systems, from consumer electronics to spacecraft and nuclear power plants. Furthermore, results of computational models are used to inform policy decisions in areas where the con-

* Correspondence to: 201 E. 24th Street, C0200, Austin, TX 78712, USA.
E-mail addresses: oliver@ices.utexas.edu (T.A. Oliver), terejanu@cec.sc.edu (G. Terejanu), csim@ices.utexas.edu (C.S. Simmons), rmoser@ices.utexas.edu (R.D. Moser).

http://dx.doi.org/10.1016/j.cma.2014.08.023
0045-7825/

sequences of inaccurate predictions and poorly-informed decisions could be catastrophic, such as disaster response and climate change. However, for computational modeling to realize its full potential in such applications, it is critical that the reliability of the results of the models be systematically characterized. In the current paper, an approach to making such reliability assessments is proposed for a broad class of problems in computational science and engineering.

Any reliability assessment of a computational model must necessarily consider the purpose for which the model is to be used. A common use of computational models is to predict the response of some complex system and thereby to inform some decision regarding the system. For example, predictions from computational models might be used to decide which of several competing system designs will be superior, to decide whether a proposed use scenario for a system will meet operational objectives, or to decide how to respond to a system as it evolves. Typically, the decisions are informed by predictions of specific quantities describing the response of the system, these are the so-called quantities of interest (QoIs). An important feature of such uses of computational models is that observational data for the QoIs in the prediction scenario are not available, since otherwise the predictions would not be needed. This is the situation of interest here. There are of course other modes in which computational models are used in science and engineering, and in those cases the reliability issues will be different.

The assessment of reliability of computational models is often divided into three aspects: verification, validation and uncertainty quantification (collectively known as V&V-UQ). Verification is concerned with assessing the discrepancy between the computer simulation and the underlying mathematical model on which it is based. Uncertainty quantification is the process of assessing uncertainties that affect simulation predictions, such as those due to uncertain model inputs or inaccuracies in the model itself, and determining the resulting uncertainty in the QoIs. Finally, following [1–3], validation is the process of determining whether a mathematical model is a sufficient representation of reality for the purposes for which the model will be used; that is, for predicting specified QoIs to inform a specific decision. Thus, while verification is a purely mathematical process concerned only with the difference between computational and mathematical models, UQ and validation are concerned with the discrepancy between the mathematical model and the real world.

While verification is vitally important and sometimes overlooked in practice, it is largely understood [4,5], with well developed techniques for estimating and controlling the impact of numerical errors on specified QoIs [6,7]. In the remainder of this paper, it will be assumed that numerical solutions have been verified to ensure that numerical errors in the predictions are small compared to other sources of uncertainty, and so, verification will not be discussed further. Instead, we focus on the validation of and quantification of uncertainty in predictions of unobserved QoIs.

1.1. Validation processes

In engineering practice, the "validity" of a computational model is often assessed by simply comparing the output of the model with experimental data. While such comparisons must be a part of any validity assessment, a straightforward process considering only the closeness of this comparison has two important shortcomings when used to assess the reliability of model predictions. First, it does not account for the uncertainties associated with the model or the data. Second, it precludes an assessment of the reliability of predictions in new situations or for unobserved quantities, which as discussed above, is the use we consider here. In essence, predictions are always extrapolations from available information, and the validation question is whether such extrapolation is justified.

In this context then, it is the validity of the prediction that is really of interest. We cannot speak of the validity of a model in general, since a model may produce valid predictions of some quantities in some circumstances, and not of others. This is different from the notion of validity of scientific theories, in which we insist that any inconsistency between a theory and observations falsifies the theory. In computational modeling, models known to be false in this strict scientific sense are used routinely (e.g., Newtonian mechanics), and the resulting predictions can nonetheless be valid, provided the inadequacies of the model have no significant effect on the prediction.

To address the shortcomings of naive comparisons with data as a technique for validating extrapolative predictions, a number of more sophisticated procedures have been proposed [4,5,8–16], and validation guidelines have been developed by professional engineering societies [1–3]. These have generally been positive developments, but they also have shortcomings. Most commonly, they do not directly address the validity of models to make predictions of unobserved QoIs. For instance, Higdon et al. [14] and Bayarri et al. [13] present similar validation frameworks for computational models, based on the work of Kennedy and O'Hagan on model calibration and model discrepancy representations [17]. These frameworks rely on statistical models—specifically Gaussian process models—to represent

the difference between the model outputs and observational data. Such representations can improve parameter inferences by accounting for discrepancies between the data and the model outputs that are caused by modeling errors. However, they are insufficient for validation of the ability to predict unobserved QoIs, because the discrepancy model is posed only for the observable quantities and there is generally no direct mapping from the observables to the QoIs.

A different approach in addressing the inadequacies in model structure is given by Strong et al. [18]. The main idea is to decompose the model into sub-functions and make judgments about the discrepancy of each sub-function. Strong et al. [18] propose a model refinement strategy based on sensitivity analysis to quantify the relative importance of different structural errors within a health economics model. The strategy of introducing discrepancy terms within the model at the source of the structural error is also embraced in the present paper. We leverage this technique to create a direct mapping between the discrepancy terms and both the observables and the unobserved QoIs. However, a number of challenges need to be addressed in using this approach for extrapolative predictions. First, since the discrepancy representation is a statistical model, it is highly dependent on calibration against observations and, hence, should not be used in situations in which it cannot be trained and tested. Thus, in general, use of this sort of discrepancy model in extrapolative predictions is suspect. Unlike Strong et al. [18] which use this strategy to guide model refinements based solely on building discrepancy models on prior knowledge, for our physics-based extrapolation problem we introduce a systematic calibration, validation and predictive assessment of these discrepancy terms to build confidence in their predictive capability. Furthermore, since there is no unique decomposition of the model, identifying the sources of error remains an issue. In our approach, we leverage the fact that the models are physics-based. Such models are generally constructed from highly reliable physical theories coupled with less reliable models to close the governing equations. This structure gives us a unique view of the modeling error. Since modeling errors are introduced entirely through the empirical models, we only need to introduce discrepancy terms where these embedded empirical models enter the formulation.

In pioneering work by Babuška et al. [11], the impact of model discrepancy on unobserved QoIs was accounted for in the validation process. They addressed a structural mechanics problem in which the primary modeling challenge was the statistical representation of the unknown, spatially varying elastic modulus. In this case, it was expected that the statistical model could be calibrated to be consistent with observations from each of several available experimental scenarios but that these separate calibrations need not be consistent with each other. The validation question in this case, was whether any such calibration discrepancies were important to the prediction of the QoI. This was tested by comparing the predictions arising from the different calibrations. This approach can be generalized to other problems in which the model parameterization is sufficiently rich that the model can always be adjusted to fit data from a single experimental scenario [16,19,20]. While this was true for the problem and models investigated by Babuška et al., with many engineering models this is not the case. Indeed, a common symptom of model inadequacy is that a model cannot be calibrated to match experimental data within experimental uncertainty, even for a single experimental scenario. In such cases, the Babuška et al. validation approach is not able to account for the impact of these discrepancies on the QoIs, and thus, a more general formulation is needed.

1.2. Predictive validation

As discussed above, there are currently no established techniques for assessing the validity of predictions of unobserved QoIs. Here we address this problem by proposing a process we call "predictive validation" by which it is possible to test the validity of such extrapolative predictions in a broad class of problems. In defining predictive validation, we will also specify the model characteristics that enable reliable predictions, a set of necessary conditions for extrapolative predictions and processes that allow satisfaction of these necessary conditions to be tested.

The predictive validation process described here was developed to evaluate predictions regarding physical systems. Such systems are commonly described by models based on highly reliable theory (e.g., conservation laws), whose validity is not in question in the context of the predictions to be made. However, these highly reliable theories typically must be augmented with one or more "embedded models", which are less reliable. The less reliable embedded models may embody various modeling approximations, empirical correlations, or even direct interpolation of data. For example, in continuum mechanics, the embedded models might include constitutive models and boundary conditions, while in molecular dynamics they would include models for interatomic potentials. We will refer to such models—i.e., highly-fidelity models with lower-fidelity embedded components—as composite models.

The fact that the composite models used for prediction are built on highly reliable models is an important ingredient enabling reliable extrapolation in the approach described here, even though less reliable embedded models are also used. Specifically, we require that the less reliable models are not used outside the range where they have been calibrated and tested. Indeed, in general, low fidelity embedded models have some finite "domain of applicability", and using them outside this domain is an error. To avoid such errors, they should ideally be tested for in software implementations. This restriction does not necessarily limit our ability to extrapolate using the composite model since the relevant scenario space for each embedded model is specific to that embedded model, not the composite model in which it is embedded.

Another important ingredient is a representation of the uncertainty. We require mathematical representations of existing or prior information, the uncertainty in the observational data that will be used to inform and test the model, and the uncertainty in model predictions resulting from model inadequacy. Since Bayesian probability [21–23] provides a powerful representation of uncertainty, probabilistic models are used to describe uncertainty in this work, and the development of these probabilistic models is the first step of the predictive validation process. In particular, as in previous work [13,14,17], our approach relies heavily on statistical models of model discrepancy. However, unlike previous work, we take advantage of the structure of the composite model described above to introduce discrepancy models that can be used to evaluate uncertainty in unobserved quantities. This advance is accomplished by posing a model for the error in the embedded physical model directly, rather than a model for the discrepancy between observations and the model output for the observables.

Such probabilistic error representations for embedded models generally require knowledge and insight regarding the physical phenomena and models being used. Thus, they require significant effort to develop. Further, such models often greatly increase the computational cost of using the composite model because evaluation of the model requires forward propagation of uncertainty. In situations where predictions of unobserved quantities are not required, the investments necessary to develop and use these embedded inadequacy representations may not be worthwhile. However, these investments are essential for validating predictions of unobserved quantities because without such representations it is not possible to connect observed discrepancies with data to expected errors in the predicted QoIs.

Prior information on parameters and experimental uncertainty is also represented using Bayesian probabilistic models. In this way, we strive to account for all significant sources of uncertainty, including uncertainty in model inputs (e.g., parameters and boundary conditions), model errors, and observational data; so that the uncertainties in any model output, whether observed or unobserved, are represented. The original composite model plus these uncertainty models provides a complete model representing both our knowledge of the physics as well as important sources of uncertainty. This complete model is then the subject of the predictive validation process.

Typically, both the embedded physical models and the accompanying uncertainty models have parameters that must be determined from observations through calibration. In our approach, calibration is the first in an integrated, three-step process for assessing the reliability of extrapolative predictions issued by the complete model described above. The three steps—calibration, validation, and predictive assessment—are designed to assess the predictive capability of the model in different ways. In calibration, the model is informed by data. Specifically, parameter values and their uncertainties are inferred from available observations by solving an inverse problem. The use of probability to represent uncertainty naturally leads to the formulation of the calibration problem as a Bayesian update.

In validation, outputs from the calibrated model are checked for consistency with available observations. While this consistency is not sufficient for predictive validation, it is necessary. Unlike calibration, the validation step is not naturally expressed in the language of Bayesian hypothesis testing because there are not well-defined alternative hypotheses. Instead, we must assess whether the validation data are plausible according to the model. There are a number of possibilities for quantifying this plausibility. Here we use highest posterior density credible sets.

The calibration and validation processes both require observational data. One issue that must be addressed regarding the data is what, if any, of the data should be excluded from the calibration process and used only for validation. The optimum approach must clearly depend on the details of the problem and available data, since validation tests need to provide confidence in the prediction of the problem-specific QoIs. A detailed discussion of possible approaches is beyond the scope of this paper, though partition strategies have been proposed for several specific applications [16,24–26]. In general, however, it is clear that calibration benefits from abundant data from observations in simple systems, and that observations of different experiments, different scenarios and/or different observable quantities than those used in calibration often yield particularly challenging and therefore useful validation tests.

Finally, predictive assessment determines whether the calibration and validation phases were sufficiently informative and challenging to provide confidence in the reliability of the predictions of the QoIs. Specifically, one must answer a number of questions about what the validation tests imply about the QoI predictions. This assessment relies on sensitivity analysis along with knowledge about the embedded models and their formulation.

Notice that in this approach to predictive validation, calibration and validation are based primarily on statistical analysis. However, all representations of uncertainty (e.g., models of uncertainty due to model discrepancy) depend heavily on the structure of the physical model and knowledge about the physical system being modeled. This knowledge is also crucial to the predictive assessment. It is this reliance on knowledge of and reliable theory about the physical system being modeled that makes extrapolative prediction possible.

The remainder of the paper is organized as follows. Important features of a composite model and associated uncertainty representations that enable predictive validation for unobserved QoIs are described abstractly in Section 2. Specific procedures used to assess the model and build confidence in its predictive capability are introduced in Section 3. The main ideas and tools of the predictive validation approach are illustrated using a simple example involving extrapolative predictions made for a spring–mass–damper system in Section 4. Concluding remarks in Section 5 address the practical challenges in carrying out the proposed validation processes for real-world applications.

2. Model inadequacy and discrepancy representations in predictive validation

In developing the predictive validation process to be described, it is important to be precise about what constitutes a prediction. Here, a prediction is the result of a computational simulation conducted to compute specific QoIs which are to be used to support some decision regarding the system being simulated. Further, there is no observational data available for the QoIs for the scenarios of interest, since otherwise predictions would not be necessary. Thus, the credibility of the prediction must be established based on available data for other quantities and/or scenarios, along with any other available information about the system. The fundamental challenge is to make credible predictions, despite the fact that the predictions are extrapolations from available information. Furthermore, in recognition of the fact that observational data have uncertainties, and that models are imperfect, we insist that predictions must be endowed with characterized uncertainties.

In general, the need to extrapolate raises concerns about the reliability of the predictions, and, indeed, we must ask: "what entitles us to make such predictions?" Part of the answer is that, unlike purely empirical models that can be used solely to represent observations, the models used to predict the behavior of physical systems are often based on theories that are known to be highly reliable within well-defined domains of applicability. However, these highly reliable theories are usually augmented with one or more "embedded models", which are less reliable. Thus, the model as a whole is a composite model, as defined in Section 1. In the predictive validation approach proposed here, reliable predictions are enabled by the fact that reliable theories, whose validity in the prediction scenario is not in doubt, form the foundation of the composite model.

To make these ideas clear, a simple abstract prediction problem is presented in Section 2.1. This is the simplest problem that has the characteristics of prediction discussed above. Critical to the predictive validation process is the representation of uncertainty due to the imperfections of the model. Background on such discrepancy representations is provided in Section 2.2, and discrepancy modeling for predictive validation is described in Section 2.3. Finally, generalizations of the abstract problem described in Section 2.1 to encompass predictions in complex physical systems are discussed in Section 2.4.

2.1. Abstract problem statement

Reliable predictions are enabled by the use of reliable physical theory whose validity in the context of the predictions to be made is not in doubt. Let this theory be written mathematically as

$$\mathcal{R}(u, \tau; r) = 0, \tag{1}$$

where $\mathcal{R}$ is an operator expressing the theory. For example, in continuum mechanics, $\mathcal{R}$ would be a partial differential operator expressing conservation of mass, momentum, and energy. In this formulation, u is the solution or state variable, and r is a set of scenario variables needed to precisely define the problem being considered. The scenario variables might include the geometry of the solution domain, boundary conditions, and other parameters that define the

problem. The final variable τ is a quantity that needs to be known to solve (1). For example, in continuum mechanics, τ could be the strain energy or the stress tensor.

If τ were known in terms of u and r, the system would be closed, and (1) would implicitly define a mapping from the scenario variables r to the solution variables u. However, it is often the case that the required relationship between τ and u and r is unknown or does not exist—i.e., u and r do not fully define τ. In either case, a model, which we call an embedded model, is required, and is written τ_m:

$$\tau \approx \tau_m(u; s, \theta), \tag{2}$$

where $\approx$ indicates that the model is imperfect; s is a set of scenario variables for the embedded model; and θ is a set of parameters required by the model in addition to the scenario. These are the calibration or tuning parameters for the model. Note that the scenario space of the embedded model may be different from that of the global model in which it is embedded, so that r and s are not necessarily the same. In particular, s often includes only a subset of the variables of r. Further, in many settings, τ_m is formulated entirely in terms of the local solution u and calibration parameters θ, in which case s is empty. The fact that the scenario spaces of the global composite model and the embedded model are different is an important feature enabling reliable extrapolative predictions.

For the purposes of model calibration and validation, we require that some observable quantities y can be measured experimentally. These observable quantities are different from the prediction QoIs q, but both y and q are determined from the model state, the scenario, and possibly the embedded model:

$$y = \mathcal{Y}(u, \tau; r), \tag{3}$$

$$q = \mathcal{Q}(u, \tau; r), \tag{4}$$

where, for simplicity of exposition, the theories underlying the operators $\mathcal{Y}$ and $\mathcal{Q}$ are presumed to be as reliable as the models embodied by $\mathcal{R}$.

2.2. Background on model discrepancy

In general, the model τ_m is less reliable than the model in which it is embedded either because its dependencies or functional form are incorrect (structural inadequacy/uncertainty) or because the parameters θ are not perfectly known (parameter error/uncertainty) or both. These errors introduce error in the model and, in turn, in both the solution and the QoI. In their seminal paper, Kennedy and O'Hagan [17] address this fact by introducing a statistical model for "the difference between the true value of the real world process and the code output at the true values of the inputs". In their approach, which has become very common in the Bayesian literature [13,14], this statistical model is posed directly for the observable quantities, such that (3) is replaced by

$$y = \mathcal{Y}(u, \tau_m; r) + \delta_y(r; \alpha),$$

where δ_y is the model for the uncertainty in the observables induced by structural inadequacy, and α are a set of hyperparameters for that model, while u is still determined by the original composite model $\mathcal{R}(u, \tau_m; r) = 0$.

In the present context, this treatment of the error is incomplete because it says nothing about the effect of model discrepancy on the QoIs. An analogous formulation for the QoIs is

$$q = \mathcal{Q}(u, \tau_m; r) + \delta_q(r; \beta),$$

and thus, the full model is given by

$$\mathcal{R}(u, \tau_m; r) = 0, \tag{5}$$

$$y = \mathcal{Y}(u, \tau_m; r) + \delta_y(r; \alpha), \tag{6}$$

$$q = \mathcal{Q}(u, \tau_m; r) + \delta_q(r; \beta). \tag{7}$$

Here, the hyperparameters α and β need to be calibrated, as well as the parameters of τ_m, θ. In the Kennedy and O'Hagan approach, the calibration of θ and α would be performed together using data for the observable y. But, because data for the QoIs q are not available, this approach cannot be used to calibrate the hyperparameters β. Furthermore, even if one were able to pose a model δ_q that did not require calibration, there would be no way to

test this model in the validation process, making it inappropriate for use in predictions. Clearly, a different approach is needed. Here we propose to take advantage of the structure of the composite model to introduce model uncertainty representations that can be informed and tested using data as well as used to quantify the uncertainty in predictions of the QoIs.

2.3. Discrepancy modeling for predictive validation

The point of predictive validation is to assess the predictive capability of the model—i.e., to characterize the accuracy of the predictions of q. A key challenge in this process is to determine what the observed discrepancies between the model outputs and observational data imply about the reliability of the QoI predictions. In the formulation of Section 2.2, one must infer δ_q from observations of y. In general, this inference is nearly if not impossible because there is no direct mapping from the observables to the QoIs—i.e., given only y, one cannot evaluate q. Thus, δ_q cannot be constructed directly from the physical model and the observations alone. Additional modeling assumptions are required.

Alternatively, in the predictive validation process, a mathematical relationship between the observables and the QoIs is constructed by formulating uncertainty models to represent errors at their sources. Such models are able to provide uncertain predictions for both the observables and the QoIs without additional assumptions. Thus, observational data can be used to inform and test these uncertainty models, and that information can directly influence the predictions of the QoIs.

To accomplish this, we recognize that the sole source of the modeling error in the current example is the embedded model τ_m. Thus, instead of modeling the effects of this error on the observables and QoIs separately as in Section 2.2, we enrich the embedded physical model (2) with a model that represents not only the physics of the original model but also the uncertainty introduced by the structural inadequacy of that model. For example, one could write

$$\tau \approx \tau_m(u, s; \theta) + \epsilon_m(u, s; \alpha), \tag{8}$$

where ϵ_m denotes the uncertainty representation, which may depend on additional parameters α. Given our choice to use probability to represent uncertainty, it is natural that ϵ_m is a stochastic model, even when the physical phenomenon being modeled is inherently deterministic. Of course, an additive model is not necessary; other choices are possible. More importantly, the form of ϵ_m must be determined. The specification of a stochastic model ϵ_m is driven by physical knowledge about the nature of error as well as practical considerations necessary to make computations with the model tractable. For example, when the enriched model (8) is introduced into (1) so that it can be solved for u, which is now stochastic, the fact that ϵ_m depends on u will in general make this solution very difficult. In practice, we have either formulated ϵ_m to be independent of u, as in the example in Section 4, or have defined ϵ_m through an auxiliary equation of the form $f(u, \epsilon_m; w) = 0$, where w is an auxiliary random variable that is independent of u. In this later case, the auxiliary equation can then be solved together with (1). Other practical formulations for introducing u dependence in ϵ_m may also be possible.

Although general principles for developing physics-based uncertainty models need to be developed, the specification of such a model is clearly problem-dependent and, thus, will not be discussed further here.

For the current purposes, it is sufficient to observe that the model ϵ_m is posed at the source of the structural inadequacy—i.e., in the embedded model for τ. The combination of the physical and uncertainty models forms an enriched composite model, which takes the following form in the current case:

$$\mathcal{R}(u, \tau_m + \epsilon_m; r) = 0, \tag{9a}$$

$$y = \mathcal{Y}(u, \tau_m + \epsilon_m; r), \tag{9b}$$

$$q = \mathcal{Q}(u, \tau_m + \epsilon_m; r). \tag{9c}$$

The inadequacy model, ϵ_m, appears naturally in the calculation of both y and q, both directly through the possible dependence of $\mathcal{Y}$ and $\mathcal{Q}$ on τ, and indirectly via the dependence of u on τ through $\mathcal{R}$. The structural uncertainty can therefore be propagated to both the observables and the QoIs. Furthermore, one can learn about ϵ_m—i.e., inform and test the model—from data on the observables and then transfer that knowledge to the prediction of the QoIs. The predictive validation process described in Section 3 is designed to address prediction problems of this type. The process involves training (calibration) embedded models and their inadequacy models, testing (validation) the

embedded model and its use in the enriched composite model (9), and assessing (predictive assessment) whether the knowledge gained through these processes can be reliably transferred to the QoI predictions.

2.4. Generalized problem abstraction for complex systems

The abstract problem statement from Section 2.1 was designed as a simple illustrative example of a broad class of problems in which validated predictions of unobserved quantities are needed. This simple formulation can be generalized in a number of ways to encompass most computational prediction problems in science and engineering, including predictions in complex physical systems, in which many interacting physical phenomena are at work. These generalizations are discussed here.

The first generalization is that many unreliable embedded models of different phenomena may be required to represent the system being studied. For example, in a combustion problem, models for thermodynamics, molecular transport, radiation transport, and chemical reactions may need to be embedded in the equations for the conservation of mass, momentum and energy. To represent this situation, we introduce a set of N quantities τ_i for $1 \leq i \leq N$ which need to be modeled to close the governing equations:

$$\mathcal{R}(u, \tau_1, \tau_2, \ldots, \tau_N; r) = 0. \tag{10}$$

An embedded model for each of these quantities would then be needed, each of which could have a set of calibration parameters θ_i, a set of model-specific scenario parameters s_i, and an inadequacy representation ϵ_{im} with hyperparameters α_i:

$$\tau_i \approx \tau_{im}(u; \theta_i, s_i) + \epsilon_{im}(u; \alpha_i, s_i). \tag{11}$$

The second generalization is that the experimental situation in which observations are made may be so different from the prediction situation that it is represented by a different reliable model $\mathcal{R}$. This experimental scenario might also depend on a number of new quantities τ that must be modeled. For example, these might represent the response of the measuring instrument or some characteristic of the experimental apparatus. In this case, the experimental scenario may be described by different parameters than the prediction scenario. There will in general be some number N_e of different experiments, each of which involves a set of modeled quantities particular to the experiment as well as at least one of the modeled quantities used in predictions. To provide the most direct information regarding the embedded models used in the predictions, each of the experimental scenarios used for calibration would involve only one of the embedded models used for prediction and a minimum number of additional embedded models required to simulate the experiment (ideally none).

This situation can be formalized by introducing N_e theoretical descriptions $\mathcal{R}^i$, where the superscript is the experiment index. Each of these models depends on a set $\{\tau\}^i$ of the embedded models used for prediction, with elements $\tau^i_j = \tau_k$ for some k, and a set of additional embedded models necessary for the experiment, denoted by $\{\sigma\}^i$ with elements σ^i_j. Furthermore, each of the experiments will in general involve a different observation model $\mathcal{Y}^i$. The experiments to be used for calibration and validation can therefore be expressed:

$$0 = \mathcal{R}^i(u^i, \{\tau\}^i, \{\sigma\}^i; r^i) \quad 1 \leq i \leq N_e \tag{12}$$

$$y^i = \mathcal{Y}^i(u^i, \{\tau\}^i, \{\sigma\}^i; r^i) \quad 1 \leq i \leq N_e \tag{13}$$

where u^i denotes the state variables, y^i represents the observation data, and r^i is the vector of scenario parameters for experiment i.

Each of the prediction modeled quantities τ_k has been modeled as $\tau_{km} + \epsilon_{km}$ for use in the prediction, as discussed above, so each vector of prediction modeled quantities $\{\tau\}^i$ is associated with a vector of models $\{\tau_m\}^i$ and a vector of error models $\{\epsilon_m\}^i$. In addition, models for the experimental modeled quantities σ must be posed. For each experiment i, there is thus a vector of models $\{\sigma_m\}^i$, and each such model may have an associated error model, so there is generally a vector of error models $\{\delta_m\}^i$ for each experiment. These models will in general have parameters which must be determined from data, and their validity will need to be assessed as with the prediction models.

The final generalization arises because the state variables u^i associated with model $\mathcal{R}^i$ for the various experiments need not be the same, or consistent with the state variables u for the prediction model. For example, in a fluid dynamics problem, the prediction state variables would generally be a three-dimensional vector field, while in the model for the

viscometer in which the parameter in the constitutive model for the internal stress (the viscosity) is determined, the state variable is a one-dimension function for the azimuthal velocity. Thus while the same embedded model is applied in both the prediction and the experiment, the dependence of the embedded model on the state must be different. We can express this by defining arguments v_k for each embedded model that are consistent for the prediction and all experiments. An operator $\mathcal{V}_k^i$ is needed that maps the state variable for each scenario to the argument of the model τ_{km}.

With these generalizations, the abstract statement of the prediction problem, analogous to (9) is

$$0 = \mathcal{R}(u, \{\tau_m\}^0 + \{\epsilon_m\}^0, r) \tag{14}$$

$$q = \mathcal{Q}(u, \{\tau_m\}^0 + \{\epsilon_m\}^0, r) \tag{15}$$

$$0 = \mathcal{R}^i(u^i, \{\tau_m\}^i + \{\epsilon_m\}^i, \{\sigma_m\}^i + \{\delta_m\}^i, r^i) \quad \text{for } 1 \le i \le N_e \tag{16}$$

$$y^i = \mathcal{Y}^i(u^i, \{\tau_m\}^i + \{\epsilon_m\}^i, \{\sigma_m\}^i + \{\delta_m\}^i, r^i) \quad \text{for } 1 \le i \le N_e \tag{17}$$

where the embedded models appearing in the composite model of the prediction are

$$\tau_k \approx \tau_{km}(\mathcal{V}_k(u), \theta_k, s_k) + \epsilon_{km}(\mathcal{V}_k(u), \alpha_k, s_k), \tag{18}$$

while when the same models appear in the composite model for experiment i they have the form:

$$\tau_k \approx \tau_{km}(\mathcal{V}_k^i(u), \theta_k, s_k) + \epsilon_{km}(\mathcal{V}_k^i(u), \alpha_k, s_k). \tag{19}$$

Clearly, if the experiments are to provide any meaningful information about the prediction models, then the errors and uncertainties associated with the embedded models for $\{\sigma\}^i$ will have to be sufficiently small, or ideally entirely absent. That is, it is preferable if there are no extra modeled quantities in the composite model for the experiment. Similarly, as mentioned above, experiments in which only one of the embedded models is exercised are particularly valuable for learning about that model, since it avoids confounding uncertainties from other models. Experiments that exercise many embedded models are generally not as useful for calibrating embedded models.

This fact leads to the idea of a validation pyramid [27]. In particular, it is helpful to organize the experimental inputs to the simulation of a complex system hierarchically. At the lowest level of the hierarchy are simple experiments that exercise only one, or few embedded models. Such experiments are generally relatively inexpensive, relatively well controlled and numerous. They therefore can provide abundant well characterized data, which is ideal for the calibration of embedded models. These experiments are at the base of the pyramid.

As one ascends the hierarchy, or pyramid, the experiments exercise more of the embedded models simultaneously; they become increasingly expensive, and they commonly become more difficult to control and instrument. Data from these more complex experiments are thus more limited and are often of lower quality, that is, they have higher uncertainty. For these reasons, experiments that are higher in the hierarchy are typically less useful for calibrating embedded models. While there are notable exceptions, these experiments are generally most critical for validation testing. For example, using data from higher levels on the pyramid, one can test that inferences based on the simple experiments at the lower levels are consistent with more complex experiments involving additional phenomena. One exception is the calibration of parameters that arise due to coupling between different embedded models; such parameters may not be present at the lowest levels of the pyramid and so calibration data from higher levels may be needed. Another exception is when parameters are recalibrated based on validation data from higher in the hierarchy as part of a validation process, such as that proposed by Babuška et al. [11].

Finally, at the highest level of the pyramid are experiments conducted on systems with complexity similar or identical to the system of interest. Experiments on the system of interest are particularly valuable because they provide an opportunity to detect unanticipated and therefore unmodeled phenomena.

The generalizations discussed above are important to understanding how the predictive validation process discussed here applies to the simulation of complex systems. However, the basic ideas underlying predictive validation are more easily discussed in the context of simpler examples, which are well characterized by the simple abstract problem described in Section 2.1. Therefore, the generalizations described above will not be discussed further.

3. Predictive validation processes

Given a composite model like (9), there are likely to be parameters—e.g., θ and α in τ_m and ϵ_m, respectively, from Section 2.3—that must be determined from observations (calibration). If the inadequacy model ϵ_m faithfully

represents the discrepancies between the model for the observables $\mathcal{Y}$ and the observations (validation), we then use it in the model for the QoIs $\mathcal{Q}$ to predict how the observed discrepancies impact uncertainty in the QoI, and assess the adequacy of the calibration and validation processes for the QoI prediction (predictive assessment). Thus, the process needed to assess the accuracy and credibility of the predictions involves three activities: calibration, validation, and predictive assessment. These activities are described briefly below.

3.1. Model calibration

The parameters in the embedded models τ_m and inadequacy models ϵ_m need to be specified in some way. Some parameters may be very well known, with either known or negligible uncertainties, and such parameters need not be calibrated (e.g., the speed of light or the acceleration due to gravity on Earth). However, in most cases, at least some of the parameters are not well-known, and so values must be determined that are consistent with existing knowledge about the phenomenon and with observational data. Generally, this can be posed as an inverse problem, where the model inputs are to be determined by requiring the model outputs to match observations under constraints imposed by prior knowledge. Furthermore, the uncertainties in the data being used and the qualitative and often imprecise nature of existing knowledge about the parameters must be accounted for to yield estimates of the resulting uncertainty in the calibrated parameters.

There are a number of different approaches that might be formulated to solve such inverse problems [28]. Provided they produce consistent estimates and uncertainties for the parameters, they can serve as calibration techniques for the predictive validation approach discussed here. Recall, however, that we have selected Bayesian probability for our mathematical representation of uncertainty. In this context, a very natural and powerful approach to calibration is Bayesian inference, which relies on Bayes' theorem:

$$p(\theta, \alpha|Y, I) = \frac{p(\theta, \alpha|I)L(\theta, \alpha; Y, I)}{\int p(\theta, \alpha|I)L(\theta, \alpha; Y, I)\, d\theta d\alpha} \tag{20}$$

where Y represents the calibration data, and $p(\cdot|\cdot)$ is a conditional probability density function (PDF). Here, $p(\theta, \alpha|I)$ is the joint prior PDF of the parameters θ and α, which is conditional on the prior knowledge I. The function $L(\theta, \alpha; Y, I)$ is the likelihood, which is given by

$$L(\theta, \alpha; Y, I) = p(y = Y|\theta, \alpha, I). \tag{21}$$

It represents the likelihood of observing the data given the model and its uncertainty, the particular values of θ and α, and the uncertainty in the measurements. Finally, $p(\theta, \alpha|Y, I)$ is the posterior PDF, the joint distribution of the parameters, conditioned on both the data and the prior knowledge. This posterior PDF is the Bayesian solution of the inverse problem, representing the desired estimate of the parameters, with uncertainties.

3.2. Validation

Calibrating the model does not guarantee that its outputs will be consistent with the calibration data, much less new data not used for calibration. Indeed, calibration can only ensure that the model matches the calibration data as well as possible, which may not be very well at all. It is therefore necessary that consistency of model outputs with the experimental observations be explicitly checked. This process of comparing models to observations is consistent with the classical notion of validation. However, in making such comparisons, determining how much discrepancy between model and observations is acceptable is subtle and important. The most relevant metric is the implied discrepancy in the QoI (i.e., the difference between the predicted QoI and reality), but this metric is inaccessible in standard validation approaches. Hence, the determination of a tolerance is generally left to expert opinion [5].

However, the consideration of uncertainties provides a natural way to define the acceptable level of discrepancy. It is simply that the data must be a plausible outcome of the model, with all its uncertainties, including those due to the model inputs, observational errors, and model inadequacy. This tests whether the uncertainty models, particularly the inadequacy model, can plausibly account for the causes of the discrepancies. If any of the available data, including data used for calibration as well as that intended only for validation, is not a plausible outcome of the model with its uncertainties, then we can conclude that the uncertainty representations are somehow insufficient and should not be used for prediction.

It remains then to define what it means for the data to be a plausible outcome of the model. This clearly depends on how uncertainties are being represented mathematically. If one represents uncertainty using probability, as we do here, then the validation criterion will clearly need to be that the data is not too improbable. To quantify this we require a metric and a tolerance.

Given the probability distributions for the model inputs θ and α obtained from the calibration process, the model (9) yields a predictive joint probability density for all of the validation observables, which we denote by z. To be clear, z includes the observables used for calibration y as well as other data that are used only for validation. To assess plausibility, we must determine whether the observed validation data Z are a plausible draw from the predictive distribution $p(z|Y, I)$. In this setting, useful metrics can be based on Bayesian credible regions, of which several can be defined.

Particularly appropriate for our use are highest posterior density (HPD) credibility regions [29]. The β-HPD ($0 \leq \beta \leq 1$) credible region S is the set for which the probability of belonging to S is β and the probability density for each point in S is greater than that of points outside S. However, because HPD credibility sets are defined in terms of the probability density, they are not invariant to a change of variables. This is particularly undesirable when formulating a validation metric because it means that one's conclusions about model validity would depend on the arbitrary choice of variables (e.g., whether one considers the observable to be the frequency or the period of an oscillation). To avoid this problem, we introduce a modification of the HPD set in which the credible set is defined in terms of the probability density relative to a specified distribution q. An appropriate definition of q would be the ignorance distribution [22] on the data space. Using this definition of the highest posterior *relative* density (HPRD) credibility set, a conceptually attractive credibility metric can be defined as $\gamma = 1 - \beta_{\min}$, where $\beta_{\min}$ is the smallest value of β, for which the validation data Z is in the HPRD-credibility set for Z. That is:

$$\gamma = 1 - \int_S p(z|Y, I)\, dz, \quad \text{where } S = \left\{ z : \frac{p(z|Y, I)}{q(z)} \geq \frac{p(Z|Y, I)}{q(Z)} \right\}. \tag{22}$$

When γ is smaller than some tolerance, say less than 0.05 or 0.01, the data are considered an implausible outcome of the model—i.e., there is an inconsistency between the model and the observations.

This metric is proposed here because it seems to be best aligned with intuitive notions of agreement between the model predictions and observations, even for skewed and multi-modal distributions. For example, for multi-modal distributions, an HPRD region may consist of multiple disjoint regions [30] around the peaks in the distribution, leaving out the low probability density regions between the peaks.

Unfortunately, while this metric appears to us to be the most appropriate way to check that the validation data are a plausible outcome of the model, as a practical matter, it suffers from the significant drawback that it is difficult to compute. Specifically, when Z involves more than a few observations, the necessary computations will generally be difficult or intractable. The best case scenario involves a high-dimensional integral over the potentially complex region S. The calculation can be further complicated by the fact that the predictive density $p(z|Y, I)$ may be difficult to sample from and/or to evaluate at a given point in data space.

For this reason, in this work, we settle for applying this HPD-credibility-based metric in a marginal sense. That is, for each individual validation observable z_i, we define the credibility metric based on the marginal predictive distribution:

$$\gamma_i = 1 - \int_{S_i} p(z_i|Y, I)\, dz_i, \quad \text{where } S_i = \left\{ z_i : \frac{p(z_i|Y, I)}{q(z_i)} \geq \frac{p(Z_i|Y, I)}{q(Z_i)} \right\}, \tag{23}$$

where, to be clear, $p(z_i|Y, I)$ represents the marginal distribution for the ith component of z, which is determined by integrating $p(z|Y, I)$ over all but the ith variable.

While this modification makes the metric simpler to compute, the cost of this simplicity is that the results are more difficult to interpret. In particular, it is clear that the tolerance that one might apply in to γ from (22) should not be naively transferred to γ_i. For instance, if we have 100 independent validation data points, on average, one would expect to observe one γ_i less than 0.01, even with data that are generated by drawing from the predictive distribution—i.e., even with data that are perfectly consistent with the model. This shortcoming could be addressed using the Bonferroni correction or a number of other procedures [31,32]. It is unclear which, if any, of the existing procedures is most appropriate for validation and so we do not advocate any specific correction here.

More generally, given the difficulties associated with interpreting the γ_i metric, in cases where results are not clear cut—i.e., when the computed γ_i values are near the established tolerance—it is probably best to investigate

further rather than simply accept or reject the model. This further investigation could, for example, include HPRD credibility region computations for low-dimensional (two or three) joint predictive distributions, where the necessary computations are tractable.

Finally, even if all available data are credible outcomes of the model by an HPRD criterion, this does not imply that the model with its associated uncertainty is valid for predicting the QoIs. Indeed it only implies that we have so far failed to invalidate the model for use in making predictions. To determine the validity of a prediction, it is necessary to assess whether the validation tests have been sufficiently thorough to give confidence in the prediction, and whether, in the prediction scenario, all the embedded models are being used in regimes in which they are expected to be reliable. This is the role of predictive assessment.

3.3. Predictive assessment

As discussed in Section 1, assessing the reliability of a prediction is difficult because prediction fundamentally requires extrapolation from available data. Such an extrapolation cannot be justified based on the agreement between the data and the model alone. Knowledge about the structure of the physical and uncertainty models and their potential shortcomings is important to determining whether the validation tests that have been performed are sufficient to warrant confidence in the predictions. There are two primary questions that need to be addressed. The first is whether the QoIs are sensitive to aspects of the embedded models that have not been effectively informed by the calibration and tested through the validation process. The second is whether any of the embedded models and their accompanying inadequacy models are used outside their "domain of applicability". Answering these questions is central to assessing the credibility of the prediction process.

Because these credibility assessments rely so heavily on knowledge of the characteristics and pedigree of the embedded models, it is not possible to provide general prescriptions for performing these assessments. Instead, in the following subsections, a number of important considerations will be discussed in the context of examples.

In addition, if a prediction is determined to be credible, there is finally the question of whether the prediction is sufficiently precise; that is, whether it has sufficiently small uncertainty for the purposes for which it is being performed. Once a tolerance on the prediction uncertainty is specified, which is clearly dependent on the nature of the decisions to be informed by the predictions, this is straightforward to assess.

3.3.1. Adequacy of calibration and validation

The fundamental issue in assessing the adequacy of the calibration and validation is whether the available data inform and challenge the model in ways that are relevant to the desired prediction. This assessment is necessarily based, at least in part, on knowledge regarding the physics of the problem. For example, in many domains, arguments based on dimensional analysis can help to understand the relevance of an experiment on a scale model to the case of interest. Whenever possible, such information must be used. To augment such traditional analyses, one must consider whether QoIs are sensitive to some characteristic of an embedded model, or the associated inadequacy model, that has not been properly informed and tested in the preceding calibration and validation phases. In particular, if the QoIs are sensitive to an aspect of the model to which the data are insensitive, then the prediction depends in some important way on things that have not been constrained by the data. In this case, the prediction can only be credible if there is other reliable information that informs this aspect of the embedded models. To assess this then requires a sensitivity analysis to identify what is important about the embedded models for making the predictions. This sensitivity analysis is necessarily concerned with the sensitivities after calibration, because it is the calibrated model that is to be used for prediction. To make this generic discussion more concrete, it is useful to consider several representative cases.

3.3.1.1. Sensitivity to an embedded model. Suppose that the prediction QoI is highly sensitive to one of the embedded models τ_m, as measured, for example, by the Fréchet derivative of the QoI with respect to τ_m at some representative θ. This indicates that perturbations of τ_m greatly affect the QoI, and thus, it is important to determine whether the embedded model and its use in the composite model has been well validated. If, for example, none of the validation quantities are sensitive to τ_m, then the validation process has not provided a test of the validity of τ_m, and a prediction based on τ_m would be unreliable. However, such a clear situation is not likely, especially when τ_m involves parameters that have been calibrated, usually with scenarios low on the validation pyramid (see Section 2.4), because at least the calibration data would then be expected to be sensitive to τ_m. A more common situation would be that validation

quantities from scenarios higher on the pyramid are not sensitive to τ_m, so that the validity of using τ_m in a composite model similar to that used in the predictions has not been tested.

To see why this could be problematic, consider an embedded model τ_m for a quantity τ, which in the prediction scenario depends on one of the system state variables u, and suppose that this dependency has not been recognized so that u is not included as an argument in τ_m. In the simplified scenarios near the bottom of the validation pyramid that are used to calibrate and validate τ_m, this dependence might be absent or unimportant. So this data could not be used to detect the missing dependence. If the validation quantities in tests that are higher on the pyramid, in which the omitted dependence is important, are insensitive to τ_m, then these tests will also not detect the omission. To guard against this and similar possible failures of τ_m, the predictive assessment process should determine whether validation quantities in scenarios "close enough" to the prediction scenario are sufficiently sensitive to τ_m to provide a good test of its use in the prediction. The determination of what is "close enough" and what constitutes sufficient sensitivity must be made based on knowledge of the model and the approximations that went into it, and of the way the models are embedded into the composite models of the validation and prediction scenarios. If there are no data for sufficiently sensitive quantities on close enough scenarios, then the resulting predictions would be unreliable, unless one had independent information that the model is trustworthy for the prediction.

3.3.1.2. Sensitivity to a model parameter. Suppose that the prediction QoI is highly sensitive to the value of a particular parameter θ in an embedded model. In this case, it is important to determine whether the value of this parameter is well constrained by reliable information. If, for example, none of the calibration data has informed the value of θ, then only other available information (prior information in the Bayesian context) has determined its value. Further, if none of the validation quantities are sensitive to the value of θ, then the validation process has not tested whether the information used to determine θ is in fact valid in the current context. The prediction QoI is then being determined to a significant extent by the untested prior information used to determine θ. This should leave us little confidence in the prediction, unless the prior information is itself highly reliable (e.g., θ is the speed of light). Alternatively, when the available prior information is questionable (e.g., θ is the reaction rate of a poorly understood chemical reaction), the predictions based on θ will not be reliable.

Alternatively, the calibration data could have been highly informative of the value of θ during the calibration process and the validation quantities could have also been very sensitive to the value of θ. This would suggest that the value of the QoI is being substantially determined by the calibration and validation data through the sensitive parameter θ. Provided the data are reliable, the determination of θ will not cause the prediction to be unreliable.

3.3.1.3. Sensitivity to an inadequacy model. Suppose that uncertainty in the prediction QoI is largely due to the uncertainty model ϵ_m representing the inadequacy of the embedded model τ_m. In this case, it is important to ensure that ϵ_m is a valid description of the inadequacy of τ_m. As with the embedded model sensitivities discussed above, validation tests from high on the validation pyramid are most valuable for assessing whether the uncertainty model represents inadequacy in the context of a composite model similar to that for the prediction. If however, the available validation data are for quantities that are insensitive to ϵ_m, then the veracity of ϵ_m in representing the uncertainty in the QoI will be suspect. Reliable predictions will then be possible only if there is independent information that the inadequacy representation is trustworthy.

Like the sensitivity to the embedded physical model, the sensitivity to an inadequacy model can be measured locally by the Fréchet derivative of the QoI with respect to ϵ_m at a particular realization of ϵ_m. When ϵ_m is highly constrained by the data, this local sensitivity may be sufficient, but in general a more global quantity is desired. To get a more global view, one may examine the statistics of the derivatives induced by the posterior distribution for ϵ_m and its hyperparameters and/or other variance-based [33] and information-theoretic [34] methods.

3.3.2. Domain of applicability of embedded models

In general, it is expected that the embedded models making up the composite model to be used in a prediction will involve various approximations and/or will have been informed by a limited set of calibration data. This will limit the range of scenarios for which the model can be considered reliable, either because the approximations will become invalid or because the model will be used outside the range for which it was calibrated. It is therefore clearly necessary to ensure that the embedded models are being used in a scenario regime in which they are expected to be reliable.

As discussed in Section 2.1, reliable extrapolative predictions are possible because the scenario parameters relevant to an embedded model need not be the same as those for the global composite model in which it is embedded.

For example, when modeling the structural response of a building, the scenario parameters include the structural configuration and the loads. However, the scenario parameters for the linear elasticity embedded model used for the internal stresses would be the local magnitude of the strain, as well as other local variables such as the temperature. For each embedded model then, we need to identify the scenario parameters that characterize the applicability of the model and the range of those parameters over which the model and its calibration is expected to be reliable. It is then a simple matter of checking the solution of the composite model to see if any of the embedded models are being used "out of range". For some embedded models, defining the range of applicability in this way is straightforward. However, for some types of embedded models—e.g., an embedded model that involves an additional equation that has non-local dependence on the state—defining the relevant scenario space and, hence, the region of scenario space that defines the domain of applicability, is significantly more difficult.

3.3.3. Other issues

There are two other issues that must be considered when performing a predictive assessment and interpreting the results of a prediction. First, the focus of the discussion to now has been on ensuring that the calibration and validation processes have been sufficiently rigorous to warrant confidence in an extrapolative prediction and its uncertainty. However, a prediction with an uncertainty that is too large to inform the decision for which the prediction is being performed is not sufficient, even if that uncertainty has been determined to be a good representation of what can be predicted about the QoI. The requirements for prediction uncertainty to inform a decision based on the prediction depend on the nature of the decision, and determination of this requirement is outside the scope of the current discussion. However, once such a requirement is known, the prediction uncertainties can be checked to determine whether these requirements are met, and therefore whether the prediction is useful. Of course, when the prediction uncertainty fails to meet the established tolerance, some action must be taken to reduce this uncertainty. While a full discussion of this process is beyond the scope of this work, we mention that the predictive validation activities previously described provide a wealth of information that can provide guidance as to how to proceed. For example, parameters that have large posterior uncertainty and that are influential to the QoIs are good candidates for further calibration based on new experiments. Alternatively, embedded models for which the associated inadequacy model introduces significant uncertainty are good candidates for new model development.

The second issue is the well-known problem of "unknown unknowns". If the system being simulated involves an unrecognized phenomenon, then clearly an embedded model to represent it will not be included in the composite model for the system. As with the examples above, the prediction QoI could be particularly sensitive to this phenomenon, while the validation observables are not sensitive. In this situation, one would not be able to detect that anything is missing from the composite model. Further one could not even identify that the validation observables were insufficient; that is, the predictive assessment could not detect the inadequacy of the validation process. This is a special case of a broader issue. The predictive validation process developed here relies explicitly on reliable knowledge about the system and the models used to represent it. This knowledge is considered to not need independent validation, and is thus what allows for extrapolative predictions. However, if this externally supplied information is in fact incorrect, then the predictive validation process may not be able to detect it.

4. Illustrative example

To illustrate some aspects of the predictive validation process, we apply it here to a simple problem involving a spring–mass–damper system. For this system, Newton's second law requires that

$$m\ddot{x} = f_d + f_s, \tag{24}$$

where m is the mass, x is the position of the mass, and f_d and f_s are the forces acting on the mass due to the damper and the spring, respectively. For this example, other forces acting on the mass (e.g., drag, gravity) are known to have negligible effect. Thus, (24) represents highly reliable theory in the context of the current problem. However, f_d and f_s must be specified—i.e., potentially inadequate embedded models are required for these forces. For the purposes of this example, a truth system, detailed in the Appendix, is used to make simulated experimental observations but is otherwise unknown to the modeler. The goal of the predictive validation exercise is to evaluate the use of simple linear models for f_d and f_s to predict the QoI, which is taken to be the maximum velocity of the mass for $m = 5$ with an initial condition of $x(0) = 4$, $\dot{x}(0) = 0$.

Execution of the predictive validation process depends on our "state of knowledge"—that is, the available information about the system. The information used here is described in Section 4.1. In the current exercise, three different states of knowledge are considered, which lead to different results. In particular, the different states of information lead to different uncertainty representations, which are discussed in Section 4.2, as well as different validation procedures and conclusions, which are described in Section 4.3.

4.1. Available information

In general, the information available in the prediction process is of several types. In the present problem, the available information consists of the high fidelity model (24), the composite physical model which will be specified below, information regarding the inadequacies of the embedded models for the forces, and observational data.

4.1.1. The composite model

The standard mathematical representations for both springs and dampers are linear, and since no information is available to suggest specific nonlinear models, linear models will be used here. This is clearly a modeling assumption, and must be considered provisional. The embedded models are thus

$$f_s = -kx,$$
$$f_d = -c\dot{x},$$

where the constants k and c are unknown. With these embedded models, the composite model is

$$m\ddot{x} + c\dot{x} + kx = 0, \qquad (25)$$
$$x(0) = 4, \quad \dot{x}(0) = 0.$$

Note that the composite model (25) is not the same as the truth model described in the Appendix. Thus, it is necessary to consider model inadequacy in the predictive validation exercise. Further, information about the model inadequacy that is known independently of the observational data that will be used for calibration and validation will be necessary.

To demonstrate how predictive validation depends on the available information, we postulate three different states of knowledge regarding the inadequacy of the model. These are labeled States of Information 0, 1, and 2 or SI0, SI1, and SI2 for short. In SI0, we are confident that physical model (25) is sufficient for the required prediction—i.e., that the embedded models are highly accurate and that there are no important neglected physics. However, the parameter values k and c are not well-known and must be calibrated from the observational data.

In SI1, it is known that all important forces are represented and that the linear spring model is highly accurate, while the constant coefficient damping model is suspected to be inadequate. However, no information is available about why the damping model might fail to represent reality.

Finally, in SI2 (as in SI1), the linear spring model is known to be adequate, and the damping model is known to be problematic. Unlike SI1, we have information regarding the cause of the inadequacy. It was noticed that the damper becomes warm when it moves, presumably because of the energy that is being dissipated by the damper. Because the viscosity of the damping fluid in the damper can reasonably be assumed to depend on temperature, this would result in a change in the damping coefficient with time. While the precise form of the model inadequacy—i.e., the temperature variation of the damping fluid or the temperature dependence of the damping coefficient—is not known, this additional information will be essential in building confidence in extrapolative predictions.

4.1.2. Observational data

To calibrate and validate the composite model, observations of the position of the mass at discrete times for two different masses, $m = 1$ and $m = 2$, are available. These observations are given in Tables 1 and 2. The tables give both the actual position, as determined by solving the truth system (see Appendix) using Runge–Kutta–Fehlberg (4, 5) time marching, and the data used here, which are perturbed by simulated observation noise. The observation noise is such that the ith observed value $\hat{x}_i$ is given by

$$\hat{x}_i = x_i + \epsilon_i,$$

where x_i is the actual position and ϵ_i are independent, identically distributed Gaussian random variables with mean zero and standard deviation 0.01.

Table 1
Observations of the actual system for $m = 1$.

Time (t)	Position (x)	Observation ($\hat{x} = x + \epsilon$)
0.0	4.0	4.0
1.0	4.025647×10^{-1}	4.056287×10^{-1}
2.0	-1.913556×10^{0}	-1.917800×10^{0}
3.0	7.536144×10^{-2}	7.331597×10^{-2}
4.0	8.219699×10^{-1}	8.176825×10^{-1}
5.0	-1.260000×10^{-1}	-1.129453×10^{-1}
6.0	-3.076154×10^{-1}	-3.011407×10^{-1}
7.0	8.109402×10^{-2}	9.303637×10^{-2}
8.0	1.062484×10^{-1}	8.884368×10^{-2}

Table 2
Observations of the actual system for $m = 2$.

Time (t)	Position (x)	Observation ($\hat{x} = x + \epsilon$)
0.0	4.0	4.0
1.0	1.718579	1.705548
2.0	−1.641053	−1.634634
3.0	−2.121425	−2.127946
4.0	-1.641898×10^{-1}	-1.818642×10^{-1}
5.0	1.278992	1.269814
6.0	8.442413×10^{-1}	8.507498×10^{-1}
7.0	-3.168699×10^{-1}	-3.259694×10^{-1}
8.0	-7.066765×10^{-1}	-7.080865×10^{-1}

4.2. Uncertainty representations for calibration and prediction

The formulation of the uncertainty representation depends on the state of knowledge. Here we develop two different models: one corresponding to SI0 and one corresponding to SI1 and SI2.

In SI0, we are confident that the physical model is adequate. Thus, no model inadequacy representation is required. However, the values of k and c are unknown. Thus, following the standard Bayesian approach, k and c are treated as random variables, which will be calibrated using Bayesian inference, using the data given in Table 1. For this, a joint prior probability of k and c must be specified, and for simplicity they are taken to be independent with $k \sim N(1, 1/2)$ and $c \sim \log N(0, 1/2)$, where $N(\mu, \sigma^2)$ denotes the Gaussian with mean μ and variance σ^2, and $\log N(\mu, \sigma^2)$ denotes the corresponding log-normal.

Since the composite model is assumed to be perfect except for the parameter uncertainty, the likelihood function is based on the experimental uncertainty alone. Thus,

$$L(k, c; \hat{x}_1, \ldots, \hat{x}_N) = \prod_{i=1}^{N} \ell(k, c; \hat{x}_i),$$

where ℓ is the likelihood for a single data point. Specifically,

$$\ell(k, c; \hat{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(\hat{x}_i - x_m(t_i; k, c))^2}{\sigma^2}\right], \tag{26}$$

where $\sigma = 0.01$ is the standard deviation of the observation error ϵ_i and $x_m(t_i; k, c)$ is the solution of (25) at t_i for parameters k and c.

In both SI1 and SI2, the damping model is known to be inadequate. There are many possible ways one might represent this inadequacy. Here we adopt a very simple model designed to represent the variability required in the damping coefficient to reproduce the departure from constant coefficient behavior. In particular, the damping coefficient is modeled as a random variable with the randomness intended to describe the variability in the damping

needed to encompass the data, not only a lack of knowledge about the true value of c. Since the damper coefficient c must be positive, the variability of the damper is modeled by a log-normal distribution: $c \sim \log N(c_\mu, c_\sigma^2)$. The parameters c_μ and c_σ are uncertain and thus must be learned from the calibration data. Note, however, that this calibration is fundamentally different from that pursued for SI0. For SI0, the assumption is that there is a unique true value of c that is to be determined. In this situation, the posterior PDF for c characterizes uncertainty about this best value due to having only a finite amount of uncertain data. As the number of independent data points increases, the posterior PDF for c narrows, and, in the limit of infinite data points, converges to a δ distribution. Alternatively, for SI1 and SI2, the goal is to find both the values of c that are most likely and how much "variation" is necessary to cover the data. Even in the limit of infinite data, c will be characterized by a log-normal distribution with non-zero variance.

Of course, one could pose more complex models that might better describe the actual inadequacy of the embedded damping model. The development of uncertainty representations is highly problem dependent, and developing a more complex model here does not further the goal of illustrating the process. Thus, we consider only this simple inadequacy model.

In the prior, the parameters k, c_μ, and c_σ are taken to be independent, and the following marginal prior distributions are chosen:

$$
\begin{aligned}
p(k) &= \log N(1, 0.5),\\
p(c_\mu) &= N(0, 0.5),\\
p(c_\sigma) &= \log N(-1, 0.4).
\end{aligned}
$$

To form the likelihood, each data point is considered an independent draw from the random model of the damping coefficient. Thus, the likelihood for each data point is given by

$$
\hat{\ell}(k, c_\mu, c_\sigma; \hat{x}_i) = \int_0^\infty \ell(k, c; \hat{x}_i)\, p(c|c_\mu, c_\sigma)\, dc,
$$

where $p(c|c_\mu, c_\sigma)$ is the log-normal PDF and $\ell(k, c; \hat{x}_i)$ is as given in (26). Since the data points are independent, the full likelihood is given by

$$
L(k, c_\mu, c_\sigma; \hat{x}_1, \ldots, \hat{x}_n) = \prod_{i=1}^{n} \hat{\ell}(k, c_\mu, c_\sigma; \hat{x}_i).
$$

4.3. Numerical results

Predictive validation is necessarily pursued in the context of our state of information, so each different state of information is considered separately below. In each case, the actions taken as part of the predictive validation process and the results observed are described.

4.3.1. State of information 0

Fig. 1 shows results of calibrating the values of k and c (with no model inadequacy representation) using Bayesian inference and the data for the $m = 1$ case given in Table 1. Clearly, the marginal posterior PDFs are very narrow, indicating that k and c are highly constrained by the data. The maximum a posteriori (MAP) value of k is roughly 2.95, which is close the true value of 3.0. The MAP value of c is approximately 0.8. There is no "true" value of c since the true damping coefficient is changing in time. However, this value is within the range of true values observed for the actual system for $m = 1$.

The first validation challenge is to simply compare the output of the calibrated model with the calibration data, as shown in Fig. 2. The figure shows the comparison in a number of different ways. Fig. 2a shows the interval corresponding to the 5th and 95th percentiles according to both the model prediction and the observation plus its uncertainty. However, the errors are quite small relative to the largest values of x, making it difficult to assess the results. Fig. 2b shows the HPRD credibility γ_i (23) for each data point, based on the marginal distribution given by the model plus the observational uncertainty, and measured relative to a uniform distribution. To be clear, the model of the observation is given by

$$
x_{\text{obs}} = x_m + \epsilon,
$$

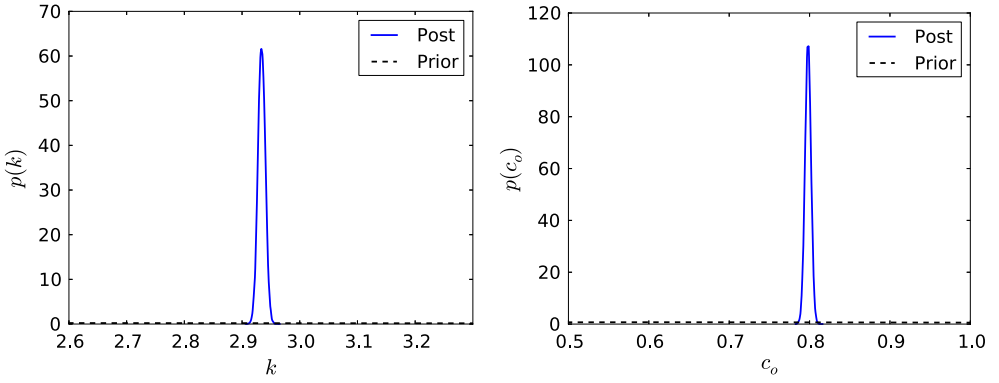
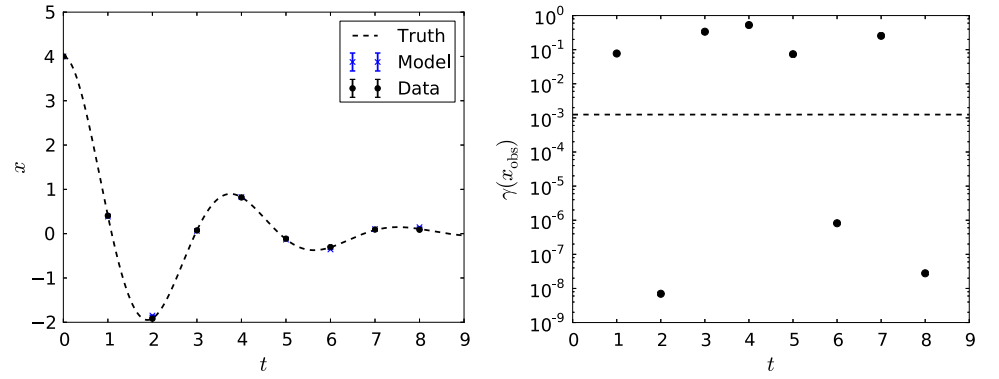
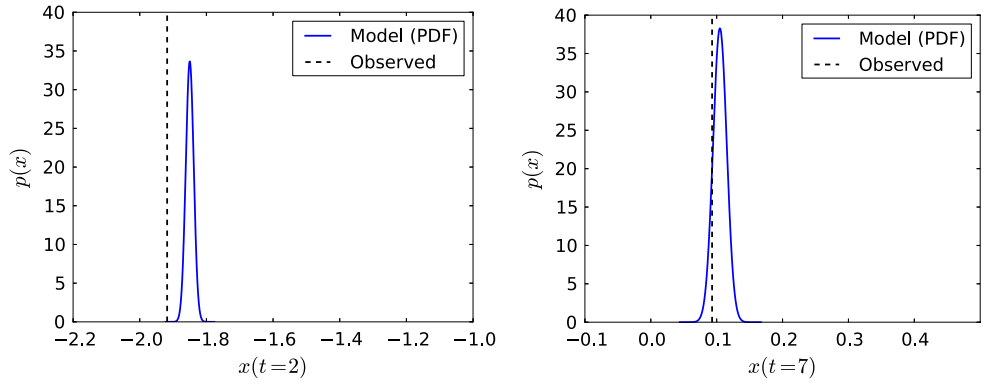


Fig. 1. Marginal distributions for parameters k and c. Shown are the posterior PDF resulting from a Bayesian calibration using the $m = 1$ data in Table 1 (solid), and the prior (dashed).



(a) All calibration data.

(b) HPRD credibility γ at x_{obs} (23).



(c) $t = 2$.

(d) $t = 7$.

Fig. 2. Comparison of output of calibrated model and observations for $m = 1$ for SI0.

where the distribution for x_m is given by forward propagation of the joint posterior distribution for k and c, and $\epsilon \sim N(0, 0.01)$ is the observational noise. Fig. 2b shows that for several of the observed data points, particularly those at $t = 2$, 6, and 8, the corresponding γ_i is less than or equal to approximately 10^{-6}, indicating that the actual observation is far out on the tail of the predictive distribution. The dashed horizontal line in Fig. 2b corresponds to $\gamma_i = 0.00125$, which is computed from the desired tolerance of $\gamma = 0.01$ using the Bonferroni correction. However, given the extreme nature of the results, the conclusion is clear for any reasonable tolerance. This is further illustrated

by Fig. 2c, which shows the marginal predictive distribution and the observation for $t = 2$. Alternatively, for $t = 7$, where γ_i is significantly larger, there is much better agreement between the predictive distribution and the observation, as shown in Fig. 2d.

However, the existence of multiple points where the actual observation gives a γ_i that is nearly zero shows that the model developed based on SI0 does not lead to plausible predictions even for the same data that was used to calibrate the model. In this situation, one must conclude that the uncertainty representation is unable to explain the differences between the physical model and the data. This fact contradicts the modeling assertion that the only important uncertainty is that due to measurement error. Since we cannot explain the observed discrepancies, we cannot confidently extrapolate to the prediction scenario using this model, even though the actual errors are not necessarily large. Despite the fact that the observed errors are not large, there is no way to characterize what the expected errors in the prediction scenario might be. Thus, the model is invalid for extrapolative prediction. Note that only the combination of the physical model and its uncertainty representation have been invalidated. It may be possible to improve either to obtain a model valid for use in the prediction. We will see that the same physical model, when equipped with a better uncertainty description, can make valid predictions.

4.3.2. State of information 1

SI1 does not provide enough information to allow extrapolation, regardless of the calibration and validation results. The fundamental issue here is that the nature of the model inadequacy is not sufficiently well understood to allow an assessment of its domain of applicability. There is therefore no way to determine the relationship between the observed model inadequacy at $m = 1$ and $m = 2$, and the model inadequacy at the prediction scenario $m = 5$. Specifically, if we are truly ignorant of the mechanism causing the inadequacy, we cannot identify the relevant model-specific scenario parameters and determine whether changing m moves the prediction outside of the parameter range that has been observed in the calibration and validation data. While we could formulate a statistical model based on the data for $m = 1$ and $m = 2$, it could only be confidently applied to these cases. This result highlights that even incomplete and/or qualitative information regarding the model and its inadequacy can be important in developing a reliable basis for extrapolation.

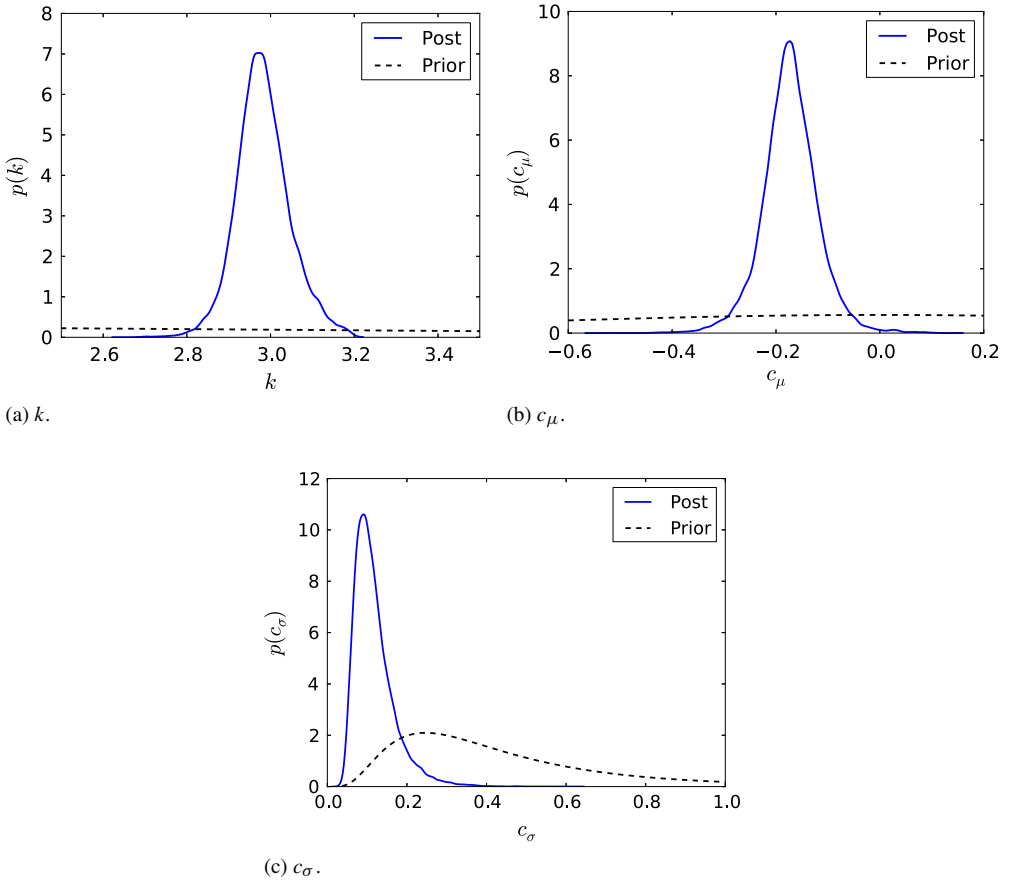
4.3.3. State of information 2

For SI2, at least there is enough information to form a hypothesis regarding the relevant model-specific scenario parameter characterizing the domain of applicability of the damper model and its inadequacy model. Based on the observation that the damper gets warm, and the expectation that the damper fluid viscosity is temperature dependent, we hypothesize that the model inadequacy is determined by the rate at which energy is dissipated by the damper. That is, the dissipation rate is the model-specific scenario parameter relevant to the inadequacy of the damper model. To arrive at this hypothesis we assumed that the temperature of the damping fluid is governed by a competition between how quickly energy is added to the fluid (by dissipation from the system) and how quickly heat is transferred to the surroundings. In this case, one would expect less temperature variation, and hence less variation in the damping coefficient, if energy is dissipated from the system more slowly.

Knowing that the energy dissipation rate is given by cv^2 where v is the velocity of the mass, a simple dimensional analysis shows that the energy dissipation rate $\dot{e}$ scales with global scenario parameters as $\dot{e} \sim ckx_0^2/m$, where x_0 is the initial displacement. Since the initial displacement, damper and spring are the same in all the scenarios considered here, the dissipation rate scales with $1/m$, so that the range over which the dissipation rate varies over time reduces with increasing m. We could also evaluate the rate of dissipation within the damper model to determine whether it varies outside the range over which it has been calibrated and validated (i.e. in $m = 1$ and 2 cases). But in this simple problem, the above dimensional analysis makes this unnecessary.

If the hypothesis is correct, then the model inadequacy representation for c formulated based on data for $m = 1$ and $m = 2$ will be conservative when used for predictions at larger masses. That is, the predicted uncertainty at higher mass should be larger than necessary to include the truth because there should be less variation of the damping coefficient. Thus, if we are satisfied with conservative uncertainty estimates, the domain of applicability of the model calibrated and validated with $m = 1$ and $m = 2$ data extends to larger mass and can therefore be used for the desired prediction.

Of course, this is a qualitative argument, and any number of phenomena could be present in the real system that would make it invalid. Thus, we must assess this hypothesis—i.e., validate it—using observational data.



(a) k. (b) c_μ.

(c) c_σ.

Fig. 3. Marginal distributions for parameters k, c_μ, and c_σ. The solid line labeled "Post" shows the marginal posterior PDF resulting from a Bayesian update of the prior (dashed line labeled "Prior") using the $m = 1$ data set (Table 1).

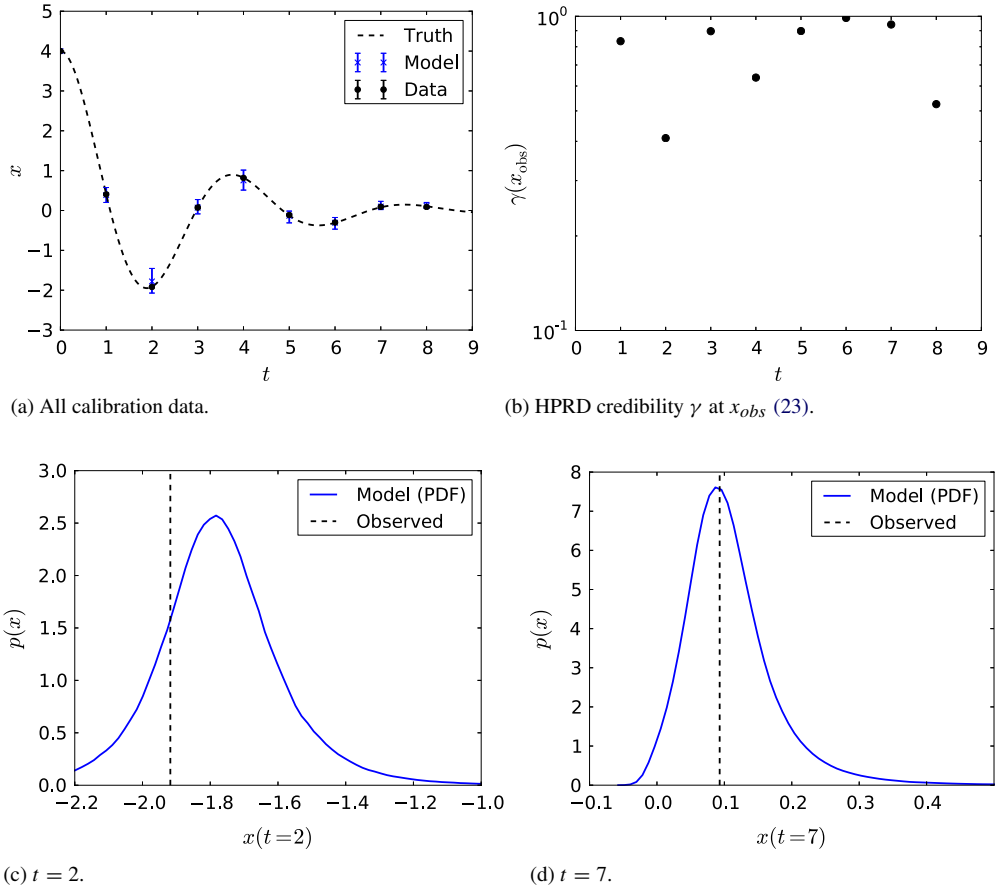
After calibration, the validation and predictive assessment processes described below will focus on assessing this hypothesis.

Fig. 3 shows the marginal PDFs resulting from the Bayesian calibration using the $m = 1$ data. Note that the marginal posterior for k is broader than in the SI0 result and that the (unknown) true value $k = 3$ is in the support of the posterior PDF.

In the validation phase, the goal is to check that the model, is capable of representing both the $m = 1$ data, with which it was calibrated, and the $m = 2$ data. If any data point is highly unlikely according to the model, it is declared invalid, as happened with the SI0 model when it failed to reproduce the calibration data.

Fig. 4 shows a comparison between the calibration data and the output of the calibrated model. The quantities shown in Fig. 4 are analogous to those shown for SI0 in Fig. 2, but unlike SI0, the predictions given are much more uncertain and agree better with the observations. In particular, the HPRD credibility, γ_i, shown in 4b are never less than approximately 0.4, indicating that the observations are near the highest probability density regions of the prediction distribution, as shown in Fig. 4c and d. The same statement holds for the $m = 2$ data, which was not used in calibration, as shown in Fig. 5. Since there is no evidence to invalidate the model after comparing the calibrated model predictions with all the available data, the validation phase is complete, and we may move on to the predictive assessment.

To begin, recall that the predictive assessment is necessarily problem specific. In the current context, a reliable prediction is dependent on the hypothesis that the energy dissipation rate defines the domain of applicability of the inadequacy model, as discussed above. This hypothesis implies that the variability of the damping coefficient will decrease with increasing mass. As part of predictive assessment, it is necessary to test this hypothesis to the fullest extent possible. Here, we test this hypothesis by performing a separate model calibration using the original prior PDF and the $m = 2$ data set. If the uncertainty in c required to fit the $m = 2$ data is smaller than that required to fit the



(a) All calibration data.

(b) HPRD credibility γ at x_{obs} (23).

(c) $t = 2$.

(d) $t = 7$.

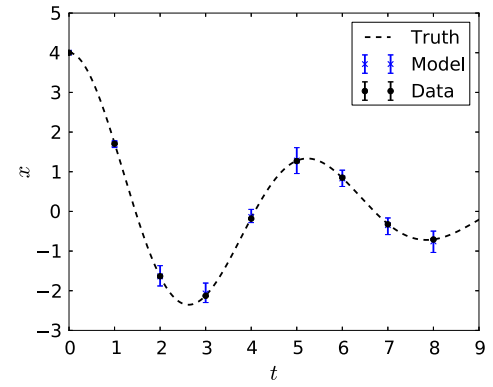
Fig. 4. Comparison of output of calibrated model and observations for $m = 1$ for SI2.

$m = 1$ data, as measured by the posterior results, then c must be varied less to match the larger mass data, and the available data supports the hypothesis.

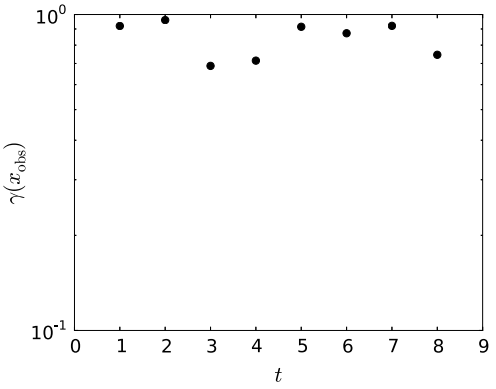
Fig. 6 shows the results of this test. The marginal posterior PDFs for k and c_μ based on the two data sets are largely consistent, although the $m = 2$ data somewhat better informs the parameters. However, c_σ, which is the standard deviation of $\log(c)$, moves significantly to the left. This result implies that the standard deviation of c decreases. This fact is demonstrated in Fig. 7, which shows the distribution of c corresponding to the maximum likelihood values of c_μ and c_σ as well as the PDF for the standard deviation of c implied by the posterior joint distribution of c_μ and c_σ. Clearly the maximum likelihood distribution of c is narrower when using the data from the larger mass. In particular, the standard deviation of the maximum likelihood model decreases from approximately 0.041 to 0.023, a decrease of nearly 44%. The distribution for σ (the standard deviation of c) shows a similar result. When using the $m = 2$ data, the probability distribution is shifted to lower values than when using the $m = 1$ data, indicating that the variability of c required to fit the data is decreasing with m. This result is consistent with the hypothesis.

Given this result, we can move to additional predictive assessments. In general, there may be many additional hypotheses to test or sensitivities to check, as discussed in Section 3.3. In this case we note that the data are quite informative about the parameters of the SI2 model (see Fig. 3), indicating that there are no uninformed aspects of the model to which the predictions could be sensitive. Further, the "domain of applicability" of the model is implicitly checked (to the extent possible with the data) by the results of the calibration with the $m = 2$ described previously. Thus, we conclude that there is good reason to trust the calibrated SI2 model to make credible predictions for the $m = 5$ case.

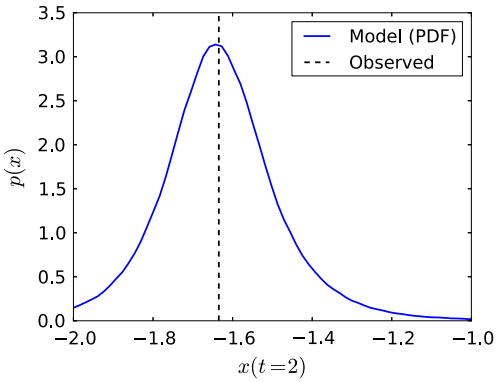
Having challenged the model and the hypothesis needed to make extrapolation to larger masses possible, we are ready to make the prediction. Recall that the QoI is the maximum velocity for $m = 5$. Fig. 8 shows the prediction



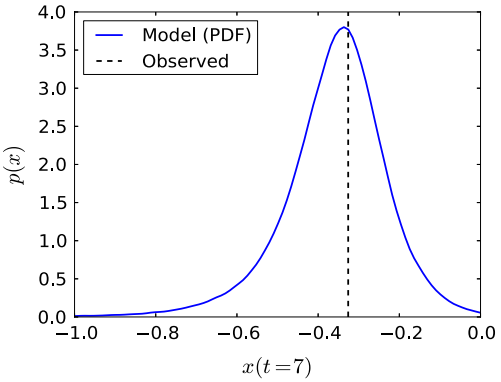
(a) All calibration data.



(b) HPRD credibility γ at x_{obs} (23).



(c) $t = 2$.



(d) $t = 7$.

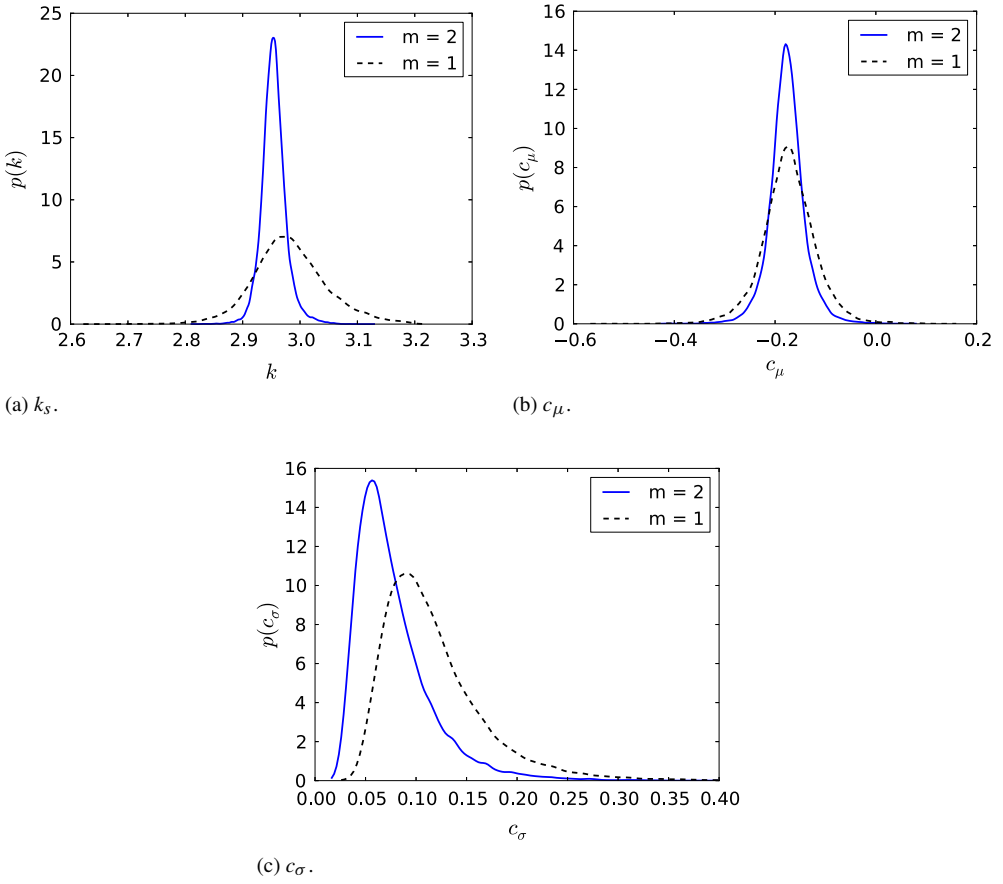
Fig. 5. Comparison of output of calibrated model and observations for $m = 2$ for SI2.

given by the model as well as the true value. Of course, in general the true value is unknown. It is given here just to show that the process has led us to the correct conclusion—i.e., the extrapolation is valid in the sense that the truth is assigned reasonable likelihood by the model.

Given this validated prediction, the next step would be to ask whether the predicted uncertainty is small enough to inform the desired decision. Since any tolerance we could specify here would be entirely contrived and artificial, we choose not to pursue this aspect. However, this is the simplest aspect of the predictive validation framework and should not cause significant difficulty. If the uncertainty is deemed too large, one would have to pose a better physical model, pose better uncertainty representations, get more data, or some combination of these, and begin the validation process again.

5. Conclusions

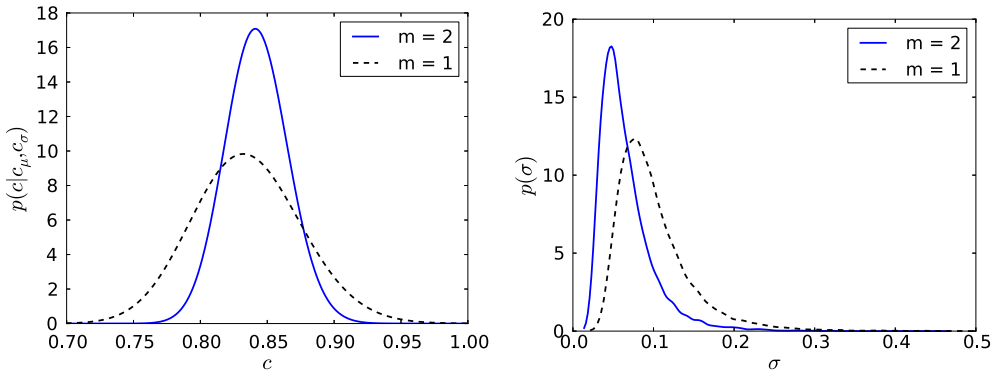
The predictive validation process proposed here provides a framework for building confidence in extrapolative predictions issued by models based on physics. There are two key ingredients enabling reliable extrapolation with such models. First, it is common that such models are based upon highly-reliable theory that is augmented with less-reliable embedded models to form a composite model. Second, the scenario dependence of the embedded models is generally different from the scenario of the full composite model, allowing the full model to be used for extrapolation without extrapolating the lower fidelity aspects. Given these ingredients, the predictive validation process requires the specification of a model inadequacy representation for the low-fidelity embedded models. This representation enables one to connect the QoI with observational data in a way that previous approaches lack. Once a physical model and inadequacy representation have been specified, the model is subjected to a calibration phase, where observations are used to inform uncertain aspects of the model. Then, in validation, the model is challenged with new observational



(a) k_s. (b) c_μ.

(c) c_σ.

Fig. 6. Comparison of marginal posterior PDFs obtained from a Bayesian update using the $m = 2$ data (solid lines) versus the $m = 1$ data (dashed lines).



(a) Maximum likelihood result. (b) Standard deviation of c.

Fig. 7. Comparison of maximum likelihood log-normal distribution of c and the PDF for the standard deviation of c for the model calibrated using both the $m = 1$ (dashed) and $m = 2$ (solid) data.

data. The primary question in this validation test is whether the observations are plausible given the uncertainty in the prediction. That is, the main objective of the validation is to test the combination of physical and uncertainty models. Finally, if the validation is satisfactory, the model is subjected to a predictive assessment to determine if the predictions of the QoIs can be trusted and whether they are sufficient from the point of view of a decision maker.

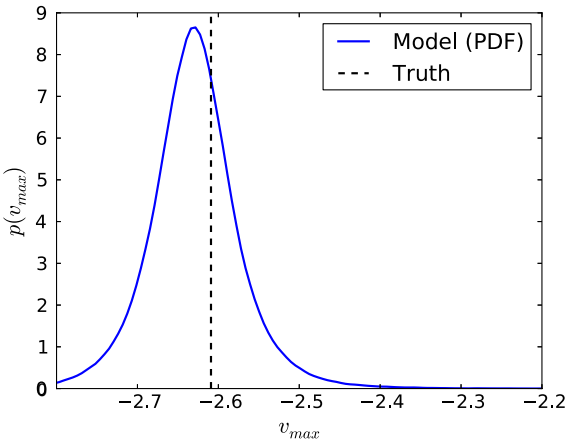


Fig. 8. Prediction of maximum velocity for $m = 5$.

The full process has been illustrated using a simple spring–mass–damper system where the true physics of the damper are not well-understood by the modeler. Despite this lack of knowledge, the process is able to build confidence in a very simple model. However, this gain in confidence is contingent on the information available at the beginning of the process. With some specific knowledge regarding the cause of the modeling error, one may be able to build confidence in an inadequate model for making a particular prediction. Without this kind of information—i.e., simply knowing that the model does not match reality and nothing more—one will generally be able to do very little.

While the global process presented here is clear, there are several research and development issues that need to be addressed to enable the validation of such predictions in a wide range of applications. These research challenges are outlined briefly below.

1. **Inadequacy models:** A critical component of the proposed process is a probabilistic model of the errors in the embedded models. Such an inadequacy model should respect all that is known about the approximations and deficiencies of the models, all that is known about the quantities being modeled, and the available data. Broadly applicable techniques for formulating these inadequacy models are needed, especially for situations where the modeled quantity is a field.
2. **Data uncertainty models:** The uncertainty in experimental data is a critical input to the process, and better characterizations of this uncertainty are needed. Of particular concern are characterizing dependencies among different data points and uncertainties arising from data reduction modeling.
3. **Representing qualitative information:** In Bayesian analysis, posing priors that faithfully represent what is known about the problem at hand is important to making reliable inference. Once the prior knowledge is expressed mathematically, one often has rigorous tools, e.g., maximum entropy, to construct the needed prior distribution. But, this knowledge is commonly qualitative and difficult to express mathematically. This process of formalizing such qualitative information is often referred to as prior elicitation or expert elicitation, and a significant research literature exists in this domain. See, for example, [35–38]. Nonetheless, more work is needed, particularly to help formalize the kinds of qualitative knowledge we commonly have regarding physical models based on reliable theory, as discussed here. Representations of qualitative information are also important in characterizing modeling inadequacy and data uncertainty.
4. **Domains of applicability:** It is critical to predictive validation to identify when an embedded model is being used under conditions for which it has not been calibrated and validated. For many models, an appropriate set of model-specific scenario parameters has not been defined. Determining such scenario parameters is part of physical modeling, and therefore dependent on the phenomena being modeled, and it is in general a significant challenge. However, generally applicable tools and techniques for developing and evaluating such parameterizations are needed.
5. **Experimental design:** Data is needed for calibration and validation, but it is critical to have data that adequately informs the QoIs in the context of the models. That is, measurements of quantities that are sensitive to the same uncertainties as the QoIs are needed under scenarios that will produce a sufficiently large domain of applicability

for the embedded models. Metrics are needed to rank potential validation cases, allowing the best experimental measurements and scenarios to be determined automatically.

6. **Computational algorithms:** While we have not discussed the computational tools needed to execute the predictive validation process discussed here, there are significant algorithmic challenges associated with high dimensional probability spaces (the curse of dimensionality), with expensive computational models and with stochastic models, which arise naturally from the inadequacy representations discussed here.

As should be clear in the above discussion, research challenges 1–4 essentially require introducing knowledge about the physical phenomena being modeled into the process. Advancing techniques to address these challenges will presumably require pursuing them in the context of a variety of specific physical systems.

Acknowledgments

This material is based on work supported by the Department of Energy [National Nuclear Security Administration] under Award Number [DE-FC52-08NA28615]. The authors are grateful to Profs. Tinsley Oden and Ivo Babuška, and Dr. David Higdon for many insightful discussions.

Appendix. Truth system

The truth system is similar to the physical model discussed in Section 4. However, instead of having a constant damping coefficient, the damping coefficient is a function of the temperature of the damper fluid. This situation is inspired by a fluid damper where the viscosity of the damper fluid, and hence the damping coefficient, varies with temperature. Here, the temperature of the fluid damper is determined by an ODE that includes the effect of energy dissipated by the damper and heat transferred from the damper to the surroundings, which are assumed to have constant temperature.

The equations are as follows:

$$m\ddot{x} + c(T)\dot{x} + kx = 0$$
$$\dot{T} = c(T)\dot{x}^2 - \frac{1}{\tau}(T - T_0)$$

where k, T_0, and τ are constants and

$$c(T) = \exp\left(\frac{T_0}{T} - 1\right).$$

For all cases, the constants are set to the following values:

$$k = 3, \quad T_0 = 20, \quad \tau = 1.$$

References

[1] AIAA Computational Fluid Dynamics Committee on Standards, AIAA Guide for Verification and Validation of Computational Fluid Dynamics Simulations. AIAA Paper number G-077-1999, 1998.
[2] ASME Committee V&V 10, Standard for Verification and Validation in Computational Solid Mechanics. ASME, 2006.
[3] ASME Committee V&V 20, Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer, ASME, 2009.
[4] P.J. Roache, Fundamentals of Verification and Validation, Hermosa Publishers, Albuquerque, 2009.
[5] W. Oberkampf, C. Roy, Verification and Validation in Scientific Computing, Cambridge University Press, 2010.
[6] M. Ainsworth, J. Oden, A Posteriori Error Estimation in Finite Element Analysis, John Wiley and Sons, Inc, 2000.
[7] S. Prudhomme, J. Oden, On goal-oriented error estimation for elliptic problems: application to the control of pointwise errors, Comput. Methods Appl. Mech. Engrg. 176 (14) (1999) 313–331. http://dx.doi.org/10.1016/S0045-7825(98)00343-0.
[8] R.G. Sargent, Verification and validation of simulation models, in: Proceedings of the 30th Winter Simulation Conference, WSC'98, IEEE Computer Society Press, Los Alamitos, CA, USA, 1998, pp. 121–130. http://dl.acm.org/citation.cfm?id=293172.293216.
[9] O. Balci, Verification validation and accreditation of simulation models, in: Proceedings of the 29th Winter Simulation Conference, WSC'97, IEEE Computer Society, Washington, DC, USA, 1997, pp. 135–141. http://dx.doi.org/10.1145/268437.268462.
[10] W.L. Oberkampf, T.G. Trucano, Verification and Validation Benchmarks. Tech. Rep. SAND2007-0853, Sandia National Laboratories, 2007, Unlimited Release.

[11] I. Babuška, F. Nobile, R. Tempone, A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria, Comput. Methods Appl. Mech. Engrg. 197 (2008) 2517–2539.
[12] P.J. Roache, Perspective: Validation—What Does it Mean? ASME J. Fluids Eng. 131 (3) (2008) 1–3.
[13] M.J. Bayarri, J.O. Berger, R. Paulo, J. Sacks, J.A. Cafeo, J. Cavendish, C.-H. Lin, J. Tu, A framework for validation of computer models, Technometrics 49 (2) (2007) 138–154. http://amstat.tandfonline.com/doi/abs/10.1198/004017007000000092.
[14] D. Higdon, M. Kennedy, J.C. Cavendish, J.A. Cafeo, R.D. Ryne, Combining field data and computer simulations for calibration and prediction, SIAM J. Sci. Comput. 26 (2) (2005) 448–466. http://dx.doi.org/10.1137/S1064827503426693.
[15] P.S. Craig, M. Goldstein, J.C. Rougier, A.H. Seheult, Bayesian forecasting for complex systems using computer simulators, J. Amer. Statist. Assoc. 96 (2001) 717–729. http://EconPapers.repec.org/RePEc:bes:jnlasa:v:96:y:2001:m:june:p:7%17-729.
[16] R.E. Morrison, C.M. Bryant, G. Terejanu, S. Prudhomme, K. Miki, Data partition methodology for validation of predictive models, Comput. Math. Appl. 66 (10) (2013) 2114–2125. http://dx.doi.org/10.1016/j.camwa.2013.09.006.
[17] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (3) (2001) 425–464. http://dx.doi.org/10.1111/1467-9868.00294.
[18] M. Strong, J.E. Oakley, J. Chilcott, Managing structural uncertainty in health economic decision models: a discrepancy approach, J. R. Stat. Soc. Ser. C Appl. Stat. (2011).
[19] J.T. Oden, E.E. Prudencio, P.T. Bauman, Virtual model validation of complex multiscale systems: Applications to nonlinear elastostatics, Comput. Methods Appl. Mech. Engrg. 266 (2013) 162–184. http://dx.doi.org/10.1016/j.cma.2013.07.011.
[20] M. Panesi, K. Miki, S. Prudhomme, A. Brandis, On the assessment of a bayesian validation methodology for data reduction models relevant to shock tube experiments, Comput. Methods Appl. Mech. Engrg. 213–216 (2012) 383–398. http://dx.doi.org/10.1016/j.cma.2011.11.001.
[21] R.T. Cox, The Algebra of Probable Inference, Johns Hopkins University Press, Baltimore, MD, 1961.
[22] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, 2003.
[23] K.S. Van Horn, Constructing a logic of plausible inference: A guide to Cox's theorem, Int. J. Approx. Reason. 34 (1) (2003) 3–24.
[24] A. Vehtari, J. Lampinen, Bayesian model assessment and comparison using cross-validation predictive densities, Neural Comput. 14 (10) (2002) 2439–2468.
[25] F. Alqallaf, P. Gustafson, On cross-validation of Bayesian models, Canad. J. Statist. 29 (2) (2001) 333–340.
[26] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Stat. Surv. 4 (2010) 40–79. http://dx.doi.org/10.1214/09-SS054.
[27] I. Babuška, F. Nobile, R. Tempone, Reliability of computational science, Numer. Methods Partial Differential Equations 23 (4) (2007) 753–784.
[28] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. van Bloemen Waanders, K. Willcox, Large-Scale Inverse Problems and Quantification of Uncertainty, John Wiley and Sons, Ltd., Chichester, UK, 2010.
[29] G. Box, G.C. Tiao, Bayesian Inference in Statistical Analysis, Wiley Classics, New York, 1973.
[30] R.J. Hyndman, Computing and graphing highest density regions, Amer. Statist. 50 (2) (1996) 120–126.
[31] R.G. J. Miller, Simultaneous Statistical Inference, second ed., Springer, 1981.
[32] J. Hsu, Multiple Comparisons: Theory and Methods, Chapman and Hall/CRC, 1996.
[33] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global Sensitivity Analysis, The Primer. Wiley-Interscience, 2008.
[34] T. Cover, J. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[35] M. Meyer, J. Booker, Eliciting and Analyzing Expert Judgment: A Practical Guide, Society for Industrial and Applied Mathematics, 2001.
[36] P.H. Garthwaite, J.B. Kadane, A. O'Hagan, Statistical methods for eliciting probability distributions, J. Amer. Statist. Assoc. 100 (2005) 680–701.
[37] A. O'Hagan, C. Buck, A. Daneshkhah, J. Eiser, P. Garthwaite, D. Jenkinson, J. Oakley, T. Rakow, Uncertain Judgements: Eliciting Experts' Probabilities, Wiley, 2006.
[38] F.A. Moala, A. O'Hagan, Elicitation of multivariate prior distributions: a non-parametric Bayesian approach, J. Statist. Plann. Inference 140 (2010) 1635–1655.