Validation of Physical Models in the Presence of Uncertainty

Robert D. Moser and Todd A. Oliver

1 Introduction

Over the last century, the field of computational modeling has grown tremendously, from virtually non-existent to pervasive. During this time, simultaneous advances in simulation algorithms and computer hardware have enabled the development and application of increasingly complicated and detailed models to represent evermore complex physical phenomena. These advances are revolutionizing the ways in which models are used in the design and analysis of complex systems, enabling simulation results to be used in support of critical design and operational decisions [25, 1]. With continued advances in models, algorithms, and hardware, numerical simulations will only become more critical in modern science and engineering.

Given the importance of computational modeling, it is increasingly important to assess the reliability, in light of the purpose of a given simulation, of the models that form the basis of computational simulations. This reliability assessment is the domain of validation. While the concept of model validation is not new, it has recently received renewed attention due to the rapid growth in the use of models as a basis for making decisions [3, 5, 29]. This article provides an overview of the state-of-the-art in validation of physical models in the presence of uncertainty.

In science and engineering, the word validation is often used to refer to simple comparisons between model outputs and experimental data such as plotting the model results and data on the same axes to allow visual assessment of agreement or lack thereof. While comparisons between model and data are at the core of any validation procedure, there are a number of problems with such naive comparisons.

Robert D. Moser
Institute for Computational and Engineering Sciences and Department of Mechanical Engineering, The University of Texas at Austin, Austin, Texas, USA e-mail: rmoser@ices.utexas.edu

Todd A. Oliver
Institute for Computational and Engineering Sciences, The University of Texas at Austin, Austin, Texas, USA e-mail: oliver@ices.utexas.edu

First, these comparisons tend to lead to qualitative rather than quantitative assessments of agreement. While such qualitative assessments are often instructive and important, they are clearly incomplete, particularly as a basis for making decisions regarding model validity. Second, in naive comparisons, it is common to ignore or only partially account for uncertainty—e.g., uncertainty in the experimental observations or the model input parameters. Without accounting for these uncertainties, it is not possible to appropriately determine whether the model and data agree. Third, by focusing entirely on the agreement in the observable quantities, such comparisons neglect the intended uses of the model and, in general, cannot on their own determine whether the model is sufficient for the intended purposes.

These drawbacks of straightforward but naive comparisons highlight the two primary difficulties in model validation. First, one must quantitatively measure the agreement between model outputs and experimental observations while accounting for uncertainties in both. This fact is widely recognized, particularly in the statistics community, and there are a number of possible approaches. Second, depending on the intended use of the model, an assessment of the agreement between model outputs and available data is not sufficient for validation. Recognizing the purpose of the model is crucial to designing an appropriate validation approach.

1.1 Measuring Agreement Under Uncertainty

While the intended uses of a model will be important in a complete assessment of the validity of the model for those uses, all validation methods rely in some way on an assessment of whether the model is consistent with some set of observational data. In general, both the observations and the model—either through input parameters, the model structure, or both—are subject to uncertainties that must be accounted for in this comparison. Indeed, if both the model and the experiments are free from any uncertainty, then they can only be consistent if the model perfectly reproduces all the data. To define consistency in the far more common situation where something is uncertain, one must supply mathematical representations of all relevant uncertainties, a quantitative method for comparing uncertain quantities using the chosen representations of uncertainty, and a tolerance defining how closely model and data must "agree" to be declared consistent.

A wide range of formalisms have been proposed to represent uncertainty [9, 13, 34, 11, 28, 21], and there is still considerable controversy in the literature regarding the most appropriate approach, especially for so-call "epistemic" uncertainty (see below). Here we choose to focus on the Bayesian interpretation of probability, where probability provides a representation of the degree of plausibility of a proposition, to represent all uncertainties [35, 21]. This choice is popular and has many advantages, including a well-defined method for updating probabilistic models to incorporate new data (Bayes' theorem) and an extensive and rapidly growing set of available algorithms for both forward and inverse UQ [10, 24, 31, 2, 23, 27]. While a full discussion of the controversy over uncertainty representations is beyond the scope of this article, for the purposes of model validation, most uncertainty representations that have been proposed are either overly simplistic—e.g., using only intervals to represent uncertainty—or reduce to probability in special cases—e.g., mixed interval/probability methods [12] or Dempster-Shafer theory [34, 33]. Thus, independent of the noted controversy regarding uncertainty representations, a method for assessing consistency between model outputs and data, where both are represented using probability, is required.

One subtlety that arises in a validation assessment is that there are two types of uncertainty that may occur in a complex physical system. First are uncontrolled variabilities in the system, which, in the context of a validation test, result in observations that differ with repetition of the test. Such uncertainties are called aleatoric (from Latin *alea* for dice game). Probabilistic representations of aleatoric uncertainties describe frequencies of occurrence. The second form of uncertainty arises from incomplete knowledge of the system, which in the context of a validation test results in no variability in the observation with repetition of the test. Such uncertainties are called epistemic, and can be represented using the Bayesian interpretation of probability. In this case, probability describes the plausibility of outcomes [9, 35, 21]. Because the interpretation of probability is different for aleatoric and epistemic uncertainties, they will need to be distinguished when formulating validation criteria (see section 2.2.2).

Note that what is considered epistemic or aleatoric depends on the details of the problem. In some validation scenarios a parameter or input could be constrained to be the same on repeated observations, while in another scenario, it is uncontrolled. A simple example is a part which has uncertainties in geometry due to manufacturing variability. In a validation scenario in which the same part is used in repeated observations, this uncertainty is epistemic. But, in a scenario in which repeated observations are made each with a different part, the uncertainty is aleatoric.

Given the choice of probability to represent uncertainty, it is natural to define consistency in terms of the plausibility of the observations arising from the probabilistic model of the experiment, which represents uncertainties in both physical model and the observation process. Of course, there are still many ways to define a "plausible outcome". Here, we use highest posterior density credibility sets and tail probabilities of the observable or relevant test quantities associated with them to evaluate the plausibility of data as an outcome of a model. These ideas are described in more detail in Section 2.

1.2 Different Uses of Models

Computational models are used for many different purposes. In science and engineering, these different purposes can be split into three broad categories: 1) investigation of the consequences of theories; 2) analysis of experimental data; and 3) prediction.

Scientific theories often lead to models that are sufficiently complex that computations are required to evaluate whether the theory is consistent with reality. When computation is used in this way, the computational model will be an expression of the theory in a scenario in which experimental data will be available. In addition to the theory being tested, the computational model may include representations (models) of, for example, the experimental facility and the diagnostic instruments. These auxiliary models should be endowed with uncertainties as appropriate. The validation question is then a simple one: given the uncertainties in the auxiliary models and the experimental data, is the data consistent with the computational model? This can be assessed using the techniques discussed in section 2. If an inconsistency is detected, then either the theory being tested or one or more of the auxiliary models is invalid. Assuming the auxiliary models have been sufficiently tested so that their reliability is not in doubt, the theory being tested must be rejected. Alternatively, a lack of detectable inconsistency implies only that the analysis has failed to invalidate the theory.

Models are also used to analyze data obtained from experiments. In particular, it is often the case that the quantity one wishes to measure, e.g., the flow velocity at a point in a wind tunnel experiment, is not directly observable. Instead, one measures a different quantity, e.g., a voltage in a hot wire anemometer circuit, which is related to the quantity of interest (QoI) through a model. As when investigating the consequences of a theory, any detectable inconsistency between the model output and a reliable reference for the quantity being inferred—to be clear, this reliable reference must be from independent source, such as a different instrument as in a calibration experiment—is cause for the model to be invalid for data analysis purposes. However, a lack of detectable inconsistency does not imply that the model is valid for data analysis. One must also ensure that the intended data analysis does not require the model to extrapolate beyond the range of independent reference data. This extra step is necessary because once the model is used in an extrapolatory mode, it is being used to make predictions, which requires substantially more validation effort, as discussed below.

The most difficult validation situation is when one wishes to use the model to make predictions. To understand this difficulty, it is necessary to be precise about what it means to make a prediction. A prediction is a model-based computation of a specific QoI for which there is no observational data, for instance because the quantity cannot be measured, because the scenario of interest cannot be produced in the laboratory or because the system being modeled has not yet been built. Indeed, the prediction is necessary precisely because the QoI is not experimentally observable at the time the information is required, e.g. to inform a decision-making process. Thus, prediction implies extrapolation.

It is well-known that a model may be adequate for computing one quantity while not another or in one region of the scenario space and not another. Thus, when extrapolation is involved, it is insufficient to simply compare the model against data to determine consistency. This consistency is necessary but not sufficient because it does not account for the fact that the prediction quantity and scenario are different from the observed quantities and scenarios. Thus, a key challenge in model validation for prediction is in determining the implications of the agreement or disagreement between the model output and the validation data on the accuracy and reliability of the desired prediction. For example, one important question is, given some observed discrepancy between the model and data, is the model likely to produce predictions of the QoI with unacceptably large error?

While this type of question has generally been left to expert judgment [3, 5], a recently proposed predictive validation process aims to systematically and quantitatively address such issues [30]. The process involves developing stochastic models to represent uncertainty in both the physical model and validation data, allowing rigorous assessment of the agreement between the model and data under uncertainty, as discussed in Section 2. However, these stochastic models, coupled with the common structure of physics-based models, allows one to pose a much richer set of validation questions and assessments to determine whether extrapolation with the model to the prediction is supported by the available data and other knowledge. This predictive validation process will be discussed further in Section 3.

2 Comparing Model Outputs and Data in the Presence of Uncertainty

Appropriate validation processes for mathematical models of physical systems depend on the purpose of the model, as discussed in Section 1. But, regardless of this purpose, the process will rely on the comparison of outputs of the mathematical models with observations of physical systems to which the model can be applied. Such comparisons are complicated by the presence of uncertainties in both the mathematical model and the observations. In the presence of uncertainty, the relatively straightforward question of whether a model and observations “agree” becomes a more subtle question of whether a model with all its uncertainties is consistent with the observations and all their uncertainties. In this section, we address the sources of uncertainty in validation tests (section 2.1) and techniques for making comparisons in the presence of uncertainty (section 2.2). Section 2.3 gives some general guidance on selecting validation data.

2.1 Sources of Uncertainty in Validation Tests

To analyze the sources of uncertainty in validation tests, it is helpful to introduce an abstract structure for such a test. Consider a mathematical model $\mathscr{U}$ of some physical phenomenon, which is a mapping from some set of input quantities x to output quantities u (in general, a model for a quantity will be indicated by a calligraphic upper case symbol). The model will in general involve a set of model parameters α_u, which had to be calibrated using data from observations of the phenomenon. The α_u are generally uncertain. In addition, in some situations, the model may be known to be imperfect, so that there is an error ε_u. Therefore,

$$u = \mathscr{U}(x;\alpha_u) + \varepsilon_u, \tag{1}$$

where the error is represented here as additive, though other choices are possible. The error ε_u is imperfectly known and may be represented by an "inadequacy model" $\mathscr{E}_u(x;\beta_u)$ [30], with inadequacy model parameters β_u that are also calibrated and uncertain.

Observations of the phenomenon modeled by $\mathscr{U}$ are generally made in the context of some larger system. This larger system has observable quantities v, which will be the basis of the validation test. The validation system must also be modeled with a model $\mathscr{V}$ that is a mapping from a set of inputs y and the modeled quantities u to the observables v. The dependence on u is necessary since the system involves the phenomenon being modeled by $\mathscr{U}$. The model $\mathscr{V}$ will in general involve model parameters α_v, which are uncertain, and $\mathscr{V}$ may itself be imperfect with error ε_v, which is modeled as $\mathscr{E}_v$ with parameters β_v. We thus have a preliminary representation of the validation system:

$$v = \mathscr{V}(u,y;\alpha_z) + \mathscr{E}_v(u,y;\beta_v) = \tilde{\mathscr{V}}(u,y;\alpha_v,\beta_v), \tag{2}$$

where $\tilde{\mathscr{V}}$ is the validation system model enriched with the inadequacy model $\mathscr{E}_v$.

To complete the validation model, u in (2) is expressed in terms of the model $\mathscr{U}$, which means that the inputs x to $\mathscr{U}$ must be determined from the inputs y to $\mathscr{V}$ using a third model $\mathscr{X}$ with parameters α_x, which may also be imperfect, introducing uncertain errors ε_x, modeled as $\mathscr{E}_x$ with parameters β_x, yielding:

$$x = \mathscr{X}(y;\alpha_x) + \mathscr{E}_x(y;\beta_x) = \tilde{\mathscr{X}}(y;\alpha_x,\beta_x). \tag{3}$$

Because the model $\mathscr{U}$ of the phenomenon is introduced into a larger model of the validation system, it is called an "embedded model" [30].

Finally, errors δ_v are introduced in the physical observations of v themselves, commonly identified as observation or instrument error. The complete model of the validation test is then

$$v = \tilde{\mathscr{V}}[\mathscr{U}(\tilde{\mathscr{X}}(y;\alpha_x,\beta_x),\alpha_u) + \mathscr{E}_u(\tilde{\mathscr{X}}(y;\alpha_x,\beta_x)),y;\alpha_v,\beta_v] + \delta_v. \tag{4}$$

Here, the $\mathscr{E}_u$ term is retained explicitly to emphasize that the validation test is directed at the physical model $\mathscr{U}$ and the associated inadequacy model $\mathscr{E}_u$, if any. In this model of the validation test there are four types of uncertainties: uncertainties in the model parameters (α_x, α_u, α_v, β_x, β_u, and β_v); uncertainties in the validation inputs y; uncertainties due to the model errors ($\mathscr{E}_u$, $\mathscr{E}_x$, and $\mathscr{E}_v$); and finally uncertainties due to the observation or instrument errors (δ_v). Note, that in some cases, it may be convenient to include the response of the measuring instrument(s) in the validation system model $\mathscr{V}$. In this case, the instrument errors are included in ε_v.

Clearly, the design of an experimental observation will seek to minimize the uncertainties not directly related to the model being studied (i.e. other than the uncertainties in α_u and $\mathscr{E}_u$). Furthermore, in the event that the model of the validation (4) is found to be inconsistent with observations of v, all that can be said is that at least one of the models involved ($\mathscr{U} + \mathscr{E}_u$, $\tilde{\mathscr{V}}$, $\tilde{\mathscr{X}}$ and/or the representation of the observation error) or some input to one of these models is inconsistent with reality. For such a validation to meaningfully test the model $\mathscr{U}$ of the phenomenon of interest, the validation problem should, if possible, be designed so that auxiliary models $\mathscr{V}$ and $\mathscr{X}$ are much more reliable than $\mathscr{U}$. In this way, any inconsistency between model and observation will strongly implicate the model $\mathscr{U}$ that is being tested.

This abstract structure of a validation test might be better understood through reference to a relatively simple example. Let $\mathscr{U}$ be a simple homogeneous linear elastic constitutive model for the stress-strain relationship in some solid part, with u being the stress tensor field and x the strain tensor field. The parameters α_u are the Lamé constants or equivalently Young's modulus and the Poisson ratio for the material. No inadequacy model is included. A validation test might be conducted by placing the part in a testing machine, which applies a specified overall load force through a fixture in which the part is mounted. The observed quantities v could be the displacement of one or more points on the part, and δ_v would represent the error in determining this displacement experimentally.

The validation system model $\mathscr{V}$ would include an equilibrium continuum equation for the part and possibly the fixture, a model for the connection between the part and the fixture, and a representation of the load characteristics of the testing machine. Parameters α_v might include those describing the material of the fixture, the connection between part and fixture and the testing machine. The inputs y could include the applied load, the geometry of the part, the load configuration and other settings of the testing machine. The error ε_v might account for uncontrolled nonidealities in the way the part is mounted in the fixture, or in the testing machine. Finally, as a consequence of determining the displacement of points on the part using a continuum representation, the displacement everywhere would be determined. The model $\mathscr{X}$ would then include the mapping from the displacement field in the continuum model used in $\mathscr{V}$ to the strain field, which because it is a simple kinematic definition would not introduce any additional modeling errors ε_x.

The validation system model (4) defines the expected relation between the generally uncertain inputs y and observations v. This model includes uncertainties due to the model parameters (the α's), the modeling errors (the ε's) and the observation or instrument errors (δ). With a mathematical characterization of these uncertainties, (4) makes an uncertain claim as to the values of the observed quantities. The validation test is then to make observations $\hat{v}$ of the physical system, and determine whether the $\hat{v}$ are consistent with the uncertain claims regarding v. Assessing this consistency is the subject of the following section.

2.2 Assessing Consistency of Models and Data

From the above discussion, it is clear that a mathematical representation of the many uncertainties in the validation system is needed. A number of such uncertainty representations have been proposed, and as discussed in Section 1, many of the issues surrounding validation under uncertainty are not unique to any particular uncertainty representation. However, in the current discussion of how one actually makes an assessment of the consistency of models and observation in the presence of uncertainty, it will be helpful to consider a particular uncertainty representation: Bayesian probability. This representation is used here for the reasons discussed in Section 1.

With representation of uncertainty as Bayesian probability, the question of consistency of the model with data falls in the domain of Bayesian model checking. There is a rich literature on this subject, which we will not attempt to recite here. Instead, we describe an approach to model checking with broad applicability, which is generally consistent with common notions of validation for models of physical systems. The ideas outlined here are most closely aligned with those of Andrew Gelman and collaborators [16, 15, 17, 18, 14], and the reader is directed to these references for a more detailed statistical perspective. In particular, see [18] for a broad perspective on the meaning and practice of Bayesian model checking.

2.2.1 Single Observation of a Single Observable

The simplest case to consider is that of a validation test with a single observable v, which we consider here to be a real-valued continuous random variable. When uncertainties are represented using Bayesian probability, the validation system model (4) yields a probability density $p(v)$ for v. If a single measurement of v is made, yielding a value $\hat{v}$, the question is then whether the model expressed as the probability density $p(v)$ is consistent with the observation $\hat{v}$. A straightforward way to assess this consistency is to plot the distribution $p(v)$, as in figure 1. Indeed graphical representations of model and data are often very informative. It is clear in this figure that if the observation is $\hat{v} = v_i$ for $i = 1$, 2 or 3, that the observation is consistent with the model because these points fall within the significant probability mass of the distribution. On the other hand, if $\hat{v} = v_i$ for $i = 4$ or 5, it is clear that the model is inconsistent with the observation. In making these assessments we are asking how likely it is for the observed valued $\hat{v}$ to arise as a sample from a random variable with distribution $p(v)$, and for the values v_i for $i = 1, \dots, 5$ the answer to this question is clear. If however, $\hat{v} = v_6$, the answer is not obvious, and for these marginal cases some criterion would be needed to decide whether model and data are consistent. More usefully, we can admit a continuum of levels of consistency, and ask for a measure of the (in)consistency between the model and data.

This is a general issue in statistical analysis. A common approach is to consider the probability of obtaining an observation more extreme than the observation in hand. That is, we can compute the "tail probability" $P_>$ as

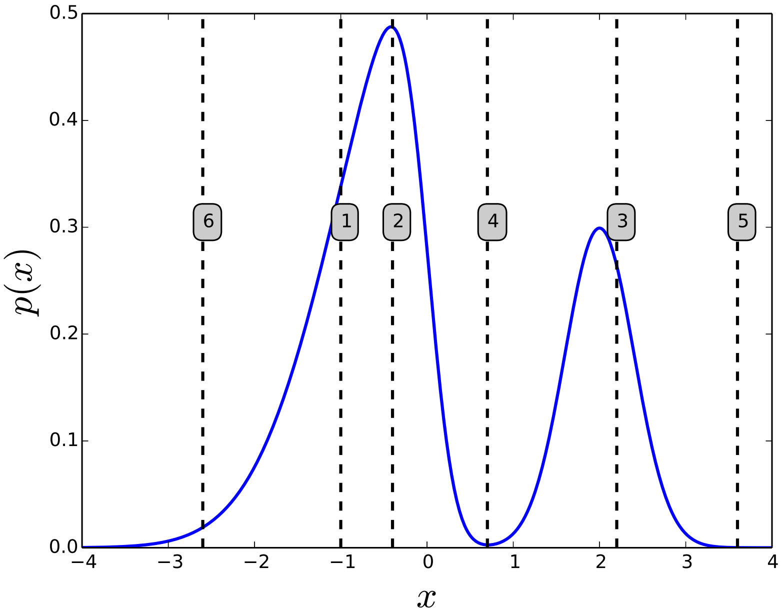

Fig. 1: Hypothetical model output distribution and a number of possible observations illustrating observations that are clearly consistent with the output distribution ($i = 1,2,3$), observations that are clearly inconsistent ($i = 4,5$), and observations where the agreement is marginal ($i = 6$).

$$P_{>} = P(v > \hat{v}) = \int_{\hat{v}}^{\infty} p(v)\, dv \tag{5}$$

which is the probability, according to the model, of an observation being greater than $\hat{v}$ ($P_{<}$ is defined analogously). This is the well-known (Bayesian) p-value, and if it is sufficiently small (0.05 or 0.01, for example), one could conclude that $\hat{v}$ is an unlikely outcome of the model, so that the validity of the model is suspect. This leads to the concept of a credibility interval, an interval in which, according to the model, it is highly probable (probability of for example 0.95 or 0.99) that an observation will fall. Given a probability distribution $p(v)$, there are many possible credibility intervals, or more generally credibility sets, with a given probability. One could, for example, choose a credibility interval centered on the mean of the distribution, or one defined so that the probability of obtaining an observation greater than the upper bound of the interval is equal to that of an observation less than the lower bound. Either of these credibility intervals has the disturbing property of including the point v_4 in figure 1, which is clearly not a likely draw from the distribution plotted.

A credibility region that is more consistent with our intuitive understanding of credible observations for skewed and/or multi-modal distributions such as that shown in figure 1 is the highest posterior density (HPD) credibility region [7]. 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. Thus, for a multi-modal distribution like that shown in Figure 1, an HPD region may consist of multiple disjoint intervals [20] around the peaks, leaving out the low probability density regions between the peaks.

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 [30]. An appropriate definition of q would be one that represents no information about v [21]. Using this definition of the highest posterior *relative* density (HPRD), 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 observation $\hat{v}$ is in the HPRD-credibility set for v according to the model. That is:

$$\gamma = 1 - \int_S p(v)\,dv, \quad \text{where} \quad S = \left\{ v \,:\, \frac{p(v)}{q(v)} \geq \frac{p(\hat{v})}{q(\hat{v})} \right\}. \tag{6}$$

When γ is smaller than some tolerance, say less than 0.05 or 0.01, $\hat{v}$ is considered an implausible outcome of the model—i.e., there is an inconsistency between the model and the observation.

2.2.2 Multiple Observations of a Single Observable

Of course, it is common to make multiple measurements of the observable v, especially if the measurement is noisy. Consider the case where the observational uncertainties represented in the model—which lead to the appearance of δ_v in (4)—are purely aleatoric and independent for each observation. Further, assume for the purposes of this discussion that any epistemic uncertainties in the model are negligible. In this case, the model implies that each observation is an independent sample of the distribution $p(v)$, and the validation question is whether a set of N observations $\hat{v}_i$ for $i = 1, 2 \ldots N$ is consistent with sampling from $p(v)$. It is clearly erroneous to check whether each individual observation is in a given credibility region, as the probability of at least one sample falling outside the credibility region will increase to 1 as N increases, even if the samples are in fact drawn according to the model distribution $p(v)$. A number of correction methods for this effect have been developed in the statistics literature [26, 19].

More generally, we are faced with a common problem in statistical hypothesis testing in which we ask whether a vector of observational values $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_N)$ is unlikely to have arisen as instances of random variables, in this case iid random variables with distribution $p(v)$. An obvious extension to the HPRD regions described above can be defined in terms of the joint distribution $p(V)$ of the vector of N iid random variables $V = (v_1, \ldots, v_N)$, which because of independence can be written:

$$p(V) = \Pi_{i=1}^{N} p(v_i). \tag{7}$$

The HPRD credibility metric can then be written:

$$\gamma = 1 - \int_S p(V)\, dV, \qquad \text{where } S = \left\{ V \ : \ \frac{p(V)}{q(V)} \geq \frac{p(\hat{V})}{q(\hat{V})} \right\}. \tag{8}$$

While this directly answers the question of how credible the observations are as samples for the model distribution, it is generally difficult to compute when N is large since it involves evaluating a high-dimensional integral over a complex region. An alternative approach is to consider one or more test quantities [15, 16]. A test quantity $T(V)$ is a mapping from an N-vector to a scalar. When evaluated for a random vector, it is a random scalar, which is designed to summarize some important feature of V. The idea then is to ask whether $T(\hat{V})$ is a plausible sample from the distribution of $T(V)$. One could for example compute p-values for this comparison. The HPRD metric could also be used, but $p(T(V))$ which is part of its definition, is usually difficult to compute.

In addition to being potentially more tractable, the use of test statistics has another advantage. With rare exceptions, the uncertainty representations leading to the stochastic model being validated are based on crude and/or convenient assumptions about the uncertainties. Indeed, iid Gaussian random variables are often used to model experimental noise, and while this is sometimes justified, it is often assumed purely out of convenience. Thus, we do not necessarily expect that, nor do we generally need, the model distribution $p(v)$ to be representative of the random processes that lead to the variability in $\hat{v}$ from observation to observation. In this case, we aspire only that the uncertainty representations will characterize what is important about the uncertainty for the purposes for which the model is to be used. While the HPRD metric given in (8) does not take this into account, validating using test quantities gives one the opportunity to choose T to characterize an important aspect of the uncertainty. For example, if the model is to be used to evaluate extreme deviations from nominal behavior or conditions, then it might make sense to perform validation comparisons based on the test quantity $T(V) = \max_i v_i$. A few example test quantities are discussed in the next subsection.

Finally, the assumption of negligible epistemic uncertainty, while useful in simplifying this discussion, is not generally applicable. This assumption can be removed by marginalizing over the epistemic uncertainties represented in the model, as described in Section 2.2.4.

2.2.3 Defining Test Quantities

To select validation test quantities, one should consider what characteristics of the aleatoric uncertainty are important in the context of the model and its planned use. One common consideration is that the mean and variance should be consistent with observations. A straightforward test quantity is simply the sample average $A(V)$; that is:

$$A(V) = \frac{1}{N}\sum_{i=1}^{N} v_i \tag{9}$$

The validation comparison then reduces to asking whether the distribution of $p(A(V))$ implied by the model is consistent with the observation $A(\hat{V})$. Since the v_i determined from the model are iid, if N is sufficiently large, the central limit theorem implies that $A(V)$ is approximately $\mathscr{N}(\mu, \sigma^2/N)$, where μ and σ^2 are the mean and variance of v. This leads to the Z-test in statistics.

To test whether the variability of v is consistent with the observed variability of $\hat{v}$, the test quantity

$$X^2(V) = \sum_{i=1}^{N} \frac{(v_i - \mu)^2}{\sigma^2} \tag{10}$$

could be used, which is a χ^2 discrepancy. Note that this test quantity is different because it depends explicitly on characteristics of the model (mean and variance). If in addition to being iid, the v obtained from the model are normally distributed, the model distribution $p(X^2(V))$ will be the χ^2 distribution, with N degrees of freedom.

When the test quantity has a known distribution, as $A(V)$ and $X^2(V)$ discussed above do, it simplifies assessing consistency using, for example, HPRD criteria or p-values. This is so because tail integrals of these distributions are known. However, it is not necessary that exact distributions be known, since the relevant integrals can be performed using Monte Carlo or other uncertainty propagation algorithms.

For example, in some problems, one is concerned with improbable events with large consequences. In this case, one is interested in the tail of the distribution of v and the extreme values that v may take. A simple test quantity that is sensitive to this aspect of the distribution of v is the maximum attained value. That is

$$M(V) = \max_i v_i \tag{11}$$

The random variable $M(V)$ that is implied by the model can be sampled by simply generating N samples of v and determining the maximum. Thus the p-value relative to the observation $M(\hat{Z})$ can be computed by Monte Carlo simulation.

There is a large literature on statistical test quantities (c.f. [22]), for use in a wide variety of applications. Texts on statistics should be consulted for more details. Among commonly used statistical tests are those which test whether a population is consistent with a distribution with known characteristics (e.g. the mean and/or variance) as with the average and χ^2 test quantities discussed above. These are also useful when the model can be used to compute these characteristics with much higher statistical accuracy than they can be estimated from the data (e.g. one can generate many more samples from the model than there are data samples). In other situations, the model may be expensive to validate, so that the number of samples of the posterior distribution of the model is limited. In this case, test quantities that test whether two populations are drawn from distributions with the same characteristics are useful. A simple example is the two sample t-test (Welch's test), which is used to test whether two populations have the same mean.

2.2.4 General Posterior Model Checks

The discussion in Sections 2.2.1-2.2.2 is instructive but does not apply to the usual situation. In particular, it is common to have multiple observations of multiple different quantities, the predictions of which are affected by both epistemic and aleatoric uncertainties. This section generalizes the validation comparisons discussed previously to this more complicated situation.

Consider the model of the observable expressed in (4). This model includes uncertain parameters, the α's and β's, representations of modeling errors, ε_u, ε_x and ε_v, and the representation of the observation errors δ_v. Generally, the uncertainties in the parameters are considered to be epistemic; that is, there are ideal values of these parameters, which are imperfectly known. Their values do not change from observation to observation in the same system. The observation error δ_v is often considered to be aleatoric, for example from instrument noise. However, there may also be systematic errors in the measurements, which are imperfectly known and thus epistemically uncertain. Similarly, depending on the nature of the phenomena involved and how they are modeled, the errors ε_u, ε_x and ε_v in the models $\mathscr{U}$, $\mathscr{X}$ and $\mathscr{V}$ may be epistemic, aleatoric or a mixture of the two. The models for these uncertainties could then include an epistemic part and and an aleatoric part. That is:

$$\varepsilon_x \approx \mathscr{E}_x^e + \mathscr{E}_x^a, \qquad \delta_v \approx \mathscr{D}_v^e + \mathscr{D}_v^a. \tag{12}$$

where superscript e and a are for epistemic and aleatoric, respectively. One challenge is then to compare the model and possibly repeated observations under these mixed uncertainties.

In addition to multiple observations, there may be multiple observables. These observables may be of the same quantity (e.g. the temperature), for different values of the model inputs y (e.g. at different points in space or time), or of different quantities (e.g. the temperature and the pressure). The observable v, should thus be considered to be a vector of observables of dimension n (say). In general, the aleatoric uncertainties in the various observables will be correlated. In the model, these correlations will arise both from the way the aleatoric inadequacy uncertainties impact the observables through the model and from correlations inherent in the dependencies of the aleatoric uncertainties (the $\mathscr{E}^a$'s and $\mathscr{D}_z^a$) on model inputs. Due to the presence of aleatoric uncertainties, a validation test may make multiple (N) observations $\hat{v}$ of the observable vector v, resulting in a set of observations $\hat{V} = \{\hat{v}_1, \ldots, \hat{v}_N\}$, which we will consider to be independent and identically distributed (iid).

To facilitate further discussion, let us consolidate the epistemic uncertainties (including those associated with $\mathscr{E}_u^e$, $\mathscr{E}_x^e$, $\mathscr{E}_v^e$ and $\mathscr{D}_v^e$) into a vector θ, which may be infinite dimensional. The validation model for the observables (4) can then be considered to be a statistical model $\mathscr{S}$ for the aleatoricly uncertain v, which depends on the epistemically uncertain θ; that is, $v = \mathscr{S}(\theta)$. The validation question is then whether the set of observations $\hat{V}$ is consistent with $\mathscr{S}$, given what is known about θ.

If θ is known precisely (i.e. no epistemic uncertainty), then the situation is similar to that discussed in section 2.2.2. For a given θ, the model $\mathscr{S}$ defines a probability distribution $p(v|\theta)$ in the n-dimensional space of observables. The observables are not independent, so this is not a simple product of one-dimensional distributions, but this introduces no conceptual difficulties. A set of N observations then defines a probability space of dimension nN. Because the observations are iid, the probability distribution is written:

$$p(V|\mathscr{S},\theta) = \Pi_{i=1}^{N} p(v_i|\mathscr{S},\theta). \tag{13}$$

Here, the probability density is conditional on the model $\mathscr{S}$ because these are the distributions of outputs v implied by the model. Consistency of $\hat{V}$ with $\mathscr{S}$ can then in principle be determined through something like the HPRD credibility metric (6). However, as discussed in section 2.2.2, this is generally not computationally tractable, nor is it generally desirable. Alternatively, one or more test quantities T may be defined to characterize what is important about the aleatoric variation of v, and as in section 2.2.2, we can test whether the observed $T(\hat{V})$ is consistent with the distribution obtained from the model $p(T(V)|\mathscr{S},\theta)$. For example, one could use the p-value $P_>$, which now depends on θ.

The parameters θ are uncertain, with uncertainty that is entirely epistemic by construction. The validation question is therefore whether there are plausible values of θ that make the observed value of the test quantity $T(\hat{V})$ plausible. In validation, it is presumed that the parameters in the models (the α's and β's in (4)), have been calibrated (e.g. via Bayesian inference), and that the epistemic uncertainties are now expressed as probability distributions $p(\theta|\mathscr{S},\hat{w})$, where $\hat{w}$ represents the data for the observables w used to calibrate the parameters. This calibration data may or may not be included in the validation data $\hat{v}$. In Bayesian inference, this is the posterior distribution, and so we are interested in the distribution of $T(V)$ or $P_>$ induced by the posterior distribution of θ.

As suggested by Box [8], Rubin [32] and Gelman *et al* [16], in this situation, the consistency of the observations $\hat{V}$ with the model can be determined by considering the distribution of $T(V)$ implied by the distribution of θ:

$$p(T(V)|\mathscr{S},\hat{w}) = \int_{\theta} p(T(V)|\mathscr{S},\theta)p(\theta|\mathscr{S},\hat{w})\,d\theta. \tag{14}$$

This is termed the posterior predictive distribution by Gelman *et al* in [16], though they were referring to the case in which $\hat{w}$ is the same as $\hat{v}$. It then can be determined whether the observed value $T(\hat{V})$ of the test quantity is consistent with the distribution $p(T(V)|w)$, using, for example p-values:

$$P_{>} = \int_{\theta} P(T(V) > T(\hat{V})|\mathscr{S},\theta)p(\theta|\mathscr{S},\hat{w})\,d\theta. \tag{15}$$

2.3 Data for Validation

Of course, to perform a validation comparison, it is necessary to have data to which to compare. The question is, what should this data be? Logically, it is required that, for a model to be valid, it must be consistent with all available relevant observational data. While true, this does not provide useful guidance for designing experimental observations for the purpose of validation. Selecting appropriate validation data requires consideration of several problem-specific issues, so it is difficult to specify generally applicable techniques for designing validation tests. Instead, listed here is broad guidance for designing validation experiments.

1. Often, the first point at which models are confronted with data is when they are calibrated. As part of the calibration process, the calibrated model should also be validated against the calibration data. Because the model has been calibrated to fit the calibration data, consistency with the data will not greatly increase confidence in the model. But if the model is inconsistent with the data with which it has been calibrated, it is a very strong indictment of the model. With parsimonious models, which describe a rich phenomenon with few parameters, failure to reproduce the calibration data is a common mode of validation failure.
2. To increase confidence in the validity of a model, it should be tested against data that was not used in its calibration. Sometimes this is done by holding back some portion of the calibration data set so that it can be used only for validation, which leads to cross-validation techniques [4]. This is generally of limited utility for physics-based models. A much stronger validation test is to use a completely different data set, from experiments with different validation models $\mathscr{V}$ in (4). For example, calibration of a model might be done using a set of relatively simple experiments in a laboratory facility, while validation experiments are in more complex scenarios in completely different facilities.
3. The development of computational models often involves various approximations and assumptions. To increase confidence in the models, one should design validation experiments that test these approximations; that is, experiments and measurements should be designed so that the observed quantities are expected to be sensitive to any errors introduced by the approximation. Sensitivity analysis applied to the model can help identify such experiments. Furthermore, test quantities used in validation assessment should also be designed to be sensitive to the approximations being tested.

4. Models of complex physical systems commonly involve sub-models (embedded models) of several physical phenomena that are of questionable reliability. When possible, the individual embedded models should be calibrated and validated separately, using experiments that are designed to probe the modeled phenomena individually. The resulting experiments will generally be much simpler, less expensive, easier to instrument and easier to simulate than the complete system. This could allow many more experiments and more measurements to be used for calibration and/or validation. When possible, further experiments involving a few of the relevant phenomena should be performed to validate models in combination. Experiments of increasing complexity and involving more phenomena can then be pursued, until measurements in systems similar in complexity to the target system are performed. This structure has been described as a "validation pyramid" [6], with abundant simple inexpensive, data-rich experiments at the bottom, and increasingly expensive and limited experiments as one goes up the pyramid.

   The role of the experiments higher in the validation pyramid is generally different from those lower down. The lower level experiments are designed to calibrate models and validate that the models provide a quantitatively accurate representation of the modeled phenomena. Experiments higher in the pyramid are intended to test the modeling of the interactions of different phenomena. Finally, measurements in systems as similar as possible to the target system, at conditions as close as possible to the conditions of interest, are used to detect whether there are unexpected phenomena or interactions that may affect predictions in the target system.
5. As discussed in section 3, when a computational model is used for predictions of unobserved QoIs, the reliability of the predictions depends on the embedded models being used in conditions that have been well tested, with validation observations that are sensitive to errors in the model in the same way as the prediction QoIs. Validation tests should therefore be designed to challenge embedded models as they will be challenged in the prediction scenarios. The conditions that the embedded model experiences during predictions can be evaluated through model simulations of the prediction scenario, and sensitivities of the QoIs to an embedded model can be determined through sensitivity analysis conducted in the system model.

3 Validating Models for Prediction

As mentioned in Section 1, when a physical model is being used to make predictions, a detectable inconsistency between the model output and experimental data is not, on its own, sufficient to invalidate the model for use in the prediction. Indeed, it is common in engineering to use models which are known to be inconsistent with some relevant data but which are sufficient for the predictions for which they are to be used. In this situation, the important validation question is not whether the model is scientifically valid—often it is known *a priori* that it is not—but rather whether a prediction made with the model is reliable. This is generally a much more difficult question since it is essentially asking whether an extrapolation from available information will be reliable. Recently, a process for addressing this question in the context of physics-based mathematical models was developed [30]. This section will outline the components of this process.

3.1 Mathematical Structure for Prediction

An abstract structure of a validation test is defined in Section 2.1. Here, we refine this mathematical structure to include features common in models of physical systems that will make reliable predictions possible. To fix the main ideas, this structure is presented first in the simplest possible context (Section 3.1.1), with a more general abstraction outlined briefly in Section 3.1.2.

3.1.1 Simplest Case

Mathematical models of the response of physical systems are generally based in part, indeed to the greatest extent possible, on physical theories that are known *a priori* to be reliable in the context of the desired prediction. The mathematical expression of these theories is a reliable, but incomplete, model of the system, which can be written

$$\mathscr{R}(u,\tau;r) = 0, \tag{16}$$

where $\mathscr{R}$ is an operator expressing the reliable theory, u is the state, r is a set of variables that defines the problem scenario, and τ is an additional quantity that must be known to solve the system of equations. For instance, in fluid mechanics, $\mathscr{R}$ could be a non-linear differential operator expressing conservation of mass, momentum, and energy with u including the fluid density, velocity, and temperature fields. In this case, τ would include the pressure, viscous stresses, and heat flux, and r would include parameters like the Reynolds and Mach numbers as well as details of the flow geometry and boundary conditions.

This structure in which the model is based at least in part on reliable theory is common in physics-based models. The foundation on theory whose validity is not in question will be important for making predictions. However, the reliable theory on its own rarely forms a closed system of equations, which is represented in (16) by τ. If τ could be determined from u and r using a model as reliable as $\mathscr{R}$ itself, then the combination of (16) with this high-fidelity model for τ would form a closed system of equations, and solutions of this system would be known to be reliable.

In general, such models—i.e., models that are both closed and known *a priori* to be highly reliable—are not available in the context of complex prediction problems. Instead a quantity τ must be represented using a lower-fidelity model or a model whose reliability is not known *a priori*. In this case, we call the model for τ, denoted $\mathscr{T}$, an embedded model and write

$$\tau = \mathscr{T}(u; s, \theta), \tag{17}$$

where s is a set of scenario parameters, possibly distinct from r, for the embedded model, and θ is a set of calibration or tuning parameters. Since the embedded model $\mathscr{T}$ is not known *a priori* to be reliable for the desired prediction, it will be the focus of the validation process.

The system of equations consisting of (16) and (17) is closed, but for calibration, validation, and prediction, some additional relationships are necessary. In particular, we require some way to express the experimental observables v and the prediction QoIs q. We assume that there are maps $\mathscr{V}$ and $\mathscr{Q}$ that determine v and q, respectively, from the model state u, the modeled quantity τ, and the global scenario r:

$$v = \mathscr{V}(u, \tau; r), \tag{18}$$

$$q = \mathscr{Q}(u, \tau; r). \tag{19}$$

This closed set of equations (16) through (19), which allows calculation of both the experimental observables and predictions QoIs, is referred to as a composite model, since it is a composition of high and low fidelity components.

Because the foundation of the composite model is a reliable theory whose validity is not questioned, it is possible to make reliable predictions despite the fact that a less reliable embedded model is also involved. All that is required is that the less reliable embedded model not be used outside the range where it has been calibrated and tested. 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. For example, an elastic constitutive relation for the deformation of a material can only be relied upon provided the strain remains within the bounds in which it has been calibrated and tested. Despite this restriction, a model for a complex structure made from the material, which is based on conservation of momentum, can reliably predict a wide range of structural responses.

3.1.2 Generalizations

The simple problem statement in Section 3.1.1 is sufficient to introduce many of the concepts that are critical in validation for prediction, including the distinction between the QoI and available observable quantities and the notion of an embedded model. However, there are several important generalizations that are required to represent the validation and prediction process in complex physical systems. These are outlined below. A more detailed description can be found in [30].

- **Multiple embedded models:** In a complex system, there will generally be multiple physical phenomena for which an embedded model is needed (e.g. thermodynamic models, chemical kinetics models and molecular transport embedded models in a composite model of a combustion system). Thus, the composite model will generally depend on N_τ quantities τ_i, each with associated models $\mathscr{T}_i$, calibration parameters θ_i and scenario parameters s_i.

- **Multiple reliable models:** The experimental systems in which measurements are made for validation and calibration are commonly different from, usually simpler than, the prediction system. For each of N_e experiments, there will in general be a different reliable model $\mathscr{R}^j$, with associated state variables u^j, observables v^j, scenario parameters r^j and set of quantities requiring embedded models $\{\tau\}^j$. There are also, therefore, N_e observation models $\mathscr{V}^j$ and sets of embedded models $\{\mathscr{T}\}^j$. For each experiment, the set of modeled quantities $\{\tau\}^j$ must include at least one member of the set of modeled quantities $\{\tau\}^0$ used in the prediction model, but may include other quantities requiring embedded models that are not relevant to the prediction (e.g. to represent an instrument, or the laboratory facility).
- **Differing state variables:** In general, the different reliable models for each experiment $\mathscr{R}^j$ have different state variables u^j. The dependence of an embedded model $\mathscr{T}_k^0$ on these different states must be represented. To this end, each embedded model $\mathscr{T}_k^0$ is formulated to be dependent on an argument w_k that is consistent for the prediction and all experiments. There is then a mapping defined by the operator $\mathscr{W}_k^j$ that maps the state variable u^j to the argument w_k.

With these extensions, we can now give a generalized version of the problem statement described in Section 3.1.1:

$$\begin{aligned} \mathscr{R}(u, \{\tau\}^0, r) &= 0 \\ q &= \mathscr{Q}(u, \{\tau\}^0, r) \\ \mathscr{R}^j(u^j, \{\tau\}^j, r^j) &= 0 \qquad \text{for } 1 \le j \le N_e \\ v^j &= \mathscr{V}^j(u^j, \{\tau\}^j, r^j) \qquad \text{for } 1 \le j \le N_e \end{aligned} \tag{20}$$

In this formulation, the embedded models required for the prediction ($j = 0$) and the validation scenarios ($1 \le j \le N_e$) are given by

$$\tau_k^j = \mathscr{T}_k^j(\mathscr{W}_k^j(u), \theta_k^j, s_k^j) \qquad \text{for } 1 \le k \le N_\tau^j \tag{21}$$

where N_τ^j is the number of embedded models in scenario j.

Like the simple problem statement from Section 3.1.1, in this generalized problem the high-fidelity theory forming the basis of the model can enable reliable predictions, despite the need to extrapolate from available data. However, additional complexity arises from confounding introduced by the presence of multiple embedded models in the validation experiments. To avoid this confounding, one would ideally use experiments where there are no extra embedded models beyond those needed for the prediction or where any such extra embedded models introduce small error or uncertainty in the context of the experiment. Of course, the assessment of any extra embedded models would itself form another validation exercise.

To further avoid confounding uncertainties, it is preferable to use experiments in which only one of the embedded models used in the prediction model is exercised. Such experiments are powerful because they provide the most direct assessment possible of the embedded model in question. However, even if experiments that separately exercise all of the embedded models necessary for prediction are available, in general these experiments alone are not sufficient for validation because they cannot exercise couplings and interactions between the modeled phenomena. This fact leads to the idea of a validation pyramid [6], as discussed in Section 2.3.

3.2 Validation for Prediction

A fundamental challenge in the validation of predictions is that even if a model is consistent with all available data, as determined by the techniques discussed in Section 2, this does not imply that the model is valid for making predictions. The reason is that the prediction QoI may be sensitive to some error or omission in the model that the observed quantities are not. To preclude this possibility and gain confidence in the prediction, further assessments of the validation process are needed. These are discussed in Section 3.2.2 below.

The opposite situation represents a different fundamental challenge in the validation of predictions; that is, even if a model is found to be inconsistent with available data, this does not imply that the model is invalid for making the desired predictions. The reason is that the prediction QoI may be insensitive to the error or omission in the model that caused the inconsistency with the observations. To determine whether a prediction can be made despite the errors in the model requires that the impact of the modeling errors on the predicted QoI be quantified. This quantification of uncertainty due to model inadequacy is discussed in Section 3.2.1.

3.2.1 Accounting for Model Error

If a discrepancy between a model and observations is detected, it may none-the-less be possible to make a reliable prediction, provided the impact of model error responsible for this discrepancy on the predicted QoI can be quantified. This can be difficult because there is no direct mapping from the observables to the QoIs—i.e., given only y, one cannot directly evaluate q. Referring to the problem statement in Section 3.1.1, it is clear that any model errors must be due to the embedded model $\mathscr{T}$, since all the other components of the model ($\mathscr{R}$, $\mathscr{V}$ and $\mathscr{Q}$) are presumed to be reliable. In essence, we need to enrich the embedded model to include a representation for the uncertainty introduced by model errors. In the simple case from Section 3.1.1 one could write

$$\tau \approx \mathscr{T}(u,s;\theta) + \mathscr{E}_\tau(u,s;\alpha), \qquad (22)$$

where $\mathscr{E}_\tau$ is an uncertainty representation of the model error ε_τ, which may depend on additional parameters α. Given our choice to use probability to represent uncertainty, it is natural that $\mathscr{E}_\tau$ 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 $\mathscr{E}_\tau$ must be determined. The specification of a stochastic model $\mathscr{E}_\tau$ 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 (22) is introduced into (16) so that it can be solved for u, which is now stochastic, the fact that $\mathscr{E}_\tau$ depends on u will in general make this solution very difficult. In practice, we have either formulated $\mathscr{E}_\tau$ to be independent of u or have defined $\mathscr{E}_\tau$ through an auxiliary equation of the form $f(u,\mathscr{E}_\tau;z)=0$, where z is an auxiliary random variable that is independent of u. In this later case, the auxiliary equation can then be solved together with (16). Other practical formulations for introducing u dependence in $\mathscr{E}_\tau$ 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 $\mathscr{E}_\tau$ 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 general case corresponding to (20):

$$\begin{aligned} \mathscr{R}(u,\{\mathscr{T}\}^0+\{\mathscr{E}_\tau\}^0,r)&=0 \\ q=\mathscr{Q}(u,\{\mathscr{T}\}^0+\{\mathscr{E}_\tau\}^0,r)& \\ \mathscr{R}^i(u^i,\{\mathscr{T}\}^i+\{\mathscr{E}_\tau\}^i,r^i)=0 &\qquad \text{for } 1\le i\le N_e \\ v^i=\mathscr{V}^i(u^i,\{\mathscr{T}\}^i+\{\mathscr{E}_\tau\}^i,r^i) &\qquad \text{for } 1\le i\le N_e \end{aligned} \tag{23}$$

The inadequacy models, $\{\mathscr{E}\}^0$ and $\{\mathscr{E}\}^i$, appear naturally in the calculation of both the observables and the QoIs, both directly through the possible dependence of $\mathscr{V}^i$ and $\mathscr{Q}$ on embedded models, and indirectly via the dependence of the state u on the embedded models appearing in $\mathscr{R}$. The structural uncertainty can therefore be propagated to both the observables and the QoIs without additional modeling assumptions. Furthermore, one can learn about the inadequacies—i.e., calibrate and test the corresponding models—from data on the observables and then transfer that knowledge to the prediction of the QoIs. This ability enables quantification of the impact of modeling inadequacies on the unobserved QoIs.

Enriching the embedded models with representations of the uncertainty due to model inadequacy is done with the goal of explaining all observed discrepancies between the model and observations. Therefore, with these enrichments included, the validation process discussed in Section 2 should reveal no inconsistencies with all relevant data. Once this is confirmed, there is no longer a validation failure, and we can proceed to evaluating whether the validation process is sufficient to warrant confidence in predictions of the QoIs.

3.2.2 Predictive Assessment

Since prediction requires extrapolation from available information, a prediction cannot be validated based on agreement between the predictive model (or some part of it) and data. This agreement alone is only sufficient to determine that the model is capable of predicting the observed quantities in the observed scenarios. To go beyond this, additional knowledge about the model and its relationship to both the validation experiments and the prediction are required. In particular we must determine whether:

1. the calibration and validation of the embedded models is sufficient to give confidence in the prediction;
2. the embedded models are being used within their domain of applicability; and
3. the resulting prediction with its uncertainties is sufficient for the purpose for which the prediction is being made.

These predictive assessments are outlined in the following paragraphs and in more detail in [30].

3.2.2.1 Adequacy of Calibration & 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 determine 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 adequately informed and tested in the preceding calibration and validation processes. 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. There are several ways in which the calibration and validation processes might be found to be insufficient. The most relevant examples are described briefly below.

1. Suppose that the prediction QoI is highly sensitive to one of the embedded models $\mathscr{T}$, as measured, for example, by the Fréchet derivative of the QoI with respect to $\mathscr{T}$ at some representative θ. If none of the validation quantities are sensitive to $\mathscr{T}$, then the validation process has not provided a test of the validity

of $\mathscr{T}$, and a prediction based on $\mathscr{T}$ would be unreliable. More plausibly, it may be that none of the validation quantities for scenarios higher in the validation pyramid are sensitive to $\mathscr{T}$. The integration of $\mathscr{T}$ into a composite model similar to that used in the predictions would then not have been tested, which would make its use in the prediction suspect [30]. To guard against this and similar possible failures of $\mathscr{T}$, the predictive assessment process should determine whether validation quantities in scenarios "close enough" to the prediction scenario are sufficiently sensitive to $\mathscr{T}$ 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.

2. 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.
3. Suppose that uncertainty in the prediction QoI is largely due to the uncertainty model $\mathscr{E}$ representing the inadequacy of the embedded model $\mathscr{T}$. In this case, it is important to ensure that $\mathscr{E}$ is a valid description of the inadequacy of $\mathscr{T}$. 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 $\mathscr{E}$, then the veracity of $\mathscr{E}$ 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.

3.2.2.2 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 §3.1, 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.

3.2.2.3 Sufficiency of the Prediction and Uncertainties

The focus of the previous assessments is 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 the current discussion, 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.

3.2.2.4 A Major Caveat

The predictive assessment process can determine whether, given what is known about the system, the calibration and validation process are sufficient to make a reliable prediction. But, the well-known problem of "unknown unknowns" remains. 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. The 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 Conclusions and Challenges

As the importance of computational simulations in science and engineering continues to grow so does the importance of validating the physical models that form the basis of those simulations. Validation is traditionally defined as a comparison between model outputs and experimental observations intended to reveal any important discrepancies between the model and reality. To make this process rigorous, one must account for uncertainties that affect the experimental observations and the computational results. Thus, in order to draw validation conclusions, it is necessary to define metrics that measure the agreement or lack thereof between uncertain experimental observations and uncertain model outputs. When these uncertainties are represented using probability, a number of such "validation metrics" are available, including highest posterior density credibility intervals and Bayesian p-values, both of which can be used in combination with appropriately chosen test quantities when necessary or desirable.

When the purpose of the computational simulation is prediction, agreement between uncertain model outputs and available uncertain data is in general necessary but not sufficient for validating the prediction because prediction requires extrapolation. In this situation, predictive validation is a process for building confidence in simulation-based predictions by exploiting typical features of physics-based models.

A number of issues remain before systematic validation methodologies like those described here can become standard in computational science and engineering. First, all of the ideas described here depend heavily on the development of probabilistic models to represent uncertainties. In some situations, such as when abundant sample data are available for an aleatorically uncertain variable, these models are straightforward to build. However, in many cases, particularly those involving complex epistemic uncertainties, this process is less clear. For example, general techniques and best practices for representing uncertainty due to model inadequacy, particularly when the modeled quantity is a field, and for representing correlations between experimental measurements when few replications are available must be developed. These difficulties are often related to the more general problem of representing qualitative information such as expert opinion that, while often crucial in accurately characterizing likely values of epistemic parameters or realistic modeling errors, can be challenging to represent quantitatively in a defensible manner.

Second, the methods discussed here require uncertainty propagation through the models being validated. When these models are computationally expensive and/or the space of uncertain variables is high-dimensional, it is well-known that typical algorithms, such as Monte Carlo sampling or stochastic collocation, often require too many forward model evaluations to be computationally tractable. Better algorithms are necessary to enable routine uncertainty analysis using complex models. The required algorithmic advances are also necessary to enable routine validation of these models.

References

1. Firm uses doe's fastest supercomputer to streamline long-haul trucks. Office of Science, U.S. Dept. of Energy, Stories of Discovery and Innovation (2011). URL `http://science.energy.gov/discovery-and-innovation/stories/2011/127008/`
2. Adams, B.M., Ebeida, M.S., Eldred, M.S., Others: Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 6.2 Users Manual. Sandia National Laboratories, Albuquerque, New Mexico (2014). URL `https://dakota.sandia.gov/documentation.html`
3. 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)
4. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Statistics Surveys **4**, 40–79 (2010). DOI 10.1214/09-SS054. URL `http://dx.doi.org/10.1214/09-SS054`
5. ASME Committee V&V 10: Standard for Verification and Validation in Computational Solid Mechanics. ASME (2006)
6. Babuška, I., Nobile, F., Tempone, R.: Reliability of computational science. Numerical Methods For Partial Differential Equations **23**(4), 753–784 (2007). DOI 10.1002/num.20263
7. Box, G., Tiao, G.C.: Bayesian Inference in Statistical Analysis. New York: Wiley Classics (1973)
8. Box, G.E.P.: Sampling and bayes' inference in scientific modeling and robustness. Roy. Statist. Soc. Ser. A **143**, 383–430 (1980)
9. Cox, R.T.: The Algebra of Probable Inference. Johns Hopkins University Press, Baltimore, MD (1961)
10. Cui, T., Martin, J., Marzouk, Y.M., Solonen, A., Spantini, A.: Likelihood-informed dimension reduction for nonlinear inverse problems. Inverse Problems **30**(11), 114,015 (2014)

11. Dubois, D., Prade, H.: Possbility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
12. Ferson, S., Ginzburg, L.R.: Different methods are needed to propagate ignorance and variability. Reliability Engineering and System Safety **54**, 133–144 (1996)
13. Fine, T.L.: Theories of Probability. Academic Press, New York (1973)
14. Gelman, A.: Comment: 'bayesian checking of the second levels of hierarchical models'. Statistical Science **22**, 349–352 (2007). DOI doi:10.1214/07-STS235A
15. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. CRC Press, Boca Raton, FL (2014)
16. Gelman, A., Meng, X.L., Stern, H.: Posterior predictive assessment of medel fitness via realized discrepancies. Statistica Sinica **6**, 733–807 (1996)
17. Gelman, A., Rubin, D.B.: Avoiding model selection in bayesian social research. Sociological Methodology **25**, 165–173 (1995)
18. Gelman, A., Shalizi, C.R.: Philosophy and the practice of bayesian statistics. British Journal of Mathematical and Statistical Psychology **66**(1), 8–38 (2013)
19. Hsu, J.: Multiple Comparisons: Theory and Methods. Chapman and Hall/CRC (1996)
20. Hyndman, R.J.: Computing and graphing highest density regions. The American Statistician **50**(2), 120–126 (1996)
21. Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press (2003)
22. Kanji, G.K.: 100 Statistical Tests, 3rd. Edition. Sage Publications (2006)
23. Le Maıtre, O., Knio, O., Najm, H., Ghanem, R.: Uncertainty propagation using wiener–haar expansions. Journal of Computational Physics **197**(1), 28–57 (2004)
24. Li, J., Marzouk, Y.M.: Adaptive construction of surrogates for the bayesian solution of inverse problems. SIAM Journal on Scientific Computing **36**(3), A1163–A1186 (2014)
25. Miller, L.K.: Simulation-based engineering for industrial competitive advantage. Computing in Science and Engineering **12**(3), 14–21 (2010). DOI 10.1109/MCSE.2010.71
26. Miller, R.G.J.: Simultaneous Statistical Inference. Springer; 2nd edition (1981)
27. Najm, H.N.: Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics. Annual Review of Fluid Mechanics **41**, 35–52 (2009)
28. Oberkampf, W.L., Helton, J.C., Sentz, K.: Mathematical representation of uncertainty. AIAA 2001-1645
29. Oden, J.T., Belytschko, T., Fish, J., Hughes, T.J.R., Johnson, C., Keyes, D., Laub, A., Petzold, L., Srolovitz, D., Yip, S.: Revolutionizing engineering science through simulation: A report of the National Science Foundation blue ribbon panel on simulation-based engineering science (2006). URL `http://www.nsf.gov/pubs/reports/sbes_final_report.pdf`
30. Oliver, T.A., Terejanu, G., Simmons, C.S., Moser, R.D.: Validating predictions of unobserved quantities. Computer Methods in Applied Mechanics and Engineering **283**, 1310 – 1335 (2015). DOI http://dx.doi.org/10.1016/j.cma.2014.08.023. URL `http://www.sciencedirect.com/science/article/pii/S004578251400293X`
31. Petra, N., Martin, J., Stadler, G., Ghattas, O.: A computational framework for infinite-dimensional bayesian inverse problems, part ii: Stochastic newton mcmc with application to ice sheet flow inverse problems. SIAM Journal on Scientific Computing **36**(4), A1525–A1555 (2014)
32. Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Statist. **12**, 1151–1172 (1984)
33. Sentz, K., Ferson, S.: Combination of evidence in Dempster–Shafer theory. Tech. Rep. SAND 2002-0835, Sandia National Laboratory (2002)
34. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton, New Jersey (1976)
35. Van Horn, K.S.: Constructing a logic of plausible inference: A guide to Cox's theorem. International Journal of Approximate Reasoning **34**(1), 3–24 (2003)