ICES REPORT 12-32

August 2012

Validating the Prediction of Unobserved Quantities

by

Robert D. Moser, Gabriel Terejanu, Todd A. Oliver, and Christopher S. Simmons

The Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, Texas 78712

Reference: Robert D. Moser, Gabriel Terejanu, Todd A. Oliver, and Christopher S. Simmons, Validating the Prediction of Unobserved Quantities, ICES REPORT 12-32, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, August 2012.

Validating the Prediction of Unobserved Quantities

Robert D. Moser, Gabriel Terejanu, Todd A. Oliver, and Christopher S. Simmons

Center for Predictive Engineering and Computational Sciences,
Institute for Computational Engineering and Sciences,
The University of Texas at Austin, Austin TX, 78712,
rmoser@ices.utexas.edu

August 7, 2012

Abstract

In predictive science, computational models are used to make predictions regarding the response of complex systems. Generally, there is no observational data for the predicted quantities (the quantities of interest or QoIs) prior to the computation, since otherwise predictions would not be necessary. Further, to maximize the utility of the predictions it is necessary to assess their reliability—i.e., to provide a quantitative characterization of the discrepancies between the prediction and the real world. Two aspects of this reliability assessment are judging the credibility of the prediction process and characterizing the uncertainty in the predicted quantities. These processes are commonly referred to as validation and uncertainty quantification (VUQ), and they are intimately linked. In typical VUQ approaches, model outputs for observed quantities are compared to experimental observations to test for consistency. While this consistency is necessary, it is not sufficient for extrapolative predictions because, by itself, it only ensures that the model can predict the observed quantities in the observed scenarios. Indeed, the fundamental challenge of predictive science is to make credible predictions with quantified uncertainties, despite the fact that the predictions are extrapolative. At the PECOS Center, a broadly applicable approach to VUQ for prediction of unobserved quantities has evolved. The approach incorporates stochastic modeling, calibration, validation, and predictive assessment phases where uncertainty representations are built, informed, and tested. This process is the subject of the current report, as well as several research issues that need to be addressed to make it applicable in practical problems.

1 Introduction

Thanks to advances in computing capabilities in recent decades and accompanying advances in the fidelity of computational models, it is now possible to simulate physical phenomena and systems of unprecedented complexity. This capability has the potential to revolutionize engineering and science by enabling detailed predictions to inform critical design and policy decisions that directly affect the safety, welfare, and security of individuals and society as a whole. For example, results of computational simulations are heavily used in the design of nearly all complex engineering systems from consumer electronics to spacecraft to nuclear power plants. Furthermore, predictions from computational models are used to inform policy decisions in areas where the consequences of inaccurate predictions and poorly-informed decisions could be catastrophic, such as disaster response and climate change. However, to maximize their utility, it is critical that the reliability of

predictions from computational models be systematically characterized and provided to decision-makers in a useful form. This reliability characterization is the goal of the processes discussed here.

The study of the reliability of predictions from computational models is often divided into three parts: verification, uncertainty quantification (UQ), and validation (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. While verification is vitally important and sometimes overlooked in practice, it is largely understood [41, 32]. In the development of the validation and UQ processes discussed here, it is assumed that all numerical solutions have been verified to ensure that uncertainties in the predictions due to numerical errors are small compared to other sources. Therefore, verification will not be discussed further in this report.

UQ is the process of assessing uncertainties that affect simulation predictions, such as uncertainties in model inputs or errors in the model form, and propagating these uncertainties to determine the resulting uncertainty in the quantities being predicted (the quantities of interest or QoIs). Finally, following [1, 2, 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.

In engineering practice, the "validity" of a computational model is often assessed by simply comparing the output of the model with experimental data. This straightforward process has several problems as a basis for determining the quality of model predictions. First, it does not account for the uncertainties associated with the model or the data. Second, such comparisons do not assess how well the model will perform in a new situation or when predicting an unobserved quantity. However, it is common to use models to make such predictions—i.e., predictions for scenarios or of quantities for which data is unavailable—because this is precisely when predictions are valuable.

To address the shortcomings of naive comparisons with data, a number of more sophisticated procedures have been proposed [44, 6, 33, 5, 40, 41, 32, 12, 20, 43, 29], and validation guidelines have been developed by professional engineering societies [1, 2, 3]. While these have generally been positive developments, 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. [20] and Bayarri et al. [12] present similar validation frameworks for computational models. These frameworks rely on statistical models, specifically Gaussian processes, to represent the difference between the model outputs and observational data. In particular, the "true" observable function is given by the model output plus a Gaussian process representing model error. Such model discrepancy models were introduced by Kennedy and O'Hagan [24] to account for model imperfections during calibration and prediction. However, such a representation is insufficient for validation. Since the discrepancy model is posed only for the observable quantities and since there is generally no direct mapping from the observables to the QoIs, this representation cannot be used directly for UQ or validation for predictions of unobserved QoIs. Furthermore, since the discrepancy model is a purely statistical model, it is highly dependent on calibration against observations. Such models should not be used in regions where they cannot be trained and tested. In general then, use of the Kennedy and O'Hagan formulation of discrepancy models is suspect in extrapolative predictions.

In a pioneering paper, Babuška et al. [5] address validation for predictions of unobserved QoIs. Their approach implicitly assumes that the model parametrization is sufficiently rich that the model parameters can always be adjusted to fit the data well. Thus, by calibrating the model parameters

with different data sets and then predicting the QoIs with the different calibrated values, one can investigate the sensitivity of the QoI to the parameter changes necessary to fit different data. However, it is often the case in engineering that the model of interest cannot be calibrated to fit all existing observations, even independently. In this situation, one is unable to assess the impact of the observed discrepancy on the QoI prediction using the Babuška et al. method. Thus, while the method is valuable for some problems, it is not sufficiently general.

Of course, the failure of the model to fit available data even after calibration is a symptom of model inadequacy, and it is tempting to conclude that such a model is invalid. However, this conclusion is not justified because failure to perfectly fit available data does not imply that the model is unable to predict the QoI with sufficient accuracy. The challenge is to determine when predictions of the QoIs are justified or not justified, in light of the observed discrepancies with available data.

To address this challenge, a broadly applicable approach to validation of predictive simulations, which we call "predictive validation", has been developed. Specifically, the approach takes advantage of the common structure of models for physical systems in which a highly reliable physical theory is augmented by less reliable "embedded models." This structure and its importance are discussed further in §2. Given such a model, predictive validation involves four distinct activities: uncertainty modeling (§3), calibration and model selection (§4), validation (§5), and predictive assessment (§6). Finally, a number of research issues that need to be addressed are discussed in §7.

2 Predictive Validation

Currently there is no consensus regarding how extrapolative predictions should be validated or if such validation is even possible. For instance, Oreskes et al. [36] argue that validation of numerical models of natural systems is impossible. While their terminology is somewhat different from that used here, they essentially argue that it is impossible to completely confirm a model—i.e., to logically prove that it is true—because one never has perfect access to reality and because of the logical fallacy of affirming the consequent. While this argument is correct, the practical goal of validation as pursued here is not to prove that a given model is absolutely correct. Rather, we aspire only to determine whether the model is a sufficiently accurate representation of reality to be useful for the intended purpose, and even this will generally be impossible to show unequivocally.

Validation of extrapolative predictions is difficult because it requires one to use existing information to draw conclusions regarding the accuracy of predictions outside the range of the data. To accomplish this, one cannot rely on the data alone; predictive validation requires information about the nature of the extrapolation and known modeling errors in addition to experimental data to enable reliable extrapolative predictions. We discuss here the specific conditions that entitle us to make extrapolative predictions for engineering applications using physics-based models, as well as some processes to gain confidence that these conditions are met. These processes are intended to assess whether the quantified uncertainties in the QoIs are consistent with observational data and existing scientific knowledge regarding the process being simulated as well as whether low-fidelity components of the model are used within their range of applicability.

To enable credible, extrapolative predictions, we will take advantage of the fact that the predictions are to be made for 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. This fact is key in the development of the proposed methodology and is often overlooked, leading to the pessimistic view that any extrapolation is suspect. Of course, if the entire model were known *a priori* to be highly reliable, there would be no validation question.

The difficulties arise because these highly reliable theories are generally 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 inter-atomic potentials. We will refer to such models—i.e., high-fidelity models with lower-fidelity embedded components—as composite models.

The fact that the composite models used for prediction are based on highly reliable theory allows the possibility of predictive simulation in the approach described here, despite the use of less reliable embedded models. In essence, it is the highly reliable theory that will be extrapolated to predict the QoIs for the scenario of interest. The composite model can therefore yield reliable extrapolative predictions, provided that the embedded models are not being used outside of the regime in which they have been calibrated and tested. It is possible to extrapolate while not leaving the calibration range of the embedded model because the scenario spaces of the composite model and the embedded models are generally different.

As an example, consider simulation models in continuum mechanics, which are based on expressions of mass, momentum and energy conservation, the reliability of which is generally not in question in the context of the predictions being made. These theories are augmented with constitutive models for internal stresses, conductive heat flux, etc. They may also require models of the boundary conditions. The constitutive models will, for example, have been formulated, calibrated and tested for some range of strains or strain rates, temperatures, and temperature gradients. Provided that for the predictions, the composite model only exercises the constitutive models under these conditions, the predictions will be reliable.

To make these ideas more concrete, let us consider the following mathematical model for a physical system:

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

where $\mathcal{R}$ is some operator, u is the solution or state, τ is a quantity for which an embedded model is required, and r is a set of scenario variables needed to precisely define the case being considered. We have in mind that in continuum mechanics, for example, $\mathcal{R}$ would be partial differential equations representing mass, momentum and energy conservation, while τ would be the internal stress for which a constitutive model is needed. In fluid mechanics, τ might also include the apparent stress due to turbulent fluctuations (the Reynolds stress). Finally, the scenario variables r would include required auxiliary data (e.g., boundary conditions) and parameters (e.g., the Reynolds number in fluid mechanics).

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. But, of course, a model for τ is needed:

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

where $\approx$ indicates that the model form is imperfect and θ denotes calibration parameters. Both the model form error and incomplete knowledge of θ lead to errors in τ and thus errors in u. However, note that the scenario variables (denoted s in (2)) describing the applicability of m may be different from those for $\mathcal{R}$. For example, if $\mathcal{R}$ is a PDE and m is a local, algebraic model, then the boundary data, which is included in r, will not appear in s. Of course, s can still depend implicitly on r since s will generally be expressed in terms of u, which depends on r. None-the-less, it will often be possible to extrapolate in some dimensions of r while interpolating in s.

For the purposes of model calibration and validation, we suppose that some observable quantities y can be measured experimentally or computed accurately from a much more reliable model (e.g.,

direct numerical simulation of turbulence, or *ab initio* quantum calculations). These observable quantities are different from the prediction QoIs q, but both y and q are determined from the solution, the embedded model, and the scenario:

$$y = \mathcal{Y}(u, \tau; r), \tag{3}$$
$$q = \mathcal{Q}(u, \tau; r), \tag{4}$$

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

This abstract problem statement is the simplest that encompasses the critical features of an unreliable embedded model and a unobserved QoI and is sufficient to describe our approach to predictive validation. Many generalizations to include the situations arising in real problems are straightforward. Some of the most obvious generalizations include the introduction of multiple imperfect embedded models for different phenomena, and the use of different scenario parameters or even different model forms $\mathcal{R}$ for calibration, validation and prediction, provided only that the same quantity τ needs to be modeled in each case.

Equations (1) through (4) form a closed model enabling prediction of both observables and QoIs for any scenario of interest. A key challenge in validating extrapolative predictions is to determine the implications of observed discrepancies between the model outputs and experimental data on the uncertainty in the QoI predictions and, ultimately, on the decision to be supported by the predictions. This is accomplished by representing the uncertainty due to the error in the inadequate model m and propagating that uncertainty to the QoI. The uncertainty due to model inadequacy can, for example, be introduced as follows:

$$0 = \mathcal{R}(u, m + \epsilon_m; r) \tag{5}$$
$$y = \mathcal{Y}(u, m + \epsilon_m; r) \tag{6}$$
$$q = \mathcal{Q}(u, m + \epsilon_m; r) \tag{7}$$

where $\epsilon_m(u; \alpha, s)$ is a model characterizing the error in the physical model m, which is uncertain. Like m, the model ϵ_m has calibration parameters α that need to be determined from observations. If the inadequacy model ϵ_m faithfully represents the discrepancies between the model for the observables $\mathcal{Y}$ and the observations, we then use it in $\mathcal{Q}$ to predict how the observed discrepancies impact uncertainty in the QoI. The VUQ process needed to assess the accuracy and credibility of the predictions then involves four activities:

1. **Uncertainty Modeling:** Uncertainties in model predictions arise from uncertainties in the scenario and model parameters r, s, θ, and α; from errors introduced by the imperfection of the embedded model m; and from errors in the measurements of the observables y. Mathematical representations of these uncertainties must be specified.

2. **Calibration and Model Selection:** The calibration parameters (θ and α) in the models m and ϵ_m, and their uncertainties, are determined from observational data. Generally, these calibration observations are taken in the simplest relevant scenarios to avoid the confounding effects of many uncertain parameters from multiple embedded models. This process also provides an opportunity to select the model that is best supported by the data from among multiple candidate models.

3. **Validation:** The selected parametrized model, along with the uncertainty models, are tested against all relevant observations. These include the calibration data, as well as new observations intended primarily for validation. The validation data are often taken in more complex

scenarios involving multiple embedded models for different phenomena. The primary validation criterion is whether each datum could plausibly have arisen from the models, given the uncertainties in the data, models, and parameters.

4. **Predictive Assessment:** Models and uncertainty models that plausibly account for all the data are an important necessary condition for making credible predictions of unobserved QoIs, but they are not sufficient. There are several questions regarding the prediction process that must also be answered. For the prediction, are the embedded models and associated uncertainty models being used outside their "domain of applicability;" that is, in scenarios for which they have not been well calibrated or tested? Are the prediction QoIs sensitive to uncertainties to which the calibration and validation observations are not? Finally, are uncertainties in the QoIs too large for the intended purpose of the prediction? If any of these are answered in the affirmative, then the physical and/or uncertainty models are insufficient for the required predictions.

There are a number of challenges to this program of predictive validation. The four primary activities, along with associated research challenges and proposed approaches, are discussed in more detail in the following sections.

3 Uncertainty Modeling

To pursue the VUQ program described here, the first requirement is a mathematical representation of uncertainty, or incomplete information. A wide range of approaches have been proposed based on such formalisms as probability theory, Dempster-Shafer evidence theory, fuzzy sets, interval analysis, worst-case scenarios, and many more. Probability, which will be used here, provides a particularly powerful representation of uncertainty. In this context, probability is used to encode the degree of certainty in (or plausibility of) a proposition, and the laws of probability, such as Bayes' theorem, constitute a "logic of plausibility," analogous to Boolean algebra for deterministic logic.

In fact, it can be shown [15, 22, 21] that any system of plausible inference in which plausibility is represented by a single real number and which satisfies three logical conditions must be re-scalable to probability. Stated informally, the conditions are: 1) that increasing the plausibility of a proposition A decreases the plausibility of the negation of A, while not decreasing the plausibility of the proposition (A and B), for any proposition B; 2) that any two valid chains of inference for the proposition A based on equivalent states of information yield the same plausibility for A; and 3) that the same logic of plausible inference apply to all problem domains. Van Horn [21] presents a formal statement of these conditions.

A second important ingredient for our representation of uncertainty is a quantitative measure of the uncertainty represented by a probability distribution. Such a measure will allow us to determine a probability distribution for a quantity that is maximally uncertain under the constraints imposed by our knowledge about the quantity. It will also allow us to measure how uncertainty is reduced when new data is introduced. Shannon [45] showed that the "information entropy" $H = -\sum_i \log(p_i) p_i$ is the unique measure of uncertainty for a discrete random variable that satisfies the desirable properties of continuity, symmetry and additivity, with maximum uncertainty occurring for a uniform probability distribution across the possible outcomes. For continuous random variables, which we will be primarily concerned with, the entropy generalizes to the entropy

of the probability distribution function p relative to a reference distribution μ:

$$H(p,\mu) = \int p(x) \log \frac{p(x)}{\mu(x)}\, dx, \tag{8}$$

which is the Kullback-Leibler divergence. It measures the reduction in uncertainty that occurs in updating the probability distribution from μ to p.

So, we are led by the logical conditions described above to a probabilistic representation of the degree of certainty; but, it remains to determine how to specify quantitative probabilities that encode our state of incomplete information in some problem domains. In what follows, there are three situations in which such probability assignments are needed: 1) specification of prior information; 2) specification of uncertainties in data; and, 3) representation of the uncertainty due to model inadequacy.

A probabilistic representation of prior information is a required input to Bayesian inference, which is used in the calibration and model selection processes described in §4. The need for priors is commonly viewed as an nuisance in Bayesian statistics. But, it is an important part of inference, providing the means to introduce a wide variety of knowledge we may have about a problem, such as physical or logical constraints, accumulated experience or expert opinion. The challenge is to express this often qualitative information probabilistically. A powerful tool for this purpose is maximization of information entropy. By maximizing the entropy subject to the constraints implied by prior information we will find prior PDFs consistent with this information, while otherwise retaining maximum uncertainty. There are however two issues in the specification of maximum entropy distributions that need to be addressed. First, for continuous variables, the entropy is defined relative to a distribution μ describing maximum ignorance [22], as shown in (8). Such ignorance distributions can, for example, be determined from the invariance properties implied by ignorance for the problem at hand [22]. Such considerations lead to the Jeffreys distribution ($\mu(x) \sim 1/x$) for a scaling parameter [23, 22]. Probabilistic expressions of ignorance need to be developed for more situations. Second, mathematically expressing prior knowledge so that it can serve as a constraint in a maximum entropy calculation is often challenging. For example, how does one express the information contained in contradictory expert opinions, or from accumulated experience in similar problems? As another example, consider inference involving uncertain fields (e.g., initial or boundary conditions), in which we must express prior constraints regarding the smoothness or other spatial characteristics of the field [46].

The second challenge in uncertainty modeling that we will need to address is characterizing the uncertainty in observational data. It is standard practice for experimentalists to assess the uncertainties in their measurements, but such assessments are commonly incomplete. In experiments on complex systems, one rarely is able to directly measure the quantities one wishes to measure (e.g., the velocity of a fluid at a point in space), rather some related quantity is measured (e.g., the voltage across a hot-wire anemometer) and the desired quantity is inferred through a "data reduction model." Unfortunately, experimental uncertainty assessments often do not include uncertainties introduced by this inference. At PECOS, in some cases we have found it necessary to develop more reliable data reduction models to reduce and better characterize these uncertainties [37, 8]. Experimental uncertainty assessments also do not generally include characterizations of the dependencies between different data points (e.g., in sampling spatial dependence). Representing such data dependencies is critically important to making valid inferences from the data.

The final uncertainty modeling challenge we face is representing the uncertainties introduced by the inadequacy of the physical models we use (e.g., ϵ_m in (5)). Here we face challenges similar to those for the prior, that is to encode our qualitative knowledge about the nature of the inadequacy. For example, this knowledge might take the form of constraints that the modeled quantity must

satisfy, or when the model is for a field quantity, it might include regularity expectations for the modeled quantity. As an example, consider models for the Reynolds stress in Reynolds Averaged Navier-Stokes simulations of turbulent flows [38, 16], which is a symmetric second-rank tensor that must be positive definite, and is known to be smooth with variations on the same length scale as the average velocity. In our experience at PECOS, even when these constraints were satisfied, we often found that the Reynolds stress model plus a draw from the inadequacy model was obviously (to a domain expert) an unrealistic Reynolds stress field. In these cases, all that was known about the model and its inadequacy was apparently not incorporated into the inadequacy models. A particularly promising approach to address these shortcomings in some contexts is to pose the inadequacy as the solution to a stochastic differential equation, which places the uncertainty modeling problem in a setting in which it is easier for physical modelers to express what is known. There is also an opportunity in some modeling domains (e.g., chemical mechanisms, turbulent combustion) for inadequacy representations to be formulated based on *a posteriori* estimates of macroscopic modeling error based on an underlying microscopic model (see e.g., [7]) or a model hierarchy.

Model inadequacy models generally have parameters, which like parameters in the physical models, must be calibrated. Furthermore, given that prior information regarding the model inadequacy is often weak, it is usually possible to formulate multiple inadequacy models that satisfy prior constraints and equally represent the knowledge of a domain expert. In this situation, model selection is crucial to determining what uncertainty representation should be used for prediction. Calibration and model selection are described in §4.

4 Model Calibration & Selection

Once a model or set of models, including model inadequacy representations, have been developed, the parameters in these models must be determined in light of the available prior information and observations. Further, if multiple models have been posed, one needs to determine which model or models are best support by available data. These are the activities of model calibration and model selection, which, consistent with our use of probability to represent uncertainty, we pursue using Bayesian inference. In this Bayesian probabilistic treatment, the free parameters in the the models are treated as "random" variables, though they are not in general inherently random. Instead, our imperfect knowledge about these parameters is characterized by their associated probability distributions. Given these PDFs (probability distribution functions), one updates one's knowledge by incorporating new data through Bayes' theorem.

In the calibration process, Bayesian inference is used to update uncertain estimates of model parameters. In this context, it provides a generalized framework for fitting the experimental data. We prefer the Bayesian approach over traditional deterministic calibration for two primary reasons. First, unlike deterministic estimates, the solution of the Bayesian inverse problem is a complete PDF describing what parameter values are consistent with prior information and experimental observations. Second, the Bayesian formulation requires a complete treatment of uncertainty (as discussed in §3). This treatment is essential in making correct inferences from data. For instance, it is common to calibrate model parameters by minimize the mean square error between the data and the model outputs. This procedure is justified provided the errors (uncertainties) in each of the data values are independent and identically distributed. In many cases however, these assumptions are not satisfied.

Measures of goodness of fit, such as mean square error, are also generally an inappropriate basis for comparison of models. In addition to the problem of correlated data uncertainties described above, comparison based on a measure of the quality of the data fit strongly favors more complex

models with many parameters that can be used to match the data. However, we are rightfully skeptical of models with many parameters because they can "fit anything." The ability of such a model to fit available data gives one little confidence that the model is in any sense "right," or that it would represent the phenomenon being modeled in any other circumstance. Further, there is a danger of "over fitting" in which the many parameters in a model allow the errors and noise in the experimental data to be fit.

Bayesian analysis again provides a better alternative: comparison of the relative probability of the models under consideration. In this approach, the relative probabilities of each model are determined via Bayesian inference. As discussed by Jaynes [22] and Muto and Beck [30], this measure favors models that fit the data well while penalizing models that rely heavily on the data to adjust parameters. This Bayesian model comparison can thus be viewed as a natural formalization of Ockham's razor, which states that given multiple models that explain the data equally well, one should prefer the simplest.

Thus, both the calibration and model comparison problems are posed here as Bayesian updates. Posterior distributions obtained in this way provide quantitative probabilistic information about what parameter values and model forms best represent the data. These procedures have been used in many modeling domains at PECOS [14, 34, 8, 35, 28].

The data is of course of primary importance in the Bayesian update process, and there are many practical issues associated with the selection of data sources for calibration and model selection. For these purposes, one commonly needs a rich set of data that covers the range of modeling scenarios expected in the predictions. For this reason experiments on simple scenarios are used in which it is relatively easy and inexpensive to make measurements, and in which embedded models for as few phenomena as possible are involved. This allows the collection of abundant data, including replicates, and facilitates calibration and model selection by avoiding the confounding effects of unreliable models for multiple phenomena. The use of simple scenarios for calibration and model selection also allows the Bayesian inference process to be pursued using model forms $\mathcal{R}_c$ that are relatively inexpensive to solve numerically. This is important because the algorithms used in Bayesian inference [27, 19, 39] require solution of the forward model at many points (often thousands or tens of thousands) in parameter space. We can think of these simple low-level observations as being at the base of a "validation pyramid" [31, 4] (see Figure 1), where they form the foundation of predictions for complex multi-physics systems, by which we mean systems that require models for multiple quantities τ.

Another practical issue that arises in calibration and model selection is choosing scenarios for calibration experiments that will maximally reduce uncertainty. This is the process of optimal experimental design. To accomplish this, we seek the set of calibration scenario parameters r_c that will maximize the expected information gain (or equivalently, expected uncertainty reduction), as measured by the relative entropy (8), that would occur when the new calibration data is included in Bayesian inference [13]. This experimental design process has been exercised at PECOS for calibration of chemical kinetics models [47], among others.

The processes of Bayesian calibration, model selection and experimental design are well established. The primary challenges to these applications are 1) consistently representing the prior information, data uncertainties (including dependencies) and model inadequacy (see §3); and 2) developing fast and efficient algorithms for Bayesian inference with models that are expensive to compute, are stochastic, and/or involve a high dimensional space of uncertain inputs (e.g., infinite dimensional for uncertain field variables).

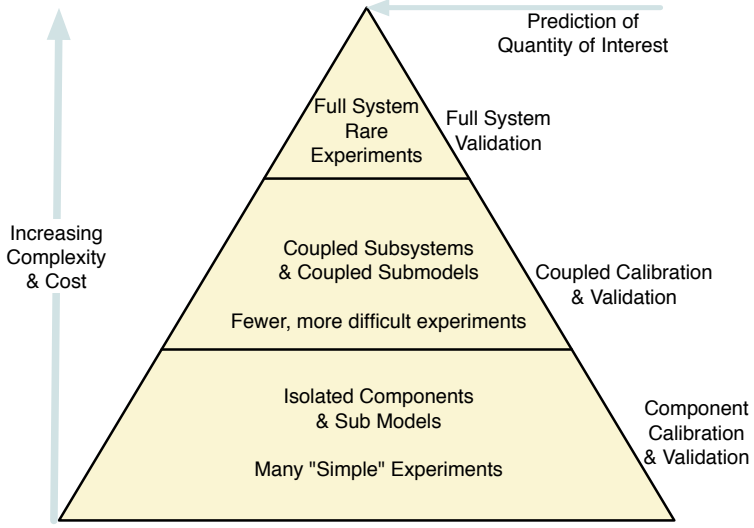

Figure 1: Representation of a "validation pyramid", showing the different type of experimental observations needed for the validation of models for complex multi-physics systems.

5 Validation

"Classical validation," or simply "validation," is a process by which models to be used in prediction are tested for consistency with observations. The modifier "classical" is introduced here to distinguish this process from the predictive validation process of which it is a part. The consistency can and should be tested with a wide range of observations relevant to the models and their use in the target predictions. These include observations in simple systems, in which just one or a few of the embedded models is active (single-physics), and observations in systems of increasing complexity in which many of the embedded models needed for prediction are important, as well as their interactions (multi-physics).

Observations, or data, used for validation can be organized into a validation pyramid [31, 4], as shown in Figure 1. As indicated in §4, the abundant data from the simple systems that make up the foundation of the pyramid are generally used to calibrate the embedded physical models and the stochastic models describing their inadequacy. Despite the fact that these observations have been used for calibration and model selection, it is none-the-less important that the consistency of the model with these observations be tested as part of the validation process. The reason is that calibration and model selection provide no guarantee that the model will represent the observations well. At PECOS, it is such validation tests using the calibration data that have most commonly failed leading to rejection of the model. Figure 2 illustrates such a case. In particular, uncertainty predictions (in the form of the $\pm$ two standard deviation interval) for mean density from a calibrated RANS turbulence model are compared against the model error for the calibration data at the maximum likelihood and posterior mean values of the parameters. Clearly, the majority of the data lies outside of the $\pm 2\sigma$ interval, indicating that the uncertainty is being underrepresented and that the model and/or its uncertainty representation are invalid.

Thus, the abundant calibration data from simple systems are important in validation also. As the process ascends the validation pyramid, increasingly scarce observations of more complex systems are used to test the consistency of coupled embedded models in scenarios more closely related to that of the predictions, and in some cases to calibrate models of the coupling. With the

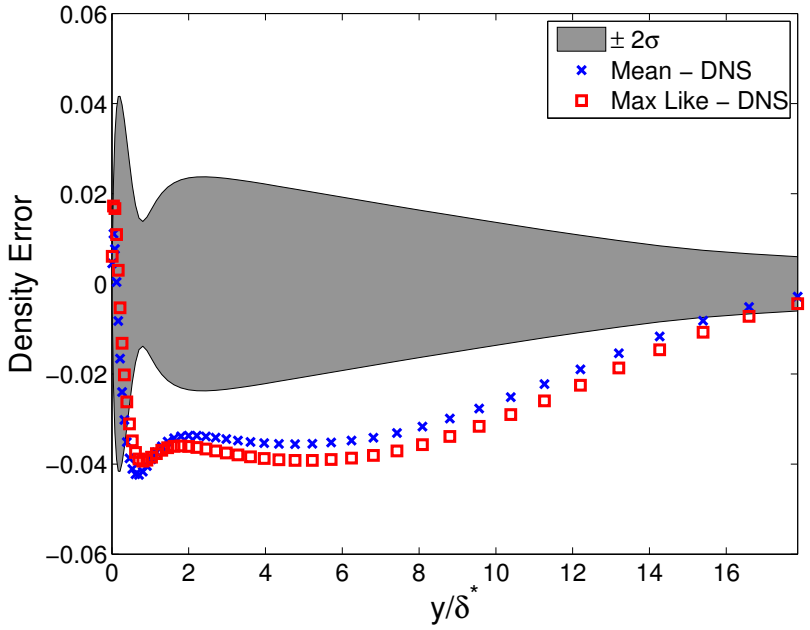

Figure 2: Predicted uncertainty ($\pm 2\sigma$) compared against observed model error at the mean and maximum likelihood parameter values.

increasing complexity of the multi-physics systems and observations, the goal of the validation tests changes. The abundant data from simple systems allows one to test whether individual embedded models provide consistent descriptions of the phenomena they are formulated to represent, under a range of conditions relevant to the prediction scenario. In contrast, scarce observations of multi-physics systems approaching the complexity and conditions of the prediction scenario can detect whether important physical phenomena have not been accounted for, or if embedded models are unexpectedly being used under conditions for which they are invalid or have not been calibrated.

One of the common challenges in conducting validation studies is deciding how much discrepancy between model predictions and observations is acceptable. Devising validation metrics and criteria is made more complicated by the need to assess the validity of the models for predicting quantities (QoIs) other than the observables. In the predictive validation approach proposed here, the usual issue of appropriate validation metrics is avoided because model discrepancies are represented explicitly through discrepancy models. The validation question is thus different; that is, we ask if the data are plausible outcomes of the combined physical and uncertainty models (including uncertainty in data, model parameters and model inadequacy)? Or restated, given the models, how likely is the observation? This is referred to as "posterior predictive model checking," and it has been the subject of much research [42, 26, 18, 9, 10, 11].

One still needs an assessment criterion, of course, but this is more straightforward. The simplest approach is to use one of a number of commonly used measures, such as p-values[17, 18, 10]. In this context, the posterior p-value of an observation is simply the probability, as determined from the calibrated model, that an observation would have lower probability than the actual observation. Thus, if the p-value of an observation is very small we would be justified in concluding that the inadequacy models are not a good representation of the observed discrepancy. Another interesting possibility is to pose the consistency question through Bayesian hypothesis testing. This, however, requires specification of a general class of alternative hypotheses appropriate for the problem at

hand. How this might be done for complex models needs to be explored.

Finally, one of the most important aspects of validation testing is the selection of observations to use for this purpose, beyond calibration data which should always be used. For this, one must consider the goals of a particular validation test. For example, for observations performed on complex systems near the top of the pyramid, the goal is generally to confirm that the embedded models are sufficient for the predictions at hand. In this case, one wants to select observations that are sensitive in the same way that the QoIs are to the embedded models and their errors, or any identified but unmodeled phenomena. Confidence can then be built that the uncertainties important to the QoIs are well represented. On the other hand, validation testing of individual embedded models at the bottom of the pyramid is intended to evaluate the fidelity of the embedded physical model, its calibration and any associated inadequacy model. Here the nature of the validation questions that must be asked, and therefore the needs for validation data, are context specific, depending on the theoretical pedigree of the model (i.e., ranging from simple data fits to highly reliable theory). For example, one may need data that can help identify missing model dependencies, test modeling assumptions, determine the range of model applicability, or test the universality of a calibration. Quantitative criteria for ranking how well potential validation cases address these requirements need to be developed.

6 Predictive Assessment

A model that has been calibrated as described in §4 and has passed all the validation tests described in §5 will not necessarily produce good predictions of the unobserved QoIs or adequately support a decision. The predictive assessment process described here addresses the question of whether predictions using the models are credible. To make this assessment, a number of questions need to be answered:

1. **Are the embedded models and associated uncertainty models being used outside their domain of applicability?** In the prediction scenario each embedded model is exercised under conditions that may or may not be different from the conditions in which the model was posed, calibrated, and/or validated. To assess this, the conditions in which the embedded model is applied need to be characterized by model-specific scenario parameters. For example, Arrhenius reaction rate models [25] have a single model scenario parameter; the temperature T. Given scenario parameters for an embedded model, the domain of applicability of the model can be defined as the range of these parameters over which the model has been calibrated and validation tested. Unfortunately, for other embedded models, such as RANS turbulence models [38, 16], the definition of appropriate model-specific scenario parameters is less clear. Further, in high dimensional scenario parameter spaces, discerning the range of applicability of a model by direct sampling will generally be impractical. Determining appropriate model-specific scenario parametrization, and defining ranges of applicability will thus need to rely on theoretical context and expertise of the modelers as well as data. Indeed, these should be considered part of the models and be subject to validation testing.

2. **Are the QoIs sensitive to uncertainties and model discrepancies to which the observables are not?** Confidence in the predictive capabilities of our models will in part be based on the assertion that the embedded models and associated uncertainties that influence the QoIs have been well calibrated and validated. If none of the observations are sensitive to a component of the model that is important to the QoI or the supported decision, then that

aspect of the model has not been informed by the available data, and thus cannot be relied on to provide good predictions of the QoIs or reliable decisions.

3. **Are the uncertainties in the QoIs sufficiently small for the predictions to be useful?** Predictions are being made in support of some decision-making process. If the uncertainty in the prediction is too large, it will not be able to inform the decision. Requirements regarding the necessary level of uncertainty clearly depend on the nature of the decisions.

To enable a complete analysis of questions (2) and (3), it may be useful to embed the prediction problem in a larger decision-theoretic framework. In decision theory, we represent (model) the action of the decision maker as minimizing the expected value of a loss function (or equivalently maximizing expected value of a utility function). This loss function is defined to express what the decision maker considers to be important or of value. Considering decision theory and specifying a loss function can be useful to the predictive assessment by providing a context in which to determine whether a decision is sensitive to the level of uncertainty in the predictions or to a poorly characterized aspect of the model. Decision theory is not considered here to automate or drive a decision process, as that would require a rigorous process to define and validate an appropriate loss function, which is out of scope for the development of predictive validation.

7 Research Challenges

The predictive validation process outlined here provides a framework in which confidence in the predictions of unobserved QoIs can be built. However, 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, maximum entropy considerations yield the needed priors. But this knowledge is commonly qualitative and difficult to express mathematically. Tools and techniques are needed to formulate 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 parameter 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 parametrizations 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) and with expensive computational models.

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 in part on work supported by the Department of Energy [National Nuclear Security Administration] under Award Number [DE-FC52-08NA28615].

References

[1] AIAA Computational Fluid Dynamics Committee on Standards. AIAA Guide for Verification and Validation of Computational Fluid Dynamics Simulations. AIAA, 1998. G-077-1998.

[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] I. Babuška, F. Nobile, and R. Tempone. Reliability of Computational Science. *Numerical Methods for Partial Differential Equations*, 23(4):753–784, 2007.

[5] I. Babuška, F. Nobile, and R. Tempone. A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria. *Computer Methods in Applied Mechanics and Engineering.*, 197:2517–2539, 2008.

[6] Osman Balci. Verification validation and accreditation of simulation models. In *Proceedings of the 29th Winter Simulation Conference*, WSC '97, pages 135–141, Washington, DC, USA, 1997. IEEE Computer Society.

[7] P. T. Bauman, J. T. Oden, and S. Prudhomme. Adaptive multiscale modeling of polymeric materials with Arlequin coupling and Goals algorithms. *Computer methods in applied mechanics and engineering*, 198(5-8):799–818, 2008.

[8] Paul T. Bauman, Jeremy Jagodzinski, and Benjamin S. Kirk. Statistical calibration of thermocouple gauges used for inferring heat flux. In *42nd AIAA Thermophysics Conference, AIAA 2011-3779*, 2011.

[9] M. J. Bayarri and J. O. Berger. Quantifying surprise in the data and model verification. In *Bayesian Statistics, 6*, pages 53–82. Oxford Univeristy Press, 1999.

[10] M. J. Bayarri and J. O. Berger. P-values for composite null models (with discussion). *J. Amer. Statist. Assoc.*, 95:1127–1142, 1157–1170, 2000.

[11] M. J. Bayarri and M. E. Castellanos. Bayesian checking of the second levels of hierarchical models. *Statist. Sci.*, 22(3):322–343, 2007.

[12] Maria J Bayarri, James O Berger, Rui Paulo, Jerry Sacks, John A Cafeo, James Cavendish, Chin-Hsu Lin, and Jian Tu. A framework for validation of computer models. *Technometrics*, 49(2):138–154, 2007.

[13] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

[14] S. H. Cheung, T. A. Oliver, E. E. Prudencio, S. Prudhomme, and R. D. Moser. Bayesian uncertainty analysis with applications to turbulence modeling. *Reliab. Eng. Syst. Safety*, 2011.

[15] R. T. Cox. *The Algebra of Probable Inference*. Johns Hopkins University Press, Baltimore, MD, 1961.

[16] P. A. Durbin. *Statistical Theory and Modeling for Turbulent Flows*. Wiley, 2001.

[17] R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh, 1925.

[18] A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica*, 6:733–807, 1996.

[19] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[20] Dave Higdon, Marc Kennedy, James C. Cavendish, John A. Cafeo, and Robert D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.*, 26(2):448–466, February 2005.

[21] K. S. Van Horn. Constructing a logic of plausible inference: A guide to Cox's theorem. *International Journal of Approximate Reasoning*, 34(1):3–24, 2003.

[22] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[23] H. Jeffreys. On the theory of errors and least squares. *Proc. Roy. Soc.*, 138(834):48–55, 1932.

[24] Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[25] Keith J. Laidler. A glossary of terms used in chemical kinetics, including reaction dynamics. *Pure & Appl. Chem.*, 68:149, 1996.

[26] X.-L. Meng. Posterior predictive p-values. *Ann. Statist.*, 22:1142–1160, 1994.

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[28] K. Miki, M. Panesi, E. E. Prudencio, and S. Prudhomme. Estimation of the nitrogen ionization reaction rate using electric arc shock tube data and Bayesian model analysis. *Phys. Plasmas*, 19:023507, 2012.

[29] Rebecca Morrison, Corey Bryant, Gabriel Terejanu, Kenji Miki, and Serge Prudhomme. Optimal data split methodology for model validation. In *Proceedings of the World Congress on Engineering and Computer Science 2011 Vol II, WCECS 2011*, pages 1038–1043, October 19-21 2011.

[30] M. Muto and J. L. Beck. Bayesian updating and model class selection of hysteretic structural models using stochastic simulation. *J. Vib. Control*, 14:7–34, 2008.

[31] W. L. Oberkampf and T. G. Trucano. Validation methodology in computational fluid dynamics. In *Fluids 2000 Conference*, Denver, CO, 2000. AIAA 2000-2549.

[32] William Oberkampf and Christopher Roy. *Verification and Validation in Scientific Computing*. Cambridge University Press, 2010.

[33] William L. Oberkampf and Timothy G. Trucano. Verification and Validation Benchmarks. Technical Report SAND2007-0853, Sandia National Laboratories, 2007. Unlimited Release.

[34] T. A. Oliver and R. D. Moser. Bayesian uncertainty quantification applied to RANS turbulence models. *J. Phys.: Conf. Ser.*, 318, 2011. 042032.

[35] T. A. Oliver and R. D. Moser. Accounting for uncertainty in the analysis of overlap layer mean velocity models. *Phys. Fluids*, 24:075108, 2012.

[36] N. Oreskes, K. Shrader-Frechette, and K. Belitz. Verification, validation and confirmation of numerical models in Earth sciences. *Science*, 263:641–646, 1994.

[37] M. Panesi, K. Miki, S. Prudhomme, and A. Brandis. On the validation of a data reduction model with application to shock tube experiments. *Computer Methods in Applied Mechanics and Engineering*, 213–216:383–398, 2012.

[38] S. B. Pope. *Turbulent Flows*. Cambridge University Press, 2000.

[39] Ernesto Prudencio and Sai Hung Cheung. Parallel adaptive multilevel sampling algorithms for the Bayesian analysis of mathematical models. *International Journal for Uncertainty Quantification*, 2(3):215–237, 2012.

[40] P. J. Roache. Perspective: Validation - What Does it Mean? *ASME Journal of Fluids Engineering*, 131(3):1–3, 2008.

[41] P. J. Roache. *Fundamentals of Verification and Validation*. Hermosa Publishers, Albuquerque, 2009.

[42] D. B. Rubin. Bayesian justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12:1151–1172, 1984.

[43] Craig P. S, M. Goldstein, Rougier J. C, and Seheult A. H. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, 96:717–729, 2001.

[44] Robert G. Sargent. Verification and validation of simulation models. In *Proceedings of the 30th Winter Simulation Conference*, WSC '98, pages 121–130, Los Alamitos, CA, USA, 1998. IEEE Computer Society Press.

[45] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July, October, 1948.

[46] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[47] Gabriel Terejanu, Rochan R. Upadhyay, and Kenji Miki. Bayesian experimental design for the active nitridation of graphite by atomic nitrogen. *Experimental Thermal and Fluid Science*, 36:178–193, Jan. 2012.