Submitted to Encyclopedia of Applied and Computational Mathematics Feb. 3, 2012

Validation

Christopher J. Roy [a] and William L. Oberkampf [b]

[a] *Aerospace and Ocean Engineering Department*
Virginia Tech
215 Randolph Hall
Blacksburg, Virginia 24061 USA
cjroy@vt.edu

[b] *Consulting Engineer*
5112 Hidden Springs Trail
Georgetown, Texas 78633 USA
wloconsulting@gmail.com

Synonyms: Model Validation, Model Form Uncertainty, Model Error

Short Definition: Validation is the quantitative assessment of a model relative to experimental observations.

1. Introduction

Mathematical *models* are used in science and engineering to describe the behavior of a system. In many cases, these models take the form of differential equations which require approximate numerical solutions (i.e., *simulations*) due to their complexity. While the focus of this article is on models based on partial differential equations, the concepts and techniques apply equally well to models based on ordinary differential equations, algebraic models, etc. Verification and validation provide a means for assessing the credibility and accuracy of models and their subsequent simulations [1, 2]. *Verification* deals with assessing the numerical accuracy of a simulation relative to the exact (but rarely known) result of the model. On the other hand, the assessment of the accuracy of the model itself is termed *validation* and requires the comparison of model predictions to observations of nature which are typically embodied in experimental measurements. While there are many approaches to model validation, we will focus on validation methods which provide a quantitative assessment of model accuracy and which also account for the presence of uncertainty in both the simulation results and the experimental data.

2. Uncertainty

There are many sources of *uncertainty* in computational mathematics including the model inputs, the form of the model (which embodies all of the assumptions in the formulation of the model), and poorly-characterized numerical approximation errors. These sources of uncertainty can be classified as 1) *aleatory* – the inherent variation in a quantity, 2) *epistemic* – uncertainty due to lack of knowledge, or 3) a mixture of the two. Aleatory uncertainty is generally characterized probabilistically by either a probability density function or a cumulative distribution function (CDF), the latter being simply the integral of the probability density function from minus infinity up to the value of interest. A purely epistemic uncertainty should be characterized as an interval (with no associated probability distribution), which is the weakest statement that one can make about the value (or distribution) of a quantity. One approach for characterizing mixed aleatory and epistemic uncertainty is a probability box, or p-box, which characterizes the infinite set of all possible probability distributions that could exist within the bounds of the p-box [3]. The two outer bounding CDFs reflect the combined aleatory and epistemic uncertainty in the quantity of interest (see Figure 1). The width of the p-box represents the range of values that are possible for a given cumulative probability, whereas the height of the p-box represents the interval range of cumulative probabilities associated with a given value.

In general, there may be one or more system outputs, which we will refer to as System Response Quantities (SRQs), that the analyst is interested in predicting with a computational mathematics model. When uncertain model inputs are aleatory, there are a number of different approaches for propagating this uncertainty through the model. The simplest approach is sampling (e.g., Monte Carlo or Latin Hypercube) where inputs are sampled from their probability distribution and then used to generate a sequence of SRQs; however, sampling methods tend to converge slowly as a function of the number of samples. Other approaches that can be used to propagate aleatory uncertainty include perturbation methods and polynomial chaos (both intrusive and non-intrusive formulations). Furthermore, when a response surface approximation of an SRQ as a function of the uncertain model inputs is available, then any non-intrusive method discussed above (including sampling) can be computed efficiently.

When all uncertain inputs are characterized by intervals, there are two popular approaches for propagating these uncertainties to the SRQs. The simplest is sampling over the input intervals in order to estimate the interval bounds of the SRQs. However, the propagation of interval uncertainty can also be formulated as a bound-constrained optimization problem: given the possible interval range of the inputs, determine the resulting minimum and maximum values of the SRQs. Thus, standard approaches for constrained optimization such as local gradient-based searches and global search techniques can be used.

When some uncertain model inputs are aleatory and others are epistemic, then a segregated approach to uncertainty propagation should be used [3,4]. For example, in an outer-loop, samples from the epistemic uncertain model inputs may be drawn. For each of these sample-values, the aleatory-uncertain model inputs are propagated assuming fixed sample-value of the epistemic-uncertain variable. The completion of each step in the outer loop results in a possible CDF of the SRQ. The total result of the segregated uncertainty propagation process will be an ensemble of possible CDFs of the SRQ, the outer bounding values of which can be used to form a p-box. An advantage of this segregated approach is that the inner aleatory propagation loop can be achieved using any of the techniques described above for propagating probabilistic uncertainty (i.e., it is not limited to simple sampling approaches).

3. Validation Experiments

A *validation experiment* is an experiment conducted with the primary purpose of assessing the predictive capability of a model. Validation experiments differ from traditional experiments used for exploring a physical phenomenon or obtaining information about a system because the customer for the experiment is the model which is generally embodied within a simulation code. There are six primary guidelines for validation experiments [2]. Validation experiments should:

1. be jointly designed by experimentalists and modelers, with the simulation code used to provide pre-test computations of the proposed experiment,

2. be designed to capture the relevant physics and measure all initial conditions, boundary conditions, and other relevant modeling data required by the simulation,
3. strive to emphasize the inherent synergism that is attainable between computational and experimental approaches,
4. be a blind comparison between simulation and experiment, i.e., the experiment should provide all required model inputs and boundary conditions, but not the measured SRQs,
5. be designed to ensure that a hierarchy of SRQs are measured, e.g., from globally integrated quantities to local quantities, and
6. be constructed to analyze and estimate the components of random (precision) and systematic (bias) experimental uncertainties in both the SRQs and the model inputs.

4. Validation Metrics

Validation metrics provide a means by which the accuracy of a model can be assessed relative to experimental observations. Liu et al. [5] proposed a classification system for validation metrics based on whether or not 1) the metric incorporates uncertainty sources in the simulation predictions and the experimental measurements (i.e., the metric is classified as either deterministic or stochastic), 2) the comparison is made for a single SRQ or multiple SRQs (i.e., univariate or multivariate), and 3) the metric provides a quantitative distance-based method that can be used to quantify modeling error. (Note that the latter criterion is also related to the mathematical requirements for a metric.) Liu et al. [5] also recommend that the metric be objective with a given set of simulations and data resulting in a single metric value (i.e., it should not depend on the analyst evaluating the metric, their preferences, or prior assumptions). The field of validation metrics is an area of active research, but for the purposes of this article, we focus only on stochastic validation metrics that provide distance-based measures of the agreement/disagreement between the model and experimental data, thus we omit any discussion of approaches such as classical hypothesis testing and Bayesian model comparison employing Bayes factors.

It is important to draw clear distinctions between the concepts of validation and calibration. While *validation* involves the quantitative assessment of a model relative to experimental data, c*alibration* (a.k.a., parameter estimation, parameter optimization, or model updating) instead involves the adjustment of input parameters to improve agreement with experimental data. For example, if all uncertain model inputs are probabilistic, then Bayesian updating can be used to update model input parameters. While calibration may be an important part of the model building and improvement process, it does not in itself provide quantitative estimates of model accuracy. The key difference is that model calibration results in a modified model that must still be assessed for accuracy when new experimental data become available.

While there are many possible validation metrics, we will focus on one implementation called the area validation metric [6] which is a mathematical metric that provides quantitative assessment of disagreement between a stochastic model and experimental data. When only aleatory uncertainties are present in the model inputs, then propagating these uncertainties through the model produces a CDF of the SRQ. Experimental measurements are then used to construct an empirical CDF of the SRQ. The area between these two CDFs is referred to as the area validation metric *d* (also called the Minkowski L_1 norm) and is given by

$$d(F, S_n) = \int_{-\infty}^{\infty} \left| F(x) - S_n(x) \right| dx \qquad (1)$$

where *F(x)* is the CDF from the simulation, $S_n(x)$ the CDF from the experiment, and *x* is the SRQ. The area validation metric *d* has the same units as the SRQ and thus provides a measure of the *evidence for disagreement* between the simulation and the experiment [6]. Note that the area validation metric represents an epistemic uncertainty since additional experiments and/or model improvements can be conducted (i.e., information can be added) in order to reduce it. This epistemic uncertainty is commonly referred to as *model form uncertainty*.

An example of this area validation metric for a case with only aleatory uncertainty occurring in the model input parameters is given in Figure 2. In this figure, the aleatory uncertainties have been propagated through the model (e.g., with a large number of Monte Carlo samples), but only four experimental replicate measurements are available. The stair-steps in the experimental CDF are due to the different values observed in each of the four experimental measurements. The stochastic nature of the measurements can be due to variability of the experimental conditions and random measurement uncertainty. This metric can also be computed for cases involving both aleatory and epistemic uncertainty in the model inputs (e.g., see Ref. [2]).

5. Extrapolation

In general, it is too expensive (or even impossible) to obtain experimental data over the entire multi-dimensional space of model input parameters for the application of interest. As a result, techniques are needed for estimating model form uncertainty at conditions where there are no experimental data. Consider a simple example when there are only two input parameters for the model: α and β (Figure 3). The validation domain consists of the set of points in this parameter space where experiments have been conducted and the validation metric has been computed (denoted by a "V" in the figure). In this example, the application domain (sometimes referred to as the operating envelope of the system) is larger than the validation domain, although many other set relationships are possible. Thus, one must choose between 1) ignoring the inaccuracy in the model, 2) using the flexibility of the model by way of calibrating the model parameters at the validation conditions, 3) extrapolating the validation metric outside of the validation domain, or 4) performing additional validation experiments (Figure 3 denotes conditions for candidate validation experiments by a "C"). The key point is that the validation domain is generally not coincident with the application domain, thus either interpolation or extrapolation of the model form uncertainty to the conditions of interest is needed.

One method for estimating the model form uncertainty at the conditions of interest is as follows [4]. First, a regression fit of the validation metric is performed using data from the validation domain. Next, a statistical analysis is performed to compute the prediction interval at the conditions of interest. This prediction interval is similar to a confidence interval, but it will be larger because we are interested in a future random deviate predicted by the regression fit of the validation metric data, i.e., the uncertainty due both to the regression fit and the variability of the validation metric evaluated at an arbitrary set of conditions. The computation of the prediction interval requires a level of confidence to be specified (e.g., 95% confidence). The model form uncertainty U_{MODEL} at the prediction conditions is then found by taking the maximum of zero and the value found from the regression fit of the validation metric $\hat{d}$ and adding in the upper value of the prediction interval, *P*, i.e.,

$$U_{MODEL} = \max\left(\hat{d},0\right) + P \tag{2}$$

A simple example showing the extrapolation of model form uncertainty in nozzle thrust (in Newtons) as a function of a single model input, stagnation pressure (in megapascals), is given in Figure 4. In this example, area validation metrics are computed from simulations and (hypothetical) experiments at stagnation pressures of 1.0, 1.5, 2.0, 2.5 and 3.0 MPa, yielding metric results of 23.0, 25.0, 24.0, 26.2, and 28.8 N, respectively. In order to extrapolate this model form uncertainty to the prediction condition, we first compute a linear regression fit of the validation metric as a function of the stagnation pressure. The resulting regression fit is

$$\hat{d} = 20.28 + 2.56 p_0 \qquad N \tag{3}$$

with p_0 given in MPa. The computed values of the validation metric, along with the above regression fit, are shown graphically in Figure 4. A prediction interval for the regression fit is then computed at the 95% confidence level as shown in the figure. The upper value of the prediction interval is then used to estimate the model form uncertainty at the prediction conditions. In this case, the regression fit of the validation metric evaluated at the prediction conditions (6 MPa) gives $\hat{d} = 35.6$ N. The magnitude of the 95% prediction interval at this location is P = ±9.7 N (i.e., $\hat{d}$ ±P), thus the estimated model form uncertainty U_{MODEL} is

$$U_{MODEL} = \max(\hat{d},0) + P = 35.6 + 9.7\ \mathrm{N} = 45.3\ \mathrm{N}.$$

Since this estimated model form uncertainty is epistemic in nature, it will be treated as an interval about the simulation prediction.

6. Predictive Capability

The total prediction uncertainty can be estimated by combining the propagated uncertainty from the model inputs (aleatory and epistemic) with the uncertainty due to the form of the model and the uncertainty due to the numerical error estimation process. For example, if *F(x)* is the CDF resulting from propagating random uncertainties through the model, then accounting for the model form and numerical uncertainties would result in the p-box $F(x \pm U_{TOTAL})$ where $U_{TOTAL} = U_{MODEL} + U_{NUM}$. In the more general case where there are both aleatory and epistemic uncertainties in the model inputs, the propagation of these uncertainties through the model results in a p-box. The uncertainties due to model form and numerical approximations simply result in a broadening of this p-box. Although uncertainties are not necessarily additive in this way [7], this approach estimates the compounding roles of the various sources of epistemic uncertainty. An example of this "extended" p-box is shown in Figure 5, where the estimated modeling and numerical uncertainties are U_{MODEL} = 45.3 N and U_{NUM} = 36.8 N (see Ref. [8] for details).

There are various ways that a decision maker can use the uncertainty information provided in Figure 5. First, if one is interested in the minimum range of the SRQ (thrust) that is predicted with a probability of 0.90, then one has an interval range of [2565, 2815] N for the SRQ at a cumulative probability of 0.05 and a range of [2695, 2930] N at a cumulative probability of 0.95. Taking the lowest possible value of the former and the highest possible value of the latter, there is a 0.90 probability that the SRQ lies in the range 2565 N ≤ SRQ ≤ 2930 N. If instead there were a requirement that the nozzle produce a thrust greater than or equal to 2600 N, Figure 5 shows that there is at most a 0.22 probability that the system would fail to achieve this required minimum thrust, i.e., the cumulative probability that the thrust is less than or equal to 2600 N is the interval [0, 0.22]. Finally, Figure 5 provides a significant amount of information to a decision maker regarding the impact of each source of uncertainty in the simulation prediction.

References

1. ASME (2006), *Guide for Verification and Validation in Computational Solid Mechanics*, American Society of Mechanical Engineers, ASME Standard V&V 10-2006, New York, NY.
2. Oberkampf, W. L. and Roy, C. J. (2010), *Verification and Validation in Scientific Computing*, Cambridge University Press, Cambridge.
3. Ferson, S., and Ginzburg, L. R. (1996), “Different Methods are needed to Propagate Ignorance and Variability,” *Reliability Engineering and System Safety*, Vol. 54, pp. 133-144.
4. Roy, C. J. and Oberkampf, W. L. (2011), “A Comprehensive Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing,” *Computer Methods in Applied Mechanics and Engineering*, Vol. 200, pp. 2131–2144 (DOI:10.1016/j.cma.2011.03.016).
5. Liu, Y., Chen, W., Arendt, P., and Huang, H.-Z. (2011), “Toward a Better Understanding of Model Validation Metrics,” *Journal of Mechanical Design*, Vol. 133, pp. 1-13.
6. Ferson, S., Oberkampf, W. L., and Ginzburg, L. (2008), “Model Validation and Predictive Capability for the Thermal Challenge Problem,” *Computer Methods in Applied Mechanics and Engineering*, Vol. 197, pp. 2408–2430.
7. Ferson, S. and Tucker, W. T. (2006), “Sensitivity in Risk Analyses with Uncertainty Numbers,” Sandia National Laboratories Report, SAND2006-2801, Albuquerque, NM.
8. Roy, C. J. and Balch, M. S. “A Holistic Approach to Uncertainty Quantification in Scientific Computing,” manuscript accepted for publication in *International Journal for Uncertainty Quantification*, January 2012.

Figures

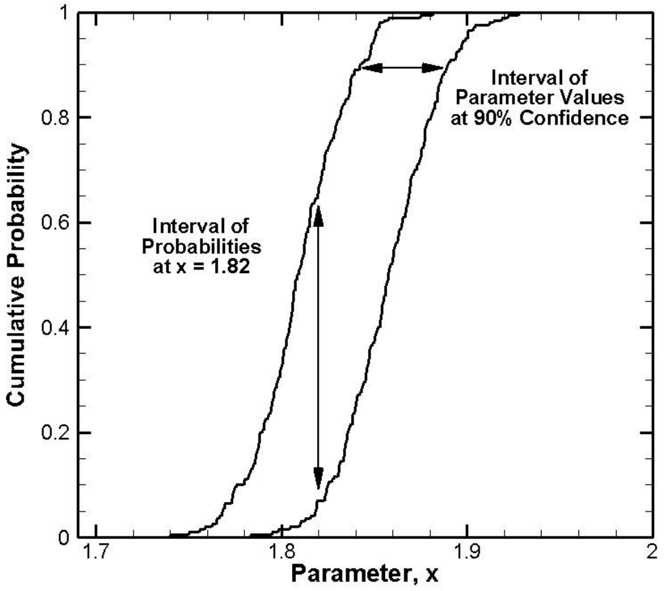

Figure 1. Simple example of a p-box (reproduced from [8]).

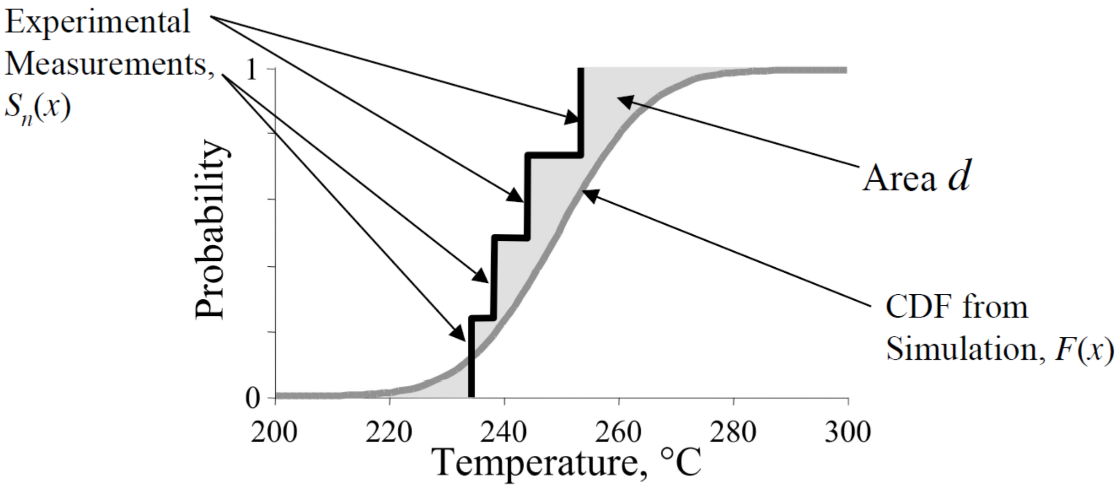

Figure 2. Area validation metric example (reproduced from [6]).

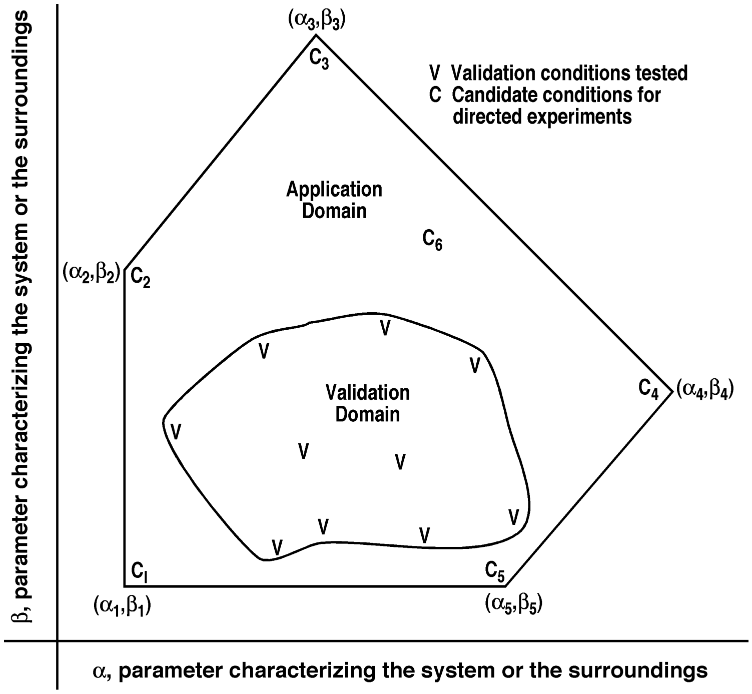
(α_3,β_3)
C_3
V Validation conditions tested
C Candidate conditions for directed experiments
Application Domain
C_6
(α_2,β_2)
C_2
V
Validation Domain
C_4
(α_4,β_4)
C_l
(α_1,β_1)
C_5
(α_5,β_5)
β, parameter characterizing the system or the surroundings
α, parameter characterizing the system or the surroundings

Figure 3. Schematic showing a possible relationship between the validation domain and the application domain (reproduced from [2]).

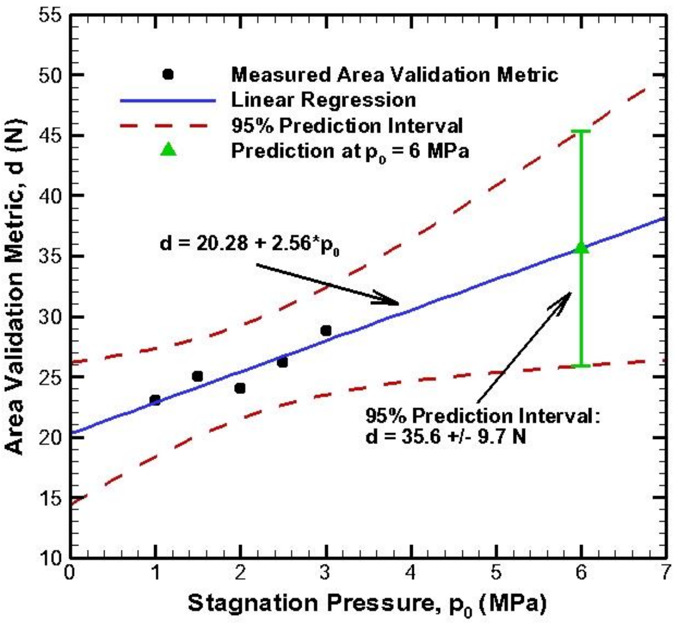

Figure 4. Example of extrapolation of validation metric to the prediction conditions (p_0 = 6 MPa) including prediction intervals (reproduced from [8]).

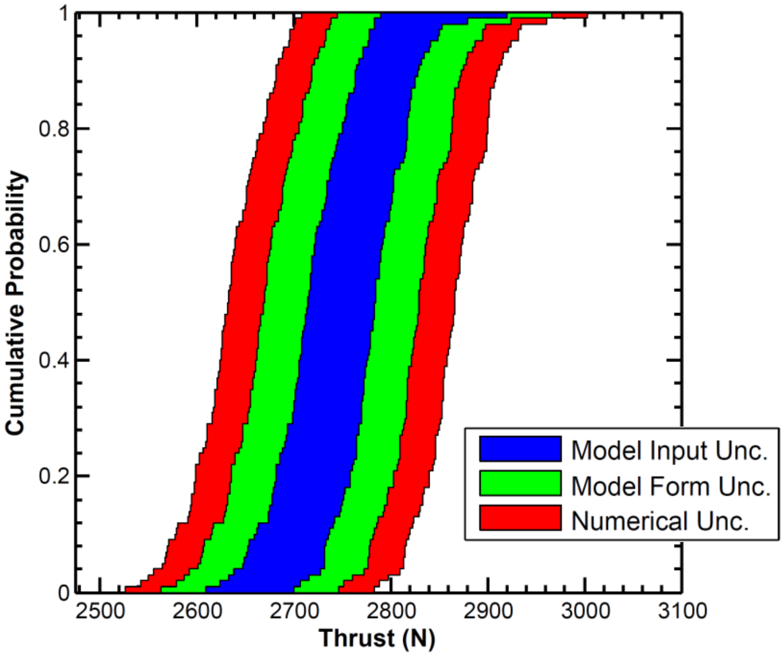

Figure 5. Example of extended p-box for the SRQ of nozzle thrust (reproduced from [8]).