# 12

# Model accuracy assessment

As has been discussed in a number of chapters, particularly Chapter 10, Model validation fundamentals, and Chapter 11, Design and execution of validation experiments, model accuracy assessment is the core issue of model validation. Our intent in model accuracy assessment is to critically and quantitatively determine the ability of a mathematical model and its embodiment in a computer code to simulate a well-characterized physical process. We, of course, are only interested in well-characterized physical processes that are useful for model validation. How critical and quantitative the model accuracy assessment is will depend on (a) how extensive the experimental data set is in exploring the important model input quantities that affect the system response quantities (SRQs) of interest; (b) how well characterized the important model input quantities are, based on measurements in the experiments; (c) how well characterized the experimental measurements and the model predictions of the SRQs of interest are; (d) whether the experimental measurements of the SRQs were available to the computational analyst before the model accuracy assessment was conducted; and (e) if the SRQs were available to the computational analysts, whether they were used for model updating or model calibration. This chapter will explore these difficult issues both conceptually and quantitatively.

We begin the chapter by discussing the fundamental elements of model accuracy assessment. As part of this discussion, we review traditional and recent methods for comparing model results and experimental measurements, and we explore the relationship between model accuracy assessment, model calibration, and model prediction. Beginning with the engineering society definitions of terms given in Chapter 2, Fundamental concepts and terminology, the perspective of this book is to segregate, as well as possible, each of these activities. There is, however, an alternative perspective in the published literature that believes all of these activities should be combined. We briefly review this alternative perspective and the associated approaches, and contrast these with approaches that segregate these activities.

Whereas model calibration has a long history, primarily in the statistical literature, quantitative model accuracy assessment has had much less development. During the last decade, model accuracy assessment has focused on constructing mathematical operators that compute the difference between the experimentally measured results and simulation results. These operators are referred to as *validation metrics*. We review recommendations

that have been proposed for the optimum construction of validation metrics. We then discuss in detail two validation metrics that have been developed. The first metric computes the difference between the estimated statistical mean of the measurements and the predictions. The second metric computes the area between the probability-boxes (p-boxes) resulting from the measurements and the predictions. Each metric is demonstrated with various examples.

## 12.1 Elements of model accuracy assessment

The task of model accuracy assessment might, at first, seem fairly simple: compare the prediction supplied by the modeler to the observation(s) made by the empiricist and see whether they match. They might match perfectly, or the model might be somewhat in error, or the model could be totally incorrect. However, in reality there are several issues that arise to complicate this comparison, as well as the model's ultimate use in predictions for which no experimental data is available. Some of the important questions are:

- How should one deal with experiment-to-experiment variability in the measured data?
- What should be done if the experimental data are given as a probability distribution or a sequence of intervals?
- What if there are statistical trends in the experimental data that do not appear in the simulation?
- What if the prediction is a probability distribution rather than a point, i.e., a deterministic, value?
- Do model accuracy requirements have any place in model accuracy assessment?
- Should the measure of model accuracy assessment represent evidence for agreement between experiment and simulation, or evidence for disagreement?
- What should be done in a comparison with experimental data if important information is not available to make the prediction?
- How should model accuracy assessment be influenced by limited experimental data obtained over a high-dimensional model input space?
- What could be done if there is only one experiment for comparison?
- What could be done to synthesize comparisons about different outputs from the model and experimental data?
- How should aleatory and epistemic uncertainty in either the simulation or the experiment be handled?
- How should an accuracy measure be constructed to penalize or reward a very precise model prediction versus a very imprecise prediction?
- What if the predictions from the model are extremely expensive to compute and only a few simulations can be computed?
- What should be done differently in model accuracy assessment if the parameters in the model are first calibrated using related experimental data?

Some of these questions were dealt with in previous chapters, but they will be revisited here in more detail.

In many research situations in traditional scientific procedures, the emphasis in validation is about deciding whether a model is right or wrong *per se*. In most engineering situations,

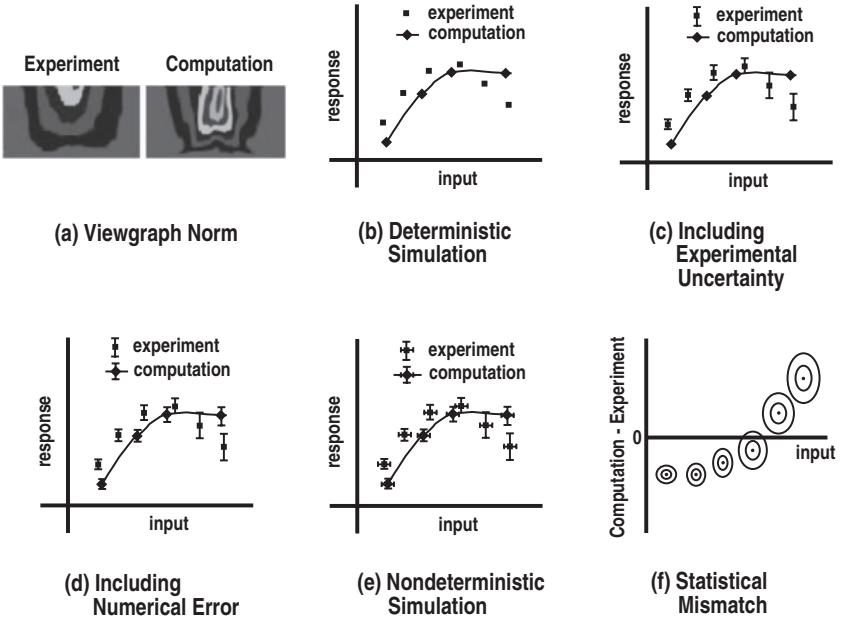


Figure 12.1 Increasing precision for comparing simulations and experiments (Oberkampf *et al.*, 2004).

however, this is *not* the case. In engineering, the emphasis is about estimating the accuracy of the model, based on comparisons with experimental data, and determining if the model is adequate for the intended use. As George Box famously asserted two decades ago, "All models are wrong. Some are useful" (Box and Draper, 1987). For a deterministic model prediction, validation can be a fairly straightforward affair. The model makes a point estimate for its prediction about some quantity. This prediction would be compared against one or more measurements about that quantity and the difference(s) would be understood as a measure of how accurate the model was. A model could be consistently inaccurate and yet close enough for its purpose. Likewise, even a highly accurate model might not be good enough if the accuracy requirements are more demanding for modeling some high-performance system.

### *12.1.1 Methods of comparing simulations and experiments*

There are a wide variety of methods for comparing simulations and experimental measurements. Figure 12.1 summarizes the methods for comparisons and identifies the direction in which more quantitative methods need to progress. This figure illustrates the conceptual increase of quantification in performing validation comparisons as increased attention is paid to experimental uncertainty, numerical error, and nondeterministic simulations. Let us consider each panel of this figure in some detail.

Figure 12.1a depicts a qualitative comparison referred to as a *viewgraph norm* comparison (Rider, 1998) of computational and experimental data often seen in practice. This comparison is typically one picture or contour plot next to another and sometimes no legend is given showing the scales involved; or similarly, the scales are adjusted so as to present the most flattering agreement between computation and experiment. Clearly, in such as case no quantitative statement about the comparison between computation and experiment can be made, but it does provide some feel of comparison at an intuitive level. Feel and intuition about agreement between computation and experiment are, of course, in the eye of the beholder. These type comparisons are commonly seen in advertising and marketing material for scientific computing software, or in proposals for potential funding.

The plot in Figure 12.1b portrays the most common type of comparison between computational results and experimental data. It shows the system response as a function of the input, or control, parameter in the experiment. While discrete experimental and computational points are shown in this plot, the concept also encompasses the display of curves for experiment and computation without any points shown. The key problem with comparisons implemented at the level of Figure 12.1b is that there is no recognition of uncertainty in the experimental or computational results, or in the quantitative comparison of the two. Conclusions drawn from this type of comparison are really only qualitative, such as "fair agreement" or "generally good agreement."

Figure 12.1c suggests that the next step for improving the method of comparison is to place estimated uncertainty bars around the experimental data. Occasionally, the meaning of the uncertainty bars is carefully justified and clearly explained. The much more common situation is where the uncertainty bars are qualitative, in the sense that: (a) they are not rigorously justified concerning what is included in terms of random and systematic experimental uncertainty, (b) a statement is made such as "Essentially all of the experimental data fell within the uncertainty bars shown," or (c) the effect of experimental uncertainty in the input quantity is not addressed with regard to the response quantity. An increasing number of technical journals are requiring some type of statement of experimental uncertainty, such as Figure 12.1c.

Figure 12.1d represents the case where there is a more quantitative estimate of experimental uncertainty and there is an estimate of numerical solution error. For example, concerning the experimental uncertainty, multiple experimental realizations could have been obtained so the experimental data point shown would represent the mean of all of the samples. In addition, it might be clarified if any "outlier" measurements were discarded and that the uncertainty bar would represent two standard deviations of an assumed normal probability distribution. Concerning the numerical solution error, an *a posteriori* numerical error estimate from the computation would be given for the specific response quantity that was plotted in the graph, as opposed to some global error norm related to the quantity shown.

Figure 12.1e suggests a further improvement in the estimation of experimental uncertainty and now, at this level, nondeterministic simulations are included. The information concerning experimental uncertainty could be improved, for example, in two ways. First, one could use a statistical design of experiments (DOE) approach with randomization and

blocking to better quantify certain systematic uncertainties. Also, one may have conducted the same experiment at separate facilities, possibly using different diagnostic techniques. Second, an estimate of the experimental uncertainty in the measured input quantity is also obtained and is shown as the lateral uncertainty bar. The uncertainty bars for the input quantity and the measured response, for instance, could represent two standard deviations for a normal probability distribution. Concerning the nondeterministic simulations, we are referring to an ensemble of computations at each of the experimental conditions. For example, multiple simulations would be made using the experimentally estimated probability distribution for the input quantity. As a result, the computational data point would be the mean of the nondeterministic simulations for both the input and the response quantity. Note that to use the experimental uncertainty distributions for the input quantity, computations would need to be made at the measured input conditions, or one must assume they could be used for other input conditions.

Figure 12.1f shows a genuine quantitative measure of the comparison between the simulations and the experimental measurements over the range of the input quantity. What is depicted in Figure 12.1f is the *mismatch* between the simulation and the experiment at each of the experimental data points. In terms of information content, one could have the same data as contained in Figure 12.1e, but now the statistical differences in the simulation and experiment are displayed. Assuming that probability distributions for both computational and experimental data are well characterized, as discussed in Figure 12.1e, comparing computations and experiments will require a difference, or more properly, a convolution of pairs of probability distributions. The elliptical symbols in Figure 12.1f are meant to signify one and two standard deviation contours of the convolutions of the simulation and experimental probability distributions. The "dot" in the center of each contour is the difference in the mean, or expected value, of the simulation and experimental distributions.

We will refer to quantitative comparisons between simulation and experiment, similar to Figure 12.1f, as a *validation metric operator*. We will formally use the definition:

*Validation metric*: a mathematical operator that measures the difference between a system response quantity (SRQ) obtained from a simulation result and one obtained from experimental measurements.

The validation metric should be an objective measure of the distance, in some sense, between simulation and experimental data. The distance measure should have the characteristic that any difference between the simulation and experimental data is a positive quantity. For example, whether the simulation is less than or greater than the experimental data, the metric is positive and additive. Objectiveness means that, given a collection of predictions and observations, a validation metric will produce the same assessment no matter what analyst uses the metric. This is a basic tenet of scientific and engineering practice; that the conclusion is reproducible and that it does not depend on the attitudes or predilections of the analysts. If some inescapable subjectivity must enter the evaluation of the metric, it would be good to keep this intrusion as small and as limited as possible so as to emphasize the objectiveness of the method and minimize the elements subject to dispute.

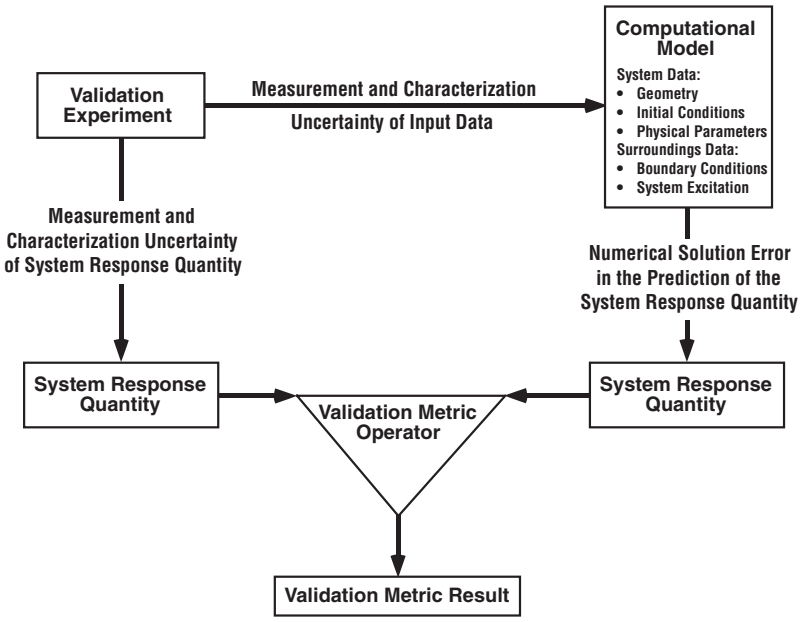


Figure 12.2 Sources of uncertainty and error in model accuracy assessment.

In general, the validation metric result is a function of all of the input parameters to the model. Commonly, however, the SRQ is primarily dependent on a handful of dominant input parameters. The validation metric should be viewed as a statistical operator because the simulation and experimental results are not single numbers but are functions; specifically they are commonly either a probability distribution or a *probability box* (p-box). A p-box is special type of cumulative distribution function (CDF) that represents the set of possible CDFs that fall within prescribed bounds (Ferson, 2002; Ferson *et al.*, 2003, 2004; Kriegler and Held, 2005; Aughenbaugh and Paredis, 2006; Baudrit and Dubois, 2006). p-boxes were first introduced in Chapter 3, Modeling and computational simulation. A p-box can express both epistemic and aleatory uncertainty in either the simulation, the experimental results, or both. p-boxes will be discussed in more detail later in this chapter.

### *12.1.2 Uncertainty and error in model accuracy assessment*

Model accuracy assessment would be fairly straightforward if we did not have to deal with uncertainties and errors clouding the issue in the experiment and in the simulation. In earlier chapters, e.g., Chapters 10 and 11, we have discussed uncertainties and errors primarily from an estimation point of view. Here we discuss them with regard to where they occur, how they are propagated, and how they affect validation metrics results.

Figure 12.2 shows a validation experiment and the corresponding computational model, the sources of uncertainty and error in experimental measurements and numerical

simulations, and the validation metric as a difference operator. First we will discuss the experimental sources of uncertainty. In Chapter 10, we discussed the two fundamental sources of experimental uncertainty: measurement uncertainty and characterization uncertainty. Measurement uncertainty is due to random and systematic (bias) uncertainties in every experimental measurement that is made. Measurement uncertainty is primarily dependent on the diagnostic techniques used, as well as the experimental measurement procedures used, e.g., replication, randomization, and blocking techniques to estimate random and systematic measurement uncertainties. Characterization uncertainty is due to a limited number of measurements of a quantity that is a random variable. Two examples are (a) an input quantity needed for the computational model was not measured in the experiment, so the experimentalists recommends an interval in which he/she believes the true value lies, and (b) a SRQ is known to be a random variable in the experiment, but too few samples of the quantity are measured because of time and cost constraints.

Also discussed in Chapter 10 were the three sources of error in the simulation: formulation of the mathematical model (i.e., model form error), mapping of the mathematical model to the discrete model (including the computer code), and numerical solution of the discrete model on a computer. The estimation of the model form error is, of course, the primary goal of model accuracy assessment. In Figure 12.2, we simply combine the second and third sources into "numerical solution error."

It is seen in Figure 12.2 that the experimental measurement and characterization uncertainties directly affect both the input data to the model and the SRQ from the experiment. Noting that the experimental measurement input uncertainties are statistically confounded with the inherent variability of the input quantities in the experiment, an important issue in model accuracy assessment is exposed. We are requiring that the model correctly map the experimental measurement input uncertainties to the SRQ, even though there is *no physics variability* that corresponds to these in the model. Stated differently, the experimental measurement uncertainties *do not* represent physics uncertainties in the input, but are purely an artifact of the measurements made in the experiment. Contrast these experimental uncertainties with true physics uncertainties in the experiment, such as (a) a parameter in the experiment that is a random variable, e.g., the material used in an experiment that is a random draw from a production-lot population, or (b) the boundary conditions are uncontrolled parameters in the experiment, e.g., weather conditions in multiple flight experiments. These later examples are *true physics uncertainties* that we expect the model to correctly propagate, whereas we are also requiring the model to correctly propagate nonphysics uncertainties, i.e., experimental measurement uncertainties. Although research needs to be dedicated to this issue, we believe it is a conceptually improper expectation of the model. If this is true, then the only route forward is to minimize the experimental measurement uncertainties relative to the input uncertainties due to physics variability.

To demonstrate this point, consider the following example. Suppose an experiment is repeated a large number of times and, in addition, every aspect of the physics is perfectly repeatable. That is, all conditions in the system and in the surroundings are precisely repeatable every time, so that the SRQ is also precisely repeatable. Further, suppose that

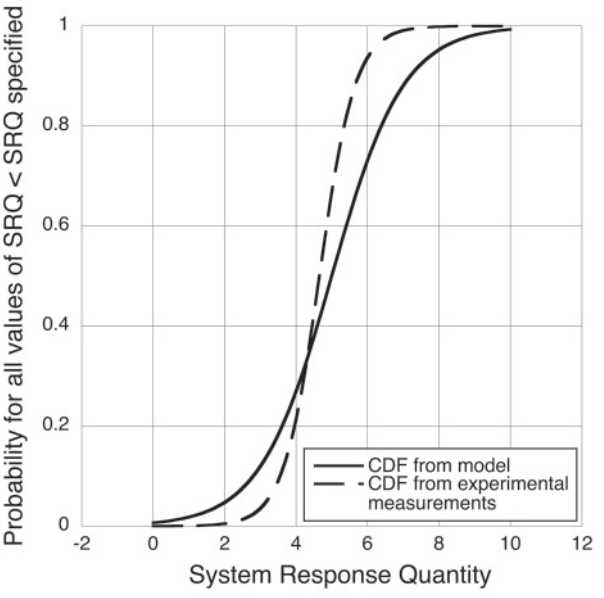


Figure 12.3 Example of cumulative distribution functions for a system response quantity from experimental measurements and from a model that is provided with apparent input physics variability.

all of the systematic (bias) errors in the experimental measurements are zero. As a result, the experimental measurement uncertainty would only consist of random error, and the experimental characterization error would be zero because all input quantities and the SRQ are perfectly characterized (i.e., known) random variables. In addition, suppose the model *perfectly* represents the relevant physics in the validation experiment. Also, suppose the numerical solution error is zero. Figure 12.3 depicts the cumulative distribution functions of the SRQ obtained from the experimental measurements and the model. They will be different because the CDF from the experiment represents the variability due to random measurement error, whereas the CDF from the model represents the variability in the SRQ due to the *apparent* physics variability in the input quantities. That is, the perfect physics model will propagate the experimental measurement uncertainty *as if* it were physics variability. When the validation metric operator measures the difference between the experimentally measured CDF and the predicted CDF from the model, the difference will *not* be zero. That is, the model will be judged to be "imperfect," when in fact it is *perfect*.

### *12.1.3 Relationship between model accuracy assessment, calibration, and prediction*

Figure 12.4 depicts several important aspects of validation, as well as features of calibration and prediction of models. The left-center portion of Figure 12.4 shows the first step in validation. The figure illustrates that the same SRQ must be obtained from both the model

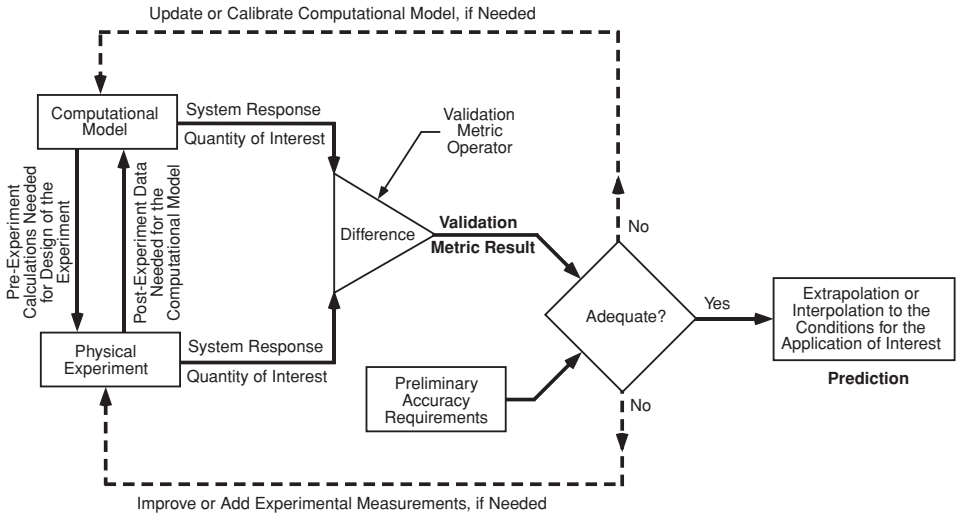


Figure 12.4 Validation, calibration, and prediction (Oberkampf and Barone, 2004).

and the physical experiment. The SRQ can be any type of physically measurable quantity, or it can be a quantity that is based on, or inferred from, measurements. For example, the SRQ can involve derivatives, integrals, or more complex data processing of computed or measured quantities such as the maximum or minimum of functionals over a domain. When significant data processing is required to obtain an SRQ, it is important to process both the computational results and the experimentally measured quantities in the same manner. The computational and experimental SRQs are input to the validation metric operator to compute a validation metric result. The SRQs are commonly one of three mathematical forms: (a) a deterministic quantity, i.e., a single value, such as a mean value or a maximum value over a domain; (b) a probability distribution; or (c) a p-box. Each of these forms can be functions of a parameter or multiple parameters in the model, such as a temperature or pressure; a function of spatial coordinates, such as Cartesian coordinates $(x, y, z)$; or a function of both space and time. If both the computational and experimental SRQs are deterministic quantities, the validation metric will also be a deterministic quantity. If either of the SRQs is a probability distribution or p-box, the result of the validation metric could be a number, a probability distribution, or a p-box; depending on the construction of the validation metric operator.

As discussed in Chapters 10 and 11, Figure 12.4 suggests the appropriate interaction between computation and experimentation that should occur in a validation experiment. To achieve the most value from the validation experiment, there should be in-depth, forthright, and frequent communication between analysts and experimentalists during the planning, design, and execution of the experiment. Also, after the experiment has been completed, the experimentalists should provide to the analysts all the important input quantities needed to conduct the simulation. What should *not* be provided to the analysts in a rigorous validation

activity is the measured SRQ. Stated differently, a blind-test prediction should be compared with experimental results so that a true measure of predictive capability can be assessed in the validation metric.

The right-center portion of Figure 12.4 shows the comparison of the validation metric result with the preliminary accuracy requirement for the application of interest. This is referred to as a *preliminary* requirement because the conditions of the application of interest, i.e., the application domain, may be different from the validation domain. If they are different, as is the usual case, then only preliminary accuracy requirements can be applied. If the model passes these preliminary accuracy requirements, then the model prediction would be made using extrapolation or interpolation of the model. Accuracy requirements for the application, i.e., final accuracy requirements, would be applied after the model has been extrapolated or interpolated to the conditions of the application of interest. Setting preliminary accuracy requirements is also useful if the preliminary model accuracy is inadequate. The following gives two examples. First, it may be concluded that the assumptions and approximations made in the conceptual model are inadequate to achieve the accuracy needed. This may necessitate a reformulation of the model, as opposed to calibration of model parameters at a later stage. Second, preliminary accuracy assessment provides a natural mechanism for adjusting and adapting the final accuracy requirements to the cost and schedule requirements of the engineering system of interest. For example, trade-offs can be made between required final accuracy, system design, and performance, in addition to cost and schedule.

If the preliminary accuracy requirements are not met, then one has two options, as shown in Figure 12.4. First, the dashed-line upper feedback loop provides for any of the activities referred to as model updating, model calibrating, or model tuning. Depending on the ability to determine parameters, either by direct measurement or by inference from a model, we have divided parameter updating into three categories: parameter measurement, parameter estimation, and parameter calibration. If the model parameters are updated, one could proceed again through the validation metric operator. It is now very likely that the validation metric result would be significantly reduced, implying much better agreement between the model and the experimental results. The parameter updating may be physically justifiable, or it may simply be expedient. How to determine the scientific soundness of the updating, and its impact on predictive capability, is a very difficult issue. This issue will be discussed in more detail in the next chapter. Second, the lower feedback loop could be taken if: (a) improvements or changes are needed in the experimental measurements, such as improved diagnostic techniques; (b) additional experiments are needed to reduce experimental uncertainty; or (c) additional experiments are needed that are more representative of the application of interest.

The right portion of the figure deals with the issue of use of the model to make a prediction for the conditions of the application of interest. It should be noted that Figure 12.4 is completely consistent with the diagram showing the three aspects of validation (Figure 2.8) in Chapter 2, Fundamental concepts and terminology. Also, Figure 12.4 is conceptually consistent with the ASME Guide diagram shown in Figure 2.5 of Chapter 2. The ASME

diagram, however, does not explicitly show the extrapolation or interpolation of the model to conditions where no data are available. Each of these figures can be applied at *any* level of the validation hierarchy. Two examples of applications requiring different types of extrapolation are given. First, suppose the physical experiment conducted is on the complete system of interest, i.e., the actual operational system hardware. Suppose the experiment, however, is conducted at a set of conditions that do not correspond to the actual operational conditions of interest because of physical size constraints, safety concerns, or environmental restrictions. The SRQs are measured in the experiment, predicted by the model, and a validation metric result is computed. Suppose further that the model satisfies the preliminary accuracy requirements. Then an extrapolation is required, using the model, to predict the SRQs of interest corresponding to the actual operational conditions of interest. This type of extrapolation is the most common in engineering applications.

Second, suppose now that the experiments in Figure 12.4 are conducted on all of the subsystems of the complete system of interest and no experiments were conducted on the complete system. Suppose further that all of the validation metrics of all of the SRQs of interest for each subsystem meet the preliminary accuracy requirements. Even if the subsystems are tested at the same operational conditions as the complete system, any real engineering system functions as a closely coupled, highly interacting group of subsystems. As a result, one could characterize this as a large extrapolation because the model of the complete system is required to predict the interactions of each of the models for each of the subsystems. This extrapolation, of course, has no parameter associated with it and the ability of the model to predict the interactions has in no way been assessed.

## 12.2 Approaches to parameter estimation and validation metrics

During the last decade, model validation (in the broad sense of the term) and validation metrics have received increased attention. Model validation has been primarily pursued using traditional approaches, specifically: parameter estimation, hypothesis testing, and Bayesian updating. These approaches, particularly hypothesis testing, are related to the concepts embodied in validation metrics as discussed here. Although we refer to parameter estimation and Bayesian updating as model calibration, many researchers and practitioners in the field refer to these approaches as model validation. In this section, parameter estimation, hypothesis testing, Bayesian updating, and newer approaches to validation metrics will be briefly discussed. It should be stressed that the development of quantitative measures for validation is a new field of research and various perspectives are being vigorously debated. A special workshop was held in May 2006 to bring together a number of researchers in the field to discuss and debate various approaches to model validation. For this workshop, three challenge problems were proposed, each one in a different engineering field: heat transfer, solid mechanics, and solid dynamics. All of the contributions to the workshop were published in a special issue of *Computer Methods in Applied Mechanics and Engineering* (Dowding *et al.*, 2008; Pilch, 2008). The reader should consult this special issue for an excellent survey of the wide ranging approaches to model validation.

### 12.2.1 Parameter estimation

In the 1960s, the structural dynamics community began developing sophisticated techniques for comparing computational and experimental results, as well as techniques for improving agreement between the results using parameter estimation techniques (Wirsching *et al.*, 1995). In many analyses in structural dynamics, certain model input parameters are considered as deterministic (but poorly known) quantities that are estimated using modal data on the structures of interest. A numerical optimization procedure is used so that the best agreement between computational and experimental results can be obtained for a single SRQ, or a group of SRQs. Multiple solutions of the model are required to evaluate the effect of different values of the model parameters on the SRQ. Although these techniques are used to compare computational and experimental results, their primary goal is to improve agreement based on available experimental data. During the last decade, more sophisticated and reliable methods have been developed for optimizing parameters in a wider range of stochastic systems (Crassidis and Junkins, 2004; Raol *et al.*, 2004; Aster *et al.*, 2005; van den Bos, 2007). In this more recent work, the parameters are considered as random variables represented as probability distributions with unknown parameters, or nonparametric distributions. The experimental data can come from multiple experiments on the same structure but under different loading conditions, or from experiments on different structures. A similar, but more complex, optimization procedure is used to determine the probability distributions for each of the uncertain parameters that maximize agreement between the model and the experimental results.

Although the terminology used in this text clearly refers to this type of procedure as calibration, as opposed to validation, we mention it here because the process begins with measuring the difference between the model results and the experimental results.

### 12.2.2 Hypothesis testing

Statistical hypothesis testing, or significance testing, is used throughout the experimental sciences (Wellek, 2002; Lehmann and Romano, 2005; Law, 2006; Devore, 2007). Hypothesis testing is a highly developed statistical method of choosing between two competing models of an experimental outcome by using probability theory to minimize the risk of an incorrect decision. In hypothesis testing, the validation-quantification measure is formulated as a "decision problem" to determine whether or not the hypothesized model is consistent with the experimental data. This technique is regularly used in the operations research (OR) community for comparing mutually exclusive models, i.e., the model is either true or false. For example, suppose the hypothesis is made that a coin is fair, i.e., in the toss of the coin it is equally likely that "heads" will appear as often as "tails." The competing hypothesis is that the coin is unfair. Experimental data are then obtained by tossing the coin a given number of times, say $N$, and recording what percentage of the outcomes is heads and what percentage is tails. Hypothesis testing then allows one to statistically determine the confidence of a fair coin. The confidence in the determination will depend on $N$, that is, as $N$ increases, the confidence in the conclusion increases.

Hypothesis testing has recently been employed in validation studies (Hills and Trucano, 2002; Paez and Urbina, 2002; Hills and Leslie, 2003; Rutherford and Dowding, 2003; Chen *et al.*, 2004; Dowding *et al.*, 2004; Hills, 2006). In this approach, the validation assessment is formulated as a decision problem of whether or not the model's predictions are consistent with the available empirical information. Typically, empirical observations are collectively compared to a distribution of realizations computed by the model, such as by Monte Carlo sampling, which together constitutes a prediction about some SRQ. The consistency between the model and the data is characterized as a probability value, with low probability values denoting a mismatch of such magnitude that it is unlikely to have occurred by chance. This probability is often considered meaningful only within the hypothesis-testing context and does not serve as a stand-alone validation metric that indicates the degree of agreement or disagreement between model and empirical results. Instead, the approach focuses on answering the yes-no question about whether a model is accurate to within a specified margin of error, or sometimes on the either-or question choosing between mutually exclusive models.

Despite its being pressed into service as a tool for validation, hypothesis testing is not ideally suited to the task. In general, the purpose of hypothesis testing is to identify statements for which there is compelling evidence of truth or falsehood. But this is a different goal from that of validation, which is more pragmatic and more focused on assessing the quantitative accuracy of a model. The model could be relatively "poor," but the question in validation could be stated as "How poor?" In hypothesis testing, an accuracy requirement could be stated, either as a preliminary or final accuracy statement, and then the result would be the probability that the model met the required accuracy. Two practical difficulties arise with this type of result. First, how should a builder of physics models or a project manager interpret the result? It is not intuitive how to interpret probability as an accuracy measure. The natural perspective of design engineers and project managers is to ask, "What is the absolute or relative error of the model?" Second, no matter what level of accuracy is specified, the model can be proven false at that level with more experimental data. That is, any model can be proven false, given enough data. As will be discussed shortly, a key difficulty with hypothesis testing is that the requirement of accuracy is built *directly into* the validation metric.

Even though the hypothesis-testing approach does not appear to be a constructive route forward for validation metrics, the approach has developed the concept of error types for incorrect conclusions drawn from hypothesis testing (Wellek, 2002; Lehmann and Romano, 2005; Law, 2006). The technique has identified two types of error in decisions that are useful and instructive for other types of validation metrics. It should be stressed that these two types of error are not limited to statistical analyses. They are actually types of error in logic. A type 1 error, also referred to as *model builder's risk*, is the error in rejecting the validity of a model when the model is actually valid. This can be caused by errors on both the computational side and the experimental side. On the computational side, for example, if a mesh is not sufficiently converged and the computational result is contaminated by numerical error, then an adverse comparison with experimental data is misleading. That is, a poor comparison leads one to conclude that a model needs to be improved or recalibrated when the source

of the poor comparison is simply an under-resolved mesh. On the experimental side, the model builder's risk is most commonly caused by a poor comparison of computational results and experimental data that is due to an unknown bias error in the experimental data. We believe that unknown bias errors in experimental results are the *most damaging* in validation because if the experimental measurement is accepted as correct, then it is concluded that the computational result is consistently in error; whereas in reality, the experimental data is the culprit. If the error is believed to be in the computation, then a great deal of effort will be expended trying to find the source of the error. Or worse, a mathematical model will be recalibrated using the biased experimental data. This results in *transferring* the experimental bias into the mathematical model and then biasing all future computations with the model.

The type 2 error, also referred to as *model user's risk*, is the error in accepting the validity of a model when the model is actually invalid. As with the type 1 error, this can be caused by errors on both the computational side and the experimental side. On the computational side, the logical reverse of the type 1 error described above can occur. That is, if a mesh is not sufficiently converged and the computational result agrees well with the experiment, then the favorable comparison is also misleading. For example if a finer mesh is used, one can find that the favorable agreement can disappear. This shows that the original favorable agreement has compensating, or canceling, errors in the comparison. We believe that compensating errors in complex simulations is a common phenomenon. Only the tenacious user of the model may dig deep enough to uncover compensating errors. Similarly, a model developer may suspect that there is a code bug that may be the cause of unexpectedly good results. In a competitive, time-constrained, or commercial code-development environment, such users or model developers as these can be very unpopular, and even muffled by co-workers and management. On the experimental side, the logical reverse of the type 1 error described above can occur. That is, if an unknown bias error exists in the experiment, and a favorable comparison between computational results and experimental data is obtained, the implication of good model accuracy is incorrect. Similarly to the type 2 error on the computational side, only the self-critical and determined experimentalist will continue to examine the experiment in an attempt to find any experimental bias errors.

Type 1 and type 2 errors are two edges of the same sword. In the operations research literature, however, it is well known that model user's risk is potentially the more disastrous. The reason, of course, is that an apparently correct model (one that has experimental evidence that it produces accurate results) is used for predictions and decision making, when in fact it is incorrect. Type 2 errors produce a false sense of security. In addition to the potentially disastrous use of the model, we contend that the model user's risk is also the more likely to occur in practice than the model builder's risk. The reason is that with experimental evidence that the model is valid, there is little or no interest by analysts, experimentalists, managers, or funding sources to expend any more time or resources pursuing possible deficiencies in either the model or the experiments. Everyone is enthused by the agreement of results and "Victory" is easily and quickly declared. Anyone who questions the results can risk loss of personal advancement or recognition within his or her

organization. Only with some possible future experimental data, or system failure, would the caution and wisdom of these individuals be recognized. But then, as is said, it's too late.

### 12.2.3 Bayesian updating

Bayesian updating, or Bayesian statistical inference, has received a great deal of attention during the last two decades from statisticians, risk analysts, and some physicists and structural dynamicists. For a modern treatment of the topic from a Bayesian perspective, see Bernardo and Smith (1994); Gelman *et al.* (1995); Leonard and Hsu (1999); Bedford and Cooke (2001); Ghosh *et al.* (2006); and Sivia and Skilling (2006). Although the process is quite involved, Bayesian analysis can be summarized in three steps. Step 1 is to construct, or assume, a probability distribution for each input quantity in the model that is chosen to be an adjustable, random variable. Step 2 involves conditioning or updating the previously chosen probability models for the input quantities based on comparison of the computational and experimental results. To update the probability models, one must first propagate input probability distributions through the model to obtain probability distributions for the SRQs commensurate with those measured in the experiment. The updating of the input probability distributions, using Bayes' Theorem to obtain posterior distributions, commonly assumes that the mathematical model is structurally correct. That is, the updating is done assuming that all of the disagreement between the model and the experimental data is due to deficient probability distributions of the parameters. Step 3 involves comparing new computational results with the existing experimental data and quantifying the changes in the updated probability distributions. The new computational results are obtained by propagating the updated probability distributions through the model. If any new experimental data becomes available, then the entire process is repeated. Much of the theoretical development in Bayesian estimation has been directed toward optimum methods for updating statistical models of uncertain parameters and in reducing the computational resources required for propagating the uncertainties through the model.

Bayesian methods have been offered as an alternative approach to validation (Anderson *et al.*, 1999; Hanson, 1999; Kennedy and O'Hagan, 2001; DeVolder *et al.*, 2002; Hasselman *et al.*, 2002; Zhang and Mahadevan, 2003; Chen *et al.*, 2006; O'Hagan, 2006; Trucano *et al.*, 2006; Bayarri *et al.*, 2007; Babuska *et al.*, 2008; Chen *et al.*, 2008). The Bayesian approach is sophisticated, comprehensive, and computationally demanding in terms of the number of model evaluations, i.e., solutions of the partial differential equations (PDEs), which are needed. In addition, it is associated with some controversy that may cause analysts to consider alternatives. One of the major criticisms, from a validation perspective, is that the updating usually assumes that the model is itself correct. Recent work has suggested Bayesian strategies to account for uncertainty in the structure of the model by way of a model bias error (Kennedy and O'Hagan, 2001; Higdon *et al.*, 2004, 2009; Rougier, 2007; Liu *et al.*, 2009; McFarland and Mahadevan, 2008; Wang *et al.*, 2009).

The Bayesian approach to validation is primarily interested in evaluating the subjective probability, i.e., a personal belief, that the model is correct. Yet, to our minds, this is not the

proper focus of validation. We are not concerned about anyone's belief that the model is right; we're interested in *objectively measuring the conformance of data with predictions*. We disavow the subjectivity that is central in the Bayesian approach. Specifically, we argue that validation should not depend on the analyst's prior opinion about the correctness of the model in question. This is, after all, a large part of what is being assessed in validation, so it seems that it would be proper to refrain from assuming a key element of the consequence. While subjectivity is perfectly reasonable for making personal decisions, it can be problematic when the methods are (a) applied to high-consequence decision-making, for example, nuclear power plant safety or environmental impact studies, and (b) applied to public safety regulations that are stated in terms of frequency of events.

Bayesian claims of individual rationality do not generalize with respect to decisions made by groups. Some have asserted that predictive models cannot be validated objectively (Hazelrigg, 2003). We strongly believe that it is possible to objectively measure the performance of a model vis-à-vis the data that is relevant to the model's performance. We recognize that there are components that influence professional judgment or other subjective elements of how validation assessments should be carried out. For example, the selection of the experimental data to be applied in validation is arguably a subjective decision, as is choosing which validation metric is to be used from among the several possible metrics. Once these issues are defined, however, the application of the metric can be an objective characterization of the disagreement between predictions and data. Even in this limited context, objectiveness is valuable. The conclusion ought not to be that we should abandon the quest for objectivity because subjectivity cannot be escaped entirely. That would be like empiricists concluding from the Heisenberg Uncertainty Principle that it's useless to measure anything at all.

The Bayesian perspective always prefers to place any analytical question in the context of decision making. Model accuracy assessment *per se* need not be a part of decision making. Asking whether or to what degree a model is supported or contradicted by available evidence is surely a legitimate question by itself. Assessing the value of a metric that measures its conformity with observations should be recognized as a reasonable solution to this question. What one does with the knowledge that different models have different metrics is not formally part of the problem. Bayesians might argue that it should be because making good decisions requires such inclusive considerations. We agree that this may well be true, but it does not thereby de-legitimize model accuracy assessments that are not subsumed as a part of decision problems. For various practical and inescapable reasons, model accuracy assessments are sometimes needed before the decision context has even been specified; for example, national security concerns and comparison of competitive system performance when other portions of the decision context have not yet be formulated.

### *12.2.4 Comparison of mean values*

A distinctly different approach, and one that is conceptually simpler, is to compare the estimated mean from both the experiment and the simulation. For the case where there

is only one experiment and only a deterministic simulation, the comparison approach is equivalent. For this type of comparison, one would intuitively think of using traditional vector norms. Let $\mathcal{S}(x_i)$ be the simulation result as a function of the control parameter $x_i$ and let $\mathcal{E}(x_i)$ be the experimental measurement, where $i = 1, 2, 3, \ldots, N$. The parameter $x_i$ could be an input parameter or it could be time in a time-dependent system. The $L_1$ and $L_2$ vector norms (normalized by $N$) have been used as validation metrics, which are given by

$$\|\mathcal{S} - \mathcal{E}\|_p = \left( \frac{1}{N} \sum_{i=1}^{N} |\mathcal{S}(x_i) - \mathcal{E}(x_i)|^p \right)^{1/p}, \tag{12.1}$$

where $p = 1$ or 2. Several researchers have constructed validation metrics based on the $L_1$ and $L_2$ norms (Coleman and Stern, 1997; Easterling, 2001, 2003; Stern *et al.*, 2001; Oberkampf and Trucano, 2002; Oberkampf and Barone, 2004, 2006). For strongly time-dependent responses or system responses with a wide spectrum of frequencies, metrics have been constructed that summed the amplitude and phases differences between the simulation and the experiment (Geers, 1984; Russell, 1997a,b; Sprague and Geers, 1999, 2004). Although the vector norm metrics and those that combine amplitude and phase errors have different perspectives, their common theme is a more engineering-oriented, less statistical approach than those discussed above. They focus only on comparing a deterministic value of the SRQ from the model with the estimated statistical mean (or a single time history) of the experimental data. Most of these investigators do not propagate uncertain input parameters through the model to obtain a probability distribution for the SRQs of interest. Rather, it is commonly assumed that the mean of any uncertain input quantities maps to the mean of the SRQs of interest.

The primary perceived advantage in deterministic computational results, as opposed to propagating input uncertainties to determine output uncertainties, is the much lower computational costs involved in deterministic simulations. Many computational analysts argue that computational resources are not available to provide both spatially and temporally resolved solutions, as well as a large number of nondeterministic solutions, for complex simulations. Risk assessment of high-consequence systems, e.g., safety of nuclear power reactors and underground storage of nuclear waste, has shown that with an adequate, but not excessive, level of physical modeling detail, one *can* afford the computational costs of nondeterministic simulations. However, we recognize that there is substantial resistance in many fields to attain both mesh-resolved, nondeterministic simulations. Consequently, there is a need to construct validation metrics that can be computed with only deterministic computational results. The perspective of many in the fields that are resistant to change is that physics modeling fidelity is *much more* important than uncertainty quantification in engineering analyses. As a result, they construct models of such physical complexity that they consume essentially all of the computer resources available for a single solution, leaving no time for nondeterministic simulations. This tradition is deeply embedded in many scientific and engineering fields and it will be very slow and difficult to change this culture.

Later in this chapter, the approach of Oberkampf and Barone (2004, 2006) for comparing mean values from computation and experiment will be discussed in depth.

### *12.2.5 Comparison of probability distributions and p-boxes*

The difference between the probability distributions of the experiment and the simulation can be characterized in many ways. Suppose $X$ and $Y$ are the CDFs from the experiment and the simulation, respectively. One could consider using the following types of differences between $X$ and $Y$: the convolved distribution $X - Y$, some type of average of $X - Y$, or some type of difference between the shapes of $X$ and $Y$. The characterization that seems to be the most useful and understandable in the context of validation of models is based on comparing the shape of the two CDFs. Random variables whose distribution functions are identical are said to be "equal in distribution." As discussed earlier, this cannot occur even with a perfect model of the physics. If the distributions are not quite identical in shape, the discrepancy can be measured with many possible measures that have been proposed in traditional statistics (D'Agostino and Stephens, 1986; Huber-Carol *et al.*, 2002; Mielke and Berry, 2007): statistical goodness of fit, probability scoring rules, and information theory.

The difficulty with these traditional measures is that they are only well developed for random variables, i.e., purely aleatory uncertainty. We are interested in determining the difference between experiment and simulation when either or both results are given by a p-box. None of the approaches discussed earlier are able to address a combination of aleatory and epistemic uncertainty in the simulation or experimental results. Later in this chapter, we will discuss an approach for computing a validation metric result using p–boxes as input (Oberkampf and Ferson, 2007; Ferson *et al.*, 2008; Ferson and Oberkampf, 2009).

## 12.3 Recommended features for validation metrics

A validation metric is a formal measure of the mismatch between predictions and experimental data. A low value of the metric means there is a good match and a higher value means that prediction and data disagree more. It should be possible to apply the validation metric when predictions are either deterministic or nondeterministic. The metric should be mathematically well behaved and intuitively understandable to engineers, project managers, and decision makers. In the following discussion we will recommend seven desirable properties of validation metrics that would be useful in assessing the accuracy of models used in science and engineering simulations (Oberkampf and Trucano, 2002; Oberkampf and Barone, 2004, 2006).

### *12.3.1 Influence of numerical solution error*

A metric should either (a) explicitly include an estimate of the numerical error in the SRQ of interest resulting from the simulation, or (b) exclude the numerical error in the SRQ of

interest only if the numerical error was previously estimated, by some reasonable means, to be negligible. The primary numerical error of concern here is the error due to lack of spatial and/or temporal resolution in the discrete solution, and secondarily the iterative solution error. Numerical error could be explicitly included in the predicted SRQ by including an upper and a lower estimated bound on the error, i.e., an epistemic uncertainty in the SRQ that would be represented by an interval. Recall that an interval is a special case of a p-box. An explicit inclusion of the numerical error in the simulation result would be appealing because it would clearly identify the numerical error contribution, exclusive of any other uncertainty, either in the experimental data or the simulation input data. However, it would add complexity to the theoretical derivation and calculation of the metric because it would require the metric to use p-boxes as input.

A simpler approach would be to quantitatively show that the numerical error is small before the metric is computed. The numerical error should be judged small in comparison to the estimated magnitude of the uncertainty in the experimental measurements and the simulation. For two-dimensional or three-dimensional unsteady PDEs, it is a formidable task to achieve mesh and time-step independence. If this can be done, however, one can eliminate the issue from the calculation and interpretation of the metric.

### *12.3.2 Assessment of the physics-modeling assumptions*

A metric should be a quantitative evaluation of predictive accuracy of the SRQ of interest, including all of the combined modeling assumptions, physics approximations, and previously obtained physical parameters embodied in the model. Stated differently, the metric evaluates the aggregate accuracy of the model, including all of its submodels, and all of the physical parameters associated with the models. Consequently, there could be offsetting errors or widely ranging sensitivities in the model that could show very accurate results for one SRQ, but poor accuracy for a different SRQ. If there is interest in evaluating the accuracy of individual submodels or the effect of the accuracy of individual input parameters within the model, one should conduct a sensitivity analysis of the SRQ to these effects. However, sensitivity analysis is a separate issue from constructing a validation metric.

### *12.3.3 Inclusion of experimental data post-processing*

A metric should include, either implicitly or explicitly, an estimate of the approximation error resulting from post-processing of the experimental and computational data to obtain the SRQ of interest. Examples of the types of post-processing of experimental data are as follows: (a) the construction of a regression function, e.g., least-squares fit, of the data to obtain a continuous function over a range of an input (or control) quantity; (b) the processing of experimental data obtained on a very different spatial or temporal scale than what is addressed, i.e., assumed, in the model; and (c) the use of complex mathematical models of the physically measured quantities to process the experimental data into a useable

SRQ. A case where the post-processing described in example (b) might be necessary is when there are localized underground measurements of a pollutant concentration and the model contains a large-scale, spatially averaged permeability model. One might require the type of post-processing defined in example (c) when additional mathematical models of the physical features measured are also needed to process and interpret the experimental data into a useable form for comparison with the computational results. Some examples of experimental data processing are (a) image processing, (b) processing the motion of markers in a sequence of images, (c) object recognition, and (d) pattern recognition.

It should be noted that this recommendation is completely separate from the features discussed in Section 12.3.1 above. Any numerical error associated with the post-processing of the numerical solution of PDEs should be considered as part of the error in the model. As discussed in Section 12.3.1, the numerical error should be either quantified using an interval or it should be demonstrated that it is small compared to the experimental and simulation uncertainty.

### *12.3.4 Inclusion of experimental uncertainty estimation*

A metric should incorporate, or include in some explicit way, an estimate of the aleatory and epistemic measurement uncertainties in the experimental data for the SRQ. The possible sources for measurement uncertainties depend on a very wide range of factors, some of which were discussed in Chapter 11. For a comprehensive discussion of experimental measurement uncertainty, see Grabe (2005); Rabinovich (2005); and Drosg (2007). At a minimum, a validation metric should include an estimate of random uncertainties, i.e., uncertainties due to random measurement errors. To the extent possible, the metric should also include an estimate of the systematic uncertainties in the experiment.

The epistemic measurement uncertainties are usually segregated into two parts: (a) characterization uncertainty due to a limited number of measurements of the SRQ, and (b) epistemic uncertainty that can exist in the measurement itself. The characterization uncertainty is due to limited knowledge of the stochastic nature of the random error in the measurements and is usually referred to as the sampling error. Epistemic uncertainty in the measurements themselves can be due to (a) uncertainty in a measurement that is reported as a plus-or-minus value, (b) uncertainty implied by specifying a certain number of significant digits in a quantity, (c) uncertainty due to intermittent measurements of a process that is known to be periodic, (d) uncertainty in measurements when the quantity being measured is less than the detectable limit of the diagnostic method, i.e., non-detects, (e) uncertainty due to statistical censoring, e.g., when the data is only known with specified ranges or bins, and (f) uncertainty due to important data that is known to be missing from a random variable (Manski, 2003; Gioia and Lauro, 2005; Ferson *et al.*, 2007). When a collection of such intervals comprise a data set, one can think of the breadths of the intervals as representing epistemic uncertainty, while the scatter among the intervals

represents aleatory uncertainty. The characterization uncertainty can be directly represented by using the individual measurements themselves, i.e., using an empirical distribution function (EDF). The epistemic uncertainty in the measurements would require a validation metric that could use a p-box as input to compute the difference between the experiment and the simulation.

If possible, the experimental uncertainty estimation method should use measurements from replications of experiments, as opposed to a propagation of uncertainty technique (ISO, 1995). If replications of the experiment are conducted, then the uncertainty in the metric should depend on the number of replicated measurements of a given SRQ of interest. Replications should be used, along with blocking and randomization, to vigorously attempt to quantify random and systematic uncertainties in measurements. In addition, if a SRQ is measured for different conditions of an input or control parameter, then techniques should be used to reduce the experimental uncertainty of the SRQ over the range of the input parameter.

### *12.3.5 Inclusion of aleatory and epistemic uncertainties*

A metric should, in a mathematically rigorous way, be able to compute the difference between the computational and experimental results for the SRQ when these results exhibit aleatory and epistemic uncertainty. This uncertainty, for example, could be due to uncertain input quantities or an estimate of numerical solution error. If the computational and experimental results both exhibit aleatory and epistemic uncertainty, then they could both be characterized as p-boxes or some other imprecise probability structure, such as belief and plausibility functions in evidence theory (Krause and Clark, 1993; Almond, 1995; Kohlas and Monney, 1995; Klir and Wierman, 1998; Fetz *et al.*, 2000; Helton *et al.*, 2004, 2005; Oberkampf and Helton, 2005; Bae *et al.*, 2006).

Consider the case when the computational and experimental results are each characterized as precise probability distributions, i.e., only aleatory uncertainty is present. If the variance in each distribution approaches zero, then the difference between the distributions should approach the difference in the mean of each distribution. As another example, suppose the computational and experimental results are each characterized as p-boxes, i.e., a combination of aleatory and epistemic uncertainty. If the aleatory and epistemic uncertainty in each distribution approach zero, then the difference between the p-boxes should reduce to the difference in the two point values, i.e., the difference between two scalar quantities.

### *12.3.6 Exclusion of any type of adequacy implication*

A metric should *exclude* any indications, either explicit or implicit, of the level of adequacy in agreement, or satisfaction of accuracy requirements, between computational and experimental results. That is, the metric should only measure the mismatch between the

computational and experimental results, separate from *any* other features or characteristics of the computational or experimental results. If any other features or characteristics are combined in the mismatch measure, then one defeats the goal of independently setting accuracy requirements, as set by the intended use of the model (see Figure 12.4, as well as Figures 2.5 and 2.8). Some examples of inappropriately combining the mismatch feature of the comparison with other features, are the following: (a) comparisons of computational and experimental results that yield or imply value judgments, such as "adequate," "good," or "excellent"; (b) computational results judged to be adequate if they lie within some stated uncertainty band or observed range of the experimental measurements; (c) a comparison that combines the mismatch measure and the confidence or probability of the mismatch; and (d) a comparison that combines the mismatch measure and an estimate of the experimentally observed uncertainty.

This last example of constructing an inappropriate validation metric (combining the metric with the experimentally observed uncertainty) is hotly debated. The most common method of constructing an inappropriate metric such as this is to scale the mismatch between the computational and experimental results by the standard deviation of the experimental measurements. For example, suppose the L1 vector norm is used as the mismatch measure between computation and experiment and this norm is normalized by the sample standard deviation of the measurements, $s$. For the validation metric one would have $\|\mathcal{S} - \mathcal{E}\|_1 / s$. We argue that this metric would be inappropriate because the metric has explicitly mixed two different types of measure; a difference measure between simulation and experiment SRQ, and a measure of the scatter of the experimental data. The scatter in the experimental data should not be mixed with the first measure because the experimental scatter is controlled by sources that have nothing to do with the ability of the model to predict the observed responses of the system. Two examples of these sources are uncertainty in the response due to uncertainty in input quantities and experimental measurement uncertainty. What would be acceptable, in our view, would be a validation metric that measures the mismatch between the predicted standard deviation and the measured standard deviation.

### *12.3.7 Properties of a mathematical metric*

A validation metric should be a *true metric* in the mathematical sense, i.e., a true distance measure. The validation metric would then measure, by some means, the distance between the simulation and experimental results. By definition, a mathematical metric $d$ has four properties (Giaquinta and Modica, 2007):

| | |
|---|---|
| non-negativity, | $d(x, y) \geq 0$, |
| symmetry, | $d(x, y) = d(y, x)$, |
| triangle inequality, | $d(x, y) + d(y, z) \geq d(x, z)$, and |
| identity of indiscernibles, | $d(x, y) = 0$ if and only if $x = y$. |

## 12.4 Introduction to the approach for comparing means

### *12.4.1 Perspectives of the present approach*

The present approach computes a validation metric by comparing the estimated mean of the computational results with the estimated mean of the experimental measurements. A statistical confidence interval is computed that reflects the confidence in the estimation of model accuracy, given the uncertainty in the experimental data. Although a comparison of mean values gives only very limited information, it is typically the first quantity that is considered when the accuracy of a prediction is considered. This type of metric would be useful for situations in which a computational analyst, a model developer, or competing model developers are interested in quantifying which model among alternative models is most accurate for a given set of experimental data. In addition, this type of metric would be useful to a design engineer or a project engineer for specifying model accuracy requirements in a particular application domain of the model. It should be noted that if the application domain is outside the validation domain, one must account for the additional uncertainty due to extrapolation of the model to the application domain.

The validation metric developed in this section satisfies most, but not all of the recommendations given in the previous section. Here we summarize the seven recommendations and comment if the present metric does not satisfy a particular recommendation.

1 Influence of numerical solution error – yes.
2 Assessment of physics-modeling assumptions – yes.
3 Inclusion of experimental data post-processing – yes.
4 Inclusion of experimental uncertainty estimation – yes, except for epistemic uncertainty.
5 Inclusion of aleatory and epistemic uncertainties – no. Since the metric only makes a comparison between the mean values of the experimental and computational SRQs, aleatory uncertainty in the SRQs due to aleatory uncertainty in the input quantities is not addressed. In addition, the metric cannot deal with any epistemic uncertainties.
6 Exclusion of any type of adequacy implication – yes.
7 Properties of a mathematical metric – no. The metric does not satisfy the symmetry property because it takes into account whether the model result is greater than or less than the experimental measurement. The metric does not satisfy the triangle inequality property because it does not measure the distance between the computational result and the experimental measurement in two dimensions; it only measures in the dimension of the SRQ.

The validation metrics developed here are applicable to SRQs that do not have a periodic character and that do not have a complex mixture of many frequencies. For example, the present metrics would not be appropriate for analysis of standing or traveling waves in acoustics or modal analyses in structural dynamics. Another example of an inappropriate use would be the time-dependent fluid velocity at a point in turbulent flow. These types of SRQ require sophisticated time-series analysis and/or mapping to the frequency domain.

The input quantities that should be used, if possible, in the simulation of the SRQ of interest are those that are *actually measured* in the validation experiment. Some of these input quantities from the experiment may not be known for various reasons, for example,

a quantity may be epistemically uncertain or it may be a random variable that is not independently measurable before the experiment. If an input quantity is a random variable and it is well characterized, then it should be propagated through the model to obtain a probability distribution for the SRQ. To avoid this computational cost, it is commonly assumed that propagating only the mean, i.e., the expected value, of all uncertain input parameters through the model can approximate the mean value of the SRQ. This approach is appropriate under two conditions. First, the response of the system is linear in the input random variables; or second, it is locally accurate in terms of the system response when (a) the coefficient of variation (COV) of the important input random variables is small, and (b) the model is not extremely nonlinear with respect to these random variables. The COV is defined as $\sigma/\mu$, where $\sigma$ and $\mu$ are the standard deviation and mean of the random variable, respectively. We will briefly discuss this assumption, however, it is addressed in many texts on propagation of uncertain inputs through a model. See, for example, Haldar and Mahadevan (2000).

A Taylor series can be written that clarifies the nature of the approximation. Let $Y_{\mathrm{m}}$ be the SRQ that is the random variable resulting from the model. Let $g(\bullet)$ represent the PDE with the associated initial conditions and boundary conditions that map uncertain inputs to the uncertain SRQ. And let $\chi_i$, where $i = 1, 2, \ldots, n$, be the uncertain input random variables. Assuming appropriate smoothness in the solution to the PDE, a Taylor series for uncorrelated input random variables can be expanded about the mean of each of the input variables, $\mu_x$, and written as (Haldar and Mahadevan, 2000)

$$E(Y_{\mathrm{m}}) = g(\mu_{\chi_1}, \mu_{\chi_2}, \ldots, \mu_{\chi_n}) + \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\partial^2 g}{\partial \chi_i^2}\right)_{\mu_{\chi_i}} \mathrm{Var}(\chi_i) + \cdots. \tag{12.2}$$

$E(Y_{\mathrm{m}})$ is the expected value, i.e., the mean, of the SRQ and $\mathrm{Var}(\chi_i)$ is the variance of each of the input variables. It is seen from Eq. (12.2) that the first term of the expansion is simply $g$ evaluated at the mean of the input variables. The second term is the second derivative of $g$ with respect to the input variables. This term, in general, will be small with respect to the first term if either $g$ is nearly linear in the input variables or the COV of all of the input variables is small. Linearity in the response of the system as a function of the input variables essentially never occurs when the mapping of inputs to outputs is given by a differential equation, even a *linear* differential equation.

In summary, the present validation metric requires the mean of the SRQ for comparison with the experimental data. The most accurate method of obtaining this mean is to propagate, usually through a sampling procedure, the uncertain inputs through the model and obtain the probability distribution of the SRQ. From a sufficiently sampled distribution, the mean can be easily computed. If the uncertainty propagation approach is not taken, then one can use the assumption discussed above. One must recognize, however, that there can be significant error in this procedure. Note that when using this approximation one could obtain poor agreement between computational and experimental results, and the cause is *not* the

model *per se*, but the inaccuracy of the computational mean caused by the assumption of the propagation of the mean of the inputs.

### 12.4.2 Development of the fundamental equations

The fundamental ideas of the present validation metric are developed for the case where the SRQ of interest is defined for a single value of an input or operating condition variable. This will allow some discussion of how the present approach implements some of the recommended features discussed above, as well as giving an opportunity to review the classical development of statistical confidence intervals. Since it may be confusing why we begin the development of validation metrics with a discussion of statistical confidence intervals, we make the following point. We are interested in obtaining a difference measure between a computational result and the *mean of a population* of experimental measurements for which only a finite set of measurements has been obtained. Once this goal is understood, it is realized that the key issue is the statistical nature of the sample mean of the measured system response, *not* the level of mismatch between the computational result and the sample mean. With this perspective, it becomes clear that the point of departure should be a fundamental understanding of the statistical procedure for estimating a confidence interval for the true (population) mean. In traditional statistical testing procedures, specifically hypothesis testing, the point of departure is the derivation for the confidence interval of the difference between two hypotheses. As a result, hypothesis testing immediately embeds a stated level of agreement, or disagreement, in the difference operator, making it impossible to satisfy the recommendation discussed in Section 12.3.6.

A short review and discussion will be given for the construction of a statistical confidence interval. The development of confidence intervals is discussed in most texts on probability and statistics. The following development is based on the derivation by Devore (2007), Chapter 7.

Let $X$ be a random variable characterizing a population having a mean $\mu$ and a standard deviation $\sigma$. Let $x_1, x_2, \ldots, x_n$ be actual sample observations from the population, which are assumed to be the result of a random sample $X_1, X_2, \ldots, X_n$ from the population. Let $\bar{X}$ be the sample mean, which is a random variable based on the random sample $X_1, X_2, \ldots, X_n$. Provided that $n$ is large, the central limit theorem implies that $\bar{X}$ has approximately a normal distribution, *regardless* of the nature of the population distribution. Then it can be shown that the standardized random variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{12.3}$$

has an approximate normal distribution with zero mean and a standard deviation of unity. $S$ is the sample standard deviation, which is a random variable, based on random samples $X_1, X_2, \ldots, X_n$. It can also be shown, provided $n$ is large, that a probability interval for $Z$

can be written as

$$P(z_{-\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha. \tag{12.4}$$

$z_{\alpha/2}$ is the value of the random variable $Z$ at which the integral of $Z$ from $z_{-\alpha/2}$ to $+\infty$ is $\alpha/2$. Since $Z$ is symmetrical and has its mean at zero, the integral of $Z$ from $-\infty$ to $z_{-\alpha/2}$ is also equal to $\alpha/2$. The total area from both tail intervals of the distribution is $\alpha$.

Equation (12.4) can be rearranged to show that the probability interval for $\mu$, the mean of the population that is the unknown quantity of interest, is given by

$$P\left(\bar{X} - z_{\alpha/2} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha. \tag{12.5}$$

For sufficiently large $n$, Eq. (12.5) can be rewritten as a confidence interval for the population mean using sampled quantities for the mean and standard deviation,

$$\mu \sim \left(\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right). \tag{12.6}$$

$\bar{x}$ and $s$ are the sample mean and standard deviation, respectively, based on $n$ observations. Note that $\bar{x}$ and $s$ are computed from the realizations $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. The term $s/\sqrt{n}$ is the standard error of the sample mean that measures how far the sample mean is likely to be from the population mean. The level of confidence that $\mu$ is in the interval given by Eq. (12.6) can be shown to be $100(1 - \alpha)\%$. The value of $\alpha$ is arbitrarily assigned and is typically chosen to be 0.1 or 0.05, corresponding to confidence levels of 90% or 95%, respectively.

The confidence interval for the population mean can be interpreted in a strict frequentist viewpoint or in a subjectivist, or Bayesian, viewpoint. Let $C$ be the confidence level chosen, i.e., $C = 100(1 - \alpha)\%$, for stating that the true mean $\mu$ is in the interval given by Eq. (12.6). The frequentist would state, "$\mu$ is in the interval given by Eq. (12.6) with confidence $C$," which means that if the experiment on which $\mu$ is estimated is performed repeatedly; for a sufficiently large number of samples $\mu$ will fall in the interval given by Eq. (12.6) $C\%$ of the time. The subjectivist would state (Winkler, 1972), "Based on the observed data, it is my belief that $\mu$ is in the interval given by Eq. (12.6) with probability $C$." The reason that it *cannot* be strictly stated that $C$ is the probability that $\mu$ is in the interval given by Eq. (12.6) is that the true probability is either zero or one. That is, the true mean $\mu$ is either in the interval or it is not; we simply *cannot know with certainty* for a finite number of samples from the population. Notwithstanding these fine points of interpretation, we will essentially use the subjectivist interpretation in a slightly different form than that presented above. We will use the interpretation: $\mu$ is in the interval given by Eq. (12.6) with confidence $C$.

Now consider the case of calculating a confidence interval for an arbitrary number of experimental observations $n$, with $n$ as small as two. It can be shown (Devore, 2007) that if it is assumed that the samples are drawn from a normal distribution, the equation analogous

to Eq. (12.6) is

$$\mu \sim \left( \bar{x} - t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} \right). \tag{12.7}$$

The level of confidence is given by $100(1-\alpha)\%$ and $t_{-\alpha/2,\nu}$ is the $1-\alpha/2$ quantile of the $t$ distribution for $\nu = n-1$ degrees of freedom. For $n$ greater than 16, the cumulative $t$ distribution and the cumulative standard normal distribution differ by less than 0.01 for all quantiles. In the limit as $n \to \infty$, the $t$ distribution approaches the standard normal distribution.

Equation (12.7) can be used for hypothesis testing in classical statistical analysis. However, our perspective in the construction of validation metrics is notably different. We wish to quantify the difference between the computational results and the true mean of the experimental results. Stated differently, we wish to measure shades of gray between a model and an experiment – not make a "yes" or "no" statement about the congruence of two hypotheses for a given accuracy level.

### 12.4.3 Construction of the validation metric for one condition

For the validation metric we wish to construct, we are interested in two quantities. First, we want to estimate an error in the SRQ of the model based on the difference between the model and the estimated mean of the population based on the experimentally measured samples of the SRQ. Let $y_{\mathrm{m}}$ be the SRQ from the model. Changing the notation used previously for the experimental measurements from $\bar{x}$ to $\bar{y}_{\mathrm{e}}$, we define the estimated error in the model as

$$\tilde{E} = y_{\mathrm{m}} - \bar{y}_{\mathrm{e}}. \tag{12.8}$$

$\bar{y}_{\mathrm{e}}$ is the estimated, or sample, mean based on $n$ experiments conducted. $\bar{y}_{\mathrm{e}}$ is given by

$$\bar{y}_{\mathrm{e}} = \frac{1}{n} \sum_{i=1}^{n} y_{\mathrm{e}}^{i}, \tag{12.9}$$

where $y_{\mathrm{e}}^{1}, y_{\mathrm{e}}^{2}, \ldots, y_{\mathrm{e}}^{n}$ are the individually measured results of the SRQ from each experiment.

Second, we wish to compute an interval that contains the true error in the model, which we do not know, at a specified level of confidence. Let the true error in the model $E$ be defined as

$$E = y_{\mathrm{m}} - \mu, \tag{12.10}$$

where $\mu$ is the true mean of the population. Writing the confidence interval expression, Eq. (12.7), for $\mu$ as an inequality relation and changing the notation as just mentioned, we have

$$\bar{y}_{\mathrm{e}} - t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{y}_{\mathrm{e}} + t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}}. \tag{12.11}$$

$s$ is the sample (not population) standard deviation given by

$$s = \left[ \frac{1}{n-1} \sum_{i=1}^{n} \left( y_{\mathrm{e}}^{i} - \bar{y}_{\mathrm{e}} \right)^2 \right]^{1/2}. \tag{12.12}$$

Multiplying Eq. (12.11) by –1 and adding $y_{\mathrm{m}}$ to each term, we have

$$y_{\mathrm{m}} - \bar{y}_{\mathrm{e}} + t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} > y_{\mathrm{m}} - \mu > y_{\mathrm{m}} - \bar{y}_{\mathrm{e}} - t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}}. \tag{12.13}$$

Substituting the expression for the true error, Eq. (12.10), into Eq. (12.13) and rearranging, one obtains

$$y_{\mathrm{m}} - \bar{y}_{\mathrm{e}} - t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} < E < y_{\mathrm{m}} - \bar{y}_{\mathrm{e}} + t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}}. \tag{12.14}$$

Substituting the expression for the estimated error, Eq. (12.8), into Eq. (12.14), we can write the inequality expression as an interval containing the true error where the level of confidence is given by $100\,(1-\alpha)\,\%$:

$$\left( \tilde{E} - t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}}, \tilde{E} + t_{\alpha/2,\nu} \cdot \frac{s}{\sqrt{n}} \right). \tag{12.15}$$

Using the level of confidence of 90%, one can state the validation metric in the following way: the estimated error in the model is $\tilde{E} = y_{\mathrm{m}} - \bar{y}_{\mathrm{e}}$ with a confidence level of 90% that the true error is in the interval

$$\left( \tilde{E} - t_{0.05,\nu} \cdot \frac{s}{\sqrt{n}}, \tilde{E} + t_{0.05,\nu} \cdot \frac{s}{\sqrt{n}} \right). \tag{12.16}$$

Three characteristics of this validation metric should be mentioned. First, the statement of confidence is made concerning an interval in which the true error is believed to occur. The statement of confidence is *not* made concerning the magnitude of the estimated error, nor concerning an interval around the computational prediction. The reason such statements cannot be made is that the fundamental quantity that is uncertain is the *true* experimental mean. Stated differently, although we are asking how much error there is in the computational result, the actual uncertain quantity is the *referent*, i.e., the true experimental value, *not* the computational result.

Second, the interval believed to contain the true error is symmetric around the estimated error. We can also state that the rate of decrease of the magnitude of the interval is a factor of 2.6 when going from two experiments to three experiments, the sample standard deviation $s$ remaining constant. For a large number of experiments, the rate of decrease of the magnitude of the interval is $1/\sqrt{n}$. Additionally, the size of the interval decreases linearly as the sample standard deviation decreases.

Third, for small numbers of experimental measurements it must be assumed that the measurement uncertainty is normally distributed. Although this is a very common assumption in experimental uncertainty estimation, and probably well justified, it is rarely *demonstrated* to be true. However, for a large number of experimental measurements, as discussed above,

the confidence interval on the mean is valid regardless of the type of probability distribution representing measurement uncertainty.

Finally, we stress the primacy we give to the experimental data. As can be clearly seen from Eq. (12.10), the referent for the error measure is the experimental data, not the model or some type of weighted average between the model and the experiment. However, our trust in the accuracy of experimental measurements is not without some risk, specifically, if an undetected bias error exists in the experimental data.

### 12.4.4 Example problem: thermal decomposition of foam

As an example of the application of the validation metric just derived, consider the assessment of a model for the rate of decomposition of polyurethane foam due to thermal heating. The model solves the unsteady energy equation for the heating of the foam and is composed of three major components: (a) thermal diffusion through the materials involved, (b) chemistry models for the thermal response and decomposition of polymeric materials due to high temperature, and (c) radiation transport within the domain and between the boundaries of the physical system. The foam decomposition model predicts the mass and species evolution of the decomposing foam and was developed by Hobbs *et al.* (1999). Dowding *et al.* (2004) computed the results for this example using the computer code Coyote that solves the mathematical model using a finite element technique (Gartling *et al.*, 1994). Three-dimensional, unsteady solutions were computed until the foam decomposed, vaporized, and escaped from a vent in the container. The container was a cylinder with a diameter of 88 mm and a length of 146 mm. Solution verification for the computational results relied on the mesh refinement studies previously conducted by Hobbs *et al.* (1999). These earlier mesh refinement studies estimated that the mesh discretization error was less than 1% for the velocity of the foam decomposition front for mesh sizes less than 0.1 mm.

The experiment designed to evaluate the model was composed of polyurethane foam enclosed in a stainless steel cylinder that was heated using high-intensity lamps (Figure 12.5). The experiment was conducted by Bentz and Pantuso and is reported in Hobbs *et al.* (1999). The position of the foam–gas interface was measured as a function of time by X-rays passing through the cylinder. The steel cylinder was vented to the atmosphere to allow gas to escape, and it was heated from different directions during different experiments: top, bottom, and side. For some of the experiments, a solid stainless steel cylinder or hollow aluminum component was embedded in the foam.

The SRQ of interest is the steady-state velocity of the foam decomposition front when the front has moved between 1 and 2 cm. The steady-state velocity was typically achieved after 5 to 10 minutes, depending on the heating temperature. The SRQ was measured as a function of the imposed boundary condition temperature. Since we are only considering one operating condition for the present validation metric example, we pick the temperature condition of $T = 750\,^{\circ}\mathrm{C}$ because it had the largest number of experimental replications. Some of the replications, shown in Table 12.1, were the result of different orientations

Table 12.1 *Experimental data for foam decomposition (Hobbs* et al.*, 1999).*

| Experiment no. | Temperature (°C) | Heat orientation | *V* (experiment) (cm/min) |
|---|---|---|---|
| 2 | 750 | bottom | 0.2323 |
| 5 | 750 | bottom | 0.1958 |
| 10 | 750 | top | 0.2110 |
| 11 | 750 | side | 0.2582 |
| 13 | 750 | side | 0.2154 |
| 15 | 750 | bottom | 0.2755 |

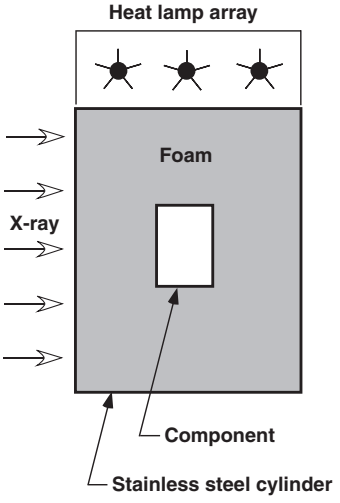


Figure 12.5 Schematic of foam decomposition experiment (Oberkampf and Barone, 2004).

of the heat lamps. No estimates of experimental measurement uncertainty were provided. Computational simulations by Dowding *et al.* (2004) showed that cylinder orientation had little effect on the velocity of the decomposition front. Since we are only interested in a single deterministic result from the model, we picked one of the Dowding *et al.* results for the computational SRQ. The computational prediction for the foam decomposition velocity was 0.2457 cm/min. With this approximation, we assigned the variability resulting from the heating orientation of the cylinder to uncertainty in the experimental measurements. In addition, there is variability in the material composition of the foam during the fabrication process resulting in variability of the decomposition front velocity for a fixed temperature. Consequently, the material variability effect is confounded with the experimental measurement uncertainty.

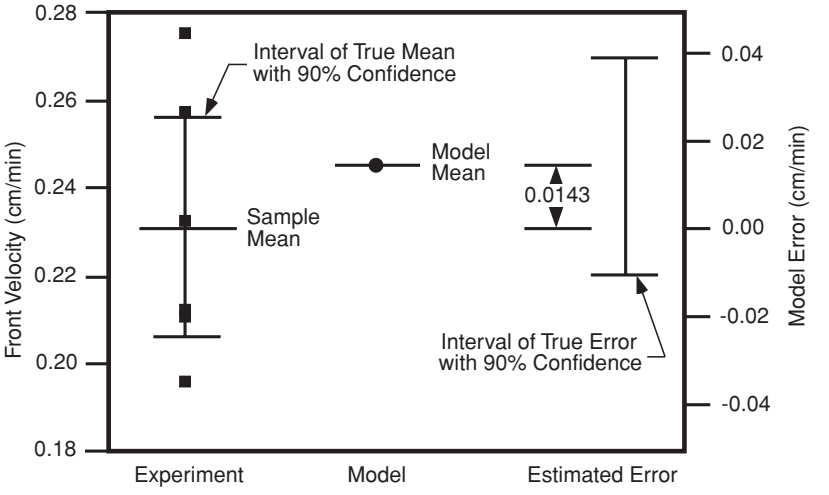


Figure 12.6 Statistical and validation metric results of foam decomposition (Oberkampf and Barone, 2006).

Using the data in Table 12.1 and Eqs. (12.7)–(12.9), (12.12), and (12.16), we obtain

number of samples $= n = 6$,
sample mean $= \bar{y}_e = 0.2314$ cm/min,
estimated error $= \tilde{E} = 0.2457 - 0.2314 = 0.0143$ cm/min,
sample standard deviation $= s = 0.0303$ cm/min,
degrees of freedom $= n - 1 = \nu = 5$,
$t$ distribution for 90% confidence $(\nu = 5) = t_{0.05,\nu} = 2.015 \pm t_{0.05,\nu} \cdot \frac{s}{\sqrt{n}} = \pm 0.0249$ cm/min,
true mean with 90% confidence $= \mu \sim (0.2065, 0.2563)$ cm/min,
true error with 90% confidence $\sim (-0.0106, 0.0392)$ cm/min.

Figure 12.6 depicts the sample mean, the model mean, the estimated interval of the true mean, and the estimated error with 90% confidence. In summary form, the result of the validation metric is $\tilde{E} = 0.0143 \pm 0.0249$ cm/min with 90% confidence. Note that the validation metric result is given in the physical units of the SRQs being compared, *not* some sort of statistical measure or a probability. Since the magnitude of the uncertainty in the experimental data is roughly twice the estimated error in the model, one cannot make any more precise conclusions than $\pm 0.0249$ cm/min (with 90% confidence) concerning the accuracy of the model.

Whether the estimated accuracy with its uncertainty is adequate for the intended use of the model is a separate step in the encompassing view of validation, as discussed earlier. If the estimated accuracy with its uncertainty is not adequate for a model use decision, then one has two options. The first option, which is the more reasonable option for this case, is to reduce the experimental uncertainty by obtaining additional experimental measurements, changing the experimental procedure, or improving the diagnostic method to reduce the experimental uncertainty. The second option would be to improve, or update, the model so

that it gives more accurate results for this situation. However, in the present case, the error in the model is small with respect to the experimental uncertainty. As a result, this option would be uncalled for.

## 12.5 Comparison of means using interpolation of experimental data

### *12.5.1 Construction of the validation metric over the range of the data*

We are now interested in the case where the SRQ is measured over a range of the input variable or the operating condition variable. For example, in the foam decomposition experiment just discussed, we would be interested in the velocity of the foam decomposition front as a function of the heating temperature of the cylinder. Another example would be the thrust of a rocket motor as a function of burn time. Here we consider the case of one input variable while all others are held constant. This is probably the most common type of comparison between computational and experimental results. The present ideas could be extended fairly easily to the case of multiple input variables as long as the input variables were independent.

The following assumption is made with regard to the computational results. The SRQ is computed at a sufficient number of values over the range of the input variable to allow an accurate construction of an interpolation function to represent the SRQ.

The following assumptions are made with regard to the experimental measurements.

(1) The input variable from the experiment is measured much more accurately than the SRQ. Quantitatively, this means that the coefficient of variation (COV) of the input variable is much smaller than the COV of the SRQ. The assumption must relate the COV of each quantity because the COV is a dimensionless statistic of the variability of a random variable. Note that this assumption allows for the case where the input variable is uncontrolled in the experiment, and could even be a random variable. However, the key is that it can be accurately measured for each replication of the experiment.
(2) Two or more experimental replications have been obtained, and each replication has multiple measurements of the SRQ over the range of the input variable. Using the terminology of Coleman and Steele (1999), it is desirable that *N*th-order replications have been obtained, and possibly even replications by different experimentalists using different facilities and different diagnostic techniques.
(3) The measurement uncertainty in the SRQ from one experimental replication to the next, and from setup to setup, is given by a normal distribution.
(4) Each experimental replication is independent from other replications; that is, there is zero correlation or dependence between one replication and another.
(5) For each experimental replication, the SRQ is measured at a sufficient number of values over the range of the input variable so that a smooth and accurate interpolation function can be constructed to represent the SRQ. By *interpolation* we mean that the function constructed to represent the data must match each measured value of the SRQ.

With these assumptions, the equations developed above are easily extended to the case in which both the computational result and the experimental mean for the SRQ are functions

of the input variable $x$. Rewriting Eq. (12.16), the true error as a function of $x$ is in the interval

$$\left(\tilde{E}(x) - t_{0.05,\nu} \cdot \frac{s(x)}{\sqrt{n}},\ \tilde{E}(x) + t_{0.05,\nu} \cdot \frac{s(x)}{\sqrt{n}}\right), \tag{12.17}$$

with a confidence level of 90%. The standard deviation as a function of $x$ is given by

$$s(x) \sim \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{\mathrm{e}}^{i}(x) - \bar{y}_{\mathrm{e}}(x)\right)^2\right]^{1/2}. \tag{12.18}$$

Note that $y_{\mathrm{e}}^{i}(x)$ is interpolated using the experimental data from the $i$th experimental replication, i.e., the ensemble of measurements over the range of $x$ from the $i$th experiment. Each experimental replication need not make measurements at the same values of $x$ because a separate interpolation function is constructed for each ensemble of measurements, i.e., each $i$th experimental replication.

### 12.5.2 Global metrics

Although these equations provide the results of the validation metric as a function of $x$, there are some situations where it is desirable to construct a more compact, or global, statement of the model accuracy. For example, in a project management review it may be useful to quickly summarize the accuracy for a large number of models and experimental data. A convenient method to compute a global metric would be to use a vector norm of the estimated error over the range of the input variable. The $L_1$ norm is useful to interpret the estimated average absolute error of the model over the range of the data. Using the $L_1$ norm, one could form an average absolute error or a relative absolute error over the range of the data. We choose to use the relative absolute error by normalizing the absolute error by the estimated experimental mean and then integrating over the range of the data. We define the *average relative error metric* to be

$$\left|\frac{\tilde{E}}{\bar{y}_{\mathrm{e}}}\right|_{\mathrm{avg}} = \frac{1}{(x_{\mathrm{u}} - x_{\mathrm{l}})}\int_{x_{\mathrm{l}}}^{x_{\mathrm{u}}}\left|\frac{y_{\mathrm{m}}(x) - \bar{y}_{\mathrm{e}}(x)}{\bar{y}_{\mathrm{e}}(x)}\right| \mathrm{d}x. \tag{12.19}$$

$x_{\mathrm{u}}$ is the largest value and $x_{\mathrm{l}}$ is the smallest value, respectively, of the input variable. As long as $|\bar{y}_{\mathrm{e}}(x)|$ is not near zero for any $x_{\mathrm{l}}$, the average relative error metric is a useful quantity.

The confidence interval that should be associated with this average relative error metric is the average confidence interval normalized by the absolute value of the estimated experimental mean over the range of the data. We define the *average relative confidence indicator* as the half-width of the confidence interval averaged over the range of the data:

$$\left|\frac{\mathrm{CI}}{\bar{y}_{\mathrm{e}}}\right|_{\mathrm{avg}} = \frac{t_{0.05,\nu}}{(x_{\mathrm{u}} - x_{\mathrm{l}})\sqrt{n}}\int_{x_{\mathrm{l}}}^{x_{\mathrm{u}}}\left|\frac{s(x)}{\bar{y}_{\mathrm{e}}(x)}\right| \mathrm{d}x. \tag{12.20}$$

We refer to $|\mathrm{CI}/\bar{y}_{\mathrm{e}}|_{\mathrm{avg}}$ as an indicator, as opposed to an interval, because the uncertainty structure of $s(x)$ is not maintained through the integration operator. Although $|\mathrm{CI}/\bar{y}_{\mathrm{e}}|_{\mathrm{avg}}$

is not an average relative confidence interval over the range of the data, it is a quantity useful for interpreting the significance of the magnitude of $|\tilde{E}/\bar{y}_e|_{avg}$. Stated differently, the magnitude of $|\tilde{E}/\bar{y}_e|_{avg}$ should be interpreted relative to the magnitude of the normalized uncertainty in the experimental data, $|\mathrm{Cl}/\bar{y}_e|_{avg}$.

There may be situations where the average relative error metric may not adequately represent the model accuracy because of the strong smoothing nature of the integration operator. For example, there may be a large error at some particular point over the range of the data that should be noted. It is useful to define a maximum value of the absolute relative error over the range of the data. Using the $L_\infty$ norm to accomplish this, we define the *maximum relative error metric* as

$$\left|\frac{\tilde{E}}{\bar{y}_e}\right|_{max} = \max_{x_l \le x \le x_u} \left|\frac{y_m(x) - \bar{y}_e(x)}{\bar{y}_e(x)}\right|. \tag{12.21}$$

If one observes a significant difference between $|\tilde{E}/\bar{y}_e|_{avg}$ and $|\tilde{E}/\bar{y}_e|_{max}$, then one should more carefully examine the trend of the model with respect to the trend of the experimental data. For example, if $|\tilde{E}/\bar{y}_e|_{max}$ is much greater than $|\tilde{E}/\bar{y}_e|_{avg}$, the model is failing to predict either a local or global trend of the experimental data.

The confidence interval that should be associated with the maximum relative error metric is the confidence interval normalized by the estimated experimental mean. Both the confidence interval and the estimated experimental mean are evaluated at the point where the maximum relative error metric occurs. Let the $x$ value where $|\tilde{E}/\bar{y}_e|_{max}$ occurs be defined as $\hat{x}$. Then the confidence interval half-width associated with the maximum relative error metric is

$$\left|\frac{\mathrm{CI}}{\bar{y}_e}\right|_{max} = \frac{t_{0.05,\nu}}{\sqrt{n}} \left|\frac{s(\hat{x})}{\bar{y}_e(\hat{x})}\right|. \tag{12.22}$$

### 12.5.3 Example problem: turbulent buoyant plume

As an example of the validation metric just derived, consider the assessment of a model for a turbulent buoyant plume that is exiting vertically from a large nozzle. Turbulent buoyant plumes, typically originating from the combustion of fuel–air mixtures, have proven to be especially difficult to model in CFD. This is primarily because of the strong interaction between the density field and the momentum field dominated by large turbulent eddies. The slowest turbulent scales are on the order of seconds in large fires, and this large-scale unsteadiness is beyond the modeling capability of a Reynolds–average Navier–Stokes (RANS) formulation. The model to be evaluated here solves the continuity equation and the temporally filtered Navier–Stokes (TFNS) equations. The TFNS equations are similar to RANS equations, but a narrower filter width is used so that large-scale unsteadiness can be captured (Pruett *et al.*, 2003). Tieszen *et al.* (2005) computed the unsteady,

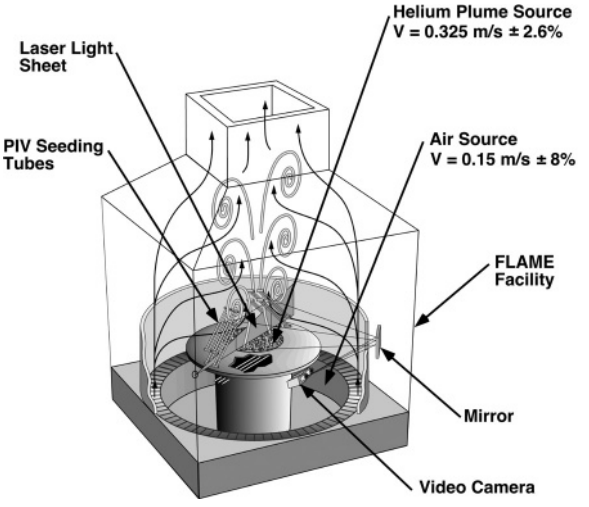


Figure 12.7 Experimental setup for measurements of the helium plume (O'Hern *et al.*, 2005). See color plate section.

three-dimensional simulations used here for a large-scale helium plume using the TFNS model and the standard $k$-$\varepsilon$ turbulence model.

The experimental data for the validation metric were obtained in the Fire Laboratory for Accreditation of Models and Experiments (FLAME) facility at Sandia National Laboratories. The FLAME facility is a building designed for indoor fire experiments, as well as other buoyant plumes, so that the plumes are not influenced by atmospheric winds, and all other boundary conditions affecting the plume can be measured and controlled. For the present experiment, a large inflow jet of helium was used (Figure 12.7) (DesJardin *et al.*, 2004; O'Hern *et al.*, 2004). The helium source is 1 m in diameter and is surrounded by a 0.51-m wide horizontal surface to simulate the ground plane that is typical in a fuel-pool fire. Inlet air is allowed in from outside the building at the bottom of the facility and is drawn in by the vertically accelerating helium plume that exits the chimney of the building.

The experimental data consist of velocity field measurements using particle image velocimetry (PIV) and scalar concentration measurements using planar-induced fluorescence (PLIF). Here we are interested in only the PIV measurements, but details of all of the diagnostic procedures and uncertainty estimates can be found in O'Hern *et al.* (2005). The PIV data are obtained from photographing the flowfield, which has been seeded with glass microspheres, at 200 images/s. Flowfield velocities are obtained in a plane that is up to 1 m from the exit of the jet and illuminated by a laser light sheet. The flow velocity of interest here, i.e., the SRQ that is input to the validation metric, is the time-averaged vertical velocity component along the centerline of the helium jet. For unsteady flows such as this, there are a number of large-scale oscillatory modes that exist within the plume, as well as turbulence scales that range down to the micron level. The SRQ of interest is time

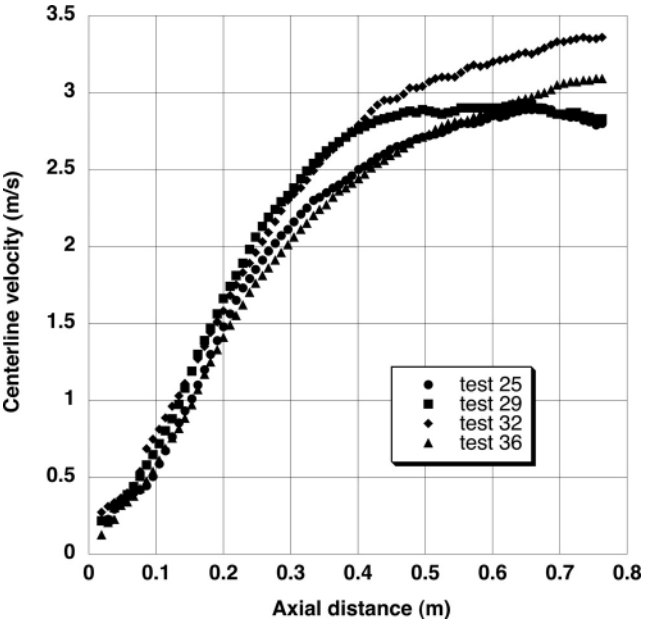


Figure 12.8 Experimental measurements of time-averaged vertical velocity along the centerline for the helium plume. Data from O'Hern *et al.* (2005).

averaged for roughly 10 s in the experiment, which is roughly seven cycles of the lowest oscillatory mode in the jet.

Shown in Figure 12.8 are four experimental measurements of time-averaged vertical velocity along the centerline as a function of axial distance from the exit of the helium jet. The experimental replications were obtained on different days, with different equipment setups, and with multiple recalibrations of the instrumentation. A large number of velocity measurements were obtained over the range of the input variable, the axial distance, so that an accurate interpolation function could be constructed.

Tieszen *et al.* (2005) investigated the sensitivity of their numerical solutions for the helium plume to both modeling parameters and numerical discretization on an unstructured mesh. The key modeling parameter affecting the TFNS solutions is the size of the temporal filter relative to the period of the largest turbulent mode in the simulation. Four spatial discretizations were investigated, resulting in the following total number of mesh points: 0.25M, 0.50M, 1M, and 2M elements ($1M = 10^6$). In order to process the vertical velocity solutions in the same way that the experimentalist processed the data, each of these solutions was time averaged over roughly seven puffing cycles. Using the method for computing the observed order of spatial convergence discussed in Chapter 8, Discretization error, and Tieszen *et al.* (2005) solutions for 0.50M, 1M, and 2M elements, it does not appear that the solutions are in the asymptotic region. A finer mesh, say, 4M elements, would greatly help in determining whether the computational results are actually converged. However,

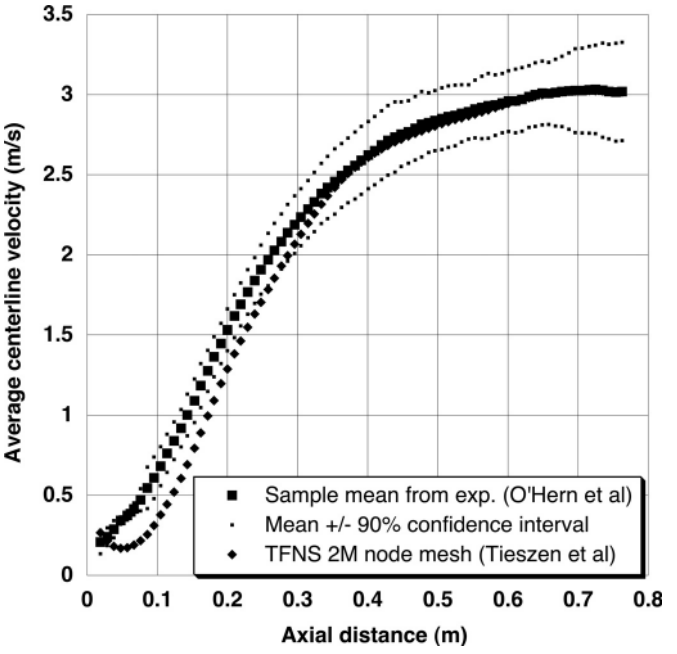


Figure 12.9 Experimental sample mean with 90% confidence interval and computational result for vertical velocity in the helium plume (Oberkampf and Barone, 2004).

computational resources were not available to compute the 4M-element solution. As a result, we will use their 2M-element solution as only representative data with which to demonstrate the present validation metric.

Using the experimental data shown in Figure 12.8, noting that $n = 4$, one obtains the sample mean of the measurements, $\bar{y}_e(x)$, shown in Figure 12.9. Also, using the interpolated function for the experimental sample mean and the confidence interval for the true mean, one obtains the interval around the estimated mean in which the true mean will occur with 90% confidence (Figure 12.9). The computational solution obtained from the 2M-element mesh is also shown in Figure 12.9.

The level of disagreement between computational and experimental results can be more critically seen by plotting the estimated error, $\tilde{E}(x) = y_m(x) - \bar{y}_e(x)$, instead of simply showing the SRQ as a function of the input variable. Figure 12.10 shows this type of plot, along with the 90% confidence interval from the experiment, as given by Eq. (12.17). Presentation of model accuracy assessment results such as Figure 12.10, even without the confidence interval, is rarely seen in practice. This type of plot very critically examines the differences between the model and the experimental results. Even though it presents the same information as Figure 12.9, the critical examination of the difference never flatters the model or the experiment. In Figure 12.10, the largest modeling (relative) error occurs very near the beginning of the plume. This error is noticeable in Figure 12.9, but it is not dominant as it is in Figure 12.10. We remind the reader that we are discussing the model

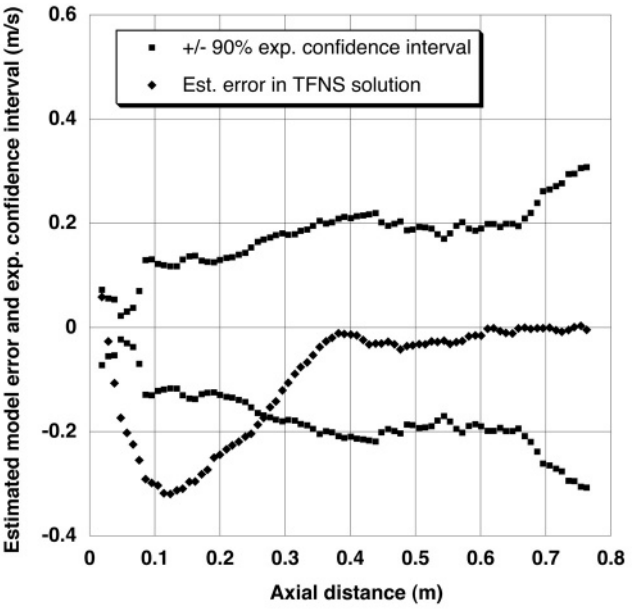


Figure 12.10 Validation metric result and 90% confidence interval for centerline velocity (Oberkampf and Barone, 2004).

error as if it were due to physics modeling, however, it may not be that at all. It may be simply due to the insufficiently mesh converged solution, but at this point one cannot be certain what the source of the error is.

The validation metric result shown in Figure 12.10 can be quantitatively summarized using the global metrics given in Eqs. (12.19)–(12.22). Over the range of the data, these results are as follows:

average relative error $= 11\% \pm 9\%$ with 90% confidence,
maximum relative error $= 54\% \pm 9\%$ with 90% confidence.

Thus, the average relative error could be as large as 20% and as small as 2% (on average) over the range of the data, with 90% confidence considering the uncertainty in the experimental data. The average relative error shows that the model accuracy (on average) is comparable to the average confidence indicator in the experimental data. Similarly, the maximum relative error could be as small as 45% and as large as 63%, with 90% confidence considering the uncertainty in the experimental data. The maximum relative error, 54%, which occurs at $x = 0.067$ m, is five times the average relative error, indicating a significant difference in the local character of the model and the experimental data. Note that, for these experimental data, the average relative confidence indicator, 9%, happens to be essentially equal to the relative confidence interval at the maximum relative error; however, this need not be the case in general. If one was using both the average relative error and the maximum relative error for a "first look" evaluation of the model, a large difference between these

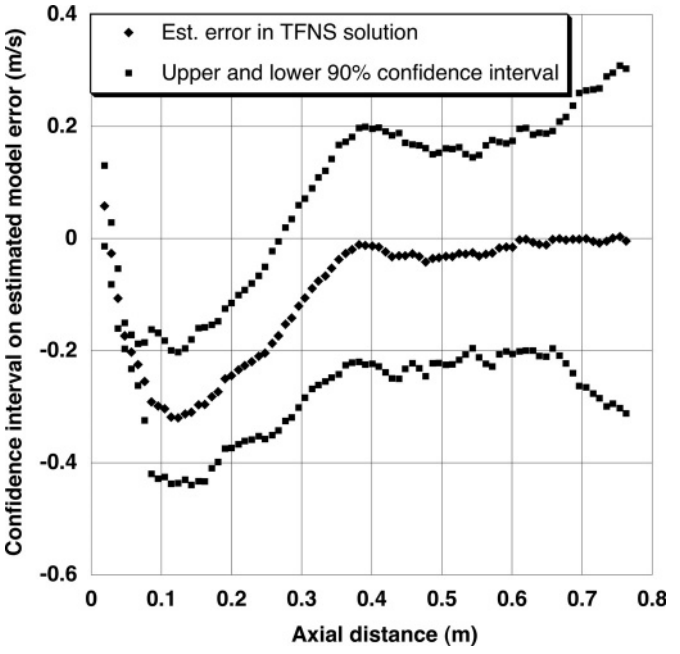


Figure 12.11 Estimated error and true error in the model with 90% confidence interval (Oberkampf and Barone, 2004).

values should prompt a more careful examination of the data, for example, examination of plots such as Figure 12.9 and Figure 12.10.

The final method of displaying the results of the validation metric is to plot the 90% confidence interval of the true error in velocity predicted by the model as a function of the axial distance from the exit of the jet. Using Eq. (12.17), one obtains the result shown in Figure 12.11. Our best approximation of the true error in the model is the estimated error. However, with 90% confidence we can state that the true error is in the interval shown in Figure 12.11.

Although Figure 12.11 displays essentially the same data as shown in Figure 12.10, Figure 12.11 allows us to consider slightly different perspectives for assessing the model. For example, we could view Figure 12.11 from the perspectives of those who might use the validation metric results to evaluate the predictive capability of the model. A model builder, for example, would likely investigate the cause of the largest error, i.e., near $x =$ 0.1 m, and explore ways to improve the model and/or compute a solution on a more highly resolved mesh. For an analyst, i.e., a person who is going to use the model for predictions of flowfields that are related to the present flowfield, the perspective is somewhat different from that of the model builder. The analyst might conclude that the accuracy of the model is satisfactory for its intended use and simply apply the model as it is. Alternatively, the analyst might decide to use Figure 12.11 to incorporate a bias-error correction directly on the SRQ, i.e., the vertical velocity on the centerline of the plume. However, this procedure for model

correction would clearly involve high risk because it completely ignores the physical and/or numerical cause of the error. Stated differently, the analyst would be treating the estimated error *as if* it was physics, but this claim is very weak because of evidence of an unresolved mesh solution.

## 12.6 Comparison of means requiring linear regression of the experimental data

### *12.6.1 Construction of the validation metric over the range of the data*

We are now interested in a case where the quantity of experimental data is not sufficient to construct an interpolation function over the range of the input variable. Consequently, a regression function (curve fit) must be constructed to represent the estimated mean over the range of the data. Some examples are lift (or drag) of a flight vehicle as a function of the Mach number, turbopump mass flow rate as a function of backpressure, and depth of penetration into a material during high-speed impact. Construction of a regression function is probably the most common situation that arises in comparing computational and experimental results when the input variable is *not* time. When time-dependent SRQs are recorded, the temporal resolution is typically high so that the construction of an interpolation function is commonly used. Time series analyses, however, must normally deal with both high-frequency characteristics in the SRQs *and* uncertainty in the experimental measurements.

Regression analysis procedures are well developed in classical statistics for addressing how two or more variables are related to each other when one or both contain random uncertainty. We are interested here in the restricted case of univariate regression, i.e., how one variable (the SRQ) relates to another variable (the input variable) when there is only uncertainty in the SRQ. The first four assumptions pertaining to the experimental measurements discussed in Section 12.5.1 are also made for the present case. In addition to these, the following assumption is made with regard to the experimental uncertainty: the standard deviation of the normal distribution that describes the measurement uncertainty is constant over the entire range of measurements of the input parameter. It should also be noted that this assumption is probably the most demanding of the experimental measurement assumptions listed.

In the present development, it was initially thought that traditional confidence intervals, as discussed above, could be applied when a regression analysis was involved. We realized, however, that commonly used confidence intervals only apply to the case of a specific, but arbitrary, value of the input parameter. That is, the traditional confidence interval is a statement of the accuracy of the estimated mean as expressed by the regression for *point values* of the input parameter $x$. The traditional confidence interval is written for $\mu$ conditional on a point value of $x$, say, $x^*$, i.e. $\mu[\bar{y}_e(x)|x^*]$, where (|) denotes that the preceding quantity is conditional on the following quantity. As a result, the traditional confidence interval analysis cannot be applied to the case of a validation metric over a range of the input variable where the determination of a regression is also involved.

A more general statistical analysis procedure was found that develops a confidence interval for the entire range of the input parameter (Miller, 1981; Draper and Smith, 1998;

Seber and Wild, 2003). That is, we wish to determine the confidence interval that results from uncertainty in the regression coefficients over the complete range of the regression function. The regression coefficients are all correlated with one another because they appear in the same regression function used to fit the range of the experimental data. This type of confidence interval is typically referred to as a *simultaneous confidence interval*, a *simultaneous inference*, or a *Scheffé confidence interval*, so that it can be distinguished from traditional (or single comparison) confidence intervals.

Let the set of $n$ experimental measurements of the SRQ of interest be given by

$$\left(y_{\mathrm{e}}^{i}, x_i\right) \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{12.23}$$

Here we consider the simplest case of representing the estimated mean of the data, $\bar{y}_{\mathrm{e}}(x)$, using a linear regression function over the range of the data. The linear regression function is written as

$$\bar{y}_{\mathrm{e}}(x) = \theta_1 + x\theta_2 + \varepsilon, \tag{12.24}$$

where $\theta_1$ and $\theta_2$ are the unknown coefficients of the regression function; and $\varepsilon$ is the random measurement error. For this case, the equations for the Scheffé confidence intervals can be analytically derived (Miller, 1981).

The estimate of the interval for the true mean $\mu(x)$ is given by

$$\mu(x) \sim (\bar{y}(x) - \mathrm{SCI}(x), \bar{y}(x) + \mathrm{SCI}(x)), \tag{12.25}$$

where $\mathrm{SCI}(x)$ is the width of the Scheffé confidence interval as a function of $x$ and is given by

$$\mathrm{SCI}(x) = s\sqrt{[2F(2, n-2, 1-\alpha)]\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right]}, \tag{12.26}$$

where $s$ is the standard deviation of the residuals for the entire curve fit, $F(\nu_1, \nu_2, 1-a)$ is the $F$ probability distribution, $\nu_1$ is the first parameter specifying the number of degrees of freedom, $\nu_2$ is the second parameter specifying the number of degrees of freedom, $1 - \alpha$ is the quantile for the confidence interval of interest, $n$ is the number of experimental measurements, $\bar{x}$ is the mean of the input values of the experimental measurements, and $s_x^2$ is the variance of the input values of the experimental measurements. One has the definitions

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left[y_{\mathrm{e}}^{i} - \bar{y}_{\mathrm{e}}(x_i)\right]^2}, \tag{12.27}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{12.28}$$

$$s_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{12.29}$$

Using the definition of the estimated error in the model, Eq. (12.8), one can write the estimated error as a function of $x$ as

$$\tilde{E}(x) = y_{\mathrm{m}}(x) - \bar{y}_{\mathrm{e}}(x), \tag{12.30}$$

Using this equation, and Eq. (12.25), the interval containing the true error with confidence level $100(1-\alpha)\%$ is given by

$$(\tilde{E}(x) - \mathrm{SCI}(x),\, \tilde{E}(x) + \mathrm{SCI}(x)). \tag{12.31}$$

One can compare Eq. (12.31) with the equation for the traditional confidence interval, Eq. (12.17). One finds that the equations are identical except for the coefficient

$$\sqrt{2F(2, n-2, 1-\alpha)}. \tag{12.32}$$

For the traditional confidence interval one has the coefficient $t_{\alpha/2,n-1}$. In general,

$$\sqrt{2F(2, n-2, 1-\alpha)} > t_{\alpha/2,n-1}. \tag{12.33}$$

As a result, the Scheffé confidence intervals are always larger, commonly by a factor of two, than traditional confidence intervals. This larger confidence interval reflects the experimental uncertainty on the entire regression function; not just uncertainty at a given value of $x$.

### 12.6.2 Global metrics

If we would like to make a quantitative assessment of the global modeling error, then we can use the global measures expressed earlier in Eqs. (12.19) and (12.21). However, the average relative confidence indicator, Eq. (12.20), and the confidence interval associated with the maximum relative error, Eq. (12.22), must be changed to take into account the simultaneous nature of the regression. Using Eq. (12.31), one can rewrite Eq. (12.20) as

$$\left|\frac{\mathrm{CI}}{\bar{y}_{\mathrm{e}}}\right|_{\mathrm{avg}} = \frac{1}{(x_{\mathrm{u}} - x_{\mathrm{l}})} \int_{x_{\mathrm{l}}}^{x_{\mathrm{u}}} \left|\frac{\mathrm{SCI}(x)}{\bar{y}_{\mathrm{e}}(x)}\right| \mathrm{d}x \tag{12.34}$$

for the average relative confidence indicator.

The confidence interval associated with the maximum relative error metric, Eq. (12.22), is

$$\left|\frac{\mathrm{CI}}{\bar{y}_{\mathrm{e}}}\right|_{\mathrm{max}} = \left|\frac{\mathrm{SCI}\,(\hat{x})}{\bar{y}_{\mathrm{e}}\,(\hat{x})}\right|, \tag{12.35}$$

where $\hat{x}$ is the $x$ value where $|\tilde{E}/\bar{y}_{\mathrm{e}}|_{\mathrm{max}}$ occurs.

### 12.6.3 Example problem: thermal decomposition of foam

As an example of use of the validation metric using linear regression, consider again the model for the thermal decomposition of polyurethane foam described in Section 12.4.4.

Table 12.2 *Experimental and computational data for a range of operating conditions, (Hobbs* et al.*, 1999).*

| Experiment no. | Temperature (°C) | Heat orientation | $V$ (experiment) (cm/min) | $V$ (computation) (cm/min) |
|---|---|---|---|---|
| 1 | 600 | bottom | 0.1307 | 0.0913 |
| 2 | 750 | bottom | 0.2323 | 0.2457 |
| 5 | 750 | bottom | 0.1958 | 0.2457 |
| 10 | 750 | top | 0.2110 | 0.2457 |
| 11 | 750 | side | 0.2582 | 0.2457 |
| 13 | 750 | side | 0.2154 | 0.2457 |
| 15 | 750 | bottom | 0.2755 | 0.2457 |
| 14 | 900 | bottom | 0.3483 | 0.4498 |
| 16 | 1000 | bottom | 0.5578 | 0.7698 |

Now, however, a more complete set of data is considered for heating the foam over a range of temperatures. The experimental data for foam decomposition obtained by Bentz and Pantuso is shown in Table 12.2. The computational data given by Easterling (2001, 2003); and Dowding *et al.* (2004) were used for a range of temperature conditions from $T = 600\,^{\circ}\text{C}$ to $T = 1000\,^{\circ}\text{C}$. Table 12.2 shows the experimental and computational results used to compute the various elements of the validation metric.

Using the experimental data in Table 12.2, standard linear regression methods, and Eqs. (12.24)–(12.29), one obtains

number of samples $= n = 9$,
$y$ intercept from the linear curve fit $= \theta_1 = -0.5406$,
slope from the linear curve fit $= \theta_2 = 0.001042$,
standard deviation of the residuals of the curve fit $= s = 0.04284$,
square of the regression coefficient $= R^2 = 0.895$,
0.9 quantile of the $F$ probability distribution for $\nu_1 = 2$ and $\nu_2 = 7$, $F(2, 7, 0.9) = 3.26$,
mean of the $x$ input values $= \bar{x} = 777.8$,
variance of the $x$ input values $= s_x^2 = 12\,569$,
Scheffé confidence interval $= \text{SCI}(x) = \pm 0.044284\sqrt{1 + 8.9 \times 10^{-5}(x - 777.8)^2}$.

Note that here we use the generic variable $x$, as developed in the equations, to represent the temperature.

To obtain a continuous function for the model, a curve fit of the computational data in Table 12.2 was made. As discussed in Section 12.4.4, the computational analysis of Hobbs *et al.* (1999) showed that the discretization error was less than 1% of the front velocity. In addition to the four computational values listed in Table 12.2, it was recommended that the model is linear for the lower ranges of temperature (Hobbs, 2003). As a result, a cubic spline curve fit was used to represent the model over the range of data.

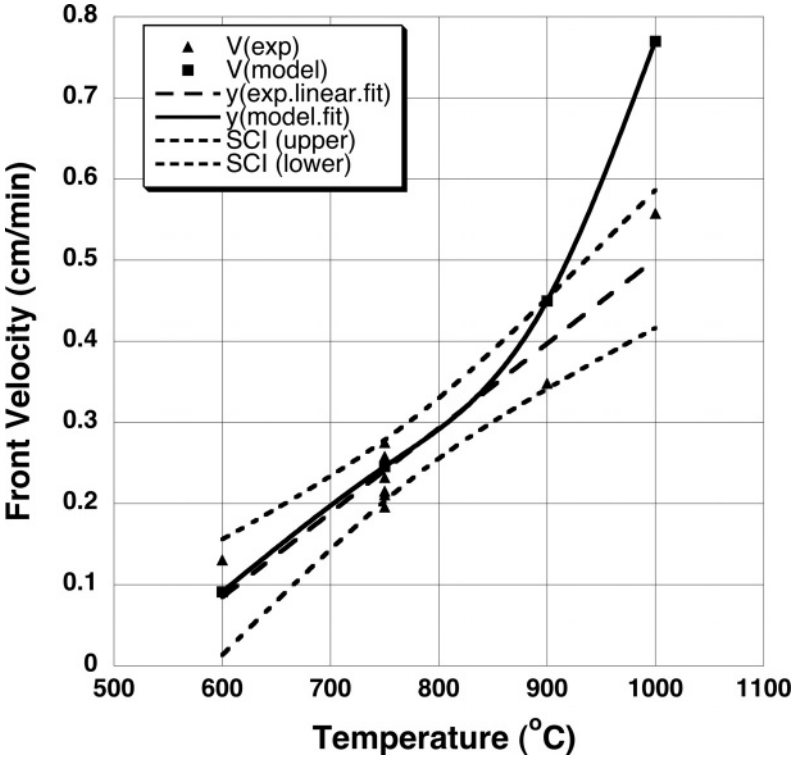


Figure 12.12 Estimated sample mean and simultaneous confidence intervals using linear regression for foam decomposition compared with the model prediction.

Figure 12.12 shows the comparison of the results for the linear regression of the experimental data, the Scheffé confidence intervals, and computational curve fit results. Two observations should be made from Figure 12.12. First, the linear regression of the experimental data seems to be a reasonably accurate representation of the data, noting that $R^2 = 0.895$. Second, the model falls within the confidence interval up to a temperature of roughly 850 °C and then departs markedly from the experimental data.

Figure 12.13 shows the estimated model accuracy over the range of experimental data, as well as the simultaneous confidence intervals for the experimental data. As mentioned above, the calculation of a validation metric result focuses directly on the mismatch between computation and experiment. As a result, any weaknesses in the model or the experimental data are much more evident than the typical comparison, such as Figure 12.12. Also note in Figure 12.13 that the simultaneous confidence intervals are symmetric with respect to zero as a result of Eq. (12.31). The upper and lower simultaneous confidence intervals are a hyperbolic conic section centered on $\bar{x} = 0$ because of the form of Eq. (12.26). One can intuitively surmise that altering the values of temperature where data were obtained could significantly decrease the magnitude of the confidence intervals. It is left as an exercise to the reader to investigate an optimum set of temperatures that minimize the magnitude of

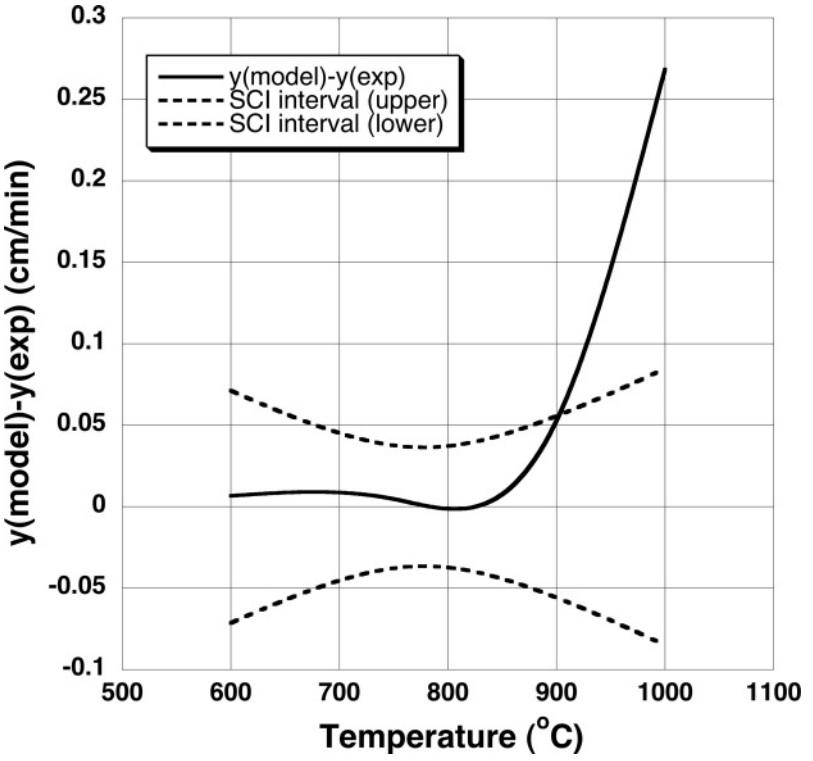


Figure 12.13 Validation metric result for foam decomposition over the range of operating conditions.

the confidence intervals, given the constraint of a fixed number of temperatures. This is the type of question that should be addressed early in the design of validation experiments.

Using Eqs. (12.19), (12.21), (12.34), and (12.35), the global validation metrics, along with their confidence indicators, can be computed as

average relative error = 11.2% ± 22.7% with 90% confidence,
maximum relative error = 53.5% ± 16.9% with 90% confidence.

Consider these global metrics as if viewing them from the perspective of not having examined the two previous graphs. For example, in a summary presentation to management, there might not be sufficient time to show the two previous graphs, so only the global metrics are given. The average relative error result is not particularly noteworthy, but it is seen that the average experimental uncertainty is roughly twice the average relative error. It is clear that the large experimental uncertainty prohibits any clear conclusions concerning the average model error. It is seen that the maximum relative error is a factor of 4.8 larger than the average relative error. When this occurs, it signals that there is a significant error in the trend of the model with respect to the trend in the experimental data. Since the relative experimental uncertainty is much smaller than the maximum relative error at that

condition, one can be certain that there is an incorrect trend in the model. With this signal of a modeling problem, management may ask for more details concerning the issues involved.

## 12.7 Comparison of means requiring nonlinear regression of the experimental data

### *12.7.1 Construction of the nonlinear regression equation*

Now consider the general case where we need to represent the estimated mean of the experimental data, $\bar{y}_e(x)$, as a general nonlinear regression function,

$$\bar{y}_e(x) = f(x;\, \vec{\theta}) + \varepsilon. \tag{12.36}$$

$f(x; \bullet)$ is the chosen form of the regression function over the range of the input parameter $x$; $\vec{\theta} = \theta_1, \theta_2, \ldots, \theta_p$ are the unknown coefficients of the regression function; and $\varepsilon$ is the random measurement error. Using a least-squares fit of the experimental data, it can be shown (Draper and Smith, 1998; Seber and Wild, 2003) that the error sum of squares $S(\vec{\theta})$ in $p$-dimensional space is

$$S(\vec{\theta}) = \sum_{i=1}^{n} \left[ y_e^i(x) - f(x_i;\, \vec{\theta}) \right]^2. \tag{12.37}$$

The vector that minimizes $S(\vec{\theta})$ is the solution vector, and it is written as $\vec{\hat{\theta}}$. This system of simultaneous, nonlinear equations can be solved by various software packages that compute solutions to the nonlinear least-squares problem. (See, for example, Press *et al.*, 2007.)

### *12.7.2 Computation of simultaneous confidence intervals for the metric*

Draper and Smith (1998) and Seber and Wild (2003) discuss a number of methods for the computation of the confidence regions around the point $\vec{\hat{\theta}}$ in $p$-dimensional space. For any specified confidence level $100(1-\alpha)\%$, a unique region envelops the point $\vec{\hat{\theta}}$. For two regression parameters, $(\theta_1, \theta_2)$, we have a two-dimensional space, and these regions are contours that are similar to ellipses with a curved major axis. For three parameters, $(\theta_1, \theta_2, \theta_3)$, we have a three-dimensional space, and these regions are contours that are similar to bent ellipsoids, i.e., shaped like a banana. A procedure that appears to be the most robust to nonlinear features in the equations (Seber and Wild, 2003) and that is practical when $p$ is not too large, is to solve an inequality for the set of $\vec{\theta}$:

$$\vec{\theta} \text{ such that } S(\vec{\theta}) \leq S(\vec{\hat{\theta}}) \left[ 1 + \frac{p}{n-p} F(p, n-p, 1-\alpha) \right]. \tag{12.38}$$

In Eq. (12.38), $F(\nu_1, \nu_2, 1-\alpha)$ is the $F$ probability distribution; $\nu_1 = p, \nu_2 = n - p, 1 - \alpha$ is the quantile for the confidence interval of interest; and $n$ is the number of experimental measurements.

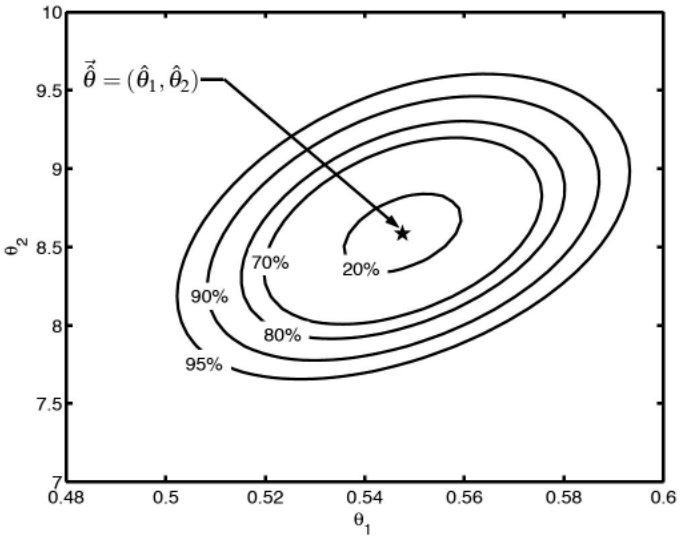


Figure 12.14 Example of various confidence regions for the case of two regression parameters (Oberkampf and Barone, 2004).

We consider a geometric interpretation of Eq. (12.38) to facilitate the numerical evaluation of the inequality. We seek the complete set of $\vec{\theta}$ values that satisfy the inequality. For a given confidence level $\alpha$, the inequality describes the interior of a $p$-dimensional hyper-surface in $\vec{\theta}$ space. Thus, for $p = 2$, it describes a *confidence region* bounded by a closed contour in the parameter space $(\theta_1, \theta_2)$. An example of a set of such contours is depicted in Figure 12.14. As the confidence level increases, the corresponding contours describe larger and larger regions about the least-squares parameter vector $\vec{\hat{\theta}}$.

The numerical algorithm recommended here discretizes the interior of the confidence region using several contour levels that lie within the highest confidence contour; for example, suppose we wish to calculate the 90% confidence interval given the confidence regions depicted in Figure 12.14. We would evaluate the regression equation at a number of points, say, 20, along the entire 90% contour. Then we would do the same along the 80% contour, the 70% contour, and so on down to the 10% contour. With all of these regression function evaluations, we would then be able to compute the maximum and minimum of the regression function over the range of the input parameter $x$. This would provide reasonably good coverage of the 90% confidence interval of the regression function. If more precision was needed, one could choose more function evaluations along each contour and compute each contour in 1% increments of the confidence level.

For a three-dimensional regression parameter space, slices can be taken along one dimension of the resulting three-dimensional surface, and each slice can be discretized in the manner described for the two-dimensional case. Generalizing to $N$ dimensions, one may generate a recursive sequence of hypersurfaces of lower dimension until a series of

two-dimensional regions are obtained and evaluation over all of the two-dimensional regions gives the desired envelope of regression curves.

To determine the upper and lower confidence intervals associated with the regression equation, Eq. (12.36), we use the solution to Eq. (12.38), i.e., all $\vec{\theta}$ lying within (and on) the desired contour. The confidence intervals are determined by computing the envelope of regression curves resulting from *all* $\vec{\theta}$ lying within the confidence region. If we think of the solution to Eq. (12.38) as given by a set of discrete vectors of $\vec{\theta}$, then we can substitute this set of parameter vectors into the regression equation, Eq. (12.36). For each element in this set of $\vec{\theta}$s, we obtain a specific regression function. If we evaluate the ensemble of all regression functions by using all of the $\vec{\theta}$s, we can compute the maximum value of the regression function, $y^{+}_{\mathrm{CI}}(x)$, and the minimum value of the regression function, $y^{-}_{\mathrm{CI}}(x)$, over the range of $x$. As a result, $y^{+}_{\mathrm{CI}}(x)$ and $y^{-}_{\mathrm{CI}}(x)$ define the upper and lower bounds on the confidence intervals, respectively, over the range of $x$. These confidence intervals need not be symmetric as they were for the interpolation and linear regression cases discussed earlier. One may ask why the regression function must be evaluated over the entire confidence region. This must be done because the nonlinear regression function can have maxima and minima *anywhere* within the confidence region.

### *12.7.3 Global metrics*

If we would like to make a quantitative assessment of the global modeling error, the global metrics expressed in Eqs. (12.19) and (12.21) can still be used. However, the equations for the average relative confidence indicator and the maximum relative error metric, Eqs. (12.34) and (12.35), must be replaced because they are based on symmetric confidence intervals. Since we no longer have symmetric confidence intervals, we approximate these by computing the average half-width of the confidence interval over the range of the data and the half-width of the confidence interval at the maximum relative error, respectively. As a result, we now have

$$\left|\frac{\mathrm{CI}}{\bar{y}_{\mathrm{e}}}\right|_{\mathrm{avg}} = \frac{1}{(x_{\mathrm{u}} - x_{\mathrm{l}})} \int_{x_{\mathrm{l}}}^{x_{\mathrm{u}}} \left|\frac{y^{+}_{\mathrm{CI}}(x) - y^{-}_{\mathrm{CI}}(x)}{2\bar{y}_{\mathrm{e}}(x)}\right| \mathrm{d}x \tag{12.39}$$

for the average relative confidence indicator. $y^{+}_{\mathrm{CI}}(x)$ and $y^{-}_{\mathrm{CI}}(x)$ are the upper and lower confidence intervals, respectively, as a function of $x$. As stated earlier, $|\mathrm{CI}/\bar{y}_{\mathrm{e}}|_{\mathrm{avg}}$ provides a quantity with which to interpret the significance of the magnitude of $|\tilde{E}/\bar{y}_{\mathrm{e}}|_{\mathrm{avg}}$.

Also, we have

$$\left|\frac{\mathrm{CI}}{\bar{y}_{\mathrm{e}}}\right|_{\mathrm{max}} = \left|\frac{y^{+}_{\mathrm{CI}}(\hat{x}) - y^{-}_{\mathrm{CI}}(\hat{x})}{2\bar{y}_{\mathrm{e}}(\hat{x})}\right| \tag{12.40}$$

for the half-width of the confidence interval associated with the maximum relative error metric $\left|\tilde{E}/\bar{y}_{\mathrm{e}}\right|_{\mathrm{max}}$. The maximum relative error point $\hat{x}$ is defined as the $x$ value where

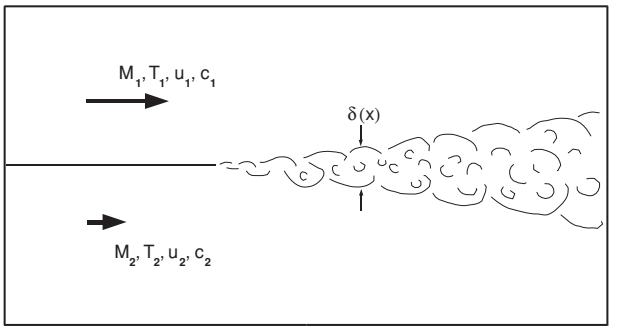


Figure 12.15 Flow configuration for the turbulent free shear layer (Oberkampf and Barone, 2004).

$|\tilde{E}/\bar{y}_e|$ achieves a maximum, that is,

$$\hat{x} = x \text{ such that } \left| \frac{y_m(x) - \bar{y}_e(x)}{\bar{y}_e(x)} \right| \text{ is a maximum for } x_l \leq x \leq x_u. \tag{12.41}$$

### *12.7.4 Example problem: compressible turbulent mixing*

The example chosen for the application of the nonlinear regression case is the prediction of compressibility effects on the growth rate of a turbulent free shear layer. An introduction to the problem is given, followed by a discussion of the available experimental data. Details of the model and verification of the numerical solutions are briefly described along with the validation metric results. Further details on this example can be found in Barone *et al.* (2006) and Oberkampf and Barone (2006).

#### *12.7.4.1 Problem description*

The planar free shear layer is a canonical turbulent flow and a good candidate for use in a unit-level validation study. Figure 12.15 shows the flowfield configuration in which a thin splitter plate separates two uniform streams (numbered 1 and 2) with different flow velocities and temperatures, but both at the same pressure. The two streams mix downstream of the splitter plate's trailing edge, forming the free shear layer within which momentum and energy are diffused. For a high-Reynolds-number flow, the boundary layers on both sides of the plate and the free shear layer are turbulent. In the absence of any applied pressure gradients or other external influences, the flowfield downstream of the trailing edge consists of a shear layer development region near the edge, followed by a similarity region. Within the development region, the shear layer adjusts from its initial velocity and temperature profiles inherited from the plate boundary layers. Further downstream in the similarity region, the shear layer thickness, $\delta(x)$, grows linearly with streamwise distance $x$, resulting in a constant value of $d\delta/dx$.

Of particular interest in high-speed vehicle applications is the behavior of the shear layer as the Mach number of one or both streams is increased. A widely accepted parameter correlating the shear layer growth rate with compressibility effects is the convective Mach number, $M_c$, for mixing two streams of the same gas (Bogdanoff, 1983). $M_c$ is defined as

$$M_c = \frac{u_1 - u_2}{c_1 + c_2}, \tag{12.42}$$

where $u$ is the fluid velocity and $c$ is the speed of sound. It has been found experimentally that an increase in the convective Mach number leads to a decrease in the shear layer growth rate for fixed velocity and temperature ratios of the streams. This is usually characterized by the compressibility factor $\Phi$, which is defined as the ratio of the compressible growth rate to the incompressible growth rate at the same velocity and temperature ratios:

$$\Phi = \frac{(\mathrm{d}\delta/\mathrm{d}x)_c}{(\mathrm{d}\delta/\mathrm{d}x)_i}. \tag{12.43}$$

#### *12.7.4.2 Experimental data*

Experimental data on high-speed shear layers are available from a number of independent sources. The total collection of experimental investigations employs a wide range of diagnostic techniques within many different facilities. Comparisons of data obtained over a range of convective Mach numbers from various experiments indicate significant scatter in the data. Recently, Barone *et al.* (2006) carefully re-examined the available data and produced a recommended data set that exhibits smaller scatter in the measurements.

The resulting ensemble of data from Bogdanoff (1983); Chinzei *et al.* (1986); Papamoschou and Roshko (1988); Dutton *et al.* (1990); Elliot and Samimy (1990); Samimy and Elliott (1990); Goebel and Dutton (1991); Debisschop and Bonnet (1993); Gruber *et al.* (1993); Debisschop *et al.* (1994); and Barre *et al.* (1997) is presented in Figure 12.16. The data are organized into groups of sources, some of which are themselves compilations of results from several experiments.

#### *12.7.4.3 Mathematical model*

The Favre-averaged compressible Navier–Stokes equations were solved using the standard $k - \varepsilon$ turbulence model (Wilcox, 2006). The low-Reynolds number modification to the $k - \varepsilon$ model of Nagano and Hishida (1987) was applied near the splitter plate. Most turbulence models in their original form do not correctly predict the significant decrease in shear layer growth rate with increasing convective Mach number, necessitating inclusion of a compressibility correction. Several compressibility corrections, derived from a variety of physical arguments, are widely used in contemporary computational fluid dynamics (CFD) codes. In this study, the dilatation-dissipation compressibility correction of Zeman (1990) was used.

The solutions were computed using the Sandia Advanced Code for Compressible Aerothermodynamics Research and Analysis (SACCARA) (Wong *et al.*, 1995a,b), which

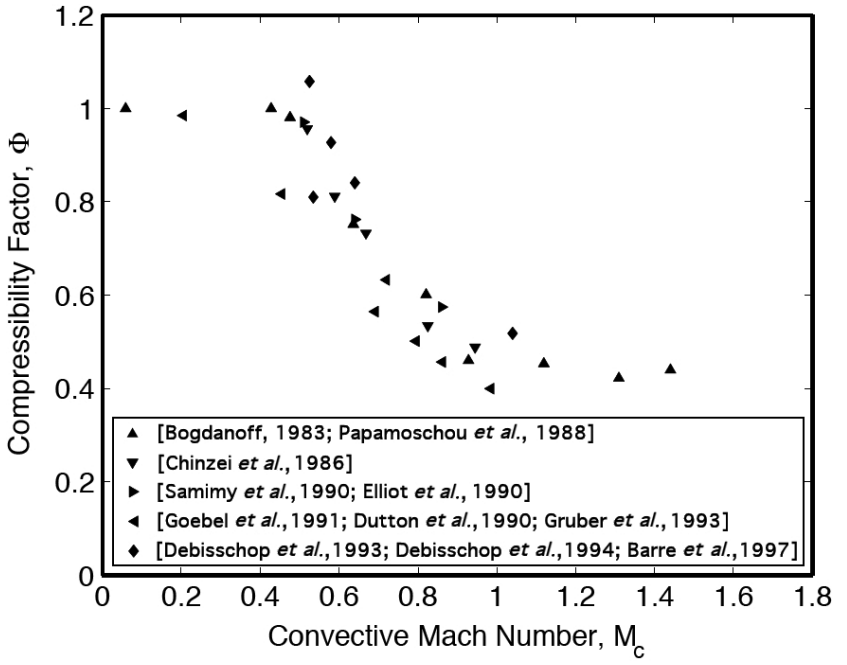


Figure 12.16 Experimental data for compressibility factor versus convective Mach number (Oberkampf and Barone, 2004).

employs a block-structured, finite volume discretization method. The numerical fluxes are constructed with the symmetric TVD scheme of Yee (1987), which gives a second-order convergence rate in smooth flow regions. The equations are advanced to a steady state using the LU-SGS scheme of Yoon and Jameson (1987). Solutions were considered iteratively converged when the $L_2$ norm of the momentum equation residuals had decreased by eight orders of magnitude. Numerical solutions were obtained over the convective Mach number range of the experimental data, from 0.1 to 1.5, in increments of 0.14.

For each convective Mach number, solutions were calculated on three meshes: coarse, medium, and fine. The meshes are uniform in the streamwise, or $x$, direction, and stretched in the cross-stream, or $y$, direction, so that mesh cells are clustered within the shear layer. The cells are highly clustered in the $y$ direction near the trailing edge and become less clustered with increasing $x$ to account for the shear layer growth. Richardson's extrapolation (Roache, 1998) was used to estimate the discretization error on $\mathrm{d}\delta/\mathrm{d}x$. The maximum error in the fine-mesh solution was estimated to be about 1% at $M_c = 0.1$ and about 0.1% at $M_c = 1.5$.

We defined $\delta$ using the velocity layer thickness definition. As mentioned previously, the thickness grows linearly with $x$ only for large $x$ due to the presence of the development region, which precedes the similarity region. Given that the growth rate approaches a constant value asymptotically, the thickness as a function of $x$ is fitted with a curve that

mimics this asymptotic character. The function used for the fit is

$$\delta(x) = \beta_0 + \beta_1 x + \beta_2 x^{-1}. \tag{12.44}$$

The coefficient $\beta_1$ is the fully developed shear layer growth rate as $x$ becomes large.

Following extraction of the compressible growth rate, $(\mathrm{d}\delta/\mathrm{d}x)_\mathrm{c}$, the incompressible growth rate, $(\mathrm{d}\delta/\mathrm{d}x)_\mathrm{i}$, must be evaluated at the same velocity and temperature ratio. Incompressible or nearly incompressible results are difficult to obtain with a compressible CFD code. Therefore, the incompressible growth rate was obtained by computing a similarity solution for the given turbulence model and flow conditions. The similarity solution is derived by Wilcox (2006) in his turbulence modeling text and implemented in the MIXER code, which is distributed with the text. The similarity solution is computed using the same turbulence model as the Navier–Stokes calculations, but under the assumptions that (a) the effects of laminar viscosity are negligible, and (b) there exists a zero pressure gradient.

#### *12.7.4.4 Validation metric results*

The quantities $\delta$ and $\mathrm{d}\delta/\mathrm{d}x$ are post-processed from the finite-volume computational solution and the MIXER code, but the SRQ of interest for the validation metric is the compressibility factor $\Phi$. Before the validation metric result can be computed, we must prescribe a form for the nonlinear regression function to represent the experimental data in Figure 12.16. It is important that the proper functional behavior of the data, established through theoretical derivation or experimental measurement, be reflected in the form of the regression function. For the compressible shear layer, we know that $\Phi$ must equal unity, by definition, in the incompressible limit $M_\mathrm{c} \to 0$. Experimental observations and physical arguments also suggest that $\Phi \to$ constant as $M_\mathrm{c}$ becomes large. These considerations lead to the following choice of the regression function, taken from Paciorri and Sabetta (2003):

$$\Phi = 1 + \hat{\theta}_1 \left( \frac{1}{1 + \hat{\theta}_2 M_\mathrm{c}^{\hat{\theta}_3}} - 1 \right). \tag{12.45}$$

Using Eq. (12.45) and the experimental data shown in Figure 12.16, we used the MATLAB (MathWorks, 2005) function *nlinfit* from the Statistics Toolbox, to calculate the following regression coefficients:

$$\hat{\theta}_1 = 0.5537, \qquad \hat{\theta}_2 = 31.79, \qquad \hat{\theta}_3 = 8.426. \tag{12.46}$$

We now compute the 90% confidence interval of the regression function in Eq. (12.45), with the $\vec{\hat{\theta}}$ values given in Eq. (12.46) and the inequality constraint given by Eq. (12.38). We use the method outlined above to compute the 90% confidence region in the three-dimensional space described by $\theta_1, \theta_2$, and $\theta_3$. The resulting confidence region, pictured in Figure 12.17, resembles a curved and flattened ellipsoid, especially for small values of $\theta_2$. The elongated shape in the $\theta_2$ direction indicates the low sensitivity of the curve fit to $\theta_2$ relative to the other two regression parameters. Evaluation of the regression function Eq. (12.45) for all $\vec{\theta}$ lying within the 90% confidence region yields the desired simultaneous confidence intervals.

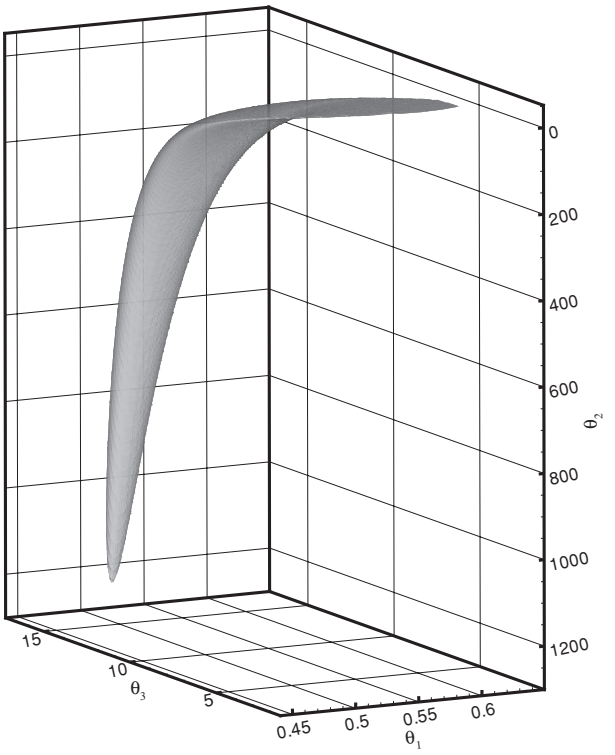


Figure 12.17 Three-dimensional 90% confidence region for the regression fit to the shear layer experimental data (Oberkampf and Barone, 2006).

Figure 12.18 shows the final result of the analysis in graphical form: a plot of the experimental data along with the regression fit, the 90% confidence intervals, and the computational simulation result. Concerning the error assessment of the $k - \varepsilon$ model, it is seen that the Zeman compressibility correction predicts a nearly linear dependence of the compressibility factor on $M_c$ over the range $0.2 \leq M_c \leq 1.35$. One could claim that the trend is correct, i.e., the Zeman model predicts a significant decrease in the turbulent mixing as the convective Mach number increases; however, the Zeman model does not predict the nonlinear dependency on $M_c$. We did not compute any simulation results for $M_c > 1.5$ and, as a result, did not determine the asymptotic value of $\Phi$ for the Zeman compressibility correction. However, the solutions for $M_c = 1.36$ and $M_c = 1.50$ suggest that the asymptotic value is near $\Phi = 0.49$.

By noting the large width of the confidence intervals for large $M_c$ in Figure 12.18, it is seen that the largest uncertainty in the experimental data occurs in this region. The parameter $\theta_1$ in the regression function, Eq. (12.45), has the main influence on the size of the confidence intervals for large $M_c$. From the viewpoint of the design of needed validation experiments, one can conclude that future experiments should be conducted at higher convective Mach numbers to better determine the asymptotic value of $\Phi$.

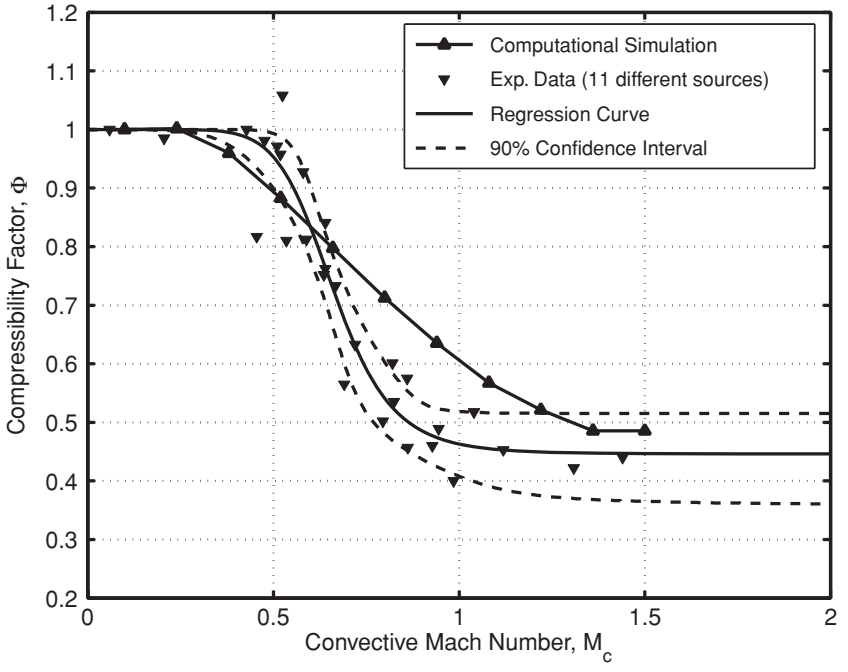


Figure 12.18 Comparison of the simulation result with the experimental data, nonlinear regression curve, and 90% simultaneous confidence interval (Oberkampf and Barone, 2006).

The estimated error, $\tilde{E}(x)$, of the model as a function of $M_c$ is plotted in Figure 12.19 along with the 90% confidence interval from the experimental data. This plot presents the validation metric result, i.e., the difference between computation and the regression fit of the experimental data, along with the 90% confidence interval representing the uncertainty in the experimental data. As pointed out previously in the helium plume example, the validation metric critically examines both the model and the experimental data. With this plot, it is seen that there is a slight under-prediction of turbulent mixing in the range $0.3 \leq M_c \leq 0.6$ and a significant over-prediction of turbulent mixing in the range $0.7 \leq M_c \leq 1.3$. Examining an error plot such as this, one could conclude that the Zeman model does not capture the nonlinear trend of decreasing turbulent mixing with increasing convective Mach number. Whether the model accuracy is adequate for the requirements of the intended application is, of course, a completely separate issue.

Note that in Figure 12.19 the confidence intervals are not symmetric with respect to zero. In the case of nonlinear regression, Eq. (12.36), the nonlinear function need not possess any symmetry properties with respect to the regression parameters. Therefore, evaluation of the nonlinear function over the set of $\vec{\theta}$ satisfying Eq. (12.38) results in asymmetric confidence intervals over the range of the input parameter. For the shear layer example, Eq. (12.45) is evaluated over the volume of regression coefficients shown in Figure 12.17.

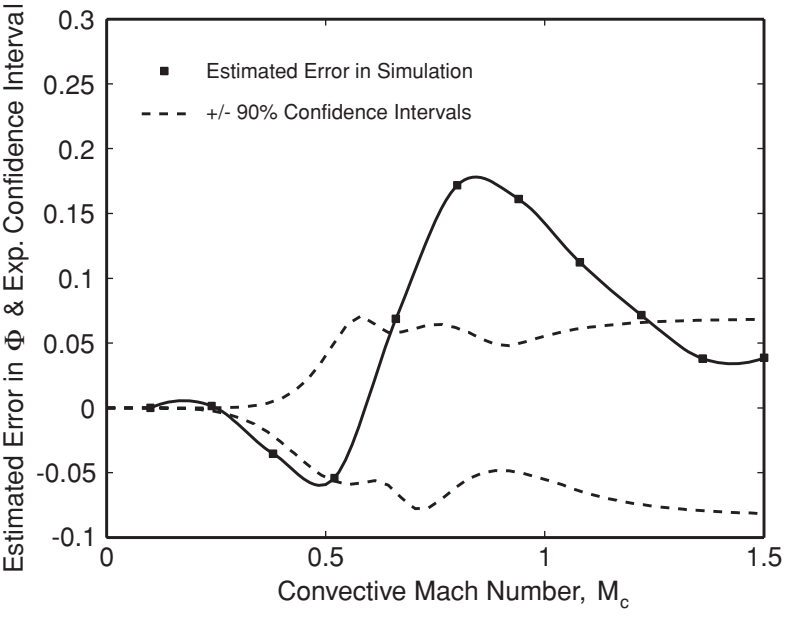


Figure 12.19 Validation metric result and 90% confidence interval for $\Phi$ (Oberkampf and Barone, 2006).

Using Eqs. (12.19), (12.21), (12.39), and (12.40), the global metric results for the $k - \varepsilon$ model with the Zeman compressibility correction over the range $0 \leq M_c \leq 1.5$ are as follows:

average relative error $= 13\% \pm 9\%$ with 90% confidence,
maximum relative error $= 35\% \pm 10\%$ with 90% confidence.

The average error of 13% seems reasonable, given Figure 12.18. As in the helium plume and the foam decomposition examples, we see a maximum error that is noticeably larger than the average error, i.e., roughly a factor of three. From Figure 12.19 it can be found that the maximum absolute error occurs at $M_c = 0.83$. The maximum relative error, however, occurs at $M_c = 0.88$. At this value of $M_c$, one determines that the 90% confidence interval is $\pm 10\%$.

### *12.7.5 Observations on the present approach*

The validation metrics derived here are relatively easy to compute and interpret in practical engineering applications. When nonlinear regression functions are required for the metric, the nonlinear regression function requires a software package, such as Mathematica or MATLAB, to perform the computations. A software package called VALMET has recently been completed that computes the validation metrics described here for both the interpolation and regression case (Iuzzolino *et al.*, 2007). Experimental and computational data

can be provided to VALMET in the form of Excel spreadsheets or text files. Either linear or spline interpolation can be used for both the experimental and computational data. For regression, fifteen different functional forms can be chosen, or user chosen functions can be programmed. The program can be run on a computer with an installed version of MATLAB, or as a stand-alone compiled code on a computer using Microsoft Windows® 2000/XP.

The interpretation of the present metrics in engineering decision making should be clear and understandable to a wide variety of technical staff (analysts, model builders, and experimentalists) and management. The metric result has the following form: estimated error of the model $\pm$ an interval that represents experimental uncertainty with specified confidence. The present metric only measures the mismatch between computational and experimental mean response of the system. The present metrics can be used to compare the modeling accuracy of different competing models, or they can help to assess the adequacy of the given model for an application of interest. It has been stressed that the manner in which the result of a validation metric relates to an application of interest is a separate and more complex issue, especially if there is significant extrapolation of the model.

The validation metrics presented here should apply to a wide variety of physical systems in engineering and science. If the SRQ is a complex time-varying quantity, then it may be possible to time average the quantity so that the present approach could be used. If the SRQ were a complex time series, such as modes in structural dynamics, then the present metrics would not be appropriate. If the response is mapped to the frequency domain, it may be possible to apply the method to the amplitudes and frequencies of the lower modes. In addition, the present metrics apply to single SRQs that are a function of a single input or control quantity. Extension of the method to multivariate analysis would require the incorporation of the correlation structure between the variables.

## 12.8 Validation metric for comparing p-boxes

When computational predictions are probability distributions, there is a great deal more information contained in how a system responds than compared to a simple deterministic response of a system. One can think of the model as a mapping of all uncertain inputs to produce a set of uncertain system responses. In this mapping, all of the uncertain inputs are convolved according to the physical processes described by the PDEs of interest. A validation metric could be viewed as asking the question: how well does the model map the inputs to the outputs, compared to the way that nature maps these?

The simplest method of comparing experimental data and predictions is in terms of their means, variances, covariances, and other distributional characteristics. The main limitation of approaches based on comparing summary statistics is that it considers only the central tendencies or other specific behaviors of data and predictions and not their entire distributions. When predictions are distributions, they contain a considerable amount of detail and it is not always easy to know what statistics are important for a particular application. While some statistical tests are certainly helpful in comparing experimental data and predictions,

most do not directly address the validation metric perspective of interest here. In addition, traditional statistical tests, as well as the Bayesian approach to validation, do not address the issue of epistemic uncertainty in either, or both, the experimental data and the prediction. If epistemic uncertainty exists in either, we will consider the representation to be given by a p-box.

In the following section, we introduce the notion of comparing imprecise probabilistic quantities, including p-boxes; describe some of the desirable properties that a validation metric should have; and suggest a particular approach that has these properties. Several simple examples are given which display some of the features of this validation metric. This section is taken from Oberkampf and Ferson (2007); Ferson *et al.* (2008); and Ferson and Oberkampf (2009).

### 12.8.1 Traditional methods for comparing distributions

There are a variety of standard ways to compare random variables in probability theory. If random numbers $X$ and $Y$ always have the same value, the random variables are said to be "equal," or sometimes "surely equal." A much weaker notion of equality is useful in the construction of validation metrics because we are interested in comparing distributions, i.e., functions as opposed to numbers. If we can only say that the expectation, i.e., the mean, of the absolute values of the differences between $X$ and $Y$ is zero, the random variables are said to be "equal in mean." If $X$ and $Y$ are not quite equal in mean, we can measure their mismatch by defining the mean metric, $d_{\mathrm{E}}$, as

$$d_{\mathrm{E}}(X, Y) = E(|X - Y|) \neq |E(X) - E(Y)|, \tag{12.47}$$

where E denotes the expectation operator. Note that this difference is not the same as the absolute value of the difference between the means. The idea can be generalized to higher-order moments, and equality in a higher-order moment implies equality in all lower-order moments.

The notion of equality for randomly varying quantities can be loosened further still by comparing only the *shapes* of the probability distributions of the random variables. Random variables whose distributions are identical are said to be "equal in distribution." This is often denoted as $X \sim Y$, or sometimes by $X = {}^{\mathrm{d}}Y$. This is really a rather loose kind of equality, because it does not require the individual values of $X$ and $Y$ to ever be equal, or even to ever be close. For instance, suppose $X$ is normally distributed with mean zero and unit variance. If we let $Y = -X$, then $X$ and $Y$ are obviously equal in distribution, but are about as far from equality as can be imagined. Nevertheless, equality in distribution is an important concept because a distribution often represents all that is known about the values of a random variable.

If the distributions are not quite identical in shape, the discrepancy can be measured with many possible measures that have been proposed for various purposes. For instance, a very common such measure is the maximal probability, i.e., vertical, difference between the two

cumulative distribution functions

$$d_S(X, Y) = \sup_z |\Pr(X \leq z) - \Pr(Y \leq z)| . \tag{12.48}$$

$d_S$ is the *Smirnov metric*, and $\sup_z$ is the supremum over the sample space of $X$ and $Y$. $d_S$ defines the Kolmogorov–Smirnov statistical test for comparing distributions (D'Agostino and Stephens, 1986; Huber-Carol *et al.*, 2002; Mielke and Berry, 2007). One of the properties of the Smirnov distance is that it is symmetric, which is to say that $d_S(X, Y)$ always equals $d_S(Y, X)$. The symmetry might be considered unnecessary or even counterintuitive as a feature for validation. We do not view predictions and observations as exchangeable with each other; in a validation metric it matters which is which. Suppose, for instance, that we inadvertently switched the predicted distribution with the experimental data distribution. One might expect to obtain a different result from having made such a mistake, but the Smirnov distance does not change whether the prediction and data are exchanged or not.

The *Kullback–Leibler divergence* is another very widely used measure of the discrepancy between distributions that is not symmetric (D'Agostino and Stephens, 1986; Huber-Carol *et al.*, 2002; Mielke and Berry, 2007). It is defined, in its discrete formulation, for a probability mass function $p$ for $X$ and a probability mass function $q$ for $Y$ as

$$\sum_z p(z) \log_2 \frac{p(z)}{q(z)}, \tag{12.49}$$

where $z$ takes on all values in the common range of $X$ and $Y$. The $p$ distribution summarizes the observations and the $q$ distribution summarizes the model prediction. The continuous formulation is similar except that the summation is replaced by integration. The term *divergence* in this metric may be misleading because this quantity has nothing to do with notions of divergence familiar from calculus as the inner product of partial derivatives or the flux per unit volume. Instead, the term is used here in its other meaning as deviation from a standard. The Kullback–Leibler divergence is commonly used in information theory and also in physics. It is interpreted as the relative entropy between $p$ and $q$, i.e., the entropy of the distribution $p$ with respect to the distribution $q$.

As we have mentioned, there are, in fact, many other measures that could be used to compare data and prediction distributions. But, given the broad acceptance and ubiquity of the Smirnov and Kullback–Leibler measures in probability and physics, it is perhaps necessary to explain why we simply don't use one of these as our validation metric. This question is addressed in the next section.

### *12.8.2 Method for comparing p-boxes*

#### *12.8.2.1 Discussion of p-boxes*

A validation metric was recently proposed by Oberkampf and Ferson (2007); Ferson *et al.* (2008); and Ferson and Oberkampf (2009) to measure the mismatch between a prediction and empirical observations. The prediction and the experimental measurements can both

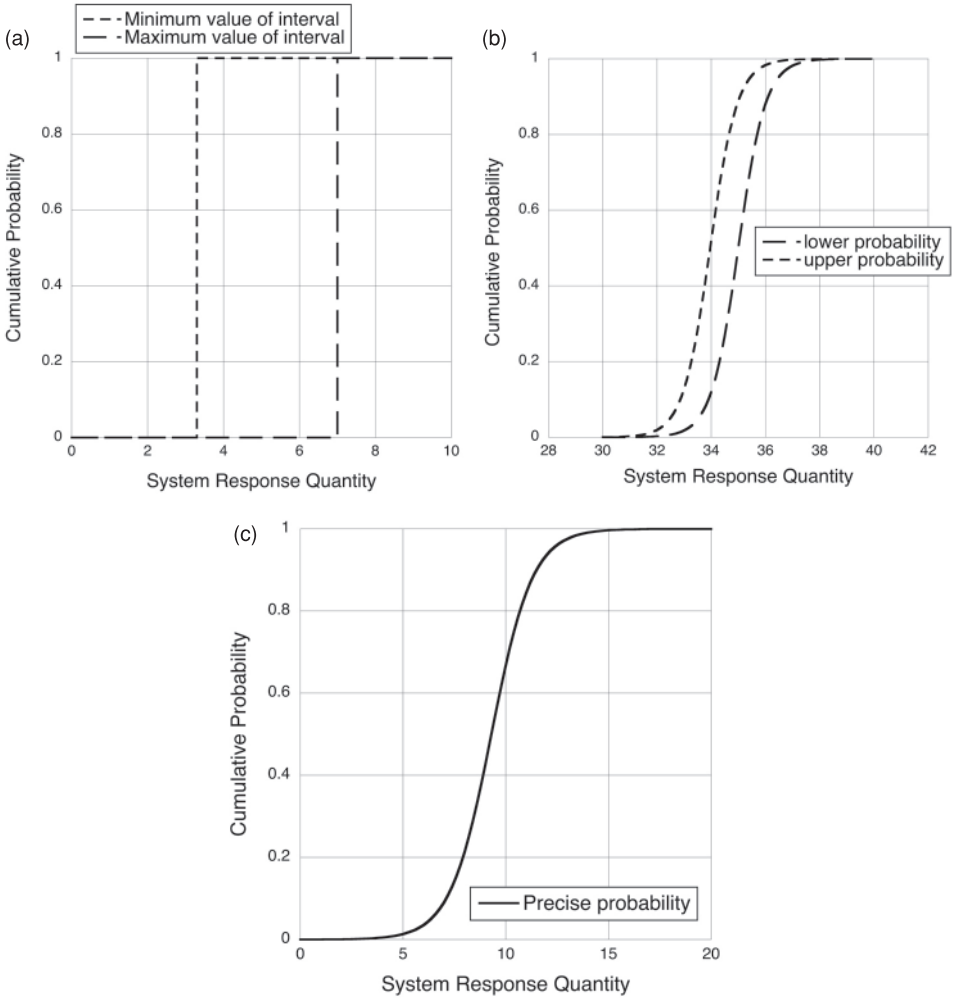


Figure 12.20 p-boxes for varying degrees of aleatory and epistemic uncertainty.
(a) p-box for purely epistemic uncertainty.
(b) p-box for a mixture of aleatory and epistemic uncertainty.
(c) Degenerate p-box for purely aleatory uncertainty.

be given by a p-box. A p-box expresses both epistemic and aleatory uncertainty in a way that does not confound the two. For a more complete discussion of p-boxes, see Ferson (2002); Ferson *et al.* (2003, 2004); Kriegler and Held (2005); Aughenbaugh and Paredis (2006); and Baudrit and Dubois (2006).

Figure 12.20 shows examples of three p-boxes with varying degrees of aleatory and epistemic uncertainty for a predicted SRQ. Figure 12.20a shows a p-box for an interval, i.e., there is no aleatory uncertainty and there is purely epistemic uncertainty. For values of the SRQ less than the minimum value of the interval, the cumulative probability is zero.

That is, there is certainty that all possible values of the SRQ are greater than the minimum value of the interval. For values of the SRQ over the range of the interval, the cumulative probability is in the interval [0, 1]. This interval-valued probability is, of course, a nearly vacuous statement in the sense that the SRQ probability could be *anything* over the range of zero to unity. For values of the SRQ greater than the maximum value of the interval, the cumulative probability is unity. That is, there is certainty that all possible values of the SRQ are less than or equal to the maximum value of the interval.

Figure 12.20b shows a p-box for a mixture of aleatory and epistemic uncertainty in the SRQ. The p-box shows the fraction, expressed as interval-valued quantity, of the sampled population that would have a value less than, or equal to, a particular value of the SRQ. Some examples are (a) for an SRQ = 32, the fraction of the population that would have a value of 32 or less would be in the range [0.0, 0.02], and (b) for an SRQ = 34, the fraction of the population that would have a value of 34 or less would be in the range [0.1, 0.5]. The interval-valued fraction is due to lack of knowledge from any source in the simulation. As is typical of a mixture of aleatory and epistemic uncertainty, the range of the interval-valued probability approaches zero for rare events of small values of the SRQ, and rare events of large values of the SRQ. It should also be noted that the horizontal breadth of the p-box expresses the magnitude of the epistemic uncertainty in terms of the SRQ, and the slope of the p-box expresses the magnitude of the aleatory uncertainty.

Figure 12.20c shows a degenerate p-box, i.e., a precise CDF, for the SRQ. The p-box is degenerate in the sense that the epistemic uncertainty is very small, or zero, compared to the aleatory uncertainty. Precise probabilities are, of course, the foundation of traditional probability theory and its application. p-boxes are also referred to as imprecise probabilities because the probability is not necessarily a unique quantity, but can be interval valued. During the last few decades, extensions of traditional probability theory have identified a number of types of imprecise probability; i.e., more complex than the p-box discussed here. For a discussion of these extensions, some recent texts are Walley (1991); Dubois and Prade (2000); Nguyen and Walker (2000); Molchanov (2005); and Klir (2006).

#### *12.8.2.2 Validation metric for p-boxes*

Any nondeterministic prediction from the model can always be characterized as a cumulative distribution function $F(x)$. In this notation, $x$ is the predicted variable, i.e., the SRQ. The observation(s), on the other hand, are usually provided as a collection of point values in a data set. The distribution function for a data set, which is referred to as an empirical distribution function (EDF), summarizes the data set as a function suitable for graphical representation. It is a function that maps $x$ to the probability scale on the interval [0, 1]. It is constructed as a non-decreasing step function with a constant vertical step size of $1/n$, where $n$ is the sample size of the data set. The locations of the steps correspond to the values of the data points. Such a distribution for data $x_i$, $i = 1, \ldots, n$, is

$$S_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i, x), \tag{12.50}$$

where

$$I(x_i, x) = \begin{cases} 1, & x_i \leq x, \\ 0, & x_i > x. \end{cases} \tag{12.51}$$

$S_n(x)$ is simply the fraction of data values in the data set that are at or below each value of $x$.

An EDF is an advantageous representation, as opposed to a continuous CDF, because it is an exact representation of the distribution of the data, regardless of the amount of data. In addition, an EDF does not require any assumptions to represent the data, for example, as is required to construct a histogram of the data set. An EDF preserves the statistical information in the data set about its central tendency or location, its dispersion or scatter, and, in fact, all other statistical features of the distribution. The only information in the original data set that is not in the distribution is the order in which the values were originally given, which is meaningless whenever the data were sampled at random. When the data set consists of a single value, then the $S_n$ function is a simple spike at the location along the $x$-axis given by that value; that is, it is zero for all $x$ less than that value and one for all $x$ larger than that value. For graphical clarity, however, it becomes convenient not to depict these flat portions at zero and one when the functions are plotted.

We propose the use of the Minkowski $L_1$ metric as a validation metric and we will refer to it as the *area metric*. It is defined as the *area* between the prediction distribution $F$ and the data distribution $S_n$ as the measure of the mismatch between them. Mathematically, the area between the curves is the integral of the absolute value of the difference between the functions

$$d(F, S_n) = \int_{-\infty}^{\infty} |F(x) - S_n(x)| \, \mathrm{d}x. \tag{12.52}$$

It is clear geometrically that this quantity is also equal to the average horizontal difference between the two functions, $\int |F^{-1}(p) - S_n(p)^{-1}| \mathrm{d}p$, but this is not the same as the average of the absolute differences between *random values* from these two distributions. (Such an average would not be zero if the distributions were coincident.) The area metric is thus a function of the shapes of the distributions, but is not readily interpretable as a function of the underlying random variables. The area measures the *disagreement* between theory and empirical evidence, as opposed to measuring the agreement. It is a metric so long as the integral exists.

It should be stressed that since this mismatch measure is always positive, it should *not* be interpreted in the same way as the common definition for error,

$$\varepsilon = y_{\text{obtained}} - y_{\text{T}}, \tag{12.53}$$

where $\varepsilon$ is the error in the quantity $y_{\text{obtained}}$ and $y_{\text{T}}$ is the true value of $y$. For example, we contrast the mismatch measure defined in Eq. (12.52) with the error measure associated with the confidence interval approach, Eq. (12.8). In the confidence interval approach,

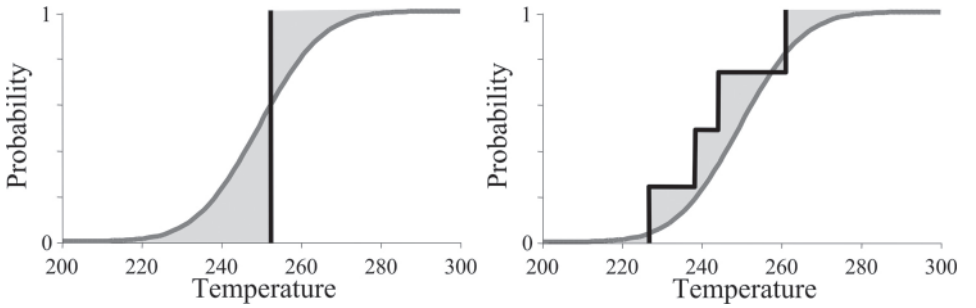


Figure 12.21 Example data sets, with $n = 1$ on the left and $n = 4$ on the right, shown as $S_n$ distributions (black) against a prediction distribution (smooth). The validation metric is the areas (shaded) between the prediction distribution and the two data sets (Ferson *et al.*, 2008).

$\tilde{E} > 0$ meant the model prediction was greater than the experimental measurement. For the area metric there is no such indication of how the model differs from the measurement.

Figure 12.21 illustrates this area measure of mismatch for two data sets against a prediction distribution of temperature, which represents an arbitrary SRQ. The prediction distribution, shown as the smooth curve, is the same in both graphs. This prediction distribution might be obtained by propagating input uncertainties through a mathematical model by using a large number of samples from a Monte Carlo simulation. Superimposed on these graphs are distribution functions $S_n$ for two hypothetical data sets. On the left graph, the data set consists of the single value 252 °C, and on the right, the data set consists of the values {226, 238, 244, 261}. In complex engineering systems it is not uncommon to have only one experimental test of the complete system. In such cases, the empirical distributions are not complex step functions, but instead single-step spikes representing point values (i.e., degenerate distributions). Note that the empirical distribution function is zero for all values smaller than the minimum of the data and unity for all values larger than the maximum of the data. Likewise, beyond the range (support) of the prediction distribution, the value of $F(x)$ is either zero or one extending to infinity in both directions. For graphical clarity, however, these flat portions at probability zero or unity are not depicted when the distributions are plotted.

The areas measuring the mismatches between the prediction and the two data sets in Figure 12.21 are shaded. In the left graph, the area consists of a region to the left of the datum at 252, and a region to right of it. In the right graph, there are four shaded regions composing the total area between the prediction distribution and the data distribution. The area metric is a generalization of the deterministic comparisons between scalar values that have no uncertainty. That is, if the prediction and the observation are both scalar point values, the area is simply equal to their difference. It is clear that the area metric reflects the difference between the full distribution of both the prediction and the observations. However, the area will tend not to be sensitive to minor discrepancies in the distribution tails because there is little probability mass in the tails of the distributions.

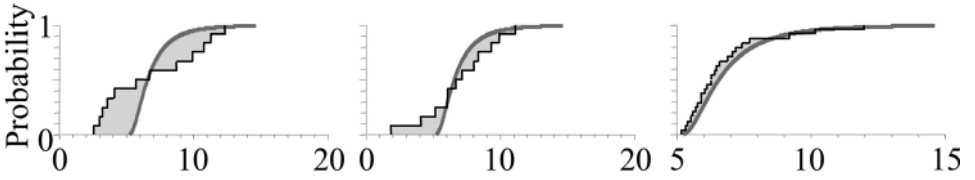


Figure 12.22 Examples of mismatch between a prediction distribution (smooth) and different empirical data sets (steps) (Ferson *et al.*, 2008).

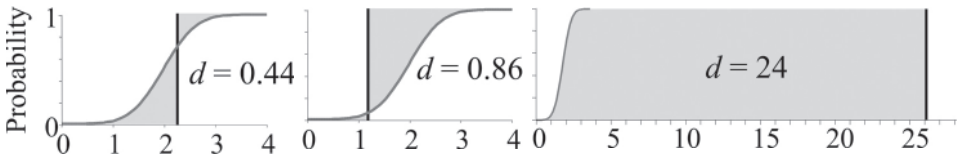


Figure 12.23 Comparisons of a prediction distribution (smooth) with three different data points (spikes) (Ferson *et al.*, 2008).

Figure 12.22 shows how the area metric differs from a validation metric based on merely matching in the mean or matching in both the mean and variance. In each of three cases, the prediction distribution is shown as a smooth curve. It is the same in all three graphs, although the scale in the third graph is a bit different from the other two. The step functions represent three different data sets as empirical distribution functions $S_n$. In the leftmost graph, the prediction distribution and the observed data have the same mean. But, otherwise, the data look rather different from the prediction; the data appear to be mostly in two clusters on either side of the mean. Indeed, so long as the average of the data balances at that mean, those data clusters could be arbitrarily far away from each other. Any validation measure based only on the mean would not detect any discrepancy between the theory and the data, even though the data might bear utterly no resemblance to the prediction, apart from their matching in the mean. In the middle graph, both the mean and the variance of the observed data and the theoretical prediction match. However, one would not claim the comparison between prediction and data were extremely good because of how the prediction deviates from the left tail of the empirical distribution. Smaller values are more prevalent in the real data than were predicted. In the third graph, the agreement between the prediction and the data is good overall. This is reflected in the smallness of the area between the prediction distribution and the data distribution. The only way for the area to be small is for the two distributions to match closely over the entire range of each. In each of these cases, the overall mismatch can be measured by the area between the two curves. It measures disagreements that the lower-order moments like the mean and variance cannot address.

Figure 12.23 shows how the area metric $d$ varies with different values for a single datum matched against a prediction distribution. In these three examples, the prediction distribution

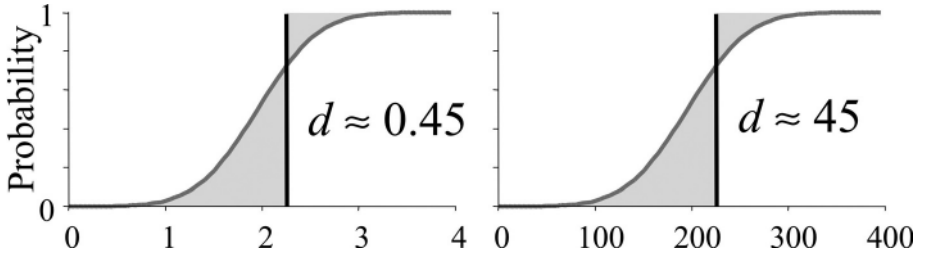


Figure 12.24 Further examples showing the metric's dependence on the scale (Ferson *et al.*, 2008).

is the same, and it is centered at 2 and ranges between 0 and 4. The most important thing to notice is that a single value can never perfectly match the entire distribution, unless that distribution is itself a degenerate point value. The first and second graphs of the figure compare the prediction distribution to a single observation at 2.25 and 1.18, respectively, yielding corresponding values for the area metric of 0.44 and 0.86. About the best possible match that a single datum could have occurs when the datum is located at the distribution's median, but, even there, the area metric will often be significant. In the case of the prediction distribution depicted in Figure 12.23, the area metric will be smallest when the observation is 2, which yields a value of 0.4 for the metric.

If, for example, the prediction distribution in Figure 12.23 were a uniform probability density function over the range $[a, b]$, a single observation can't be any "closer" to it than $(b - a)/4$, which is the value of the area metric if the point is at the median. That's the best match possible with a single data point. Stated differently, the mismatch between theory and experiment *could* be much better if more experimental measurements were available, but, due to sampling error in the experiment, it is the smallest mismatch that can exist. How *bad* could the match be? The match could be very bad; indeed, it can be bad to an arbitrarily large degree. The rightmost graph in Figure 12.23 shows another example of a single datum compared to the same prediction distribution. In this case, the data point is at 26, which means that it is about 24 units away from the distribution. The area metric can be arbitrarily large, and it reduces to the simple difference between the datum and the prediction when both are point values. Because probability is dimensionless, the units of the area are always the same as the units of the abscissa.

The area metric depends on the scale in which the prediction distribution and data are expressed. The two graphs in Figure 12.24 depict a pair of comparisons in which the corresponding shapes are identical but the scales are different, as though the left graph were expressed in meters and the right graph in centimeters. Although the shapes are the same, the area metric is different by 100 fold. It would, of course, be possible to normalize the area measure, perhaps by dividing it by the standard deviation of the prediction distribution, but we do not believe this would be a good idea because the result would no longer be expressed in the physical units of the abscissa. Such normalization would destroy the *physical meaning* of the metric.

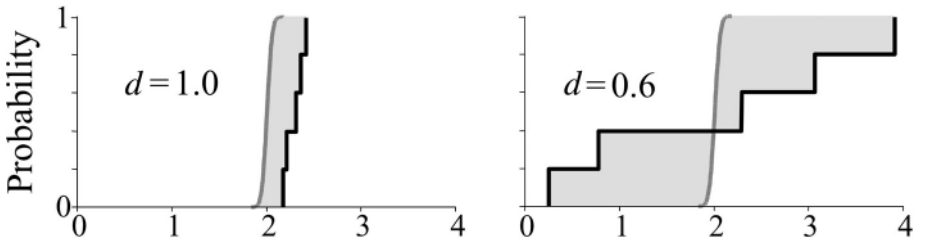


Figure 12.25 Why physical units are important for a validation metric (Ferson *et al.*, 2008).

Figure 12.25 illustrates why retaining the scale and physical units (degrees, meters, newtons, newtons/m$^2$, etc.) of the data is important for the intuitive appeal of a validation metric. The two graphs are drawn with the same $x$-axis, and they depict the same prediction distribution as a gray curve concentrated around 2. Two data sets are summarized as the $S_n$ distributions shown as black step functions on the graphs. From a statistical perspective it might be argued that the comparison in the right graph reveals a better match between the theory and the data than the comparison in the left graph. In the left graph, the two distributions do not even overlap, whereas in the right graph the distributions at least overlap and are similar in their means. Using a traditional Kolmogorov–Smirnov test for differences between the two distributions, one would find statistically significant evidence that the distributions in the left graph are different ($d_S = 1.0$, $n = 5$, $p < 0.05$), but would *fail* to find such evidence for the distributions in the right graph ($d_S = 0.6$, $n = 5$, $p > 0.05$). But this is not at all how engineers and analysts would understand these two comparisons. For engineers and scientists, the main focus is on the difference between the two distributions, in units along the $x$-axis. In this sense, the comparison on the left is a much better match between theory and data than the comparison on the right. Engineers and scientists have a strong intuition that the data–theory comparison on the left might be attributed to a small bias error in the theory or the data, but the theory does a good job of capturing the variability of the physics. The discrepancy on the left is never larger than half a unit along the $x$-axis, whereas the discrepancy on the right could be larger than two units. It's this physical distance measuring the theory–data disagreement that really matters to engineers and analysts, not some arcane distance measured in terms of a probability. This is the reason why the validation metric should be expressed in the original units, as is the case for the area metric.

Finally, consider the behavior of the area metric as theory and evidence diverge further and further. Figure 12.26 shows two graphs, each with a prediction distribution drawn in gray and data distribution drawn in black. The traditional and commonly used Smirnov's distance (which is the maximum vertical distance between the two distributions) cannot distinguish between these two comparisons. The maximal vertical distance in both cases is just unity, so the distributions are both as far apart as they can be according to the Smirnov metric. Under this measure, each data distribution is simply "far" from its prediction distribution. The area metric, on the other hand, is about 2 for the left graph and about 40

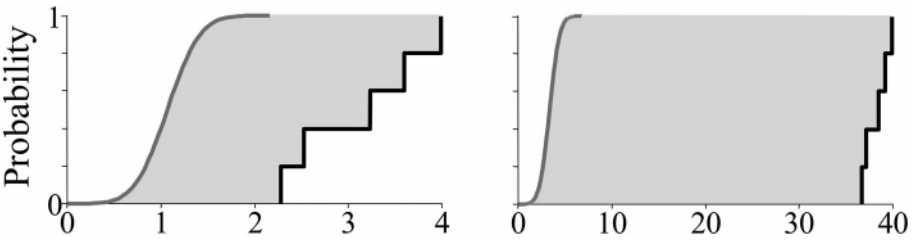


Figure 12.26 Distinguishing nonoverlapping data and prediction distributions (Ferson *et al.*, 2008).

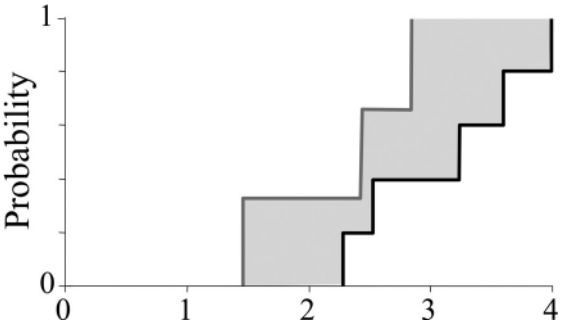


Figure 12.27 The area metric when the prediction distribution (gray) is characterized by only three simulations (Ferson *et al.*, 2008).

for the right graph. The area metric therefore identifies the left graph as having considerably more concordance between data and prediction than the right graph. If the criterion for an acceptably accurate prediction is that it is within 10 units of the actual data, then it might be that the prediction in the left graph is acceptable for the intended purpose, even though the prediction in this case does not overlap with the data. Likewise, given the same accuracy requirement, the prediction in the right graph is not acceptable for the intended use.

The area metric proposed here is applicable even when the predictions are sparse. Suppose, for instance, that it is practical to compute only a small number of simulations of a complex model to produce a handful of quantitative predictions. Although these computed results cannot produce the smooth prediction distributions shown earlier, it may be reasonable to consider the values computed to be *samples* from that smooth distribution. If they are random samples (as they would be if inputs are selected randomly), then the "empirical" distribution function formed from these values is an unbiased nonparametric estimator of the true distribution that would emerge with asymptotically many runs.

Figure 12.27 illustrates the idea of constructing $S_n$ functions for both data and predictions, the latter being the values from the sample runs of the model. In this case, there were only three simulations of the model conducted, whose values are random samples from the underlying distributions of the input. They are to be compared against a data set consisting of five values. Having very few simulation runs really means that analysts can construct only a vague picture of what the model is actually predicting. This implies the prediction

will have substantial epistemic uncertainty arising from sampling error. When only a few random samples are available from either the model or the measurements, the area metric could actually be *much smaller* if a larger number of samples were available. As a result, the present area metric should be viewed as the *evidence for the mismatch between the model and the measurements*, instead of the evidence for matching.

### *12.8.3 Pooling incomparable CDFs*

#### *12.8.3.1 u-pooling*

The previous section described how several observations of a physical process can be collected into an empirical distribution $S_n$ for comparison against a single prediction distribution. In practice, however, a model is often used to make several *different* predictions. For instance, a heating model might be used to predict the time-dependent temperature at a given location on an object. At one point in time, we have one predicted distribution of temperatures representing uncertainty about temperature then, and at another point in time, the predicted temperature distribution is different. Sometimes a model makes predictions about multiple SRQs. For instance, a single model might predict temperature, but also electrical resistivity and material stress. This would imply that we would have multiple values of the validation metric to compute. One could certainly compute all the areas separately for each pair of prediction distribution and its observation(s). Even if the data relate to a single SRQ, it would be improper to pool all of them into an empirical distribution function if they are to be compared against different prediction distributions. Each datum must be compared to the prediction distribution to which it corresponds.

A strategy was recently introduced, referred to as *u*-pooling, to deal with this situation (Oberkampf and Ferson, 2007; Ferson *et al.*, 2008). Multiple system response quantities, even if they are physically unrelated physical quantities, are combined by converting them to a universal scale via probability transformations. This strategy allows us to pool fundamentally incomparable data in terms of the relevance of each datum as evidence about the model's mismatch with experimental data. It also allows us to quantitatively answer questions like "Is the mismatch for temperature similar to the mismatch for material stress?"

To determine whether the data are generally drawn from distributions that are the same as their corresponding prediction distributions, a strategy must overcome the problem of incomparability among data and express the conformance of theory and data on some universal scale. Probability is an appropriate scale for this purpose for probabilistic models. Each datum $x_i$ is transformed by the prediction distribution $F_i$ with which it corresponds to obtain a variate $u_i = F_i(x_i)$ on the universal probability scale, which ranges on the unit interval [0, 1]. Figure 12.28 shows such transformations for three hypothetical cases. Each case is a pair of an observation depicted as a spike, i.e., single measurement, and its corresponding prediction distribution shown as a smooth curve. The prediction for the first graph is an exponential distribution on a temperature scale. In the middle graph, it is a roughly normal distribution on a scale measuring resistivity in ohms. The third graph

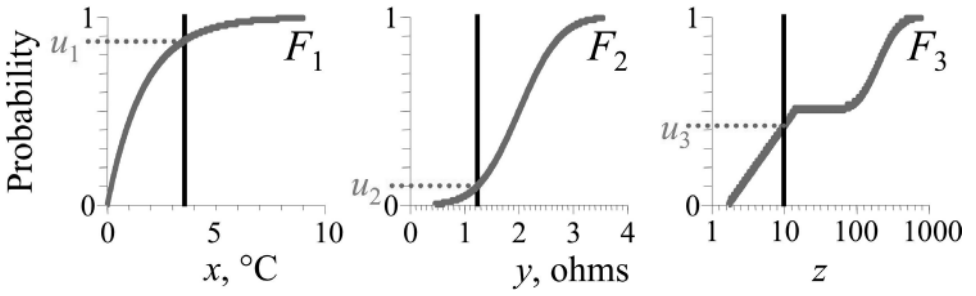


Figure 12.28 Translation of single observations (spikes) through prediction distributions (smooth) to a probability scale (ordinates) from three dimensionally inconsistent scales (abscissas) (Oberkampf and Ferson, 2007).

depicts an unusual distribution of a dimensionless quantity $z$. The intersections of the spikes and their respective distribution functions identify values on the probability scale for each $u$-value. The prediction distributions $F_i$ can be any shape at all, and they need not be the same for different observations. The $u$-values are always defined because $F(x) = 1$ for any value of $x$ larger than the largest value in the distribution, and $F(x) = 0$ for any value smaller than the smallest value in the distribution.

The various resulting $u$-values produced by these transformations can then be pooled to obtain an *overall* summary metric for the mismatch of the model's predictions to the data, even when the various individual comparisons are in different dimensions. Under the assumption that the $x_i$ are distributed according to their respective distributions $F_i$, these $u_i$ will have a uniform distribution on [0, 1]. This fact is called the *probability integral transform theorem* in statistics (Angus, 1994). This is what it means for a random variable to be "distributed according" to a distribution. The converse of this fact is perhaps more familiar to engineers and scientists because it is often used to generate random deviates from any specified probability distribution: given a distribution $F$ and a uniform random value $u$ between zero and one, the value $F^{-1}(u)$ will be a random variable distributed according to $F$. Conversely, as is needed here, if $x$ is distributed according to $F$, then $u = F(x)$ is distributed according to a uniform distribution over [0, 1]. None of this changes if there happen to be multiple $x$- and $u$-values and, in fact, none of it changes if there are multiple distribution functions so long as the $x$-values are properly matched with their respective distributions. Each $u$-value tells how the datum from which it was derived compares to its prediction distribution. The $x$-values are made into compatible $u$-values by this transformation.

Because all the $u$-values are randomly and uniformly distributed over the same range, pooling them together yields a set of values that are randomly and uniformly distributed over that range. If, however, we find that the $u_i$ are not distributed according to the uniform distribution over [0, 1], then we can infer that the $x$ observations must *not* have been distributed according to their prediction distribution functions.

In principle, the area metric can be applied directly to a pooling of all of the $u$-values. They would be collected into an empirical distribution function $S_n$ that would be compared

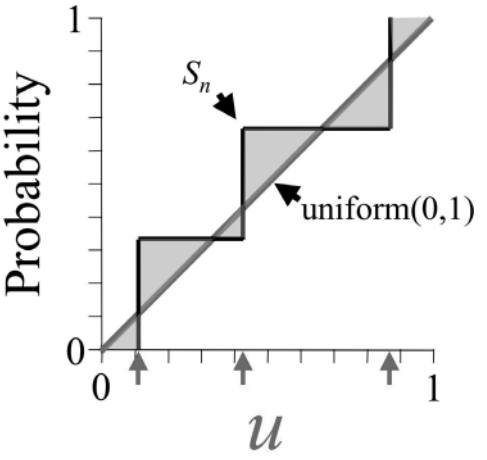


Figure 12.29 Comparison of three pooled $u$-values (black step function) against the standard uniform distribution (45° line).

against the standard uniform distribution, which is the prediction for all the $u$-values. Because both the standard uniform distribution and an empirical distribution function of the transformed values are constrained to the unit square (i.e., zero to one for cumulative probability and zero to one for the universal scale of the $u$-values), the largest possible value of the area metric is 0.5. This is the largest discrepancy between the 45° line of the uniform distribution and any distribution whose range is limited to the interval [0, 1]. The smallest possible discrepancy is zero, which would correspond to the empirical distribution function being identical to the standard uniform distribution. This would occur if and only if: (a) the data points are indeed distributed according to their respective predictions, and (b) there are sufficient data so that their step function $S_n$ approaches the continuous uniform distribution. Such a value for the validation metric would be strong evidence that there is little mismatch between the model and the measurements.

Figure 12.29 shows an example application of the area metric comparing the $u$-values from Figure 12.28, which have been synthesized into the black three-step empirical distribution function $S_n$, against the standard uniform distribution depicted as the diagonal line. The shaded region between the two functions has an area of about 0.1. The distribution of pooled $u$-values can be studied to infer characteristics of the overall match between the $x$-values and their respective prediction distributions. For instance, the area metric can be applied directly to the $u$-values compared against the standard uniform distribution. Also, the model's mismatch for different predictions, generated from their particular observations, can also be compared to each other. This would allow one to conclude that, for example, a model predicts well for, say, high temperatures but not for low temperatures. The reason this is possible is that we have transformed all the observations into the same universal probability scale for the comparisons.

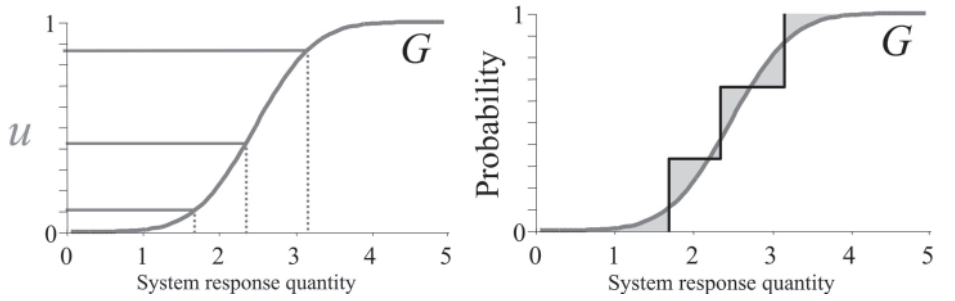


Figure 12.30 Back-transformation from the $u$-scale to an archetypical scale determined by a distribution $G$ (left) and the area metric for the pooled back-transformed values against the $G$ distribution (right) (Oberkampf and Ferson, 2007).

Transforming the observations into a universal probability scale is useful for aggregating incomparable data and comparing evidence collected in incommensurable dimensions, but, by itself, it has the disadvantage of abandoning the original physical units of the comparisons and retreating to a bounded metric constrained to the range [0, 0.5]. The deficiency can be repaired by back-transforming the $u$-values through a suitable distribution function $G$ that restores the units, scale, and their interpretation to the $u$-values. The area metric can then be computed in the physically meaningful units after this back-transformation. The justification for this transformation back to SRQs can certainly be questioned for the case where the back-transform $u$-values were pooled from evidence expressed in different SRQs. However, back-transformation seems very defensible for pooling of the same SRQ that was predicted and measured for different times in a time-dependent simulation, or different spatial locations over the domain of the PDEs.

Figure 12.30 shows how this would work for the three $u$-values considered in the previous two figures. The result in the right graph is the area between the distribution $G$ and a data distribution of back-transformed data values,

$$y_i = G^{-1}(u_i) = G^{-1}(F_i(x_i)). \tag{12.54}$$

All of these $y_i$ have the same units, which are inherited from $G$. This back-transformation, and any physical meaning associated with the units it reattaches to the $u$-values, depends on the specification of the $G$ distribution. The right graph in Figure 12.30 also shows, with shading, the area metric between the back-transformed $y_i$ and the prediction distribution $G$. This metric is in the physically meaningful units as a result of the back-transformation.

What distribution should be used to define the back-transformation? In some cases, the scientific or application context of the model accuracy assessment will specify the distribution to use for the back-transformation. This is the distribution, after all, that spells out *where* we are specifically interested in the model's predictive capability. Using the specified prediction distribution as the $G$ distribution allows all the available observations germane to *any* predictions made by the model to be used to characterize the uncertainty about this most important prediction. In general, one would want to use a distribution that

expresses where the interest in the model's predictive capability lays. This might be the prediction distribution associated with the prediction that is most important in some sense. Specifying some $G$ distribution to use as the back-transformation allows all the available observations germane to any predictions made by the model to be used to characterize the uncertainty about this most important prediction.

In validation activities for which there is no particular application requirement, there are many choices for the back-transformation distribution available to an analyst, and almost any choice one might make could yield reasonable results. For instance, the back-transformation $G$ depicted in Figure 12.30 yields a value for the area metric of almost 0.3 units. Had we instead used a uniform distribution for $G$ ranging, say, from 0 to 100 seconds, the corresponding shaded area would have the same shape as that shown in Figure 12.29, but it would have an area of about 10 units. If the $G$ were an exponential distribution, the back-transformation would emphasize deviations in the right tail of the distribution. Naturally, different $G$ distributions will express the mismatch of the data to the model in different units. But being able to express the error in different units is exactly what we want to be able to do whenever we extrapolate forecasts from models.

We note that $u$-pooling is not the only possible way to obtain an overall validation measure for data observations that address different prediction distributions. An obvious alternative is to use a multi-variate approach to the problem of validation. Such an approach could take account of any dependence information there might be in the data when different SRQs are measured simultaneously. If, for example, large deviations in temperature are typically associated with large deviations in resistivity, the correlation may be relevant to a multi-variate assessment of validation. Of course, any advantage would be at the cost of an increase in methodological complexity.

#### *12.8.3.2 Statistical significance of a metric*

As mentioned earlier, the proposed validation metric can be viewed as the evidence for mismatch between measurements and predictions. The question arises: "Is the magnitude of the metric primarily due to model error, or is it primarily due to limited samples of either or both the measurements or the predictions?" Consider, for instance, two situations. In the first, experimental observations have been exhaustively collected so that there is essentially no sampling uncertainty about the data distribution, and likewise the function evaluations are cheap so the prediction can also be specified without any sampling uncertainty. Suppose that we compute the validation metric in this situation to be $d = 1$. In the second situation, we compute the validation metric to be $d = 10$, but it is based on a very small sample size of empirical observations, or a small number of function evaluations, or both. In the first situation, the disparity between the predictions and the data is statistically significant in the sense that it cannot be explained by randomness arising from the sampling uncertainty (because there is none), but must rather be due to inaccuracies in the model. In the second situation, however, it is not clear that the disagreement between the predictions and the data is significant, even though it is ten times larger. The computed discrepancy might be

entirely due to the vagaries of random chance that were at play when the observations and function evaluations were made. Some kind of statistical analysis is required to give a *context* for these two $d$ values to understand when a value is statistically significant. Note that here we are only discussing statistical significance of the computed $d$ values, *not* the model accuracy requirement for an intended use that was discussed in Figure 12.4.

We suggest that statistical methods to detect evidence of significant mismatch between a model and its validation data can be constructed by applying standard statistical tests to the $u$-values or the $y$-values derived in the previous section. These methods can be used by an analyst to formally justify an impression or conclusion that the experimental observations disagree with the model's predictions. Transforming the $x$-values to $u$-values and pooling all the $u$-values together can substantially increase the power of the statistical test because the sample size is larger in a single, synthetic analysis. That is, $u$-pooling is a powerful tool to reduce sampling uncertainty even if it is applied to multiple measurements of the *same* SRQ.

Standard statistical tests for departures from uniformity, such as the traditional Kolmogorov–Smirnov test (D'Agostino and Stephens, 1986; Mielke and Berry, 2007), applied to the $u$-values can identify significant overall failure of the model's predictive capability. This test assumes that the experimental data values are *independent* of one another, which is not always true in practice, especially when observations have been collected for a single SRQ as a function of time. There are other statistical tests that can be applied in this situation, including the traditional chi-squared test and Neyman's smooth tests (D'Agostino and Stephens, 1986; Rayner and Rayner, 2001). The test can also be applied to compare the $y_i$ values against the predicted distribution $G$ in the physically meaningful scale. One could also define statistical tests of whether the discrepancy between data and theory is larger than some threshold size. Providing a statistical significance indicator with the validation metric result would provide the analyst, and the decision maker using the results, more information concerning model accuracy assessment. This addition would be in the same vein as the information provided by the confidence interval associated with the comparison of means approach discussed earlier in this chapter.

### *12.8.4 Inconsistency between experimental and simulation CDFs*

The technique of $u$-pooling to synthesize data corresponding to different prediction distributions has a technical limitation that is important to address (Oberkampf and Ferson, 2007). If a prediction distribution is bounded and a datum falls completely outside its range, the prediction is asserting that the datum is impossible. That is, the datum's value is in a region the prediction characterizes as having zero probability density. This is not a problem if all the data are in the same units because we can simply use the mixture $S_n$ to pool the data as described above, but if we need $u$-pooling to transform them to a common scale, then all values outside the range of the prediction distribution are transformed to zero or one (depending on whether they are below or above the range). This means that values just outside the range are considered to be the same as those very far outside the range.

Consequently, we would not be penalizing the model enough for any experimental data values that are far afield from the prediction distribution.

There are two ways to react to the situation of a prediction saying some of the data are impossible. We could conclude that the model is patently false and not bother with trying to compute any validation metric to describe its performance with respect to the data. This would, in essence, remand the problem back to the modeler who would need to develop a more reasonable model. While this might be an entirely appropriate reaction in many cases, it is perhaps not the most helpful reaction. We think that a practically robust approach to validation would tolerate very large discrepancies, even when they amount to a logical contradiction between data and prediction. The whole point of validation is to assess the performance of a model's predictive accuracy. One obvious way in which it might be wrong is if it claims that some observations are impossible, when they are not. Such a failure should be duly cataloged and appropriately incorporated into the calculation of the validation metric. Of course, data are not immune from imperfection themselves. If perchance the contradiction arises because of outlier or otherwise erroneous measurements rather than flaws in the model, then the remedy would not be a better model. In either of these cases, it would be reasonable to have some way to finesse impossibility and still make the validation metric calculation in a reasonable way.

There are different possible cases of impossibility in an attempted comparison of a model and data. Is the datum outside the prediction's range, but just outside? Is it many units away from the largest or smallest values characterized as possible? Is it orders of magnitude away? Clearly, this degree of mismatch, whatever it is, could be quantified by a reasonably designed validation metric. This section describes a strategy for designing such a metric. Although it might at first seem ad hoc, it addresses the underlying problem of large disagreements in an effective way with minimal assumptions needed to resolve them.

The robust validation metric we seek will somehow preserve *how far away* the datum is from the prediction range. The computational problem that prevents this occurs in the mapping of the data to their $u$-values. If the distribution function is infinite in both directions, then the impossibility situation will never occur, and the mapping from the data to their $u$-values is always defined in a natural way. If, however, the predicted range is bounded, then there can be data values $x$ for which $F(x)$ is not defined or is defined trivially to be zero or one. In such cases, we should like to define an *extended function* $F^*$ to use to make $u$-values. There are many ways to form a suitable extended function $F^*$ from a probability distribution. $F^*$ should be a left-continuous (but not necessarily differentiable) and monotonically increasing combination of three functions:

$$F^*(x) = \begin{cases} F^<(x), & x < \max F^{-1}(0), \\ F(x), & \max F^{-1}(0) \le x \le \min F^{-1}(1), \\ F^>(x), & \min F^{-1}(1) < x, \end{cases} \tag{12.55}$$

where $F^<$ and $F^>$ are extrapolation mappings chosen in some way that captures how far away "impossible" data are.

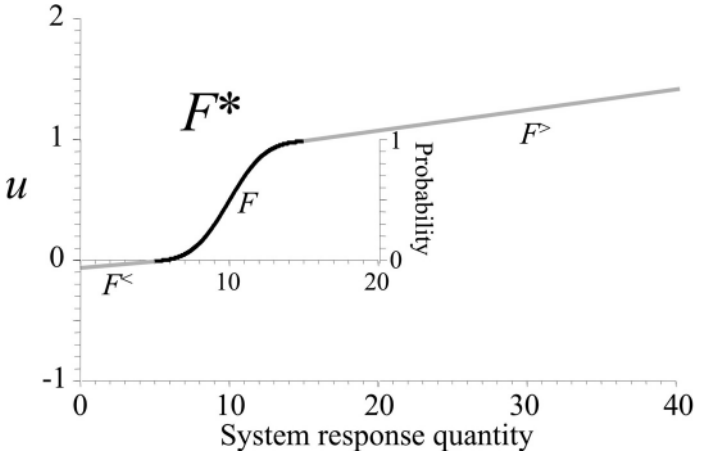


Figure 12.31 An extended function $F^*$, consisting of a distribution function $F$ (black) and extensions $F^<$ and $F^>$ (gray) to the left and right respectively.

Figure 12.31 gives an example of an extended function for a triangular distribution ranging over the interval [5, 15]. The extension mappings might be chosen by the validation analyst, or perhaps by the modeler or by some agreement between them. They might be constructed as lines whose slopes are functions of the standard deviation or interquartile range of the prediction distribution, as exemplified in the figure, or they might be more complicated nonlinear functions. The sole purpose of these mappings is to quantify the degree of the mismatch between the data and the predicted distribution in regions where the distribution itself no longer quantifies that disagreement. It is proper that these extensions are defined in an ad hoc way for a particular application whenever analysts want to accommodate the influence of the impossibility on the validation metric.

Alternatively, the extensions can be simply defined as relocated 45° lines:

$$F^<(x) = x - \max F^{-1}(0) \tag{12.56}$$

and

$$F^>(x) = x - \min F^{-1}(1) + 1. \tag{12.57}$$

In addition to being objective, these definitions have the advantage of preserving the dimensional units of the $x$-axis.

We emphasize that the extended function $F^*$ encodes a distribution function, but it is not a probability distribution itself. $F^*$ can have values less than zero and larger than one, so it is certainly no distribution function. It is merely an artifice that will allow us to quantify the extreme disagreement between a prediction distribution and data that would be considered impossible by that distribution.

Extended functions allow arbitrary data values to be translated into extended $u$-values and pooled into a common scale. These extended $u$-values range on the reals rather than only on

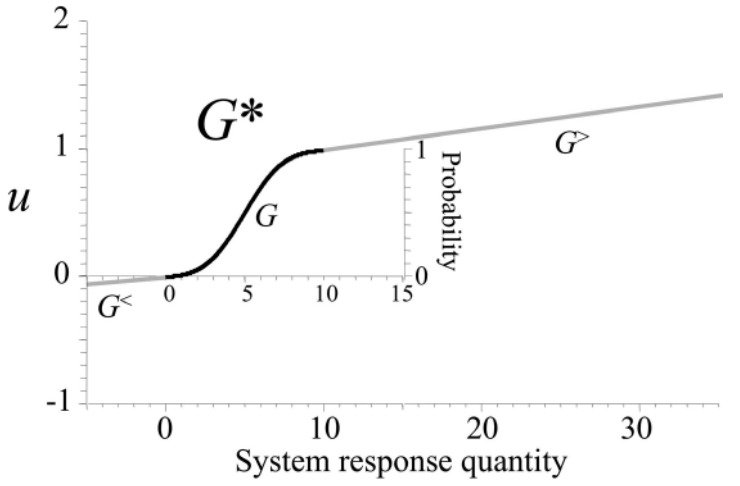


Figure 12.32 An extended back-transformation function $G^*$, consisting of a distribution function $G$ (black) and extensions $G^<$ and $G^>$ (gray) on the left and right.

the unit interval [0, 1]. The back-transformation distribution can similarly be extended to a $G^*$ function such as shown in Figure 12.32 to accept these $u$-values with values below zero or above one. This extension of the distribution function $G$ in Figure 12.30 parallels the extension of $F$ in Figure 12.31. This simple maneuver of extending the $F$ and $G$ distributions allows values considered impossible by a particular prediction to be represented, combined with other predictions and re-expressed on a common scale for calculation of a general validation metric. It can be used to combine comparisons made against bounded prediction distributions with comparisons made against prediction distributions that are infinite.

The impossibility issue does not arise when prediction distributions are infinite in both directions, like normal distributions. But of course, in practice, many prediction distributions are bounded. For instance, Weibull, exponential, and Poisson distributions cannot be negative; and beta, binomial, uniform, and triangular distributions are constrained to a range bounded above and below. If the prediction distribution is computed by Monte Carlo simulation, its range will be bounded in both directions. Even seemingly trivial bounds can lead to the problem identified here when data contain experimental uncertainties. For instance, slight leaks in a fluid transfer system composed of tubing can lead to a measurement of fluid mass that would appear to be impossible. Likewise, unsuspected addition or loss of heat in an assumed adiabatic system could lead to unrealistic measurements of thermal conductivity.

Clearly, the selection of the $G$ distribution is subjective. Another significant limitation of the approach outlined in this section is that it does not seem applicable when the back-transformation is based on a $G$ distribution that already has infinite tails such as a normal distribution. If any of the prediction distributions had to be extended, there will be some $u$-values outside the range [0, 1]. Back-transforming these values will require inverting an

extended function $G^*$ that can accept values outside [0, 1], otherwise, the back-transformed values will be undefined, or located at either plus or minus infinity. If any back-transformed values are placed at plus or minus infinity, the resulting value of the area metric will of course be infinite.

### *12.8.5 Dealing with epistemic uncertainty in the comparisons*

#### *12.8.5.1 Epistemic uncertainty in the prediction and measurements*

As discussed in Chapters 10 and 11, high-quality validation experiments should minimize, if not eliminate, epistemic uncertainties in the input quantities for the model. There are a number of situations, however, in which it cannot be avoided. Some examples are (a) the experiment was not conducted as a high quality validation experiment so that several important inputs were not measured; (b) information concerning certain inputs was measured, but it was not documented; (c) certain input quantities were quantified as interval-valued quantities based on expert opinion; and (d) the experimentalist never expected that the fidelity of physics models would be developed to the point that extremely detailed information would be needed as input data in the future. For most of these situations, the unknown information should be treated as a lack of knowledge, i.e., as an epistemic uncertainty, in the prediction.

For many of these situations, the lack of knowledge of input data needed for simulations should be represented as an interval. Giving an interval as the representation of an estimated quantity is asserting that the value (or values) of the quantity lie somewhere within the interval. Note that the interval-valued quantity can be used to represent an uncertain input quantity, or it can be used to represent the uncertainty in a parameter in a parameterized family of probability distributions. In the latter case, the uncertain quantity would be a mixture of aleatory and epistemic uncertainty that would be represented as a p-box. As discussed earlier, when these intervals, precise probability distributions, and/or p-boxes, are propagated through the model, the model prediction is a p-box for the SRQ of interest. An example of such a p-box was shown in Figure 12.20b.

Empirical observations can also contain epistemic uncertainty. A number of examples of where these occur in experimental measurements were discussed in Section 12.3.4. Again, the simplest form of this is an interval. When a collection of such intervals comprise a data set, one can think of the breadth of each interval as representing epistemic uncertainty while the scatter among the intervals represents aleatory uncertainty. Recent reviews (Manski, 2003; Gioia and Lauro, 2005; Ferson *et al.*, 2007) have described how interval uncertainty in data sets also produce p-boxes. When empirical observations have uncertainty of this form that is too large to simply ignore, these elementary techniques can be used to characterize it in a straightforward way.

#### *12.8.5.2 Epistemic and aleatory uncertainty in the metric*

The comparison between two fixed real numbers reduces to the scalar difference between the two. Suppose that, instead of both being scalar numbers, at least one of them is an interval range representing an epistemic uncertainty. If the prediction and the observation overlap, then we should say that the prediction is *correct*, in a specific sense, relative to the

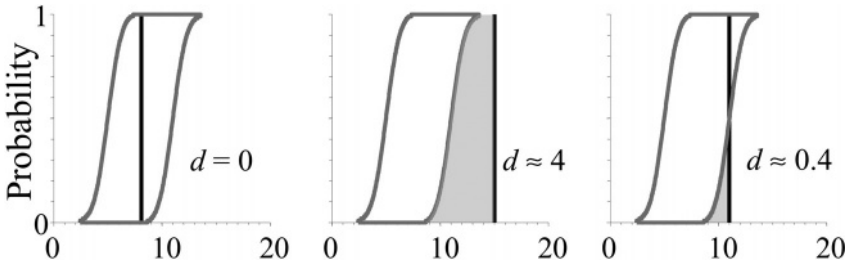


Figure 12.33 Comparison of a prediction characterized as p-boxes (smooth bounds) against three separate single observations (black spikes).

observation. If the prediction is an interval, this means that the model, for whatever reason, is making a weaker claim about what is being predicted. For example, the assertion that some system component will record a maximum operating temperature between 400 and 800 °C is a much weaker claim than saying it will be exactly 600 °C. It is also a stronger claim than saying the temperature will be between 200 and 1200 °C. In the extreme case, a prediction with extraordinarily large bounds, while not very useful, is certainly true, if just because it isn't claiming anything that might be false. For example, predicting that some probability will be between zero and one doesn't require any foresight, but at least it is free from contradiction. It is proper that a prediction's express uncertainty be counted toward reducing any measure of mismatch between theory and data in this way because the model is admitting doubt. If it were not so, an uncertainty analysis could otherwise have no epistemological value. From the perspective of validation, when the uncertainty of prediction encompasses the actual observation, there is *no evidence* of mismatch because *accuracy is distinct from precision*. Both are important in determining the usefulness of a model, but it is reasonable to distinguish them and give credit where it is due.

A reciprocal consideration applies, by the same token, if the datum is an interval to be compared against a prediction that's a real number. If the datum is an interval, and the prediction falls within the measured interval, there is no evidence of mismatch between the two. For instance, if the prediction is, say, 30% and the observation tells us that it was measured to be somewhere between 20% and 50%, then we would have to admit that the prediction might be perfectly correct. If on the other hand the evidence was that it was between 35% and 75%, then we would have to say that the disagreement between the prediction and the observation might be as low as 5%. We could also be interested in how bad the comparison might be, but a validation metric should not penalize the model for the empiricist's imprecision. In most conceptions of the word, the "distance" between two things is the length of the shortest path between them. Thus, the validation metric between a point prediction and an interval datum is the shortest difference between the characterizations of the quantities.

Figure 12.33 gives three examples of how a single point observation might compare with a prediction that is expressed as a p-box. The prediction is the same in all three graphs and is shown as smooth bounds representing the p-box for some SRQ. Against this prediction, a single observation is to be compared, and the area between this datum and the prediction

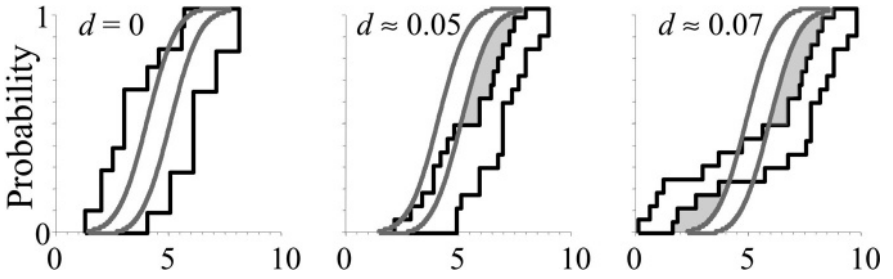


Figure 12.34 Three comparisons of predictions (smooth bounds) against empirical observations (black step function bounds), both containing aleatory and epistemic uncertainty.

is shaded. In the leftmost graph, the datum happens to fall entirely inside the bounds on the prediction. The comparison in this graph evidences no discrepancy at all between the datum and the uncertain prediction. At this value of $x$, the spike fits entirely inside the graph of the p-box, which tells us that a (degenerate) point distribution at 8 is perfectly consistent with the predictions made by the model. Thus, the area between the datum and the prediction is zero, i.e., there is no evidence for mismatch. In contrast, the data value at 15 in the middle graph is completely outside the range of either bounding distribution. The area between 15 and the prediction is about 4 units, which is the area between the spike and the right bound of the p-box. In the rightmost graph, the observation is located at the intermediate value of 11, which is within the range of the right bound of the prediction distribution. The area of mismatch in this case is only about 0.4 because the area only includes the small shaded region between 9 and 11. These comparisons are qualitatively unlike those between a scalar observation and a well-specified probability distribution. We see now that a single observation *can* perfectly match a prediction so long as the prediction has epistemic uncertainty.

Figure 12.34 illustrates three more examples, this time with epistemic uncertainty in both the prediction (shown as smooth lines) and the data (black step functions). The leftmost comparison has a distance of zero because there is at least one distribution from the prediction that can be drawn simultaneously inside both empirical distribution functions from the measurements. The area of mismatch exists only when there are no possible probability distributions that lie within both the bounds on the prediction distribution and the bounds on the data distribution. For instance, there is no such distribution consistent with both data and prediction in the middle or rightmost graph.

It should be noted that the area between the prediction and data no longer constitutes a true mathematical metric when at least one is an interval or a p-box. The reason is that the area can fall to zero without the prediction and data becoming identical (as in the leftmost graph of Figure 12.34), which violates the identity-of-indiscernibles property of a true metric. There may be ways to generalize the area metric from probability distributions to p-boxes that are mathematical metrics. However, these have not yet been developed.

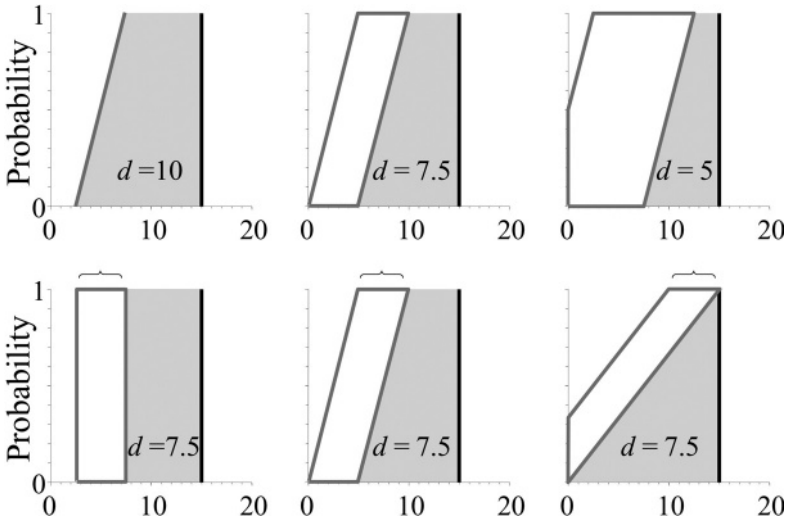


Figure 12.35 Increasing epistemic uncertainty (breadth) of predictions in the top panel and increasing variance (slant) of predictions in the lower panel.

In the cases just discussed in Figure 12.33 and Figure 12.34, the shaded regions indicate the mismatch. The validation metric now is defined by the integral

$$\int_{-\infty}^{\infty} \Delta([F_{\mathrm{R}}(x),\, F_{\mathrm{L}}(x)],\, [S_{n\mathrm{R}}(x),\, S_{n\mathrm{L}}(x)])\, \mathrm{d}x, \tag{12.58}$$

where $F$ and $S_n$ denote the prediction and the data distributions, respectively. The subscripts L and R denote the left and right bounds for those distributions, and

$$\Delta(A,\, B) = \min_{\substack{a \in A \\ b \in B}} |a - b| \tag{12.59}$$

is the shortest distance between two intervals, or zero if the intervals touch or overlap. This measure integrates the regions of nonoverlap between the two sets of bounds, for every value along the probability axis. This validation metric accepting epistemic uncertainty in either, or both, the model and the measurements is still the measure of mismatch between the model and the measurements.

Figure 12.35 illustrates another feature of this approach to assessing mismatch between prediction and measurement. As before, predictions are shown in gray and the datum is a black spike. Increasing the *breadth* of an uncertain prediction so that it possesses larger epistemic uncertainty and wider bounds can result in lowering the mismatch between the theory and data, as illustrated in the upper panel of three graphs. This breadth is a measure of the epistemic uncertainty in the prediction. It is not the same as the dispersion or variance of a distribution, which measures aleatory uncertainty. In contrast, the lower panel of three

graphs in the figure shows that increasing variance in the prediction – reflected in the *slant* of the p-box – does *not* by itself reduce the mismatch.

These behaviors of the area distinguish it from another commonly used measure of disagreement between a data value and a probability distribution expressed as the datum's displacement in standard deviation units. That measure would suggest that the three graphs in the lower panel of Figure 12.35 depict increasing agreement because the datum is progressively closer to the prediction in terms of standard deviation units. This contrast suggests that accounting for epistemic uncertainty of predictions and observations with p-boxes and measuring their mismatch as the area between those p-boxes is quite different from the common statistical idea of measuring the disagreement as displacement in standard deviation units. We think that our approach has the distinct advantage of distinguishing between aleatory and epistemic uncertainties. These uncertainties are confounded by the validation metric based on displacement in standard deviation units. For example, although the displacement decreases in the three lower graphs of Figure 12.35, the differences between small realizations from the prediction in the rightmost graph on the lower panel are necessarily larger than the corresponding difference in the graphs on the left.

## 12.9 References

Almond, R. G. (1995). *Graphical Belief Modeling*. 1st edn., London, Chapman & Hall.

Anderson, M. C., T. K. Hasselman, and T. G. Carne (1999). Model correlation and updating of a nonlinear finite element model using crush test data. *17th International Modal Analysis Conference (IMAC) on Modal Analysis*, Paper No. 376, Kissimmee, FL, Proceedings of the Society of Photo-Optical Instrumentation Engineers, 1511–1517.

Angus, J. E. (1994). The probability integral transform and related results. *SIAM Review*. **36**(4), 652–654.

Aster, R., B. Borchers, and C. Thurber (2005). *Parameter Estimation and Inverse Problems*, Burlington, MA, Elsevier Academic Press.

Aughenbaugh, J. M. and C. J. J. Paredis (2006). The value of using imprecise probabilities in engineering design. *Journal of Mechanical Design*. **128**, 969–979.

Babuska, I., F. Nobile, and R. Tempone (2008). A systematic approach to model validation based on bayesian updates and prediction related rejection criteria. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2517–2539.

Bae, H.-R., R. V. Grandhi, and R. A. Canfield (2006). Sensitivity analysis of structural response uncertainty propagation using evidence theory. *Structural and Multidisciplinary Optimization*. **31**(4), 270–279.

Barone, M. F., W. L. Oberkampf, and F. G. Blottner (2006). Validation case study: prediction of compressible turbulent mixing layer growth rate. *AIAA Journal*. **44**(7), 1488–1497.

Barre, S., P. Braud, O. Chambres, and J. P. Bonnet (1997). Influence of inlet pressure conditions on supersonic turbulent mixing layers. *Experimental Thermal and Fluid Science*. **14**(1), 68–74.

Baudrit, C. and D. Dubois (2006). Practical representations of incomplete probabilistic knowledge. *Computational Statistics and Data Analysis*. **51**, 86–108.

Bayarri, M. J., J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C. H. Lin, and J. Tu (2007). A framework for validation of computer models. *Technometrics*. **49**(2), 138–154.

Bedford, T. and R. Cooke (2001). *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge, UK, Cambridge University Press.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*, New York, John Wiley.

Bogdanoff, D. W. (1983). Compressibility effects in turbulent shear layers. *AIAA Journal*. **21**(6), 926–927.

Box, E. P. and N. R. Draper (1987). *Empirical Model-Building and Response Surfaces*, New York, John Wiley.

Chen, W., L. Baghdasaryan, T. Buranathiti, and J. Cao (2004). Model validation via uncertainty propagation. *AIAA Journal*. **42**(7), 1406–1415.

Chen, W., Y. Xiong, K.-L. Tsui, and S. Wang (2006). Some metrics and a Bayesian procedure for validating predictive models in engineering design. *ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Philadelphia, PA.

Chen, W., Y. Xiong, K.-L. Tsui, and S. Wang (2008). A design-driven validation approach using bayesian prediction models. *Journal of Mechanical Design*. **130**(2).

Chinzei, N., G. Masuya, T. Komuro, A. Murakami, and K. Kudou (1986). Spreading of two-stream supersonic turbulent mixing layers. *Physics of Fluids*. **29**(5), 1345–1347.

Coleman, H. W. and W. G. Steele, Jr. (1999). *Experimentation and Uncertainty Analysis for Engineers*. 2nd edn., New York, John Wiley.

Coleman, H. W. and F. Stern (1997). Uncertainties and CFD code validation. *Journal of Fluids Engineering*. **119**, 795–803.

Crassidis, J. L. and J. L. Junkins (2004). *Optimal Estimation of Dynamics Systems*, Boca Raton, FL, Chapman & Hall/CRC Press.

D'Agostino, R. B. and M. A. Stephens, eds. (1986). *Goodness-of-Fit-Techniques*. New York, Marcel Dekker.

Debisschop, J. R. and J. P. Bonnet (1993). Mean and fluctuating velocity measurements in supersonic mixing layers. In *Engineering Turbulence Modeling and Experiments 2: Proceedings of the Second International Symposium on Engineering Turbulence Modeling and Measurement*. W. Rodi and F. Martelli (eds. New York, Elsevier.

Debisschop, J. R., O. Chambers, and J. P. Bonnet (1994). Velocity-field characteristics in supersonic mixing layers. *Experimental Thermal and Fluid Science*. **9**(2), 147–155.

DesJardin, P. E., T. J. O'Hern, and S. R. Tieszen (2004). Large eddy simulation of experimental measurements of the near-field of a large turbulent helium plume. *Physics of Fluids*. **16**(6), 1866–1883.

DeVolder, B., J. Glimm, J. W. Grove, Y. Kang, Y. Lee, K. Pao, D. H. Sharp, and K. Ye (2002). Uncertainty quantification for multiscale simulations. *Journal of Fluids Engineering*. **124**(1), 29–41.

Devore, J. L. (2007). *Probability and Statistics for Engineers and the Sciences*. 7th edn., Pacific Grove, CA, Duxbury.

Dowding, K. J., R. G. Hills, I. Leslie, M. Pilch, B. M. Rutherford, and M. L. Hobbs (2004). *Case Study for Model Validation: Assessing a Model for Thermal Decomposition of Polyurethane Foam*. SAND2004–3632, Albuquerque, NM, Sandia National Laboratories.

Dowding, K. J., J. R. Red-Horse, T. L. Paez, I. M. Babuska, R. G. Hills, and R. Tempone (2008). Editorial: Validation challenge workshop summary. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2381–2384.

Draper, N. R. and H. Smith (1998). *Applied Regression Analysis*. 3rd edn., New York, John Wiley.
Drosg, M. (2007). *Dealing with Uncertainties: a Guide to Error Analysis*, Berlin, Springer-Verlag.
Dubois, D. and H. Prade, eds. (2000). *Fundamentals of Fuzzy Sets*. Boston, MA, Kluwer Academic Publishers.
Dutton, J. C., R. F. Burr, S. G. Goebel, and N. L. Messersmith (1990). Compressibility and mixing in turbulent free shear layers. *12th Symposium on Turbulence*, Rolla, MO, University of Missouri-Rolla, A22–1 to A22–12.
Easterling, R. G. (2001). *Measuring the Predictive Capability of Computational Models: Principles and Methods, Issues and Illustrations*. SAND2001–0243, Albuquerque, NM, Sandia National Laboratories.
Easterling, R. G. (2003). *Statistical Foundations for Model Validation: Two Papers*. SAND2003–0287, Albuquerque, NM, Sandia National Laboratories.
Elliot, G. S. and M. Samimy (1990). Compressibility effects in free shear layers. *Physics of Fluids A*. **2**(7), 1231–1240.
Ferson, S. (2002). *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Setauket, NY, Applied Biomathematics.
Ferson, S. and W. L. Oberkampf (2009). Validation of imprecise probability models. *International Journal of Reliability and Safety*. **3**(1–3), 3–22.
Ferson, S., V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz (2003). *Constructing Probability Boxes and Dempster-Shafer Structures*. SAND2003–4015, Albuquerque, NM, Sandia National Laboratories.
Ferson, S., R. B. Nelsen, J. Hajagos, D. J. Berleant, J. Zhang, W. T. Tucker, L. R. Ginzburg, and W. L. Oberkampf (2004). *Dependence in Probabilistic Modeling, Dempster-Shafer Theory, and Probability Bounds Analysis*. SAND2004–3072, Albuquerque, NM, Sandia National Laboratories.
Ferson, S., V. Kreinovich, H. Hajagos, W. L. Oberkampf, and L. Ginzburg (2007). *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. Albuquerque, Sandia National Laboratories.
Ferson, S., W. L. Oberkampf, and L. Ginzburg (2008). Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*. **197**, 2408–2430.
Fetz, T., M. Oberguggenberger, and S. Pittschmann (2000). Applications of possibility and evidence theory in civil engineering. *International Journal of Uncertainty*. **8**(3), 295–309.
Gartling, D. K., R. E. Hogan, and M. W. Glass (1994). *Coyote – a Finite Element Computer Program for Nonlinear Heat Conduction Problems, Part I – Theoretical Background*. SAND94–1173, Albuquerque, NM, Sandia National Laboratories.
Geers, T. L. (1984). An objective error measure for the comparison of calculated and measured transient response histories. *The Shock and Vibration Bulletin*. **54**(2), 99–107.
Gelman, A. B., J. S. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*, London, Chapman & Hall.
Ghosh, J. K., M. Delampady, and T. Samanta (2006). *An Introduction to Bayesian Analysis: Theory and Methods*, Berlin, Springer-Verlag.
Giaquinta, M. and G. Modica (2007). *Mathematical Analysis: Linear and Metric Structures and Continuity*, Boston, Birkhauser.

Gioia, F. and C. N. Lauro (2005). Basic statistical methods for interval data. *Statistica Applicata*. **17**(1), 75–104.
Goebel, S. G. and J. C. Dutton (1991). Experimental study of compressible turbulent mixing layers. *AIAA Journal*. **29**(4), 538–546.
Grabe, M. (2005). *Measurement Uncertainties in Science and Technology*, Berlin, Springer-Verlag.
Gruber, M. R., N. L. Messersmith, and J. C. Dutton (1993). Three-dimensional velocity field in a compressible mixing layer. *AIAA Journal*. **31**(11), 2061–2067.
Haldar, A. and S. Mahadevan (2000). *Probability, Reliability, and Statistical Methods in Engineering Design*, New York, John Wiley.
Hanson, K. M. (1999). A framework for assessing uncertainties in simulation predictions. *Physica D*. **133**, 179–188.
Hasselman, T. K., G. W. Wathugala, and J. Crawford (2002). A hierarchical approach for model validation and uncertainty quantification. *Fifth World Congress on Computational Mechanics*, wccm.tuwien.ac.at, Vienna, Austria, Vienna University of Technology.
Hazelrigg, G. A. (2003). Thoughts on model validation for engineering design. *ASME 2003 Design Engineering Technical Conference and Computers and and Information in Engineering Conference*, DETC2003/DTM-48632, Chicago, IL, ASME.
Helton, J. C., J. D. Johnson, and W. L. Oberkampf (2004). An exploration of alternative approaches to the representation of uncertainty in model predictions. *Reliability Engineering and System Safety*. **85**(1–3), 39–71.
Helton, J. C., W. L. Oberkampf, and J. D. Johnson (2005). Competing failure risk analysis using evidence theory. *Risk Analysis*. **25**(4), 973–995.
Higdon, D., M. Kennedy, J. Cavendish, J. Cafeo and R. D. Ryne (2004). Combining field observations and simulations for calibration and prediction. *SIAM Journal of Scientific Computing*. **26**, 448–466.
Higdon, D., C. Nakhleh, J. Gattiker, and B. Williams (2009). A Bayesian calibration approach to the thermal problem. *Computer Methods in Applied Mechanics and Engineering*. In press.
Hills, R. G. (2006). Model validation: model parameter and measurement uncertainty. *Journal of Heat Transfer*. **128**(4), 339–351.
Hills, R. G. and I. Leslie (2003). *Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application*. SAND2003–0706, Albuquerque, NM, Sandia National Laboratories.
Hills, R. G. and T. G. Trucano (2002). *Statistical Validation of Engineering and Scientific Models: a Maximum Likelihood Based Metric*. SAND2001–1783, Albuquerque, NM, Sandia National Laboratories.
Hobbs, M. L. (2003). Personal communication.
Hobbs, M. L., K. L. Erickson, and T. Y. Chu (1999). *Modeling Decomposition of Unconfined Rigid Polyurethane Foam*. SAND99–2758, Albuquerque, NM, Sandia National Laboratories.
Huber-Carol, C., N. Balakrishnan, M. Nikulin, and M. Mesbah, eds. (2002). *Goodness-of-Fit Tests and Model Validity*. Boston, Birkhauser.
ISO (1995). *Guide to the Expression of Uncertainty in Measurement*. Geneva, Switzerland, International Organization for Standardization.
Iuzzolino, H. J., W. L. Oberkampf, M. F. Barone, and A. P. Gilkey (2007). *User's Manual for VALMET: Validation Metric Estimator Program*. SAND2007–6641, Albuquerque, NM, Sandia National Laboratories.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B – Statistical Methodology*. **63**(3), 425–450.
Klir, G. J. (2006). *Uncertainty and Information: Foundations of Generalized Information Theory*, Hoboken, NJ, Wiley Interscience.
Klir, G. J. and M. J. Wierman (1998). *Uncertainty-Based Information: Elements of Generalized Information Theory*, Heidelberg, Physica-Verlag.
Kohlas, J. and P.-A. Monney (1995). *A Mathematical Theory of Hints – an Approach to the Dempster-Shafer Theory of Evidence*, Berlin, Springer-Verlag.
Krause, P. and D. Clark (1993). *Representing Uncertain Knowledge: an Artificial Intelligence Approach*, Dordrecht, The Netherlands, Kluwer Academic Publishers.
Kriegler, E. and H. Held (2005). Utilizing belief functions for the estimation of future climate change. *International Journal for Approximate Reasoning*. **39**, 185–209.
Law, A. M. (2006). *Simulation Modeling and Analysis*. 4th edn., New York, McGraw-Hill.
Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. 3rd edn., Berlin, Springer-Verlag.
Leonard, T. and J. S. J. Hsu (1999). *Bayesian Methods: an Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge, UK, Cambridge University Press.
Liu, F., M. J. Bayarri, J. O. Berger, R. Paulo, and J. Sacks (2009). A Bayesian analysis of the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2457–2466.
Manski, C. F. (2003). *Partial Identification of Probability Distributions*, New York, Springer-Verlag.
MathWorks (2005). *MATLAB*. Natick, MA, The MathWorks, Inc.
McFarland, J. and S. Mahadevan (2008). Multivariate significance testing and model calibration under uncertainty. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2467–2479.
Mielke, P. W. and K. J. Berry (2007). *Permutation Methods: a Distance Function Approach*. 2nd edn., Berlin, Springer-Verlag.
Miller, R. G. (1981). *Simultaneous Statistical Inference*. 2nd edn., New York, Springer-Verlag.
Molchanov, I. (2005). *Theory of Random Sets*, London, Springer-Verlag.
Nagano, Y. and M. Hishida (1987). Improved form of the k-epsilon model for wall turbulent shear flows. *Journal of Fluids Engineering*. **109**(2), 156–160.
Nguyen, H. T. and E. A. Walker (2000). *A First Course in Fuzzy Logic*. 2nd edn., Cleveland, OH, Chapman & Hall/CRC.
Oberkampf, W. L. and M. F. Barone (2004). Measures of agreement between computation and experiment: validation metrics. *34th AIAA Fluid Dynamics Conference*, AIAA Paper 2004–2626, Portland, OR, American Institute of Aeronautics and Astronautics.
Oberkampf, W. L. and M. F. Barone (2006). Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics*. **217**(1), 5–36.
Oberkampf, W. L. and S. Ferson (2007). Model validation under both aleatory and epistemic uncertainty. *NATO/RTO Symposium on Computational Uncertainty in Military Vehicle Design*, AVT-147/RSY-022, Athens, Greece, NATO.
Oberkampf, W. L. and J. C. Helton (2005). Evidence theory for engineering applications. In *Engineering Design Reliability Handbook*. E. Nikolaidis, D. M. Ghiocel and S. Singhal (eds.). New York, NY, CRC Press: 29.
Oberkampf, W. L. and T. G. Trucano (2002). Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences*. **38**(3), 209–272.

Oberkampf, W. L., T. G. Trucano, and C. Hirsch (2004). Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*. **57**(5), 345–384.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: a tutorial. *Reliability Engineering and System Safety*. **91**(10–11), 1290–1300.

O'Hern, T. J., E. J. Weckman, A. L. Gerhart, S. R. Tieszen, and R. W. Schefer (2005). Experimental study of a turbulent buoyant helium plume. *Journal of Fluid Mechanics*. **544**, 143–171.

Paciorri, R. and F. Sabetta (2003). Compressibility correction for the Spalart-Allmaras model in free-shear flows. *Journal of Spacecraft and Rockets*. **40**(3), 326–331.

Paez, T. L. and A. Urbina (2002). Validation of mathematical models of complex structural dynamic systems. *Proceedings of the Ninth International Congress on Sound and Vibration*, Orlando, FL, International Institute of Acoustics and Vibration.

Papamoschou, D. and A. Roshko (1988). The compressible turbulent shear layer: an experimental study. *Journal of Fluid Mechanics*. **197**, 453–477.

Pilch, M. (2008). Preface: Sandia National Laboratories Validation Challenge Workshop. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2373–2374.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2007). *Numerical Recipes in FORTRAN*. 3rd edn., New York, Cambridge University Press.

Pruett, C. D., T. B. Gatski, C. E. Grosch, and W. D. Thacker (2003). The temporally filtered Navier-Stokes equations: properties of the residual stress. *Physics of Fluids*. **15**(8), 2127–2140.

Rabinovich, S. G. (2005). *Measurement Errors and Uncertainties: Theory and Practice*. 3rd edn., New York, Springer-Verlag.

Raol, J. R., G. Girija and J. Singh (2004). *Modelling and Parameter Estimation of Dynamic Systems*, London, UK, Institution of Engineering and Technology.

Rayner, G. D. and J. C. W. Rayner (2001). Power of the Neyman smooth tests for the uniform distribution. *Journal of Applied Mathematics and Decision Sciences*. **5**(3), 181–191.

Rider, W. J. (1998). Personal communication.

Roache, P. J. (1998). *Verification and Validation in Computational Science and Engineering*, Albuquerque, NM, Hermosa Publishers.

Rougier, J. (2007). Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climate Change*. **81**(3–4), 247–264.

Russell, D. M. (1997a). Error measures for comparing transient data: Part I, Development of a comprehensive error measure. *Proceedings of the 68th Shock and Vibration Symposium*, Hunt Valley, Maryland, Shock and Vibration Information Analysis Center.

Russell, D. M. (1997b). Error measures for comparing transient data: Part II, Error measures case study. *Proceedings of the 68th Shock and Vibration Symposium*, Hunt Valley, Maryland, Shock and Vibration Information Analysis Center.

Rutherford, B. M. and K. J. Dowding (2003). *An Approach to Model Validation and Model-Based Prediction – Polyurethane Foam Case Study*. Sandia National Laboratories, SAND2003–2336, Albuquerque, NM.

Samimy, M. and G. S. Elliott (1990). Effects of compressibility on the characteristics of free shear layers. *AIAA Journal*. **28**(3), 439–445.

Seber, G. A. F. and C. J. Wild (2003). *Nonlinear Regression*, New York, John Wiley.

Sivia, D. and J. Skilling (2006). *Data Analysis: a Bayesian Tutorial*. 2nd edn., Oxford, Oxford University Press.

Sprague, M. A. and T. L. Geers (1999). Response of empty and fluid-filled, submerged spherical shells to plane and spherical, step-exponential acoustic waves. *Shock and Vibration*. **6**(3), 147–157.

Sprague, M. A. and T. L. Geers (2004). A spectral-element method for modeling cavitation in transient fluid-structure interaction. *International Journal for Numerical Methods in Engineering*. **60**(15), 2467–2499.

Stern, F., R. V. Wilson, H. W. Coleman, and E. G. Paterson (2001). Comprehensive approach to verification and validation of CFD simulations – Part 1: Methodology and procedures. *Journal of Fluids Engineering*. **123**(4), 793–802.

Tieszen, S. R., S. P. Domino, and A. R. Black (2005). *Validation of a Simple Turbulence Model Suitable for Closure of Temporally-Filtered Navier Stokes Equations Using a Helium Plume*. SAND2005–3210, Albuquerque, NM, Sandia National Laboratories.

Trucano, T. G., L. P. Swiler, T. Igusa, W. L. Oberkampf, and M. Pilch (2006). Calibration, validation, and sensitivity analysis: what's what. *Reliability Engineering and System Safety*. **91**(10–11), 1331–1357.

van den Bos, A. (2007). *Parameter Estimation for Scientists and Engineers*, Hoboken, NJ, Wiley-Interscience.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, London, Chapman & Hall.

Wang, S., W. Chen and K.-L. Tsui (2009). Bayesian validation of computer models. *Technometrics*. **51**(4), 439–451.

Wellek, S. (2002). *Testing Statistical Hypotheses of Equivalence*, Boca Raton, FL, Chapman & Hall/CRC.

Wilcox, D. C. (2006). *Turbulence Modeling for CFD*. 3rd edn., La Canada, CA, DCW Industries.

Winkler, R. L. (1972). *An Introduction to Bayesian Inference and Decision*, New York, Holt, Rinehart, and Winston.

Wirsching, P., T. Paez and K. Ortiz (1995). *Random Vibrations: Theory and Practice*, New York, Wiley.

Wong, C. C., F. G. Blottner, J. L. Payne, and M. Soetrisno (1995a). Implementation of a parallel algorithm for thermo-chemical nonequilibrium flow solutions. *AIAA 33rd Aerospace Sciences Meeting*, AIAA Paper 95–0152, Reno, NV, American Institute of Aeronautics and Astronautics.

Wong, C. C., M. Soetrisno, F. G. Blottner, S. T. Imlay, and J. L. Payne (1995b). *PINCA: A Scalable Parallel Program for Compressible Gas Dynamics with Nonequilibrium Chemistry*. SAND94–2436, Albuquerque, NM, Sandia National Laboratories.

Yee, H. C. (1987). *Implicit and Symmetric Shock Capturing Schemes*. Washington, DC, NASA, NASA-TM-89464.

Yoon, S. and A. Jameson (1987). An LU-SSOR scheme for the Euler and Navier-Stokes equations. *25th AIAA Aerospace Sciences Meeting*, AIAA Paper 87–0600, Reno, NV, American Institute of Aeronautics and Astronautics.

Zeman, O. (1990). Dilatation dissipation: the concept and application in modeling compressible mixing layers. *Physics of Fluids A*. **2**(2), 178–188.

Zhang, R. and S. Mahadevan (2003). Bayesian methodology for reliability model acceptance. *Reliability Engineering and System Safety*. **80**(1), 95–103.