# 10

# Model validation fundamentals

Because of the well-developed methodologies and techniques developed in metrology, experimental measurements are usually viewed as the best way to estimate a true value, if one exists. This level of trust and credibility in experimental measurements has been built over at least four millennia of learning, both from mistakes and successes. This is not meant to imply that experimental measurements *always* yield an accurate estimate of the true value. Experimental measurements can be inaccurate, or downright wrong, for many reasons. What is meant by the trustworthiness of an experimental measurement is that techniques for investigating its limitations, weaknesses, and uncertainties are generally well understood. When new experimental diagnostic techniques are developed they must be carefully investigated and understood. Where possible, comparisons need to be made with measurements from existing, better understood, techniques so that measurement uncertainty can be better quantified. As science and technology progresses, new measurement techniques can increase the confidence in measurement accuracy and also begin to measure physical quantities that were previously immeasurable.

An enlightening way of thinking about the trustworthiness of experimental measurements is to think of an experimental measurement as "asking a question of nature" (Hornung and Perry, 1998). When a measurement result is obtained, the result can be thought of as nature's answer to a question we have asked. We tend to believe that the answer obtained is the answer to the question we think we asked. However, this is not actually the case because there are always assumptions involved on our part. For example, when we measure the fluid velocity in a flow field, we believe we asked the question: given the flow field of interest, what is the velocity of the fluid at a certain point? Our intent is to ask nature the question that excludes or minimizes the effect of random or systematic errors in the measurement. If, however, there is significant random measurement error or if there is an unknown systematic error either in the measurement itself or in the data reduction procedure, then the question asked of nature is different from what we thought we asked. That is, nature answered the question that *includes* the random and systematic errors, whether large or small.

In addition, nature can use any vagueness or ambiguity in our question to deceive us. The deceit is not in any sense trickery or maliciousness on the part of nature. It is the questioner who can easily deceive him/herself because either it is not recognized that the question is vague, or because our preconceived notions or agendas led us astray. The goal of science, whether discoveries are through experiment, theory, or simulation, is to refine our understanding and description of nature, regardless of our present confusion or beliefs. Consider the following example. Suppose we are interested in measuring some local quantity in a physical domain, e.g., the local strain in a solid, the local velocity in a flow field, or the total energy at a point. We frame our question of nature within a certain mindset, such as the strain is elastic, the flow field is steady, or certain components of energy are unimportant. Then we make our measurements and interpret the results within our mindset. Often the interpretation makes perfect sense and is fully consistent within our framework, but our interpretation could be *completely* wrong. Our mindset most often involves assumptions concerning the relevant physics and the measurement technique, but it also includes our presumption that the theory or the simulation is correct. Any disagreements or oddities in the measurements that don't quite fit our theory or simulation are relegated to experimental error or dissolved into our model using calibration of parameters. Human nature places great value on success, but nature has no agenda.

From a validation perspective, we could think about asking our simulation the same question that is asked of nature. The philosophical foundation of validation experiments is to ask precisely the *same* question of our model that we ask of nature. From our discussion above, we see that this presents a dichotomy. From an experimental measurement viewpoint, we strive for precision in the question of nature while minimizing (a) the constraints of our measurement assumptions, and (b) the uncertainty of our measurements. From a simulation viewpoint, we are fundamentally constrained by the assumptions embedded in the model. To bridge this gulf, we must steadfastly seek to ensure that the experiment provides all the input information needed for the simulation. In this way, we can critically test the accuracy of the assumptions in our model. For example, the experimentalist should provide all the boundary conditions (BCs), initial conditions (ICs), system excitation, geometric features, and other input data needed for the simulation. If our knowledge of the needed conditions is poor or pieces of information are missing, then our simulation is answering a somewhat different question than was asked of nature. Or, if there is a systematic uncertainty in the measurements, then the question asked of nature is different from what was asked of our simulation.

## 10.1 Philosophy of validation experiments

### *10.1.1 Validation experiments vs. traditional experiments*

Experimentalists, computational analysts, and project managers ask: what is a validation experiment? Or: how is a validation experiment different from other experiments? These are appropriate questions. Traditional experiments could be grouped into three general categories (Oberkampf and Trucano, 2002; Trucano *et al.*, 2002; Oberkampf *et al.*, 2004). The first category comprises experiments that are conducted primarily to improve the fundamental understanding of some physical process. Sometimes these are referred to as physical discovery or phenomena discovery experiments. Examples are (a) experiments that measure fundamental fluid turbulence characteristics; (b) experiments that investigate crack propagation in solids; (c) experiments in high-energy density physics; and (d) experiments probing the onset and stability of phase changes in solids, liquids, and gases.

The second category of traditional experiments consists of those conducted primarily for constructing, improving, or determining parameters in mathematical models of fairly-well understood physical processes. Sometimes these are referred to as model calibration or model updating experiments. Examples are (a) experiments to measure reaction rate parameters in reacting or detonating flows, (b) experiments to measure thermal emissivity of material surfaces, (c) experiments to calibrate parameters in a model for predicting large plastic deformation of a structure, and (d) experiments to calibrate mass diffusion rate parameters in a mass transport chemistry model.

The third category of traditional experiments includes those that determine the reliability, performance, or safety of components, subsystems, or complete systems. These experiments are sometimes referred to as acceptance tests or qualification tests of engineered components, subsystems, or systems. Examples are (a) tests of a new combustor design in gas turbine engines, (b) pressurization tests of a new design of a filament-wound composite pressure vessel, (c) safety test of the emergency cooling system in a nuclear power reactor, and (d) a limit load test of an aircraft wing structure.

Validation experiments constitute a new type of experiment. (See Oberkampf and Aeschliman, 1992; Marvin, 1995; Oberkampf *et al.*, 1995; Aeschliman and Oberkampf, 1998, for a discussion of the early concepts of a validation experiment.) A validation experiment is conducted for the primary purpose of determining the predictive capability of a mathematical model of a physical process. In other words, a validation experiment is designed, executed, and analyzed for the purpose of quantitatively determining the ability of the model and its embodiment in a computer code to simulate a well-characterized physical process. In a validation experiment *the model builder is the customer* or similarly *the computational analyst is the customer*. Only during the last two decades has scientific computing matured to the point where it could even be considered as a separate customer in experimental activities. As modern technology increasingly moves toward engineering systems that are designed, certified, and even fielded based primarily on scientific computing, then scientific computing itself will increasingly become the customer of experiments.

One other aspect of validation experiments that is different from traditional experiments is that traditional experiments place a great deal of emphasis on measurements of processes in a controlled environment. Only with a controlled environment can measurements of physical processes be reliably repeated by other experimentalists; models carefully calibrated; and the reliability, performance, and safety of systems assessed. In validation experiments, however, *characterization* of the experiment is the more important goal. By *characterization* we mean measuring all of the important characteristics of the experiment that are needed for the simulation, both within the system and in the surroundings. Stated differently, control and repeatability of the experiment are less important in a validation experiment than precisely *measuring* the conditions of an uncontrolled experiment. Variability in the surroundings of a validation experiment, for example due to weather conditions, is

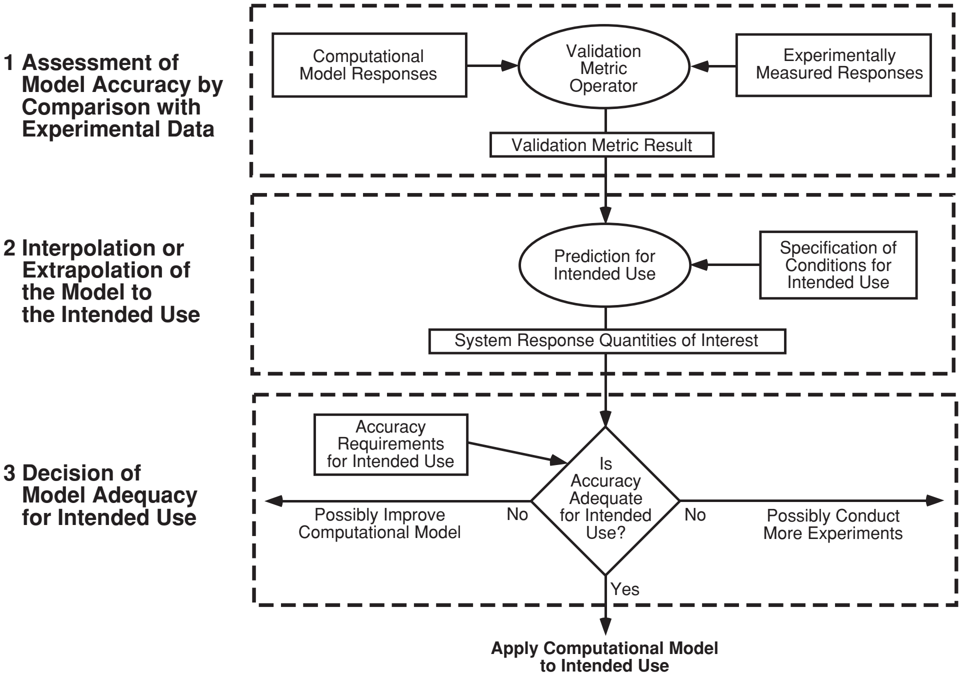

1 Assessment of Model Accuracy by Comparison with Experimental Data
Computational Model Responses
Validation Metric Operator
Experimentally Measured Responses
Validation Metric Result
2 Interpolation or Extrapolation of the Model to the Intended Use
Prediction for Intended Use
Specification of Conditions for Intended Use
System Response Quantities of Interest
3 Decision of Model Adequacy for Intended Use
Accuracy Requirements for Intended Use
Is Accuracy Adequate for Intended Use?
No
No
Possibly Improve Computational Model
Possibly Conduct More Experiments
Yes
Apply Computational Model to Intended Use

Figure 10.1 Three aspects of the encompassing view of model validation (Oberkampf and Trucano, 2007; Oberkampf and Trucano, 2008).

not critical, as long as the conditions of the surroundings are precisely measured. For experiments with uncontrolled conditions, however, a number of experimental realizations are necessary to carefully characterize the variability of the system and surroundings so that this information can be provided to the computational analyst.

### *10.1.2 Goals and strategy of validation*

In Chapter 2, Fundamental concepts and terminology, the fundamental concepts of validation were introduced. In Section 2.2.3, three aspects of the encompassing view of model validation were discussed with regard to Figure 10.1. Aspect 1 involves the quantitative comparison of computational responses and experimentally measured responses. The mathematical operator used to make the comparisons is a *validation metric operator*. The operator is usually formulated as a difference operator so that the validation metric result is a measure of the disagreement between the computational and experimental responses. Aspect 2 deals with the use of the model to make predictions, in the sense of interpolation or extrapolation, for the conditions of the intended use of the model. Aspect 3 deals with (a) the comparison of the estimated accuracy of the model relative to the accuracy requirements of the model for the domain of the model's intended use, and (b) the decision of adequacy/inadequacy of the model over the domain of the model's intended use.

Aspect 1 can be thought of in two ways: scientific validation and project-oriented validation. *Scientific validation* is a quantitative assessment of model accuracy without regard to any specific accuracy requirements or engineering project needs. This is the most common type of validation activity that is published in the literature. *Project-oriented validation* is a quantitative assessment of model accuracy with the needs of the project taking priority. The following sections will discuss each type of validation activity in detail. We remind the reader that this book uses the restricted view of the term *validation*, meaning model accuracy assessment as depicted in Aspect 1.

#### *10.1.2.1 Scientific validation*

A number of authors have written papers and articles concerning the general strategy of conducting model validation. Some of the key contributors were Marvin (1995); Rykiel (1996); Aeschliman and Oberkampf (1998); Barber (1998); Benek *et al.* (1998); Kleijnen (1998); Kleindorfer *et al.* (1998); Murray-Smith (1998); Roache (1998); Sargent (1998); Balci *et al.* (2000); Refsgaard (2000); Anderson and Bates (2001); Oberkampf and Trucano (2002); Trucano *et al.* (2002); Oberkampf *et al.* (2004); and Oberkampf and Barone (2006). Most published work dealing with comparing model and experimental results is directed toward what is referred to here as *scientific validation*. By using this term, we are *not* contrasting validation in scientific endeavors versus engineering endeavors. We are using the term in the broad sense to mean any type of quantitative comparison between computational results and experimental measurements for the purpose of accuracy assessment of a physics-based model. Discussed here are some of the important strategies for improved accuracy assessment that have been learned through the years.

Descriptions and documentation of experiments that are subsequently used for model validation seldom provide all of the important input information needed for the model. For documentation of experiments that appear in journal articles, there is considerable pressure to limit the length of the article and avoid including detailed information. The biggest reason, however, for the lack of documented information is that many experimentalists do not understand or care about the input information needs of an analyst who wants to use the experiment for model validation. In addition, the experimentalists are at a disadvantage because they can only guess at all of the types of information that might be needed for different modeling approaches. Given this lack of information, computational analysts almost always choose to adjust unknown parameters or conditions so that the best agreement is obtained between their results and the experimental measurements. Any type of adjustment of parameters, conditions, or modeling approaches degrades the primary goal of validation: assessment of predictive accuracy of a model. Sometimes the adjustment of parameters or conditions is explicit, for example, using calibration or parameter estimation procedures. This, of course, significantly decreases critical assessment of model predictive accuracy. Sometimes, the adjustment procedures are not explicitly mentioned or explained, either intentionally or unintentionally. For example, certain modeling approaches were attempted, gave poor agreement with the experimental data, and were abandoned. The experience, however, was used to guide new modeling assumptions that yielded modeling results with improved agreement with the data.

For experiments that are designed and conducted in the future, a much more constructive procedure is for both the computational analysts and the experimentalists to jointly design the experiment. Many of the difficulties just mentioned can be eliminated in a joint effort. This topic will be discussed in depth in Chapter 11, Design and execution of validation experiments, but here, two aspects of a joint effort will be mentioned. First, in a joint effort, the analyst must inform the experimentalist of the input information required by the model. In addition, the analyst must communicate to the experimentalist what system response quantities (SRQs) are of particular interest so that they can be measured. Second, the experimental measurements of the SRQs should be withheld from the analysts so that a blind-test computational prediction can be made. Although the value of blind predictions in scientific computing is not uniformly accepted, we believe it is crucial to withhold the measured SRQs so as to critically assess the predictive accuracy of a model. Some fields of science, such as drug testing in medicine, have long recognized that without blind or double-blind testing, conclusions are commonly distorted and misleading.

In comparing computational and experimental results, emphasis should be placed on quantitative comparison methods. Color contour levels of a SRQ over a two-dimensional region, one for a computational result and one for an experimental result, can be qualitatively useful. However, little quantitative information can be gained, especially if the quantitative values of the color scale are not given. The most common comparison method is a graph where computationally predicted and experimentally measured SRQs are shown over a range of an input, or control, parameter. Two quantitative shortcomings of this method are the following. First, agreement or disagreement between computational and experimental results is not quantified, sometimes not even stated. The focus is on the trends in the SRQ, not quantifying the computation-experiment disagreement. Observations and conclusions are commonly made about the claimed agreement, such as "good" or "excellent" or "the model is validated," but all of these are in the eye of the beholder in this approach.

Second, these types of graph rarely have any type of information concerning the effect of the uncertainties on the results. For example, the computational results are commonly deterministic results. Additionally, no information is shown concerning the influence of the numerical solution errors on the results. For the experimental results, it is still common to see experimental results without any experimental uncertainty estimates shown. Significant improvement in quantitative comparisons is made if the effects of the uncertainties on both the computational and experimental results are shown. One method is to show the mean value and the uncertainty bars with plus or minus two standard deviations for both the computational results and experimental measurements. The topic of quantitative comparison of computational and experimental results will be discussed in detail in Chapter 12, Model accuracy assessment.

Computational results and experimental results produce many different SRQs. Validation is concerned with the comparison of the same SRQ under the same conditions from both the computation and experiment. There is a range of difficulty in predicting and experimentally

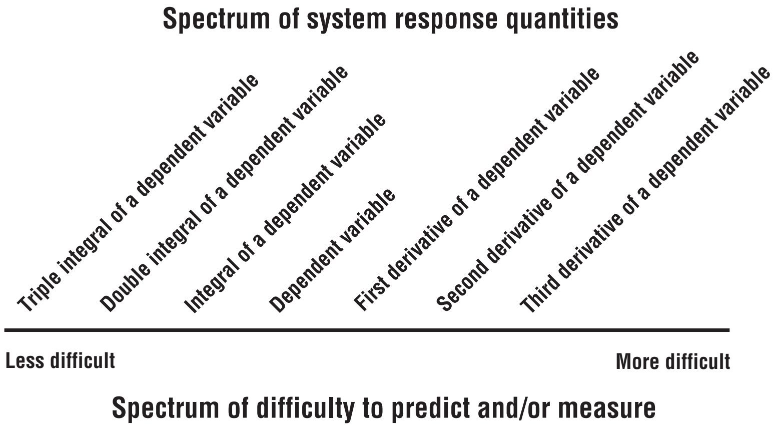

Spectrum of system response quantities
Triple integral of a dependent variable
Double integral of a dependent variable
Integral of a dependent variable
Dependent variable
First derivative of a dependent variable
Second derivative of a dependent variable
Third derivative of a dependent variable
Less difficult
More difficult
Spectrum of difficulty to predict and/or measure

Figure 10.2 Spectrum of various types of SRQs and the difficulty to predict and measure each.

measuring different SRQs. By *prediction difficulty* we mean aspects such as (a) the fidelity of the physics model that is required to accurately predict an SRQ; (b) the range of spatial and/or temporal scales exhibited by an SRQ; and (c) the spatial, temporal, and iterative convergence characteristics required to computationally resolve multiple physical scales and physical phenomena. By *experimental difficulty* we mean that there is a wide range of difficulty in measuring different SRQs. In experiments, this range of difficulty is primarily due to large differences in spatial and/or temporal scales either within an SRQ or across various SRQs. With modern digital electronic equipment, temporal scales do not generate as great a degree of difficulty as measurement of a wide range of spatial scales. Difficulty in experimental measurements commonly translates to increased experimental uncertainty for smaller spatial and temporal scales, both in bias and random uncertainty.

Figure 10.2 depicts the range of difficulty in predicting and measuring different SRQs. The spectrum could have a different type of ordering for some complex physics cases, but the conceptual ordering of interest here is shown in the figure. The scale of difficulty is ordered in terms of derivatives and integrals of the dependent variables in the partial differential equations (PDEs) of interest. For example, consider steady-state heat conduction through a homogenous solid with constant thermal conductivity. The PDE for the temperature distribution through a two-dimensional solid is given by Laplace's equations,

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0. \tag{10.1}$$

The dependent variable in the PDE is the temperature $T(x, y)$. This quantity is represented in the middle of the range in Figure 10.2. If one is interested in predicting and measuring the heat flux through a vertical line at $x =$ constant, the heat flux $q$ as a function of $y$ is given by

$$q(y) = -k \left( \frac{\partial T}{\partial x} \right)_{x=\text{constant}}. \tag{10.2}$$

From this equation it is seen that the heat flux is an SRQ that is dependent of the first derivative of the dependent variable in the PDE. This quantity is represented just to the right of the dependent variable in Figure 10.2. Various types of integral of dependent variables are also of interest as SRQs. These are referred to as functionals because they operate on a function and produce a real number. These are shown to the left of the middle in Figure 10.2 and they are ordered in terms of the number of integrals of the dependent variable of the PDE.

In validation activities, computational analysts will commonly compare their computational results with experimental measurements at one level of difficulty shown in the spectrum in Figure 10.2 and then claim the model accurate at all levels in the spectrum. Because there exists a spectrum of difficulty in prediction, a validation claim at one level of difficulty does *not necessarily* translate into accuracy at higher levels of difficulty. For example, if one were to show good agreement for the temperature distribution through a solid, it is more demanding of the model to accurately predict the heat flux because of the derivative operator. However, demonstrated accuracy at high levels of difficulty *does imply* accuracy at lower levels of predictive difficulty because of the integral operator. For example, if heat flux can be accurately predicted over the entire domain of the PDE, then the temperature distribution would be expected to be at least as accurate as the heat flux prediction.

#### *10.1.2.2 Project-oriented validation*

*Project-oriented validation* is similar to scientific validation in that one is still interested in a quantitative assessment of model accuracy as measured with respect to experimental data. However, in project-oriented validation, the focus shifts to how validation activities contribute to assessing the predictive accuracy of the model when the model is applied to the specific system of interest. Project-oriented validation is interested in evaluating the model accuracy with project-relevant experimental data, in addition to where predictions are needed in the application domain. As discussed in Chapter 2, the validation domain may or may not overlap with the application domain. For regions of overlap, interpolation is required to estimate the model accuracy for points away from the experimental data. For nonoverlapping regions, extrapolation of model accuracy is required for the application domain. (See Figure 2.10 for a discussion of overlapping and nonoverlapping regions of the validation and application domains.) Chapter 13, Predictive capability, will deal with the issue of extrapolation of model accuracy using validation metrics and alternative plausible models.

During the last decade, large scientific computing projects, such as the US National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) program, have developed methods to improve the planning and prioritization of project-oriented validation experiments (Pilch *et al.*, 2001; Trucano *et al.*, 2002; Oberkampf *et al.*, 2004; Pilch *et al.*, 2004). Improved planning and prioritization were found to be critical for managing the allocation of time, money, facilities, and talent resources to optimally advance

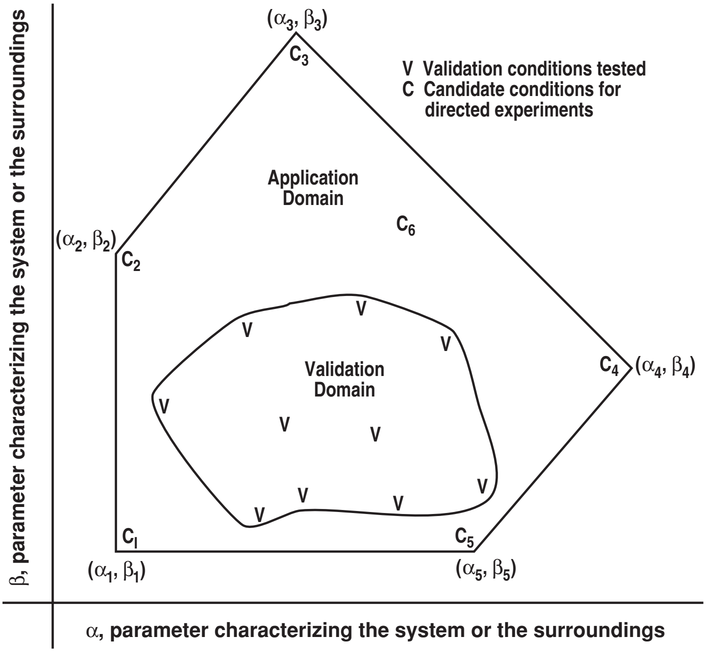

$(\alpha_3, \beta_3)$
$C_3$
V Validation conditions tested
C Candidate conditions for directed experiments
Application Domain
$C_6$
$(\alpha_2, \beta_2)$
$C_2$
Validation Domain
V
$C_4$
$(\alpha_4, \beta_4)$
$C_l$
$C_5$
$(\alpha_1, \beta_1)$
$(\alpha_5, \beta_5)$
β, parameter characterizing the system or the surroundings
α, parameter characterizing the system or the surroundings

Figure 10.3 Application domain, validation domain, and candidates for directed validation experiments (adapted from Trucano *et al.*, 2002).

the predictive capability of a project-oriented computational effort. Project-oriented validation experiments, referred to here simply as *directed experiments*, are those designed purposefully with specific goals linked to the project or the objectives of the system of interest. Occasionally, experimental data directly relevant to the system of interest can be found in the published technical literature or proprietary company reports, but this is rare because of the specificity of project needs. Because directed experiments are required to allow quantitative comparisons of computations with experimental data, important requirements are placed on directed experiments to create the greatest opportunities for performing these comparisons. In addition, directed experiments should be designed to assist, if needed, computational analysts and model builders to understand why the model may have performed poorly.

Figure 10.3 shows a two-dimensional space defined by the two parameters $\alpha$ and $\beta$, each one characterizing some feature of the system or the conditions imposed by the surroundings. The validation domain is shown as the region in which various validation experiments have been conducted, denoted by V. The application domain shows the region of interest, from a project perspective, in which the predictive capability is needed. As is typical of operating conditions of a system, the corners of the operating envelope are specified in terms of pairs of coordinates $(\alpha_i, \beta_i)$, $i = 1, 2, \ldots, 5$. The relationship between the application domain and the validation domain shown in Figure 10.3 is part of the class of relationships shown in Figure 2.10b in Chapter 2. That is, the application domain is *not* a subset of the validation domain, but there is overlap between them.

Presume that a validation metric result has been computed at each of the conditions marked with a V. The boundary of the validation domain would represent the apparent limit where the model accuracy has been assessed. The validation metric result can be thought of as the model error $E$ interpolated over the validation domain, $E(\alpha, \beta)$. The application domain for the system of interest is shown as the polygon. Let the conditions $\mathrm{C}_i$, $i = 1, 2, \ldots, 6$, denote the points in the parameter space that are candidates for future directed experiments. The coordinates of $\mathrm{C}_i$, $i = 1, 2, \ldots, 5$ usually correspond to the corners of the operating envelope of the system, $(\alpha_i, \beta_i)$, $i = 1, 2, \ldots, 5$. For each of these five conditions, an estimate should be made of the accuracy of the model. In addition, an estimate of the model accuracy should be made over the entire application domain. Estimation of model accuracy over the validation and application domains requires interpolation and extrapolation of the model and its observed accuracy, which will be discussed in Chapter 13. The condition corresponding to $C_6$ is shown to suggest that this condition was found to have the largest estimated model error over the application domain. The largest inaccuracies usually occur on the boundaries of the application domain, but this need not be the case because of complex physics interactions that can occur.

As is typical of project-oriented validation, system designers and project managers usually try to jump from the question of model accuracy assessment to the question of required accuracy of the model for their system. This is understandable because they are usually focused on their system, but the distinction between model accuracy assessment and accuracy requirements should always be kept in mind. Comparing the estimated model accuracy with the accuracy requirements of the project will help guide where future validation experiments may be needed. If there are multiple conditions for potential validation experiments, then these must be prioritized in some way. Chapter 14, Planning and prioritization in modeling and simulation, provides a detailed discussion of various methods for planning and prioritization of not only directed experiments, but also other activities of scientific computing.

The mathematical model should not only be applied in the decision of what conditions should be used in possible directed experiments, but also in other key factors in the design of these experiments. Some examples are (a) determination of key geometric features of the component or system to be tested, (b) guidance in determining the BCs and ICs, (c) guidance about where to locate diagnostic instrumentation, and (d) estimation of what magnitudes of SRQs could be expected for different sensors so that the proper sensors can be employed. The overarching theme of the design of the experiment is achieving system characteristics and responses that are of interest to the application driver. For example, if the successful operation of the system required that a certain design feature not fail, then special instrumentation would be designed to be certain this feature was captured during the experiment.

In a directed experiment, there must be a balance between the project needs and the needs of a high-quality validation experiment. As mentioned above, the primary purpose of

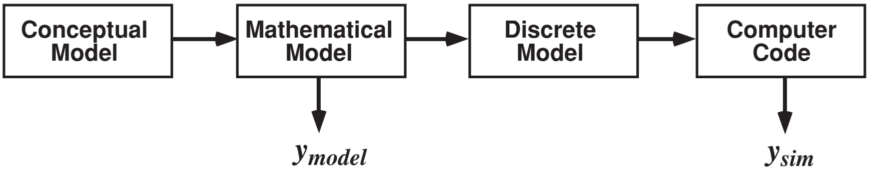


Figure 10.4 Sequence of model mappings to produce a simulation result.

a validation experiment is determining the predictive accuracy of a model. In engineering applications, there is usually some level of conflict between the goals of the project and the goals of a validation experiment. This tension is exacerbated if the project is funding the design, execution, and analysis of the directed experiment. For example, in the design of the directed experiment, the project leader typically would like the system geometry and functionality of the hardware to be similar to the actual system being designed. The computational analysts, however, would like the experiment to be as focused as much as possible on the physics processes that are of concern in the modeling. The viewpoint of the experimentalists also enters the discussion because their goals are to obtain the highest accuracy measurements possible, given the time and resource available. There are no straightforward cures to this three-way tension. The only general recommendation that can be made is to openly discuss the logic and priorities of each perspective so that a reasonable compromise can be attained. Our experience is that the project perspective commonly dominates any debate concerning design trade-offs of the validation experiment to the detriment of the validation goals of the experiment. Project managers who have personal experience with scientific computing, as well as an understanding of the role of scientific computing in helping the system achieve its performance goals, can provide enlightened decision-making on needed compromises. Chapter 11 will discuss these types of issue in more detail.

### *10.1.3 Sources of error in experiments and simulations*

We now describe the fundamental sources of error in experimental measurements and scientific computing. For this discussion it is more useful to consider errors instead of uncertainties because we begin by using the definition of error in a quantity, either experimental or computational. This discussion expands on the development in Oberkampf *et al.* (2004). Let $y_{\text{sim}}$ be an SRQ that results from a computational simulation. As discussed in Section 3.4, $y_{\text{sim}}$ is a final result of the various mappings shown in Figure 10.4. Let $y_{\text{nature}}$ be the true value of the SRQ from nature that, of course, can never be known exactly. Using the common definition of error discussed in Section 2.4, we define the error in the simulation, $E_{\text{sim}}$, as

$$E_{\text{sim}} = y_{\text{sim}} - y_{\text{nature}}. \tag{10.3}$$

Both terms on the right side of Eq. (10.3) can be separated into additional terms in order to explicitly identify various contributors to error. We rewrite Eq. (10.3) as

$$E_{\text{sim}} = (y_{\text{sim}} - y_{\text{Pcomputer}}) + (y_{\text{Pcomputer}} - y_{\text{model}}) + (y_{\text{model}} - y_{\text{exp}}) + (y_{\text{exp}} - y_{\text{nature}}). \tag{10.4}$$

$y_{\text{Pcomputer}}$ is the SRQ that could be theoretically computed on a perfect computer with infinite speed, precision, and memory such that we are able to take the limit as the discretization error and iterative error approach zero. Note that the same mathematical model, the same algorithms, and the same computer code used in $y_{\text{sim}}$ are also used in $y_{\text{Pcomputer}}$. $y_{\text{model}}$ is the SRQ resulting from the exact solution to the mathematical model; i.e., the model given by the PDEs, BCs, ICs, system excitation, geometric features, and all other input data needed for the simulation (Figure 10.4). $y_{\text{exp}}$ is the value of the SRQ that is measured in an experiment.

A more compact form of Eq. (10.4) can be written as

$$E_{\text{sim}} = E_1 + E_2 + E_3 + E_4, \tag{10.5}$$

where

$$\begin{aligned} E_1 &= (y_{\text{sim}} - y_{\text{Pcomputer}}), \\ E_2 &= (y_{\text{Pcomputer}} - y_{\text{model}}), \\ E_3 &= (y_{\text{model}} - y_{\text{exp}}), \\ E_4 &= (y_{\text{exp}} - y_{\text{nature}}). \end{aligned} \tag{10.6}$$

$E_1$ through $E_4$ represent all of the fundamental sources of error in a comparison of a computational result and an experimental measurement. The only two quantities in Eq. (10.5) that are always known are $y_{\text{sim}}$ and $y_{\text{exp}}$. Writing the simulation error in this way clearly demonstrates the possibility of error cancellation in the sum of the errors resulting in $E_{\text{sim}} = 0$. When model calibration is introduced, it is seen that error cancellation becomes the *goal*. For special cases where $y_{\text{model}}$ is known, it is seen that $y_{\text{model}}$ provides an important benchmark in the midst of this featureless landscape. The processes of verification and validation attempt to estimate each error contributor so that increased confidence can be gained concerning the magnitude of the sum. These processes are depicted in Figure 10.5 with respect to each of the error terms identified in Eq. (10.5). These terms will now be discussed.

$E_1$ represents all numerical errors resulting from the difference between the discrete solution, $y_{\text{sim}}$, (obtained using a finite discretization size, finite iterative convergence, and finite precision computer) and the exact solution to the discrete equations obtained on a perfect computer as the discretization size approaches zero, $y_{\text{Pcomputer}}$. $E_1$ is referred to as the solution or calculation error and is estimated by solution verification procedures. Chapters 7 and 8 discussed a number of methods for estimating the magnitude of the solution error. In the limit as the discretization size approaches zero, we still use the same numerical algorithms and computer code for $y_{\text{Pcomputer}}$ as is used for $y_{\text{sim}}$. If the numerical algorithm is deficient in some way, or the computer code contains programming errors,

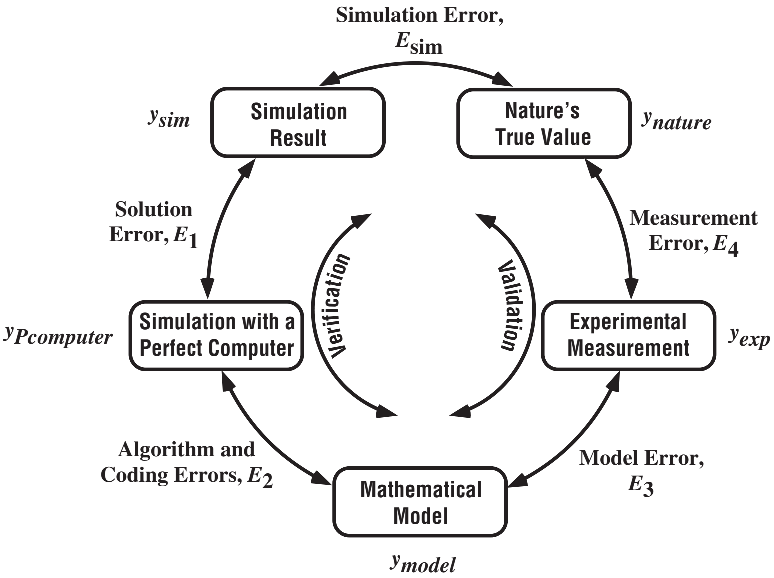


Figure 10.5 Error sources and verification and validation.

$E_1$ would be contingent on these errors but they would not technically be present in $E_1$. For example, suppose an algorithm error or a coding mistake was present in the computer code. Further, suppose that these errors caused the numerical solution to converge to an incorrect solution. $E_1$ would still represent the solution error due to a finite discretization size, finite iterative convergence, and a finite precision computer, exclusive of the algorithm and coding errors.

$E_2$ represents all errors resulting from the difference between the exact solution to the discrete equations as the discretization size approaches zero, $y_{\text{Pcomputer}}$, and the exact solution of the mathematical model, $y_{\text{model}}$. These errors are due to algorithm and coding errors and they are addressed by code verification procedures. Chapters 4 through 6 discussed a number of methods for detecting and removing these errors. Since we are not typically able to compute the exact solution of the discrete equations in the limit, we must rely on a systematic mesh and time-step convergence study with highly converged iterative solutions. To briefly summarize, the strategy is to compute the observed order of accuracy in the asymptotic region obtained during the convergence study. If an algorithm error (or deficiency) or a coding error causes an unexpected observed order of accuracy, we can be assured something is amiss. However, the reverse is not true, i.e., if the observed order of error matches the expected order of error it is not a proof that the algorithms are perfect and the coding is without error. The key to computing the observed order of accuracy is the availability of an exact solution to the mathematical model. The most demanding exact solutions have been found to be manufactured solutions, i.e., solutions obtained by picking $y_{\text{sim}}$ and manufacturing the mathematical model that reproduces $y_{\text{sim}}$.

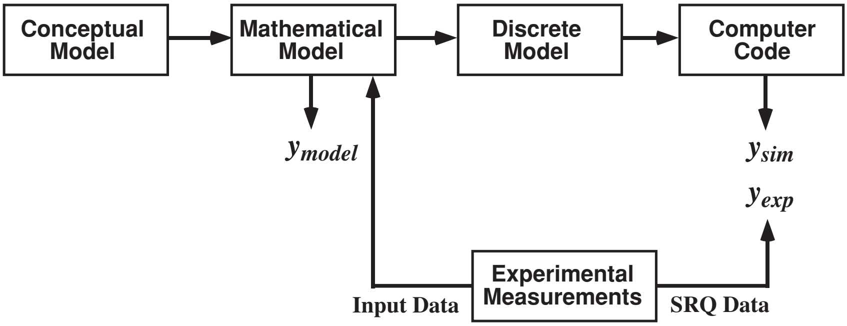


Figure 10.6 Model mappings with experimental data.

$E_3$ represents all errors resulting from the difference between the exact solution to the mathematical model, $y_{\text{Pcomputer}}$, and the experimental measurement, $y_{\text{exp}}$. $E_3$ is referred to as the model error or model form error although, as we will discuss, it is convolved with experimental measurement error. Estimating model error is conceptually and mathematically more difficult that estimating $E_1$ and $E_2$ for two reasons. First, experimental measurements always contain both random and systematic uncertainty since the true value from nature, $y_{\text{nature}}$, is never known. Examples of random and systematic experimental measurement uncertainties are multiple measurements of a quantity when the same experiment is repeated and improper calibration of a measurement instrument, respectively. Second, the experimental measurement provides not only $y_{\text{exp}}$, but also the necessary input data for the mathematical model; e.g., BCs, ICs, system excitation, and geometric features (see Figure 10.6). As a result, the mathematical model now depends on uncertain experimental data. It can be seen from Figure 10.6 that if unknown or poorly known parameters in the model are considered as adjustable parameters, then model calibration can be seen as a closed feedback loop. If one is also interested in quantifying model error during the model calibration process, model error has become inextricably convolved with calibration.

There are two different extremes when convolving model error and experimental error. First, suppose the experimental uncertainty is perfectly characterized and assume the systematic uncertainty is zero. This means that both the input data needed for the mathematical model from the experiment and $y_{\text{exp}}$ are perfectly characterized as purely aleatory uncertainties. Then the uncertainty in $y_{\text{sim}}$ can be characterized as a purely aleatory uncertainty, for example using Monte Carlo sampling of the mathematical model. Then one can quantitatively compare $y_{\text{sim}}$ and $y_{\text{exp}}$ using a validation metric operator that computes the difference between two probability distributions. Even though the experimental uncertainty is convolved with the model error, the model error can still be separately quantified in the validation metric result.

Second, suppose the experimental uncertainty is not perfectly characterized. For example, suppose very few experimental samples are obtained for an uncertain input quantity, or the experimentalist simply did not measure some of the needed input quantities. One can either characterize this lack of knowledge as a probability box (p-box) or one can use the poor characterization as flexibility in the mathematical model to calibrate model parameters. When using a validation metric operator, the first route will result in epistemic uncertainty in quantifying model error, while the second route will inextricably convolve model error and experimental uncertainty. Chapters 12 and 13 will provide an in-depth discussion of these issues.

$E_4$ represents all errors due to the difference between the true, but unknown value of nature, $y_{\mathrm{nature}}$, and the measurement of a physical quantity, $y_{\mathrm{exp}}$. The true value from nature can be viewed as either a deterministic quantity or a nondeterministic quantity. If $y_{\mathrm{nature}}$ is considered as a deterministic quantity, then we are saying that a fixed set of physical conditions is exactly reproduced in an experiment, resulting in the same physical quantity from nature. Only in very simple experiments, such as measurement of the mass of a fixed object, can $y_{\mathrm{nature}}$ be considered as an unknown deterministic quantity. In general, this viewpoint is not very constructive because in most experiments there are uncontrollable factors that change the measured quantity of interest. It is more useful to consider $y_{\mathrm{nature}}$ as a nondeterministic quantity because exactly the same physical conditions cannot be reproduced from one physical realization to the next. For example, even in a very well-controlled experiment using the same physical system, it is common to have slightly different BCs, ICs, or system excitation. $y_{\mathrm{exp}}$ is always nondeterministic because of random and systematic uncertainties in the experimental measurement. As a result, only in special situations is $E_4$ considered a fixed but unknown error, as opposed to an uncertain quantity composed of both aleatory and epistemic uncertainties.

### *10.1.4 Validation using data from traditional experiments*

Most researchers in the field of validation methodology have learned, often the hard way, why the various goals, strategies, and procedures discussed here are fundamentally constructive and necessary for assessment of model accuracy. However, there is commonly significant resistance to implementing these strategies and procedures in practice. Often the resistance is simply due to the inertia of human nature to respond to change. Sometimes, technical or practical arguments are presented as to why validation experiments, as described here, should not be embarked on. Practical arguments against conducting new validation experiments typically center on schedule and money constraints of the underlying project. The resistance is usually voiced, in its most low-key form, as: “We have been collecting experimental data for decades on similar systems, why can’t you use existing data to validate your models?” A technically sound and defensible answer to this recurring question is so important that we will provide a detailed discussion of possible responses. These responses are a compilation of experiences from several researchers (Marvin, 1995; Porter, 1996; Aeschliman and Oberkampf, 1998; Barber, 1998; Rizzi and Vos, 1998; Oberkampf and Trucano, 2002; Trucano *et al.*, 2002; Oberkampf *et al.*, 2004; Oberkampf and Trucano, 2008). Although these responses do not apply in some situations, they give the reader an idea of typical difficulties in using traditional experiments for validation.

The most common reason for not being able to use existing experimental data for validation is that important information needed for defining input for the simulation is not available or documented as part of the description of the experiment. The general types of information not documented, or poorly quantified, are system characteristics, BCs, ICs, and system excitation. Examples of system characteristics not documented are (a) mechanical, electrical, thermal, chemical, magnetic, optical, acoustical, radiological, and atomic properties of materials or components; (b) spatial distribution of these characteristics throughout the system, if this is needed as input data; (c) detailed geometric features of the system, such as material gaps and as-tested geometric inspection data; and (d) system assembly details, such as the preload torque on bolts and information of friction-fit assemblies. Examples of BCs are (a) Dirichlet, Neumann, Robin, mixed, periodic, and Cauchy, and (b) any information needed by the PDEs concerning how the surroundings affect the domain of the PDEs, including possible time dependence of the BCs. Examples of ICs are (a) knowledge of the spatial distribution of all dependent variables of the PDEs over the domain of interest, and (b) knowledge of all the required temporal derivatives of the dependent variables over the domain of the PDE. And finally, examples of system excitation are (a) knowledge of the excitation over the spatial and/or temporal domain of the PDEs, (b) knowledge of how the excitation may change over time, and (c) knowledge of the excitation when the system is deformed or in a damaged state due to response to the excitation.

The level of characterization of the input information can range from: (a) a precisely known (deterministic) value or function; (b) a precisely known random variable, i.e., a pure aleatory uncertainty; (c) a random variable characterized by a family of probability distributions, but the parameters of the distribution are imprecisely known; (d) expert opinion that characterizes the uncertain quantity by an interval, i.e., a pure epistemic uncertainty; or (e) an opinion of the form “my best recollection is.” The causes of the poor information conditions can range from (a) quantitative information was recorded, documented, and archived, but no one can find it now; to (b) the person who knew all about the experiment has retired and is skiing in Idaho. Of course, not all needed input information is important to the prediction of the SRQs of interest. The poorer the knowledge of important input information, the poorer the ability to quantitatively assess the accuracy of the model.

As an example, suppose a traditional experiment was conducted and all details of the experiment were well documented, including all input data needed for a simulation. As part of the documentation of the experiment, suppose all input data needed for the solution of the PDEs was deterministic and it was exactly known, save one parameter. That parameter was only known to be in a specified interval with no likelihood information known. Further, suppose that there is only one SRQ of interest, and that in the experiment it was perfectly measured, i.e., the experimental measurement uncertainty was zero. Using all of the information from the experiment, a nondeterministic simulation was computed because of the one interval-valued parameter. Then, the single SRQ of interest from the model would also be an interval-valued quantity. When a quantitative comparison is made between the computational and experimental results, one is comparing an interval from the computation with an exactly measured quantity from the experiment. If the interval is large, for example, because the input interval is very important in predicting the SRQ, then what can be concluded from the comparison? If the measurement value falls anywhere within the computational interval, one could say “That’s good.” However, very little can be quantitatively concluded concerning model accuracy, particularly if the interval is large.

The example just described is the *best result* that could occur in a validation exercise. What more commonly happens is that the computational analyst determines the value of the parameter within the interval that gives the best agreement with the experimental measurement, and then he/she declares the model validated. Nothing substantive is usually said concerning what was uncertain in the input data, nor how uncertain parameters were chosen. Our observation, and that of many other researchers in validation methodology, is that either: (a) the uncertainties in the simulation are so large because of missing information that nothing quantitative can be learned, or (b) uncertainties in the simulation due to missing input information are used as free parameters to optimize agreement between computation and experiment. The first result contributes little to assessing predictive accuracy, and the second result is misleading at best and fraudulent at worst.

The second most common deficiency in attempting to use data from traditional experiments for validation is that very few measurements of the SRQs of interest are made in experiments. With limited experimental data, only a limited number of statements can be made concerning quantitative accuracy of the model. Two types of situation commonly occur. First, experimental measurements of a local SRQ are made over a very limited portion of the domain of the PDE. For example: (a) flow field velocity measurements are made only over a very small region near a vehicle geometry of interest, (b) temperature measurements are made for a small number of points on the surface of an component in a heat transfer simulation, and (c) a small number of vibrational modes of a structure are measured over a portion of the structure. Second, experimental measurements are made of some SRQs of interest, but not the most important SRQs; for example: (a) surface pressure measurements are made on a vehicle, but the SRQ of interest is the predicted region of separated flow; (b) the strain is measured at various locations on the internal and external surface of a composite structure, but the SRQ of interest is delamination between layers of the structure; and (c) the material recession rate is measured on an ablating heat shield of a high-speed vehicle, but the SRQ of interest is the heat transfer rate to the heat shield. When little experimental data are available, the inferences concerning validation are weak. This deficiency can be magnified when the SRQ of interest is one or two derivatives removed from the experimentally measured quantity. (See the discussion in Section 10.1.2.1 concerning the spectrum shown in Figure 10.2.)

The third most common deficiency in trying to use traditional experiments in validation is that little or no estimation of experimental uncertainty is given for the SRQs measured. Even though uncertainty estimation has been a long tradition in experimental measurements, it is surprising how often measurements are reported without an uncertainty estimate. If an estimate of measurement uncertainty is provided, many times it underestimates the true uncertainty. Some reasons for a possible underestimate are (a) the uncertainty estimate provided is an estimate of the repeatability of the measurement, not an estimate of a broader class of random and systematic uncertainties in the experiment; (b) the estimate is simply a guess based on experience or what *appears* to look right; (c) systematic uncertainties due to particular diagnostic techniques, specific experimental procedures, or experimental facilities have not been quantified by conducting experiments using different diagnostics, procedures, or experimental facilities; and (d) only one experimental realization is conducted so that any estimate of uncertainty is based primarily on optimistic conjecture.

The common situation of underestimation of experimental measurement uncertainty was highlighted in a famous article by metrologist William Youden, "Enduring Values" (Youden, 1972), as well as by (Morgan and Henrion, 1990). In his paper Youden states: "Why do results obtained by different investigators characteristically disagree by more than would be expected by their estimates of uncertainty?" Youden notes that everything gets changed in another laboratory, whereas the investigator can (or does) make only minor changes within his/her laboratory. Youden gives two examples of systematic uncertainties in measurements of fundamental physical constants. He shows 15 values of the Astronomical Unit measured by investigators over the period 1895 to 1961. He points out that each investigator's reported value of uncertainty is *outside the limits* reported by his immediate predecessor. Youden also references the article by McNish (1962) concerning measurements of the speed of light. McNish (and Youden) show 24 measurements of the speed of light along with each investigator's estimate of their experimental uncertainty. The graph of the measurements shows that the estimated value (the experimenter's best estimate for the speed of light) is changing at a rate of about 1 km/s for every seven investigators who measure it! For the 24 investigators, the estimated value has changed 3.5 km/s, and half of the investigators estimated their experimental uncertainty as *under* 0.5 km/s. Youden points out that the problem is that the experimenters significantly underestimated unknown systematic uncertainties in their measurements. Even though Youden and other metrologists have called attention to the problem of underestimation of measurement uncertainty, human nature, saving face among investigators, and competition between commercial laboratories never change.

## 10.2 Validation experiment hierarchy

During the 1980s and 1990s, a number of researchers in fluid dynamics were struggling with the fundamental issue of: how should validation be undertaken for complex engineering systems? The *building block approach* that is used today was developed by Lin *et al.* (1992); Cosner (1995); Marvin (1995); Sindir *et al.* (1996); Sindir and Lynch (1997); and AIAA (1998). A similar hierarchical structure was developed in the nuclear reactor safety community during the same time frame for the purpose improving the understanding of

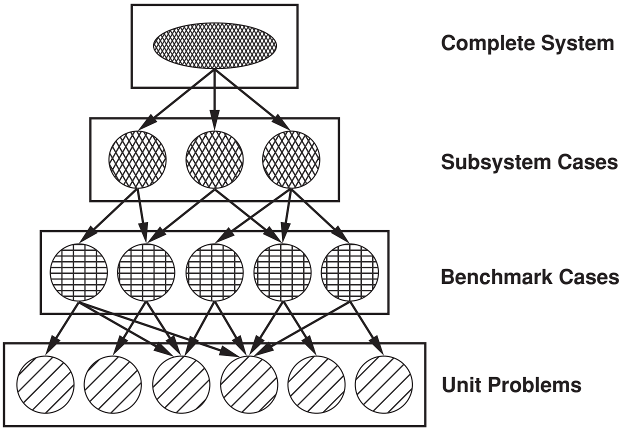


Figure 10.7 Validation hierarchy tiers (AIAA, 1998).

mass, momentum, and energy exchange in an accident environment (Zuber *et al.*, 1998). The building block or hierarchical approach is shown in Figure 10.7, as depicted in the AIAA Guide (AIAA, 1998). This approach divides the complex engineering system of interest into an arbitrary number of progressively simpler tiers. If the complex system is divided into three tiers, they are usually referred to as *subsystem*, *benchmark*, and *unit*. The strategy of the tiered approach encourages assessment of the accuracy of the model at multiple levels of complexity and physics coupling. The approach is clearly constructive in that it: (a) recognizes there is a hierarchy of complexity in real systems and simulations, and (b) recognizes that the quantity and accuracy of information that is obtained from experiments vary radically over the range of the hierarchy. The arrows from a higher tier to a lower tier indicate the primary impact of that element on a lower tier element.

The hierarchical view of validation is distinctly a system-oriented perspective. The system of interest is placed at the pinnacle of the validation hierarchy. The purpose of the validation hierarchy is to help identify a range of lower level tiers where experiments can be conducted for assessment of model accuracy for simpler systems and physics. A validation hierarchy is typically constructed by large-scale projects, with their system of interest as the focus. As one passes from the top tiers to the lower tiers of the hierarchy, the emphasis moves from system-oriented engineering to a physics-assessment orientation. Scientific validation, as discussed earlier, typically occurs at the lowest tier of the hierarchy. It should be pointed out that most validation hierarchies differ at the top tiers because the system or subsystems of interest are different, but they share many common elements at the lower tiers. For the common elements at the lower tiers, scientific validation can have a widespread impact on many projects.

Characteristics of each of the four tiers are discussed below. Additional tiers could be added to the validation hierarchy, but it would not significantly alter the discussion or the recommended methodology.

| Complete System Tier | |
|---|---|
| **Physical Characteristics** | **Measured Data for Validation** |
| Actual system hardware | Very limited measurement of model inputs |
| Actual geometry, materials, and features | Very limited measurement of model outputs |
| Complete physics and chemistry | Very few experimental realizations |
| Actual BCs, ICs, and system excitation | Little or no estimate of experimental uncertainty |

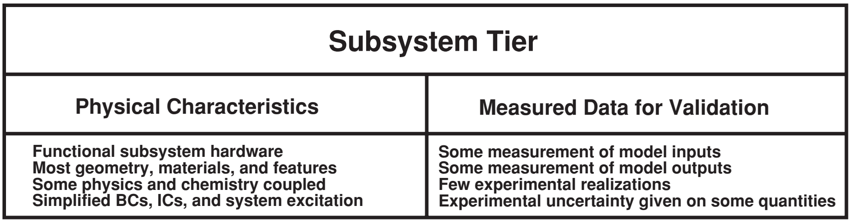
Figure 10.8 Characteristics of validation at the system tier (adapted from AIAA, 1998).

### *10.2.1 Characteristics of the complete system tier*

The *complete system tier* consists of the actual engineering hardware or system of interest (Figure 10.8). Thus, by definition, it is functioning hardware with all the geometric characteristics, materials, and features associated with manufacturing and assembly of the system. For typical complex engineering systems such as a gas turbine engine, multi-disciplinary, coupled physical phenomena occur in the system. The BCs, ICs, and system excitation would commonly correspond to those that are of interest for realistic operating conditions of the system. One may also be interested in poorly defined or poorly controlled abnormal or hostile environments.

Experimental data are measured on the engineering hardware under operating conditions. The quantity and quality of these measurements, however, are always very limited because: (a) the diagnostic and instrumentation systems must have minimal impact on operational systems, and (b) the test programs are typically conducted on a rigid schedule and with a tightly constrained budget. It is difficult, and sometimes impossible, for complex systems' tests to quantify even a small percentage of the input conditions required for computational simulation. As a result, little or no information is available for most of the input needed for simulation, including the uncertainty of the input quantities. Such tests generally provide only data that are related to engineering parameters of clear design interest, system functionality, and high-level system performance measures. Referring to Figure 10.2, these high-level measures would correspond to multiple integrals of dependent variables appearing in the PDEs of the model. Sometimes, the performance measures of a complete system test are simply: Did it work? Did it meet contract specifications? Did it fail safely?

Experimental data from complete systems are always specific to existing operational hardware and are available mainly through large-scale test programs. Existing data from these tests have traditionally focused on issues such as the functionality, performance, safety, or reliability of the system. Often, competition between alternative system designs underlies large-scale tests. If the competition is between outside organizations or suppliers of hardware, then the ability to obtain complete and unbiased information for validation becomes essentially impossible. The test programs typically require expensive ground-test facilities, full-scale flight testing, or testing under dangerous conditions, such as unpredictable weather, or abnormal or hostile environments. Also, there are certain situations where it is not possible to conduct a validation experiment of the complete system. Such



| Subsystem Tier | |
|---|---|
| **Physical Characteristics** | **Measured Data for Validation** |
| Functional subsystem hardware<br>Most geometry, materials, and features<br>Some physics and chemistry coupled<br>Simplified BCs, ICs, and system excitation | Some measurement of model inputs<br>Some measurement of model outputs<br>Few experimental realizations<br>Experimental uncertainty given on some quantities |

Figure 10.9 Characteristics of validation at the subsystem tier (adapted from AIAA, 1998).

situations could involve public safety or environmental safety hazards, unattainable experimental testing requirements, or international treaty restrictions.

### *10.2.2 Characteristics of the subsystem tier*

The *subsystem tier* represents the first decomposition of the actual system hardware into subsystems or subassemblies (Figure 10.9). Each of the subsystems or subassemblies is composed of actual functional hardware from the complete system. Subsystems usually exhibit three or more types of physics that are coupled. The physical processes of the complete system are partially represented by the subsystem tier, but the degree of coupling between various physical phenomena in the subsystem tier is typically reduced. For example, there is reduced coupling between subsystems as compared to the complete system. Most geometric characteristics are restricted to the particular subsystem and its attachment or simplified connection to the complete system. Essentially all of the materials, features, and capabilities of the subsystems are present. During subsystem testing, the BCs, ICs, and excitation are usually simplified compared to the operation of the complete system.

In subsystem tests there is commonly a significantly increased opportunity for all types of experimental measurement. There is usually much more willingness of test managers for installation of instrumentation and there is more interest in better understanding the details and operating conditions of the subsystem. In addition, there is usually less pressure from the project schedule, cost constraints, and high-level management attention. A much wider range of test facilities are possible for subsystems, with an accompanying improvement in control of the test conditions. There is an increased percentage of measured inputs and outputs for simulation, relative to complete system tests. There are commonly more experimental realizations than complete system tests. Experimental uncertainty estimates are given on some of the measured outputs, but few uncertainty estimates are commonly provided on measured input quantities.

### *10.2.3 Characteristics of the benchmark tier*

The *benchmark tier*, sometimes referred to as the component tier, represents the next level of decomposition and simplification beyond the subsystem tier (Figure 10.10). For the benchmark tier, special hardware is fabricated to represent the main features of each subsystem.

**Benchmark Tier**

| Physical Characteristics | Measured Data for Validation |
|---|---|
| Special, nonfunctional, hardware fabricated | Most model inputs measured |
| Simplified geometry, materials, and features | Many model outputs measured |
| Little coupling of physics and chemistry | Several experimental realizations |
| Very simple BCs, ICs, and system excitation | Experimental uncertainty given on most quantities |

Figure 10.10 Characteristics of validation at the benchmark tier (adapted from AIAA, 1998).

By special hardware, we mean hardware that is specially fabricated with simplified materials, properties, and features. For example, benchmark hardware is normally *not* functional or production hardware, nor is it fabricated with the same materials as actual subsystems. For benchmark cases, typically only two or three types of physical phenomenon are considered. The benchmark cases are normally simpler geometrically than those cases at the subsystem level. The only geometric features that are retained from the subsystem tier are those critical to the types of physical phenomenon that are considered at the benchmark tier. In addition, at this tier there is a distinct shift from project focused goals and schedules to those that are aimed at improved understanding of the physics involved and also the accuracy of the mathematical models that are used.

For this tier, most of the inputs needed for simulation are measured, or at least most of the important inputs are measured. Many of the model outputs that are measured correspond to dependent variables in the PDEs, or possibly one integral removed from the dependent variables. Since the benchmark tier uses nonfunctional hardware and special materials, the ability to instrument the hardware is significantly improved. Most of the experimental data obtained have associated estimates of measurement uncertainties. The experimental data, both model input data and output data, are usually documented with moderate detail. Examples of important experimental data that are documented include (a) detailed inspection of all hardware, (b) characterization of the variability of materials used in the experiment, (c) detailed information concerning assembly of the hardware, and (d) detailed measurement of BCs and excitation that were produced by the experimental apparatus or testing equipment.

### *10.2.4 Characteristics of the unit problem tier*

*Unit problems* represent the total decomposition of the complete system into isolated physical processes that are amenable to high-quality validation experiments (Figure 10.11). At this level, high precision, special-purpose hardware is fabricated and inspected. This hardware may only vaguely resemble some features of the subsystem or benchmark tier, especially in the view of the system project manager. Unit problems are characterized by very simple geometries that are accurately characterized. The geometry features are commonly two-dimensional, either planar or axisymmetric, or they can be very simple three-dimensional geometries with important geometric symmetry features. One element

**Unit Problem Tier**

| Physical Characteristics | Measured Data for Validation |
|---|---|
| Very simple, nonfunctional, hardware fabricated<br>Very simple geometry and features<br>No coupled physics<br>Very simple BCs, ICs, and system excitation | All model inputs measured<br>Most model outputs measured<br>Many experimental realizations<br>Experimental uncertainty given on all quantities |

Figure 10.11 Characteristics of validation at the unit problem tier (adapted from AIAA, 1998).

of complex physics is allowed to occur in each of the unit cases that are examined. The purpose of these cases is to isolate elements of complex physics so that critical evaluations of mathematical models or submodels can be evaluated. In fluid dynamics, for example, unit problems could individually involve (a) fluid turbulence for single phase flow, (b) fluid turbulence for two phase flow, (c) unsteady laminar flows, or (d) laminar diffusion flames. If one were interested in turbulent reacting flow, which combines turbulence and chemical reactions, it is recommended to consider this as one tier above the unit problem tier since it combines two aspects of complex physics.

For this tier, all of the important model inputs needed for simulation must be measured or well characterized. This, of course, is a highly demanding requirement for an experiment; a requirement that may not be attainable in many complex physics modeling situations. The following is a procedure that can be used to address this situation. First, experiments are conducted for the purpose of calibrating those parameters in the model that cannot be directly measured independently of the model. Second, follow-on experiments are conducted in which a few deterministic input parameters, i.e., accurately measured parameters with very small aleatory and epistemic uncertainty, are changed and the SRQs are remeasured. Third, new simulations are computed with the new values for the deterministic input parameters that were changed. And fourth, a validation metric result is computed by comparing the new computational and experimental results.

By changing a few deterministic input parameters, and *not* conducting any new calibrations, one can critically test the predictive capability of the model for new, well-defined conditions. This procedure is basically a two-step process; calibration and then validation on a closely related system. It requires a fairly large number of experiments for both steps of the process. A large number of experiments is needed in the calibration step so that very well characterized probability distributions can be determined for the calibrated parameters. In the second step, a large number of experiments is needed to precisely characterize the SRQs measured in the new experiment. With a correspondingly large number of nondeterministic simulations for both steps, a critical comparison can be made between the predicted and measured probability distributions for the SRQs.

As an example of this procedure, consider the prediction of the vibrational modes in a production lot of bolted-frame structures. First, a large number of bolted-frame structures are assembled using a fixed specification for the pre-load torque on all of the bolts in the structure. Experimental measurements are made on all of the structures to determine all of the vibrational modes of interest. Using a parameter optimization procedure, the estimated stiffness and damping in each of the bolted joints for all of the structures is determined using the mathematical model. Well-characterized probability distributions can then be computed for the stiffness and damping in all of the joints. The probability distributions represent the effect of variability in manufactured parts, the assembly of the structures, and the experimental measurement uncertainty.

Second, a large number of new structures are assembled such that the pre-load torque on all the bolts is changed. Suppose the torque on half of the bolts was doubled and half of the torques were decreased by a factor of two. The new structures are assembled using the same production lot of frame components and the same technicians assembling the structures. Experiments are conducted on the new structures to measure the vibrational modes of interest. A sufficient number of structures must be tested in order to construct a well-characterized probability distribution for the SRQs of interest. In addition, the new measurements are *not* shown to the computational analysts.

Third, simulations are computed using the new preload information on the bolt torques for the new structures. It is presumed that a submodel is included that can predict the change in stiffness and damping in joints due to bolt torque. Presuming a Monte Carlo sampling procedure is used, a sufficient number of simulations must be computed so that a well-characterized probability distribution can be computed for the SRQs of interest.

Fourth, using the new simulations and the new experimental data, a validation metric result is computed for SRQs of interest. The validation metric operator must be able to accept probability distributions as input from the simulation and the experiment. The validation metric operator computes the difference between the probability distribution from the simulation and the experiment for the SRQs of interest. The validation metric result is a quantitative measure of the predictive accuracy of the model. For example, if the metric result was small, then the predictive accuracy is high. Chapters 12 and 13 discuss combined calibration and validation procedures in more detail.

In the unit problem tier, highly instrumented, highly accurate experimental data are obtained and an extensive uncertainty analysis of the experimental data is conducted. If possible, experiments on unit problems should be repeated using different diagnostic techniques, and possibly in separate experimental facilities, to ensure that systematic uncertainties (bias) in the experimental data are quantified. These types of experiment could be conducted in any organization that has the experimental facilities, the technical expertise, *and* the willingness to embrace this type of critical evaluation of both the experiment and the computation. Commonly, these experiments are conducted at universities or research laboratories.

### 10.2.5 Construction of a validation hierarchy

The underlying concept of system engineering in the validation hierarchy is not new. What is new is the consistent theme of model accuracy assessment over the spectrum from full-scale engineering systems to experiments focusing on physics models at the lower levels of the hierarchy. Stated differently, hierarchical model validation is application driven, *not* physics driven. The construction of a hierarchical validation structure and the identification of the types of experiments that could be conducted at each tier are formidable challenges. There are many ways of constructing the tiers; no single construction is best for all cases. In addition, one validation hierarchy would not be appropriate for systems in different environments; for example, normal, abnormal, and hostile. In fact, there could be different validation hierarchies for different scenarios *within* each environment condition.

A good hierarchical tier construction is one that accomplishes two tasks. First, the construction carefully disassembles the complete system into tiers in which each lower-level tier has one fewer level of physical complexity. For complex engineered systems, this will require more than the four tiers shown in Figure 10.7. The types of physical complexity that could be uncoupled from one tier to the next are spatial dimensionality, temporal nature, geometric complexity, and physical process coupling, including multi-scale coupling. The most important of these types of physical complexity to decouple or segregate into separate effects experiments, from one tier to the next, is physical process coupling. The reason is that physical process coupling generally produces the highest nonlinear response due to the various contributors. It is important to recognize the nonlinear nature of all of the contributors in the construction of the tiers because the philosophy of the tier construction rests heavily on linear system thinking. That is, it is assumed that confidence in predictive capability for the complete system can be built from assessment of predictive accuracy of each of its parts. The complete system of interest clearly does not have to be linear, but the philosophy of the hierarchical validation approach loses some of its utility and strength when strong nonlinear coupling occurs from one tier to the next.

The second task accomplished by a good hierarchical tier construction is the selection of individual experiments in a tier that are practically attainable and able to produce validation-quality data. In other words, the individual experiments should be (a) practically achievable given the experimental test facilities, budget, and schedule, and (b) capable of producing quantitative experimental measurements of all the important input quantities and multiple SRQs that can critically assess the model. As discussed earlier, the ability to conduct a true validation experiment at the complete system tier is extremely difficult, if not impossible, for complex systems. At the subsystem tier, it is feasible to conduct validation experiments, but it is still quite difficult and expensive. One usually chooses a single hardware subsystem or group of subsystems that are closely related in terms of physical processes or functionality. For complex subsystems, one might want to add a new tier below subsystems called *subassemblies*. As with subsystems, this tier would consist of actual operational hardware.

When one defines the individual experiments at the benchmark-tier level, then special hardware, i.e., nonoperational, nonfunctional hardware would be fabricated. The benchmark tier is probably the most difficult to formulate because it represents the transition from a hardware focus in the two top tiers to a physics-based focus in the bottom tiers of the hierarchy. At the bottom tier, unit problems, one should identify simple geometry experiments that have a single element of physical process complexity. For high-quality

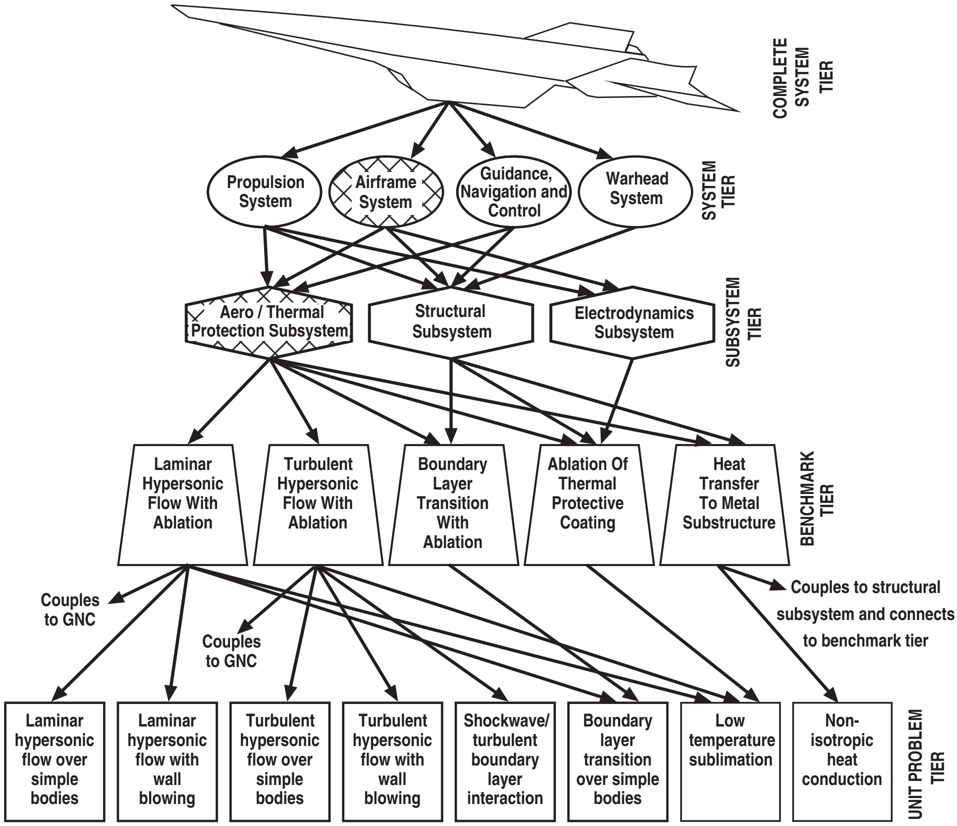

COMPLETE SYSTEM TIER
Propulsion System
Airframe System
Guidance, Navigation and Control
Warhead System
SYSTEM TIER
Aero / Thermal Protection Subsystem
Structural Subsystem
Electrodynamics Subsystem
SUBSYSTEM TIER
Laminar Hypersonic Flow With Ablation
Turbulent Hypersonic Flow With Ablation
Boundary Layer Transition With Ablation
Ablation Of Thermal Protective Coating
Heat Transfer To Metal Substructure
BENCHMARK TIER
Couples to GNC
Couples to GNC
Couples to structural subsystem and connects to benchmark tier
Laminar hypersonic flow over simple bodies
Laminar hypersonic flow with wall blowing
Turbulent hypersonic flow over simple bodies
Turbulent hypersonic flow with wall blowing
Shockwave/ turbulent boundary layer interaction
Boundary layer transition over simple bodies
Low temperature sublimation
Non-isotropic heat conduction
UNIT PROBLEM TIER

Figure 10.12 Example of a validation hierarchy for a hypersonic cruise missile (Oberkampf and Trucano, 2000).

validation experiments (a) one must be able to provide the necessary, highly characterized, input data to the simulation and (b) the experiments must be designed and conducted so that experimental uncertainty can be estimated in both inputs and outputs. High-quality validation experiments are practically attainable at the benchmark and unit-problem tiers, but they are not simple to conduct, nor inexpensive.

## 10.3 Example problem: hypersonic cruise missile

Here, we consider an example of a validation hierarchy for a hypersonic cruise missile (Oberkampf and Trucano, 2000). Assume the following characteristics of the missile system: (a) it is launched by an aircraft toward a ground target; (b) it is powered by a conventional gas turbine propulsion system, without a ducted fan; (c) it has an autonomous guidance, navigation, and control (GNC) system with an on-board millimeter-wave target seeker; and (d) it is functioning in a normal operating environment. Figure 10.12 shows a five-tier validation hierarchy for the hypersonic cruise missile that forms the basis of this discussion.

### 10.3.1 System tier

The entire missile is referred to as the *complete system* and the following as systems: propulsion, airframe, GNC, and warhead. These systems are as expected in the engineering design of such a vehicle. Additional elements could be added, however, depending on the perspective of the system engineer or the potential customer for the vehicle. The launch aircraft is not included at the system tier because its location in the hierarchy would be at the next higher tier, i.e., above the cruise missile. The hierarchical structure shown is not unique and it is not necessarily optimum for every system aspect that may be of interest, e.g., with regard to performance, reliability, or safety. In addition, the hierarchy shown emphasizes the airframe and aero/thermal protection subsystem (with cross-hatching), as will be discussed shortly.

### 10.3.2 Subsystem tier

At the subsystem tier, we have identified the following elements: aerodynamic performance and thermal protection, structural, and electrodynamics. The electrodynamics subsystem deals with the aspects of electromagnetic detectability of the cruise missile. This would range from radio frequencies used to detect the missile by radar, to detection in the visible spectrum. Only three elements are identified at the subsystem tier because they are the primary engineering design features that deal with the airframe system. Arrows drawn from the system-tier elements to the subsystem-tier elements indicate the primary elements that influence the subsystem tier. Recall at the subsystem tier that each element should be identified with functional hardware of the cruise missile. Consider how one would begin to conduct validation experiments at the subsystem tier depending on the computational discipline of interest. For example, the aero/thermal protection subsystem would contain subsystem hardware that is functionally related to the aerodynamic performance and heat transfer protection of any portion of the cruise missile. Some examples are (a) the as-manufactured thermal protective coating over the metal skin of the missile; (b) the as-manufactured metallic skin of the vehicle; (c) the metal or composite substructure under the skin of the vehicle; (d) all lifting surfaces, such as the vertical tail, and any control surfaces with their actuators; and (e) internal flow paths through the propulsion system, particularly the inlet and the exhaust ducts. However, the aero/thermal subsystem probably would not contain any other hardware inside the vehicle, unless some particular fluid flow path or heat conduction path was important. Note that there is commonly hardware that is in multiple subsystem-tier elements. For example, essentially all of the aero/thermal protection and the electrodynamics subsystem elements are also in the structural subsystem. The types of validation experiment and measurement, however, would be very different for each subsystem.

Suppose one were interested in validation experiments for the structural subsystem. If one were interested in the static deflection and stress of the various structural components, one would need to consider the mechanical coupling with nearby components and the aerodynamic loading, thermally induced loads, and any high-intensity acoustic excitation. For example, if one were interested in the deflection of the horizontal tail, the vertical tail and the nearby mechanical components of the fuselage would need to be included. If one were interested in the structural dynamic response of the various components, one would need to include essentially every piece of hardware from the missile because every part of the structure is dynamically coupled to every other part of the structure. Certain simplifications of the hardware, however, would be appropriate. For example, one could substitute mass-mockups for certain components, such as the completely functional propulsion system and the warhead, with little loss in fidelity. However, the structural dynamic and acoustic excitation caused by the propulsion system would be quite important in the validation of the structural subsystem.

### *10.3.3 Benchmark tier*

At the benchmark tier, Figure 10.12 shows only the elements that would be functionally related to the aero/thermal protection subsystem. Although additional benchmark-tier elements could be shown, only the following elements are identified: (a) laminar hypersonic flow with ablation, (b) turbulent hypersonic flow with ablation, (c) boundary-layer transition with ablation, (d) ablation of the thermal protective coating, and (e) heat transfer to the metal substructure. The arrows drawn from the structural and electrodynamics subsystems to the benchmark tier only show coupling from these two subsystems to elements depicted at the benchmark tier.

At the benchmark tier one fabricates specialized, nonfunctional hardware. For example, the laminar, turbulent, and boundary-layer-transition elements may not contain the as-manufactured ablative coating of the missile. Instead, a simpler material might be used – one that would produce wall blowing and possibly gases or particles that may react within the boundary layer, yet simpler than the typically complex gas and particle chemistry that results from actual ablative materials. The element for ablation of the thermal protective coating may use the actual material on the missile, but the validation experiment may be conducted, for example, at conditions that are attainable in arc-jet tunnels. The arrow from the structural subsystem to the boundary-layer-transition element is drawn to show that structural vibration modes of the vehicle surface can influence transition. An arrow is drawn from each of the elements for hypersonic flow with ablation that are marked "Couples to GNC." These arrows indicate a coupling of the boundary layer flow field to the millimeter-wave seeker in the GNC hierarchy (not shown here). The element for the heat-transfer-to-metal substructure shows an arrow that would connect to elements at the benchmark tier in the structural subsystem hierarchical tree. This arrow indicates the coupling that will result in thermally induced stresses and cause temperature-dependent material properties to be considered in the structural simulation.

### *10.3.4 Unit-problem tier*

The following elements are identified at the unit-problem tier: (a) laminar hypersonic flow over simple bodies, (b) laminar hypersonic flow with wall blowing, (c) turbulent hypersonic flow over simple bodies, (d) turbulent hypersonic flow with wall blowing, (e) shock wave/turbulent boundary layer interaction, (f) boundary layer transition over simple bodies, (g) low temperature sublimation, and (h) non-isotropic heat conduction. Many other elements could be identified at this tier, but these are representative of the types of validation experiment that should be conducted at the unit-problem tier. The identification of elements at this tier is easier than at the benchmark tier because unit-problem elements are more closely related to traditional mathematical model building experiments and model calibration experiments in fluid dynamics and heat transfer.

A point of clarification should be made concerning experiments at the lower tiers of the hierarchy, particularly at the unit-problem tier. Some researchers and system designers refer to experiments at the lower tiers, such as laminar hypersonic flow in a wind tunnel, as a *simulation* of the flight vehicle in the atmosphere. From the perspective of a project engineer interested in performance of the *real* vehicle, this view about experimentation is appropriate. From the perspective of conducting a validation experiment, however, this view only confuses the issue. That is, an experiment conducted at any tier is a physical realization of a process whose results can be used to assess the accuracy of a simulation. The relationship of the physical experiment to the performance of some engineering system is not the critical issue with regard to the *reality* of the validation experiment. *Any* experiment is reality. However, the project engineer who is interested in performance of the *real* vehicle may not appreciate how certain experiments relate to goals of his/her project.

### *10.3.5 Validation pyramid*

To explain better how the validation hierarchy of the airframe system is related to the validation hierarchy of the propulsion, GNC, and warhead systems, see Figure 10.13. The validation hierarchy of each of these four systems could be viewed as the primary facets of a four-sided pyramid. In the earlier discussion, the airframe facet was divided into three additional facets, each representing the three subsystems: aero/thermal protection, structural, and electrodynamics. The propulsion system could be divided into four additional facets to represent its subsystems: compressor, combustor, turbine, and thermal signature. Similarly, the GNC and the warhead systems could be divided into subsystems appropriate to each. On the surface of this multifaceted pyramid, one could more clearly and easily indicate the coupling from one facet to another. For example, we discussed the coupling of laminar and hypersonic flow with ablation to the millimeter-wave seeker of the GNC system. This coupling would be shown by an arrow connecting these hypersonic flow elements to appropriate elements on the GNC facet of the pyramid.

The validation pyramid stresses the system engineering viewpoint, as opposed to a scientific discipline viewpoint sometimes used in simulation-based design. Each facet of the

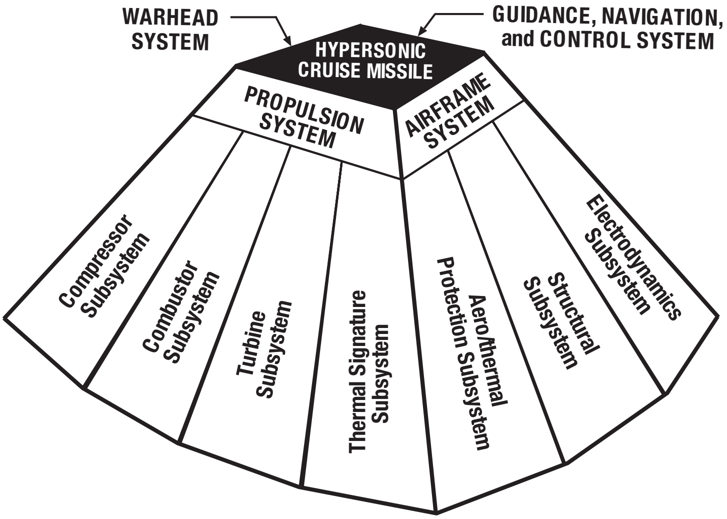

WARHEAD SYSTEM
GUIDANCE, NAVIGATION, and CONTROL SYSTEM
HYPERSONIC CRUISE MISSILE
PROPULSION SYSTEM
AIRFRAME SYSTEM
Compressor Subsystem
Combustor Subsystem
Turbine Subsystem
Thermal Signature Subsystem
Aero/thermal Protection Subsystem
Structural Subsystem
Electrodynamics Subsystem

Figure 10.13 Example of a validation pyramid for a hypersonic cruise missile (Oberkampf and Trucano, 2000).

pyramid can then be devoted to identifying validation experiments for each computational model responsible for part of the design of the system. As one traverses around the top of the pyramid, the number of facets is equal to the number of systems that are identified. For example, in the hypersonic cruise missile, this would total four. As one traverses around the bottom of the pyramid, the number of facets is equal to the total number of elements that have been identified around the entire pyramid at the unit-problem tier. For a complex system, the total number of facets would likely be larger than a hundred. For example, in the hypersonic cruise missile, we identified eight elements at the unit-problem tier that were primarily due to the aero/thermal protection subsystem. However, several additional elements could be identified at the unit problem tier; all primarily related to the aero/thermal protection subsystem. We strongly believe this type of system engineering thinking is necessary to increase the confidence in complex systems that are designed, manufactured, certified, and deployed with higher reliance on scientific computing and reduced levels of testing.

### *10.3.6 Final comments*

Two comments are in order concerning the construction of a validation hierarchy. First, the location of a particular validation experiment within the hierarchy must be determined relative to all of the surrounding elements in the hierarchy; i.e., it must be appropriately related to all of the experiments above it, below it, and in the same tier. Stated differently, the same validation experiment can be at different tiers for validation hierarchies that are constructed for different complex systems of interest. For example, the same turbulent-separated-flow experiment could be at the unit-problem tier in a complex system and at the benchmark tier in a simpler engineering system.

Second, a validation hierarchy is constructed for a particular engineered system operating under a particular class of operating conditions; for example, normal operating conditions. A new validation hierarchy would be constructed if one were interested in computationally analyzing other classes of system operating conditions or environments. Suppose one was interested in abnormal or hostile environments and the particular scenario of interest is that certain subsystems failed. Two examples are (a) certain weather environments or battle damage caused loss of a portion of the thermal protection system, and (b) certain electrical components of the GNC system failed or were damaged due to a microwave pulse of energy from defensive weapon systems. For these types of scenario, one would construct a different pyramid because different mathematical models would come into play.

## 10.4 Conceptual, technical, and practical difficulties of validation

### *10.4.1 Conceptual difficulties*

In Chapter 2, the philosophical issues underlying the concept of validation of models were touched on. It was pointed out that philosophers of science generally agree that theories and laws of nature can only be *disproved* or *failed to be disproved*. But, it was also stated that this perspective is unproductive and even debilitating for assessing the credibility of models in engineering and some natural science fields. The greatest debate over model validation seems to come from the field of hydrology, specifically surface water flow and subsurface water transport (Oreskes *et al.*, 1994; Chiles and Delfiner, 1999; Anderson and Bates, 2001; Morton and Suarez, 2001; Oreskes and Belitz, 2001). Hydrologists have well-justified concerns of validation concepts because of the nature of their models. Their models are dominated by parameters that are calibrated based on measurements of system responses. The calibrated parameters typically are not just a few scalar quantities, but also scalar fields in two and three dimensions, as well as tensor fields. As a result, there is astounding flexibility in hydrological models to match observations. From a conceptual viewpoint, one is still left with the question: can calibrated models be validated? We give two answers to the question, depending on the definition of validation one chooses.

First, suppose we use the restricted view of the term *model validation* as we have done throughout this book. That is, model validation refers to Aspect 1 in Figure 10.1 shown earlier in this chapter: assessment of model accuracy by comparison with experimental data. The unequivocal answer to the question would be *yes*. The accuracy of a model can be assessed regardless of whether the model has been calibrated or not. If the model uses very closely related experimental data (from a physics perspective) for accuracy assessment as used for calibration, however, the test of accuracy adds little value in the sense of new knowledge about the shortcomings of the model.

Second, we refer back to the discussion in Section 2.2.3 of Chapter 2 concerning the *encompassing view* of model validation. This view of validation includes all three aspects of validation as shown in Figure 10.1, above. First, assessment of model accuracy by comparison with experimental data; second, interpolation or extrapolation of the model to the intended use; and third, decision of model adequacy for the intended use. If one uses this view, as is done by the ASME Guide, one would also answer the question *yes*. Each of the three aspects of validation can be accomplished regardless of whether the model had been calibrated or not. That is, the restricted view and the encompassing view of the validation make *no presumptions* concerning *how* the model was built.

We argue that the more important question to be asked is: how can the predictive capability of a calibrated model be assessed? This question intuitively reflects the science-based perspective that is our foundation. Zeigler *et al.* (2000) discusses validation in terms of a sequence of more demanding requirements for the model. They define *replicative validity* to mean "for all of the experiments possible within the experimental frame, the behavior of the model and system agree within acceptable tolerance." By *predictive validation* they require "not only replicative validity, but also the ability to predict as yet unseen system behavior." By *structural validity* they require "that the model not only is capable of replicating the data observed from the system, but also mimics in step-by-step, component-by-component fashion the way in which the system does it transitions." Bossel (1994) touches on the same issues by defining two types of model: descriptive models and explanatory models. *Descriptive models* are those that can imitate the behavior of the system based on previous observations of the system. *Explanatory models* are those that represent a system's structure and its components, and their connections so that one can understand the future system behavior even under conditions never before experienced.

As discussed in Section 3.2.2, examples of descriptive models are regression and empirical models that relate the inputs to outputs. They can also be very sophisticated stochastic Markov chain Monte Carlo models to deal with nondeterministic features of not only the inputs and outputs, but also some presumed internal features of the system. Descriptive models require a great deal of data to gain confidence that the mathematical mapping of inputs to outputs is properly characterized. Explanatory models require a great deal of knowledge concerning the fundamental relationships and interactions occurring within the system. Input data are only needed to make specific predictions of outputs. The goal of science is clearly explanatory models because the strength of the model inference is built on detailed knowledge. Computational simulation of complex physical systems must live in the world between descriptive and explanatory models. We prefer to think of our models as scientific, but many times the reality is that they are more descriptive than we would like to admit.

We suggest a framework for providing a graded answer to the question of predictive capability, depending on the degree of calibration of the model. Figure 10.14 depicts the notional ability to assess the predictive capability of a model as a function of the number of free parameters in the model. *Free* parameters are those that *cannot* be independently measured separate from the model being assessed. As suggested in Figure 10.14, the ability to

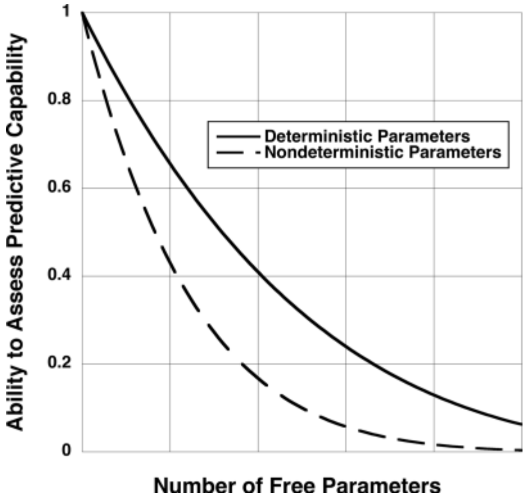


Figure 10.14 Ability to assess predictive capability as a function of the number of free parameters.

assess the model's predictive capability depends on whether the free parameters are deterministic (scalars) or nondeterministic (functions). If one is dealing with nondeterministic parameters, there is much greater flexibility in calibrating because one has probability distributions as opposed to just being numbers. The free parameters may have some physical justification based on knowledge of the processes occurring within the system, for example, they may be effective physical properties. However, their quantification is fundamentally dependent on the assumptions and mental constructs of the model, as well as the observed responses of the system. The numbers on the ordinate of the graph (Figure 10.14) are only notional in the following sense. First, the ability to assess a model is simply represented as a scale based on unity as a crisp *yes*, and zero as a crisp *no*. Second, the ability to assess a model does not only depend on the number of free parameters, but also on how sensitive the model is to the various parameters. For example, a model may have a hundred free parameters, but only five are important in the prediction of a given SRQ.

By answering the predictive capability question in this way, we suggest the following framework. First, the ability to assess a model's predictive capability rapidly decreases as the number of free parameters increases. For a model with a large number of free parameters, the calibrated parameters *become* the essence of the model, as opposed to the posited knowledge in the assumptions of the model. Hydrological models, for example, are clearly in this domain. The number of free parameters in a two- or three-dimensional scalar field typically depends on the spatial discretization that is chosen in the computational simulation. This number can total into the millions on modern supercomputers. Additionally, the free parameters are usually given by probability distributions, which gives additional flexibility to the model.

Second, most models in scientific computing are physics-based models. These models conserve mass, momentum, and energy so the associated conservation equations (which are typically PDEs) can be viewed as constraints on the calibration process. As a result, calibration of parameters in these models can be viewed as a PDE constrained optimization problem. When the number of free parameters becomes large, even in a physics-based model, the effect of the physics constraints becomes imperceptible. That is, a physics-based model becomes equivalent to a descriptive model that has little internal structure and is adapted to fit experimental observations of the system.

And third, our postulated framework gives no credit or loss for models that approach model accuracy assessment as a two-step process: calibration then validation. For example, suppose a deterministic model contained a handful of free parameters that were calibrated based on ten or twenty observations of the system. The parameters could be estimated using an optimization procedure to minimize some type of error measure based on the difference between model output and the observations. The two-step process is very useful, and often necessary, in complex physics modeling. But the framework argues that the ability to assess a model's predictive capability depends primarily on how many free parameters are available, not on the degree of independence between the calibration data and the validation data. Depending on the closeness of the calibration and validation data, as argued above and in Chapter 12, the ability to assess a model's predictive capability is additionally diminished.

As a result, the suggested framework does not address the issue of how demanding is the test of model accuracy assessment. For example, suppose a model has been calibrated with one set of observed system responses, and then the model accuracy is assessed with a very closely related set of observations. This test of model accuracy is very weak compared to a test where the model is asked to predict the system response for a substantially different set of conditions. The issue of rigorously understanding *how demanding* is a model accuracy test, or *how dissimilar* are new conditions from calibration conditions, is an open research topic. Section 14.5 discusses this issue in more detail.

### *10.4.2 Technical and practical difficulties*

There are situations where technical and practical difficulties in obtaining experimental measurements either hinder or eliminate the possibility of model validation. These difficulties can be generally grouped into (a) the technology is presently unavailable to make the measurements needed, (b) it is technically impossible to obtain the data regardless of the technology, and (c) it is impractical or not cost effective to obtain the measurements. Some examples where the technology is unavailable at the present time are the following. First, experimental data on hypervelocity impact is extremely limited in terms of spatial detail or the measurement of a variety of SRQs. Typical experimental results are photographs of the impact crater or hole through a specimen after the impact. In some facilities, high speed imaging of the penetration event is available. These data are useful in validation, but they greatly limit the ability to quantitatively assess the accuracy of various SRQs from the model. Second, the measurement of underground transport of substances, primarily due to water flow through porous media, is very limited. The typical procedure is to inject tracer substances at locations where wells have been drilled, then monitor the tracer concentration at nearby wells as a function of depth and time. The porosity and permeability of the surrounding subsurface material that appear in the PDEs are then adjusted so that the observed tracer concentration record is matched. As can be seen, this results in the solution of an inverse (calibration) problem. That is, given the observed output, what field characteristics should the system have for the assumed model to produce the observed output. Third, the ability of measuring both simulation input quantities and SRQs at the micrometer scale and smaller is very limited. As mathematical models continue to develop for spatial scales in these ranges, the ability to validate these models will be a pacing item. As a result, confidence in predictions at these spatial scales for materials science, biochemistry, and biophysics will be significantly impeded.

An example of a situation where it is conceptually impossible to obtain the needed experimental data for validation is in modeling physical phenomena with very long time scales, on the order of centuries or tens of centuries, or very large physical scales. Some examples are (a) long-term prediction of the underground storage of toxic or nuclear wastes; (b) long-term prediction of the effect of various contributors to global warming; and (c) response of the global environment to a large-scale atmospheric event, such as a volcanic eruption or the impact of a sizeable asteroid.

There are also situations where it is not cost effective, impractical, or not allowed to obtain experimental data for validation, even though experiments are technically feasible. Some of these are (a) conducting an experiment on the explosive failure of a full-scale reactor containment building, (b) conducting an experiment on the earthquake or explosive failure of a large-scale dam, (c) obtaining certain experimental data for the physiological response of humans to toxic chemicals or substances, and (d) hazardous or environmentally damaging tests that are banned by international treaties.

## 10.5 References

Aeschliman, D. P. and W. L. Oberkampf (1998). Experimental methodology for computational fluid dynamics code validation. *AIAA Journal*. **36**(5), 733–741.

AIAA (1998). *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*. AIAA-G-077–1998, Reston, VA, American Institute of Aeronautics and Astronautics.

Anderson, M. G. and P. D. Bates (2001). Hydrological science: model credibility and scientific integrity. In *Model Validation: Perspectives in Hydrological Science*. M. G. Anderson and P. D. Bates (eds.). New York, John Wiley.

Anderson, M. G. and P. D. Bates, eds (2001). *Model Validation: Perspectives in Hydrological Science*. New York, NY, John Wiley.

Balci, O., W. F. Ormsby, J. T. Carr, and S. D. Saadi (2000). Planning for verification, validation, and accreditation of modeling and simulation applications. *2000 Winter Simulation Conference*, Orlando FL, 829–839.

Barber, T. J. (1998). Role of code validation and certification in the design environment. *AIAA Journal*. **36**(5), 752–758.

Benek, J. A., E. M. Kraft, and R. F. Lauer (1998). Validation issues for engine–airframe integration. *AIAA Journal*. **36**(5), 759–764.

Bossel, H. (1994). *Modeling and Simulation*. 1st edn., Wellesley, MA, A. K. Peters.

Chiles, J.-P. and P. Delfiner (1999). *Geostatistics: Modeling Spatial Uncertainty*, New York, John Wiley.

Cosner, R. R. (1995). CFD validation requirements for technology transition. *26th AIAA Fluid Dynamics Conference*, AIAA Paper 95–2227, San Diego, CA, American Institute of Aeronautics and Astronautics.

Hornung, H. G. and A. E. Perry (1998). Personal communication.

Kleijnen, J. P. C. (1998). Experimental design for sensitivity analysis, optimization, and validation of simulation models. In *Handbook of Simulation: Principles, Methodology, Advances, Application, and Practice*. J. Banks (ed.). New York, John Wiley: 173–223.

Kleindorfer, G. B., L. O'Neill, and R. Ganeshan (1998). Validation in simulation: various positions in the philosophy of science. *Management Science*. **44**(8), 1087–1099.

Lin, S. J., S. L. Barson, and M. M. Sindir (1992). Development of evaluation criteria and a procedure for assessing predictive capability and code performance. *Advanced Earth-to-Orbit Propulsion Technology Conference*, Marshall Space Flight Center, Huntsville, AL.

Marvin, J. G. (1995). Perspective on computational fluid dynamics validation. *AIAA Journal*. **33**(10), 1778–1787.

McNish, A. G. (1962). The speed of light. *Institute of Radio Engineers, Transactions on Instrumentation*. **I-11**(3–4), 138–148.

Morgan, M. G. and M. Henrion (1990). *Uncertainty: a Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. 1st edn., Cambridge, UK, Cambridge University Press.

Morton, A. and M. Suarez (2001). Kinds of models. In *Model Validation: Perspectives in Hydrological Science*. M. G. Anderson and P. D. Bates (eds.). New York, John Wiley.

Murray-Smith, D. J. (1998). Methods for the external validation of continuous systems simulation models: a review. *Mathematical and Computer Modelling of Dynamics Systems*. **4**, 5–31.

Oberkampf, W. L. and D. P. Aeschliman (1992). Joint computational/experimental aerodynamics research on a hypersonic vehicle: Part 1, Experimental results. *AIAA Journal*. **30**(8), 2000–2009.

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 T. G. Trucano (2000). Validation methodology in computational fluid dynamics. *Fluids 2000 Conference*, AIAA Paper 2000–2549, Denver, CO, American Institute of Aeronautics and Astronautics.

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. and T. G. Trucano (2007). *Verification and Validation Benchmarks*. SAND2007–0853, Albuquerque, NM, Sandia National Laboratories.

Oberkampf, W. L. and T. G. Trucano (2008). Verification and validation benchmarks. *Nuclear Engineering and Design*. **238**(3), 716–743.

Oberkampf, W. L., D. P. Aeschliman, J. F. Henfling, and D. E. Larson (1995). Surface pressure measurements for CFD code validation in hypersonic flow. *26th AIAA Fluid Dynamics Conference*, AIAA Paper 95–2273, San Diego, CA, American Institute of Aeronautics and Astronautics.

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.

Oreskes, N. and K. Belitz (2001). Philosophical issues in model assessment. In *Model Validation: Perspectives in Hydrological Science*. M. G. Anderson and P. D. Bates (eds.). New York, John Wiley.

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*. **263**, 641–646.

Pilch, M., T. G. Trucano, J. L. Moya, G. K. Froehlich, A. L. Hodges and D. E. Peercy (2001). *Guidelines for Sandia ASCI Verification and Validation Plans – Content and Format: Version 2*. SAND2000–3101, Albuquerque, NM, Sandia National Laboratories.

Pilch, M., T. G. Trucano, D. E. Peercy, A. L. Hodges, and G. K. Froehlich (2004). *Concepts for Stockpile Computing (OUO)*. SAND2004–2479 (Restricted Distribution, Official Use Only), Albuquerque, NM, Sandia National Laboratories.

Porter, J. L. (1996). A summary/overview of selected computational fluid dynamics (CFD) code validation/calibration activities. *27th AIAA Fluid Dynamics Conference*, AIAA Paper 96–2053, New Orleans, LA, American Institute of Aeronautics and Astronautics.

Refsgaard, J. C. (2000). Towards a formal approach to calibration and validation of models using spatial data. In *Spatial Patterns in Catchment Hydrology: Observations and Modelling*. R. Grayson and G. Bloschl (eds.). Cambridge, Cambridge University Press: 329–354.

Rizzi, A. and J. Vos (1998). Toward establishing credibility in computational fluid dynamics simulations. *AIAA Journal*. **36**(5), 668–675.

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

Rykiel, E. J. (1996). Testing ecological models: the meaning of validation. *Ecological Modelling*. **90**(3), 229–244.

Sargent, R. G. (1998). Verification and validation of simulation models. *1998 Winter Simulation Conference*, Washington, DC, 121–130.

Sindir, M. M., S. L. Barson, D. C. Chan, and W. H. Lin (1996). On the development and demonstration of a code validation process for industrial applications. *27th AIAA Fluid Dynamics Conference*, AIAA Paper 96–2032, New Orleans, LA, American Institute of Aeronautics and Astronautics.

Sindir, M. M. and E. D. Lynch (1997). Overview of the state-of-practice of computational fluid dynamics in advanced propulsion system design. *28th AIAA Fluid Dynamics Conference*, AIAA Paper 97–2124, Snowmass, CO, American Institute of Aeronautics and Astronautics.

Trucano, T. G., M. Pilch, and W. L. Oberkampf (2002). *General Concepts for Experimental Validation of ASCI Code Applications*. SAND2002–0341, Albuquerque, NM, Sandia National Laboratories.

Youden, W. J. (1972). Enduring values. *Technometrics*. **14**(1), 1–11.

Zeigler, B. P., H. Praehofer and T. G. Kim (2000). *Theory of Modeling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems*. 2nd edn., San Diego, CA, Academic Press.

Zuber, N., G. E. Wilson, M. Ishii, W. Wulff, B. E. Boyack, A. E. Dukler, P. Griffith, J. M. Healzer, R. E. Henry, J. R. Lehner, S. Levy, and F. J. Moody (1998). An integrated structure and scaling methodology for severe accident technical issue resolution: development of methodology. *Nuclear Engineering and Design*. **186**(1–2), 1–21.