Final version appeared as

Easterling, R.G. and Berger, J. O. (2002). Statistical foundations for the validation of computer models. In *Proceedings of the Workshop on Foundations for V&V in the 21st Century*, D. Pace and S. Stevenson (Eds.). Society for Modeling and Simulation International.

Statistical Foundations for the Validation of Computer Models

Robert G. Easterling*
Statistical Consultant
51 Avenida del Sol
Cedar Crest, NM 87008
505-286-8796
rgeaste@comcast.net

James O. Berger**
Duke University
PO Box 90251
Durham, NC 27708
919-684-4531
berger@stat.duke.edu

Abstract

Confidence in computational predictions is enhanced if the potential 'error' in these predictions (the difference between the prediction and nature's outcome in the situation being simulated) can be credibly bounded. The "model-validation" process by which experimental or field results are compared to computational predictions to produce this confidence provides the raw material for characterizing a computational model's predictive capability in terms of such error limits. In general, the goal is to evaluate predictive capability, first for predictions in the region of experimentation, then, if possible, for predictions in untested regions of applications. This whole process is fundamentally statistical because it requires the acquisition and careful analysis of appropriate data. We establish a statistical model for characterizing predictive-capability and discuss various experimental design and statistical data analysis issues and approaches for resolving them   Analyses based on both 'frequentist' and Bayesian statistical paradigms are discussed in general in this paper and illustrated in accompanying papers presented at this workshop.

[*] The work of the first author was supported by Sandia National Laboratories and the United States Department of Energy under Contract DE-AC04-97AL85000. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockhed Martin Company, for the United States Department of Energy.
**The work of the second author was supported by General Motors and the National Science Foundation, Grant DMS-0103265.

Table of Contents

Introduction

Computer models, of increasing sophistication, are being used increasingly in a wide variety of contexts to predict the outcomes of physical events. The credibility and utility of computational predictions requires meaningful answers to questions such as:

> *How well does the computer model represent reality?*
>
> *How well can the computer model predict reality under new, untried conditions?*

The process by which such questions are addressed is called model validation [AIAA 1998]. The answers to these questions provide an evaluation of a model's predictive capability. The model-validation process, it is generally recognized, involves the collection of field or experimental data and a comparison of those results to corresponding computational predictions of those outcomes. While this objective is easy to state, implementation raises a number of difficult issues that have only recently begun to be addressed. [See, e.g., Trucano, Pilch and Oberkampf 2002.] Many of these issues are statistical. The purpose of this paper is to discuss and illustrate the statistical foundations of model validation.

The validation process is fundamentally statistical. It involves the acquisition of data (from designed experiments, field experience, surveys, or sampling plans), statistical data analysis (to characterize systematic and random patterns in the experimental and computational results), and inference (characterizing, subject to and reflecting the amount and nature of the available data, the predictive capability of the model in unobserved situations). Such activities and analyses are not done in a vacuum. They must be closely tied to the scientific understanding of the process being computationally simulated.

a. The Process of Evaluating Predictive Capability

Figure 1 is a view of the process of answering the above questions, set in the context of comparing a computational prediction to a system requirement. The top ellipse in Fig. 1 depicts the intended use of the computational tool: system requirements specify various performance goals and the computational model will be used to predict system performance in scenarios that embody these requirements. Comparing the prediction to the requirement requires an uncertainty yardstick, or frame of reference, depicted by the uncertainty 'cloud' surrounding the prediction. To develop such a yardstick, experiments and computations must be conducted – depicted by the bottom ellipse. The design of these experiments should be driven by the system scenarios and the structure of the computational model. These experiments and computations provide first for an evaluation of prediction capability in the situations tested. Next, and most importantly, the ensemble of observed differences are potentially the basis of an inference about prediction uncertainty in the system applications of interest -- the connection to the upper ellipse.

Figure 1 provides a glimpse into the difficulty of the model-validation problem. It's not enough just to compare experiment and computation, where possible and perhaps convenient, and make a assessment of the degree of agreement. The inference bridge to the application has to be constructed. The distance between the two ellipses may make this scientifically difficult, difficult to justify and defend.

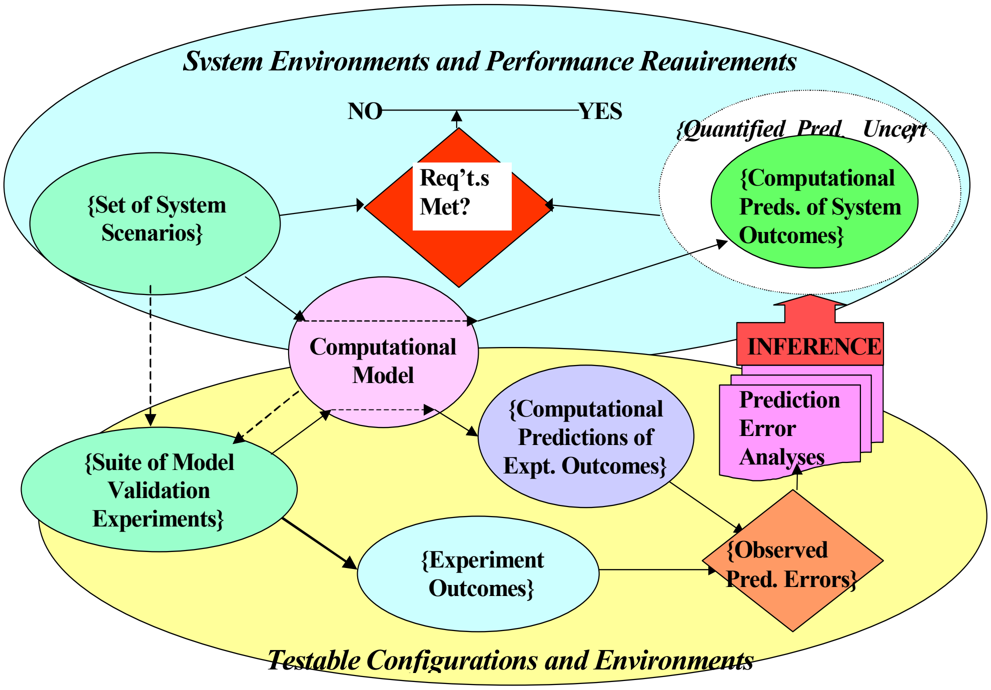
Svstem Environments and Performance Reauirements
NO
YES
{Quantified Pred. Uncert
Req't.s Met?
{Set of System Scenarios}
{Computational Preds. of System Outcomes}
Computational Model
INFERENCE
Prediction Error Analyses
{Computational Predictions of Expt. Outcomes}
{Suite of Model Validation Experiments}
{Experiment Outcomes}
{Observed Pred. Errors}
Testable Configurations and Environments

Figure 1. Schematic for Characterization of Prediction Capability

Though we will focus on the design and analysis of one set of model-validation experiments, the process in Fig. 1 is generally iterative. Both the experimental database and the computational model will evolve. There will be instances in which the analysis of observed prediction errors in the lower ellipse will detect flaws in either the computational model or the experimental procedures and data, so they will need to be corrected before the inferential loop is completed. There will also be situations in which the empirical and scientific bases are not adequate for the required inference to prediction error in the system application. What then?

There are several approaches to solving this dilemma. The experimentalist's solution would be to seek to expand the space of testable configurations and environments to be more "application-like," to move the bottom ellipse closer to the upper ellipse. An engineering solution would be to redesign the system to make it less susceptible to phenomena that are difficult to model. The modeler's solution might be either to develop a deeper model – put more physics into the model – or to simplify the model and replace difficult-to-validate components of the model by simplified, bounding sub-models. The program manager, who of course wants this process to end because of cost and schedule requirements, might seek to resolve the dilemma by convincing the customer to change the requirements, thereby moving the application ellipse closer to the testable space.

In spite of all these efforts, it must be admitted that in some situations we may not be able to develop credible, defensible statements of a computational model's predictive capability for the outcomes of system-application events. The statistical framework advanced in this paper will identify the gaps and obstacles to successful inference. Because of the importance of computational models, and the importance of characterizing their predictive-capability in some way, the resolution then may be some form of quantified informed opinion, such as: "In a wide variety of experimental contexts, we never saw a prediction error greater than 20%. The differences between application and experimental conditions, though, are substantial enough that we think an additional factor of 2x is reasonable. Thus, our judgment is that system outcomes can be predicted to within 40%." The experience and reputation underlying such statements will determine their credibility. Methods have been developed to enhance the credibility of quantified informed opinion, but such are not addressed in this paper.

b. Mathematical Framework

To frame this paper's discussion we mathematically represent a prediction generated by a computational model as:

$$y^M(x) = M(x{:}\varphi), \tag{1}$$

where $M(x{:}\varphi)$ represents the computational model of the phenomenon of interest; x is model input variables that define the event of interest, φ is model parameters; and y^M is the model output or *prediction*. All the terms in expression (1), namely x, φ, and y^M, could be vectors or fields. The distinction between x and φ is discussed in the next paragraph.

In general, the model's input vector x is a set of variables whose values define a physical entity and the environment to which it is subjected. This vector will include physical dimensions, materials, environmental variables, and initial and boundary conditions. For example, x could be the temperature to which a given material specimen is subjected. The numerical model parameter vector φ includes parameters that are needed to specify physical responses in the model. Think generally of the vector φ as constants such as transfer coefficients in the set of equations on which M is based. For example, the reaction rate of a chemical process is often modeled as depending on temperature via the "activation energy" parameter in an Arrhenius model [Hammes 1978]. Thus, e.g., *x* would specify a material, φ would be its (assumed or estimated) activation energy. The particular parameter-values, φ, used to generate a prediction may be estimates based on handbooks, other experimentation, or judgment. The "uncertainty" of such estimated parameters will be considered below.

To focus on the validation problem, as opposed to the model development and verification problem, we further assume that it has been 'verified' that the code will adequately produce the intended mathematical result and that all numerical aspects of $M(x{:}\varphi)$, such as mesh size, time steps, and convergence criteria, have been satisfactorily resolved and are fully specified in *M*. The computer model, $M(x{:}\varphi)$, is thus an operator that transforms input *x* into the predicted result, y^M. This transformation is assumed to be deterministic in this paper in the sense that for a given specification of *x* and φ the code always gives the same y^M. Repeated runs of a deterministic code, as in a Monte Carlo analysis, however, will be considered.

c. Statistical Framework

Now, corresponding to the prediction, $y^M(x)$, consider an experiment conducted at the specified *x* and represent its outcome by $y(x,w)$. In this expression, *w* represents variables not included in the model that influence nature's experimental outcome. For example, a container might be modeled as a perfect cylinder and a 2-D model could be used to predict its behavior. Actual containers, however, are not perfect cylinders, so the out-of-roundness characteristics would be this situation's *w*'s. These *w*'s could affect performance and they would vary among nominally identical containers. in general, the *w*'s may not be recognized, or if recognized, may not be measured and they may not be controlled in the experiment or in events for which we desire to make predictions. We treat this "extra-model" contributor to nature's outcome statistically by modeling *w* as a random variable (with an unknown probability distribution). This means that the outcome of an experiment at *x*, say $y(x)$, is a realization of the random variable, $y(x,w)$, which has a probability distribution induced by the probability distribution of *w*. We define the *prediction error* of the model at *x* as the random variable,

$$e_x = y(x) - y^M(x). \qquad (2)$$

The probability distribution of e_x will in general depend on x. That is, the predictability of an event defined by *x* is apt to differ as one moves around the *x*-space of events. For example, both the bias and variance of e_x may depend strongly on *x*. This is not a desirable state, so such a finding is apt to lead to efforts to improve the computational model or to find functions of the experimental and computational results that do not have this dependence. For example, the standard deviation of prediction error may depend on *x* when dealing with a selected $y(x)$, but not when the data are analyzed using $\ln(y(x))$.

In general, $y(x)$ is observed with measurement error, so we express the observed experimental or field result as

$$y^E(x) = y(x) + \delta_x, \qquad (3)$$

where δ_x is a random variable representing measurement error. The probability distribution of measurement error may also depend on *x*. By combining (2) and (3), the relationship between experimental data and model predictions is

$$y^E(x) = y^M(x) + e_x + \delta_x. \qquad (4)$$

This relationship can be further complicated in situations in which the experimental and computational *x*'s do not match. For example, measured temperature in an experiment, used to calculate y^M, might be x = 300C, but the actual temperature in the experiment might have been 301C. When the differences between experimental and computational x's are small, the resulting error can often be folded into the measurement error in $y^E(x)$, namely δ_x. Good instrumentation is vital in model-validation experiments in order to prevent extraneous sources of error from distorting the evaluation of prediction error.

Equation (4), though written as a sum, in general represents the conceptual relationship that nature's outcome differs from the computational prediction because of prediction errors (e_x) and experimental measurement errors (δ_x). The relationship is not necessarily additive, but one goal of an analysis is to find a transformation of *y* that will linearize the relationship.

From (2) it can be seen that if the probability distribution of e_x was known at an *x*-value of interest, then, given a computational prediction, $y^M(x)$, one could probabilistically bound nature's outcome, $y(x)$. We could then answer the two questions about computational predictions posed in the first paragraph: How well does the model predict nature's outcome, first at conditions that can be tested, then at untested conditions?. The problem, of course, is that the distribution of prediction-error is not known; it must be estimated from model-validation experiments and predictions and ancillary data pertaining to measurement error. For reasons of cost and high-dimensionality of the *x*-space, the data from which to estimate the prediction-error distributions at *x*-values or over *x*–regions of interest are apt to be quite sparse. Hence, estimation, particularly of tail-percentiles of the distribution, will be quite imprecise, if even realistic. Statistical methods are aimed at characterizing the imprecision of data-based estimates. One can see from this framework, however, that the too-common notion that validation can be accomplished via a few well-chosen validation experiments is not apt to provide an adequate characterization of prediction error for complex, high-dimensional computational models.

Evaluating model predictive capability means estimating the probability distribution of e_x at selected *x*-points or over selected *x*-regions. This evaluation requires selecting a set of *x*-points, then obtaining computational predictions and experimental results for each. The results of a suite of model-validation experiments and corresponding computational predictions is thus a set of (x, y^M, y^E) values. This is the raw material from which estimates of the probability distribution of e_x must be constructed. Note that e_x and δ_x

in (4) cannot be separated using only the (x, y^M, y^E) data. Their effects are "confounded." Ancillary data pertaining to measurement error, as a function of x, are needed in order to isolate the probability distribution of prediction error, e_x. Again, one can see the importance of good instrumentation in model-validation experiments in order that the observed prediction errors, $y^E(x) - y^M(x)$, will predominantly reflect e_x, not δ_x.

[Note. It is also possible to analyze predictive capability when the experimental and computational results are not obtained oncommon x-values. Let (x^E, y^E) denote a set of experimental results and (x^M, y^M) denote a set of computational results. Then one can fit response surfaces (y as a function of x) to each set of results and use the difference between fitted values as an estimate of prediction error at selected x-values. In this paper we will focus on the analysis of (x, y^M, y^E) data to avoid the complications associated with separate fitting of experimental and computational results.]

One can also see from the preceding framework that the only way to learn about prediction error is to run experiments (or collect field data) and compare the results to computational predictions. Monte Carlo simulation on $M(x{:}\varphi)$ can only provide information on how the computational prediction, y^M, would vary as x or φ vary. Because such simulations cannot address the variability of the w's, they cannot provide information on the difference between nature and computational prediction. We mention this because there has been a tendency in some work to claim that such simulations provide a measure of prediction uncertainty.

Viewing the differences between experiment and model as statistical has engineering precedent. For example, in bridge design, civil engineers use a mathematical model for "scour" – the erosion of soil around a bridge's foundation due to river flooding [Johnson 1995]. This model is a function of soil type, flood magnitude, river velocity and other pertinent variables. For predictions civil engineers incorporate an additional "modeling factor" to represent the deviation of actual scour depths from the theoretical model predictions. This modeling factor corresponds to e_x in (4).

Implementing the process represented by Fig. 1 and the analysis based on eq. (4) leads to a variety of issues. The design of experiments is critical in the generation of data that can potentially yield a meaningful characterization of predictive-capability and the statistical analysis used to extract that information from the data is also important. Following a brief literature review, the next two sections discuss some of the problems that are likely to be encountered in these areas and indicate directions to take. More concrete illustrations of methodology are given in subsequent papers, [Bayarri et al. 2002] and Easterling [2002].

d. Brief Literature Review

There has been limited recognition in the computational modeling community of the statistical nature of model-validation and the evaluation of predictive capability. The authors of a National Academy of Sciences review of defense acquisition [Cohen et al. 1998] stated,

> "Given the critical importance of model validation.. ., it is surprising that the constituent parts are not provided in the (DoD) directive concerning … validation. A statistical perspective is almost entirely missing in these directives."

While it is generally recognized that model-validation involves the comparison of data to model predictions, Trucano et al. [2002] have characterized the typical analysis as being based on a "viewgraph norm" – data and model predictions are overlaid on a transparency and a judgment – good enough? – is made. We can and must do better. The organizers of this workshop have recognized the need to address the statistical aspects of model-validation in a more complete and fundamental way and we hope this paper is useful in this regard.

The preceding subsections have identified some of the difficulties in evaluating predictive capability. In a philosophical paper, Oreskes et al. [1994] argue that, "Verification and validation of numerical models of natural systems is impossible." Rather, in their view, the best we can hope for is a demonstration of "empirical adequacy." This goal is in the spirit of our statistical perspective – using data to evaluate adequacy. We don't expect perfection. The view of Oreskes et al. [1994], as they acknowledge, does not mean that numerical models have no value. We prefer the pragmatic view expressed by (University of Wisconsin statistician) George Box [1980]: "All models are wrong, but some are useful." Our goal is to use statistical methods to characterize a model's usefulness.

There has been prior work on statistical comparisons of experimental data and computational predictions. For example, Kleijnen [1995] addresses the comparison of binary (success/failure) outcomes from R runs of a simulation model and K field trials. Note that this is an aggregate comparison, not a test-by-test comparison as we primarily address here. The performance of military systems is the context for this analysis. Fries [2000], in a similar context, considers a combined analysis of a suite of comparisons of single field trials to a large number of simulation runs. In the area of Department of Energy applications, examples of statistical analysis of physics models and experiments are given in Hills and Trucano [2001] and Easterling [2001]. An extensive discussion of validation literature is given by Oberkampf and Trucano [2000].

Experimental Design

In broad terms, experimental design for model-validation means selecting a set of *x*-points (that define, e.g., test hardware and environments) at which to do experiments and computational predictions. This set constitutes the suite of experiments on which to build an evaluation of predictive capability. In detail, this specification of experiments also means determining experimental plans that specify the test hardware, methods, conditions, instrumentation, data collection, and post-processing techniques used to obtain information required for subsequent data analyses. All of these elements have different nuances for experiments that are designed for model validation studies as opposed to phenomena discovery or exploration. It is critical to emphasize this point. It is also important to recognize at the outset that measuring predictive capability has profound implications for the experimental sciences, not just the analytic.

The role of experimental design in the inference problem is illustrated in Fig. 2 in which the space of validation experiments and system applications is defined by two meta-variables, configuration and environment. Because of economic and other reasons it may not be possible to test actual systems in their required environments. (For this reason, Fig. 2 depicts an extrapolation situation; intuitively, interpolation should be easier.) For example, the application of interest might be the performance of a sophisticated electro-mechanical component in a severe radiation environment, but experimentation will be conducted using greatly simplified component mock-ups in less severe radiation environments. Thus, we have to extend what we can learn about predictive capability (represented by the prediction errors, $\{y(x)-y^M(x)\}$, in Fig. 2) at the selected *x*-points where we can evaluate it to an inference about predictive capability where we cannot. This inference requires an extension of the model itself *plus* an extension of what we know about unmodeled phenomena, as represented by the observed prediction errors. It requires merging prediction error data from tests of a variety of single and multiple phenomena into an inference about prediction error in the application's environment and configuration. Making this extension successfully and credibly requires subject-matter knowledge about the axes along which we can make such extensions and it requires a suite of experiments suitably located in the configuration-environment space to provide the data necessary to make such extensions. The design of this set of experiments thus has to be driven by the ultimate applications for which computational predictions and a model's predictive capability are required, as was discussed above.

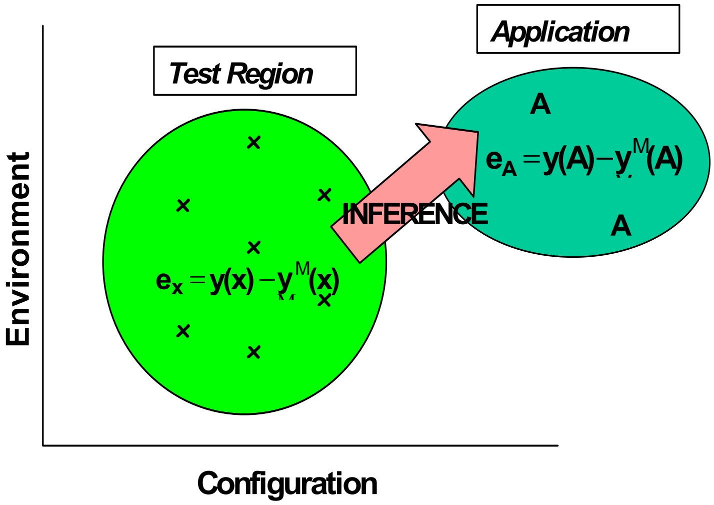

Figure 2. Inferring predictive capability

Clearly, extrapolation is a matter of judgment and potential disagreement. Experimental design should strive to minimize the need for extrapolation, to the extent possible. For example, if the application of interest is defined by a specified thermal profile – temperature ranges, ramp rates, and dwell times, say – then the experiments should duplicate this profile, if possible, rather than be conducted with perhaps more convenient and less model-stretching temperature profiles. That is, scientific validation of a computational model is not the same as estimating a model's predictive-capability in particular applications and this difference should influence the design and conduct of model-validation experiments.

a. Experimental Objectives

Meaningful validation experiments are designed to meet one or more explicit objectives. In general, the experiments conducted (1) should provide a sufficient test of predictive capability for the selected experimental situations and (2) the collective set of experiments and associated computational predictions should provide a basis for making the desired inference of predictive capability in application conditions.

There are various ways to translate the first objective into a basis for experimental design. For example, one measure of predictive capability at x is the standard deviation of prediction error, e_x, at that point. One could define the objective to estimate this standard deviation, call it sigma, within 100P% (at a specified confidence level) and then derive the number of experiments required to achieve that precision. These experiments could either be n replications at the selected x-point or n total experiments at different x-points within a region within which it is reasonable to expect a constant standard deviation. Under a distributional assumption for e_x, such as normality, one determine n so that, e.g., the ratio of an upper 95% statistical confidence limit on the standard deviation to the point estimate provided by the data is 1+P. Extending this sort of analysis to the simultaneous design of a suite of experiments, ranging from single-phenomenon to multi-phenomena system-like tests, is a problem-specific research problem. If the prediction errors in individual experiments can be mathematically linked to prediction error for the application, then potentially we can link the precision with which the application sigma is estimated to the precisions with which the constituent sigmas are estimated and thus arrive at a basis for specifying the suite of experiments. Also, we can work the problem sequentially, determining where additional experimentation would most improve the precision with which the application prediction-error sigma can be estimated.

The conduct of a validation experiment also influences how well predictive capability can be measured. As mentioned above, a variety of random and systematic factors can contribute to the difference between computational prediction and nature. Validation experiments need to be conducted in ways that allow these factors (nature's *w*'s) to be manifested as they would in an application of interest. For example, predictive capability measured in a tightly-controlled, pristine lab environment may not be appropriate for inferring predictive capability for predictions for a much less controlled, noisier application environment. The objective of assessing predictive capability in a specific application influences experimental design in terms of both what is controlled and what is not controlled in the experiments.

Another situation that will arise is when in an application some of the *x*'s, such as environmental conditions, will vary, but they will be fixed in any particular experiment. The ultimate objective will be to predict characteristics of the distribution of system response over some probability distribution of these *x*'s. The analysis problem in this case is discussed below. The experimental design objective in this situation is to conduct experiments over a suitable set of specified *x*'s to support the required distributional predictions and to characterize the precision of such predictions.

Time, resources, and experimental capability constrain validation experimental design and conduct. Such constraints must be balanced against the experimental objectives in arriving at a plan for model validation experimentation. A difficult decision will have to be made as to whether a meaningful evaluation of predictive capability is possible under existing constraints in any given situation. Of course, there are other reasons for experimentation and model-building besides characterizing the precision of application-level computational predictions.

b. Experiment-Model Compatibility

The computational and experimental elements of the model validation process cannot be executed in isolation. The vector *x* needs to be meaningful to both the experimentalist and modeler in order to align experiment and model so that both computational predictions and experiments at selected *x*-points can be run and compared. Further, this alignment needs to be meaningful in terms of the system scenarios for which computational predictions are required.

The discussion so far has assumed that the full *x*-vector could be controlled or measured in an experiment. If the modeler's *x*-vector contains variables that have no experimental meaning, this is not the case and it may not be possible to make meaningful comparisons. If the modeler's *x*-vector requires measurements that cannot be made, the result will be increased prediction uncertainty. To avoid this misalignment, there may be a need to develop new experimental and instrumentation capabilities. The definition of the variables in the *x*-vector is not just a modeling issue. The experimenter, the requirements-setter, and the decision-maker have to be able to operate and communicate in terms of this *x*-space.

c. Simplification

The objective of characterizing predictive capability over some high-dimensional *x*-space can quickly require an experimental design that exceeds available or reasonable resources. One way to avoid this problem is to vary only a subset of the variables in *x* while holding the others fixed at nominal or bounding values. Statistical experimental design methods [e.g., Box, Hunter, and Hunter 1978] should be used to efficiently and adequately explore the specified *x*-space.

Model simplification is another route to reduce the cost of predictive capability measurement. For example, suppose a model contains high-order effects or phenomena that cannot be controlled or measured in an experiment. It may be more appropriate to make computational predictions without those effects in the model and then capture those effects experimentally through the observed prediction errors. Where computational resources are constraining factors, model simplification increases in importance and attractiveness, but may also increase the complexity of inferring a computational model's predictive capability from the validation process.

Analysis

After conducting a suite of experiments and computational predictions the next task is to analyze the resulting data, $\{x_i, y^E(x_i), y^M(x_i) : i = 1, 2, \ldots n\}$. It is important to note that the subscript *i* refers to distinct experiments. Both *y* and *x*, though, may be fields or vectors containing thousands of measurements or calculations. It is decidedly not the case, however, that thousands of measurements, e.g., of temperature over a fine grid of space and time, for one experiment is equivalent to thousands of separate experiments. The number and nature of the experiments conducted will determine the precision with which predictive capability is measured, not the number of measurements per experiment. It is the variability of nature's w's from experiment to experiment (or among replications of the application) that determines the variance of prediction error; multiple measurements on one experiment do not capture this variability. (Incidentally, awareness and proper treatment of multiple sources of variation is a characteristic of a careful statistical data analysis.) Given the computational and experimental outcomes from the suite of experiments, the objective of the analysis of these results is to measure and/or estimate predictive capability. The following subsections address issues that arise in this analysis.

a. "Metrics"

Predictive capability at an *x*-point can be characterized by a variety of "parameters" (in the statistical sense of being a characteristic of a probability distribution) pertaining to the probability distribution of e_x. The expected value and the standard deviation of e_x are two important possibilities. Others might be the square root of the expected squared error, the 99th percentile of the distribution of absolute error; the lower and upper 95th percentiles on the distribution of e_x; and others. If the computational model was designed to be conservative on the high-side (i.e., e_x is intended to be negative), the metric of interest might be Prob($e_x <$ 0). When e_x has a normal distribution all of these distributional characteristics (parameters) are functions of the two parameters that characterize a normal distribution, the mean and the standard deviation.

Any of these measures of predictive capability must be estimated from the experimental and computational results. With limited data, estimation uncertainty will be appreciable. Statistical methods account for estimation uncertainty by methods such as confidence, prediction, and tolerance intervals [see, e.g., Hahn and Meeker 1991]. For example, a conclusion might be stated as: with 90% confidence the upper 95th percentile of the distribution of e_x for a specified *x* is no more than $U_{90/95}$. Hence, with 90% confidence, there is at least a 95% probability that nature's response, $y(x)$, at *x* will be less than $y^M(x) + U_{90/95}$. Comparing such a limit against a requirement provides an assessment of margin. The essential analysis point is that any "metric" of predictive capability derived from the model validation process will be a statistical estimate and the reliability of that estimate must also be evaluated and communicated.

b. Hypothesis Testing Metrics

It is common [e.g., Hills and Trucano 2001] to treat model-validation as a hypothesis testing problem. The approach is to create "uncertainty" probability distributions for $y^E(x_i)$and $y^M(x_i)$, separately. Let Σ^E and Σ^M denote the assumed/estimated covariance matrices for the vectors of experimental and model results, respectively. Then, under the further assumption that the experimental and model "uncertainties" are independent random variables, the covariance matrix for the difference between the experimental and computational results, $d = y^E - y^M$, is

$$\Sigma^d = \Sigma^E + \Sigma^M.$$

A dimensionless metric that measures the distance between the experimental and computational results is

$$X^2 = d^T(\Sigma^d)^{-1}d,$$

where the T superscript denotes the vector transpose.

Now let μ^E and μ^M denote the expected values of y^E and y^M. Under the hypothesis that $\mu^E = \mu^M$, and the assumption that the assumed/estimated covariance matrix, Σ^d, is the actual covariance matrix of *d*, the metric, X^2, has a chi-squared probability distribution with degrees of freedom equal to the dimension of the vector of differences. Comparing the observed value of X^2 to percentiles of this distribution provides a test of the hypothesis: $\mu^E = \mu^M$. Under the further assumption that measurement error is unbiased (has expected value = 0) this test is a test of whether the model predictions are unbiased.

The question asked via this hypothesis test is whether the observed difference, $y^E - y^M$, is satisfactorily within the combined estimated uncertainties. When this is not so, the hypothesis is rejected and the model is declared invalid. If the agreement is satisfactory, it is noted that, at least, the data don't rule out equality of μ^E and μ^M. If the agreement is satisfactory, then at least the data don't rule out equality of μ^E and μ^M. .Of course, failure to reject an hypothesis does not imply that the hypothesis is true. Indeed, for very uncertain y^M or very noisy data, a statistical test will typically not reject the hypothesis unless the model is egregiously bad. Because of this, it is only permissible in classical testing to accept a null hypothesis if a careful power analysis has been performed, and this can be a difficult undertaking in model validation. In the Bayesian approach discussed in the accompanying paper [Bayarri, *et al.* 2002], one can directly assess the posterior probability that the null hypothesis is true, but this is also a difficult computation

Finally, rejecting the hypothesis of equal underlying expectations, however, does not mean that the model is not useful for predictive purposes. An ensemble of differences might show that y^M predicts y (suppose there is reason to assume measurement error is negligible) consistently within 15%, which might be perfectly tolerable in the context of interest, even though the combined assumed separate uncertainties were, speaking heuristically, only 5% in magnitude. Conversely, if the hypothesis of equal expectations is not rejected, this result does not guarantee that useful predictions of y are provided by y^M. In fact, the more uncertain y^M is or is assumed to be, the more unlikely it is that the hypothesis of equal expectations will be rejected. Thus, hypothesis-test results, or refinements such as P-values, discussed in the following paragraph, associated with the test, are inadequate and inappropriate tools for characterizing predictive capability.

Classical hypothesis testing requires the prior determination of α, the probability of falsely rejecting the null hypothesis, and then an acceptance criterion is established that achieves this α. Rather than the binary pass/fail outcome, an alternative is to summarize the test by finding the largest α-value for which the test would fail. This threshold value is termed the P-value and is continuous on the [0, 1] interval. Stated another way, a P-value is the probability of observing a result that contradicts the hypothesis by as much or more than does the observed result. Thus, the greater the disagreement between model and experiment, the smaller the P-value is. But, the P-value does not provide a measure of predictive capability.

The estimates of the covariance matrices, Σ^E and Σ^M, are generally based on limited data and other information. Passing or failing the hypothesis test of equal expectations can occur due to errors in estimating these covariance matrices, so one does not, in general, achieve a reliable test of the hypothesis of unbiasedness.

One further flaw in this hypothesis-test metric is that the assumed covariance matrix, Σ^d, created by adding the separately constructed covariance matrices for y^E and y^M, does not capture what may be a major contributor to the difference, $y^E - y^M$. That contributor is the effect of the *w*'s that influence nature's outcome but are not in the model. The effect of this omission, all else being the same, is to increase the probability of declaring the two expectations to be significantly different.

Approaching model-validation as a pass/fail test of a computational model has led to treating model-validation as a statistical hypothesis-testing problem. However, the problem of measuring a model's predictive capability calls for a statistical estimation analysis, which is the objective here.

c. Choice of Analysis Variables.

In both experiments and computations there are a large number of response, or output, variables that can be observed and compared. Making the analysis manageable and the results meaningful and communicable requires a careful selection of outcome variables for which to evaluate predictive capability.

The selection of variables should first be driven by system requirements. If the requirement is that peak strain at a given location should not exceed some value, for example, then the model validation objective is to measure the predictive capability pertaining to calculated peak strain at that location. While it would add confidence in the computational model to know that the complete strain vs. time history at various sites in the test device can be reasonably well predicted, it is really not appropriate in the given situation to devote a lot of analysis effort to measuring predictive capability over an extensive time and space grid. This requirements focus is also a way to greatly reduce the dimensionality of the data, which in general may be time-histories of responses such as acceleration, strain, or temperature in time and space, to a small number of 'integral' variables such as peak acceleration or the time to reach critical temperatures at selected points in a system or component.

d. Statistical Models

Statistical analyses are generally carried out by modeling observed data as observations from some family of probability distributions, often, but not only, the normal distribution family. This family is characterized by two parameters, the mean and standard deviation, and the objective of the analysis is to estimate these two parameters or pertinent functions of them, such as the probability that a characteristic of interest exceeds its required lower limit. Statistical tolerance limits, such as $U_{90/95}$ in the previous subsection, are based on such models. For a given set of data, different values of $U_{90/95}$ would be obtained using normal distribution theory than would be obtained based on, say, the Weibull or logistic distributions. The choice of distribution family, however, is not ad hoc or blind. The data themselves can be used to guide the selection and to assess the aptness of a selected distribution family. Limited data may be consistent with different distribution families in which case one could choose a family based on mathematical convenience, precedence, or other subjective grounds. Or, one could conduct an analysis under a variety of plausible models to illustrate the sensitivity or insensitivity of the results to the choice of statistical model and to envelope the results.

For the sake of illustration, suppose the normal distribution family is used to analyze the observed prediction error data, $\{x_i, e(x_i) = y^E(x_i) - y^M(x_i)\}$. Suppose further, for the time being, that in the experimental region it has been established that measurement error is statistically negligible relative to prediction error. The statistical model for the observed error data will be that at *x*, *e(x)* is normally distributed with a mean *β(x)* and standard deviation *σ(x)* (the subscripts are suppressed for ease of notation). If replicate experiments are conducted at a given x_i (allowing the *w*'s to vary appropriately), then $\beta(x_i)$ and $\sigma(x_i)$ can be estimated directly from the data at x_i. More likely, because of limited data, and more appropriately, because there are apt to be smooth patterns in *β(x)* and *σ(x)* over regions in the *x*-space, the analysis objective would be to use the ensemble of data to estimate the bias and standard deviation functions of *x*. To this end, the fitting functions could range from simple linear models to spatial or statistical process models [e.g., Chiles and Delfiner 1999], depending on the nature and the amount of data. The analysis would be done in an exploratory, adaptive mode as different models for the bias and standard deviation functions are tried. For high-dimensional *x* and limited data, it may not be possible to obtain meaningful estimates. Hence, there is a need, as discussed above, to reduce the dimensionality of *x* in order to obtain useful results.

e. Statistical Analyses

The basic objective of statistical data analysis is to extract and convey what the data have to say about various issues, the resolution or clarification of which is the reason the data were obtained. There are a variety of statistical paradigms that have been developed to meet this objective. We focus on the two most prominent and their application to the problem of characterizing predictive capability.

Frequentist

As noted in the previous section, a statistical analysis starts by modeling data as realizations of random variables, generally with some underlying structure. For example, in a situation in which a response, y, is observed at various values of a possible explanatory variable, x, the "simple linear regression" model for such data is:

$$y = \alpha + \beta x + e; \quad e \sim \mathrm{N}(0, \sigma^2). \qquad (5)$$

Thus, (5) means that observed *y* is modeled as a linear function of *x* plus "random error" generated by a Normal distribution with mean zero, variance σ^2. (The observed $\{x,y\}$ data and diagnostics calculated from them can be used to assess the appropriateness of this model, so the adoption of the model, (5), is not done blindly.) The three parameters in this model, *α, β,* and *σ,* are unknown and the analysis objective is to identify plausible values of these parameters, given the data. (With this information about the model's parameters, one can address issues such as how large might *y* be for *x* within some specified range and the probability it is within requirements.) Finite data cannot uniquely identify the parameters, but can identify parameter regions that are consistent with the observed data to some specified extent.. To this end, the frequentist approach is to derive estimators (functions of the data) of the model parameters that have known statistical relationships with the parameters, relationships that permit the analysis objectives to be met. For example, for the linear regression model, the least squares estimator of the slope, β, is

$$b = \Sigma(x-x^{-})(y-y^{-})/\Sigma(x-x^{-})^{2},$$

where x^{-} and y^{-} are the means of the observed x and y data. This estimator has the "frequentist" property that in repeated realizations of data from the model (5) the expected value of *b* is β. Thus *b* is an unbiased estimator of β. Further, the precision of the estimator is given by the variance of *b*, which is $\sigma^{2}/\Sigma(x-x^{-})^{2}$. The unknown variance, σ^{2}, is estimated by the "residual mean square," say s^{2}, and the quantity, $s/[\Sigma(x-x^{-})^{2}]^{.5}$, which is the square root of the estimated variance of *b*, is termed the standard error of *b,* denoted here by *se(b)*. This standard error is important because the "pivotal quantity,"

$$t = (b - \beta)/se(b),$$

has a known probability distribution: the Student's t distribution with, in this case, n-2 degrees of freedom (df). This frequentist property enables one to bound β, given *b* and *se(b)*. For example, a 95% confidence interval on β is those values of β for which *t* will fall in the middle 95% of the t(n-2) distribution. Thus, to be consistent with the data, as summarized by *b* and *se(b)*, at the 95% level, β would have to be in the derived confidence interval. (See any statistical text on regression for more details of this analysis.)

The frequentist approach encounters difficulty in complex situations for which exact variances and pivotal relationships cannot be obtained. For the case at hand, the prediction-error data may exhibit a strong nonlinear relationship with several x-variables and non-Normal patterns of variability. To work these sorts of problems, various approximations are used. For example, Taylor's series expansions can lead to approximate standard errors of complex estimates and further analysis [Satterthwaite's method; see e.g., Ostle 1963] can associate an approximate df associated with the standard error. Normal distribution-theory results then provide approximate confidence intervals, perhaps after data-transformations that enhance the accuracy of such approximations. Another approach is to estimate the probability distributions of pivotal-like quantities via parametric or nonparametric Bootstrap methods [Efron and Tibshirani 1993]. The trouble with these approximate methods is that they are 'guaranteed' to be sufficiently accurate only with large enough data sets, and the definition of 'large enough' is highly situation-dependent. Their performance in small-data set situations is highly situation-dependent. The only way to know how well "truth" is approximated is to know "truth," in which case one wouldn't need an approximation. When the available data are limited, as is common in model validation scenarios, large-scale simulations, spanning a variety of "truth-states," are often required to provide adequate insight into the adequacy of an approximation. In complex situations, such simulations are often not feasible.

Bayesian

The Bayesian approach adds further probabilistic structure to the data model by assuming that the fixed but unknown parameters underlying the data are themselves random realizations from assumed "prior distributions." Bayes Theorem is then used to "update" these prior distributions, which means to obtain the posterior distribution of the parameters, given the data. For some standard problems, such as the simple linear regression model, and well-chosen priors, closed form solutions are possible. Modern computing capabilities, however, permit more general Bayesian analyses to be well-approximated in complex situations. (See [Bayarri et al. 2002] for details and discussion.)

Bayesian analysis does not require large data sets for implementation in complex situations, as do the approximations discussed earlier in the frequentist approach. On the other hand, when there is only a small amount of data, the choice of prior distribution can be critical to the analysis and influential on the results. Whereas the adequacy of a frequency model for the data can be evaluated via the data, the data provide very little information regarding the adequacy of the assumed prior distribution. (Because we have data from only one value of β, for the linear model example, it is hard to evaluate, from the data, how well the assumed underlying variability of β is represented by the assumed prior distribution.) There are two approaches to this issue.

The subjective Bayesian approach is to use prior distributions to represent degree of belief or state of knowledge about the parameters, prior to collecting the data. This can be both a blessing and a curse. The blessing is that expert opinion and/or partial prior scientific knowledge can easily be incorporated into the Bayesian analysis, allowing for predictive accuracy assessment to be performed based on a mixture of prior knowledge and data; this can be especially valuable when it is impossible to perform a complete suite of validation experiments. The curse is that such statements will often be treated more skeptically by others than will statements based primarily on data. One device for overcoming such skepticism is to conduct sensitivity studies with respect to the choice of priors but, in complex situations, this can become unmanageable.

The objective Bayesian approach is to choose innocuous priors, priors that will minimally influence the message in the data, then use the Bayesian machinery to obtain results that can be regarded as useful approximations to unobtainable exact frequentist results. For example, a 95% posterior probability interval on a parameter may be nearly the same as a 95% statistical confidence interval. In the linear regression example, for a suitable objective prior, the posterior distribution of β, given the data, will exactly satisfy the t-distribution relationship in the previous subsection.

An advantage of the objective Bayesian approach is that one set of machinery can be used to work all problems. One can write down the defining relationships between data and parameters and adequate computing power can work out the implications. A problem with sparse data and complex relationships is that selected prior distributions can still be influential, so sensitivity analyses are required to try to discern how much of the message is data and how much is artificially introduced by the prior.

Comment.

There can be sharp (and entertaining) philosophical and technical disagreements between Bayesian and frequentist adherents, although the two schools seem to have been growing closer in recent years. In any case, it is our view that such issues are secondary to those that must be addressed in order to conduct the right experiments and generate enough of the right kind of data to permit a meaningful evaluation of predictive capability by whatever method.

f. Model Tuning

When the analysis of prediction error data shows evidence of a bias in the computational model, one can potentially either incorporate that bias into subsequent prediction error limits, in essence calibrating out the model's bias, or one can modify the model in an attempt to remove the bias. One mode of modification is to adjust the φ parameters, which may often be uncertain estimates of, e.g., materials properties. Such 'tuning' can be suspect, but there are legitimate analyses that compensate for parameter estimation in characterizing the uncertainty of subsequent predictions.

Consider the case of a simple linear model, $y^M = \alpha + \beta x$. If an experiment is done at x_1, yielding y_1, then there are infinite ways to adjust α and β to achieve perfect agreement between y^M and y_1. No rational statement could be made, however, about predictive capability for the adjusted model. If a second experiment is done at x_2, then a unique α and β can be found to achieve perfect agreement at both points, but no statement about subsequent predictive capability can be made (obviously, a claim of perfect predictions is bogus). For three or more experiments, however, we can use standard statistical methods to estimate α and β and characterize the prediction error for subsequent predictions based on these estimates. The following case study [Easterling 2002] demonstrates this analysis. This sort of prediction-error analysis that accounts for tuning can be extended to the situation of more complex, higher-dimensional models, as in the accompanying paper [Bayarri, *et al.* 2002].

For complex codes and corresponding experiments, one computation and one experiment can each yield thousands of data-values – traces of multiple response variables over time and space. There may be many parameters in φ that could be adjusted to improve the agreement between computation and data. Even when there is a scientific basis for selecting the parameters on which to tune the computation, the residual errors over time and space after tuning to one experimental outcome do not contain any information about predictive capability. One could only infer at best that: If another similar experiment were run *and tuned*, the resulting residual errors should look like the post-tuning errors obtained in the first experiment. One could not infer: If we used the tuned model to make a prediction in a similar experiment, the error of that prediction should be in line with the post-tuning errors we obtained in the initial experiment.

g. Dealing With Bias

The finding of (possibly *x*-dependent) bias can lead down several paths: i. Bias could be evidence of correctable flaws in either the computational model or the experiments. Tuning the model parameters is one potential fix, though as just discussed, tuning can lead to misleading impressions of predictive capability. Making fundamental changes in the computational model's structure is another possible fix. The maturity of the model would be a factor in whether to pursue this fix. If the model is modified, additional experimentation, essentially another loop through the validation process, may be required to "validate" the model changes. Bias in the experiments' conduct or measurements is a source of apparent prediction-error bias that should be eliminated (rendered negligible), to the extent possible. Otherwise, we will be making predictions of a biased measurement of nature's outcome, not the outcome itself.

ii. If there are no (affordably) correctable flaws, and one still wants to use the computational model to make predictions, then another way to deal with bias is to adjust model predictions by adding the estimated bias function to them. Such an empirical fix is regarded as bad science by some, bit it only seems prudent to take advantage of the superior predictions that are available by bias-correction, until the source of the bias in the model can be identified and corrected.

Bias-correction of a model prediction for a single input often results in roughly replacing the computational model by an empirical model built from the validation-experiments' data. To see this, let $b(x)$ denote the estimate of the prediction-error bias function, $\beta(x)$ (= expected value of $e(x)$). This estimate, by whatever means it is obtained, can be regarded as a "smooth" of the observed prediction error data, $\{x_i, y^E(x_i) - y^M(x_i)\}$. Let $y^{M\wedge}$ denote the bias-corrected prediction. Then,

$$y^{M\wedge} = y^M + smooth(y^E - y^M)$$

$$\approx smooth(y^E).$$

That is, the model essentially cancels out so that prediction is based on a possibly science-guided, but nevertheless empirical function based on the validation data. Thus, bias-correction can effectively reject the computational model in favor of the data. . In the Bayesian approach, bias-correction is often less extreme, with the answer being a weighted average of the model-prediction and the data-prediction, with the weights (typically themselves a function of *x*) reflecting the variabilities and uncertainties in the models and data.

In certain situations, the model-predictions can dominate the data-predictions. One such situation is when the model is used to predict outside the range of the data. For instance, in the Bayesian approach, the weight that is given to the data-prediction will typically sharply decline as one moves away from the range of the data. A second important situation is when predictions for a small change in input values is desired. Indeed, if one desires to predict the difference between reality at x and $x + \Delta$, the bias-corrected answer will typically behave as

$$
\begin{aligned}
y^{M\wedge}(x) - y^{M\wedge}(x+\Delta) &= y^{M}(x) - y^{M}(x+\Delta) + smooth(y^{E}(x) - y^{M}(x)) - smooth(y^{E}(x+\Delta) - y^{M}(x+\Delta)) \\
&\approx y^{M}(x) - y^{M}(x+\Delta),
\end{aligned}
$$

so that the result from the model-prediction then dominates. Note that this is consistent with the commonly heard folklore that even globally biased models are often useful for predicting small changes.

There are situations in which bias is deliberately introduced into a computational model: e.g., a simplified, conservative mathematical model is used for a complex relationship that is difficult to model more accurately. The analysis of the observed prediction error data would quantify the degree to which this conservative strategy was successful. The choice of whether to use model predictions directly, or to bias-correct them would depend on the results of the analysis.

Regardless of how bias is dealt with, this discussion highlights an important point: Having enough data to estimate the bias function adequately means having enough data to build an empirical model of the phenomena of interest, at least within the experimental region. This observation is not to downplay the value of computational model, but it does indicate that data-based modeling still has a role to play.

h. Dealing With Variation

In addition to bias, the variance of prediction error is an important measure of predictive-cpability. The statistical analysis of the observed prediction-error data can result in an estimate of the variance of observed prediction error as a function of x. Denote this estimate by $s^2(x)$. Under the general statistical model, (4), $s^2(x)$ estimates the variance of the sum of prediction error, e_x, and measurement error, δ_x. Under the generally plausible assumption that measurement error is independent of prediction error, the variance of this sum, $e_x + \delta_x$, is $\sigma_e^2(x) + \sigma_\delta^2(x)$, the sum of the individual variances. Unless individual experiments have been measured more than once, these two “variance components” cannot be separated using the results of the model-validation suite of experiments. The recourse in this case is to separately estimate the variance of measurement error via gauge studies or other evaluations of measurement processes, then subtract that estimated variance from the estimated total variance. The accompanying implementation paper, [Easterling 2002] illustrates this analysis.

Suppose that a consideration of the estimated prediction-error variance in the appropriate context leads to a conclusion that this variance is "too large." This is an indication that the unmodeled, uncontrolled *w*'s in nature are causing more variation than is acceptable, either in terms of how well the experiments can be predicted or in terms of the ability to infer predictive-capability in untested applications. That is, large observed prediction-error variation.may mean that there are just too many unknowns in controlled situations to risk extrapolation to less controlled situations. The recourse, in terms of our framework, would then be to attempt to incorporate some of these *w*'s into the model, i.e., convert some of the *w*'s to *x*'s. For example, replace a 2-D model of a 3-D phenomenon by a 3-D model. This means that the 3-D characteristics of an experimental unit, such as a map of thicknesses and diameters (assumed constant in a 2-D model), would need to be measured so that unit-specific model predictions could be computed and compared to each unit's experimental outcome.

i. Inference

While it is valuable to know, thorough statistical evidence, how well a computational model can predict the outcomes of observable situations, computational models are particularly valuable if we "know" their predictive capability in situations that cannot be tested. Such situations occur, e.g., when the objective is to predict the outcomes of abnormal, or catastrophic events involving major systems such as transportation or weapons. To characterize predictive capability in these situations requires extending information about predictive capability where it can be evaluated to the applications of interest. This is the "inference bridge" in Fig. 2.

The inference bridge can be constructed, first of all, if the underlying scientific relationships can be assumed to extend over a region containing both the x_A and the x_E. Secondly, there needs to be a credible basis for similarly extending the prediction-error distribution. The scientific basis for this extension, however, is more tenuous because the prediction-error distribution, after all, reflects factors in nature not captured by the scientific model. Nevertheless, an informed judgment can sometimes be made. When there is a mathematical connection, statistical methods can account for and reflect the 'distance' between these points. The greater the distance, the greater the prediction uncertainty is.

An example is simple linear regression, *y* vs. *x*. The data in the experimental region may support the assumptions that the expected value of *y* is a linear function of x and that deviations from this relationship are randomly distributed according to a Normal distribution with mean zero and a variance that is constant across the experimental region. Given the assumptions that both the linear model and the homogeneous-variance, unbiased, Normally-distributed extra-model variability can be extended from the experimental region to the application region, statistical methods exist for characterizing the precision of application predictions based on the experimental data [see any statistical text that includes regression analysis]. The statistical analysis is conditional on those assumptions; it does not characterize how well they extrapolate. Theory may support extending the linearity assumption, but assumptions about the extra-model variability are much more ad hoc. There is no guarantee, statistical or otherwise, e.g., that unbiasedness will hold outside of the experimental region, so inferences will be conditional on such assumptions. Subject-matter knowledge about the differences between experimental and application conditions, however, can lend credence to the assumptions. As with a Bayesian analysis, the sensitivity of inferences to untestable assumptions should be investigated.

As mentioned earlier, the spatial representation of the experimental design (in x-space) and inference problems suggests that spatial statistical methods [Chiles and Delfiner 1999], such as kriging, can be used to model a metric, such as the estimated standard deviation at x, as a function of x, then estimate the value of that metric at x_A and estimate the uncertainty of that estimate.

There are several different inference situations (and a careful taxonomy of these is a research need). In one category, predictions of system performance may be made during design or development. Then, when the system goes into operation it provides field or system-test experience that can confirm or deny the assumptions on which the development-based inferences, say, were drawn. Such experience provides prediction-error data that may be pertinent for the next round of model and system development. In another category, such as predictions of abnormal events such as nuclear power plant accidents, it is unlikely and undesirable that data will be obtained to compare against model-based predictions. This situation puts a premium on transparency and communicability of the experimental evidence and its analysis in order that users of computational predictions have a clear view of their limitations and risks.

One other inference situation is discussed in the following subsection. In experiments, the conditions, x, may be held fixed at specified levels, while in applications, they will vary. Examples are temperatures, velocities, impact conditions – boundary conditions, in general. Given assumptions about the nature of that variability in the application, inferences about the resulting distribution of y can be obtained. Extrapolation can still be a concern here if, e.g., one controlled experimental temperatures in the range of 600C to 1000C, then sought to predict the distribution of application outcomes when temperature is assumed to vary randomly between 1200C to 1500C.

If no credible inference is possible, one may have to re-examine everything from requirements to system design to test program. More system-like testing may be required to reduce the inferential gap. A system may have to be redesigned so as not to be vulnerable to an environment whose effect cannot be well-predicted computationally. The sort of framework proposed here provides a vehicle for addressing such fundamental issues.

j. Distributional Predictions

A deterministic code calculates a prediction for a single, completely specified situation. Predictions of interest, though, are often ‘statistical,’ or distributional predictions, not single point predictions, as considered up to this point. For example, in a weapon systems context, delivery and target conditions, such as impact angle, impact velocity, and target hardness, vary from mission to mission. In such situations the objective may be to predict the resulting probability distribution of some characteristic of weapon-performance, such as maximum shock on a key component, over some probability distribution of environmental conditions, and then to predict characteristics of this distribution. These characteristics could be the distribution's mean, its upper two-sigma point, or the probability of exceeding a failure threshold.

Suppose that x_r, a subset of the variables in x, is to be treated as random to obtain a distributional prediction. Suppose further, as a starting point, that the probability distribution of x_r is a given. Our objective is to estimate the resulting distribution of y and parameters associated with it. The statistical model specified above in (3) provides the means to do this, given appropriate experiments and data.

Consider now the model relating nature's outcome at x with the model prediction:

$$y(x) = y^M(x) + e_x \tag{6}$$

The law of total variance[8] says that

$$var(y) = \ var_x[E(y|x)] + E_x[var(y|x)], \tag{7}$$

where $var(.)$ denotes variance, $E(.)$ denotes expectation, and | denotes conditioning. The subscript indicates the random variable over which these moments are calculated. In words, (6) says that the unconditional variance of y is the sum of the variance of the conditional expectation of y, given x, and the expected value of the conditional variance of y, given x. Applying this relationship to the problem at hand leads to:

$$var_r(y) = var_r[y^M(x) + \beta_x] + E_r[var(e_x)], \tag{8}$$

where the subscript r denotes that the indicated variance or expectation is with respect to the distribution of x_r and β_x is the bias function, the expected value of e_x. .

Suppose, to simplify things for this discussion, that $\beta_x = 0$, for all x in the x-region of interest. Then (8) becomes

$$var_r(y) = var_r(yM) + E_r[\sigma_x^2]. \tag{9}$$

Propagation of the assumed distribution of x_r through $M(x:\phi)$, by methods such as Monte Carlo, provides an estimate of the first right-hand term in (9). Model-validation experiments and data analysis, if successful, provide an estimate of σ_x^2, as a function of x. The expectation of this function with respect to the distribution of x_r could then be calculated or approximated to estimate the second right-hand term in (9). In the ideal situation in which σ_x is independent of x in the region of interest, the second right-hand term is simply σ^2, the variance of the difference between nature and computation. In either case I call σ_x^2 the 'extra-model' variability. Similarly to (9), other parameters of the distribution of y, such as an exceedance probability, would have to be estimated by folding in the extra-model variability represented by the distribution of e_x.

Equation (9) shows that the role of the extra-model variability is not to provide bounds on the computational prediction, as was the case for point predictions. Rather, it is to add an additional variance component to the analysis; the effect of this addition is to inflate the variance one would get from propagation through the code. By itself, the code propagation variance, the first right-hand term in (9), underestimates the variance of nature's y, the left-hand term. If the code propagation variance, $var_r(y^M)$, was used as an estimate of nature's variation, then, e.g., failure probabilities would tend to be underestimated, sometimes drastically, even if the model has been deemed valid via a hypothesis test. To obtain valid distributional predictions it is necessary to combine the estimated 'extra-model variability' with the estimated model-propagated variability.

Traditional code uncertainty-propagation analyses work the first right-hand term in (9), in various manifestations. Much research has been and continues to be conducted trying to wring out one more significant digit in approximations to this first term, all the while ignoring the second term (sometimes of necessity in situations in which meaningful model-validation experiments cannot be run). The only way to know whether the second term is ignorable is to run the model-validation experiments and perform analyses to evaluate it. Estimating the second term and the bias function, β_x, should be the objective of model-validation programs. This is a much harder problem to work. It requires designing and running experiments, not just conducting computer exercises. It requires test facilities. It requires collaboration with experimentalists. It is messy. But it is necessary if credible measures of predictive capability are to be obtained. See [Aeschliman and Oberkampf 1997] for discussions and illustrations on this point in the context of fluid dynamics.

Summary

This paper has attempted to lay out the statistical foundations of model-validation, by which we mean the process of evaluating the predictive capability of a computational model. Issues pertaining to the design, conduct, and data analysis of suites of model-validation experiments were discussed. Our primary message is that a substantial amount of experimentation may be required to develop a credible evaluation of predictive capability. Success is not always assured, particularly when findings pertaining to predictive-capability in an experimental regime must be extrapolated to a system-application environment. The real test of the proposed statistical approach will come through attempts to implement the ideas and concepts presented. We believe that the appropriate statistical theory and methods exist, research should be aimed at situation-specific implementation of these methods. To this end, implementation methodology is illustrated in two accompanying papers, [Bayarri et al. 2002 and Easterling 2002].

Computational models are and will (have to) be used to make predictions of complex, unobservable events in a wide variety of applications, whether or not a credible statistical evaluation of predictive-capability can be accomplished. Even if the ideal outcome we set forth is not achieved, the reality checks provided by a robust suite of comparisons of experiments to computational prediction increases credibility. Furthermore, a careful statistical evaluation of predictive capability within the experimental region is a valuable step beyond much current practice. We hope that the framework and examples we present encourage producers and consumers of computational predictions to increase the statistical content of their efforts to construct and communicate the credibility of computational predictions.

References

Aeschliman, D. P., and Oberkampf, W. L. *Experimental Methodology for Computational Fluid Dynamics Code Validation*, Sandia National Laboratories report SAND95-1189, September 1997.

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

Bayarri, M. J., Berger, J. O., Higdon, D., Kennedy, M. C., Kottas, A.,Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H., and Tu, J. A Framework for Validation of Computer Models, V&V Foundations 2002.

Box, G. E. P. Sampling and Bayes' Inference in Scientific Modeling and Robustness, *Journal Statist. Soc. A.* v. 143, 383-430, 1980.

Box, G. E. P., Hunter, W.G., and Hunter, J. S. *Statistics for Experimenters*, John Wiley and Sons, Inc., New York (1978).

Chiles, J. P., and Delfiner, P. *Geostatistics, Modeling Spatial Uncertainty*, John Wiley and Sons, Inc., (1999).

Cohen, M. L., Rolph, J. E., and Steffey, D. L., eds. *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements,* National Academy Press, Washington, DC 1998.

Easterling, R. G. *Measuring the Predictive Capability of Computational Models: Principles and Methods, Issues and Illustrations*, Sandia National Laboratories Report SAND2001-0243, February, 2001.

Efron, B., and Tibshirani, R. *An Introduction to the Bootstrap*, Chapman and Hall, New York. 1993.

Fries, A., Another "New" Method for "Validating" Simulation Models, *Proceedings of the Army Conference on Applied Statistics*, Houston, TX, , ACAS CD Proceedings, October 2000.

Hahn, G. J., and Meeker, W. Q. *Statistical Intervals*, John Wiley & Sons, Inc., New York (1991).

Hammes, G. G., *Principles of Chemical Kinetics,* Academic Press, NY, 1978.

Hills, R. G., and Trucano, T. G. *Statistical Validation of Engineering and Scientific Models with Application to CTH.* Sandia National Laboratories Report SAND2001-0312, September, 2001.

Johnson, P.A. Comparison of Pier Scour Equations Using Field Data, *ASCE Journal of Hydraulic Engineering*, v. 121, 626-629 (1995).

Kleijnen, J. P. C. Cast Study: Statistical Validation of Simulation Models, *European Journal of Operational Research* v. 87, 21-34 (1995).

Oreskes, N., Shrader-Frechette, K., and Belitz, K. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, *Science*, v.263, 641-646, February 1994.

Ostle, B. *Statistics in Research,* Iowa State University Press (1963)

Parzen, E. *Stochastic Processes,* Holden-Day, San Francisco (1962).

Pilch, M., Trucano, T., Moya, J. L., Froehlich, G. Hodges, A., and Peercy, D. *Guidelines for Sandia ASCI Verification and Validation Plans – Content and Format: Version 2.0.* Sandia National Laboratories Report SAND2000-3101 (January, 2001).

Satterthwaite, F. E. An Approximate Distribution of Estimates of Variance Components, *Biometrics*, 110, 1946.

Trucano, T., Pilch, M., and Oberkampf, W. O. *General Concepts for Experimental Validation of ASCI Code Applications.* Sandia National Laboratories Report SAND 2002-0341, March, 2002.

Author Notes

1. Robert G. Easterling received a PhD in statistics from Oklahoma State University and spent most of 34 years at Sandia National Laboratories as a consulting statistician and department manager. He retired from Sandia in 2001 from the position of Senior Statistical Scientist and taught at the University of Michigan, fall 2001. During the last five years, his primary research and project interest has been in the application of statistical methods in model-validation experimentation and analysis.

2. James O. Berger received a Ph.D. in mathematics from Cornell University in 1974. After 23 years at Purdue University, he moved to Duke University in 1997 as the Arts and Sciences Professor of Statistics. He is currently also the Director of the Statistical and Applied Mathematical Sciences Institute, located in Research Triangle Park in North Carolina. His research focuses on development and application of Bayesian statistical methods to areas such as astronomy and model-validation.