SAND REPORT
SAND2003-0287
Unlimited Release
Printed February 2003

Statistical Foundations for Model Validation: Two Papers

Robert C. Easterling

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/ordering.htm

SAND2003-0287
Unlimited Release
Printed February 2003

Statistical Foundations for Model Validation: Two Papers

Robert G. Easterling
Statistical Consultant
51 Avenida del Sol
Cedar Crest, NM 87008

Abstract

This report reprints two papers that were presented at the Foundations for V&V (Computer Model Verification and Validation in the 21st Century Workshop, October 22-23, 2002. This workshop was sponsored by the Defense Modeling & Simulation Office and other government agencies and hosted by the Johns Hopkins University Applied Physics Laboratory. Its first objective was to produce a clear & comprehensive description of the state of the verification and validation art for modeling and simulation. Complete conference proceedings on CD-ROM can be obtained from the DMSO.

Overview

Validation of a computer model means the comparison of computational predictions to physical outcomes of the events being predicted. Conventional validation practice is to use these comparisons to make a pass/fail, valid/not-valid decision about the computer model. I take the view that the purpose of these comparisons is to characterize how well the computer model predicts, first within the realm of conditions in which validation-experiments are conducted, then for real-world applications in which computational predictions are required. Because of limitations on experiments, this latter objective may require an extrapolation, or inference. Basically, the objective is to take what we can learn about a model's predictive capability in an experimental region and use that information, buttressed by scientific understanding of the phenomena being modeled and predicted, to make credible, defensible, communicable statements about predictive capability for pertinent applications of the model that may be outside the realm of experimental or other available data. This is not easy.

What we learn about predictive capability depends on the nature and extent of the experiments conducted for that purpose. This means that experimental design, the specification of a suite of experiments and corresponding computational predictions, defines the breadth and depth of what we can potentially earn about predictive capability. Then, the analysis of the predictive-capability data that are obtained from a possibly quite diverse set of experiments summarizes and communicates what we have learned from these experiments and computations. These two elements, experimental design and data analysis, are fundamentally statistical in nature. For this reason the V&V02 Workshop sponsors organized a session on the statistical foundations of model-validation. The first paper, co-authored by James Berger, Duke University, and myself sets up a statistical model for the predictive-capability problem and identifies and discusses various issues that need to be addressed in experimental design and statistical data analysis. To provide concrete illustrations of the methods involved in addressing these issues, two implementation papers, or case studies, were also presented. One, by James Berger and associates from Duke University and the National Institute of Statistical Sciences defines and illustrates a Bayesian analysis of a suite of computational predictions and experimental data. That paper is in the Workshop Proceedings or is available from the authors. The second implementation paper, included in this report, presents my analysis, from what is sometimes called a "classical" or "frequentist" statistical perspective, of a set of Sandia experiments and predictions pertaining to the degradation of polyurethane foam in a thermal environment.

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.

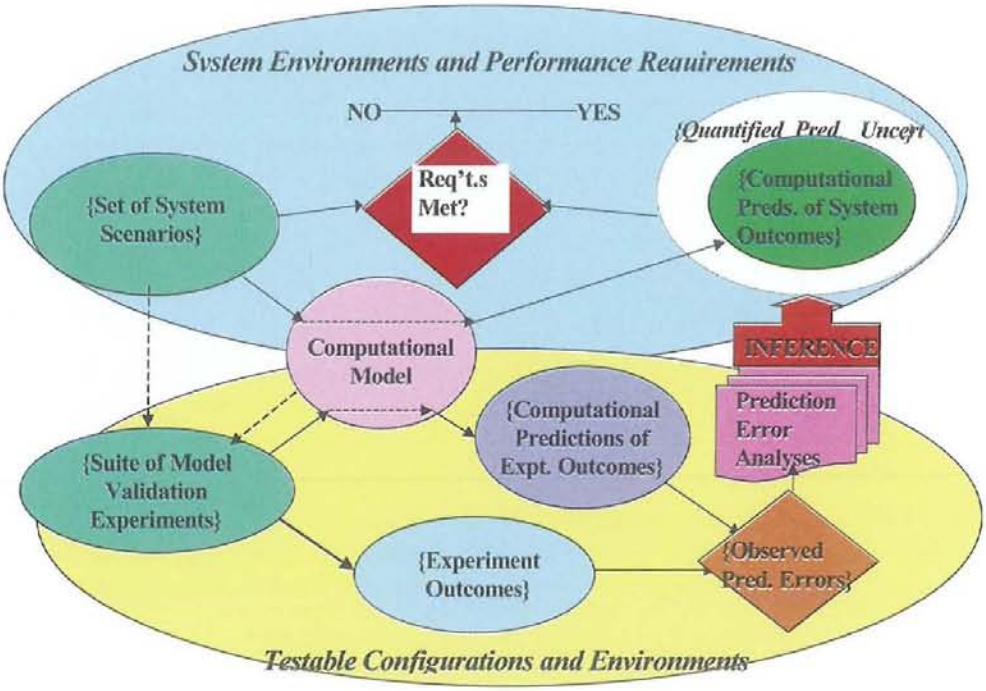

Figure 1. Schematic for Characterization of Prediction Capability

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,

$$e_x = y(x) - y^M(x). \tag{2}$$

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

$$y^E(x) = y(x) + \delta_x, \tag{3}$$

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

$$y^E(x) = y^M(x) + e_x + \delta_x. \tag{4}$$

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.

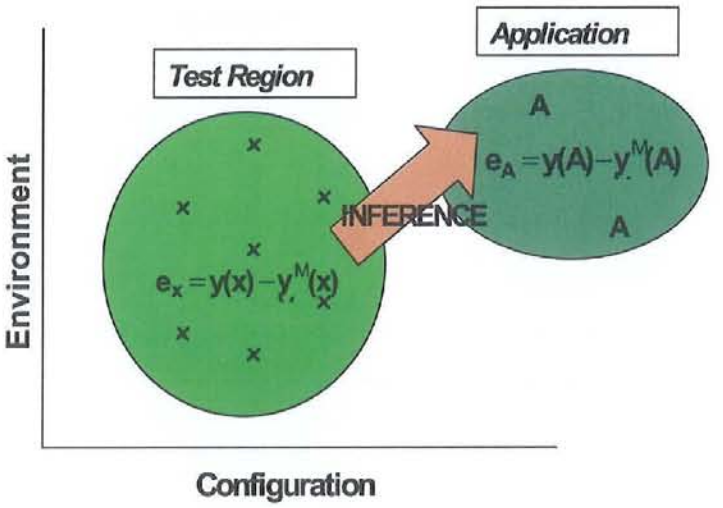

Figure 2. Inferring predictive capability

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 $\beta(x)$ and standard deviation $\sigma(x)$ (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 $\beta(x)$ and $\sigma(x)$ 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:

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

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

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

where $\bar{x}$ and $\bar{y}$ 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 $\sigma^2/\Sigma(x-\bar{x})^2$. The unknown variance, σ^2, is estimated by the "residual mean square," say s^2, and the quantity, $s/[\Sigma(x-\bar{x})^2]^{.5}$, 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

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

$$\approx y^{M}(x) - y^{M}(x+\Delta),$$

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.

Measuring Predictive Capability of Computational Models: Foam Degradation Case Study

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

Abstract

Statistical methods for evaluating the predictive capability of computational models are tested and illustrated for a Sandia National Laboratories case study pertaining to the degradation of polyurethane foam in a thermal environment. A newly developed computational model of this phenomenon is compared to a suite of nine experiments. The statistical analysis focuses on characterizing prediction-error as a function of experimental variables, primarily temperature. It is found that both predicted degradation-front velocity and the experimental data exhibit an approximate Arrhenius relationship, but with different slopes ("activation energies"). Statistical prediction intervals are obtained in each case and compared. The need for additional experimentation in order to resolve ambiguities is also discussed.

Introduction

In a 'foundations paper' presented at this conference, Easterling and Berger [2002, abbreviated as EB02 hereafter] present a statistical framework for designing, conducting, and analyzing the results from suites of model-validation experiments and computations. The goal of such programs, in the authors' view, is to characterize the predictive capability of the computational model – how close predictions are apt to be to nature's outcomes of the events being computationally simulated, especially of events that cannot be realized experimentally. That paper also identifies issues and challenges involved in arriving at credible, defensible, communicable evaluations of the predictive capability of computational models. To move from the abstract to the concrete, my purpose in the present paper is to illustrate an implementation of the foundational framework in EB02. This case study pertains to a set of experiments and computations that were conducted at Sandia National Laboratories to evaluate the predictive capability of a computational model of polyurethane foam decomposition in a thermal environment.

* This work 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.

It is important to note that the purpose of this case study is to emulate the process of measuring predictive capability and thereby test and illustrate methodology, not to arrive at a definitive evaluation of the predictive capability for the particular computational model in this study. My involvement in this project began two years after a suite of model-validation experiments had been designed and run. The in-depth interactions among modelers, experimenters, and analysts that are envisioned in EB02 to design and analyze a suite of experiments in the "real" case were thus not possible here. Additionally, only a portion of the experimental data was available at the time of this analysis. Thus, the analysis gives only an interim view of what can be said about the predictive-capability of the case study's computational model in the experimental context and perhaps in related applications. Model-validation is a process and it is anticipated that the collection of data pertaining to predictive capability will continue as more is learned about a model's predictive capability.

The constraints of this study, which translate into limited data, mean that there is a fair amount of ambiguity in the interim findings. This makes the study representative of what may be the real situation when we need to evaluate predictive-capability in complex situations perhaps well out of reach of experimental capabilities and resources. One major goal of my analysis is to communicate the limitations of the inferences that can be drawn from the data and to identify areas in which further experimentation would solidify the evaluation of predictive-capability.

Foam Experiments

Polyurethane foam is used to encapsulate nuclear weapon components and thereby provide structural support in shock environments. In an abnormal thermal environment, however, the insulating properties of foam can delay the failure of safety-critical components. Current safety analyses model this function of the foam somewhat crudely. To do better, a detailed chemical computer model of temperature-induced foam decomposition, termed CPUF [Hobbs, Erickson, and Chu 1999], was developed and incorporated into Sandia's Coyote finite element thermal code.

To "validate" the Coyote/CPUF computational model a set of 15 experiments in Sandia's Radiant Heat Facility were conducted [Bentz and Pantuso 1999]. Figure 1 depicts the experimental set-up. In these experiments a foam-encapsulated simulated component was exposed to a thermal environment produced at a base-plate interface to the foam. Factors that were varied in the suite of experiments across the indicated levels were:

- base plate temperature (600, 750, and 900C)
- heating orientation (overhead, side, bottom)
- foam density (low, high)

- internal component (none, stainless steel, aluminum)

When exposed to the specified thermal environments, the foam decomposes, starting at its interface with the heated base plate and progressing through the foam. X-ray imagery was used to track the advance of the decomposition-front vs. time over the course of the experiment.

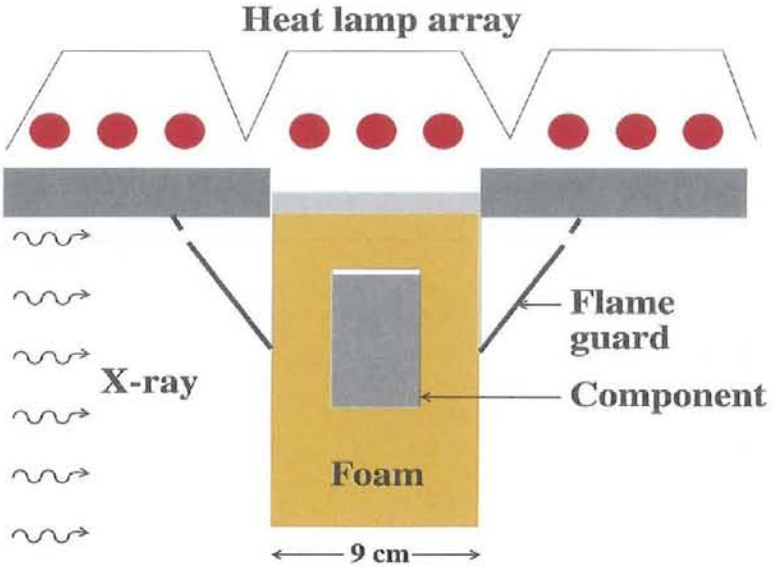

Figure 1. Foam Decomposition Experiments Schematic

Corresponding to each experiment, a "pure" Coyote/CPUF computational prediction of front dynamics was obtained. In these computations, boundary thermal conditions measured in the experiment were used as input because the intended temperature profile, which is to ramp from room to target temperature in 1.5 mins., then hold temperature constant for the duration of the experiment, cannot be achieved exactly. There are fluctuations above and below the target steady-state temperature. It was therefore deemed more appropriate to use the measured temperature profile as the boundary condition rather than the target profile as a better representation of the environment that the foam is subjected to. Other than this linkage between the experiment and the computation, the prediction is "pure" because the experimental results were not used to choose or adjust constitutive parameters in the model, such as the "activation energies" associated with some 16 chemical bonds in the foam. These model parameters, elements in the parameter vector φ (following the notation in EB02), were estimated from a separate set of foam-decomposition experiments designed for this "parameter-identification" purpose. At the time of this analysis, predictions and experimental results were available for nine experiments with high-density foam.

Figure 2 illustrates the computational and experimental results for three of the experiments. In the computational results, front position was defined as the axial distance from the base plate at which the calculated solid fraction of foam was 50%. Experimentally, measured front position was determined by

digitized gray-scale measurements of the x-ray images. Yogie Berra has been quoted as saying, "You can see a lot just by looking." The analysis will focus on the slope of the response curves and 'just looking' at Fig. 2 suggests that the computational model pretty well matches the slope at 750C but over-estimates it at 900C, under-estimates it at 600C. The rest of this analysis confirms and quantifies that impression and discusses what to do about it.

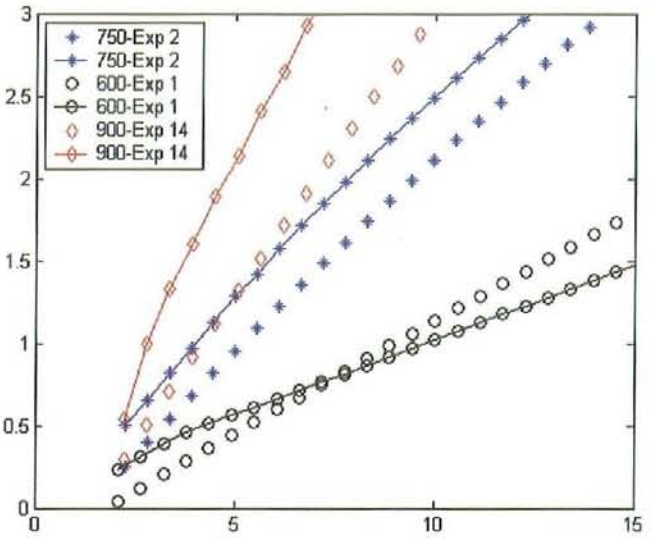

Figure 2. Results From Three Foam Decomposition Experiments: Front Position (cm.) vs. Time (min.). Computational (connected plotting symbols) and Experimental (unconnected symbols) Results.

Statistical Model

Let x be a vector that defines an experiment. In this study, x is the set of experimental variables listed above and includes other experiment-defining variables such as the dimensions of the experimental apparatus. The measured temperature-time histories at selected boundary points are also elements of x. Denote the experimental and computational results at x by $y^E(x)$ and $y^M(x)$. My approach to the analysis of predictive capability, as developed in EB02, is based on the statistical model:

$$y^E(x) = y^M(x) + e_x + \delta_x. \qquad (1)$$

where e_x is an unobservable random variable representing the difference between model prediction and nature's outcome at x and δ_x is an unobservable random variable representing measurement error in the experimental results. Both of these random variables have unknown probability distributions that possibly depend on x. This set-up is the conventional statistical paradigm: data equal signal plus noise, where the "plus" is in general a conceptual merging, not necessarily addition. Transformations of both y and x, however, can improve the linearity of the relationship and facilitate the analysis.

The rationale for this statistical approach, as discussed in EB02, is that there are a wide variety of random and systematic effects at play to make nature's outcome different from the predicted outcome. The vector x is a simplified representation of a complex situation in nature. Thus, other variables, not captured in the model, contribute to nature's outcome and these variables in general, depending on how the experiments are conducted, vary randomly from experiment to experiment. The nature of that variability is a consequence of the experimental materials, conditions, and controls. It is this extra-model variability that we want to characterize through model-validation experimentation and analysis. Further, we want to use what we learn about extra-model variability in the experimental program to infer the extra-model variability present in applications for which predictions will be made. In this case study, differences in foam composition from specimen to specimen are a source of extra-model variability. Additionally, the computational model itself is a mathematical approximation to nature and this is apt to be a source of systematic differences between model and nature. There is no guarantee that the expected values of the prediction and measurement errors will be zero.

To continue the EB02 framework and notation, there is another class of model arguments, called model parameters and denoted by the vector φ. These are constants in the equations used in calculating $y^M(x)$. For example, in CPUF the thermal effect on various chemical bonds in the foam is characterized via associated parameters called "activation energies." Such φ-parameters may be estimated experimentally or obtained from other sources such as handbooks of material properties. Estimation errors in φ can affect the data in different ways. For example, if actual φ varies from experiment to experiment, but the corresponding calculations all use the same φ-estimate, then the differences between actual and estimated φ contribute to the random variation of e_x across experiments If the same variability carries over to applications, then the analysis will capture this source of variation, but perhaps not isolate it. If actual φ is measured for each experiment, then predictions for each experiment would be calculated using each experiment's measured φ-value. That φ-value is in essence a measured boundary condition, as some of the x's may be. On the other hand, if actual φ and estimated φ are constant from experiment to experiment, as, e.g., one might reasonably presume for activation energies associated with particular chemical bonds (these parameters are nominal material properties, not specimen properties), then the difference between estimated φ and actual φ contributes to systematic prediction error. One can see that sorting all this out based on a small number of experiments may be difficult or impossible.

Under the statistical model, (1), measuring predictive capability becomes the estimation of characteristics of the probability distribution of e_x, such as its mean and standard deviation, as a function of x. Such estimates are statistical in nature – functions of limited, variable data, so the statistical properties of these estimates are important, both in conveying the reliability with which we can measure predictive capability and in designing experiments that will provide adequate reliability.

Measurement error associated with both $y^E(x)$ and measured x-values contribute to observed prediction errors. Such sources of both bias and variability are extraneous to the measurement of predictive capability – we want to characterize how well the model predicts nature, not how well it predicts a measurement of nature. It is apparent from expression (1) that prediction error, e_x, and measurement error, δ_x, cannot be separated on the validation set of experimental and computational results only. Additional data are needed on the measurement process that characterize the distribution of δ_x as a function of x in order to estimate the distribution of prediction error cleanly. Adjusting for the effect of measurement error introduces additional uncertainty in the characterization of prediction error so, again, it is important that the model-validaiton experiments be designed, conducted, and data-processed in ways that minimize extraneous measurement variability.

Statistical Analysis

a. Choice of Prediction Variable

When experimental and computational results are dynamic responses, such as in Fig. 2, the first choice to be made with respect to measuring predictive-capability is the choice of prediction variable (the "predictand" I'll call it). What characteristics of this response do we want to predict? One answer is: the whole curve. From a scientific perspective this is a worthy objective, in the sense that once we characterize how well the model predicts front position as a function of time, for any given set of experimental conditions, the predictive capability for any other predictand derived from the whole curve can be characterized. From an application-driven and practical perspective, however, this is excessive. For one thing, prediction errors, as in Fig. 2, are obviously highly correlated across time. Modeling that correlation, then estimating the model parameters (principally, variances and covariances) on quite limited data is apt to be both difficult and disappointing. Primarily, though, (and here I am simulating the customer's perspective) foam characteristics that are important in the system application do not require a detailed characterization of how well the whole curve can be predicted.

The choice of predictand should start by asking what characteristics of the foam decomposition front dynamics are important to system performance. For this illustrative case study, it seems reasonable to argue that the time to expose a component to an uninsulated thermal environment is a property of high interest. Thus, in an initial analysis [Easterling 2001b] I considered prediction of the time for the decomposition front to reach distance d, for $d = 0.5, 1.0, 1.5, 2.0$ cm, this set of distances corresponding roughly to possible system applications. This selection amounts to selecting four points off the response curve (Fig. 2), which is a considerable simplification compared to attempting to analyze the whole curve. Graphical analysis and further discussions with project personnel, however, indicated that apparent biases in the predictions could be due to the difficulty of lining up 'time-zero' values between the experimental

and computational results. To reduce the effect of this extraneous source of error the second predictand chosen was the front travel-time from 0.5 to 1.5 cm. The distance range selected corresponds to the typical amounts of foam protection in system applications. This travel-time is the reciprocal of front velocity over that one-centimeter range, a variable that is intuitively valuable in characterizing foam performance in applications of interest. The near-linearity of front progression, as shown in Fig. 2, further supports this choice of predictand. (I might have chosen velocity instead, but since the analysis led to taking the log transformation, the question is moot.) At any rate, the choice of predictand also has attributes of simplicity and communicability, which are important in any analysis of the prediction capability of sophisticated computational models of complex processes. Note also that this choice eliminates having to consider the complex, scientifically interesting, and computationally challenging interactions between the retreating foam and an exposed component. Because of initial concerns about some of the data, the illustrative analysis in [Easterling 2001] considered only five experiments.

Subsequent to my initial analysis [Easterling [2001], slight changes were made to the parameter estimates used in the model at that time (these estimates were not prompted by the results of that analysis, so we do not have a "tuned" model) and data from additional experiments became available. Thus, at this writing, results [Dowding 2002] are available for nine experiments, all with high-density foam. The prediction variable of interest used by Dowding [2002] is front velocity over the 1 – 2 cm range, so this analysis follows that precedent. Over the 1-2 cm range, the presence of a simulated component buried in the foam does not come into play, so data from three such experiments are included in the analysis.

b. Front Velocity Data

Table 1 gives the computational predictions and experimental results for front velocity, denoted by v^M and v^E. The observed prediction error, labeled e, is the difference between the experimental and computational outcomes. Because my previous analysis led to using the logarithmic transformation, the logarithmic errors, namely the natural log of the ratio of the experimental to computational outcomes, are also given in Table 1, denoted *lne*. Logarithmic errors are approximately equal to relative errors.

Table 1. Experimental Conditions, Computational Predictions, Experimental Results, Prediction Errors

Exp.	Temp.	Heat Orient.	Int'l. Comp	v^M	v^E	e	lne
2	750	bottom	none	0.246	0.232	-0.013	-0.056
5	750	bottom	SS slug	0.284	0.196	-0.088	-0.372
10	750	overhead	none	0.234	0.211	-0.023	-0.105
11	750	side	none	0.262	0.258	-0.004	-0.014
13	750	side	none	0.228	0.215	-0.012	-0.056
15	750	bottom	AL cyl.	0.284	0.275	-0.009	-0.030
1	600	bottom	none	0.091	0.131	0.039	0.358
14	900	bottom	none	0.450	0.349	-0.100	-0.253
16	1000	bottom	AL cyl.	0.770	0.558	-0.212	-0.322

The objective of the following analysis, as set forth in EB02, is to see if prediction error is related to the x-variables (temperature, orientation, internal component) defining the experiment and to characterize that relationship. I also want to simulate the inference process, so I will first analyze the data at 750C (the first six rows of Table 1, then make a leap-of-faith inference that the same error patterns apply from 600C to 1000C. I can then check these inferences against the data at other temperatures (the last three rows of Table 1). I will then do an analysis of all the Table 1 data and discuss possible inference beyond this data base.

The computational model does not model the effect of orientation and any effect of an internal component would not be manifested in the front velocity over the 1-2cm range. Thus, the variability of the predicted velocities reflects variability in the measured boundary conditions for nominally identical experiments. Qualitatively, the observed variability of v^M seems large. One potential contributor, which has not been resolved at the time of this analysis, is that the plate temperature was measured at two locations and some model predictions were based on using the temperature data from one location as the input boundary condition for the CPUF calculation, the other predictions were made using the other.

At 750C, the observed prediction error for experiment 5 is a distinct outlier; its ln-error is -.372 compared to the other five experiments that are tightly grouped around -.05. Discussions with project personnel indicate there may be problems with the model prediction or experiment in this case. Pending the resolution of these problems, I will exclude experiment 5 from this illustrative analysis. A statistical test for outliers would probably support rejecting experiment 5 as being inconsistent with the other 750C data, but the existence of a potential assignable cause is the real driver for this decision. (I would note that in a previous iteration with these data, experiment 11 was a distinct outlier also, but in the other direction. A problem was found in the model prediction and corrected, so that point, no longer an outlier and no longer having an assignable cause, is included in the analysis. It has been my experience that statistical eyeball examination of data often turns up problems that experimenters were either not aware of or thought were inconsequential.)

At 750C experiments 11 and 13 are replicates – the same nominal conditions were run twice. The measured boundary conditions for these two experiments differed and this difference is reflected in the different predicted velocities for these two experiments. Given the variability of the two *lne*'s for these two experiments, there does not appear to be any systematic (x-dependent) variability among the five 750C experiments. That is, neither orientation of the presence of a component appear, on these limited data, to have a systematic effect on prediction errors pertaining to the front-velocity over the 1 – 2 cm range.

c. Predictive Capability Analysis.

Based on the preceding discussion, and for the sake of illustration, I will treat the variability of the *lne*'s for the five 750C experiments as random, extra-model variability—that is, as indicative of the actual, variable difference between nature and model in these situations. Measurement error may contribute to these observed differences, so adjustment of the observed variability for the effect of measurement error will be addressed at the end of the analysis (subsection h). Also, because of experience on previous analyses of data from these experiments, I am going to deal with the logarithmic errors from the start.

Summary statistics for the 750C log-error data are:

$$\text{average} = -.052 \quad \text{std dev} = .034.$$

All five log-errors are negative, so there is some evidence that the model tends to over-predict front velocity. I first consider the extent to which this apparent bias (relative to an expected error of zero) is statistically significant. The test statistic for testing unbiasedness (under the assumption that the five observed prediction errors can be treated as a sample from a Normal distribution – an assumption not contradicted by these limited data) is

$$t = \sqrt{n} * (\text{ave/stdev}) = -3.40$$

where n = 5, which is the number of observations. Comparing this t-value to the Student's t distribution based on 4 degrees of freedom (df) (see any basic statistics text for a discussion of the Student's t distribution and its role in significance tests for the mean of a normal distribution;) shows that this result is fairly unusual (the critical value of t corresponding to the two-tail 5% significance level is 2.776 and there is only a 3% chance of a t-value, in absolute value, being as large as 3.40 under the hypothesis of unbiasedness). I will consider this finding to warrant the consideration of bias in the subsequent analysis. The finding of bias should lead to an investigation of possible causes. In this case, the inconsistent measurement of boundary conditions is a candidate, but is not resolved at this writing. The assumed φ-values used in the calculation may also be a source of bias. It should be noted that a single set of φ-values

was used for all the calculations. Thus, any randomness or "uncertainty" in the estimated φ-values does not contribute to the variability of the observed prediction errors.

Statistical Prediction Intervals

There are various ways to characterize predictive capability, given the preceding results. The most straightforward is a statistical prediction interval [Hahn and Meeker 1991]. A statistical prediction interval is a confidence interval on a single future prediction error. For the present case, a 90% prediction interval for a future error is derived from the 'pivotal' relationship,

$$(e - \text{ave})/\text{stdev}(1 + 1/n)^{.5} \sim t(4),$$

where e denotes a future observed error, ave and stdev denote the sample average and standard deviation, the symbol ~ means 'is distributed as' and $t(4)$ denotes the t distribution with 4 degrees of freedom. From this relationship, a 90% prediction interval, e.g., is given by ave +/- t(.05, f)* stdev$(1 + 1/n)^{.5}$, where t(.05, f) is the upper 5th percentile on the t distribution with f degrees of freedom. In this case, t(.05, 4) = 2.132, so 90% prediction limits on a single logarithmic error are -.052 +/- 2.132*.034*sqrt(6/5) = -.052 +/-..08 = (-.13, .03). The inference is that with 90% confidence, if another experiment and computational prediction were done for another 750C experiment like the five for which we have data, the logarithmic error for these two results will be in this interval. It would be an unusual event (probability less than .10) if it were not. On the velocity scale, by exponentiating these limits, this interval translates into multiplicative limits of (.88, 1.03). The ratio of the experimental result to the model prediction would fall between .88 and 1.03, with 90% confidence.

In terms of predicted front velocities, these prediction error results at 750C mean that for this situation, in which the computational prediction, v^M, was about .25 cm/min. (see Table 1), that with 90% confidence we would predict that the (measured) velocity in a future experiment like these (we have no basis for broadening the inference at this point) would fall within multiplicative factors of (.88, 1.03) of the predicted value of .25cm/min., namely .22 to .26cm/min. Whether this information is "good-enough" depends on requirements. If, hypothetically, satisfactory performance required only that velocity at 750C, for boundary conditions for which v^M = .25cm/min, be less than or equal to .3cm/min., then we have good evidence supporting a conclusion that the requirement is met. If the required velocity at 750C was to be less than or equal to .25cm/min., then we don't have that assurance. More data or arm-waving would be required. Or, a different foam. Or a design change to provide more insulation. The point to make is that if we did not have this statistical predictive-capability 'yardstick,' and all we had was a computational prediction of v^M = .25cm/min., trust-me, then we could not distinguish between the two cases and would be inclined to conclude that the requirement was met in both cases.

The ultimate goal of a predictive-capability analysis is to use what we learn about prediction error in experimental situations to infer prediction errors in untested conditions. To emulate this process, suppose we make a leap-of-faith inference, armed only with the assumption that variances will be rendered consistent across the temperature range by the log transformation, that this prediction-error interval applies over the whole experimental temperature range, 600C – 1000C. Because experiments have been done over that range, we can test this assumption. Figure 3 compares this prediction interval to the observed prediction errors in Table 1 and shows that the observed prediction errors at 600, 900, and 1000C are substantially outside of the 750C-based 90% prediction interval. Thus, if we had made an inference that logarithmic prediction errors over this temperature range would fall within (-.133, .028) we would be grossly in error. In Fig. 3, as remarked at the start of this analysis, there appears to be a temperature-dependent pattern to the prediction errors and this should be investigated. This problem swamps the problem of bias at 750C.

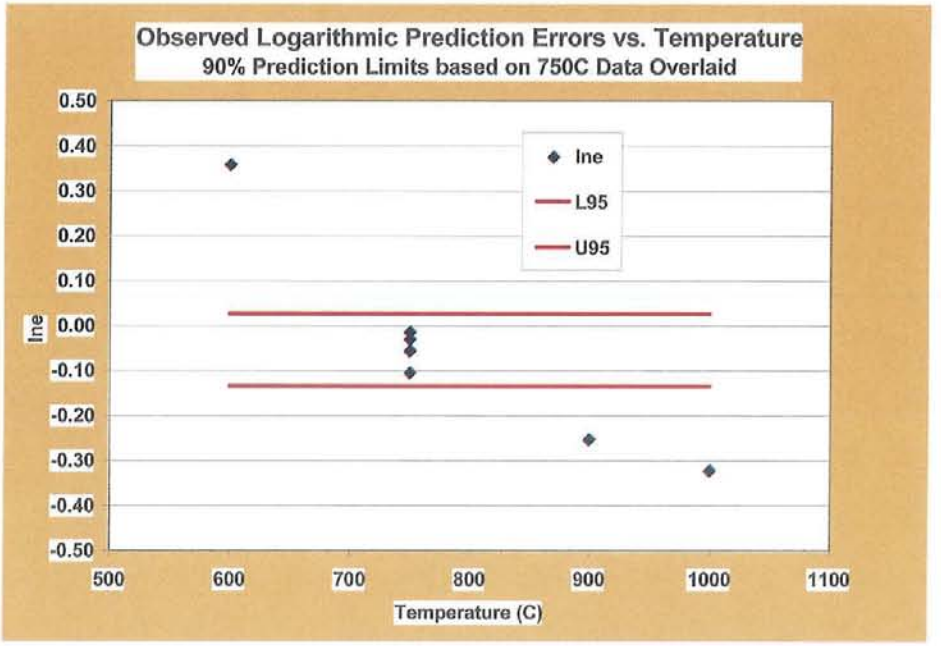

Figure 3. Comparison of Inferred Prediction Error to Observed.
(The upper and lower ends of the 90% prediction interval are denoted by U95 and L95)

An alternative way to characterize prediction capability, rather than a prediction interval for a single outcome, is by a statistical tolerance interval on percentiles of the error distribution. For example, methods exist [Hahn and Meeker 1991] to determine an interval such that, e.g., with 90% confidence 95% of the error distribution will fall within the interval. An example is given in [Easterling 2001a]. Bounding the center 95% of a distribution is more difficult than bounding a single observation, so these intervals will be wider than the prediction intervals at the same confidence level. The choice of which interval to use

depends on whether it is more pertinent to make an inference about a single outcome or a distribution of outcomes.

d. Analysis of the Temperature-Effect.

In any "real" predictive-capability data analysis it is important to incorporate subject-matter knowledge into the analysis. In this case, because front-velocity is related to a temperature-dependent reaction rate of the foam, my limited chemistry knowledge suggests an Arrhenius relationship [Hammes 1978]. Under this theoretical relationship, front velocity would be related to temperature by:

$$v \propto \exp[E/(\text{abs. Temp})],$$

where $\propto$ denotes "is proportional to." Taking logarithms means that *ln(v)* is a linear function of inverse absolute temperature.

To see whether this Arrhenius-based relationship is appropriate for the foam-decomposition phenomenon, I plotted all of the computational and experimental results in Table 1 (excluding expt. 5) on Arrhenius coordinates of *ln(v)* vs. 1/(abs. Temp) – see Figure 4. Note that this plot is based on the target steady-state temperature and so the variability of the actual boundary conditions is not accounted for. The variability of the v^M values for the nominal 750C experiments shows that this variability is not negligible (side calculations indicate that the range of v^M values corresponds to roughly a range of 40C in nominal base plate temperature). In principle, the measured base plate temperatures could be used to obtain a more accurate nominal steady-state temperature to use in the analysis, but such an analysis has not been done at this writing. Figure 4 shows that the v^M computational predictions are fairly well-fitted by a straight line on this scale; the fitted line is shown in Fig. 4. The deviations of the outer temperature *ln(v)* results from the fitted line are consistent with the variability of the predictions observed at 750C. Figure 4 also suggests a linear relationship for the experimental results, but with apparently a different slope (activation energy, in Arrhenius terminology). The analysis in this subsection is aimed at characterizing these patterns.

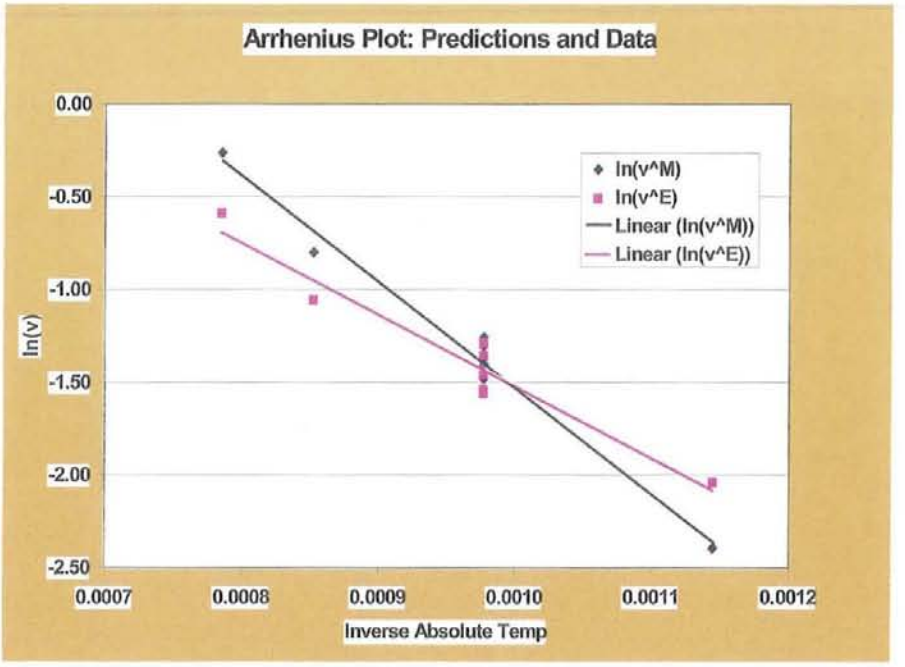

Figure 4. Computational Predictions and Experimental Results Plotted on Arrhenius Coordinates.

As mentioned in the above discussion of the statistical model for model-validation data, the objective of an analysis of predictive capability is to estimate the distribution of prediction error as a function of the *x*-variables in the model and experiment. The single x-variable of interest in this illustrative case study is temperature. Figure 5 plots the logarithmic errors for the eight experiments vs. inverse absolute (target) temperature. The plot, supported by regression, as well as eyeball, analysis, indicates that logarithmic prediction error is strongly and nearly linearly temperature-dependent. (There is some indication of curvature in Fig. 5, but that will not be pursued here.)

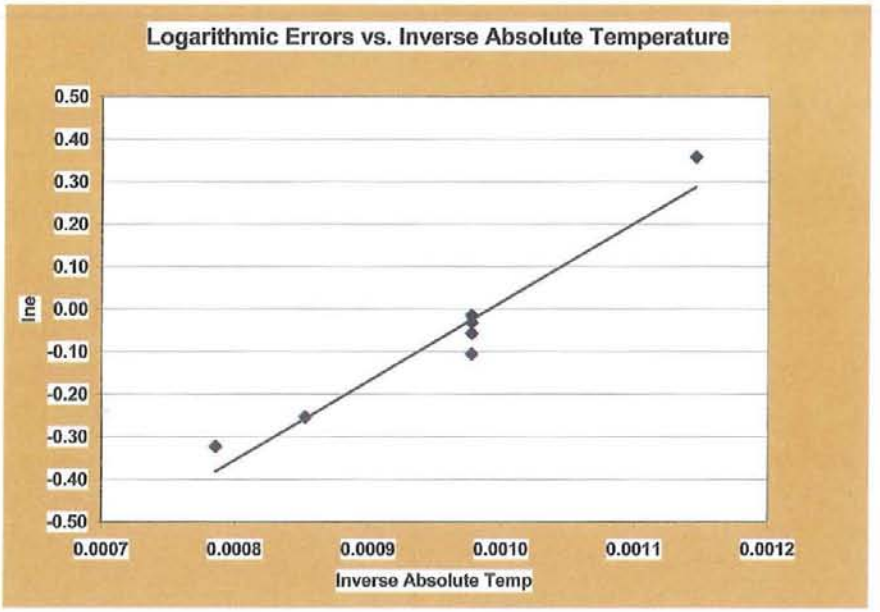

Figure 5. Logarithmic Prediction Errors vs. Inverse Absolute Temperature, with Fitted Straight Line

The regression analysis of the Fig. 5 data yields the following results:

Fitted line:	ln-error = -1.8 +1858*(1/Temp(K))
Standard error of the slope:	191, on 6 df.
Residual standard deviation:	s = .053

The slope is significantly different from zero at much less than the 0.1% level. Thus, by these data, taken as a whole, the observed deviations from predictions at the temperature extremes of 600C, 900C, and 1000C are systematic, not random. What next?

When a substantial bias between experimental results and computational predictions is found, the likely course of action is to look for and possibly remove the cause. The source of bias could be either in the experiment, the computational model, or both. If the experiments are absolved, then it is natural to focus on the model. A potential source of this sensitivity is the parameter values used in the computations. The parameters for the Coyote/CPUF model include 16 activation energies, for different chemical structures in the foam (another reason for my Arrhenius-based analysis), plus additional material properties. These parameters were estimated from a separate set of experiments conducted for that purpose. At this writing, there has been no attempt to remove the bias by modifying the parameter estimates or by making other changes in the CPUF model.

In general, for long-running computational models, it is desirable to conduct various analyses using a simpler approximate code. The near-linearity of the model predictions in Fig. 4 suggests that, for this set of

experiments and the selected predictand, the 24 hr., finite element CPUF calculation can be reasonably approximated by a straight line, which has two parameters. The regression analysis in the next subsection re-estimates these two (pseudo-) parameters and properly accounts for the tuning of the parameters to the data in deriving prediction limits that characterize the predictive capability of the tuned model.

Another potential reaction to bias is to make bias-corrected predictions at temperatures between 600C and 1000C by calculating the Coyote/CPUF prediction, $ln(v^M)$, then adding the estimated mean error at that temperature by the above fitted line, then adding and subtracting suitably calculated error bounds. Because of the near-linearity of the Coyote/CPUF predictions in the realm of these experiments, this prediction/correction process amounts to essentially ignoring Coyote/CPUF (except for its support for the Arrhenius relationship) and doing a regression analysis on the experimental results. This analysis follows next.

e. Regression Analysis.

Regression analysis of $ln(v^E)$ vs. inverse (target) absolute temperature yields the following results:

Fitted line: $ln(v) = 2.34 - 3858*(1/\text{Temp(K)})$

Residual standard deviation: s = .113, on 6 df.

The details will not be presented here, but statistical prediction intervals for regression can be obtained by methods given by [Hahn and Meeker 1991]. The results of that calculation, plotted on the original temperature-velocity scale of the data, are shown in Figure 6. The strong underlying assumption for Fig. 6 is that the logarithmic variance is constant across the experimental temperature range. Single observations at the extreme temperatures do not provide much statistical power for testing this assumption. Under the underlying statistical assumptions, the interpretation of Fig. 6 is that at a particular temperature within the experimental range, there is 90% confidence that the observed v^E in a future experiment like these would fall within the indicated interval. If these assumptions can be assumed to hold outside the experimental temperature range, then this sort of inference could be applied to extrapolation situations.

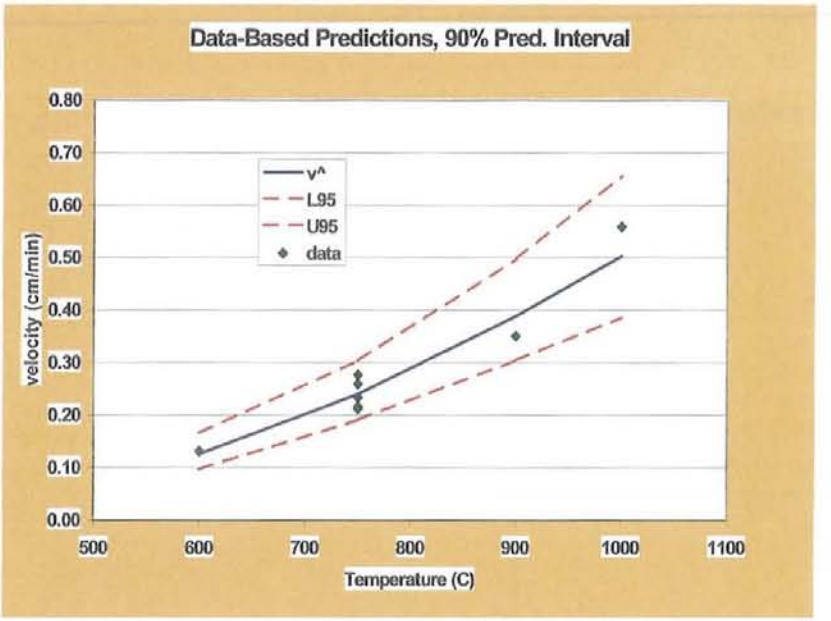

Figure 6. 90% Prediction Intervals Based on Regression Analysis of Experimental Data

f. Alternative Bias Correction Regression Analysis

The preceding analysis associates the experimental data with the target steady-state base plate temperature and thus does not account for the variability in the actual base plate temperature profile. It is also based on an assumed Arrhenius relationship. An alternative bias-correction method is to consider the regression of v^E on v^M. This model assumes that prediction error depends on x only through v^M. For high-dimensional x, this can be a highly-simplifying model, if the data support this simplification. For the present case of a single x-variable, no such simplification is provided, but this relationship will account for the variability of actual boundary conditions and it has the potential of smoothing out the slight nonlinearity seen in the previous analysis. Figure 7 plots v^E vs. v^M and also shows the fitted line. Clearly, the data do not fall along the 45 degree line, which again reflects the bias under discussion,. The regression analysis results are:

Fitted line:	$v^E = -.47 + .68v^M$
residual stdev:	.066 on 6 df
standard error of slope:	.041 on 6 df

Thus, by accounting for the variability of boundary conditions, the residual standard deviation is reduced by nearly 40%. Qualitatively, comparing Fig. 7 to Fig. 5 shows the advantages of this alternative model, in this case.

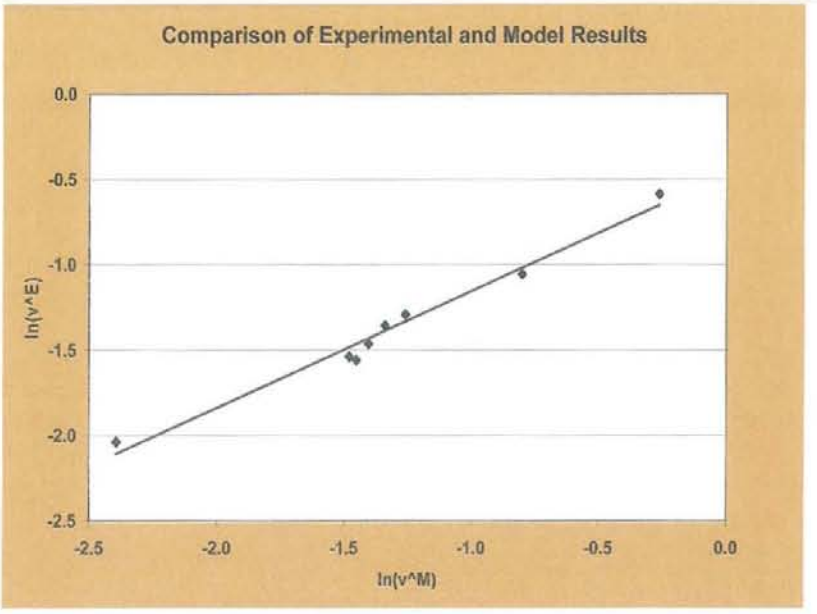

Figure 7. v^E vs. v^M and Fitted Line

Bias-corrected predictions would be obtained by calculating v^M, then substituting v^M into the fitted line equation. Statistical prediction intervals, obtained by the same procedure referred to in the previous subsection, are shown in Fig.8 in the original velocity scale. By comparing Fig. 8 to Fig. 6 one can see the improved precision achieved.

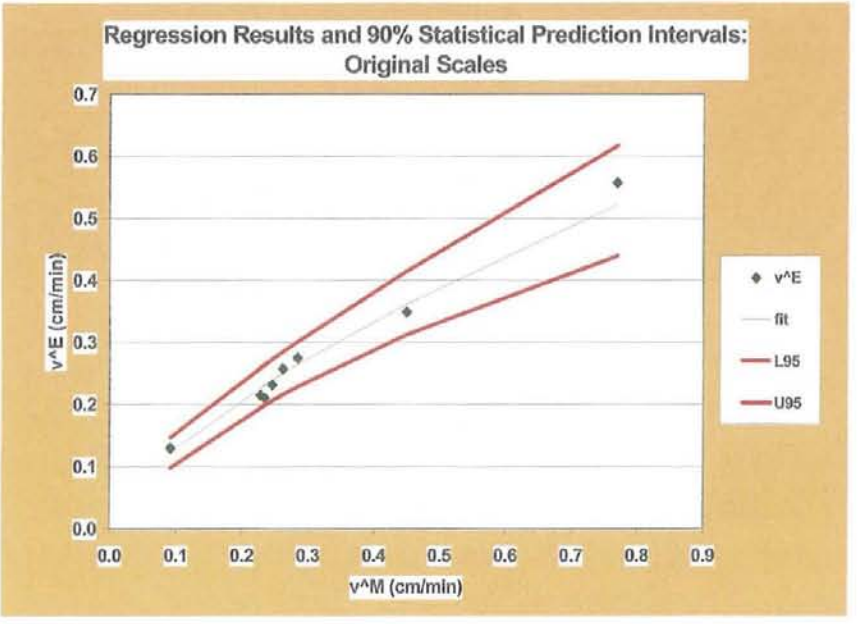

Figure 8. 90% Statistical Prediction Interval for v^M Model

It should be noted, though, that by using the measured base plate temperature profile, it might be possible to determine an effective base plate temperature for each experiment that could be used in the Arrhenius-based regression analysis, in place of the target base plate temperature, and improve its fit comparably to that achieved here.

g. Adjustments for Measurement Error

The variance of the observed prediction errors, as calculated in the preceding analyses, can be inflated by measurement error. We can get both more appropriate and tighter limits on actual prediction error (nature – model) if we can quantify and remove these sources of variation.

If x-variables measured in the experiment and then used in the computation of v^M have measurement-error variability in them, then this variability is transmitted into variability in v^M. For these experiments, measured temperature is used in obtaining the computational prediction. The sensitivity of $ln(v^M)$ to temperature measurement error can be gauged as follows. At 900C, $ln(v^M) = -.80$ and at 600C, $ln(v^M) = -2.39$. Thus, the temperature sensitivity coefficient, over this range, is $(-.80 + 2.39)/300C = .0053$/deg.C. Now, suppose the standard deviation of temperature measurement error is σ_T. The resulting measurement-error induced standard deviation of $ln(v^M)$ is $.0053\sigma_T$. Suppose, for the sake of illustration, based on thermocouple measurement-capability data, that a reasonable estimate is $\sigma_T = 2$ deg.C. Then the resulting standard deviation of $ln(v^M)$ due to temperature measurement error is .011.

Now consider measurement error for the experimental outcome, $ln(v^E)$. At this writing, an informed evaluation of the sort of measurement error variability associated with interpreting x-ray imagery has not been done, but suppose that such an analysis led to the conclusion that this $ln(v^E)$ measurement error standard deviation associated with log front-velocity was thought to be about .02 (i.e., measured front velocity is assumed to have a relative standard deviation of about 2%). Because velocity is measured by taking a difference between two front position measurements one would expect major sources of error to cancel, thus leaving what should be a fairly precise velocity measurement. Then, under this assumption, the combined measurement-error variance is $.011^2 + .02^2 = .023^2$. In the regression analysis in the previous section, the estimated residual variance is $.066^2$. This variance estimates the sum of the variances of prediction error and measurement error. Thus, an adjusted prediction error standard deviation estimate, adjusting for the effects of measurement error, would be

$$s_{adj} = \sqrt{(..066^2 - .011^2 - .02^2)} = .062.$$

The previous prediction intervals could be appropriately scaled down to reflect this adjustment, but the effect would be small. Thus, for this illustrative case, measurement error is a negligible contributor to the

variation of observed prediction errors. Of course, if there was reason to assume that the $ln(v^E)$ measurement-error standard deviation was .06, then experimental measurement error would essentially account for all of the observed prediction error variability. We would then conclude prediction error was all bias, with negligible variability. Because of the wide range of possible effects of adjusting for measurement error, it is important to have good information on the accuracy and precision of the measurement process.

Further Experimentation

Several questions arose in the previous section's analyses that might be answered with further experimentation. The experimental design of the eight high-density foam experiments on which the preceding analysis of predictive-capability is primarily built is given by Table 2.

Table 2. Experimental Design: HD Foam

Temperature (degC)

Orient.	600	750	900	1000
Bottom	1	2	1	1
Side		2		
Top		1		

This is a 'classic' one-factor-at-a-time (OFAT) experimental design in which all but one factor is held fixed while one factor is varied, sequentially, across a set of factors. In Table 2, the base condition is 750C, bottom heating, from which we horizontally consider other temperatures, then vertically, other orientations. An additional factor, not shown because it apparently had no effect on early front-velocity is the presence or absence of an internal component. Unfortunately, this design is an inefficient way to evaluate the effects of multiple factors.

One way to build on the existing experimental results is to run experiments at the six untested combinations of temperature and orientation in Table 2. By this less than doubling of the number of experiments information about the effects of temperature and orientation on decomposition-front characteristics will increase substantially. For example, instead of having only two experiments by which to evaluate the difference between 600C and 900C there will be six, three at each temperature. Under the assumption that the effects of temperature and orientation on $ln(v)$ are linear, this set of 15 experiments will also provide a means of checking the assumption made in the above analysis that the observed orientation differences were random.

In considering additional experimentation it is appropriate to consider additional variables. For example, situations in which the container for a foam-encapsulated component is hermetically sealed, thus leading to pressurization effects on the decomposition process, may be of more applications-related interest. If so, further experiments should include a pressurization variable, included in the experimental design as a whole, not added as a further OFAT set of experiments.

Discussion

The statistical analysis of the results of eight experiments and corresponding computational predictions, aided and abetted by subject-matter-motivated Arrhenius modeling, have led to the conclusion that the Coyote/CPUF model, at least as parameterized for these calculations, gives predictions in which there is a temperature-related bias. Beyond characterizing predictive-capability in experiments such as these, our ultimate interest is in characterizing predictive-capability for predictions made for foam (and ultimately component) performance in a system-in-a-fire environment. Whether the sort of extra-model variability observed in these isothermal, unconfined, end-on exposure experiments can be extended directly, or via some sort of scaling, to more complex system-in-a-fire induced environments for foam-encapsulated components will require careful study. The nature and possible effect of the added dimensions of a system-in-a-fire environment will need to be analyzed. More predictive-capability experiments, computational predictions, and analyses may be required. The experience here, in which predictive capability observed at 750C could not reliably extended to other temperatures, is not encouraging. On the other hand, an intuitive argument might be made that foam is foam and any fluctuating temperature profile can be enveloped by an isothermal environment, so the predictive capability exhibited in these and perhaps additional similar experiments is applicable to system-in-a-fire environment predictions of foam performance. Inference beyond the temperature range of these experiments, though, is problematical. Again, the purpose of this case study is to identify issues that need to be addressed in a real predictive-capability analysis, not to resolve them all in this particular case.

This case study illustrates the thought processes and statistical analysis tools that should be involved in using model-validation experiments to characterize the predictive capability of computational models. It also illustrates the imprecision and ambiguity that can occur when only a limited amount of experimentation is available for analyzing and evaluating predictive capability for computational models of complex phenomena. In this regard, the case study is a good model for the sort of high-level experiments that may be possible for complex, system-level performance prediction.

This case study also illustrates that the trade-off between potential computational and experimental efforts needs to be considered in setting priorities and funding. The data used to evaluate predictive capability can also be used directly to develop semi-empirical models of the phenomena that a computational science-

based model is designed to address. While a science-based model provides a stronger basis for predicting the outcome of untested applications, it must be recognized that the attendant extension of measures of prediction error, which are due to unmodeled phenomena, is also empirical and not science-based. To be more concrete, if system engineers cared only about the early front-velocity of a particular foam in various environments, would it be more cost-effective to run a set of experiments designed strictly to characterize front-velocity as a function of important environmental variables or to develop a science-based computational model, then do a suite of model-validation experiments to characterize its predictive capability over the same range of environmental variables? With limited resources, one might opt for the former in some situations.

Acknowledgments

I am grateful to T. Y. Chu, who sponsored the foam experimentation program and made its results available for this case study, Mike Hobbs, who developed the CPUF model and provided the computational and experimental results for analysis, and Kevin Dowding for data analysis assistance and discussions pertaining to the measurement of predictive capability in this situation.

References

Bentz, J. and Pantuso, J., *Letter Report for the Thermal Degradation of Polyurethane Foam at Radiant Heat Facility*, November 1999

Dowding, K. personal communication. August, 2002.

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

Easterling, R. G., Measuring Predictive Capability of Computational Models: Foam Case Study (internal Sandia report), July 2001

Easterling, R. G., and Berger, J., *Statistical Foundations for the Validation of Computer Models*, V&V Foundations, October, 2002

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

Hobbs, M. L., Erickson, K. L., and Chu, T. Y., *Modeling Decomposition of Unconfined Rigid Polyurethane Foam,* SAND99-2758, November 1999.

Owen, D. B., *Factors for One-Sided Tolerance Limits and for Variables Sampling Plans*, SCR-607 (Sandia Corporation Monograph), March 1963.

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

This page is intentionally left blank

External Distribution:

1 Charles E. Anderson
Southwest Research Institute
P. O. Drawer 28510
San Antonio, TX 78284

1 Mark Anderson
ACTA
2790 Skypark Dr., Suite 310
Torrance, CA 90505-5345

1 Bilal Ayyub
Department of Civil Engineering
University of Maryland
College Park, MD 20742

1 Osman Balci
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

1 Steven Batil
Dept. of Aerospace &
Mechanical Engr.
University of Notre Dame
Notre Dame, IN 46556

1 Charles K. Bayne
Oak Ridge Nat'l. Lab
Comp. Sci. and Math. Div.
PO Box 2008, Bldg 6012
Oak Ridge, TN 37831-6367

1 David Belk
WL/MNAA
101 W. Eglin Blve., Suite 219
Eglin AFB, FL 32542-6810

1 Ted Belytschko
Department of Mechanical
Engineering
Northwestern University
2145 Sheridan Road
Evanston, IL 60208

1 James Berger
Institute of Statistics and Decision
Science
Duke University
Box 90251
Durham NC 27708-0251

1 Pavel A. Bouzinov
ADINA R&D, Inc.
71 Elton Avenue
Watertown, MA 02472

1 John A. Cafeo
General Motors R&D Center
Mail Code 480-106-256
30500 Mound Road
Box 9055
Warren, MI 48090-9055

1 K. Casteel
University of Texas at El Paso
600 W. University Ave.
El Paso, TX 79968-0521

1 James C. Cavendish
General Motors R&D Center
Mail Code 480-106-359
30500 Mound Road
Box 9055
Warren, MI 48090-9055

1 Thomas Chwastyk
U.S. Naval Research Lab.
Code 6304
4555 Overlook Ave., SW
Washington, DC 20375-5343

1 Hugh Coleman
Department of Mechanical &
Aero. Engineering
University of Alabama/Huntsville
Huntsville, AL 35899

1 Raymond Cosner
Boeing-Phantom Works
MC S106-7126
P. O. Box 516
St. Louis, MO 63166-0516

1 Frank Dean
Strategic Systems Programs
Nebraska Avenue Complex
287 Somers Court NW,
Suite 10041
Washington, DC 20393-5446

1 David Dolling
Department of Aerospace
Engineering
& Engineering Mechanics
University of Texas at Austin
Austin, TX 78712-1085

1 Isaac Elishakoff
Dept. of Mechanical Engineering
Florida Atlantic University
777 Glades Road
Boca Raton, FL 33431-0991

1 John Fortna
ANSYS, Inc.
275 Technology Drive
Canonsburg, PA 15317

1 Mike Giltrud
Defense Threat Reduction
Agency
DTRA/CPWS
6801 Telegraph Road
Alexandria, VA 22310-3398

1 James Glimm
Dept. of Applied Math &
Statistics
P138A
State University of New York
Stony Brook, NY 11794-3600

1 James Gran
SRI International
Poulter Laboratory AH253
333 Ravenswood Avenue
Menlo Park, CA 94025

1 Bernard Grossman
Dept. of Aerospace &
Ocean Engineering
Mail Stop 0203
215 Randolph Hall
Blacksburg, VA 24061

1 Barbara Guttman
Stat. Eng. Division
Room 353 Bldg 820
NIST
Gaithersburg, MD 20899

1 Sami Habchi
CFD Research Corp.
Cummings Research Park
215 Wynn Drive
Huntsville, AL 35805

1 Raphael Haftka
Dept. of Aerospace and
Mechanical Eng. and Engineering
Science
P. O. Box 116250
University of Florida
Gainesville, FL 32611-6250

1 Achintya Haldar
Dept. of Civil Engineering
& Engineering Mechanics
University of Arizona
Tucson, AZ 85721

1 Tim Hasselman
ACTA
2790 Skypark Dr., Suite 310
Torrance, CA 90505-5345

1 George Hazelrigg
Room 508N
4201 Wilson Blvd.
Arlington, VA 22230

1 David Higdon
Institute of Statistics and
Decision Science
Duke University
Box 90251
Durham NC 27708-0251

1 Richard Hills
College of Engrg., MSC 3449
New Mexico State University
P. O. Box 30001
Las Cruces, NM 88003

1 F. Owen Hoffman
SENES
102 Donner Drive
Oak Ridge, TN 37830

1 Leo Kadanoff
Research Institutes Building
University of Chicago
5640 South Ellis Ave.
Chicago, IL 60637

1 Alan Karr
Nat'l. Inst. of Statist. Sciences
PO Box 14006
Res. Triangle Park, NC 27709

1 Hyoung-Man Kim
Boeing Company
M/S: ZC-01
502 Gemini Ave.
Houston, TX 77058

1 W. K. Liu
Northwestern University
Dept. of Mechanical Engineering
2145 Sheridan Road
Evanston, IL 60108-3111

1 Robert Lust
General Motors R&D
MC 480-106-256
30500 Mound Road
Warren, MI 48090-9055

1 Sankaran Mahadevan
Dept. of Civil &
Env. Engineering
Vanderbilt University
Box 6077, Station B
Nashville, TN 37235

1 Hans Mair
Institute for Defense Analysis
Operational Evaluation Division
1801 North Beauregard Street
Alexandria, VA 22311-1772

1 W. McDonald
Naval Surface Warfare Center
Code 420
101 Strauss Avenue
Indian Head, MD 20640-5035

1 Gregory McRae
Dept. of Chemical Engineering
Massachusetts Institute of
Technology
Cambridge, MA 02139

1 Michael Mendenhall
Nielsen Engrg. & Res., Inc.
510 Clyde Ave.
Mountain View, CA 94043

1 Sue Minkoff
Dept. of Mathematics and
Statistics
University of Maryland,
Baltimore Co.
1000 Hilltop Circle
Baltimore, MD 21250

1 Max Morris
Department of Statistics
Iowa State University
304A Snedecor-Hall
Ames, IW 50011-1210

1 Paul Muessig
Naval Air Warfare Center
Weapons Division, Code
418000D
1 Administration Circle
China Lake, CA 93555-6100

1 Vijay Nair
Department of Statistics
4062 Frieze Bldg
University of Michigan
Ann Arbor, MI 48109

1 NASA/Ames Research Center
Attn: U. B. Mehta
MS: T27 B-1
Moffett Field, CA 94035-1000

1 NASA/Glen Research Center
Attn: Chris Steffen, MS 5-11
Cleveland, OH 44135

6 NASA/Langley Research Center
Attn: Michael Hemsch, MS 280
Jim Luckring, MS 280
Ahmed Noor, MS 369
Sharon Padula, MS 159
Manuel Salas
J. Weilmuenster, MS 408
Hampton, VA 23681-0001

1 Robert Nelson
Dept. of Aerospace & Mech.
Eng.
University of Notre Dame
Notre Dame, IN 46556

1 Efstratios Nikolaidis
MIME Dept.
4035 Nitschke Hall
University of Toledo
Toledo, OH 43606-3390

1 Tinsley Oden
Texas Institute of Comp.
Mechanics
University of Texas at Austin
Austin, TX 78712

1 Michael Ortiz
Graduate Aeronautical Labs
California Inst. of Technology
1200 E. California Blvd.
MS105-50
Pasadena, CA 91125

1 Dale Pace
Applied Physics Laboratory
Johns Hopkins University
111000 Johns Hopkins Road
Laurel, MD 20723-6099

1 Allan Pifko
George Court
Melville, NY 11747

1 Cary Presser
Process Measurements Div.
NIST
Bldg. 221, Room B312
Gaithersburg, MD 20899

1 Chris Rahaim
SAIC
14 E.Washington St., Suite 401
Orlando, FL 32801-2320

1 Pradeep Raj
Computational Fluid Dynamics
Lockheed Martin Aeronautical
Sys.
86 South Cobb Drive
Marietta, GA 30063-0685

1 John Ramberg
Systems and Ind. Engineering
Engrg. Building #20, Room 111
P.O. Box 210020
Tucson, AZ 85721-0020

1 J. N. Reddy
Dept. of Mechanical Engineering
Texas A&M University
ENPH Building, Room 210
College Station, TX 77843-3123

1 John Renaud
Dept. of Aerospace & Mech. Engr.
University of Notre Dame
Notre Dame, IN 46556

1 Patrick J. Roache
1108 Mesa Loop NW
Los Lunas, NM 87031

1 Tim Ross
Dept. of Civil Engineering
University of New Mexico
Albuquerque, NM 87131

1 Jerry Sacks
Nat'l. Inst. of Statist. Sciences
PO Box 14006
Res. Triangle Park, NC 27709

1 Sunil Saigal
Carnegie Mellon University
Department of Civil and Environmental Engineering
Pittsburgh, PA 15213

1 Len Schwer
Schwer Engrg. and Consulting
6122 Aaron Court
Windsor, CA 95492

1 Paul Senseny
Factory Mutual Research Corp.
1151 Boston-Providence Turnpike
P.O. Box 9102
Norwood, MA 02062

1 Mark Shephard
Rensselaer Polytechnic Institute
Scientific Computation Research Center
Troy, NY 12180-3950

1 T. P. Shivananda
Bldg. SB2/Rm. 1011
TRW/Ballistic Missiles Division
P. O. Box 1310
San Bernardino, CA 92402-1310

1 Don Simons
RDA Lobicon
6053 W. Century Blvd.
P.O. Box 92500
Los Angeles, CA 90009

1 Munir Sindir
Boeing Company, MC IB39
Rocketdyne Propulsion & Power
P. O. Box 7922
6633 Canoga Avenue
Canoga Park, CA 91309-7922

1 Ashok Singhal
CFD Research Corp.
Cummings Research Park
215 Wynn Drive
Huntsville, AL 35805

1 Bill Spencer
Dept. of Civil Engineering and Geological Sciences
University of Notre Dame
Notre Dame, IN 46556-0767

1 D. E. Stevenson
Computer Science Department
Clemson University
442 Edwards Hall, Box 341906
Clemson, SC 29631-1906

1 Ben Thacker
Southwest Research Institute
6220 Culebra Road
Postal Drawer 28510
San Antonio, TX 78228-0510

1 Simone Youngblood
DOD/DMSO
Technical Director for
VV&A1901 N. Beauregard St.,
Suite 504
Alexandria, VA 22311

1 M. A. Zikry
North Carolina State University
Mechanical & Aerospace
Engineering
2412 Broughton Hall, Box 7910
Raleigh, NC 27695

Foreign Distribution:

1 Yakov Ben-Haim
Dept. of Mech. Engineering
Technion-Israel Institute of Tech.
Haifa 32000
ISRAEL

1 Graham de Vahl Davis
CFD Research Laboratory
University of NSW
Sydney, NSW 2052
AUSTRALIA

1 Luis Eca
Instituto Superior Tecnico
Dept. of Mec. Engineering
Av. Rovisco Pais
1096 Lisboa CODEX
PORTUGAL

1 Charles Hirsch
Dept. of Fluid Mechanics
Vrije Universiteit Brussel
Pleinlaan, 2
B-1050 Brussels
BELGIUM

1 Igor Kozin
Systems Analysis Dept.
Riso National Laboratory
P.O. Box 49, DK-4000 Roskilde
DENMARK

1 Malcolm Wallace
National Engrg. Laboratory
East Kilbride
Glasgow G&% 0QU
UNITED KINGDOM

4 Department of Energy
Attn: Kevin Greenaugh, DP-53
Juan Meza, DP-51
William Reed, DP-51
John Garcia, DP-53
Department of Energy
1000 Independence Ave., SW
Washington, DC 20585

20 Lawrence Livermore National Laboratory
7000 East Ave.
P. O. Box 808
Livermore, CA 94550
ATTENTION:
Thomas F. Adams, MS L-095
Steven Ashby, MS L-561
John Bolstad, MS L-023
Peter N. Brown, MS L-561
T. Scott Carman, MS L-031
Randy Christensen, MS L-160
Richard Klein, MS L-023
Byung S. Lee, MS L-550
Kirk Levedahl, MS L-016
Roger Logan, MS L-125
C. F. McMillan, MS L-098
Christian Mailhiot, MS L-055
James F. McEnerney, MS L-023
G. Michael Murphy, MS L-650
Daniel Nikkel, MS L-342
Cynthia Nitta, MS L-096
Douglas Post, MS L-038
Peter Raboin, MS L-125
Peter Terrill, MS L-125
Charles Tong, MS L-560

1 Argonne National Laboratory
Attn: Paul Hovland
MCS Division
Bldg. 221, Rm. C-236
9700 S. Cass Ave.
Argonne, IL 60439

38 Los Alamos National Laboratory
Mail Station 5000
P.O. Box 1663
Los Alamos, NM 87545
ATTENTION:
Peter Adams, MS B220
Dick Beckman, MS F600
Terrence Bott, MS K557
Dominic Cagliostro, MS F645
David Crane, MS P946
John F. Davis, MS B295
Scott Doebling, MS P946
Stephen Eisenhawer, MS K557
Dawn Flicker, MS F664
George T. Gray, MS G755
Michael Hamada, MS F600
Ken Hanson, MS P940
Rudolph Henninger, MS D413
Brad Holian, MS B268
Kathleen Holian, MS B295
Darryl Holm, MS B284
James Hyman, MS B284
Michael E. Jones, MS B259
Cliff Joslyn, MS B265
James Kamm, MS D413
Jeanette Lagrange, MS D445
Sallie Keller-McNulty, MS F600
Ken Koch, MS F652
Len Margolin, MS D413
Harry Martz, MS F600
Mike McKay, MS F600
Mark P. Miller, MS P946
John D. Morrison, MS F602
Maysa Peterson-Schnell, MS B295
William Rider, MS D413
Tom Seed, MS F663
David Sharp, MS B213
Richard N. Silver, MS D429
Ronald E. Smith, MS J576
Christine Treml, MS H851
Daniel Weeks, MS B295
Morgan White, MS F663
Alyson G. Wilson, MS F600

Sandia National Laboratories

1	MS	0136	Carl F. Oliver, 9000
1	MS	0139	Pete Wilson, 9902
1	MS	0143	John Kelly, 9904
1	MS	0145	Ron Bentley, 9700
1	MS	0301	J. L. McDowell, 15400
1	MS	0318	G. S. Davidson, 9201
1	MS	0318	P. D. Heermann, 9215
1	MS	0318	C. F. Diegert, 9215
1	MS	0321	W. J. Camp, 9200
1	MS	0321	A. L. Hale, 9224
1	MS	0405	T. R. Jones, 12333
1	MS	0412	Bob Paulsen, 9817
1	MS	0415	Keith Almquist, 9811
10	MS	0829	R. Easterling, 12323
1	MS	0417	Bill Ling, 9813
1	MS	0417	J. D. Rogers, 9813
1	MS	0421	R. J. Detry, 9800
1	MS	0421	W. C. Hines, 9810
1	MS	0425	Stan Fraley, 9814
1	MS	0428	D. D. Carlson, 12300
1	MS	0428	V. J. Johnson, 12301
1	MS	0429	J. S. Rottler, 2100
1	MS	0434	R. J. Breedin, 12334
1	MS	0435	M. L. Abate, 2102
1	MS	0447	J. O. Harrison, 2111
1	MS	0453	P. A. Sena, 2104
1	MS	0475	R. C. Hartwig, 2105
1	MS	0481	M. A. Rosenthal, 2114
1	MS	0481	W. C. Moffatt, 2167
1	MS	0481	K. D. Meeks, 2131
1	MS	0481	K. Ortiz 2168
1	MS	0483	T. Hernandez, 2112
1	MS	0490	J. A. Cooper, 12331
1	MS	0490	P. E. D'Antonio, 12331
1	MS	0492	D. R. Olson, 12332
1	MS	0501	G. E. Boettcher, 2616
1	MS	0516	J. P. Brainard, 2564
1	MS	0521	R. A. Damerow, 2561
1	MS	0553	R. A. May, 9126
1	MS	0555	M. S. Garrett, 9122
1	MS	0557	T. J. Baca, 9125
1	MS	0601	Eric D. Jones, 1123
1	MS	0632	J. C. Hogan, 2907
1	MS	0638	M. Blackledge, 12326
1	MS	0638	D. E. Peercy, 12326
1	MS	0708	P. S. Veers, 6214
1	MS	0718	C. D. Massey, 6141
1	MS	0737	P. A. Davis, 6114
1	MS	0744	Dana A. Powers, 6400
1	MS	0746	R. M. Cranwell, 6411
1	MS	0747	R. L. Camp, 6410
1	MS	0747	G. D. Wyss, 6410
1	MS	0750	S. E. Minkoff, 6116
1	MS	0751	L. S. Costin, 6117
1	MS	0752	James Gee, 6201
1	MS	0755	Bernard Zak, 6233
1	MS	0779	J. C. Helton, 6849
1	MS	0806	L. G. Pierson, 9336
1	MS	0819	T. G. Trucano, 9211
1	MS	0819	E. A. Boucheron, 9231
1	MS	0820	P. Yarrington, 9232
1	MS	0825	W. H. Rutledge, 9115
1	MS	0826	W. Hermina, 9113
1	MS	0826	D. K. Gartling, 9100
1	MS	0827	J. E. Johannes, 9114
1	MS	0827	K. S. Chen, 9114
1	MS	0827	R. O. Griffith, 9117
1	MS	0827	J. D. Zepper, 9131
1	MS	0828	R. K. Thomas, 9102
1	MS	0828	K. V. Chavez, 9132
1	MS	0828	J. L. Moya, 9132
1	MS	0828	S. N. Burchett, 9132
1	MS	0828	T. Y. Chu, 9132
1	MS	0828	M. Pilch, 9133
1	MS	0828	B. F. Blackwell, 9133
1	MS	0828	K. J. Dowding, 9133
1	MS	0828	A. R. Lopez, 9133
1	MS	0828	K. E. Metzinger, 9133
1	MS	0828	W. L. Oberkampf, 9133
1	MS	0828	T. L. Paez, 9133
1	MS	0828	C. Romero, 9133
1	MS	0828	V. J. Romero, 9133
1	MS	0828	A. Urbina, 9133
1	MS	0828	W. R. Witkowski, 9133
1	MS	0829	J. M. Sjulin, 12323
1	MS	0829	F. W. Spencer, 12323
1	MS	0829	B. M. Rutherford, 12323

1	MS	0829	K. V. Diegert, 12335
1	MS	0834	A. C. Ratzel, 9112
1	MS	0834	Tze Yao Chu, 9100
1	MS	0835	S. N. Kempka, 9111
1	MS	0835	R. J. Cochran, 9111
1	MS	0835	J. S. Peery, 9121
1	MS	0836	E. S. Hertel, 9116
1	MS	0836	W. Gill, 9116
1	MS	0836	S. R. Tieszen, 9116
1	MS	0836	Melvin Baer, 9100
1	MS	0836	C. W. Peterson, 9100
1	MS	0839	Dennis Engi, 16000
1	MS	0841	T. C. Bickel, 9100
1	MS	0841	J. A. Fernandez, 9102
1	MS	0847	H. S. Morgan, 9123
1	MS	0847	A. F. Fossum, 9123
1	MS	0847	D. R. Martinez, 9124
1	MS	0847	K. F. Alvin, 9124
1	MS	0847	T. B. Carne, 9124
1	MS	0847	R. V. Field, 9124
1	MS	0847	S. Y. Chakerian, 9211
1	MS	0847	M. S. Eldred, 9211
1	MS	0847	J. R. Red-Horse, 9211
1	MS	0847	R. W. Leland, 9226
1	MS	0886	D. M. Haaland,1812
1	MS	0977	S. M. DeLand, 6524
1	MS	0978	R. E. Spalding, 5901
1	MS	1003	B. L. Spletzer, 15211
1	MS	1056	Samuel Myers, 1112
1	MS	1071	T. A. Dellin, 1700
1	MS	1109	R. J. Pryor, 9201
1	MS	1109	S. S. Dosanjh, 9221
1	MS	1109	R. Benner Jr., 9224
1	MS	1109	R. J. Pryor, 9212
1	MS	1110	D. E. Womble, 9222
1	MS	1110	N. D. Pundit, 9223
1	MS	1111	G. S. Heffelfinger, 9225
1	MS	1111	H. P. Hjalmarson, 9225
1	MS	1135	D. B. Davis, 9134
1	MS	1137	A. L. Hodges, 6534
1	MS	1152	M. L. Kiefer, 1642
1	MS	1170	L. E. Larsen, 15300
1	MS	1179	J. R. Lee, 15340
1	MS	1179	L. J. Lorence, 15341
1	MS	1188	Craig L. Olson, 1600
1	MS	1202	William Boebert, 5901
1	MS	1203	A. P. Zelicoff, 5327
1	MS	1221	R. D. Skocypec, 15002
1	MS	1349	Jeff Brinker; 1841
1	MS	1349	John Curro, 1834
1	MS	1349	Ron Loehman, 1843
1	MS	1393	J. R. Garcia, 1902
1	MS	1393	F. F. Dean, 2106
1	MS	1413	Peter J. Feibelman, 1114
1	MS	1423	G. C. Osbourn, 1118
1	MS	1425	Robert Hughes, 1744
1	MS	1479	W. J. Tedeschi, 2113
1	MS	9003	K. E. Washington, 8900
1	MS	9003	D. L. Crawford, 9900
1	MS	9003	Jim Costa, 9903
1	MS	9012	P. E. Nielan, 8920
1	MS	9032	Dave Havlik, 9700
1	MS	9042	W. A. Kawahara, 8725
1	MS	9042	C. D. Moen, 8728
1	MS	9051	C. A. Kennedy, 8351
1	MS	9053	David Chandler, 8353
1	MS	9161	E. P. Chen, 8726
1	MS	9202	W. P. Ballard, 8418
1	MS	9202	Rob Rinne, 8100
1	MS	9217	P. T. Boggs, 8950
1	MS	9405	T. M. Dyer, 8700
1	MS	9405	R. A. Regueiro, 8743
1	MS	9409	Dan Tichenor, 8422
1	MS	9951	M. F. Melius, 8130
1	MS	9018	Central Technical Files, 8945-1
2	MS	0899	Technical Library, 9616
1	MS	0612	Review & Approval Desk, 9612, for DOE/OSTI

1 MS 0428 12300 D. D. Carlson