

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.

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

* 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 Lockheed Martin Company, for the United States Department of Energy.

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.

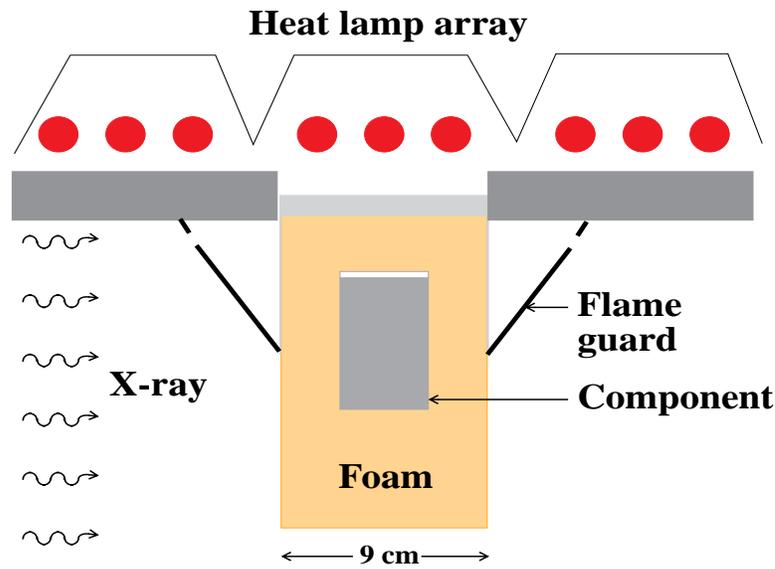


Figure 1. Foam Decomposition Experiments Schematic

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.

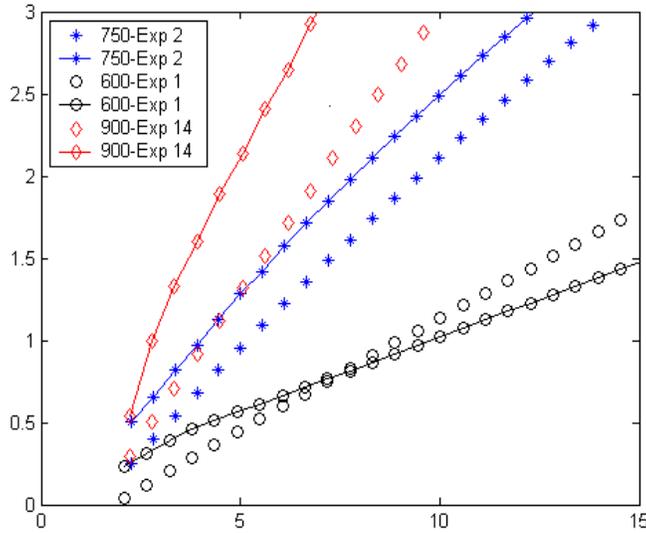


Figure 2. Results From Three Foam Decomposition Experiments: Front Position (cm.) vs. Time (min.). Computational (connected plotting symbols) and Experimental (unconnected symbols) Results.

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:

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

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-validation 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

<i>Exp.</i>	<i>Temp.</i>	<i>Heat Orient.</i>	<i>Int'l. Comp</i>	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 *ln*e'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 *ln*e'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

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

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 \sim 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 $\text{ave} \pm t(.05, f) * \text{stdev}(1 + 1/n)^{.5}$, 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 $-.052 \pm 2.132 * .034 * \sqrt{6/5} = -.052 \pm .08 = (-.13, .03)$. 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 $v^M = .25\text{cm/min.}$, 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 $v^M = .25\text{cm/min.}$, 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.

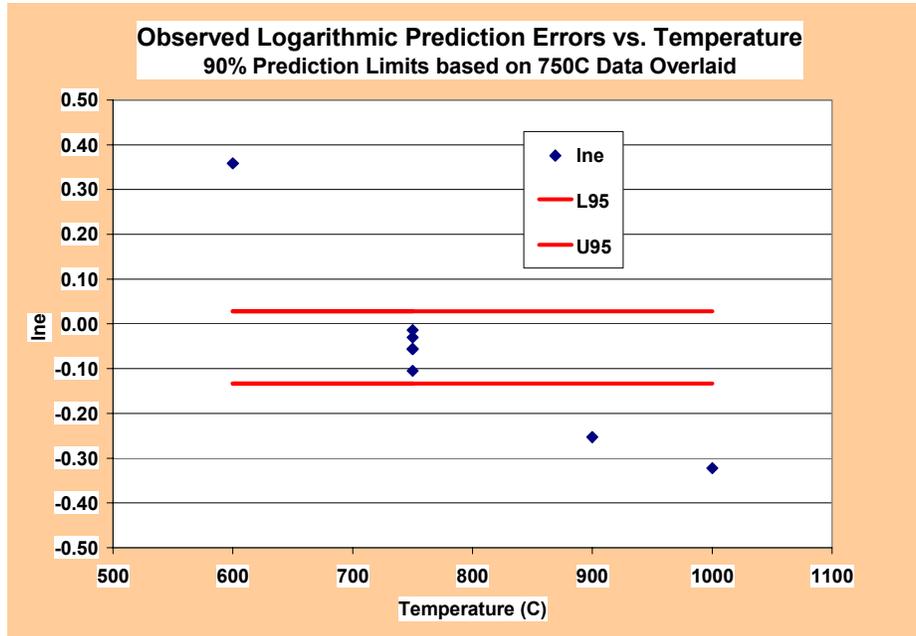


Figure 3. Comparison of Inferred Prediction Error to Observed.
 (The upper and lower ends of the 90% prediction interval are denoted by U95 and L95)

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. $1/(\text{abs. Temp})$ – 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.

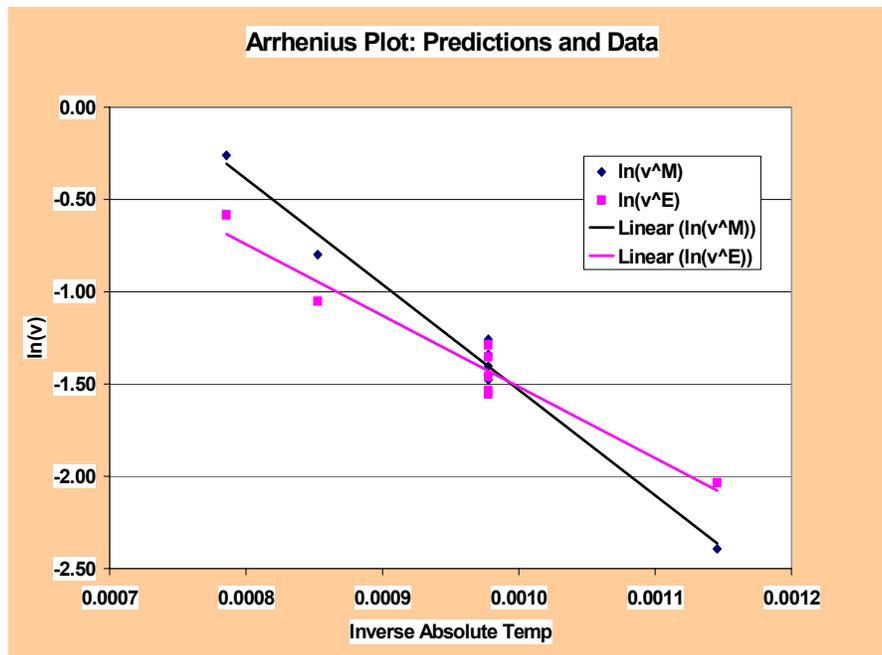


Figure 4. Computational Predictions and Experimental Results Plotted on Arrhenius Coordinates.

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.)

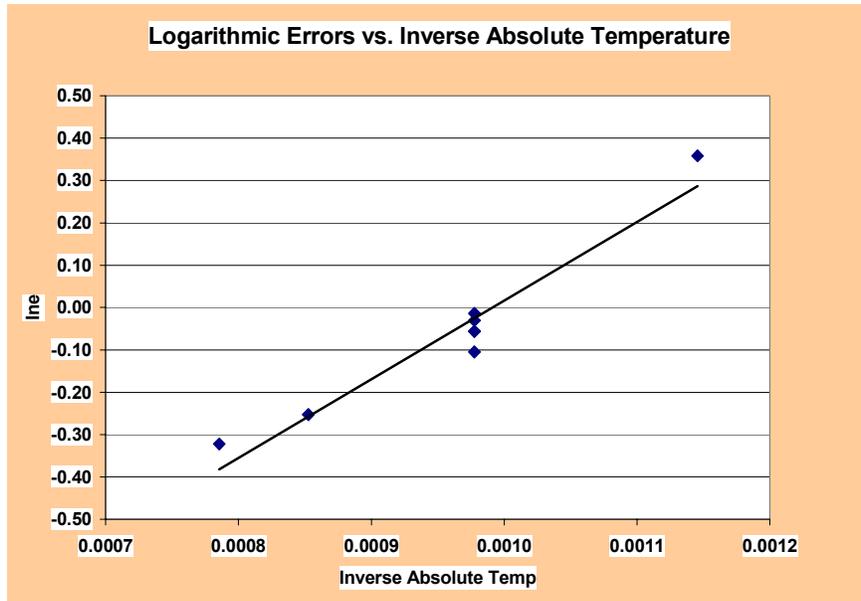


Figure 5. Logarithmic Prediction Errors vs. Inverse Absolute Temperature, with Fitted Straight Line

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

Fitted line:	$\ln\text{-error} = -1.8 + 1858 \cdot (1/\text{Temp(K)})$
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:

$$\begin{aligned} \text{Fitted line:} \quad & \ln(v) = 2.34 - 3858*(1/\text{Temp(K)}) \\ \text{Residual standard deviation:} \quad & s = .113, \text{ on } 6 \text{ df.} \end{aligned}$$

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.

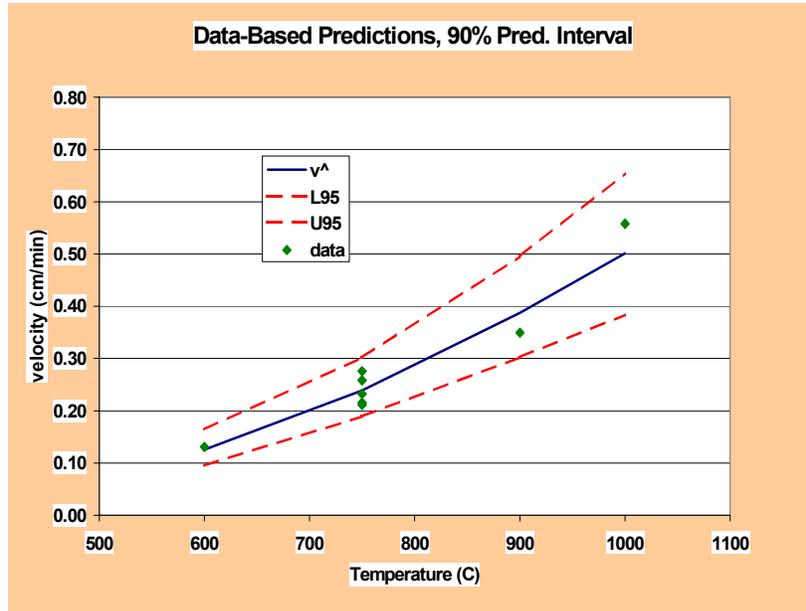


Figure 6. 90% Prediction Intervals Based on Regression Analysis of Experimental Data

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.

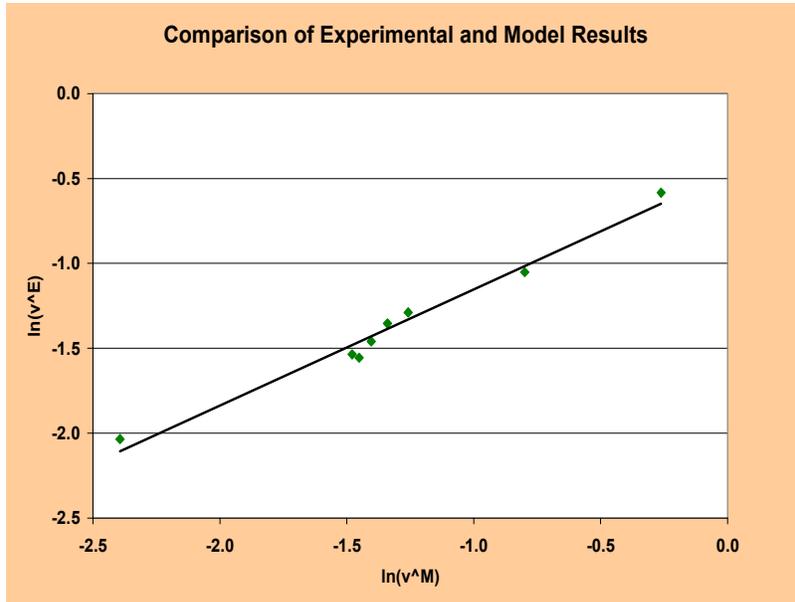


Figure 7. v^E vs. v^M and Fitted Line

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.

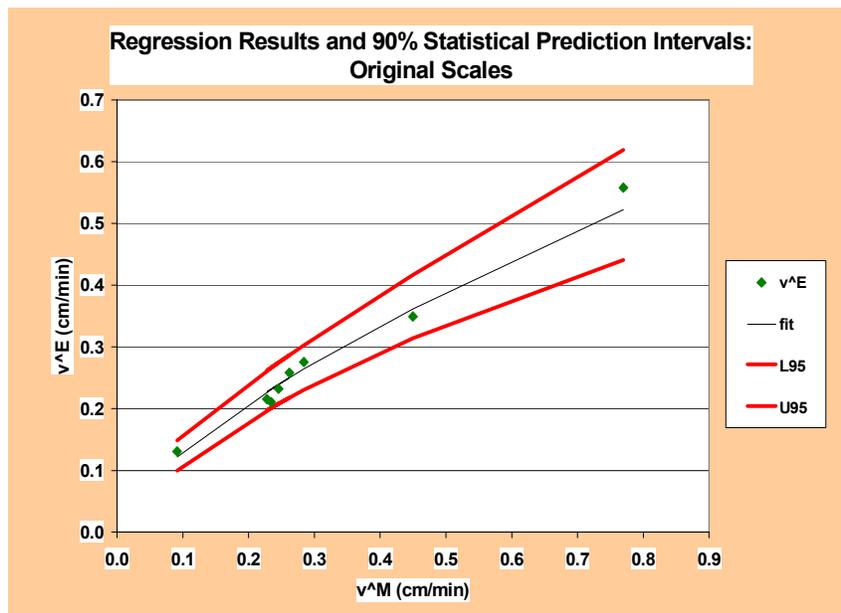


Figure 8. 90% Statistical Prediction Interval for v^M Model

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 $(-.80 + 2.39)/300C = .0053/\text{deg.C}$. 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 $\sigma_T = 2 \text{ deg.C}$. 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

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

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.