SANDIA REPORT

SAND2005-1826
Unlimited Distribution
Printed April 2005

Statistical Validation of Engineering and Scientific Models: Bounds, Calibration, and Extrapolation

Richard G. Hills and Kevin Dowding

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's
National Nuclear Security Administration under Contract DE-AC04-94AL85000

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)576-5782
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

SAND2005-1826
Unlimited Release
Printed April 2005

Statistical Validation of Engineering and Scientific Models: Bounds, Calibration, and Extrapolation

Richard G. Hills
Department of Mechanical Engineering
New Mexico State University
Las Cruces, New Mexico 88003

Kevin Dowding
Optimization and Uncertainty Estimation
Sandia National Laboratories
Albuquerque, New Mexico 87185-0819

Numerical models of complex phenomena often contain approximations due to our inability to fully model the underlying physics, the excessive computational resources required to fully resolve the physics, the need to calibrate constitutive models, or in some cases, our ability to only bound behavior. Here we illustrate the relationship between approximation, calibration, extrapolation, and model validation through a series of examples that use the linear transient convective/dispersion equation to represent the nonlinear behavior of Burgers' equation. While the use of these models represents a simplification relative to the types of systems we normally address in engineering and science, the present examples do support the tutorial nature of this document without obscuring the basic issues presented with unnecessarily complex models.

(Page left blank)

Acknowledgements

This report is an account of contract research (Doc. # AX-0620) performed by the first author in cooperation with the second author. We appreciate the critical reviews of this manuscript by Tim Trucano, Vicente Romero, and Laura Swiler, all of Sandia National Laboratories.

(Page left blank)

Table of Contents

List of Figures

(Page left blank)

List of Tables

(Page left blank)

1.0 Introduction

1.1 Previous Reports

This report is the sixth in a series presenting issues related to model validation methodology. In the first report (Hills and Trucano, 1999), the conceptual ideas behind model validation in the presence of experimental and model parameter uncertainty were presented. We discussed the use of statistical methodology to develop model validation metrics for linear and nonlinear models. Examples were presented showing the application of these metrics to several physical applications.

The second report (Hills and Trucano, 2001) further demonstrated the use of these metrics for one-dimensional shock data. We also introduced the idea of a metric that relates the anticipated target application of a model to the measurements taken from validation experiments. This linkage is important since the validation experiments generally do not exactly represent the target application. Validation experiments are typically more carefully controlled so that the sources of potential differences between observation and prediction can be better resolved. For the same reason, validation experiments are typically designed to test a subset of the physics important to the system. Suites of validation experiments are used to cover the range of physics and the range of anticipated conditions (or parameters) for the target application. Mathematically defining the link between the validation experiments and the target application is important if we wish to provide quantitative evidence as to how well our suite of experiments represent the anticipated application of the model. The application-based metric presented in the second report was designed to weight the experimental data so that they better represent the application. More specifically, data that does not have as direct of an impact on the target application were weighted less. This modification was based on eliminating the linear combinations (or the directions in the validation space – see Hills and Trucano, 2001) that are not important to the application. An example was presented relating a two-dimensional shock application to the one-dimensional shock physics data.

The third report (Hills and Trucano, 2002) focused on the application of the Maximum Likelihood method to the non-application based validation metrics developed in the first two reports. The use of Maximum Likelihood allows highly nonlinear problems with non-normally distributed uncertainties in the measurements and the model parameters to be more easily handled.

The fourth report (Hills and Leslie, 2003) further developed the relationship between the component or unit level validation experiments and the system level target application. Specifically, the relationship between the decision variables that are important to the target application and the measurements obtained from the suite of supporting validation experiments was investigated. In this context, we consider a decision variable to be a predicted variable that is important to the application. It is that quantity that defines

whether a design is successful or not. A decision variable may be the temperature in a component, the probability that a component will detonate, or the stress at a critical location. It is not unusual for the decision variable to be different than the quantities measured in the validation experiments. For example, one may not be able to directly measure maximum stress in a component because the location of maximum stress is not accessible. The methodology presented in the fourth report uses first order sensitivity analysis 1) to assess whether the suite of validation experiments adequately represents the physics of an anticipated target application; 2) to evaluate how to weight the measurement data to best represent the sensitivities of the target application to the physics tested by these experiments; and 3) to evaluate the sensitivity of the reconstructed decision variables to uncertainties in the experimental measurements, and to uncertainties in the model predictions due to uncertainties in the model parameters for the validation experiments and the target application.

The fifth report (Hills, et. al., 2004) focused on the application of model validation metrics developed by Hills and Trucano (1999, 2001) to a particular application, the thermal decomposition of foam when heated from one side. Rigid polyurethane foams are used in weapons systems to isolate and support sensitive components. Abnormal thermal environments, such as fire, can cause foam decomposition and subsequent exposure of the components to undesirable thermal input. The ability to model and predict this decomposition is a critical aspect of the behavior of engineered components in abnormal environments. To this end, we investigated the uncertainties in the model predictions due to uncertainties in the model parameters, and used the metrics developed by Hills and Trucano to evaluate consistency between the model predictions and the experimental observations. The model predictions of foam-decomposition front location as a function of time were obtained using CPUF/COYOTE (Hobbs et al., 2003, Gartling et al., 1994) . We used a first order sensitivity analysis combined with a Monte Carlo approach to estimate uncertainty in these model predictions.

The present report focuses on issues related to model validity when the model is to be extrapolated from the conditions under which the model was tested. Unlike previous work by the present authors, this work addresses the use of approximate models in some detail; including issues associated with calibration, and with bounds on predicted behavior. We begin with a conceptual discussion of models and their use in extrapolation and interpolation.

1.2 Models

It is characteristic of humans to model the environment around them. Models can range from purely subjective (i.e., intuitive models for human behavior) to models based on physical and mathematical principles. Many of these models are used to provide insight as to the anticipated or possible behavior of a system under various external influences, and do not claim to accurately predict behavior. Other models, such as engineering models, are based on well established principles (theoretical basis with considerable supporting experimental observation). Predictions from such models can be used for

engineering design, when safety factors are used to compensate for possible model errors. In other cases, the ability to model the physical phenomena (response of an elastic solid, thermal heat conduction in a simple solid) exceeds our ability to accurately characterize the external influences on the parameters that appear in the model (i.e., boundary conditions, initial conditions).

Many models used for engineering design possess some type of approximation. As the complexity of the system increases, the uncertainty in the predictions generally increases. The basic theory (mathematical equations) modeling the phenomena is often approximate. The numerical schemes used to solve the mathematical equations contain approximations and add uncertainty to the model predictions. The appropriate values for the parameters defining the constitutive equations, and the boundary and initial conditions, are often uncertain due to the lack of knowledge (i.e., not sufficiently measured or controlled), or due to natural variability in the system. Many models represent approximate extensions of well-established principles outside the range of the application originally anticipated during the development of the principle. For example, Darcy's law for single phase flow through a porous media is well established. When multiple phases are present and one of the phases is a wetting phase, capillary effects can become important. An example is the flow of water through unsaturated soils. Darcy's equation is often applied to such cases, but only after the proportionality constant between pressure gradient and flow is empirically modified to account for capillary tension.

For the case of well established system level models, such as those based on conservation principles, the uncertainties in the application of such models to different scenarios is not due to the uncertainty in the system level model, but due to the uncertainties in the calibration of the constitutive models, the boundary and initial conditions, and in the numerical approximations. If we can characterize the uncertainty in the constitutive models (e.g., the spatial variability in geological media), the boundary and initial conditions, and bound the uncertainty due to the numerical approximations, then we can propagate these uncertainties through system level models to predict the uncertainty in the system due to these effects.

In the present work, we present a series of modeling scenarios and discuss the extrapolation/validation issues associated with these models. We look at the impact of approximate bounding models, calibration, and other practical issues associated with complex engineering systems. The intent of this document is to provide a series of simple, but instructive examples.

(Page left blank)

2.0 Background

2.1 Introduction

We begin with a brief discussion of terminology, followed with a literature review of model extrapolation and validation.

2.1.1 Data Based Model

For the purpose of discussion, we define a data-based model as one which represents the behavior of a system, based strictly on fitting non-physically based functions to observed data. Regression models (linear and polynomial) and many calibration models are examples of such models. Many constitutive models are also based on simple linear regression (such as thermal conductivity as a function of temperature). These models make no assumption as to the physics of the phenomena being modeled and simply attempt to represent the behavior of the system over the restricted range of data used to calibrate the model. These models may be in the form of look-up-tables based on direct experimental observation, direct measurements of scalar quantities, or regression. These data-based models may be deterministic in the sense that single values are returned, or they may include the effect of uncertainty in their representation of the behavior through error bounds or other representations of uncertainty.

Extrapolation of these models to domains different than those used to calibrate the model should be a cautious endeavor unless one has evidence that the assumed behavior (linear for example) is appropriate outside the data support range. Interpolation can also be fraught with danger if the calibration data is sparse. For example, fitting a high order polynomial through a limited number of data points can lead to oscillatory behavior between the data points and serve as a poor interpolation scheme.

2.1.2 Physical Based Models

Here we consider physically based models to be models that are based on fundamental principles. These models can be based on physical laws, such as the conservation of mass, momentum, and energy. These models often contain constitutive parameters to represent some sort of integrated behavior of the material, such as conductivity in Fourier's law. These models can also be well established approximations, such as the k-epsilon model in turbulence, which have physical justification for its form.

The advantage of physically based models relative to data-based models is we generally have more confidence in the extrapolation of these models outside the range of conditions tested. This is based on expert subject matter knowledge. For example, we fully expect the thermodynamic laws to extrapolate to an application.

We also expect physically based models, for which there are calibration parameters to account for material properties, to apply to various applications and geometries, as long as we use the correct calibration parameters for that material. Our mathematical expression for Fourier's law - heat flux is equal to the negative of the product of the thermal conductivity and the temperature gradient - is generally considered invariant for heat conduction in many engineering applications.

In contrast, fitting a mathematical surface or curve to the measurements of heat flux over a range of conditions, can represent heat flux only over the range of conditions defined by the geometry, initial conditions, and boundary conditions for the generating experiments. A model for heat flux based on Fourier's law is much more useful as we expect to be able to apply this law to heat conduction under conditions other than those tested, as long as we properly account for the thermal conductivity of the materials.

Caution must be exercised in some cases when making the distinction between data-based and physical based models. Many constitutive models may be more data-based than they at first appear. As mentioned earlier, Darcy's law relates the flux of a single phase fluid through a porous media to the pressure gradient, just as Fourier's law relates the flux of thermal energy to a temperature gradient. In fact, the proportionality constant for Darcy's law is called the hydraulic conductivity. In some applications in hydrology, Darcy's law is extended to the transport of a two component system (water and air) through porous media. For unsaturated media, the primary driving potential is often capillary tension rather than pressure. Darcy's law is extended to such cases by using multi-parameter models for hydraulic conductivity developed though experimental observation and parameter estimation. Because of the nature of capillary tension (very high tensions when the soil is dry, low tensions when the soil is near saturated), the representation of hydraulic conductivity is often a highly non-linear function of saturation and very dependent on the soil, due to the complexity of pore geometry. While there sometimes is a fundamental phenomenological principle for the choosing a function to relate hydraulic conductivity to water saturation, these functions are often chosen simply because they simulate the empirical data well. Even though the extension of Darcy's law to unsaturated flow is an approximation, we still expect mass to be conserved at the system level.

2.1.3 Model Interpolation

Most researchers have the intuitive sense that model interpolation means application of a model to conditions bounded by the calibration and validation experiments. At first glance, this description seems rather straight forward. However, the application of this idea to complex models quickly reveals that this definition is somewhat nebulous.

While calibration of the constitutive models used in the system model can often be performed over a range of conditions associated with the application of a model, the validation of a system level model against system level measurements may not be as straight forward. This is especially true for multidimensional models with a high number

of constitutive model parameters. For example, consider a model with 25 parameters. If we were to test this model at three values for each parameter (the min, the max, and the median), we would require $25^3 = 15{,}635$ experiments. Clearly, this is not practical. Another approach could be to test the model at the median values of the parameters (or the anticipated values), and then perturb one parameter at a time and retest. This would result in 26 experiments which also may not be practical. Even if this approach were practical, this experimental scheme ignores potential correlation that can occur in the 25 parameter model, and cannot adequately cover the parameter space. Another procedure is to use some form of experimental design to randomly pick the values of the model parameters to test at according to some sampling scheme. This approach does allow the observation of some correlation structure in the experiment, and provides some representation over the space of the samples. However, this approach still cannot fully cover the space of the parameters for the 25 parameter example discussed above. There will be some regions in the parameter space (often near the boundaries) that are not well represented by the series of experiments. Thus the use of the model near the boundary will represent an extrapolation simply because the experiments did not span this region. So what appears at first glance to be model interpolation may, in-fact, be model extrapolation due to the high dimension of the model parameter space.

2.1.4 Model Extrapolation

Model extrapolation has different meanings to different investigators. It can mean the extrapolation of the model outside the range of model parameters tested, or the extrapolation of the model to conditions not tested (i.e., different geometries or boundary conditions), or in the most extreme case, the extrapolation of the model to different physical phenomena for which the model acts as a surrogate.

In this present work, we consider extrapolation to include the following:

1. Prediction of a quantity by a model that was not measured directly during the experiment even for similar test conditions: For example, we may be interested in the stresses in a bolt, but can only measure strain in the components surrounding the bolts. In this case, a model will be used to relate the strain measured at one location to infer stress at another. Because we are using a model to relate the response variable tested during the experiment to one that was not tested, we consider this to be extrapolation.

2. Prediction of a quantity outside the range of model parameters or conditions tested: For example, we may perform validation experiments on a response variable over some temperature range, but then use the model to predict behavior at higher temperatures.

3. Use of the model as a surrogate for another type of system: This represents an extreme form of the case discussed in item 2.

Our primary focus in this work is on extrapolation of the type described in item 2. Extrapolation of the type described in item 3 is not normally accepted in engineering, unless we have collaborating experimental evidence that the system we are extrapolating to is well represented by the surrogate model.

2.1.5 Approximate Physics-Based Models and Impact on Model Validation

In the previous work by this author (Hills and Trucano, 1999, 2001, 2002; Hills and Leslie, 2003; Hills et al., 2003), we focused on the development of model validation methodology which finds a model valid if the errors introduced by non-modeled or approximately modeled physics, are less significant than those introduced by the uncertainties in the validation experiments. This approach does provide useful insight as to the methodology required to account for the effects of uncertainty in the model parameters and the experimental measurements on model validation. However, engineering models that approximate or bound behavior are often more practical to develop and implement. For some complex systems, approximate or bounding models may be the only feasible approach. For example, while we may not be able to accurately estimate the amount of diffusion that takes place in a system, we may be able to estimate a lower bound or a range for diffusion. This may allow us to use these bounds to bound predicted behavior of the system. Another example is the use of approximate models to represent sub-grid behavior. If there is a consensus in the scientific community that such a model consistently provides an estimate of the behavior that is conservative in some sense (however, this may be defined), then we may be able to use this model to bound the behavior of the system.

The concept of a model validation for bounding models is different than that for models that we claim accurately represent the physics within validation experiment uncertainty. Rather than answering the question – are model predictions consistent with experimental observations, given the uncertainty in the validation exercise; here we ask – do the model predictions bound (from above or below) the experimental observations, given the uncertainty in the validation exercise? In the case of bounding approximate models, our validation tests will often take on a one-sided nature, which complicates the corresponding statistical inference for multivariate data. This feature of bounding models will be addressed in a later chapter.

2.2 Literature on Model Extrapolation/Validation

A search of the SciSearch Plus scientific article database (ISI, 2003) for the phrase “model validation”, yields a significant number of hits. Investigation of these hits suggests that to most authors, model validation is the comparison of model predictions to experimental observations through graphical means. The impact of experimental and/or model uncertainty on these comparisons is generally ignored. A general literature review on model validation is provided by Oberkampf and Trucano (2000) and Oberkampf et al. (2003). Additional literature on statistical methods in model validation is discussed by Hills and Trucano (1999). In recent years, there has been an increased emphasis on using

statistical methodology to characterize experimental and/or parameter uncertainty in the comparisons between experimental data and physics-based model predictions. Examples of such methodology include the work by Dowding et al. (2004), Easterling (2003), Rutherford and Dowding (2003), and the series of reports by Hills and Trucano (1999, 2001, and 2002), Hills and Leslie (2003), and Hills et al., (2003).

A search of the scientific article databases for the term model extrapolation yields significantly fewer hits. Since engineering/scientific models are generally developed for the purpose of providing insight for conditions other than those tested (i.e., model extrapolation), this observation is not surprising. The technical areas for which we found the largest number of hits for the phrase "model extrapolation" were in the environmental and medical fields. For example, extrapolation is an important aspect in environmental risk assessment and health fields. This includes the extrapolation over different physical scales (from patch scale to landscape scale; Landis, 2002, Munns, 2002), across different temporal scales (Munns, 2002, Kalberlah et al., 2002), and biological extrapolation (across levels of biological organization - Munns, 2002; between animals and human – Bernillon and Bois, 2000, Kalberlah et al., 2002, Vermeire et al., 2001).

More specifically, Bernillon and Bois (2000) present statistical issues associated with toxicokinetic modeling. As discussed in their paper, toxicokinetic models can be classified into two broad categories, classical compartmental toxicokinetic (TK) models and physiologically based toxicokinetic (PBTK) models.

The classical models represent the body by several compartments which may or may not represent the anatomy of the species. Chemical transport between the compartments, each of which is assumed to contain a uniform distribution of chemicals, is modeled through differential equations. The parameters appearing in the differential equations are estimated through parameter estimation based on empirical data. Because of this, TK models are often referred to as data-based or empirical models. These models are found to be reliable when used to interpolate from the calibration data. Because they are not physiologically based, they are not appropriate for the extrapolation to other species, exposure conditions, and routes.

The physiologically based toxicokinetic models are based on compartments corresponding to specific organs or lumped tissue and organ groups. The transport between compartments is based on blood flow, lymphatic circulation and chemical transfers between these compartments; and is represented by differential equations for the corresponding mass balances. The model parameters have physical significance (solubility, tissue volumes, diffusion, etc.). Because these models are more closely representative of actual physiological systems, they are felt to better model the linear and nonlinear dynamics associated with the metabolism of the toxic compound under study. As a result, the extrapolation to species with the same structures (between mammals for example) is appropriate with the proper choice of parameters for that species. In addition, extrapolation to different forms of exposure (between air, ingestion, and intravenous injection) can be made with the appropriate application of source terms. Because these

models can require a substantial number of parameters (20 or more) to characterize the underlying phenomena, and because the values of these parameters have uncertainty due to both estimation and natural variability within and between species, the predictions of PBTK models will contain uncertainty. Bernillon and Bios (2000) discuss the uncertainties associated with both TK and PBTK models and present a Bayesian based approach to the calibration of these models. They provide an example of the calibration of TK models and illustrate the pitfalls associated with ignoring the probabilistic structure, such as correlation of the parameters. They also discuss the ability of the Bayesian approach to use prior knowledge to help constrain parameter estimates when sufficient calibration data is not available.

Even though the literature previously cited acknowledges uncertainty and provides methodology for the assessment of uncertainty due to parameter uncertainty, the concept of model validation and model validation methodology is not addressed with the rigor desired here (see Trucano et al., 2001, 2002). Model validation has different meanings to different disciplines. For example, the statistical concept of cross-validation can be used to discriminate between competing models. Lockwood, et al. (2001) looks at competing models for arsenic occurrence in source waters across the United States community water systems. Because the corresponding observation data is sparse and uncertain, Bayesian approaches to model calibration are used. Lockwood, et al., consider models of increasing complexity, calibrate the various models to subsets of data, and rank the predictive ability of the models using the remaining data through the use of predictive density. By repeating selecting subsets of the data to calibrate and to validate against, one can select the model that has high predictive ability over a large percentage of the cross-validation samples. Note that this approach uses a relative metric to choose the best model, but does not necessarily set a criteria for predictive density for which a model is declared invalid (or valid).

Warren-Hicks, et al., (2002) explicitly discuss the role of uncertainty in model validation. They argue that prediction uncertainty must be used in the comparison of prediction to observation. Otherwise, complex models can easily fail validation tests simply due to the uncertainty in the model's parameters rather than due to some structural defect in the model. Warren-Hicks, et al., point out that typical estimators of model accuracy are mean squared error, paired *t*-statistics, and correlation statistic. They state that

> *"While these statistics may or may not be valid indicators of statistical accuracy, a larger issue arises in that these statistics do not reflect the uncertainty in model use, such as the decisions made in model calibration, model structure, or choice of time step."*

They argue that

> *"… a simple comparison of observations and predictions is a naïve approximation of the usefulness of the model or the expected inferences that can be drawn from the model output."*

As Warren-Hicks et al. discuss, the use of classical statistical indicators, such as the paired *t* test, use the variance in the differences between measurements and predictions as the test statistic. This estimate of uncertainty is appropriate if the experimental data truly represents all sources of uncertainty, such as multiple builds of the experimental apparatus, manufacturing lot-to-lot differences of the materials, and uncertainties in boundary conditions over repeated independent experiments. They argue that an appropriate method to account for these effects, if not reflected in the data, is to use the Monte Carlo method of propagating these uncertainties through a model. This allows one to model important sources of uncertainty that may not be reflected in the validation experiments. Of course, one must have sufficient knowledge to adequately characterize these sources for use in a Monte Carlo analysis.

Warren-Hicks et al. (2002) quantify model validity by measuring the percentage of the probability density function for the predicted measurement that lies below and above the experimental observations. When the measured value is "near the center" of the predicted distribution, the model is considered to be accurate. When the measured value is in the lower or upper portions of the predicted distribution, the model is considered to be less accurate. If the entire distribution is below or above the measured value, the model is considered to be inaccurate. No guidance as to what constitutes "near the center" of the distribution is given.

A thoughtful tutorial of model validation is provided by Robinson (1999). He breaks model validation into several components as follows: Conceptual Model Validation is the determination "*that the scope and level of detail of the proposed model is sufficient for the purpose at hand, and that any assumptions are correct*." Data Validation is evaluating "*that the data required for model building, validation and experimentation are sufficiently accurate*." White-Box Validation is evaluating whether "*the constitutive parts of the computer model represents the corresponding real world elements with sufficient accuracy*." Black-Box Validation is "*determining that the overall model represents the real world with sufficient accuracy*." Experimental Validation is "*determining that the experimental procedures adopted are providing results that are sufficiently accurate*." Solution Validation is "*determining that the results obtained from the model of the proposed solution are sufficiently accurate*." Solution Validation takes place after the designed product is complete. It is a comparison of the final performance of the product to the predicted performance. Guidance of what constitutes or how to measure "sufficiently accurate" is not addressed.

While the concept of model extrapolation and model validation is discussed in the literature, we are not aware of work, other than that by Hills and Trucano (2001) and Hills and Leslie (2003), which directly tie model validation metrics to extrapolation. The work by Hills et al. focuses on the modification of validation metrics at the unit or subsystem level to reflect the target application. In this work, the relationship (extrapolation) between the unit or subsystem level to the system level and the effect of uncertainty is investigated. This work assumes that the underlying models, if proven

valid for the intended application, represent the physics with an error that is within the effect of measurement and model parameter uncertainty.

In the present work, we focus on a different aspect of model extrapolation, that of temporal extrapolation. We also relax the emphasis that a valid model be consistent with the experimental data, within the uncertainty induced by the model parameters and the experimental data. Here we explicitly address the use of approximate and approximate bounding models that need not be statistically consistent with the data - those which do not fully represent the physics, but approximate or bound the physical behavior.

3.0 Simulated Data and Approximate Model

Two models will be used in this work. The "true" physics will be defined by a two parameter form of Burgers' equation. This equation is non-linear, transient, and can produce fully developed moving fronts. This equation will be used to generate simulated measurements of front velocity by taking random realizations of the parameters, approximating the resulting solution to Burgers' equation and the associated front movement, and adding random noise to the predicted measurements to represent experimental error.

To illustrate the effect of non-modeled physics, we will use the convective-dispersive equation as the approximate predictive model. This equation does not possess the non-linearity in the convective term that Burgers' equation does, and cannot accurately represent the physics for both short and long times. However, this equation can be used to approximate the nonlinear behavior of Burgers' equation over a limited time period, and can be effectively used to bound the behavior (predict upper and lower bounds) if properly conditioned.

We begin with Burgers' equation.

3.1 Simulated Measurements (Burgers' Equation)

We use Burgers' equation to represent the correct model of the true physics of the validation experiment. A two parameter form of Burgers' equation can be written as

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - U\frac{\partial c^2}{\partial x}, \quad 0 \le x \le 10, \quad 0 \le t \le 2.5 \tag{3.1}$$

We define the initial condition to be a pulse defined by

$$c(x,0) = \begin{cases} 1, & 3 \le x \le 4 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

c is the dependent variable, D the dispersion coefficient, and U is a parameter that affects front speed. We take the range of x large enough (i.e., $0 \le x \le 10$) so that the leading edge of the front does not reach a boundary during the first 2.5 times units, given the values of our parameters and the form of the initial condition. We take our validation variable to be the front location as a function of time. Here we define the front location X as that location for which $c = 0.25$ on the leading edge of the $+x$ moving front.

We consider two forms of uncertainty in the simulated data. First, we consider the uncertainty associated with the true model parameters for the validation experiment. For illustrative purposes, we take the mean and standard deviation for these parameters to be

$$D_{\text{mean}} = 0.08, \quad U_{\text{mean}} = 1.0 \tag{3.3a, b}$$

$$\sigma_D = 0.008, \quad \sigma_U = 0.1 \tag{3.4a, b}$$

We assume that the uncertainty between these two parameters is not correlated and the uncertainty in each parameter is normally distributed. Note that the standard deviations are 10% of the mean values. We also consider measurement noise, which we take to be normally distributed, uncorrelated, with zero mean and a standard deviation of 0.05.

$$\sigma_{\text{meas}} = 0.05 \tag{3.5}$$

To approximate the uncertainty in the model parameters due to the uncertainty in the model predictions, we use a multivariate linear sensitivity analysis presented by Hills and Trucano (2001). We begin by relating changes in predicted front location $X(t)$ at the discrete times $t = t_1, t_2, \ldots, t_n$, to perturbations in the vector $\boldsymbol{\alpha}$ of model parameters from its mean.

$$\mathbf{X} = \mathbf{X_0} + \nabla_{\boldsymbol{\alpha}}\mathbf{X} \ \Delta\boldsymbol{\alpha} \tag{3.6}$$

where

$$\mathbf{X} = \begin{bmatrix} X(t_1) \\ X(t_2) \\ \vdots \\ X(t_n) \end{bmatrix}, \quad \mathbf{X_0} = \begin{bmatrix} X_{\text{mean}}(t_1) \\ X_{\text{mean}}(t_2) \\ \vdots \\ X_{\text{mean}}(t_n) \end{bmatrix}, \quad \Delta\boldsymbol{\alpha} = \begin{bmatrix} D - D_{\text{mean}} \\ U - U_{\text{mean}} \end{bmatrix} \tag{3.7}$$

The sensitivity matrix (composed of the sensitivity coefficients) is given by

$$\nabla_{\boldsymbol{\alpha}}\mathbf{X} = \begin{bmatrix} \dfrac{\partial X(t_1)}{\partial D} & \dfrac{\partial X(t_1)}{\partial U} \\ \dfrac{\partial X(t_2)}{\partial D} & \dfrac{\partial X(t_2)}{\partial U} \\ \vdots & \vdots \\ \dfrac{\partial X(t_n)}{\partial U} & \dfrac{\partial X(t_n)}{\partial U} \end{bmatrix} \tag{3.8}$$

We can now estimate the covariance matrix for the model predictions using first order sensitivity analysis (see Hills and Trucano, 2001):

$$\text{cov}(\mathbf{X}_{\text{model}}) = \nabla_{\boldsymbol{\alpha}}\mathbf{X} \ \text{cov}(\boldsymbol{\alpha})\left[\nabla_{\boldsymbol{\alpha}}\mathbf{X}\right]^{\text{T}} = \nabla_{\boldsymbol{\alpha}}\mathbf{X}\begin{bmatrix} \sigma_D^{\,2} & 0 \\ 0 & \sigma_U^{\,2} \end{bmatrix}\left[\nabla_{\boldsymbol{\alpha}}\mathbf{X}\right]^{\text{T}} \tag{3.9}$$

The covariance matrix of the differences between the simulated measurements and the experimental data, including the effect of measurement uncertainty, is given by

$$\mathrm{cov}(\mathbf{X}_{\mathrm{total}}) = \mathrm{cov}(\mathbf{X}_{\mathrm{exp}}) + \mathrm{cov}(\mathbf{X}_{\mathrm{model}}) \tag{3.10}$$

where

$$\mathrm{cov}(\mathbf{X}_{\mathrm{exp}}) = \sigma_{\mathrm{meas}}{}^{2}\ \mathbf{I} \tag{3.11}$$

and **I** is an *n*x*n* dimensional identity matrix corresponding to the *n* discrete times of the measurements.

To minimize the effect of numerical diffusion, Eq. (3.1) is solved numerically using an operator splitting technique. This equation was split into a strictly advective equation and a dispersive equation for each time step (Hills et al., 1994). The resulting advective equation is solved using the second order TVD scheme of Roe and Sweby combined with a Superbee limiter (Roe, 1985, 1986 and Sweby, 1984). Comparison of this method with other shock-capturing methods, as applied to Burgers' inviscid equation, is presented by Yang and Przekwas (1992). In solving Eq. (3.1), we used $\Delta x = 0.1$, $\Delta t = 0.02$. Reducing the time step by a factor of 50 and the spatial step by a factor of 2 resulted in a shift in the predictions that were within 0.2% at t =0.1 and 0.04% at t=1.0. We did not perform a more rigorous convergence analysis on the numerical algorithm since the purpose of this exercise was to generate simulated measurements that contained uncertainty. Any bias that may result from the lack of convergence simply adds non-modeled uncertainty to our examples, increasing the possibility that a model will be rejected as valid.

We use a multinormal random number generator with a mean of zero and the covariance defined by Eq. (3.10) to generate simulated realizations of the data. These realizations thus incorporate uncertainty in both the model parameters and the simulated measurements. The uncertainty in the model parameters is important because it reflects the uncertainty in the execution of the validation experiment (i.e., uncertainty in the true parameters for our particular realization of the validation experiment). Because we included parameter uncertainty in our realization, our simulated measurements will contain bias from the predictions of Burgers' equation using mean parameters

Simulated data were generated on 0.05 time increments out to t = 2.5. This gives 51 measurements. The random realization of these simulated measurements used here is tabulated in Table 3.1. The second random realization is tabulated in Table 3.1 for later use as "independent" validation data. We use finite differences to estimate the components of the sensitivity matrix. Specifically, we use forward differences with an increment that is 5% of the mean value of the model parameter of interest for each term in Eq. (3.8).

Table 3.1 Simulated Front Locations

Time	X_mean	X_exp_1	X_exp_2	Time	X_mean	X_exp_1	X_exp_2
0.00	4.000	3.945	3.985	1.30	5.268	5.456	5.233
0.05	4.113	4.136	4.148	1.35	5.304	5.423	5.201
0.10	4.168	4.237	4.197	1.40	5.337	5.531	5.267
0.15	4.224	4.242	4.207	1.45	5.375	5.578	5.315
0.20	4.281	4.309	4.256	1.50	5.408	5.547	5.314
0.25	4.329	4.426	4.315	1.55	5.440	5.543	5.290
0.30	4.386	4.406	4.325	1.60	5.476	5.615	5.405
0.35	4.432	4.429	4.402	1.65	5.507	5.653	5.390
0.40	4.489	4.647	4.434	1.70	5.538	5.712	5.512
0.45	4.533	4.621	4.540	1.75	5.572	5.783	5.485
0.50	4.589	4.666	4.530	1.80	5.602	5.811	5.557
0.55	4.632	4.748	4.582	1.85	5.632	5.726	5.542
0.60	4.686	4.735	4.716	1.90	5.664	5.813	5.608
0.65	4.729	4.757	4.616	1.95	5.693	5.831	5.612
0.70	4.779	4.803	4.854	2.00	5.722	5.828	5.574
0.75	4.822	4.828	4.810	2.05	5.751	5.894	5.640
0.80	4.867	4.953	4.893	2.10	5.781	5.964	5.669
0.85	4.913	5.127	4.819	2.15	5.808	6.043	5.855
0.90	4.949	5.125	4.902	2.20	5.836	6.066	5.782
0.95	4.998	5.195	4.916	2.25	5.864	6.027	5.780
1.00	5.034	5.131	4.924	2.30	5.891	6.089	5.751
1.05	5.079	5.213	5.078	2.35	5.917	6.183	5.837
1.10	5.117	5.184	5.223	2.40	5.944	6.094	5.969
1.15	5.153	5.295	5.157	2.45	5.971	6.272	5.985
1.20	5.195	5.348	5.053	2.50	5.996	6.153	5.942
1.25	5.229	5.300	5.280				

Figure 3.1 illustrates the front location as a function of time, using the mean values of the parameters, Eq. (3.3), as well as, two sets of simulated measurements containing the effect of uncertainty in both the parameters and the measurements. As discussed above, the two sets of simulated measurements represent the effect of two different realizations of *D* and *U*. Note that the inclusion of both measurement noise and parameter uncertainty resulted in significant bias of the simulated measurements from the predictions of Burgers' equation using the mean values of the parameters. Note also that the second realization of the data shows a slower front motion than that predicted using the mean parameter values. In contrast, the first realization of the data shows a faster front motion. These two realizations were intentionally chosen to illustrate issues related to calibration later in the report. The 95% measurement uncertainty intervals, based the measurement noise defined by Eq. (3.5), are also shown for both sets of simulated measurements.

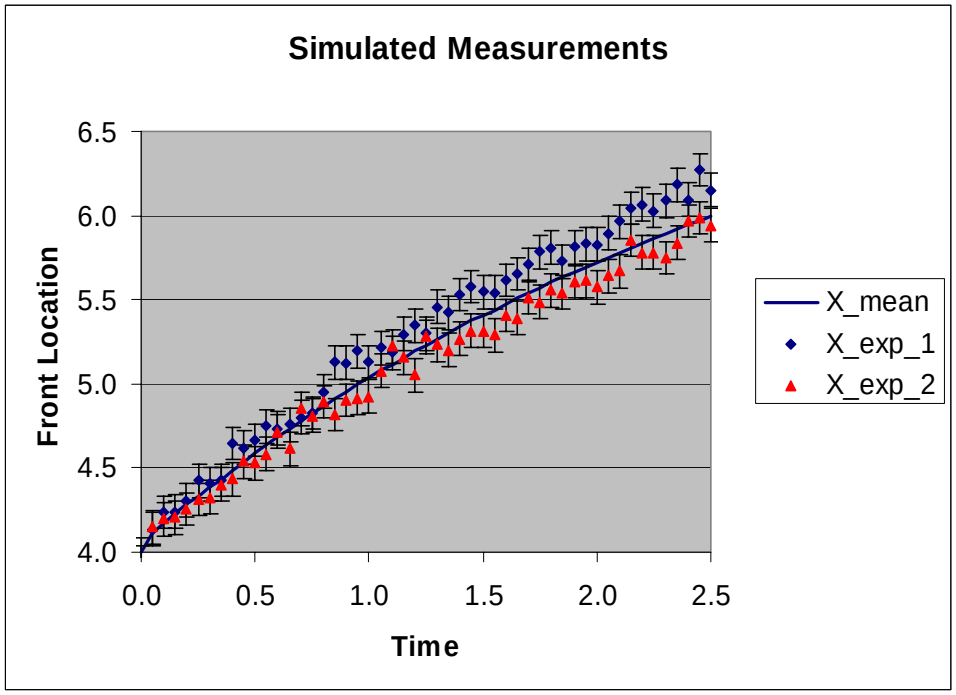

Figure 3.1 Simulated Measurements of Front Location

3.2 Approximate Model (Convective-Dispersion Equation)

To illustrate the effect of non-modeled physics, we use predictions from the following convective-dispersive (C-D) equation to approximate the behavior of Burgers' equation.

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - V\frac{\partial c}{\partial x}, \quad 0 \le x \le 10, \quad 0 \le t \le 2.5 \tag{3.12}$$

with

$$c(x,0) = \begin{cases} 1, & 3 \le x \le 4 \\ 0, & \text{otherwise} \end{cases} \tag{3.13}$$

where c is the dependent variable, D the dispersion coefficient, and V is convective velocity. Note that we used the same initial pulse condition as was used for the simulated measurements, and thus assume that we have good knowledge of these conditions. As before, the time period of the simulation was chosen such that the front does not reach either boundary. We also measure front location as that location for which c = 0.25 at the leading edge of the front. We use a simple explicit finite difference algorithm to model

Eq. (3.12), with the convective term approximated using an upwinded difference. The use of this algorithm results in the presence of some numerical diffusion. As in the case of the approximation for Burgers equation, we did not perform a detailed convergence study. As a result, our model and algorithm contain both non-modeled physics and numerical approximation errors (uncertainties). Since the purpose of the present work is to develop validation metrics that apply to engineering models that invariably contain both numerical approximation and non-modeled physics, we suggest that the inclusion of numerical diffusion in our approximate model is not unrealistic. Here we used $\Delta x = 0.1$ and $\Delta t = 0.02$. We found that the results were within a line width of the results shown in Figure 3.2 (for X_mean) when we reduced the spatial and time step by a factor of two.

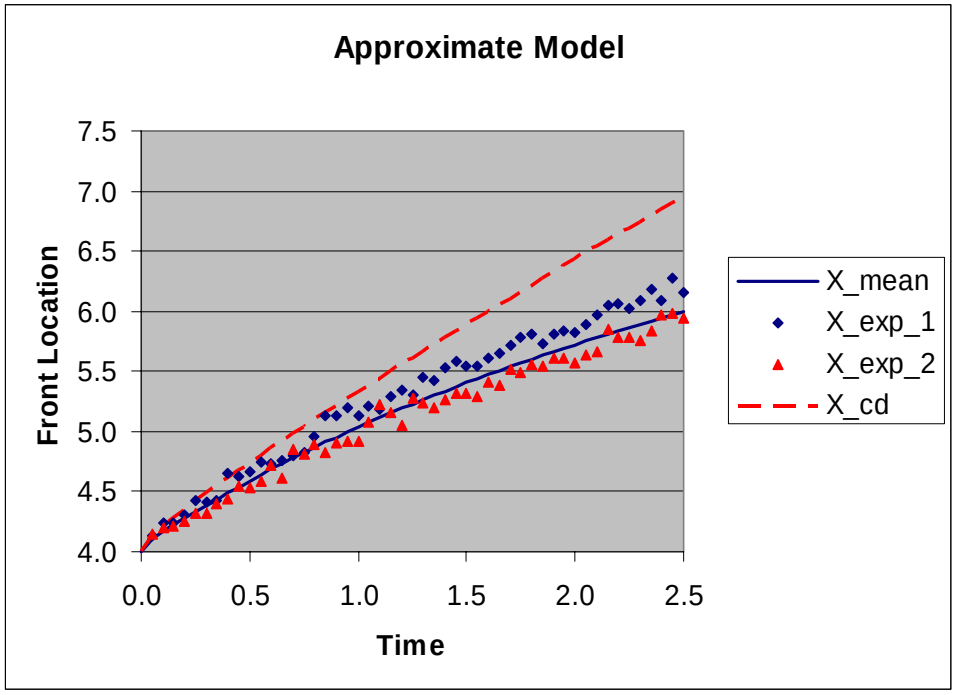

Figure 3.2 **Predicted and Experimental Measurements for Front Location**: X_mean – prediction base on mean model parameters; X_exp_1, X_exp_2 – simulated experimental data for experiments 1 and 2, respectively: X_cd – prediction based on convective-dispersive equation.

A comparison of the simulated measurements from the previous section and the predictions using the following somewhat arbitrary choice for the parameters in the C-D equation is illustrated in Figure 3.2.

$$D = 0.08, \quad V = 1.0 \qquad (3.14a, b)$$

We will discuss less arbitrary choices for these parameters in Chapter 5. Note that due to

the nonlinearity in Burgers' equation, the front moves at increasing slower speeds, relative to that for the C-D equation. Clearly, the non-modeled physics of Burgers' equation is not well approximated by the C-D equation for this measure of system response. However, as will be shown later, the two parameters in the C-D model can be calibrated to obtain improved results over a limited time interval. We will demonstrate this process and discuss its implications on validation in Chapter 5.

(Page left blank)

4.0 Complete Physics-Uncertainty Model

4.1 Introduction

The procedures and results discussed in this chapter focus on the validation of models that we claim accurately represent the physics, but contain uncertainty in their parameters. We use the phrase "accurately represent the physics" to mean that the model resolves physical behavior within the characterized uncertainty of the validation experiments. The metrics developed previously by this author (Hills and Trucano, 1999, 2001, 2002; Hills and Leslie, 2003; and Hills et al., 2003) evaluates whether this accuracy is obtained. While the presentation of this approach does not provide any new concepts, it does provide some insight as to how one might test a model for it's ability to extrapolate from the test conditions and provides background for the evaluation of approximate models in the next chapter.

4.2 Validation – Complete Data Set

We begin with the use of the r^2 metric of Hills and Trucano (1999, 2001) applied to the data of Figure 3.1, and the predictions of Burgers' equation. Since Burgers' equation was used to generate the simulated data, we expect that a validation metric should find this data to be consistent with the model predictions.

While we could perform a full Monte Carlo analysis to evaluate the effect of parameter uncertainty on Burgers' equation, we will use the locally linear approximation employed by Hills et al. (2003) to estimate the prediction uncertainty. This approximation, Eq. (3.6), is based on a first order sensitivity analysis discussed in the previous chapter and requires considerably fewer function evaluations of Burgers' equation than does a Monte Carlo analysis for this two parameter problem. We assume previous experience indicates that the model predictions are normally distributed with a know distribution, and will derive the covariance matrix derived from a first order sensitivity analysis, Eq. (3.9). Because we assumed that the uncertainty in the measurements is normally distributed, the difference between model predictions and experimental observations will also be normally distributed. Hills and Trucano (2001) illustrate methodology based on Monte Carlo analysis to handle non-normal distributions. Hills and Trucano (2002) provide an alternative metric based on maximum likelihood, which does not require a sensitivity analysis or as many function evaluations as a Monte Carlo analysis for nonlinear, non-normally distributed systems.

The r^2 metric used here inversely weights the differences relative to their uncertainty. The added advantage of this metric is the resulting r^2 has a χ^2 distribution for normally distributed differences, which is tabulated in most statistics text books (see for example, Brownlee, 1965). This metric is given by

$$r^2 = (\mathbf{X}_{model} - \mathbf{X}_{exp})^T [\mathbf{cov}^{-1}(\mathbf{X}_{model} - \mathbf{X}_{exp})] (\mathbf{X}_{model} - \mathbf{X}_{exp}) \quad (4.1)$$

where

$$\mathbf{cov}(\mathbf{X}_{model} - \mathbf{X}_{exp}) = \mathbf{cov}(\mathbf{X}_{model}) + \mathbf{cov}(\mathbf{X}_{exp}) \quad (4.2)$$

$\mathbf{X}_{model}$ is a vector of model predictions using mean values for the parameters, and $\mathbf{X}_{exp}$ is the vector of experimental observations developed at the beginning of the previous chapter. The covariance matrix for the model predictions is estimated using Eq. (3.9) where the mean values for *D* and *U* are given by Eq. (3.3). For our case, the covariance matrices of the model parameters and the experimental observations are (see Eqs. (3.4) and (3.5))

$$\mathrm{cov}(\boldsymbol{\alpha}) = \begin{bmatrix} 0.008^2 & 0 \\ 0 & 0.1^2 \end{bmatrix} = \begin{bmatrix} 0.000064 & 0 \\ 0 & 0.01 \end{bmatrix} \quad (4.3)$$

$$\mathrm{cov}(\mathbf{X}_{exp}) = \begin{bmatrix} 0.05^2 & 0 & \cdots & 0 \\ 0 & 0.05^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0.05^2 \end{bmatrix} \quad (4.4)$$

We use a finite difference approximation to estimate the components of the sensitivity matrix. Specifically, we use a forward difference with an increment that is 5% of the mean value of the model parameter of interest. Using the above results to evaluate Eq. (4.1) gives

X_exp_1: $r^2 = 56.85$ (4.5a)
X_exp_2: $r^2 = 64.68$ (4.5b)

The significance of this value for r^2 can be evaluated from the $\chi^2(df)$ distribution where df is the degrees of freedom. In our case, we have 51 differences with known variances. Since we did not estimate any parameters, the total degrees of freedom is 51. Given the $\chi^2(51)$ distribution, the probability of obtaining an $r^2 = 56.85$, 64.68 or larger is

X_exp_1: $P(r^2 > 56.85) = 0.266$ (4.6a)
X_exp_2: $P(r^2 > 64.68) = 0.094$ (4.6b)

Thus, given our models for uncertainty in the model predictions due to parameter uncertainty, and in the uncertainty in the measurements, the probability of a valid model given this large or larger value of uncertainty in the weighted distance squared, Eq. (4.1), is 26.6% for data set 1 and 9.4% for data set 2. This is more significant than the 5% that we typically require to reject a model. We can thus say that the data do not provide

sufficient evidence to reject the model as valid. This is not surprising since this model was used to generate the simulated data. But it is a sanity check on our methodology.

4.3 Validation – Partial Data Set

If the model of interest will be used to extrapolate over time, then we will not have data over the full range of time. We can simulate this by using only the data for $t \leq 1.0$, (i.e., the first 21 data points) from the first data set. We use only the first 21 rows of the sensitivity matrix, Eq. (3.8), and the upper left 21x21 submatrix of the measurement covariance matrix, (4.4), to evaluate a distance squared for the corresponding 21 degrees of freedom.

$$r^2 = 30.175, \quad df = 21, \quad P(r^2 > 30.175) = 0.0885 \tag{4.7}$$

Compared to the previous case for data set 1, there is less evidence that the model is correct. A valid model would have only a 9% chance of possessing this r^2. Because the level of significance for this test is approaching the 5% level, we should investigate the validity of this model further before it is used for extrapolation to later times. There are several reasons why one may obtain a low level of significance, even though the model is valid. These include the following:

1. The numerical model does not accurately represent the mathematical model (algorithm or convergence issues).
2. The uncertainty in either the model parameters and/or the measurements has been underestimated, and/or we are using the wrong distribution, and/or the method used for the uncertainty propagation may not be accurate.
3. There is significant bias in the measurements.
4. We simply have a set of measurements or model parameters that were realized near the tails of their distributions.

In the present work, we intentionally choose a realization for data set 1 whose joint probability for the first 21 differences between the model predictions and the simulated measurements was near the tail of the predicted distribution (item 4). This is reflected in Figure 3.1 and provides a more interesting case to study than a realization for which the data aligns with the model predictions when using the mean values for the model parameters.

4.4 Extrapolation

If we are confident that the model can simulate the correct physics and that the models for uncertainty capture the true uncertainty, then we may feel confident that this model can be used to extrapolate outside the parameter or time range tested. One advantage of characterizing uncertainty in the model parameters, and testing a model against this uncertainty, is this provides us some confidence in not only extrapolation of the model predictions, but also the extrapolation of the predicted uncertainty.

We illustrate this extrapolated predicted uncertainty by using the sensitivity analysis performed at the beginning of this section, applied to all 51 data points, to estimate prediction intervals on extrapolated model predictions. The square root of the diagonal elements of the covariance matrix for the model predictions, Eq. (3.9), gives the standard deviations for the corresponding prediction times. Figure 4.1 illustrates the prediction intervals at the 5% significance level ($\pm 1.96\,\sigma$) for each measurement time. Note that the parameter-induced uncertainty increases from zero at time zero, to larger values at later times. Our current model for the uncertainty assumes no uncertainty in the initial conditions. If uncertainty in the initial conditions is significant, we could incorporate this through the parameterization of the initial conditions and the incorporation of the respective sensitivities of the front motion to these additional parameters. Note also that the variation in the interval width appears to not be entirely smooth. We believe that this is due to numerical noise generated by the flux correctors in the numerical scheme. The effect of this noise is amplified when using finite differences to estimate the sensitivity matrices, as required by Eq. (3.9).

The simulated measurements from Figure 3.1 are also shown in Figure 4.1 for reference. Because we have not included the effects of measurement uncertainty in the error bars of this figure, one should not judge the ability to extrapolate until such uncertainty is included. This will be evaluated in the next section.

We wish to emphasis that these prediction intervals are only as good as our underlying model. Based on the validation test for the first 21 data points performed in the previous section, the uncertainty in the experimental data and model predictions can account for the differences between model predictions and experimental observations at the 9% significance level. While we do not consider this a significant reason to reject the model, we suggest that our standards of performance should be higher if a model is to be used for extrapolation rather than interpolation.

4.5 Validation of the Extrapolated Model

If we have sufficient data, we can use a bootstrap approach to march our way up the front location versus time curve to see if the consistency of the model with the data decreases with time. More specifically, we can look at the significance of the agreement between the data and the model predictions over time, using a sliding window. A decrease in significance over time can serve as an indicator that the model's ability to extrapolate also decreases over time. While this approach does not provide a fail safe check for the validity of model extrapolation (there is no fail safe check for this other than through data at the extrapolated conditions), it can show trends relative to the known validation exercise uncertainty.

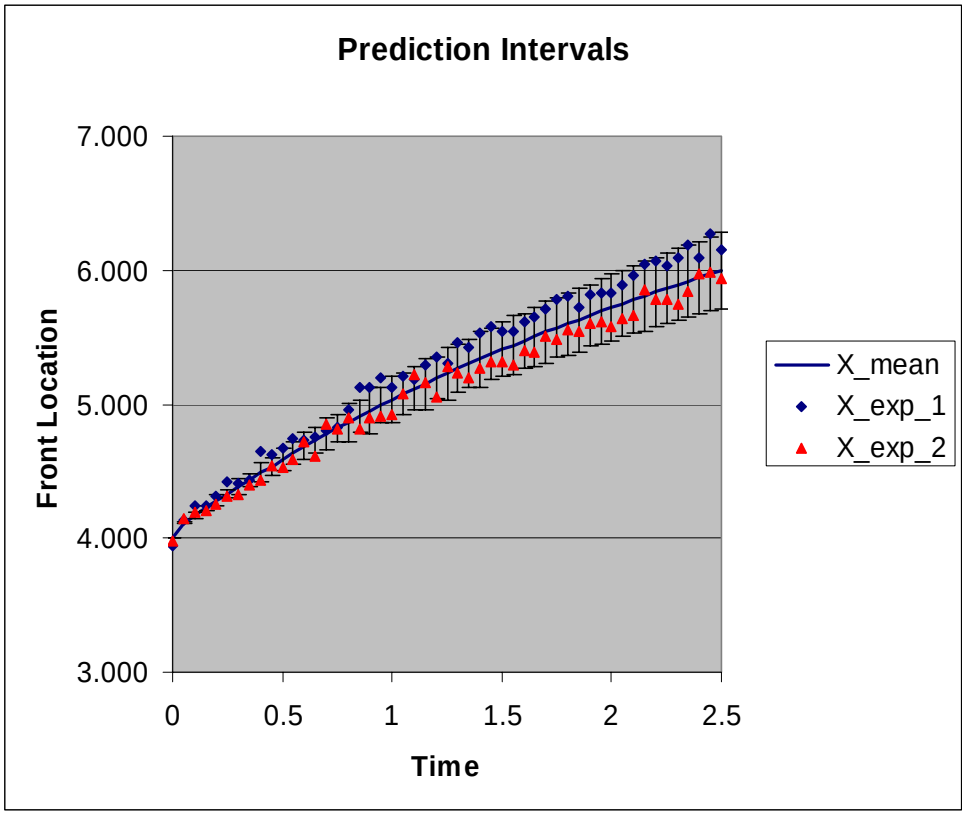

Figure 4.1 Prediction Intervals for Burgers' Equation due to Parameter Uncertainty

To demonstrate the use of significance for the present model, we use the following 5 time windows of the data from Table 3.1.

$$0 < t \leq 0.5, \quad 0.5 < t \leq 1.0, \quad 1.0 < t \leq 1.5, \quad 1.5 < t \leq 2.0, \quad 2.0 < t \leq 2.5 \qquad (4.8)$$

Applying the metric defined at the beginning of this chapter to the model predictions and the simulated experimental observations listed in Table 3.1, we obtain the results for significance (df = 10 for each case) as a function of time shown in Table 4.1.

Note that the significance is quite large for all the time windows greater than $t = 1.0$ for the X_exp_1. This is in contrast to the significance for the first 21 data points indicating that the agreement between predictions and measurements, relative to the uncertainty in the validation experiment, is better at later times than for the first 21 data points ($t \leq 1$). Because this is a random process, we would expect that significance to vary somewhat randomly over time. If we find that the trend in significance at later time is approaching small values (say ≤ 0.1), then we should question whether the model is appropriate for further extrapolation without further investigation of the reason for these small values of

significance. Note that the last time interval for both X_exp_1 and X_exp_2 has a significance that is less than the previous interval. This could be due to a reduction in the ability of the model to extrapolate to future times, or due to the random nature of the process. In the present case, we know that this reduction is due to the random nature of the process because we used Burgers' equation to generate the data. We would not have the luxury of knowing the data came from a known model in a real world case. This effect is also illustrated by the X_exp_2 data in the interval $1 \leq t \leq 1.5$. This data contains a large quantity of noise, and its trend does not appear to be consistent with the model as indicated by the metric.

Table 4.1 Significance of Agreement as a Function of Time

Time	r^2	$P(\rho^2 > r^2)$
X_exp_1		
$0.0 < t \leq 0.5$	10.94	0.362
$0.5 < t \leq 1.0$	20.18	0.028
$1.0 < t \leq 1.5$	9.44	0.491
$1.5 < t \leq 2.0$	8.62	0.569
$2.0 < t \leq 2.5$	11.93	0.290
X_exp_2		
$0.0 < t \leq 0.5$	3.41	0.970
$0.5 < t \leq 1.0$	15.29	0.122
$1.0 < t \leq 1.5$	20.28	0.027
$1.5 < t \leq 2.0$	6.58	0.765
$2.0 < t \leq 2.5$	15.30	0.122

The advantage of using significance to measure agreement over time is that it accounts for uncertainty in the model parameters and experimental observations. Thus even though we may see a systematic degradation in agreement between measurement and prediction over time, this degradation may simply be due to the increase in uncertainty of the validation experiments as a function of time, and not due to decreasing consistency of the model as it is extrapolated. This approach also properly accounts for parameter-induced correlation in the model predictions as one extrapolates over time.

The previous results do suggest a weakness of using data in this fashion to test the ability of a complete physics model to extrapolate. The present methodology gives the benefit of doubt to a model in that we require a small level of significance before we reject the model. This methodology has a small probability of rejecting a good model, at the expense of increasing the probability of accepting a bad model. In the next chapter, we discuss the use of bounding models. Since we do not use these models to represent behavior, but only to bound behavior, we are able to apply higher standards of acceptance that the bounding models are consistent with the data. This will be illustrated in the next chapter.

5.0 Approximate Physics Models

5.1 Introduction

In contrast to the previous chapter, here we assume (or acknowledge) that our model is approximate and that we use it to represent or bound behavior. As an example, we will illustrate the use of both a data-based, linear correction to predictions from the convective-dispersive equation (C-D equation); and the use of a calibrated C-D equation to approximate the behavior of a non-linear system (modeled by Burgers' equation).

We also develop several types of bounding models. Bounding models are sometimes used when 1) we do not have knowledge of the appropriate model for the full physics, but we are confident in our ability to bound the physics; or 2) we do have knowledge of the appropriate model for the full physics, but practical considerations do not allow full numerical resolution of the physics. Modeling at less than full resolution can occur when we have multi-scale physical phenomena, such as turbulence, chemical reactions in a mass transport system, and joint friction (typically modeled with one or a few finite element cells) in a complex structure.

5.2 Linear Correction Model

In this section, we develop a linear correction based on the differences between experimental observations and model predictions of front motion from the C-D equation. The purported advantage of this approach is one can develop corrections using simple linear regression on the differences between the experimental results and the model predictions, which require that the model be run only once. Using the following parameter values in Eq. (3.12) result in the predictions for front location as a function of time illustrated by the curve labeled X_cd in Figure 5.1.

$$D = 0.08,\ V = 1.0 \qquad \text{(5.1a, b)}$$

We used the same diffusivity and velocity as was used for Burgers' equation for demonstration purposes. Inspection of Figure 5.1 indicates that the resulting predictions are somewhat consistent for early time, but too high for later time. We develop a linear correction model for the differences between the first 21 measurements from data set 1 and the first 21 predictions of the C-D equation, as follows. The correction equation is defined as:

$$X_{\text{exp}}(t) - X\_\text{cd}(t) \approx \text{a} + \text{b}t \qquad (5.2)$$

a and b are estimated using simple least squares. Applying Eq. (5.2) to the first 21 data points gives

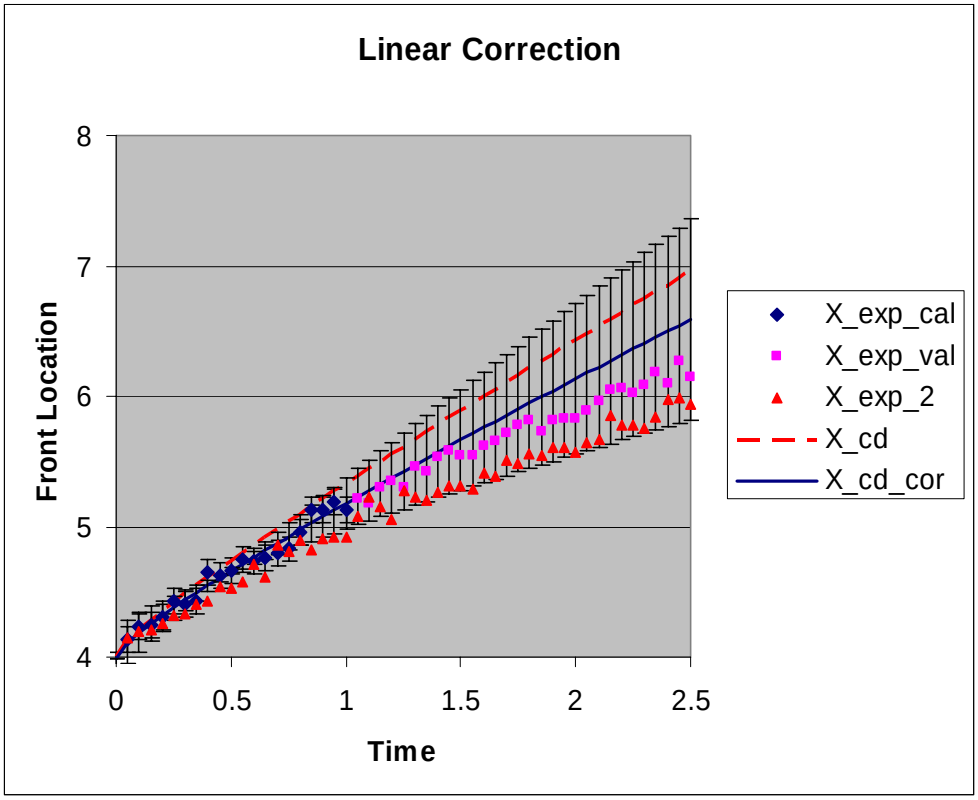

Figure 5.1 **Linear Correction**: The data labeled X_exp_cal was used for calibration

$$\begin{bmatrix} X_{\exp}(t_1) - X\_\mathrm{cd}(t_1) \\ X_{\exp}(t_2) - X\_\mathrm{cd}(t_2) \\ \vdots \\ X_{\exp}(t_{21}) - X\_\mathrm{cd}(t_{21}) \end{bmatrix} \approx \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_{21} \end{bmatrix} \begin{bmatrix} \mathrm{a} \\ \mathrm{b} \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mathrm{a} \\ \mathrm{b} \end{bmatrix} \tag{5.3}$$

where **A** is the sensitivity matrix (second matrix in Eq. (5.3)). The least squares solution is given by

$$\begin{bmatrix} \mathrm{a} \\ \mathrm{b} \end{bmatrix} = \mathbf{Y}_{\mathrm{ls}} \begin{bmatrix} X_{\exp}(t_1) - X\_\mathrm{cd}(t_1) \\ X_{\exp}(t_2) - X\_\mathrm{cd}(t_2) \\ \vdots \\ X_{\exp}(t_{21}) - X\_\mathrm{cd}(t_{21}) \end{bmatrix} \tag{5.4}$$

where

$$\mathbf{Y}_{ls} = \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T \tag{5.5}$$

The covariance matrix of the estimated parameters is

$$\mathrm{cov}\begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{Y}_{ls}\ \mathrm{cov}(\mathbf{X}_{exp})\ \mathbf{Y}_{ls}^{\ T} \tag{5.6}$$

The corresponding corrected equation is

$$X\_\mathrm{cd\_cal}(t) = X\_\mathrm{cd} + (a + bt) \tag{5.7}$$

with the associated prediction uncertainty given by (assuming no uncertainty in X_cd)

$$\mathrm{cov}\left(\mathbf{X}\_\mathrm{cd\_cal}\right) = \mathbf{A}\ \mathrm{cov}\begin{bmatrix} a \\ b \end{bmatrix} \mathbf{A}^T \tag{5.8}$$

Application of Eq. (5.4) to the first 21 measurements (i.e., measurements for which $t \leq 1.0$) results in the following least squares estimate for the linear correction parameters:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -0.00650 \\ -0.1442 \end{bmatrix} \tag{5.9}$$

Adding this correction (see Eq. (5.7)) to the predicted X_cd results in the curve labeled X_cd_cor shown in Figure 5.1. Note that the corrected results do appear to provide good predictions over the times of calibration, $t \leq 1$, when compared to the data from X_exp_1. While the corrected model provides predictions closer to the experimental observations than the uncorrected model, the corrected model still over predicts the results for $t \geq 1$. The 95% confidence prediction intervals ($\pm 1.96\ \sigma$) due to estimated parameter uncertainty are also shown in Figure 5.1 where the σ are given by the square roots of the diagonal elements in Eq. (5.8). The corresponding $\pm 1.96\ \sigma_{meas}$ uncertainty intervals for the measurements are shown in Figure 3.1. The expanding height of the uncertainty intervals over time reflects the sensitivity of the linear correction model to uncertainty in the linear correction model parameters, as one extrapolates from the time range of the calibration data. Note that little of the data for $t > 1$ falls outside these prediction intervals.

The natural question is – are the differences between the calibrated model predictions and the experimental data significant? To answer this question, we must account for the uncertainty in both the calibration parameters and in the measurements. Because we have assumed normal distributions throughout, and because our correction model is linear in the parameters, the uncertainty in the corrected predictions will also be normally distributed. Using the r^2 metric defined by Eq. (4.1) gives

$$r^2 = (\mathbf{X}\_cd\_cal - \mathbf{X}_{exp})^T [\mathbf{cov}^{-1}(\mathbf{X}\_cd\_cal - \mathbf{X}_{exp})] (\mathbf{X}\_cd\_cal - \mathbf{X}_{exp}) \quad (5.10)$$

where

$$\mathbf{cov}(\mathbf{X}\_cd\_cal - \mathbf{X}_{exp}) = \mathbf{cov}(\mathbf{X}\_cd\_cal) + \mathbf{cov}(\mathbf{X}_{exp}) \quad (5.11)$$

Because we estimated two parameters from the data when developing the calibration, r^2 is distributed as a $\chi^2(n-2)$ distribution with n - 2 degrees of freedom. Applying the above equations to our data results in the significances listed in Table 5.1 for several time windows. Note that in all cases, except for the case when all measurement times are used, the significance is greater than 5% for data set 1. Also note that as time increases past the calibration time ($t \leq 1$), the significance decreases over time. This suggests that the extrapolation of the model much past $t = 2.5$ is questionable. The reason for the low significance (zero to 3 significant figures) when using all measurement times is the ability of the metric to resolve a bad model increases with more data, assuming correlation is properly accounted for. This indicates that the linear correction model is not consistent with the data, within the uncertainty defined by our estimates of the linear correction parameters and the uncertainty associated with measurement error, over the full range of data.

Table 5.1 Significance of Agreement: Linear Calibration

Time	r_o^2	$P(r^2 > r_o^2)$
X_exp_1		
$0.0 < t \leq 2.5$	88.53	0.000
$0.0 < t \leq 1.0$	24.63	0.173
$1.0 < t \leq 1.5$	6.29	0.614
$1.5 < t \leq 2.0$	9.88	0.274
$2.0 < t \leq 2.5$	11.69	0.166
X_exp_2		
$0.0 < t \leq 2.5$	97.29	0.000
$0.0 < t \leq 1.0$	48.80	0.000
$1.0 < t \leq 1.5$	20.27	0.010
$1.5 < t \leq 2.0$	9.11	0.333
$2.0 < t \leq 2.5$	16.13	0.041

The results of Table 5.1 and Figure 5.1 assume that there was no variability in the values of the parameters used in the convective-dispersive equation. To account for such variability, one must propagate this variability through the convective-dispersive equation. Doing so negates the purported advantages of the linear correction method, that of requiring only one evaluation of the model. The effect of another experimental realization is illustrated by comparisons between the calibrated model and X_exp_2. We see visually that the comparisons are much worse. The lower significances tabulated in Table 5.1 for X_exp_2 relative to X_exp_1 also support the statement that the calibrated model is only relevant to the particular realization of the experiment represented by the data from data set 1. This last result illustrates the dangers of calibrating a model which

contains significant prediction uncertainty, due to unit to unit variability, to results from a single experiment. A necessary but not sufficient condition to mitigate this danger is to use of data from an ensemble of experiments that truly reflect this unit to unit variability.

5.3 Calibration of the Convective-Dispersive Equation

Rather than developing a linear correction to the C-D equation, we can calibrate the C-D equation directly. In either case, we are estimating two parameters. The disadvantage of calibrating the C-D equation directly, relative to the linear correction, is this approach requires multiple evaluations of the physics-based model. The advantage is that we are using a calibrated model that incorporates some of the true physics. To perform the resulting nonlinear calibration, we use the IMSL (1997) function **bconf**. This function is a nonlinear optimization routine that allows for simple bounds. Specifically, we select D and V in Eq. (3.12) that minimizes the sum of the square of the differences between the measured front location and the predicted front location for the first 21 measurements for X_exp_1. The resulting calibrated parameters are listed below and the resulting model predictions are shown in Figure 5.2. Twenty four function evaluations were required by **bconf**.

$$D = 0.0808, \quad V = 0.853 \tag{5.12a, b}$$

We can show the effect of uncertainty in the estimated parameters as we did in the previous section. However, we prefer to show the effect of model parameter uncertainty on predicted front location, to account for the uncertainty from realization of the experiment to experiment. We use Eq. (3.9) to estimate the covariance of the predictions due to the covariance of the parameters from experiment to experiment. For illustrative purposes, we assume that this covariance is given by

$$\mathrm{cov}(\boldsymbol{\alpha}) = \begin{bmatrix} 0.008^2 & 0 \\ 0 & 0.1^2 \end{bmatrix} = \begin{bmatrix} 0.000064 & 0 \\ 0 & 0.01 \end{bmatrix} \tag{5.13}$$

The effect of the resulting parameter uncertainty is illustrated through 95% prediction intervals in Figure 5.2. In contrast to the previous case, we see that the later time X_exp_2 data lies outside the model-parameter induced prediction intervals. We also see that the prediction intervals are somewhat smaller at later times than those shown in the previous figure. This is because we are accounting for the uncertainty in the model parameters based on our prior knowledge of this uncertainty, not on the uncertainty in the estimates of the calibrated model parameters due to the uncertainty in the calibration data. If we were to account for both, the intervals would be even larger than those shown in Figure 5.2.

Because we have assumed normal distributions throughout, and because our first order sensitivity analysis is linear in the parameters, we will model the corrected predictions as

normally distributed. Hills and Trucano (2001) demonstrate the use of Monte Carlo methods to account for non-normal distributions and nonlinearity in the model parameters. Using the r^2 metric defined at the beginning of Chapter 4.1 gives

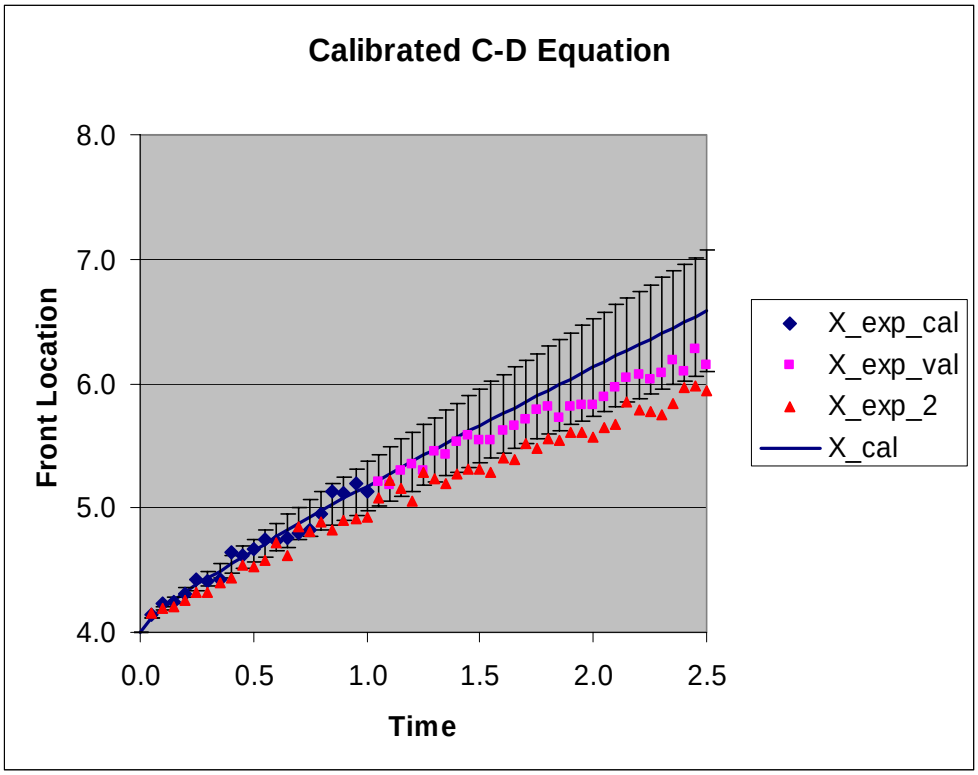

Figure 5.2 Calibrated Convective Dispersion Equation

$$r^2 = (\mathbf{X}\_cal - \mathbf{X}_{exp})^T [\mathbf{cov}^{-1}(\mathbf{X}\_cal - \mathbf{X}_{exp})] (\mathbf{X}\_cal - \mathbf{X}_{exp}) \quad (5.14)$$

where

$$\mathbf{cov}(\mathbf{X}\_cal - \mathbf{X}_{exp}) = \mathbf{cov}(\mathbf{X}\_cal) + \mathbf{cov}(\mathbf{X}_{exp}) \quad (5.15)$$

We evaluate **cov**(**X_cal**) by applying the first order sensitivity analysis developed for Burgers' equation in Chapter 3 (see Eq. 3.9) to the convective dispersion equation. Since we estimate two parameters, we have lost two degrees of freedom. In this case, the appropriate distribution is the $\chi^2(n-2)$ distribution with the degrees of freedom equal to the number of measurements, n, minus the number of estimated parameters, 2. The corresponding significances of the calibrated model are listed in Table 5.2 for various time windows. Note that the significances are somewhat smaller than those found using linear calibration. As in the case for linear calibration, we see that the significance decreases for time greater than 1.0 for X_exp_1. This indicates that, relative to the uncertainty in the measurements and the parameters, we have evidence that the calibrated

model becomes less effective as time increases. The zero significance (to 3 significant figures) for the calibrated model when using all measurement times indicates that even a calibrated C-D equation, when calibrated to the first 21 points, cannot predict front movement at all times within the uncertainty of the model parameters and the experimental measurements. As mentioned earlier, the more data one uses in the

Table 5.2 Significance of Agreement: Calibrated Convective-Dispersion Equation

	Time	r_o^2	$P(r^2 > r_o^2)$
X_exp_1			
	$0.0 < t \leq 2.5$	113.7	0.000
	$0.0 < t \leq 1.0$	24.92	0.163
	$1.0 < t \leq 1.5$	6.435	0.599
	$1.5 < t \leq 2.0$	11.94	0.154
	$2.0 < t \leq 2.5$	13.51	0.096
X_exp_2			
	$0.0 < t \leq 2.5$	88.58	0.000
	$0.0 < t \leq 1.0$	22.91	0.242
	$1.0 < t \leq 1.5$	26.23	0.001
	$1.5 < t \leq 2.0$	12.57	0.128
	$2.0 < t \leq 2.5$	18.84	0.016

evaluation of the above metric, the better the ability to resolve a bad model. Comparison of the results for X_exp_1 and X_exp_2 indicates that the significance of the model for several of the time windows is well less than the 5% at which we typically reject a model. So we see that a model calibrated to data from one realization of an experiment, may not represent the experimental behavior for another realization of the experiment. Proper calibration requires data from an ensemble of experiments that are sufficient to represent this unit to unit variability.

5.4 Bounding Models

In this section, rather than develop corrected or calibrated models, we develop bounds on the model predictions in the form of two bounding models, one for the lower bound and one for the upper bound. We begin by considering the case for which we have good knowledge of the physics and can develop bounding models based on this physics.

Consider the use of the convective-dispersive (C-D) equation to bound experimental observations from a system well modeled by Burgers' equation with the parameter values provided in the previous chapter. The governing equations for each of these models are repeated here:

Burgers' equation: $$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - U\frac{\partial c^2}{\partial x} = D\frac{\partial^2 c}{\partial x^2} - 2cU\frac{\partial c}{\partial x} \quad (3.1)$$

C-D equation: $$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - V\frac{\partial c}{\partial x} \quad (3.12)$$

Assume that we know and can represent the full physics mathematically (i.e., Eq. 3.1), but do not wish to do so with our numerical approximations. Comparison of these two equations indicates that the dispersive term in Burgers' equation is correctly modeled by the C-D equation, but the convective term is not. Rather than attempting to accurately model Burgers' equation with the C-D equation, we will attempt to use the C-D equation to bound predictions by Burgers' equation. This is equivalent to bounding non-modeled physics with good knowledge of this non-modeled physics.

If we wish to over-predict the front movement of Eq. (3.1), we should choose the last term in Eq. (3.12) so that it provides a faster moving front than the last term in Eq. (3.1). We can do this by setting $V = 2c_{\max}U$ and recognizing that the maximum c we can observe in the system (see Eq. (3.2)) represented by Eq. (3.1) is unity. Thus we take $V = 2U$ and use the following form of the C-D equation to bound the maximum front velocity of Burgers' equation from above:

$$\frac{\partial c_{ub}}{\partial t} = D\frac{\partial^2 c_{ub}}{\partial x^2} - 2U\frac{\partial c_{ub}}{\partial x} \quad (5.16)$$

where D and U are the corresponding values for Burgers' equation (we will discuss the case where we use different values in following sections).

Likewise, we can bound front velocity from below by using a value for V that is always smaller in magnitude than $2c_{\min}U$. Since the lowest value for c, for which we can observe the front for the current metric, is $c = 0.25$ (i.e., that value at which we define the front location, see Chapter 3), we can bound the front velocity from below by using $V=2(0.25\ U) = 0.5\ U$. The corresponding C-D equation is

$$\frac{\partial c_{lb}}{\partial t} = D\frac{\partial^2 c_{lb}}{\partial x^2} - 0.5U\frac{\partial c_{lb}}{\partial x} \quad (5.17)$$

The resulting prediction of front motion as a function of time and the experimental data are illustrated in Figure 5.3. Note that the model does, in-fact, bound front movement over the times of interest for both data sets, especially at later times. Unfortunately, the bounds are very broad and may not be very useful to the modeler. This is especially true for the upper bounds.

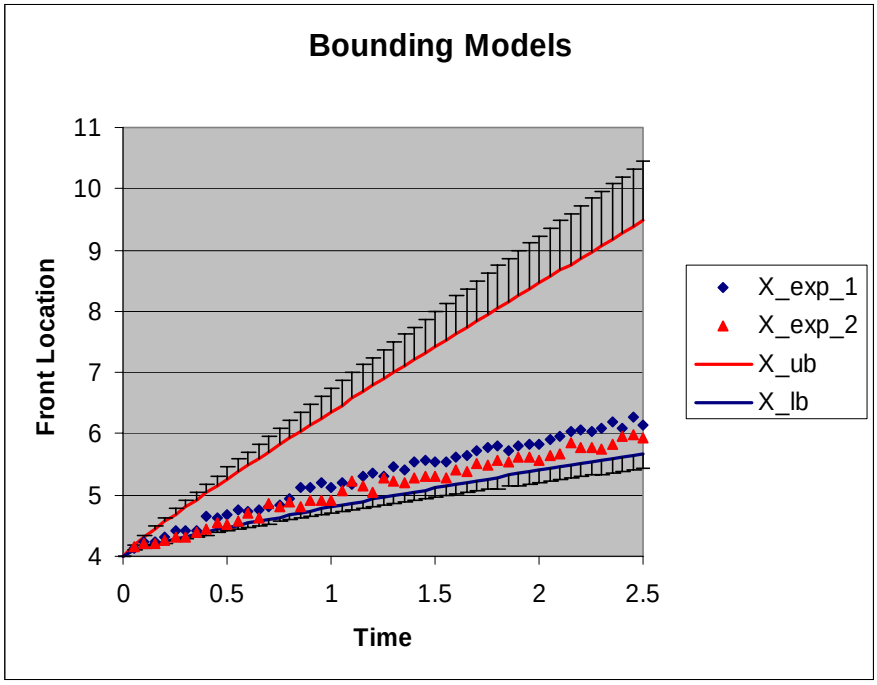

Figure 5.3 Front Movement: Bounding Models

5.4.1 Prediction Intervals

Figure 5.3 also contains prediction intervals on the predictions due to model parameter uncertainty. Unfortunately, the evaluation of these prediction intervals does not follow from the methodology presented previously. Here we discuss the approximate technique used to generate these intervals.

The difficulty in developing these intervals are three fold: 1) we actually have two models, one for the upper bound and one for the lower bound with uncertainty around each of these models due to parameter uncertainty, 2) the prediction interval on each bound represent one-sided intervals, and 3) the upper and lower bounds are correlated to each other since they both are derived from the same model parameters, D and U (see Eqs. (5.16) and (5.17)). If we wish to develop overall upper and lower prediction intervals that account for parameter uncertainty, then we need some method to develop the joint probability of these two distributions for each time. We will use an approximate method to develop these limits.

We wish to develop a lower prediction interval on the lower bound and an upper prediction interval on the upper bound such that the cumulative probability of obtaining a parameter set that leads to predictions outside these intervals is less than α_0. Specifically,

we wish to define significance levels on each of these two bounding models so that the overall significance is given by the following:

$$\alpha_0 = 1 - \text{P(no Type 1 error on } H_{o1} \text{ and no Type 1 error on } H_{o2}) \tag{5.18}$$

where P(no Type 1 error on H_{o1}) is the probability that a parameter set will result in predicted measurements above the lower prediction interval on the lower bounding model (i.e., due to parameter uncertainty), and P(no Type 1 error on H_{o2}) is the probability that the same parameter set will result in predicted measurements below the upper prediction interval on the upper bounding model. Since these two hypotheses are not independent, we cannot assume that the joint probability is equal to the product of their probabilities. Rather, we will use the Bonferroni inequality (see Miller, 1980) to estimate these intervals.

When applied to two events, Bonferroni's inequality states that the joint probability of the two events, E_1 and E_2, is related to the probability of the individual events by the following:

$$P(E_1 \text{ and } E_2) \geq 1 - P(E_1^C) - P(E_2^C) \tag{5.19}$$

where E_1^C and E_2^C are the complementary events to E_1 and E_2. If we require that the prediction intervals are defined so that the probabilities of these complementary events are both equal to α, then Eqs. (5.18) and (5.19) lead to

$$1 - \alpha_0 \geq 1 - \alpha - \alpha = 1 - 2\,\alpha \tag{5.20}$$

If we wish the overall probability of both of these events to be 95% or greater, then the significance of each event follows from Eq. (5.20).

$$\alpha \geq \alpha_0 / 2 = 0.05 / 2 = 0.025 \tag{5.21}$$

So we see that the significance of each event must be greater than or equal to 0.025. The broadest prediction intervals are obtained with the use of 0.025 (i.e., the smallest value of significance).

For the case considered here, we note that $V=2U$ for the upper bound and $V=0.5U$ for the lower bound. Because of this, the variance of V will be 4 times (i.e., 2^2) that of U for the upper bound and 0.25 (i.e., 0.5^2) that of U for the lower bound. The covariance matrices for the model parameters become (see Eq. (3.4))

$$\text{cov}(\boldsymbol{\alpha}) = \begin{bmatrix} 0.008^2 & 0 \\ 0 & 0.05^2 \end{bmatrix} = \begin{bmatrix} 0.000064 & 0 \\ 0 & 0.0025 \end{bmatrix}, \qquad \text{lower bound} \tag{5.22}$$

$$\mathrm{cov}(\boldsymbol{\alpha}) = \begin{bmatrix} 0.008^2 & 0 \\ 0 & 0.2^2 \end{bmatrix} = \begin{bmatrix} 0.000064 & 0 \\ 0 & 0.04 \end{bmatrix}, \qquad \text{upper bound} \qquad (5.23)$$

The sensitivity analysis method demonstrated for Burger's equation (see Eq. (3.8)) is applied to the present model for each bound, and the results are used in Eq. (3.9), along with the corresponding covariance matrices for the parameters, Eqs. (5.22) and (5.23), to evaluate the variances of the differences between the measurements and the model predictions.

The prediction intervals plotted in Figure 5.3 are at 1.96 σ for the lower bounds and 1.96 σ for the upper bounds. Note that for a one-sided prediction interval, 1.96 σ represents a significance of 2.5% for a normal distribution as defined by Eq. (5.21). We see from Figure 5.3 that the effect of parameter uncertainty is zero at time zero, but increases with time. This is due to the accumulative effect of error in front velocity over time and because we assumed that the initial conditions are known perfectly. Also note that the prediction interval height for the upper bound is much higher than that for the lower bound. This is due to the significantly larger uncertainty in the upper bound, velocity parameter, as reflected in Eq. (5.23) relative to Eq. (5.22).

5.4.2 Validation

As discussed above, the previous analysis serves to show the effect of parameter uncertainty on *D* and *U* on the upper and lower bounding models. Here we consider the development of a validation metric for bounding models. In this case, we need to consider both the effects of measurement uncertainty and model parameter uncertainty. As was the case for the prediction intervals due to parameter uncertainty, the development of a validation metric is complicated by the fact that we are dealing with two correlated, bounding models. In the spirit of the multivariate metric developed previously by Hills and Trucano (2001), we develop a multivariate metric for bounding models which accounts for correlation. Unfortunately, we cannot use the r^2 approach applied previously since we are interested in testing whether the models bound behavior rather than represent behavior. Thus we are interested in two one-sided tests rather than one two-sided test as discussed previously.

Here we use an approximate Monte Carlo approach which allows us to perform one-sided, multivariate tests while accounting for correlation. Specifically we evaluate the percentage of time that the Monte Carlo realizations for upper and lower bounds (including the effect of measurement error) are not violated by the observed measurements. The procedure is as follows:

1. The covariance matrix of the measurements is used to generate a perturbation vector, containing one element for each measurement time, for a mean of zero, using our chosen measurement distribution. While we used a normal distribution for the present case, other distributions can be just as easily handled with the

Monte Carlo method described here.

2. The probability distributions of the model parameters are used to generate a realization for these parameters. Again, we assumed a normal distribution.

3. Each realization of the model parameters is used to estimate the corresponding upper and lower bounding models for front histories though a first order approximation to the solution of Eqs. (5.16) and (5.17) given by the following equation.

$$\mathbf{X} = \mathbf{X}_{\text{bound}} + \nabla_{\boldsymbol{\alpha}} \mathbf{X} \left(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\text{bound}} \right) \tag{5.24}$$

   where the subscript "bound" indicates that the quantity is evaluated at the parameter values established in Section 5.4 for the lower or the upper bounds. The sensitivity matrix is evaluated through the application of Eq. (3.8) to the model predictions for front locations, based on Eqs. (5.16) and (5.17). As before, we use forward differences with an increment that is 5% of the "bound" values of the model parameters to estimate the elements in the sensitivity matrix.

4. We add the measurement perturbation vector generated in step 1 to the upper and lower bound predictions found in step 3 to account for the additional uncertainty of the measurements.

5. Steps 1 through 4 are repeated 10,000 times and the fraction of times the individual observed measurements fall within the realizations of the bounds (including measurement uncertainty) are counted. We use this fraction as an estimate for probabilities that the bounds contain the measurements, which, in-turn, are used as a measure of consistency of the bounding models with the observations.

The results of applying the above procedure are listed in Table 5.3 for several time intervals for both data sets. Because the solutions of Eqs. (5.16) and 5.17) are inexpensive, we can solve these equations directly for each realization rather than use the first order approximation defined by Eq. (5.24). We choose to use the approximate method in this report to illustrate the methodology. However, to check adequacy of our approximate method, we performed the Monte Carlo analysis (using the same initial random number seed and 10,000 realizations of each bounding model) using Eqs. (5.16) and (5.17) directly. We found that the results agreed within 0.001 of the results shown in Table 5.3 for all cases shown.

The results of Table 5.3 indicate that despite the uncertainty in the model parameters and the experimental measurements, the resulting bounding models bound the data 93% to 100% of the time for X_exp_1 and 75% to 99% of the time for X_exp_2. This indicates that the bounds are sufficiently conservative. The variability of these bounds, due to parameter uncertainty, is not sufficient to significantly overlap the variability in the

measurement, due to measurement uncertainty. Also note that the bounding models bound 98% of the data for time intervals greater than $t = 0.5$. This is a much stronger statement than that for the non-bounding models of the previous section. In the previous case, we gave the model the benefit of doubt because we required a very high certainty that a valid model could not produce the observed differences between experiment and model prediction before we would reject a model as valid. In the present case, we show a very high level of certainty that our bounds do bound the model. Of course, if we wish to narrow the bounds so that the bounds are within model parameter and measurement uncertainty of the data, then our level of significance will be smaller and our confidence that the model bounds the behavior, as reflected by the validation data, will be less.

Table 5.3 Significance of Bounding Models

Time	$P(X_{lb} \leq data)$	$P(data \leq X_{ub})$	$P(X_{lb} \leq data \leq X_{ub})$
X_exp_1			
$0.0 < t \leq 0.5$	0.929	0.982	0.955
$0.5 < t \leq 1.0$	1.000	1.000	1.000
$1.0 < t \leq 1.5$	1.000	1.000	1.000
$1.5 < t \leq 2.0$	1.000	1.000	1.000
$2.0 < t \leq 2.5$	1.000	1.000	1.000
$0.0 < t \leq 2.5$	0.969	0.994	0.981
X_exp_2			
$0.0 < t \leq 0.5$	0.755	0.981	0.868
$0.5 < t \leq 1.0$	0.953	1.000	0.977
$1.0 < t \leq 1.5$	0.989	1.000	0.994
$1.5 < t \leq 2.0$	0.977	1.000	0.988
$2.0 < t \leq 2.5$	0.972	1.000	0.986
$0.0 < t \leq 2.5$	0.918	0.989	0.954

5.5 Calibration of Bounding Models

The results from the previous section provided good confidence in the ability of our bounding models to bound the experimental measurements, for the data sets tested. This is not surprising considering how wide the bounds are. What can we do to narrow these bounds? One approach is to use calibration of the bounding models with the hope that the bounds will be narrower, but still acceptable. However, we must understand the assumptions associated with calibration, and the impact on validation.

5.5.1 Issues

Model calibration is typically performed using data which does not represent the full variability of the underlying system. For example, we often take data from one or more

experiments performed on a single experimental apparatus. Variability from test unit to test unit is not represented. An example of a case for which such variability is important is vibration damping due to slippage (including micro slippage) in mechanical joints. Mechanical slippage can be a very effective damping mechanism under high shock loads, and has been found to provide significant protection to electronics and other components under such conditions. Unfortunately, the modeling of such damping is very difficult due to the importance of fastener configuration and load, and changes in surface contact conditions due to fretting, chemical changes, and thermally induced stresses. Even the process of disassembly and reassembly of the same components can result in considerable differences in damping characteristics. For such cases, we must be very careful about assuming that the calibration constants are appropriate if the model is to be extrapolated to other shock loading conditions; applied to a different assemblage of components (or even reassembly); or applied to other environmental conditions that could affect joint surface friction, such as temperature, or changes in force distribution over the joint. The extrapolation to other conditions is fraught with danger due to the sensitivity of the model (in the case of structural vibration – predictions at high frequency) to external loading and assembly conditions. If, on the other hand, one can develop distributions or bounds on reasonable values for the parameters, then this information can be used to bound behavior.

If we do have sufficient experimental resources to run multiple, independent tests at the system level, we may be able to calibrate the model to the individual tests, resulting in a distribution of calibration constants. This approach was used by Hobbs et al. (2003) at the unit level to develop a distribution of activation energies for very small samples of foam. This distribution can then be propagated through the system level model to evaluate the corresponding uncertainty in the system predictions.

Various types of calibration can be used to condition a model. The worst case scenario, from a validation perspective, is pure calibration with a total lack of independent knowledge of the parameters and their distributions. In this case, we simply fit the model to a fixed set of data. If we have more measurements than the number of fitted model parameters, we can estimate uncertainty intervals on the fitted model parameters through the variance of the data about the fitted model. These, in-turn, can be used to estimate the corresponding uncertainty in the evaluation of a model at a given set of independent variables, due to the uncertainty in the estimated parameters. However, this uncertainty is the uncertainty in the mean or fitted values of the parameters, and is based exclusively on the variability of the data about the fitted model for that particular set of data. A single calibration experiment cannot tell us anything about the variability of these parameters from experiment to experiment, or the extrapolation from validation experiment to target application. We must have additional knowledge, either in the form of additional independent experiments to develop additional calibrations so that distribution of model parameters can be estimated, or prior knowledge of the uncertainty in the parameters and the measurements themselves. We take the position that a calibrated model, with the lack of additional independent data, cannot be used for model validation. If we must calibrate, then we must have other independent knowledge to validate against. An example of a

method that uses calibration with independent data for validation, is the Maximum Likelihood method discussed by Hills and Trucano (2002). In this case, calibration was used to estimate likely values of the parameters; and the significance of the estimated parameters, based on prior independent knowledge of the uncertainty distribution of the parameters, was used to evaluate validity. For example, if the values for the estimated parameters are unlikely, given our prior knowledge of their distribution; we reject the combined model for the physics, and the associated uncertainties in the model parameters and in the measurements.

If one does not possess complete knowledge of the distribution of model parameters, but does have some estimate of the bounds of these parameters, then calibration can be used to further condition the model for this particular realization of the tested system. This approach is useful if one desires to bound behavior in interpolation, or to bound extrapolation **if and only if** one has confidence that the calibrated constants apply to the extrapolated conditions without change. This later statement implies that the calibration constants are not sensitive to different realizations of the experiment, or to the extrapolate variables. Because this is an approach used (i.e., sometimes called dial turning or parameter tuning) for some systems of interest to the DoE, we provide an example of this approach below. As will be shown, one can obtain useful information from this approach if the purpose is to bound behavior rather than to predict behavior, and if we have some assurances that the calibrated parameters are appropriate for the extrapolation.

5.5.2 Under-Constrained Calibration

We begin with a case of model calibration when we have more unknown model parameters than measurements. We assume that we have prior knowledge of the bounds of the uncertain model parameters, but no knowledge of their distributions. We will also assume that the calibrated bounding parameters are appropriate for extrapolation, if the model is to be used for extrapolation. The applicability of this last assumption to a specific application is largely a matter of engineering judgment.

Consider the use of the linear C-D equation to model front movement well represented by the non-linear Burgers' equation. Also assume that we have knowledge that the following bounds on the parameters for the approximate model are reasonable (perhaps conservative).

$$0.04 \leq D \leq 0.12 \tag{5.25}$$

$$0.5 \leq V \leq 1.5 \tag{5.26}$$

Note that the range of these bounds is quite large. Now assume that we have measured the system response at a single time, $t = 1$, from X_exp_1, and obtained the results listed in Table 3.1.

$$X_{\exp}(1) = 5.131 \tag{5.27}$$

Our hope is that by requiring the model for front movement to go through this data point (within experimental error), combined with the parameter bounds of Eqs. (5.25) and (5.26); useful information can be provided. To implement these constraints, we define the following objective functions.

$$f_{lb} = 10.0 \mid X_{pred}(1) - X_{exp}(1) + 2.58\,\sigma_{exp}\mid + X_{pred}(1) \tag{5.28}$$

$$f_{ub} = 10.0 \mid X_{pred}(1) - X_{exp}(1) - 2.58\,\sigma_{exp}\mid - X_{pred}(1) \tag{5.29}$$

Note that σ_{exp} is the standard deviation due to measurement uncertainty. In defining the objective functions, we have subtracted and added the normally distributed measurement uncertainty at the 1% significance level (i.e., 2.58 σ_{exp}), which define the upper and lower bounds of the calibrated model. While this approach to defining the bounds is appropriate for this application, a random sampling approach will often be required when there is no evidence of a monotonic relationship between the measurement and the corresponding bound (i.e., monotonic in the sense that a lower value for the measurement results in a lower bound on the prediction). The IMSL (1997) function **bconf** was used to find the sets of parameters that minimize each of these equations for each time of interest. This is the poor man's approach to minimization of nonlinear functions with simple and nonlinear constraints. The use of this approach does require some trial and error to find the value of the weight (10.0 in Eq. (5.28) and (5.29)) that results in good agreement between the predicted and measured value of front location at $t = 1$, offset by the measurement uncertainty, while minimizing X_{pred} and $-X_{pred}$. In our case, we found the agreement to be within three significant figures. The preferred approach would be to use an algorithm that allows for nonlinear constraints, such as those provided in DAKOTA (Eldred, 2002), or a sampling method for more complex problems. The resulting parameter sets (all developed from the measurement at $t = 1$) as a function of time for the lower and upper bounds are listed in Table 5.4. Note that we define parameter bounds for times before as well as after the time of the calibration measurement. Also note that for the case of this model, the resulting diffusivity is at either its upper (late time) or lower bound (early time), with the estimated velocity taking on values that allow the bounding curves to satisfy the constraints given by Eqs. (5.28) and (5.29).

The bounding curves for the bounding parameters defined by the measurement at $t = 1$ are shown in Figure 5.4. We must evaluate the upper and lower bounds by evaluating Eqs. (5.16) and (5.17), up to the time of interest, for the parameter set evaluated for that time. Since the corresponding parameter sets are essentially constant over large periods of time for this case, we could use larger time increments and fewer function evaluations to evaluate the results of Figure 5.4 if computational resources were limited. We also show the lower and upper bounds of the model predictions given the original bounds on

Table 5.4 Bounding Parameters

t	D_{lb}	V_{lb}	D_{ub}	V_{ub}
0.00	0.040	0.746	0.040	0.991
0.05	0.040	0.746	0.040	0.991
0.10	0.040	0.746	0.040	0.991
0.15	0.040	0.746	0.040	0.991
0.20	0.040	0.746	0.040	0.991
0.25	0.040	0.746	0.040	0.991
0.30	0.040	0.746	0.040	0.991
0.35	0.040	0.746	0.040	0.991
0.40	0.040	0.746	0.040	0.991
0.45	0.040	0.746	0.040	0.991
0.50	0.040	0.746	0.040	0.991
0.55	0.040	0.746	0.040	0.991
0.60	0.040	0.746	0.040	0.991
0.65	0.120	0.641	0.040	0.991
0.70	0.120	0.641	0.040	0.991
0.75	0.120	0.641	0.040	0.991
0.80	0.120	0.641	0.040	0.991
0.85	0.120	0.641	0.040	0.991
0.90	0.120	0.641	0.040	0.991
0.95	0.120	0.641	0.040	0.991
1.00	0.120	0.641	0.040	0.991
1.05	0.120	0.641	0.040	0.991
1.10	0.120	0.641	0.040	0.991
1.15	0.120	0.641	0.040	0.991
1.20	0.120	0.641	0.040	0.991
1.25	0.120	0.641	0.040	0.991
1.30	0.120	0.641	0.040	0.991
1.35	0.120	0.641	0.040	0.991
1.40	0.120	0.641	0.040	0.991
1.45	0.120	0.641	0.040	0.991
1.50	0.120	0.641	0.040	0.991
1.55	0.120	0.641	0.040	0.991
1.60	0.120	0.641	0.040	0.991
1.65	0.120	0.641	0.040	0.991
1.70	0.120	0.641	0.040	0.991
1.75	0.120	0.641	0.040	0.991
1.80	0.120	0.641	0.040	0.991
1.85	0.120	0.641	0.040	0.991
1.90	0.120	0.641	0.040	0.991
1.95	0.120	0.641	0.040	0.991
2.00	0.120	0.641	0.040	0.991
2.05	0.120	0.641	0.040	0.991
2.10	0.120	0.641	0.040	0.991
2.15	0.120	0.641	0.040	0.991
2.20	0.120	0.641	0.040	0.991

2.25	0.120	0.641	0.040	0.991
2.30	0.120	0.641	0.040	0.991
2.35	0.120	0.641	0.040	0.991
2.40	0.120	0.641	0.040	0.991
2.45	0.120	0.641	0.040	0.991
2.50	0.120	0.641	0.040	0.991

the model parameters. The results are labeled X_lbt and X_ubt and they occur at

$$D_{lb} = 0.04,\ V_{lb} = 0.5 \qquad (5.30a,\ 5.30b)$$

$$D_{ub} = 0.12,\ V_{ub} = 1.5 \qquad (5.31a,\ 5.31b)$$

which represent the overall bounds for each parameter (see Eqs. (5.25), (5.26)). Note that the calibrated bounds are significantly narrower than the overall bounds. This suggests that if our calibration holds for the anticipated application of the model, then calibration can increase precision of the extrapolation of the bounding model. If calibration does not hold because we expect experiment to experiment variability due to parameter variability, then calibration to this one experiment does not help us and we must use the overall bounds.

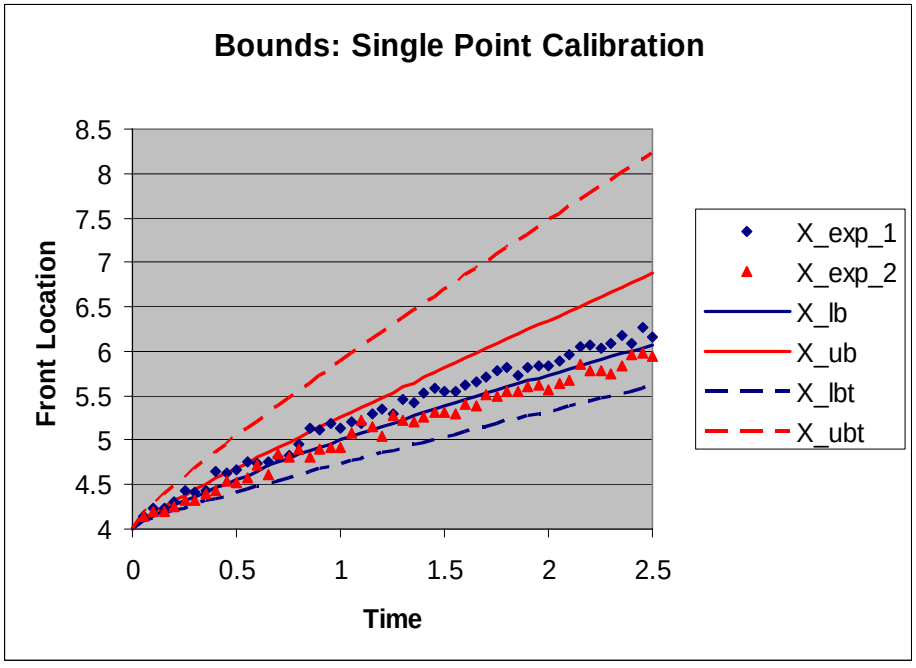

Figure 5.4 Bounding Models - Calibrated at Single Point (t = 1.0)

The calibrated bounds are clearly narrower than those shown in Figure 5.3, which indicate that calibration to this one point does allow us to narrow the bounds. We can evaluate the significance of these results using the procedure previously defined. In contrast to the previous case, here we hold the parameter values of the upper and lower bounds fixed at the values listed in Table 5.4, since we do not assume that we have uncertainty in these bounding parameter values. This allows us to use the bounding predictions shown in Figure 5.4 directly rather than generating Monte Carlo realizations of these predictions. We account for only the uncertainty in the measurements in the Monte Carlo analysis described in Section 5.4.2. The results of this analysis are presented in Table 5.5.

The results show that the combined bounds bound 99% or more of the measurements for each time interval for X_exp_1, including the effect of measurement uncertainty. We see that the lower bound is not as conservative as the upper bound, as illustrated in Figure 5.4, nor is the lower bound as conservative as it was for the previous case for later times. The calibrated bounds show less significance for X_exp_2. Inspection of Figure 5.4 and Table 5.5 illustrates that this lower significance is due to the decreased performance of the lower bound. This result is not surprising considering the data used for the calibration was from X_exp_1 rather than from X_exp_2.

Table 5.5 Significance of Bounding Models: Single Point Calibration

Time	$P(X_{lb} \leq data)$	$P(data \leq X_{ub})$	$P(X_{lb} \leq data \leq X_{ub})$
X_exp_1			
$1.0 < t \leq 1.5$	0.996	0.997	0.997
$1.5 < t \leq 2.0$	0.993	1.000	0.996
$2.0 < t \leq 2.5$	0.989	1.000	0.994
$0.0 < t \leq 2.5$	0.992	0.999	0.996
X_exp_2			
$1.0 < t \leq 1.5$	0.469	1.000	0.735
$1.5 < t \leq 2.0$	0.110	1.000	0.555
$2.0 < t \leq 2.5$	0.106	1.000	0.553
$0.0 < t \leq 2.5$	0.228	1.000	0.614

The advantage of this approach, relative to the previous case, is we did not require knowledge of the full physics model to develop these bounds (i.e., we did not assume a relationship between *U* and *V*). We also did not require full knowledge of the distribution of the model parameters of the approximate model, but rather only conservative bounds on these parameters. We do have to assume that the bounds, in-fact, bound behavior of the true physics, and that the single measurement, plus and minus the uncertainty, is representative of the range of possible realizations at this measurement time. Such was clearly not the case as the performance of the bounding models for X_exp_2 was considerably less than for X_exp_1. The validation metric presented here provides a

statistic measure to evaluate whether the bounding models bound the observations over the time range tested.

5.5.3 Calibration Using the First 21 Data Points

We now extend the case represented in Figure 5.4 to calibration using multiple data in an attempt to further narrow the upper and lower bounds. We calibrate the C-D equation to the first 21 measurements, and then extrapolate the results using the bounding methodology discussed earlier in this chapter. More specifically, we use the following model parameters:

$$D = D_{\text{cal}}, \quad 0 \le t \tag{5.32}$$

$$V = \begin{cases} V_{\text{cal}}, & 0 \le t \le 1 \\ V_{\text{cal}}, & 1 \le t, \text{ lower bound} \\ 0.25V_{\text{cal}}, & 1 \le t, \text{ upper bound} \end{cases} \tag{5.33}$$

where the subscript "cal" indicates calibrated values using the data up to $t = 1$ (i.e., the first 21 data points). Note that we maintain the ratio between the multipliers on U in Eqs. (5.16) and (5.17). The calibrated parameters are defined by finding the set of parameters that minimizes the sum of squared residuals between the first 21 measurements and the corresponding predictions. Again, we use the IMSL (1997) function **bconf**. The resulting estimated parameters were found to be (24 function evaluations required)

$$D_{\text{cal}} = 0.0808, \quad V_{\text{cal}} = 0.853 \tag{5.34a, b}$$

The resulting upper and lower bound predictions are shown in Figure 5.5 along with the corresponding estimates of the prediction intervals. We use the covariance matrices defined by Eqs. (5.22) and (5.23) for the bounding parameters and do not include the effect of parameter estimation uncertainty. As the results of the first example presented at the beginning of this chapter suggests, the uncertainty due to parameter estimation may be significant when we use only the first 21 points for calibration. This added uncertainty could result a significant increase in the width of the intervals for $t > 1.0$. Note that the uncertainty in the upper bounds is significantly larger than in the lower bounds. This is because we assumed that the standard deviation of the velocity for the upper bound model is 4 times that for the lower bound as indicated by Eqs. (5.22) and (5.23).

The corresponding significances of the observed data are listed in Table 5.6. These significances are less than those obtained from the previous case since we did not solve for the maximum and minimum bounds consistent with the data, but used a least squares estimate of the parameters in the bounding models. We also have considerable uncertainty in the tolerance intervals, especially the upper interval, as indicated by Figure 5.5. Note that the prediction intervals are actually two sided for each bounding curve. The true bounding curve can be either above or below the curves shown in the Figure.

The results of Table 5.6 indicate that the uncertainty in the upper bound is a significant contributor to the lower significance of the results for X_exp_1. However, note that the significance for the sliding time window increases with increasing time. This indicates that the bounds are becoming more and more conservative. This process illustrates the importance of evaluating the effect of parameter uncertainty when testing bounding models for validity. For such a model to be useful at the 95% confidence level, we would either need to reduce the uncertainty in the model parameters for the upper bound, or

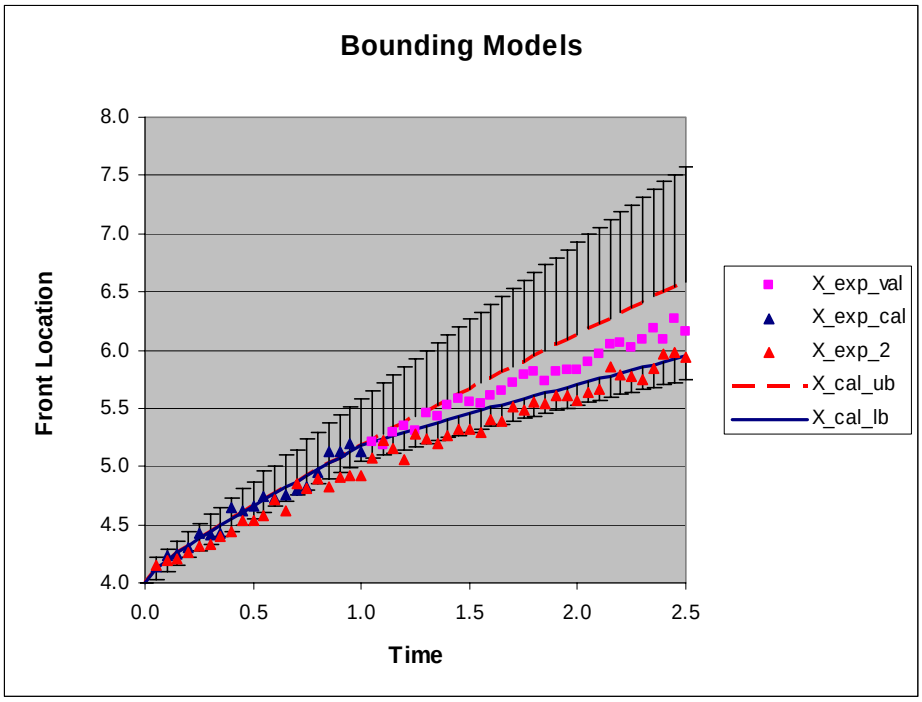

Figure 5.5 Bounding Models - Calibrated using 21 Points ($t \leq 1.0$)

Table 5.6 Significance of Bounding Models: 21 Point Calibration

Time	$P(X_{lb} \leq data)$	$P(data \leq X_{ub})$	$P(X_{lb} \leq data \leq X_{ub})$
X_exp_1			
$1.0 < t \leq 1.5$	0.695	0.583	0.639
$1.5 < t \leq 2.0$	0.926	0.693	0.809
$2.0 < t \leq 2.5$	0.987	0.739	0.863
$0.0 < t \leq 2.5$	0.869	0.672	0.770
X_exp_2			
$1.0 < t \leq 1.5$	0.109	0.791	0.450
$1.5 < t \leq 2.0$	0.172	0.870	0.521
$2.0 < t \leq 2.5$	0.447	0.880	0.663
$0.0 < t \leq 2.5$	0.243	0.847	0.545

settle for a more conservative model for this bound (i.e., shift the upper bounding curve upward).

If on the other hand, we use the prediction intervals shown in Figure 5.5 themselves as the bounds (i.e., more conservative), then we obtain the following results:

Table 5.7 Significance of Bounding Models: 21 Point Calibration, Intervals

Time	$P(X_{lb} \leq data)$	$P(data \leq X_{ub})$	$P(X_{lb} \leq data \leq X_{ub})$
X_exp_1			
$1.0 < t \leq 1.5$	0.996	1.000	0.998
$1.5 < t \leq 2.0$	1.000	1.000	1.000
$2.0 < t \leq 2.5$	1.000	1.000	1.000
$0.0 < t \leq 2.5$	0.999	1.000	0.999
X_exp_2			
$1.0 < t \leq 1.5$	0.678	1.000	0.839
$1.5 < t \leq 2.0$	0.872	1.000	0.936
$2.0 < t \leq 2.5$	0.993	1.000	0.997
$0.0 < t \leq 2.5$	0.848	1.000	0.924

Since we do not estimate the uncertainty in the location of the prediction interval, we do not include this source of uncertainty in the above results. The only source considered is the uncertainty due to measurement noise. Clearly, the bounds that are defined at the prediction intervals of Figure 5.5 include more of the data. These broader bounds are more conservative as indicated by both the figure and the significances listed in Table 5.7. However, as was the case for the previous calibrations, the calibration of a model to data from one realization of the experiment (X_exp_1) does not insure that the bounds will perform as well for another experimental realization (X_exp_2).

5.6 Summary

In this chapter, we introduced the concept of using and testing approximate, bounding, and calibrated models. While we present multiple types of bounding and calibrating models, there are many other combinations that can be considered. We also demonstrated methodology to test these models against data. We will discuss the relative merits of these various approaches in the next chapter.

6.0 Discussion

Our original plan was to focus on model extrapolation. As work progressed, we quickly realized that the impact of approximation on modeling, which exists in most complex system models, has a critical impact on extrapolation. For this reason, we broadened our focus to extrapolation and validation of approximate and bounding models.

In the review of the literature related to model validation and extrapolation, we did not find contributions that closely align with our interests. Generally, the concept of model extrapolation is implicit in modeling, when the model is based on fundamental principles (conservation principles, etc.), and issues associated with extrapolation are not explicitly discussed. The intent of most physical and engineering models is to incorporate sufficient fundamental knowledge so that models can be extrapolated over a reasonable range of parameters.

The issue of extrapolation complicates the concept of model validation. The simple use of the term extrapolation suggests that we are extrapolating outside the range of conditions for which the model is tested. Doing so puts increased dependence on our confidence that a model is valid at the test conditions, and that the fundamental principles incorporated in the model are sufficient for the extrapolation. Because we do not have data for the extrapolated conditions to test the model, we are ultimately depending on our confidence in the model to extrapolate to these conditions.

The acknowledgement that many (if not all) models for complex engineered systems contain approximation further complicates the concept of model validation and model extrapolation. In previous work by the first author and others (Hills and Trucano, 1999, 2001, 2002; Hills and Leslie, 2003; Hills et al., 2003), we focused on the case where the model, including any approximation that might have been incorporated, could predict behavior within the uncertainty of the validation experiments. We found that many of the models tested in this previous work could, in fact, predict behavior within this uncertainty. However, as the systems being modeled become more complex, the need to approximate some of the components of behavior increases due to the increased computation resources required to model all of these components. The idea of approximation invariability enters the modeling process.

The acknowledgement of approximation affects our approach to model validation. We no longer require that the model predict behavior within the uncertainty of the validation exercise (this uncertainty includes the effects of both the parameter uncertainty and the measurement uncertainty), but only that we test whether the model provides useful bounds on expected behavior. These bounds may be one sided in that we may wish to evaluate whether a critical component does not exceed a specified temperature, or may be two sided in that we bound temperature from above and below. In the case of approximation, we modify our concept of model validation to answer the question – are the observed experimental data consistent with the bounds defined by the approximate

models? Note that this statement leaves little room for models for which we acknowledge are approximate, but do not quantification of the level of this approximation, such as provide by prediction intervals or bounding models. While comparisons of model predictions to experimental observations, without a quantifiable sense of uncertainty, can be useful for the characterization of these differences, for this particular realization of the experiments; such comparisons do not provide useful results for the extrapolation of this model outside the range of the experiments. We contend that to extrapolate approximate models, we must not only compare the model predictions against the experimental observations, but we must also test the ability to bound the differences between prediction and observation (e.g., through bounding models or prediction intervals).

This last observation introduces an interesting question. Do we prefer to develop models that appear to represent behavior over the range of data tested and hope that the model extrapolates to the conditions required, or do we prefer to develop models that are conservative in the sense that we have high confidence that they bound behavior when extrapolated, rather than accurately represent behavior? As is the case with most engineering and scientific applications, the answer depends on our application, and on how confident we are that our model truly represents physical behavior.

The acknowledgement of approximation introduces another issue, that of calibration. Since approximate models generally have parameters that are calibrated to condition the approximation on observed data, the impact of calibration on approximate models is an issue that cannot be ignored.

In the present work, we presented a potpourri of ideas on possible methodology that acknowledges the effect of approximation and calibration on model validation and extrapolation. No attempt was made to provide a comprehensive overview of these ideas. Rather we intended to provide examples to suggest possible approaches to model validation, in the presence of approximation and extrapolation. We discuss these approaches below.

6.1 Full Physics Model

The first type of model we considered were those that we claim resolves the physics of the validation experiments to a level of uncertainty, sufficiently less than the uncertainty in the validation experiments (Chapter 4). In this case, we used Burgers' equation to model data that was, in-fact, generated by Burgers' equation. To provide simulated data that is somewhat realistic, we included different realizations of system behavior though the use of off-mean realizations of the model parameters, as well as the addition of measurement noise with zero mean.

To provide consistency with previous work (Hills and Trucano, 1999, 2001, 2002; Hills and Leslie, 2003; Hills et al., 2003), we first demonstrate one of their r^2 model validation metrics. We then looked at the use of time windows to investigate the temporal variation of the validation metric as an indicator of the ability of a model to extrapolate over time.

Because the model used was valid, we expect that the comparisons to support validity. However, as the example illustrated, we should also expect the agreement between the model and the experiment to be somewhat random for different time windows. Because of this expectation, it can be difficult to infer whether the model can correctly extrapolate, using trends in r^2 with time. A decrease in r^2 over the last two time windows may simply be a result of the randomness of the process and not an indication that the model is not valid for extrapolation. If, on the other hand, the value r^2 over any of the windows has a very low significance, then we must question the validity of the model over this window. One thing to keep in mind is that in this process, we are giving the benefit of doubt to the model. We require significant evidence (i.e., a low significance of agreement) before we reject the model as valid. Because of this, we are less likely to reject a good model, but more likely to fail to reject a bad model.

6.2 Approximate Model

The second set of examples provided was based on the calibration of predictions of the linear convective-dispersive (C-D) equation. In the two cases considered, we used the first 21 data points (i.e., the early time data) for calibration, and investigated the impact of this calibration on extrapolation over time using the remaining 30 data points. We consider two types of calibration. In the first, we developed a simple linear correction to the model predictions, as a function of time. In this case, we assumed that this linear correction over time, developed using the first 21 data points, extrapolates to later times. In the second case, we calibrated the dispersion coefficient and the velocity in the convection-dispersion equation to the first 21 data points, and used these estimated parameters in the C-D equation to extrapolate over time. We then demonstrated the time-windowed r^2 validation metrics used previously, to evaluate the consistency of the calibrated predictions to the experimental observations at later time.

It is an issue with both of these approaches that we know that calibrated models only approximate behavior, and we fully expect them to add bias to the later time predictions simply because these are approximate models. In fact, unless we understand something about the physics not being modeled, we will not know if the corrected or calibrated model will over or under predict behavior in extrapolation to later time without later time data.

A second issue with our implementation of the linear correction approach is that we did not incorporate known uncertainty in the model parameters into the correction. While we can bound the estimated linear model parameters due to estimation error resulting from the characterization of measurement uncertainty, we cannot incorporate the impact of experiment to experiment parameter uncertainty without either running the multiple realizations of the experiments, or by propagating this uncertainty through the model for the convective-dispersive equation.

A more rigorous approach in both cases would be to account for both parameter estimation uncertainty and experiment to experiment uncertainty. This would expand the

bounds further and make it less likely to reject a good model when using multiple realizations of the data, but also less likely to reject a bad model.

6.3 Bounding Models

The focus of the last set of examples was on the use of bounding models. Bounding models, by their nature are approximate. One advantage of bounding models, if they are valid, is they have a sense of direction from the true physics (i.e., they bound from above or below) and they can be developed to become increasingly conservative in the extrapolation. In contrast, models that we hope accurately represent the physics within the uncertainties of the validation exercise (Chapter 4), cannot become more conservative in extrapolation. The ability to become increasingly conservative over the extrapolation increases our confidence in bounding extrapolated behavior. Unfortunately, the development of bounding models that are truly conservative is not always an easy task, especially for models that must be conservative over multiple design variables. There is also a real danger of developing bounding models with bounding intervals so large that the model is not sufficiently predictive for the anticipated application.

One advantage of bounding models, from a model validation point of view, is these models represent an alternative hypothesis from the model that represents the data accurately. We can test at a very high level of significance, say 95% rather than 5%, that a model bounds behavior, given a set of observations. This is because we can require that the bounding models bound behavior, regardless of the uncertainty introduced due to parameter uncertainty and measurement uncertainty. Thus we are much more likely to reject an invalid bounding model. This does come at the expense of broader bounds.

Two types of bounding models were considered. The first model was developed through the use of known physics to develop bounding models without calibration. The second was developed through the use of single point and multiple point calibration. One advantage of calibration is that if the calibrated parameters are appropriate for extrapolation (i.e., there is no need to re-calibrate at later times), then calibration provides a mechanism to narrow the bounds of bounding models. If, on the other hand, we expect the realization of the application to possess different calibration parameters than the validation experiments, the use of calibration will lead to incorrect bounds at later times due to the bias introduced by the incorrect values for the parameters. If we possess a model for variability of the calibration parameters from experiment to experiment, then we can incorporate this into the bounding models to increase the bounding width. Unfortunately, the typical approach assumes that model predictions, using the calibrated parameters, represent the mean behavior. In-fact, our estimated parameters are likely to be biased from their true means. Unless we have multiple, independent realization of the experiments, the assumption of mean behavior is very weak.

6.4 Further Comments on Extrapolation and Calibration

In presenting the results of Chapters 4 and 5, we glossed over some important issues of model extrapolation. These are discussed below:

1. We considered only extrapolation of a single variable (front location) over time. We did not look at multi-variable extrapolation. Because of this, our prediction intervals were one-dimensional (i.e., represented as upper and lower intervals). Multiple variable extrapolation will possess multidimensional prediction regions rather than the simple upper and lower intervals illustrated here. While the methodology presented can be extended to such cases, it is more difficult to visualize this extension graphically.

2. We considered only the extrapolation of a variable to itself at later time. We did not look at the effect of extrapolation to another variable (e.g., from front location to maximum temperature). The methodology that was based on sensitivity analysis or on Monte Carlo analysis can be extended to these cases since we are using the assumed form of the underlying model to related one variable to another. While this assumed form of the model may be in error, it does contain more of the modeled physics that the simple linear correction method shown at the beginning of Chapter 5.

3. Our use of calibration of the approximate physics-based model (i.e., in contrast to the linear correction term), coupled with our model of uncertainty of the model parameters, provided estimates of prediction intervals on the bounding models, assuming the calibrated parameters are appropriate for the extrapolation. These estimates did not incorporate possible prior knowledge of the mean values of the parameters. We assumed that the prediction intervals were symmetric about the estimated parameters rather than the true means for the parameters. If we have prior knowledge of the parameter distributions, this assumption can be relaxed by using a maximum likelihood approach, such as was discussed by Hills and Trucano (2002).

4. The prediction intervals developed for the calibration example of Section 5.2 did not include the effect of test unit to test unit parameter variability. The intervals included only the effects of estimation uncertainty, based on the uncertainties in the measurements for that one experiment. The inclusion of both types of uncertainty will broaden the prediction intervals further.

5. The prediction intervals developed for the calibration examples of Sections 5.3 and 5.5 did not include the effect of estimation uncertainty in the parameters. The intervals did include the effects of test unit to test unit parameter variability. The inclusion of parameter estimation uncertainty will broaden the prediction intervals

further.

Finally, we wish to reemphasize several points discussed in previous chapters related to the calibration of models to be used for extrapolation.

1. We suggest that the model validation and model calibration can coexist if and only if we have validation data that is independent of that used for the calibration. Examples of independent data include prior knowledge of the model parameter distributions, or additional independent experimental data.

2. The acknowledgement that a predictive (i.e. to be used for extrapolation) model is approximate, in the sense that it is not consistent with the data within measurement and model parameter uncertainty, requires that we not only bound the model predictions, but that we also test or validate these bounds. The simple calibration of models to differences between model prediction and experimental observation does not provide an independent test of the ability of bounding models to predict bounded behavior.

3. We demonstrated the effect of independent realizations of the data in the previous chapters. As was clearly shown, the calibration of a model or model bounds, to a single realization of the data does not ensure that the results will be valid for another realization of the data (or of the experiment).

4. If an extrapolated model is sensitive to the values of its parameters and we have little confidence that the parameter set for the validation experiment applies to the target application, then we must either have 1) independent knowledge of the distributions (or bounds) on the parameters, or 2) sufficient independent realizations of the validation experiments which adequately represent the full uncertainty in the parameters, to perform extrapolation. We should have little confidence in our ability to extrapolate if we have little confidence in the appropriate range of values (or distributions) of the parameters under the extrapolated conditions.

6.5 Summary

The work presented here represents an overview of issues related to model calibration, extrapolation, and validation. While we do not claim that this work provides a comprehensive or fully rigorous description of the issues addressed, we suggest that this work does provide some insight into the issues that calibration and extrapolation bring to model validation.

7.0 References

Bernillon P., and F. Y. Bois (2000), “Statistical Issues in Toxicokinetic Modeling: A Bayesian Perspective,” Environmental Health Perspectives, V. 108(#5), October.

Brownlee, K. A. (1965), “Statistical Theory and Methodology in Science and Engineering,” John Wiley & Sons, New York.

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

Easterling, R. C. (2003), “Statistical Foundations for Model Validation: Two Papers,” SAND2003-0287, Sandia National Laboratories, Albuquerque.

Eldred, M. S, A. A. Giunta, B. G. van Bloemen Waanders, S. F. Wojtkiewicz, W. E. Hart, and M. P. Alleva (2002), “DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis, Version 3.0 Users Manual,” SAND2001-3796, Sandia National Laboratories, Albuquerque.

Gartling, D. K., R. E. Hogan, and M. W. Glass (1994), "Coyote - A Finite Element Computer Program for Nonlinear Heat Conduction Problems Part I – Theoretical Background," Report number SAND94-1173, Sandia National Laboratories, Albuquerque.

Hills, R. G., K. A. Fisher, M. R. Kirkland, and P. J. Wierenga (1994), “Application of Flux-Corrected Transport to the Las Cruces Trench Site,” Water Resources Research, Vol. 30, No. 8, pp. 2377-2385.

Hills, R. G. and T. G. Trucano (1999), **“**Statistical Validation of Engineering and Scientific Models: Background,” SAND99-1256, Sandia National Laboratories, Albuquerque.

Hills, R. G. and T. G. Trucano (2001), “Statistical Validation of Engineering and Scientific Models with Application to CTH,” SAND2001-0312, Sandia National Laboratories, Albuquerque.

Hills, R. G. and T. G. Trucano (2002), “Statistical Validation of Engineering and Scientific Models: A Maximum Likelihood Based Metric,” SAND2001-1783, Sandia National Laboratories, Albuquerque.

Hills, R. G., and I. H. Leslie (2003), "Statistical Validation of Engineering and Scientific Models: Validation Experiments to Application," SAND2003-0706, Sandia National Laboratories, Albuquerque.

Hills, R. G., I. H. Leslie, and K. Dowding, (2004), "Statistical Validation of Engineering and Scientific Models: Application to the Abnormal Environment," SAND2004-1029, Sandia National Laboratories, Albuquerque.

Hobbs, M. L., K. L. Erickson, T. Y. Chu, T. T. Borek, K. R. Tompson, K. J., Dowding, D. Clayton, and T. H. Fletcher (2003), "CPUF - A Chemical-Structure-Based Polyurethane Foam Decomposition and Foam Response Model," SAND2003-2282, Sandia National Laboratories, Albuquerque.

IMSL (1997), "IMSL, Math Library, Vols. 1 and 2," Visual Numerics, Inc., Houston.

ISI (2003), SciSearch Plus, www.isinet.com, Philadelphia.

Kalberlah, R., U. Fost, and K. Schneider (2002), "Time Extrapolation and Interspecies Extrapolation for Locally Acting Substances in Case of Limited Toxicological Data," Annals of Occupational Hygiene, V. 46(#2), pp. 175-185, March.

Landis, W. G. (2002), "Uncertainty in the Extrapolation from Individual Effects to Impacts upon Landscapes," Human and Ecological Risk Assessment, V. 8(#1), pp. 193-204, January.

Lockwood, J. R., M. J. Schervis, P. Gurain, and M. J. Small (2001), "Characterization of Arsenic Occurrence in Source Waters of U.S. Community Water Systems," Journal of the American Statistical Association, V. 96(#456), pp. 1184, December.

Miller, R. G. Jr. (1980), Simultaneous Statistical Inference, Springer-Verlag, New York.

Munns, W.R. (2002), "Axes of Extrapolation in Risk Assessment," Human and Ecological Risk Assessment, V. 8(#1), pp. 19-29, January.

Oberkampf, W. L., and T. G. Trucano (2000), "Validation Methodology in Computational Fluid Dynamics," AIAA 2000-2549, presented at Fluids 2000, Denver, June.

Oberkampf, W. L., T. G. Trucano, and C. Hirsch (2003), "Verification, Validation, and Predictive Capability in Computational Engineering and Physics," SAND2003-3769, Sandia National Laboratories, Albuquerque.

Robinson, S. (1999), "Simulation Verification, Validation and Confidence: A Tutorial," Transactions of the Society for Computer Simulation International, V. 16(#2), pp. 63-69, June.

Roe, P. L. (1985), "Some Contributions to the Modelling of Discontinuous Flows, Lectures in Applied Mathematics," Vol. 22, pp. 163-193.

Roe, P. L. (1986), "Characteristic-Based Schemes for the Euler Equations," Annual Review of Fluid Mechanics, Vol. 18, pp. 337-365.

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

Sweby, P. K. (1984), "High Resolution Schemes using Flux Limiters for Hyperbolic Conservation Laws," SIAM Journal on Numerical Analysis, Vol. 21, No. 5, pp. 995-1011.

Trucano, T. G., R. G. Easterling, K. J. Dowding, T. L. Paez, A. Urbina, V. J. Romero, B. M. Rutherford, and R. G. Hills (2001), "Description of the Sandia Validation Metrics Project," SAND2001-1339, Sandia National Laboratories, Albuquerque.

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

Vermeire T, T. Jager, G. Janssen, P. Bos, and M. Pieters (2001), "A Probabilistic Human Health Risk Assessment for Environmental Exposure to Dibutylphthalate," Human and Ecological Risk Assessment, V. 7(#6), pp. 1663-1679, November.

Warren-Hicks, W., J. P. Carbone, and P. L. Havens (2002), "Using Monte Carlo Techniques to Judge Model Prediction Accuracy: Validation of the Pesticide Root Zone Model 3.12," Environmental Toxicology and Chemistry, V. 21(#8), pp. 1570-1577.

Yang, H. Q. and A. J. Przekwas (1992), "A Comparative Study of Advanced Shock-Capturing Schemes Applied to Burgers' Equation," Journal of Computational Physics, Vol. 102, pp. 139-159.

External Distribution

Ivo Babuska
TICAM
Mail Code C0200
University of Texas at Austin
Austin, TX 78712-1085

James Berger
Inst. of Statistics and Decision Science
Duke University
Box 90251
Durham, NC 27708-0251

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

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

Department of Energy (5)
Attn: Kevin Greenaugh, NA-115
D. Kusnezov, NA-114
Jamileh Soudah, NA-114
K. Sturgess, NA-115
J. Van Fleet, NA-113
Forrestal Building
1000 Independence Ave., SW
Washington, DC 20585

Robert G. Easterling
51 Avenida Del Sol
Cedar Crest, NM 87008

Ashley Emery
Dept. of Mechanical Engineering
Box 352600
University of Washingtion
Seattle, WA 98195-2600

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

George Hazelrigg
Division of Design, Manufacturing
& Innovation
Room 508N
4201 Wilson Blvd.
Arlington, VA 22230

Richard Hills (10)
Mechanical Engineering Dept.
New Mexico State University
P. O. Box 30001/Dept. 3450
Las Cruces, NM 88003-8001

George Karniadakis (2)
Division of Applied Mathematics
Brown University
192 George St., Box F
Providence, RI 02912

Sankaran Mahadevan (2)
Dept. of Civil &
Environmental Engineering
Vanderbilt University
Box 6077, Station B
Nashville, TN 37235

Tinsley Oden (2)
TICAM
Mail Code C0200
University of Texas at Austin
Austin, TX 78712-1085

Patrick J. Roache
1215 Apache Drive

Socorro, NM 87801

Tim Ross (2)
Dept. of Civil Engineering
University of New Mexico
Albuquerque, NM 87131

Len Schwer
Schwer Engineering & Consulting
6122 Aaron Court
Windsor, CA 95492

Fred Stern
Professor Mechanical Engineering
Iowa Institute of Hydraulic Research
The University of Iowa
Iowa City Iowa 52242

D. E. Stevenson (2)
Computer Science Department
Clemson University
442 Edwards Hall, Box 341906
Clemson, SC 29631-1906

Department of Energy Laboratories

Los Alamos National Laboratory (10)
Mail Station 5000
P.O. Box 1663
Los Alamos, NM 87545
Attn: Mark C. Anderson, MS T080
Cuauhtemoc Aviles-Ramos, MS P946
Katherine Campbell, MS F600
Scott Doebling, MS P946
Ken Hanson, MS B250
R. Henninger, MS D413
S. Keller-McNulty, MS F600
David Higdon, MS F600
Mike McKay, MS F600
James Peery, MS F652

Lawrence Livermore National Laboratory (6)
7000 East Ave.
P.O. Box 808
Livermore, CA 94550
Attn: Evi Dube, MS L-095
Henry Hsieh, MS L-229
Richard Klein, MS L-023
Roger Logan, MS L-125
Cynthia Nitta, MS L-096
Charles Tong, MS L-560

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

Sandia Internal Distribution

1	MS 9014	8242	A. R. Ortega
1	MS 9042	8752	C. D. Moen
1	MS 9217	8962	P. D. Hough
1	MS 9217	8962	M. L. Martinez-Canales
1	MS 0384	9100	A. Ratzal
1	MS 1393	9100	T.Y. Chu
1	MS 0825	9100	C. W. Peterson
1	MS 0826	9100	D. K. Gartling
1	MS 0825	9110	W. Hermina
1	MS 0834	9112	S. J. Beresh
1	MS 1310	9113	S. N. Kempka
1	MS 0834	9114	J. E. Johannes
1	MS 0825	9115	M. Barone
1	MS 0836	9116	E. S. Hertel
1	MS 0836	9116	D. Dobranich
1	MS 0836	9116	R. E. Hogan
1	MS 0836	9117	R. O. Griffith
1	MS 0139	9120	P. J. Wilson
1	MS 0847	9124	R. V. Field
1	MS 0821	9132	L. A. Gritzo
1	MS 1135	9132	J. T. Nakos
1	MS 0836	9132	C. Romero
1	MS 1135	9132	S. R. Tieszen
3	MS 0828	9133	M. Pilch
1	MS 0828	9133	A. R. Black
5	MS 0828	9133	K. J. Dowding
1	MS 0828	9133	A. A. Giunta
1	MS 0779	9133	J. C. Helton
5	MS 0828	9133	W. L. Oberkampf
1	MS 0557	9133	T. L. Paez
1	MS 0828	9133	J. R. Red-Horse
1	MS 0828	9133	V. J. Romero
1	MS 0828	9133	M. P. Sherman
1	MS 0828	9133	A. Urbina
1	MS 0847	9133	W. R. Witkowski
1	MS 0370	9211	S. A. Mitchell
1	MS 0370	9211	M. S. Eldred
1	MS 0370	9211	L. P. Swiler
5	MS 0370	9211	T. G. Trucano
1	MS 1111	9215	R. Heaphy
1	MS 0829	12323	J. M. Sjulin
1	MS 0829	12323	B. M. Rutherford
1	MS 9018	8945-1	Central Technical Files
2	MS 0899	9616	Technical Library