

Statistical Foundations for the Validation of Computer Models

V&V Foundations Workshop

October 22, 2002

Robert G. Easterling¹
consulting statistician

James O. Berger²
working statistician

¹work supported by Sandia National Laboratories, Albuquerque, NM

²work supported by GM and NSF

Contents

Statistical Framework

Issues

experimental design

data analysis

inference

Case Study -- RGE

Introduction

- **Computational Predictions -- Inquiring minds want to know:**
 - *How well does the computer model represent reality?*
 - *How well can the computer model predict reality under untried conditions?*
- **Answers from: "MODEL VALIDATION"**
 - comparisons of computations to data
 - » *(field or experimental)*

Statistical View of Model Validation

- Essence of Model Validation
 - Comparison of computations to data
- This implies
 - design of experiments - to generate the right data
 - data analysis - to extract and communicate the information contained in the data
- *Thus, model-validation is fundamentally statistical*
 - (that's why we're here)

Some Additional Motivation

National Academy of Sciences report on statistics, testing, and defense acquisition, [Cohen et al. 1998]:

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

Goal:

- Our goal is to be able to characterize a model's 'predictive capability' with statements like --
 - Our understanding of the underlying science, our ability to translate that understanding to a computational model, and an analysis of a robust set of experiments and corresponding calculations indicate that actual system performance is quite likely to be within $P\%$ of the computational prediction for the application of interest.
 - » *Then, e.g., if the computational prediction plus $P\%$ is less than the failure threshold, we can "confidently" declare that the system meets its requirement.*

Goal, cont.

- Use the results of a suite of model-validation experiments and computations to evaluate a computational model's "predictive capability"
(“the degree to which a model represents the real world”)
- **Constraint: The experimental region may not be the same as the application region, for which predictions are the objective.**
 - *Example. lab expts. on mock-ups vs. real device in field*
- **Evaluation should be:**
 - **credible, defensible, communicable, ...**
 - » (“Don't give me no statistics, Meathead. I want facts!” Archie Bunker)
- **How do we (hope to) achieve the goal?**
 - **process: following slides**

Mathematical Set-up:

Let x be a vector that defines an event of interest, often a system and the environment to which it is subjected; e.g.,

- **experiment**

- » *x : hit an instrumented missile nose cone with a 500lb. hammer*

- **application**

- » *x : subject a missile to hostile in-flight environment*

- Let y be event outcome

- *e.g., stress on key missile parts*

Mathematical Set-up, cont.

Computational Model:

- $y^M(x) = M(x:\varphi)$, where
- x = event-defining variables
- φ = model parameters (constants in the equations within M):
 - » *e.g., material properties of nose cone and hammer, damping coeffs., ...*

Notes.

x , y^M , and φ are all possibly vectors or fields.

Focus on deterministic M , but for stochastic M , y^M could be vector of realizations from a probability distribution

Computational parameters (e.g., grid size, convergence criteria) are included in the specification of M .

Assume that M has been 'verified.' It is deemed *validation-ready*

Statistical Set-up:

- Conduct experiment at x
- Experimental outcome: $y(x) = y(x,w)$,
 - where w = unmodeled variables that influence nature's outcome
 - statistical model: w varies randomly across expts.
 - » *w has unknown probability distribution*
 - $y(x)$ is a "realization" of the random variable, $y(x,w)$
 - » *("true" outcome, not measured -- see next slide)*

Prediction Error

- “Prediction Error” at x :

$$e_x = y(x) - y^M(x) \quad (\text{nature} - \text{model})$$

- Contributors to (random variable) e_x :
 - random effects, w , in nature, not in M
 - » Example: M is 2-D model; nature is 3-D
 - systematic differences between nature and M
 - » Example: Model is linear in x ; Nature is not

Statistics, cont.

- Measurement Error
 - $y(x)$, the "true" experimental outcome is, in general, not observable.
 - Observed experimental result:
$$y^E(x) = y(x) + \delta_x,$$

where δ_x is measurement error, a random variable with an unknown probability distribution that may depend on x

» *separate "gage studies" or trustworthy instrumentation manufacturers provide estimate of distribution of δ_x at selected x -points or regions*

- **THUS ...**

Statistical Framework, Bottom Line

- The resulting statistical relationship between $y^E(x)$ and $y^M(x)$ is:

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

Data = Signal + Noise

where e_x and δ_x are random variables with unknown dist'ns. that, in general, depend on x

- The Task (should you choose to accept it) is to conduct a suite of experiments and computations that provide for a credible, defensible, communicable, ... characterization of the probability distribution of e_x (for pertinent x -values or x -regions)

Notes

- The addition in the statistical model is conceptual, not necessarily arithmetic.
 - transformations of x and y can enhance additivity
- It is convenient, but not essential, to pair the computational and experimental results on x . For unpaired results, you could fit separate “response surfaces,” then compare the fitted response surfaces at x 's of interest.

Evaluating Predictive Capability: Process

- Experimental Design: select a set of x 's at which to conduct experiments - $\{x\}$
 - Note: ultimate objective is to get good idea of e_x in application environment; nature of exp't. important
- Run the computational model to predict the outcomes of these experiments - $\{y^M\}$
 - objective is "pure" prediction, but boundary conditions from experiment may be required as input to the computation
- Run the experiments: $\{y^E\}$
- Analyze the data: $\{x, y^E, y^M\}$ in order to:
 - estimate the distribution of e_x in the experimental region
 - predict the distribution of e_x in the application region

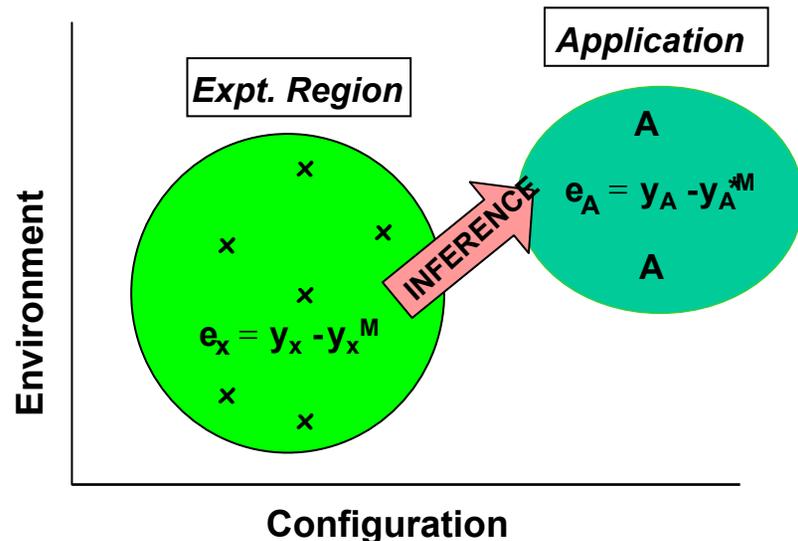
Graphically, ...

1. Experimental Design

Design and conduct a set of experiments and corresponding calculations (the x-points in Test Region, which is defined by two meta-variables: configuration and environment)

2. Data Analysis

Evaluate predictive capability $\{e_x\}$ for the experiments in the Test Region



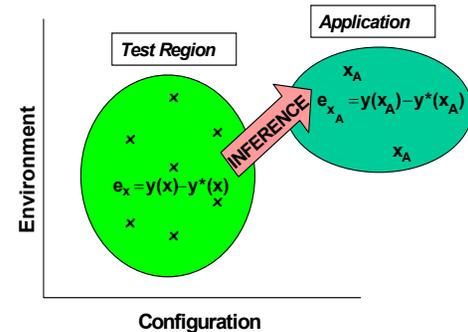
3. Inference

Estimate predictive capability at untested situations (A-points in the Application region)

Issues -- Experimental Design

Suite of Experiments -- Objectives

- **building blocks for inference to application**
 - » *single phenomenon -- multi-phenomena*
 - walk before you run
 - synthesis
 - » *application- and model-driven*
 - esp. w.r.t. environments
- **adequately characterize distribution of e_x**
 - » *give application-like w 's a chance to act*



Statistical Design Considerations:

- **distribution of experiments**
 - *explore the X -space efficiently*
- **number of experiments**
 - *precision: choose n to estimate σ_x within $P\%$*
 - *power: choose n to have $Q\%$ chance of detecting bias of Δ at $x = x_0$*
 - *some replication*

Experimental Design -- Short Takes

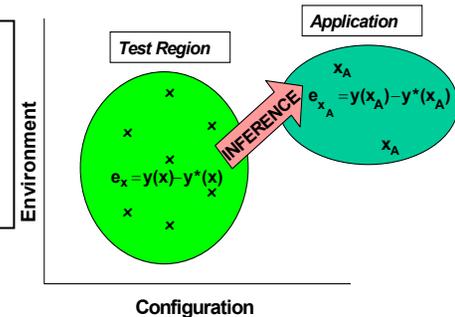
- **Model-Experiment Compatibility**
 - x 's in model have to be measurable/controllable in experiment
- **The need for simplification**
 - reduce dimensionality of experimentally manipulated x 's
 - » *use computational model to identify (apparently) important x 's, x -regions*
- **The battle of the x 's and w 's**
 - **Modeler:** if we put enough x 's into model (i.e., convert w 's to x 's), we can drive prediction error to zero
 - **Experimenter:** if you put too many x 's into model, I can't do enough experiments to "validate" zero-error

Experimental Design -- Research

- Design of Suites of Experiments?
 - single phenomenon -- multi-phenomena
 - resource-constrained
 - test capability-constrained
 - appropriately controlled x's, appropriately uncontrolled w's
- Simplification?
 - but not over-simplification
- Feasibility?
 - linkage of application-space to experiment-space

Technical Issues -- Data Analysis

- Data: $\{x_i, y^E(x_i), y^M(x_i) : i = 1, 2, \dots, n\}$
 - n experiments in the Test Region



Objective: characterize (estimate) the probability distribution of prediction error, e_x [$= y(x) - y^M(x)$]:
in the test region

Limited, variable data mean that any characterization has statistical 'uncertainty'

Use statistical concepts such as standard errors of estimates and confidence limits on parameters to convey the reliability of estimated characteristics of e_x

Issues -- Data Analysis

1. Choice of analysis variable ("predictand")

- e.g., y may be a time-history of temperatures at various locations
- analysis-variable possibilities
 - *complete temperature vs. time and location profile*
 - **NO: excessive, complex, resource-draining**
 - *"integral" measures, e.g.,*
 - max Temperature at critical location
 - ΔT over critical time period at critical location
 - critical $\Delta T/\Delta t$, ...
 - **YES, if focused on application and requirements**

Issues -- Data Analysis

2. Measures ("metrics") of predictive capability

- estimated characteristics of distribution of e_x , such as estimates of:
 - » *bias function, $\beta(x)$ ($=E\{y(x) - y^M(x)\}$)*
 - » *standard deviation function, $\sigma(x)$*
- **outcomes of tests of hypotheses such as**
 - » *$H_0: b(x) = 0$, for x in X_0*
- **Model-Validation is an estimation problem, not a hypothesis-testing problem.**
- **The analysis outcome is not binary -- e.g., "pass/fail"**

Statistical Data Analysis -- Basics

- Model data as observations from some family of probability distributions
- Use data to guide choice of statistical model and to evaluate goodness of modeling assumptions
- Develop data-based estimates of parameters (unknown constants) in the statistical model
- Characterize the precision of such estimates

Statistical Paradigm 1

- **Frequentist**
 - use “frequency” properties of statistical models to derive and evaluate estimates

Example.

{x,y} data

statistical model (simple linear regression):

$$y = \alpha + \beta x + e; \quad e \sim N(0, \sigma)$$

Analysis leads, e.g., to:

b = least squares est. of β

se(b) = standard error of b

Frequentist 'pivotal' relationship:

In repeated realizations of data from this model:

$(b-\beta)/se(b)$ has a t-distribution with n-2 degrees of freedom

Use this relationship to identify plausible β values

Statistical Paradigm 2

- Bayesian
 - Add further probabilistic assumptions that the statistical model's parameters (e.g., α , β , σ in lin. regression model) are realizations from known 'prior' probability distributions
 - Derive or approximate the parameters' 'posterior' distribution, given the data
- Two Varieties:
 - Objective Bayesian -- use innocuous priors and the Bayesian machinery to obtain or closely approximate frequentist results
 - Subjective Bayesian -- use 'informative' priors that connote (someone's) degree of belief in the parameters

Illustrations of Frequentist and Objective Bayesian statistical paradigms to follow

Dealing with Bias

- If there is an analysis finding of appreciable bias:
 - search for, then fix assignable causes in:
 - » *experiment*
 - instrumentation, controls, protocol, data reduction, ...
 - » *model*
 - ϕ values, equations, numerics, ...
 - » *an additional round of experiments and computations may be required to 'validate' fixes.*
 - **If unfixable (with available resources) bias remains:**
 - » *ignore bias (esp. if in conservative direction)*
 - » *do bias-corrected predictions*
 - » *scrap the model*

Adjustment for Measurement Error

- If, at x , e_x and δ_x are (statistically) independent, then

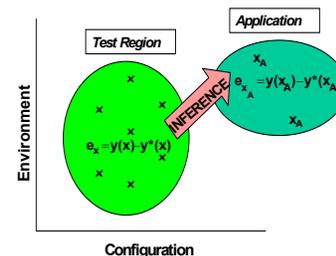
$$\begin{aligned} & \text{var}(\text{observed prediction errors}) \\ &= \text{var}[y^E(x) - y^M(x)] = \text{var}(e_x) + \text{var}(\delta_x) \end{aligned}$$

- Thus, data analysis provides estimate of this sum
- Gauge study or trustworthy mfg. data can provide estimated variance of measurement error: $\hat{\text{var}}(\delta_x)$
- Estimate $\text{var}(e_x)$ by subtraction

Issues -- Inference

Inference -- the BIG question:

Does the prediction-error structure have legs?



Extrapolation requires extending:

a. *modeled physics, via y^M*

theory-supported

b. *unmodeled physics - the prediction errors*

empirical, judgment-based

Conditional inference:

If the error structure we found in the test region (e.g., unbiased, Normally distributed, homogeneous log-prediction-error variance), holds in the application space, then the following prediction-error limits can be inferred: ...

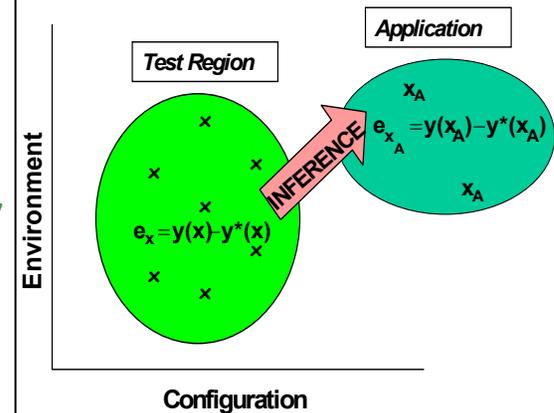
Special Inference Situations

- Inference in design/development:
 - application field data eventually become available
 - error data applicable to subsequent modeling cycles
- Some x 's are controlled in experiments, variable in application, over the same ranges (e.g., impact velocity)
 - no extrapolation required
 - 'propagation' methods discussed below
- Less ambitious objectives
 - e.g., Does the model get the sign of the relationship between y and x_1 right?

Issue: Success is Not Guaranteed

What if we cannot bridge the inference gap? Possible solutions:

- test in more application-like configurations and environments (science)
 - » *extreme example: resumption of underground nuclear testing*
- redesign system (engineering)
 - » *design out features that are most difficult to model*
- improve the comp. model (modeling)
- rework the requirements or scenarios (program mgt.)
- "softer" methods -- expert opinion:
 - E.g. *We never saw more than a 25% prediction error in the experiments we could do, but differences between those conditions and the application lead us to think that an additional factor of two would be prudent -- i.e., 50% prediction error limit. Trust us.*



Even if we fail on one loop, knowing why and what the obstacles are is useful in deciding what next to do.

Comments

- The possibility of not making the desired inferences should not deter us from doing a disciplined suite of model-validation experiments and computations
 - builds knowledge, confidence
 - if nothing else, de-bugs models vs. nature
- **Model-validation experimentation has important implications in re experimental capability**

Issue: Simplification

- It takes a substantial experimentation and analysis effort to build a meaningful “prediction-error map,” e.g., $\sigma_x^{\hat{}}$ vs. x .
- **Research Agenda: SHORTCUTS!**
 - **dimension reduction:** leave some of the x 's in w and capture their effects experimentally in e_x
*we design hardware for testability;
we need to design computer models for 'validatability'*
 - **simplified X-space:**
 - » *focus on subspace of interest; use code to help find interesting subspaces*
 - **simplified error maps**
 - » *e.g., envelope -- prediction errors are generally less than P%*
 - ...

Extension: Distribution Prediction

- Suppose x has an *assumed* probability distribution over some set of scenarios
- Problem is to predict resulting dist'n. of y

- Under the statistical model for y ,

$$y_x = y_x^M + e_x; \quad e_x \sim (\beta_x, \sigma_x),$$

by the law of total variance:

$$\text{var}_x(y_x) = \text{var}_x(y_x^M) + E_x(\sigma_x^2) \quad (\text{when } \beta_x = 0)$$

- In words:

nature's variance = model-based var. + extra-model var.

Comment

For this relationship:

$$\text{var}_x(y_x) = \text{var}_x(y_x^M) + E_x(\sigma_x^2) \quad (\text{when } \beta_x = 0)$$

- Stochastic propagation techniques - estimate the first right hand term
- Model-Validation experiments and analyses - estimate the second right hand term
- Many “uncertainty” analysts work the first term; ignore the second (*and claim they're evaluating prediction uncertainty!*), thereby underestimating variability, thereby overestimating reliability, ...
- Both terms are needed for distributional predictions

Analysis Issue: Putting it all together

- **Research Issue: How to combine prediction error data/models from different levels to infer prediction capability for application?**
- **One possibility:**
 - $y_A^M = M(y_1, y_2, \dots, y_k)$
 - $y_i^M = m_i(x_i : \varphi_i)$
 - $y_i = y_i^M + e_i$ (from predictive-capability expts. on m_i)
 - **Analysis: propagate estimated e_i distributions through M ; estimate resulting distribution of e_A and characterize precision of that estimate**
- **Example: Separate models for:**

$$y_1 = \text{stress}; y_2 = \text{strength}$$

Combined model:

$$y_A = \text{margin} = y_2 - y_1$$

Transition to Case Study

- Abstract concepts need to be made concrete via implementation of the statistical approach advocated above (*"facts, not statistics!"*)
- This case study (Easterling only):
 - polyurethane foam degradation in thermal environment
 - Sandia computational model and suite of experiments

Foam Degradation Case Study

Polyurethane foam -

1. requirement: structural support, normal environments
2. role (not requirement): insulates system components in accident-induced thermal environments

Applications:

system: systems in fires --
x = fuel source, temperature profile, orientation, duration, weather, system description, system-damage state, ...
component: x = system-fire induced thermal environment

Models :

CPUF (foam decomposition)

- Newly developed comp. model,
- better accounting for foam effects

Validation test program -- to date

simulated-component experiments - decomp., diffusion, radiation

Analysis:

Evaluate predictive capability

Foam Vaporization Experiments

Nature: Eight experiments in Sandia's Radiant Heat Facility:

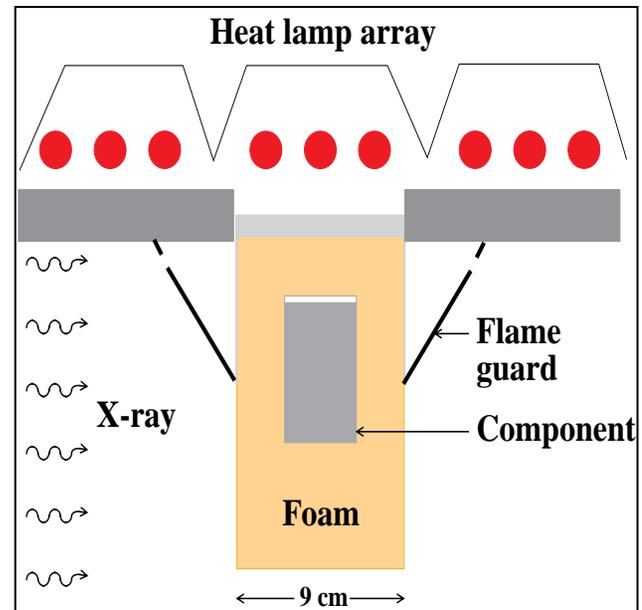
y^E = decomposition-front position vs. time, measured via x-ray imagery (unconnected dots, plot)

Model: $M(x:\varphi) = \text{CPUF}$, where

x = experimental factors, especially:
base plate temp. (600, 750, 900, 1000C
-- after 1.5 min. ramp)

φ = activation energies for foam decomposition, emissivity, ...
(obtained from other sources)

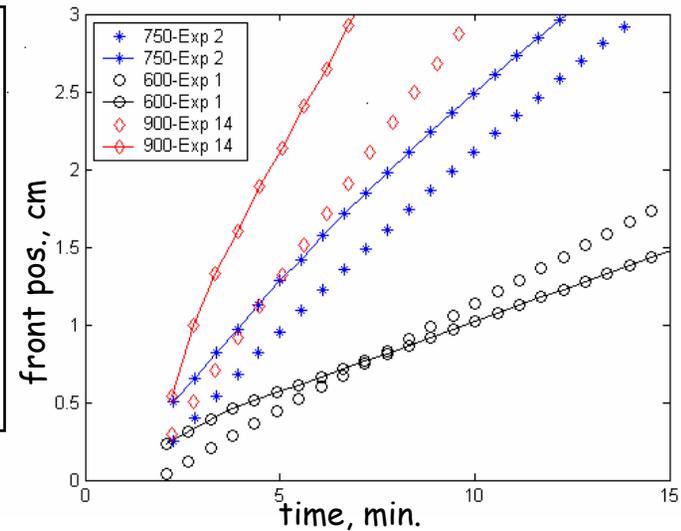
y^M = calculated decomposition-front position vs. time (connected dots)



Subset of Results

Plot shows:

computational predictions
(connected points) and
experimental results (unconnected)
for experiments at 600, 750, and
900C.



Analysis: focus on front velocity (slope of curves) between heat source and insulated component (1-2 cm)

Eyeball Analysis: Model is OK at 750C, over-predicts velocity at 900C, under-predicts at 600C. Issue:

"real" model error or "in the noise?"

The following analysis will substantiate the eyeball analysis

The Data

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

Table shows logarithmic error ($\ln[v^E/v^M]$, denoted *lne*) because preliminary analysis led to this transformation based on theoretical and potential variance-stabilizing properties

First five expts. are nominally the same; variability of v^M results reflects variability of measured boundary conditions

Analysis 1 --

Characterize Predictive Capability at 750C

Working assumption: the variability of the observed prediction log-errors among the 5 experiments at 750C is random extra-model variability (primarily specimen-to-specimen; measurement error info not available, assumed to be negligible for sake of illustration):

Analysis:

summary statistics, lne: ave = $-.05$ stdev = $.034$

evidence of bias?

$t = \text{ave}/(\text{stdev}/\sqrt{5}) = -3.40$, on 4df

$P(2\text{-tail}) = .03^*$

Fairly strong statistical evidence of bias; may not be practically significant, and bias is in conservative direction

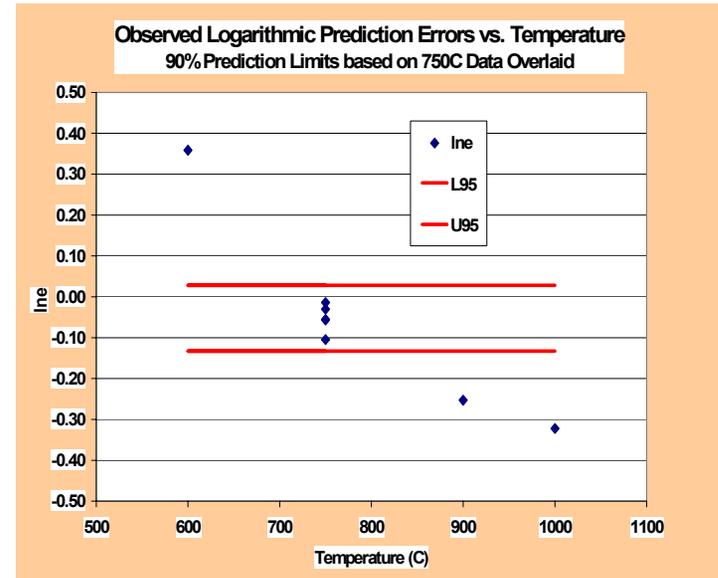
* $P = \text{Prob}(|t_4| > 3.40)$, based on Normal distribution assumption for lne

Analysis 2 -- Inference to Temp. Extremes

Emulate the inference process by extending 750C findings to 600C, 900C, 1000C

90% prediction interval for future log-error:

$$\begin{aligned} & \text{ave} \pm t_{.05}(4) * \text{stdev} * \sqrt{(1+1/n)} \\ & = -.05 \pm .080 \\ & = (-.13, .03) \end{aligned}$$



Inference, based on (leap-of-faith, judgment-based) assumption that the lne distribution is independent of temperature over the experimental range: to be consistent (@90% level) with 750C data, lne at temp. extremes should be in (-.13, .03)

Finding (see plot): Inference grossly in error!

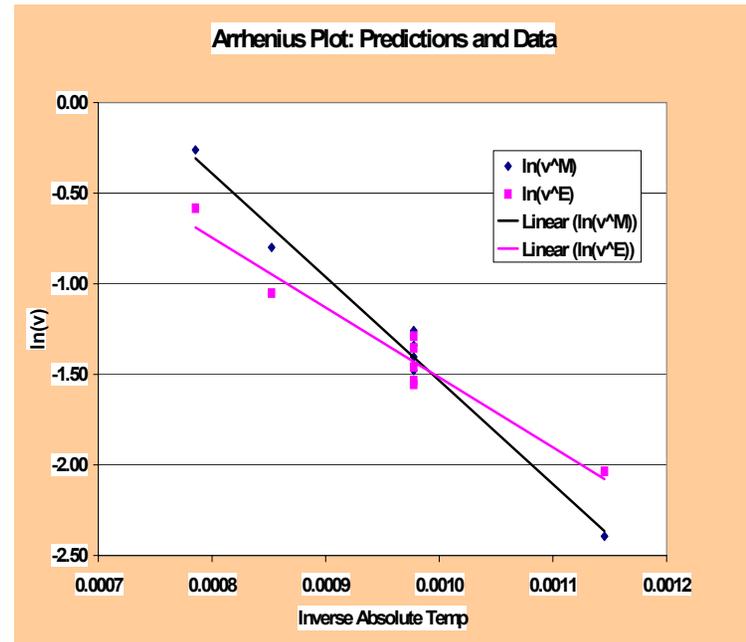
Analysis 3 -- Look at all the data

- Use subject-matter insight -
 - Arrhenius model:
$$v \propto \exp(E/\text{abs. Temp})$$

Chart: $\ln(v)$ vs. $1/(\text{abs. Temp})$

Both the model predictions and the experimental data exhibit fairly good linearity on these scales, BUT with different slopes.

Whatcha gonna do?



Analysis of Prediction Errors:

Linear model: $b(x)$ vs. x

Possible Further Actions

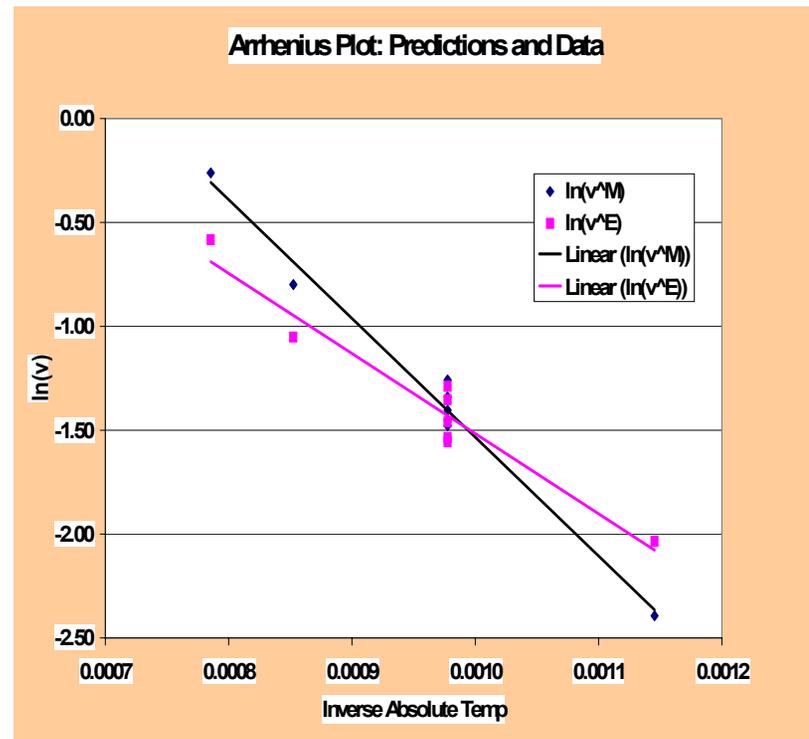
- Do bias-corrected prediction:
 - $y^M(x) + b(x) +/-$ prediction-error limits, where $b(x)$ is estimated bias-correction function
 - » *weak science - unappealing*
 - » *not feasible in predictions for applications*
- Fix/improve the model
 - modify parameter estimates (activation energies)
 - put more, or better, physics/chemistry into model
 - » *incorporate specimen-specific variables (w 's) into model*
 - *requires measuring those variables on each specimen*
 - » *expensive, time-consuming*
- Abandon theoretical model; use semi-empirical model (for limited purpose of predicting front velocity in experimental region)

"Pseudo-Fixing" CPUF

CPUF has been neither fixed nor its parameter estimates updated, but we can do analyses that illustrate the inferences that would result from these fixes.

1. (updated parameter estimates):

Use a linear model of $\ln(v)$ vs. $1/K$ (inv. target absolute temp.) as an approximate CPUF; use data to estimate slope and intercept



Analysis 4 --

Updated Parameter Estimates

• Regression analysis of $\ln(v^E)$ vs. $1/K$ provides “updated” estimates of slope and intercept

• Results*:

- Fitted line: $\ln(v) = 2.34 - 3858*(1/K)$
- Residual standard deviation: $s = .11$, on 6 df

• Statistical prediction intervals in regression situation (which can be obtained by standard methods) -- next slide

*Notes. Call this a semi-empirical model. Theory is the basis for assumed linearized Arrhenius relationship. Data support assumption of linearity and provide parameter estimates.

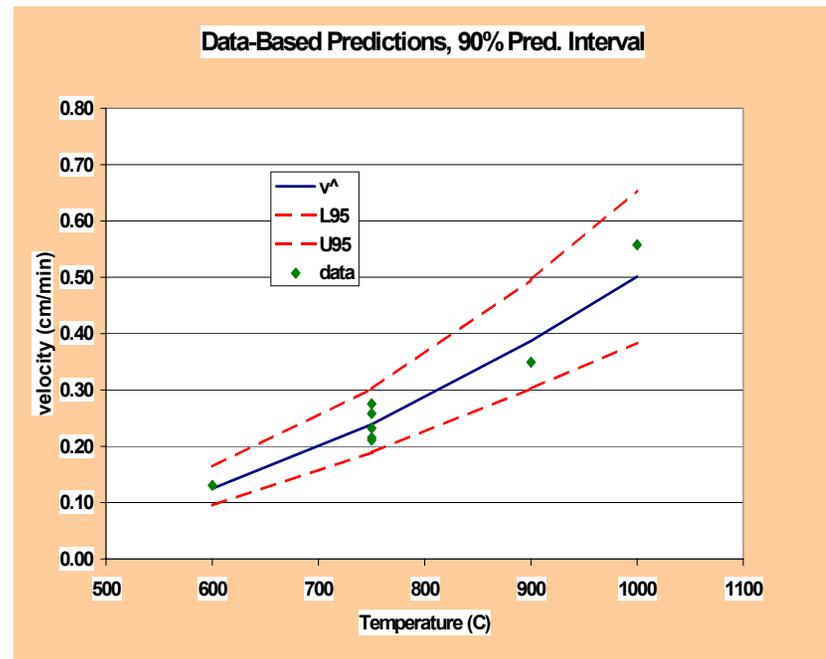
A better fit might be obtained if we used average measured base plate temp., rather than the target temp.

Statistical Prediction Intervals

Plot shows regression fit and 90% statistical prediction intervals*, on original velocity and temperature axes.

Prediction interval width is temperature-dependent: there are wider logarithmic error limits the greater the distance from the center of the data.

*assump.: log-errors Normally distributed, homogeneous variances



Interpretation: At a given temperature, with 90% confidence the measured velocity in a future experiment like these would fall within the indicated limits

Inference to 1500C? 2000C? Other geometries? ... ?

Requires foam-expert judgment

Analysis 5 -- Alternative Bias Correction

Regress $\ln v^E$ on $\ln v^M$:

Assumption (testable): prediction error depends on x only via v^M .
(Note. Great dimension reduction!)

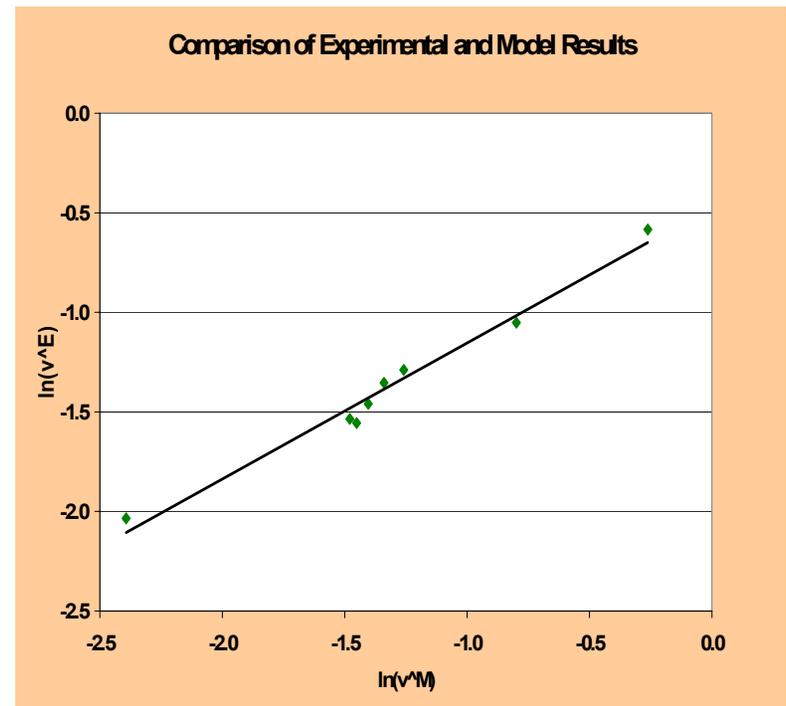
Results:

fit: $\ln v^E = -.47 + .68 \cdot \ln v^M$

resid. stdev. = .07, 6df

In contrast to Arrhenius-based linear approximation, this analysis makes direct use of the model.

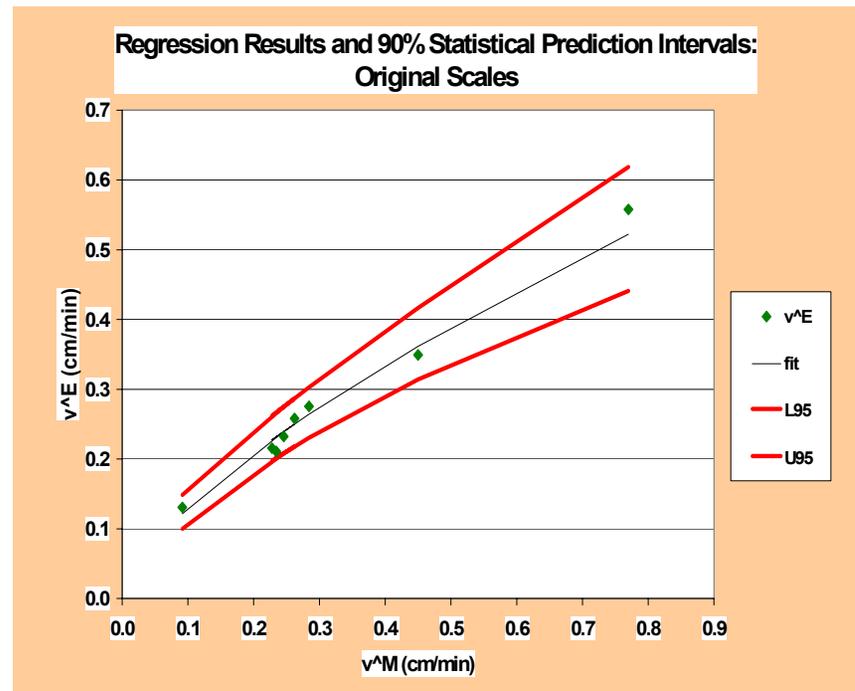
Bias-corrected prediction:
substitute calculated v^M into above linear model



Statistical Prediction Interval

Better fit and tighter prediction intervals than Approach 1 (slide 13) because:

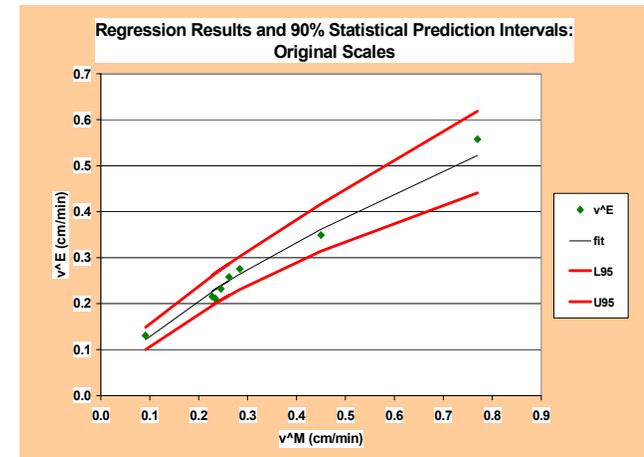
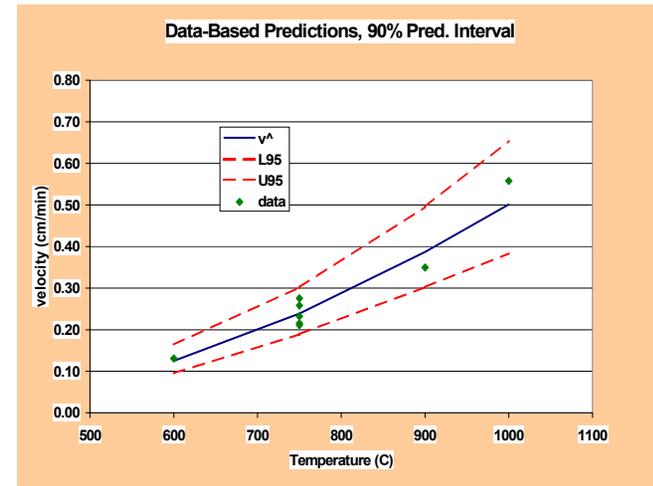
- variability of boundary conditions is accounted for and
- there is better log-linearity in the relationship.



Comment: A "fixed" CPUF (approach 2) might lead to similar results but with a near 45deg line relationship between v^E and v^M .

Issue: Model vs. Experiment

- The estimated predictive capability of a semi-empirical model constructed from the validation experiments can be comparable to the estimated predictive capability of a computational model plus error data obtained from the same validation experiments.
- A computational model should extrapolate better for nominal predictions, but *extrapolation of prediction errors is the same problem in both cases.*
- Kinda makes you wonder - is it all worthwhile?



Model vs. Experiment, cont.

- Fact of Life: Linearity won't necessarily hold in other situations for which we need predictions - e.g., confined, pressurized, ... environments
 - It's easier to do model calculations in new situations than to do physical experiments, ...
- **BUT we will still need experiments in such conditions to evaluate predictive capability in those situations, BECAUSE ...**
- We generally cannot assume that predictive capability will travel (from explored x-space to unexplored).
- **MESSAGE: *When you have a choice, think about it!***

Adjustment for Measurement Error

- There has been no informed evaluation of measurement error w.r.t. $\ln(v)$.
- Velocity, v , should be measured fairly precisely because major sources of error should cancel out
- Hypothetically, suppose v is measured with relative std dev of 2%. Then, std dev of $\ln(v)$ is .02.
- For the regression analysis, $\ln v^E$ vs. $\ln v^M$, the resulting adjusted standard deviation of prediction error would be:
 - $s_{\text{adj}} = \sqrt{(.07^2 - .02^2)} = .067$
 - negligible effect, in this hypothetical case

Case Study - Comments

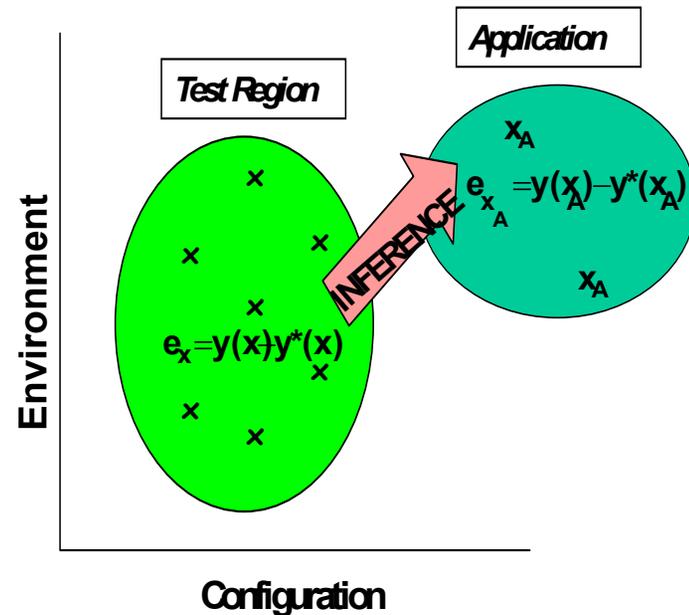
- Small no. of experiments, limited exploration of x-space: representative of what will be possible in high-level testing
 - A more efficient experimental design would have helped some
- Predictive capability is not too good
 - CPUF doesn't get the temperature effect right
 - » *Model's sensitivity to temperature is about 2x that of Nature*
- The data used to evaluate predictive capability can be used to construct a semi-empirical model directly

Case Study Conclusions - Methodology

Statistical analysis methodology for evaluating predictive-capability is:

- workable
- inexpensive
- illuminating
- communicable

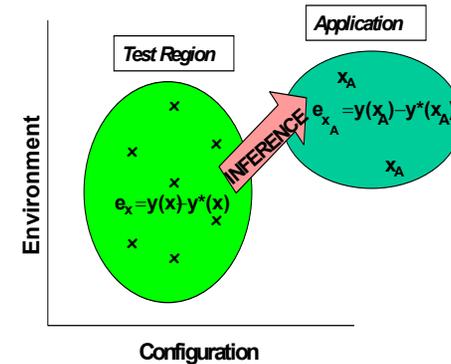
But the results may not be as broadly applicable or precise as we would like or need them to be



Summary

Measuring predictive-capability poses numerous difficult problems:

- scientific, experimental, statistical, organizational, management



Statistical ideas and methods can contribute to successful resolution of these problems, or clear understanding of why they cannot be solved

We all gotta work together

