A FRAMEWORK FOR MODEL VALIDATION

Robert G. Easterling
Sandia National Laboratories
Albuquerque, NM 87185

ABSTRACT

Computational models have the potential of being used to make credible predictions in place of physical testing in many contexts, but success and acceptance require a convincing model validation. In general, model validation is understood to be a comparison of model predictions to experimental data, but there appears to be no standard framework for conducting this comparison. This paper gives a statistical framework for the problem of model validation that is quite analogous to calibration, with the basic goal being to design and analyze a set of experiments to obtain information pertaining to the 'limits of error' that can be associated with model predictions. Implementation, though, in the context of complex, high-dimensional models, poses a considerable challenge for the development of appropriate statistical methods and for the interaction of statisticians with model developers and experimentalists. The proposed framework provides a vehicle for communication between modelers, experimentalists, and the analysts and decision-makers who must rely on model predictions.

INTRODUCTION

Mathematical models of phenomena, processes, products, and their performance, accompanied with high-performance computing, have the potential of reducing the amount, or changing the nature, of the physical testing required to design and produce complex components and systems, predict their performance in various environments, and certify their safety and reliability. This is the premise behind a tremendous amount of national and international research and code development and is manifested in DOE (Department of Energy) plans for science-based stockpile stewardship and such programs as ASCI (the Advanced Scientific Computing Initiative). Similarly, modeling and simulation of weapon systems and military operations are becoming increasingly important in DoD acquisitions and decision-making. The value of such models (existing or future) is realized when model calculations, at points in the parameter space that have not been or cannot be physically tested to provide confirmation, can be trusted for the purpose of drawing conclusions or making important decisions at these untested points. (For example, the DOE may be required to certify a new component's performance when subjected to radiation fields only achievable in underground nuclear tests, now precluded.) This trust comes when physical phenomena are well-understood and expressed mathematically, then accurately converted to code, and when empirical support is obtained at some points in the model's parameter space.

Developing this empirical support, the process of which is generally and generously termed 'model validation,' is well-recognized as critical to the success of modeling and simulation. Yet, there does not appear to be any consistent approach to model validation, no overarching guidelines or framework for linking model objectives with validation efforts. The expense and difficulty of testing, and the absence of clear validation objectives, can lead to ad hoc approaches to validation, which in turn can be unconvincing to potential users of the model. A more formal approach to validation is needed.

Model-based prediction is a combination of scientific and statistical inference, because it is based on both theory and the accompanying model-building or validation data. Determining the nature and amount of validation testing is a problem in statistical experimental design -- a systematic approach to determining a suite of tests that can efficiently estimate or predict characteristics of interest with predetermined precision. Hence, the problem of model validation is to a major extent a statistical problem. Statistics provides methods and approaches for stating the validation objectives and determining the data requirements for meeting those objectives. These methods have typically been applied to model validation problems only in fairly simple mathematical situations. Extending statistical methods to models of

1
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
complex interactions and geometries, captured, e.g., in finely meshed, finite element, massively parallelized codes is a major challenge.

Statistical methods also pertain to the analysis of data that result from model validation experiments and the corresponding model calculations. Model-based predictions are uncertain and the user or decision-maker who will rely on the predictions needs some idea of credible ‘limits of error’ to associate with a prediction. Thus, the objective pursued here is to design and analyze model validation experiments in ways that permit any model-based prediction to be accompanied by a credible ‘limit of error.’ This is an extension beyond typical model validation analyses which overlay computations with experimental results and invite the judgment: close enough, or not. The model validation situation is comparable to calibration, in which measurements from two methods of measuring physical phenomena are compared, the objective being to be able to use one method to predict the other and to characterize how good the predictions are.

MODEL VALIDATION IN THE NEWS

A 1996 news item pertaining to model validation stated something like:

DoD says comparison of computer simulations versus live-fire tests of the effect of gunfire on helicopter blades shows that even the most sophisticated computer models cannot accurately mirror real life. On a scale of 1 to 10, the models scored a 7 in predicting how the shell would penetrate the blade, a 3 in predicting the destruction of the blade, and a 2 in predicting the loss of a helicopter.

This brief item demonstrates several important facts about modeling and model validation. First, note the modeling chain: phenomenon, component, system. Accurate modeling gets more difficult as one moves up the chain from single phenomena to multiple, interacting phenomena. The physics of shell penetration can be modeled fairly accurately, but predicting whether that penetration will destroy the blade is more difficult and predicting whether the helicopter will crash as a result is more difficult still. This example also points out a problem in resource allocation. Where should future resources be spent? Improving the shell model to a 9? Improving the system model to a 4? And, how to decide? Within an overall budget, how should resources be divided between modeling and testing? What is the most cost-effective and scientifically sound way to develop and certify a survivable helicopter?

Modern advances in modeling and computing capabilities, plus pressure to reduce the amount of money spent on test facilities and testing, have led present decision-makers to lean toward the modeling and simulation side. The cost of developing these tools, particularly in situations in which important, real decisions are going to ride on them, is being found to be substantial, though. These situations also call for the most care in model validation testing and, as will be illustrated here, this, too, can be a substantial undertaking. There is a clear need for careful analyses of the evolution of testing and modeling and simulation lest we lose confidence in the decisions that have to be made on their combined basis as this evolution progresses.

This example uses scoring on a 1 to 10 scale as a means of measuring agreement of computer models and ‘real life.’ Doing so is a useful communication device, but it does point to the problem of developing definitive, informative measures of this agreement.

TERMINOLOGY

If ‘validation’ is interpreted as meaning ‘to establish the truth of,’ then it is clear that such a goal is unattainable. A model is, after all, an approximation. Even if it is found to closely approximate reality in selected circumstances, there is no guarantee that it will do so in all circumstances. This is particularly
true when the model has been calibrated, or ‘tuned.’ Oreskes et al.\(^1\) discuss these and related issues in depth and provide a healthy antidote to glib claims of validated models.

The most prominent definition in use today appears to be that adopted by the DoD’s Defense Modeling and Simulation Office:\(^2\):

Validation. The process of determining the degree to which a model or simulation is an accurate representation of the real-world from the perspective of the intended uses of the model or simulation.

Thus, ‘determining/estimating the degree of accuracy’ recognizes that the goal is not the establishment of agreement, but the measurement of disagreement between the model and the situation it is approximating. This definition leads naturally to the above-stated goal of being able to derive credible limits of error from validation testing and data analysis.

Model validation is sometimes set up as a test of the hypothesis: MODEL = NATURE. Then, based on the test data, the decision is made either to reject or accept this hypothesis. The alternative view here is that the basic problem is one of estimation: What is the magnitude of (MODEL – NATURE) within the set of situations in which it is desired to use the model to approximate nature? With that information, the user can decide whether the approximation is adequate for “the intended uses.”

**MATHEMATICAL SET-UP**

“All models are wrong, but some are useful\(^\text{3}\)” is a statement by George Box, University of Wisconsin, that succinctly captures the essence of the preceding section. This statement and the problem of model validation can be expressed mathematically, as follows:

**Model.**

\[
y^* = h(x), \text{ where} \\
h(x) \text{ is the (computer) model of the phenomenon of interest,} \\
x = \text{model input (in general, a vector),} \\
y^* = \text{model output, a prediction of a characteristic, } y \text{ (possibly a vector)}
\]

**Nature.**

\[
y = g(x, w), \text{ where} \\
w = \text{additional parameters or variables that influence nature’s outcome,} \\
g(x, w) = \text{nature’s function (generally unknown),} \\
y = \text{nature’s outcome, at } x, w.
\]

In words, the computer model, \(h(x)\), takes input \(x\) and calculates output \(y^*\). I assume here that the computer model is fully specified – grid size, convergence criteria, internal parameters, etc. – and that the vector \(x\) describes, in general, a physical entity and the environment to which it is subjected. The predicted outcome is \(y^*\). The vector \(x\) is an abstraction of nature, so nature’s variables at work in the simulated situation will generally involve variables beyond what the modeler has chosen. These are the \(w\)’s. The functional relationship is consequently different, and might be different even if there were no \(w\)’s. For example, nature may be nonlinear where linearity is assumed in \(h(x)\). Also, nature’s relationship might not even involve some \(x\)’s the modeler has chosen. Nature’s function, \(g(x,w)\) is unknown and some or all of the \(w\)’s are also unknown. An example, given below, may help clarify these ideas.

**ERROR**
Consider now the difference between model and nature:

\[
\text{error}(x,w) = y^* - y
\]

The degree-of-accuracy of the model is characterized by error(x,w), abbreviated as e(x,w) below, as x and w range over spaces, X and W. Trust in a model prediction comes from knowing how large e(x,w) might be, i.e., how far y* might be from nature's result, and knowing that the decision to be made is unaffected within this range. The only way to learn anything about e(x,w) is to run tests, or collect data, at selected x's, allowing nature to generate w's at will. There is no substitute.

EXAMPLE

Chambers et al. developed a model for the stresses and strains produced in a structure by the curing of an encapsulant. A comparison of model and experiment was obtained from instrumented tubes that were filled with a resin, then cured. The 2-D model assumed the tube was a perfect cylinder, the geometry of which is specified by three parameters: diameter, length, and wall thickness. These are x's in the model. Experimental results, generated as a check on the model, showed deviations from the predicted strains and these deviations were found to be due to tube out-of-roundness. Thus, actual tube dimensions, which could take a 3-D map of hundreds of variables to characterize, are nature's w's in this example, and they produced noticeable prediction errors. Whether such errors would preclude use of the model for certain situations would be a matter to investigate.

Enhanced computing capabilities make it possible to develop a 3-D model. A heuristic that helps describe this situation is that advanced code development would move some of the w's into the x's for a more detailed computer model. Validation testing would be used to check to see if the remaining w's, and modeling error, lead still to appreciable prediction errors. It is possible that the code might predict well for some types of out-of-roundness, poorly for others. To validate that the code adequately predicted tube strains for a wide variety of geometries might take a considerable amount of experimentation.

Alternatively, one could treat the 2-D model as an expected value model -- the input parameters would be the average tube diameter and wall thickness -- and test a random sample of instrumented tubes to provide a statistical characterization of deviations from expected strain. One would not have to take detailed dimensional measurements on the tubes -- only assure that they were representative of the tubes for which predictions are desired.

Either of the preceding options is appropriate if the objective of the model and experimentation is the prediction of the range of strains that might be produced in encapsulated tubes. Alternatively, if one were just interested in choosing the best curing temperature, comparative calculations using the original 2-D model might provide a valid basis for making that process decision, even though the model doesn’t capture the effect of tube out-of-roundness, an effect that might largely cancel out in a comparison of curing temperatures. Establishing such validity, though, brings us back to the basic problem of model validation with respect to predicting differences. If validity of predicted differences is established, a further inferential question to consider, though, is whether one can also infer that the selected temperature is also (near) optimum for other geometries, such as a more complex shape. Other refinements could also be pursued. For example, resin is not necessarily homogeneous within a tube and its properties can vary from lot to lot and pour to pour. The effect of these variabilities can either be incorporated into the model, or estimated experimentally.

This example points out an important conundrum that will have to be resolved. The modeler believes that by putting more physics into the model, which generally means putting more x’s into the model, model predictions will improve and the error can be driven to negligible. This is science. The experimentalist, however, recognizes that the more parameters are included in the model, the more difficult it will be to carry out validation tests that adequately explore the parameter space and characterize prediction ability. The potential user wants some of both -- adequately sophisticated models adequately supported by data.
The sponsor cannot afford unconstrained modeling or testing. The proposed mathematical framework offers at least the start of a context within which this battle of the x’s and w’s can be debated and resolved.

STATISTICAL CHARACTERIZATION OF ERROR

Because not all of nature’s w’s are known, or measured, in a validation experiment, it is not possible to observe e(x,w) directly. For any situation described by x, nature may generate a variety of outcomes because of variability in the w’s that also come into play in a situation the modeler describes by x. This situation will be described statistically: the random variable, e_x, has a probability distribution indexed by x. That is, at any point x, in the X-space, there is a distribution of possible errors – deviations between nature and y* at x. The nature of that distribution could, and is likely to, depend on x; predictions are apt to be more accurate in some parts of X than in others. An ideal situation, beloved by statisticians, would be for e_x to be normally distributed, with mean zero and constant standard deviation throughout X. In general, though, one can consider a mean function, μ(x), and a standard deviation function, σ(x), as characterizing error. The goal of validation testing is to shed some useful light on these two functions – we can’t hope to know them well over a high-dimensional, complex X-space, and we can’t hope to say anything definitive about the nature of the distribution of e_x.

VALIDATION EXPERIMENTS

EXPERIMENTAL CONDITIONS

Designing a validation experiment, in the framework of the preceding section, means to select some points in the X-space, then conduct tests at those points. Because our goal is inference about how nature would respond at those x-conditions, it is important that the experiment be conducted in such a way that nature’s variability in the w’s is given full rein. For example, if the tube-strain tests had been conducted with high-precision tubes, the results might have been better in terms of how well the computational results matched the test results, but we would not have a good estimate of how much variability in strain rates was incurred by standard issue tubes and therefore would not have obtained data on which to base credible limits of error for predictions about that population.

This statistical perspective is contrary to a common view of validation testing, which is that well-controlled, subscale lab experiments can be used to validate models that will be applied to full-scale systems under use conditions. While certain physical laws may scale, it is also necessary to know how residual, or error, variability will scale, if validation testing is done at a subscale level. The main point to make is that the nature of the experimentation, not just the points in X-space, needs to be addressed in planning model validation experiments.

EXPERIMENTAL DESIGN

Objectives. Now, consider the selection of test points in the X-space. The number and location of those points depends on objectives and resources. At one extreme, if validation is desired at a single x-point, and experimentation is possible at that point, then the obvious thing to do is to test at that point. The only statistical issue would be to decide how precisely the bias (mean error) and/or the sigma at that point need to be estimated and run the appropriate number of tests, resources permitting. Of course, in this situation, the model is somewhat superfluous – the test results themselves can be used directly to predict performance at the selected x-point. A slight expansion is the case in which subject-matter knowledge enables one to say that error variability exhibited at the selected x also applies in some specified region about x, so that error limits for subsequent model predictions in that region are subject to the same limits of error as determined at the test point. At the other extreme, validation is desired throughout a high-dimensional X-space, thus requiring experimental ‘coverage’ of that space, to some degree.
To illustrate the sort of objectives that might drive validation testing, consider a simple linear model: Theory, and the computational model built on that theory, says that a single response variable, y, is a linear function of a single input variable, x. Thus, \( y^* = \alpha + \beta x = h(x) \) is the computer model. Theory might provide the coefficients, \( \alpha \) and \( \beta \), or theory might say that they are functions of materials or environments, in which case experimentation would be required to estimate them, as well as ‘validate’ the linear model. Possible testing objectives in this situation include:

a. test the linearity assumption
b. estimate the model coefficients
c. test the agreement of nature with theoretical model coefficients
d. test that the slope is positive
e. estimate the error standard deviation, \( \sigma(x) \), as a function of x.
f. some subset or all of the above

**Statistical Power.** Because of nature’s variability – induced by the w’s not captured in the model – the sorts of questions embodied by the preceding list of objectives can be answered only with some uncertainty. That is, there is some chance that with limited data, chance variations could mislead us, for example, by indicating that the slope is positive when it really is negative. Statistical power is the method by which risks can be controlled. For example, we could design a test such that there is a 90% probability of detecting that \( y^* \) at a particular x-point differs from nature’s mean outcome at that point by more than 20%. Or, design a test such that there is only a 5% chance of concluding the slope is positive when in fact it is really negative by some specified amount. Solutions to these sorts of design problems generally depend on a prior estimate or assumption of error-variability. Preliminary experimentation may be required to provide this information.

**Uncertainty Analysis.** Computer models can contain physical constants, such as transfer coefficients and material properties, that are estimates from limited data and so are uncertain approximations of these constants. In the set-up here, these estimated constants are considered to be part of the model, \( h() \), not the model input, \( x \). That is, these constants have to be specified to run the model, just as grids and convergence criteria have to be specified. A common analysis in this situation is to assume ‘uncertainty distributions’ that probabilistically represent the uncertainty of the estimated constants, then, by Monte Carlo or other methods, propagate these distributions through \( h(x) \) to generate an uncertainty distribution of \( y^* \). There is a temptation to interpret this uncertainty distribution in terms of how well \( y^* \) predicts nature, but such is not justified. The uncertainty distribution of \( y^* \) just characterizes how close the chosen model might be to the best model, the one with the correct physical constants in it. It does not, and cannot, capture the effects of nature’s w’s that are not even in the model and which contribute to prediction error.

**Conceptual Outcome of a Validation Experiment.** Table 1 illustrates the conceptual outcome of a suite of validation experiments. In general, \( n \) points in X-space are selected for testing and \( r_i \) tests are done at point \( i \). As discussed, the design issues are how many x points to select, where should they lie in the X-space, and how many replications of the experiment should be run at the selected x-points. Figure 1 shows the conceptual outcome for the simple case of a single x-variable, and tests conducted at two levels of this variable. This figure provides a glimpse into how one might model \( e(x) \) as a function of x and to quantify prediction errors at untested \( x \), within physically-supported reason. That is, errors appear to be centered on zero, but with a standard deviation that increases with increasing \( x \). Statistical analysis could characterize this pattern and also characterize the uncertainty associated with estimates based on such limited data.

<p>| Table 1. Conceptual Model-Validation Experiment |
|-----------------|-----------------|-----------------|-----------------|
| ( x ) | Model | Experiment | Errors |
| ( x_1 ) | ( y^*<em>1 ) | ( y</em>{11}, y_{12}, \ldots, y_{1r_1} ) | ( e_{11}, e_{12}, \ldots, e_{1r_1} ) |</p>
<table>
<thead>
<tr>
<th>$x_2$</th>
<th>$y^*_2$</th>
<th>$y_{21}, y_{22}, \ldots, y_{2n}$</th>
<th>$c_{21}, c_{22}, \ldots, c_{2n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$x_m$</td>
<td>$y^*_m$</td>
<td>$y_{m1}, y_{m2}, \ldots, y_{mn}$</td>
<td>$c_{m1}, c_{m2}, \ldots, c_{mn}$</td>
</tr>
</tbody>
</table>
To conduct validation experiments, as just described, it is necessary to be able to control the x’s experimentally. For high-dimensional models of complex phenomena, such experimental control may not be possible. The model may contain x’s that cannot be controlled, or even measured in a test environment. As will be discussed, this poses problems to a validation test data-analysis and will likely lead to more uncertain predictions than would be achievable if the x’s can be controlled. There is a clear need for communication between the model developer and experimentalist. Building the most sophisticated model possible, then throwing it over the wall for the experimentalist to validate will not lead to a model with the best predictability.

Uncontrollable x’s. In many contexts, even with the best efforts of modeler and experimentalist, some x’s will be uncontrollable, either directly or by controlling other variables that determine particular x’s. Suppose the input vector, x, is separated into its controllable and uncontrollable variables, x_c and x_u. The experiment will then consist of selecting points in X_c and conducting the experiment so that nature can randomly deal x_U’s as well as the w’s. On the computational side, one could generate a set of y*’s at a given x_c by drawing Monte Carlo samples of x_U, from an assumed or estimated conditional distribution of x_U given x_c. Thus, rather than the point-to-point comparison of test results and computations, as illustrated in Table 1, the analysis will involve comparisons of collections of test results (samples) and computations at the selected x_c points. It is beyond the scope of this paper to work out analysis details by which one would obtain credible limits of error in this situation, but clearly there are more sources of variability in the data in this case than above, so broader error limits, than would be obtained if all the x’s are controllable, are the likely result.

Uncontrollable x’s are also a problem for the potential user of a model. Suppose one has a model of a production process and wants to use it to optimize the design of the process – determine the times, rates, and temperatures, e.g., at which the process will be run. If key variables in process control are uncontrollable x’s, the knowledge is of little use to the process engineer. In general, the model developer, experimentalist, and model user all need to be involved in developing meaningful, usable, validated models. While this may be know at least in the abstract, the framework developed here illustrates the importance of such communication.

Shortcuts. If one adopts it that a goal of model validation testing is the development of a statistical characterization of the error function, e(x), over a specified X-space, then it is clear that for high-
dimensional $X$, a considerable amount of testing would be required—on the same order of the amount of testing required to build an empirical model, $h(x)$. In general, this is grossly impractical and defeats the purpose of developing models that can be used in place of testing. Either this goal has to be abandoned, or scaled back. Some possible shortcuts are:

a. reduce dimensionality. Focus on a subset of the $x$'s and set the remaining $x$'s at bounding values. The selection of the subset could be based on a sensitivity analysis of $h(x)$, the underlying assumption being that the code is adequate for this purpose.

b. focus on a sub-space of interest. Use $h(x)$ to help find interesting subspaces.

c. do worst-case testing. Through subject-matter knowledge, identify points or regions in the $X$-space where the model/test disagreement is expected to be maximized.

d. test at the sub-model or single phenomenon level. This approach would be justified in cases in which the interactions and interfaces between phenomena are known not to be sources of prediction error.

CONCLUDING COMMENTS

There is no single validation problem to be worked, but rather a collection of problems. The focus here has been on point predictions, $y^*$, but there are contexts in which ‘integral’ predictions are of interest. For example, in the tube encapsulant example, one may be interested in predicting the extreme strains in some population of tubes whose geometry, as characterized by $x$, varies. Assumed probability distributions for $x$ could be propagated through $h(x)$ to predict a parameter such as the 99th percentile of strain, or the probability that strain exceeds a defined failure threshold. Testing of a sample of tubes would lead to estimates of the same parameters. Differences between these estimates could either reflect modeling error, such as the effect of $w$'s not included in the model, or erroneous assumed distributions of the $x$'s. The possibility of either compounding or offsetting errors is apparent.

It is one thing to write about model validation in the abstract; quite another to apply the concept. Just as a computer model needs to be tested against reality, theories of model validation need to be tested in real applications. It can be expected that such testing will lead to refinements and improvements, and an increased appreciation of the difficulty of the task. The nature of both modeling and validation testing may change as a result of this testing.

REFERENCES


