SAND2005-7678C

On Model Validation and Extrapolation for Best-Estimate-Plus-Uncertainty Predictions[1]

Vicente J. Romero[2]
Model Validation and Uncertainty Quantification Department
Sandia National Laboratories[3], Albuquerque, NM

The intent of this paper is to point out a connecting bridge between several disparate views of model validation, and to argue for an associated split that I believe should occur in the conceptualization and articulation of model validation. Such a split fairly recently occurred with model verification, where it was recognized that Code Verification is different from "Solution or Calculation" Verification. The former is "code-centric" and the latter is "application-centric". Both are legitimate types of verification, but are very different in nature and implementation. In the following I propose a similar split for model validation, into "model-centric" and "application-centric" types of validation. The discussion hinges on the expectation that the model is to be validated before use in an extrapolative prediction away from the validation point(s) in the parameter space. The cited references give deeper support to the brief presentation here.

I argue that there are at least three distinct aspects to model validation. These are:

1. model **accuracy** characterization — model results comparison against relevant experimental data to quantify the bias (and uncertainty thereof) of the model predictions relative to the data
2. model **adequacy** assessment — implies the further step of comparing the characterized model error or bias against some prescribed accuracy requirement for the purpose of accepting or rejecting the model for the intended application
3. **augmentation of the original model** with a representation of the *resolution uncertainty* of the validation experiment(s).

Most people active in model verification and validation are familiar with the first 2 items. The third item is discussed in detail below. With regard to Item 1, many technical details still remain to be worked out, e.g., for statistical assessment of agreement between time-varying vector-field outputs of model and experiment. While these are currently being worked on, many examples can already be cited where the simpler case of accuracy comparison for scalar outputs of model and experiment have been accomplished with reasonable rigor. So, Item 1 seems to already be practical to apply in some simple cases, and will probably quickly become more widely applicable in more complex cases.

In a particular validation exercise, let us assume that the accuracy characterization aspect can be accomplished in a reasonable and practical manner. Let us also assume that the model adequacy assessment, Item 2 above, can also be accomplished, and that this results in an indicated model accuracy that lies within stated accuracy requirements. Accordingly, the model is "accepted". Can it be pronounced 'validated' at this point? I argue 'NO', in the following context. We are not yet free to proceed with the model for use in interpolation or extrapolation and to claim that we used a 'validated model' in the prediction. The reason is that, before interpolating or extrapolating with

---

[2] contact: vjromer@sandia.gov, 505-844-5890
[3] 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.

the model, we must first accomplish Item 3 above. That is, we must first augment the model with a representation of the *resolution uncertainty* of the validation experiment(s). The resolution uncertainty of the validation experiment(s) sets the uncertainty level to within which the model can be ascertained to agree with reality.

The resolution quality of the validation activity is governed by the tightness of control on boundary conditions, materials, and geometry; accuracy and precision of the experimental apparatus, measurement equipment, and measurements; number of repeat experiments, etc. The resolution uncertainty to be defined below is somewhat different from, but in the same spirit as, the "validation uncertainty" that Coleman and Stern ([1]) present in their validation theory. However, I advocate using the concept of a resolution limit in a different way than Coleman et al. Instead of reporting that this is the uncertainty level within which the model can be validated, I advocate use of the quantity as an uncertainty term that is mapped back into the model to form what is called the "**combined or augmented model"** with which interpolatory or extrapolatory predictions are to be performed. The combined model is a {deterministic model + uncertainty representation} that implicitly embodies the resolution uncertainty of the validation experiments. (For deterministic models that are significantly biased from the data, the combined model can also be augmented with a bias correction as discussed in [2].)

This "validation requirement" on the model, before it can be used in prediction, is analogous to requiring that, e.g., experimental uncertainties be included in the representation of material property data. This is something that especially those performing model validation recognize as essential. Consider a simple example. Figure 1 shows a straight line through the data determined by a Least-Squares regression. The straight line can be recognized as a deterministic model of the nominal material-property behavior over the tested temperature range. However, the deterministic model alone does not suffice to define material-property value at some given value of the state variable. The actual material-property data deviates significantly from the regression line, i.e., from the deterministic model. That is, the deterministic model does not sufficiently reflect the actual measured values, but must be augmented as follow to more accurately represent the data.

The random deviations of the data about the regression line are deemed to be caused by random deviations in the experiments and measurements in going from point to point. (Otherwise, the deterministic model would be formulated to go from point to point to more appropriately represent the precisely known material-property behavior.) Such random deviations derive from various sources of uncertainties that incur different realizations at different measurement points. Such sources include deviations of applied boundary condition from experimental set points; use of different *nominally identical* measurement sensors and/or reading instrumentation that are really only within certain accuracy tolerances of each other; and unit-to-unit differences in nominally identical experimental units, geometries, materials, etc. Point to point random deviations of such uncertainty sources are manifested in the observed random scatter in the experimental data as it rises with a distinct upward trend as the value of the state variable increases. The upward trend of the data is captured by the slope of the regression line. The line also serves as a deterministic reference function about which the random deviations of the data can be parameterized by the standard deviation or 'standard error' of the data about the regression line.

To use only the deterministic trend line and neglect the deviations of the data about the linear model is to imply exact knowledge of the material property value as a function of the state-variable value. This misrepresents our actual state of knowledge. The material-property model must also reflect the uncertainty of the property measurements. This is set by the quality of the material characterization experiments, manifested in the scatter of the data about the reference

trend line, and modeled by the standard deviation of the data about the trend line. This uncertainty can be mapped into the deterministic model through its ordinate-intercept parameter. This parameter can be made to vary with a standard deviation and probability density function (PDF) given by the standard error and PDF of the data about the regression line. Hence, the deterministic model of nominal property behavior (the line) is augmented with an uncertainty model to obtain a *combined model* that best represents the state of knowledge of the material behavior.

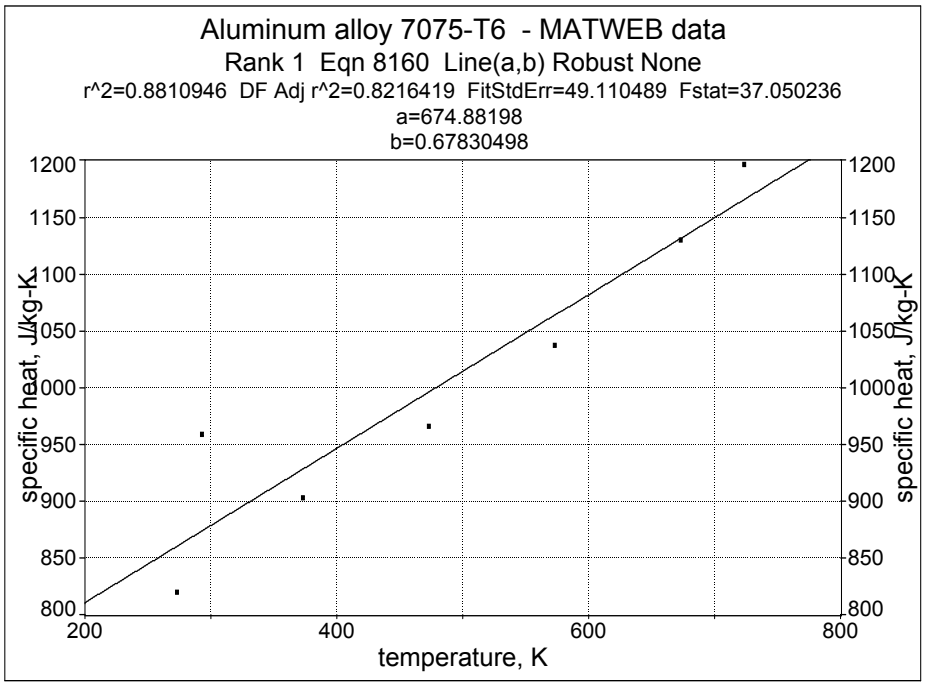

Figure 1: Material property measurements and Least-Squares regression line through the data.

The experimental **resolution uncertainty** that must be carried in the combined model is the "**extraneous"** uncertainty that exists in the validation experiments but not in the application space, and is not separately characterized elsewhere. For example, the random variability in the data will have a contribution due to sensor accuracy uncertainty if different sensors are used at the different measurement points. If the particular sensors used have random bias errors (random accuracy errors) that are reasonably represented by the manufacturer-published accuracy uncertainty of the sensors, then the associated variance may be subtracted from the variance of the data about the mean trend line discussed above.[4] The result is called the **reduced**

[4] A straight subtraction of variances presumes a linear convolution of independent contributing uncertainties, which is often adequate for most engineering purposes. The subtraction is accomplished by first performing a linear forward propagation of the applicable uncertainty, such that $\text{Var[response]}_{\text{due to model input factor}} = \text{Var[model input factor]} * [\partial\text{response}/\partial\text{input\_factor}]$. The involved partial derivatives are usually approximated by a one-sided or central difference, but care should be taken to assess derivative accuracy with respect to interactions between difference step size, model noise, and model solver tolerances ([3]). The propagated variance is then subtracted from the total experimental variance of the data about the mean trend line. (Because we are differencing, the model being used to form the derivatives does not need to be accurate in an absolute sense, but only in a much weaker relative sense with respect to trend information. This is somewhat comforting because the model being used is the subject of validation in the first place.)

experimental variance. Because some of the original variance in the data would be due to sensor variability, and we have this independently characterized, we can use this knowledge to decrease the resolution-uncertainty "penalty" from the validation experiments that must be carried along with the model. If a separate characterization of sensor bias uncertainty from the manufacturer or some other source was not available, then we would not be able to commensurately reduce the experimental variance.

Now consider a contributor to the data variance *that also exists* in the application setting. For example, assume that the different data points are obtained from different but nominally identical experimental units that embody the unit-to-unit variability that would be seen in the application setting. Let the "unit" be a component or subsystem of a larger system. The validated model of the unit is eventually to be part of a larger system model. Let experimental response of the component depend on the spatial-average density of packaging foam in the unit. If we have some way of measuring foam average density, then the variations in the response data due to variations in foam average density could be subtracted from the observed response variance. However, when the resulting augmented model of the component is merged with the rest of the system model, the effect of the foam density uncertainty would then have to be **explicitly** added back in. This could be accomplished by treating the foam density parameter as an uncertain variable in the system model.

A better approach is probably to not subtract out the foam's effect on the experimental variance in the first place. This preserves the "Top-Down" uncertainty characterization ([4]), where the foam uncertainty effects are **implicitly** carried in the augmented model. Therefore the effects would not have to be separately modeled in the system model. This approach **consolidates** the uncertainty of all the factors sampled in the validation experiments into a single variance representation that is carried in the augmented model. In several situations with complex models having many such input factors, the author has carried the experimental variance through a single parameter in the augmented model, much like the variance of the data in Figure 1 is carried via the variance of the y-intercept parameter of the linear property model. Mapping each submodels' elemental uncertainties into the system model through one or a few parameters per submodel allows better modularization and management of complexity in system-level uncertainty models and calculations. Then substantially fewer uncertain parameters must be manipulated and managed at the system level.

Of course, any uncertain factors that exist in the application setting that are not sampled in the validation experiments must be explicitly modeled with a Bottom-Up uncertainty propagation. Consider a case where a custom batch of specially controlled low-variance packaging foam is produced for use in the validation experiments. This might be done to isolate for examination the effects of other variance contributors in the model validation activity. The response variance associated with the (known) density variance of the custom foam would first be subtracted from the total experimental variance. Then, the effect of the actual density variance of the standard production foam (used in the application setting) would be incorporated into predictions by either: A) propagating the effect of the actual foam variance in through the variance representation of the augmented unit-level model; or B) separately modeling foam density effects via an explicit uncertain parameter in the system-level model.

Finally, consider a different case than has been assumed in Figure 1. Consider the case where the same sensor is used at all the measurement points of Figure 1. Then the data points would each have uncertainty bars about them. The uncertainty derives from the fact that the sensor has some

accuracy uncertainty. The sensor output at each data point would therefore have a common or *systematic* bias error that is associated with the particular sensor. Since we do not know what the particular sensor's bias error is, we assign a characteristic uncertainty magnitude to the sensor, from manufacturer specifications or some other characterization activity. This sensor uncertainty (variance) at the data points is then linearly convolved (summed) with the point-to-point variance about the regression line. The total variance is then carried in the augmented model.

We now return to considerations of model validation. Clearly, the regression line (deterministic model) in Figure 1 falls within the uncertainty of the experimental data. A **model-centric approach to model validation**, which combines Items 1 and 2 above, would therefore tend to judge the linear property model as having been validated, or at least as having being validated within the resolution allowed by the quality (uncertainty) in the experiments. However, such a pronouncement is not sufficient for the model to be appropriately extended to other settings, e.g., in the analysis of components containing the material, or in material-property extrapolation outside the range of the experiments. For the model to be used in other modeling settings, it must also embody the uncertainty that represents the experimental resolution to within which the model can be claimed to agree with the data. Thus, we arrive at the reasonable criterion stated above that, in order to claim use of a 'validated model' in a prediction, the extraneous experimental uncertainty in the validation experiment(s) must be carried along with the original model, in a *combined* or *augmented* representation to be used for predictions.

If the physics being modeled is effectively deterministic[5], then as the experimental measurements become noisier or more uncertain, the resolution uncertainty within which the model can be said to agree with the data increases. Since the resolution uncertainty is mapped into the model before extrapolating with it, less resolution (greater uncertainty) in the validation experiments is reflected in greater uncertainty in the augmented model, that is carried forward into predictions. Hence, there is no "free lunch" where lower resolution or quality in the validation experiments (i.e., larger error bars on the data) make it easier to claim a "validated" model by the criterion that it produces results that lie within the error bars of the data.

To reiterate, **application-centric model validation** requires that Items 3 be performed as a **minimum requirement in order for models to be validated for use**. Note that this is just the minimum differentiator between illegitimate and potentially legitimate claims that "validated models are being used in these calculations". Others requirements may apply in various circumstances. By this **working criterion for model validation**, to date I have seen perhaps hundreds of illegitimate claims that validated models were being used in predictions. Note also that *this criterion for model validation paradoxically requires at least some degree of adjustment or augmentation of the model* in order to prepare it for extrapolation or interpolation. That is, the deterministic model has to pick up an uncertainty contribution from the validation experiment(s), as well as perhaps a bias-correction component if thought to improve the predictive capability of the model.

---

[5] The stochastic case is slightly more complex. Here there are stochastic interactions between the assembled elements of the system or subsystem being modeled. Even if all the models of the separate system elements have been validated in the sense of Item 3, a Bottom Up propagation of uncertainty would still not contain the uncertainty at system level from stochastic interactions between the elements. Whether the uncertainty propagated from Bottom-Up has a mean and variance that resemble the actual system statistics can only be ascertained through a Top-Down validation. Top-Down validation presents an opportunity to confirm, or if necessary calibrate, the calculation of system statistics from a Bottom Up approach that would then be used in extrapolative predictions ([2]).

Note that this working criterion for model validation includes statistical and highly calibrated models. This may be a point of contention for some. However, reference [5] argues that even purely calibrated models, like statistical regression models of material property behavior or component failure, do indeed sensibly fit into the class of validatable models, especially if they are being used in interpolation (the intended use). Certainly, it seems contrary to common sense that a partial-differential-equation (PDE) based model could be said to be validated (by the model-centric criterion, Item 2), while a regression model is not validatable even though it better agrees with the data over the interpolation parameter space of intended use. Such is the case in [5]. It makes sense to recast the definition of model validation to recognize the application-centric criterion for validation, which admits calibrated models to the class of models that can be validated. This is important because then we can talk about validating models for use from an exceedingly relevant and useful class of models that includes material property models, constitutive models, failure models, statistical regression models, etc.

Besides, one cannot escape the fact that effective models tend to have some amount of calibration in them. Show me a model that performs well on a real problem, and I am confident I can point out where it has been calibrated. Even "first principles based" PDE models are calibrated. Material-property parameters in PDE models are free parameters. Some of these parameters are obtained by iterating the properties values until 1-D predictions best match 1-D experimental results. Hence, a self-consistent matched set of PDEs and parameter values arises that best replicate the experimental data. The presence of non-constant material properties, such as temperature-dependent thermal conductivity, signifies that the extent of the equations' predictive capability is limited. The applicability of the equations is extended in state-variable space through re-calibration via the mechanism of state-dependent material properties.

The issue of predictiveness in extrapolation appears to come down to the rate in the parameter space at which the model must be recalibrated, versus how far in parameter space the extrapolation lies ([2]). Models that need recalibration at slower rates in the parameter space are thought of as "more predictive" or better in extrapolation. Those that need more frequent re-calibration are thought of as "less predictive" or worse in extrapolation. However, there does not appear to be a quantitative methodology for measuring the rate at which the applied equations must be recalibrated in the parameter space, nor for gauging predictiveness by assessing this rate against various extrapolation directions and/or distances.

Some comments are now made regarding Item 2 listed above. My personal observation is that it is not an applicable type of model validation in most real cases. A statistical comparison to determine if an accuracy requirement is met with some acceptable degree of statistical significance, even if some uncertainty exists in the requirement, is easy enough to do if a scalar requirement exists. One problem, however, is that both the accuracy requirement for the model, and the level of significance set in a hypothesis test of whether the accuracy requirement is met, can only be based on subjective criteria instead of quantitatively-derived unique criteria. Theoretical and pragmatic reasons for this are cited in [4], but the empirical evidence is also convincing. In my personal experience, the numerous well-funded high-fidelity validation activities I am aware of have never involved a rigorously traceable requirement for the acceptable level of model accuracy.

Even in cases where a unique mapping could exist in principle, a practical impediment exists in mapping accuracy requirements from the application space to the validation space where the hypothesis test would be applied. I have considered such a mapping in several projects. It does not appear to be possible unless there is a continuous, parameterized mapping between the validation and application domains. That is, any changes in the conditions from one domain to the

other have to be recoverable by smoothly morphing the model from the application domain to the validation domain. This means that you must have the degrees of freedom and associated parameters in the model to span both domains. In other words, you must have the "same" model for both domains, with only the values of the parameters being different. Needless to say, current modeling technology and practice do not support models that can morph between the different geometries and boundary conditions that normally exist between validation and application domains.

Another problem with Item 2 is the issue of multiple accuracy requirements. Even when considering a scalar accuracy requirement, many problems exist even in principle. Worse yet are cases where multiple accuracy requirements are put on the model. "It must be y% accurate in this region of the fluid domain and w% accurate in that region." Or, "It must calculate total body drag to within u% and total heat transfer gain to within v%," etc. How does one deal with competing accuracy requirements, some of which are met and some of which are not? Is the model of no use or value if it does not meet all the prescribed accuracy requirements to the required degree of significance? How would one formulate and interpret a weighted hypothesis test for multiple accuracy objectives?

Thus, it is argued that hypothesis tests for model accuracy (Item 2) are very subjective and volatile measures of model quality/value/usefulness. In the end, the answer to whether a model is good enough is impossible to rigorously determine in most realistic cases. In some cases the circumstances are so loaded that the answer is obvious, but when the question becomes close the arbitrariness issue surfaces. Even if the judgment is made, this says nothing about whether the model is good enough for the extrapolative predictions it will be used for. Hence, I do not accept that Item 2 is normally a viable aspect of model validation.

What about the morphing problem between validation and application spaces? If accuracy requirements cannot be rigorously mapped from the application domain to the validation domain, how can augmented models, {deterministic model + uncertainty representation}, be accurately extrapolated from the validation to the application domain? The answer unfortunately appears to be that they cannot be guaranteed to be accurately extrapolated. This conclusion changes the hope of being able to perform extrapolative predictions with rigorous statistical confidence assignable to the predictions.

Although we cannot guarantee accuracy of predictions or accompanying uncertainty bands, we can still set about the objective of contextualize and improving our estimates as much as possible through appropriate model validation procedures. We can also attempt to maximize accuracy potential through optimized design of validation experiments and optimized model development, calibration, and extrapolation procedures for a given prediction task. This is an engineering science still in the very early stages of development.

In summary, it does not appear to be determinable with any particular certainty whether a model meets accuracy intentions in the validation space, much less whether it will meet them in the application space. However, this does not preclude the objective of making the best possible predictions and associated uncertainty estimates. Items 1 and 3 are essential in this regard, and comprise an "application-centric" aspect of model validation. The model-centric form (Items 1 and 2) appears to currently define model validation for most of the model validation community. However, I argue that claiming use of a "validated model" in predictions requires the application of Items 1 and 3. This makes them fundamental aspects of model validation, which together comprise a recognizably different facet of model validation than the model-centric facet. This is as Code and Calculation verification are recognized as different facets of verification.

REFERENCES

[1] Coleman, H.W., and Stern, F., "Uncertainties in CFD Code Validation," *Journal of Fluids Engineering* Dec. 1997, vol. 119, pp. 795-803.

[2] Romero, V.J., "An Emerging Theory of Model Validation and Its Relationship to Model Calibration and Extrapolation", Sandia National Laboratories SAND report in progress.

[3] Romero, V.J., "Characterization, Costing, and Selection of Uncertainty Propagation Methods for Use with Large Computational Physics Models," paper AIAA2001-1679 presented at the 42nd Structures, Structural Dynamics, and Materials Conference, Seattly, WA, April 16-19, 2001. Updated and extended version available from the author at vjromer@sandia.gov.

[4] Hasselman, T., Yap, K., Wathugala, G., "A Top-Down Method for Uncertainty Quantification and Predictive Accuracy Assessment," paper AIAA-2005-1903, 46th Structures, Structural Dynamics, and Materials Conference, Austin, TX, April 18-21,2005.

[5] Romero, V.J., M.P. Sherman, J.F. Dempsey, J.D. Johnson, L.R. Edwards, K.C. Chen, R.V. Baron, C.F. King, "Development and Validation of a Component Failure Model," paper AIAA-2005-2141 presented at the 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, April 18-21, 2005, Austin, TX. Refined version with corrections available from the author at vjromer@sandia.gov.