SAND2007-0678C
2007-01-1746

A Paradigm of Model Validation and Validated Models for Best-Estimate-Plus-Uncertainty Predictions in Systems Engineering[1]

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

ABSTRACT

What constitutes a validated model? What are the criteria that allow one to defensibly make the claim that they are using a validated model in an analysis? These questions get to the heart of what model validation really implies (conceptually, operationally, interpretationally, etc.), and these details are currently the subject of substantial debate in the V&V community. This is perhaps because many contemporary paradigms of model validation have a limited modeling scope in mind, so the validation paradigms do not span different modeling regimes and purposes that are important in engineering. This paper discusses the different modeling regimes and purposes that it is important for a validation theory to span, and then proposes a validation paradigm that appears to span them. The author's criterion for validated models proceeds from a desire to meet an end objective of "best estimate plus uncertainty" (BEPU) in model predictions. Starting from this end, the author works back to the implications on the model validation process (conceptually, operationally, interpretationally, etc.). Ultimately a shift is required in the conceptualization and articulation of model validation, away from contemporary paradigms. Thus, this paper points out weaknesses in contemporary model validation perspectives and proposes a conception of model validation and validated models that seems to reconcile many of the issues.

INTRODUCTION

Loosely speaking, Model Validation is concerned with the accuracy of models and model predictions as compared to reality, or more commonly, some subset or filter of reality that is important to predict for some purpose. That is, in as far as we can (through appropriately designed and controlled experiments and comparison against model predictions) ascertain the degree of agreement at specific validation point/s in the parameter space.

There seems to be a fairly uniform agreement in the validation community on what model validation implies at a *vague conceptual* level, i.e., at the level of various one-sentence expressions like the first sentence in this section. However, the *details* of what model validation really implies (conceptually, operationally, interpretationally, etc.) are subject to substantive disagreements among many in the V&V community. The intent of this paper is to outline the relevant issues and propose a conception of model validation and validated models that seems to reconcile many of the issues. This paper develops and extends ideas first presented in Ref. [27].

SOME PRELIMINARIES

What are the criteria for assessing and answering whether one has, or is using, a validated model? Because many contemporary paradigms of model validation have a limited modeling scope in mind, their answers for this question do not appear to be viable across the various modeling regimes and purposes that are important in engineering. After outlining in this section the different modeling regimes and purposes that it is important for a validation theory to span, in the next section the author proposes a validation paradigm that appears to span them.

To begin understanding the issues, a first partitioning of the contextual space of model validation can be made in terms of **"physics-field" validation** on one hand, and **"effect" validation** on the other. For example, we might be interested in how well a fire CFD simulation represents a hydrocarbon pool fire in calm wind conditions. The fire CDF model is the validation object of attention. We could contemplate performing validation comparison of the simulation's spatial-temporal *field* predictions (e.g., pressure, three components of velocity, species concentrations, etc., all as a function of time and space) against identical quantities measurable in the experiment (e.g., with laser sheets and particle image veloci-

[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.

metry). Such field-based validation is very difficult and expensive, perhaps nearly impossible—the author knows of no instances where this has been fully accomplished yet. More practically, we could simply compare the *effect* that the fire has on a number of sensors or transducers of the field. Such sensors could be, e.g., flux gauges distributed throughout the fire, and/or suitably located and oriented calorimetric plates, objects, and walls outfitted with thermocouples. Because the fire is stochastically fluctuating in time and space, these sensors would have to provide a sufficient amount of time and space averaging to get well behaved quantities for comparison. By measuring and computing such quantities, the computed *effects* of the fire can be compared to measured *effects*.

Of course, even if predicted effects compare favorably with measured effects, this is necessary but not sufficient to conclude that the predicted physics field details would compare favorably with the actual fire field details. In fact, more detailed comparisons might compare quite poorly. If the modeled physics is not precisely correct at a detailed level, but the aggregate *effect* is sufficiently accurate, the model can still prove quite useful. Because no model is a perfect representation of reality, it is unavoidable that at some level of examination, agreement will break down. We can only strongly conclude, then, that the particular *effects* we have quantified are similar, and not that the full fields are similar. Nonetheless, we would like some reassurance that the full fields are not drastically dissimilar but somehow coincidentally yield similar effects. The more spatially and temporally diverse (well distributed) sensors used, the greater the chances quickly go up of identifying any such coincidence. This drives toward using as many diverse sensors as affordable in validation experiments. Investigating a diversity of conditions/scenarios in a validation matrix is another way to perhaps identify any such coincidence. (A hierarchical validation approach would not work well here to incrementally build up to the full system (fire) to try to affirm that in the end, at the system level, any agreement of effects is occurring "for the right reasons". It is not possible experimentally to build up to the full fire by incrementally adding different mechanisms, e.g., combustion, then turbulence, then soot generation/agglomeration, then radiative participation, etc.)

The next item to consider regarding contextual characterization of model validation is concerned with **interpretation and usage** of the results from model validation activities. It is not worthwhile to go through the expense and rigors of a validation exercise unless the results can be interpreted in some useful manner, where salient and meaningful conclusions can be drawn, and/or what is learned can be used to quantitatively caveat and/or improve the model. The key point here is one of information generation versus knowledge/value generation. If a large amount of information is generated but cannot be interpreted or used meaningfully, then the utility of the exercise is questionable and difficult to justify in terms of opportunity cost of the funding and resource allocation. Therefore, for a model validation activity to be relevant beyond just qualitative insights, there needs to be some plan for how the information gained will be interpreted and packaged to provide a quantitative characterization of model accuracy. It would be further desirable if this accuracy characterization transferred well, or at least optimally well, to other applications of the model beyond the validation experiment conditions.

Consider modeling at the stage where a pool-fire model is being developed by a group of researchers. The ultimate objective is to employ the fire model in engineering applications where prediction of object heating (e.g., weapon or structural member heating) is the engineering purpose. Before getting to the application-modeling stage, the fire dynamics model is usually developed in relative isolation, say at a university or fire research laboratory. Here, the modeling context is devoid of a particular engineering application for which the model is to help resolve certain issues, answer specific questions, and/or make decisions. Thus, the modeling goal at this stage is simply that the model replicate an actual fire as well as possible. Beyond this, there appears to be no absolute accuracy requirements against which the model might be judged to be "sufficiently accurate to be validated", as many contemporary paradigms of model validation would frame it. What criteria are applied, then, in characterizing the quality and usefulness of the model? In more stark validation terms, if one wants to put a stamp of acceptance (validation) signifying that the model adequately replicates the data and is therefore suitable for use in engineering applications, then what quality/accuracy hurdle or threshold must the model pass?

Since the objective in this regime of modeling is simply that the model replicate an actual fire "well", one approach might be to apply a hypothesis test to determine whether the model is different from the data—within the ability to determine this as allowed by the quality of the validation experiments. Experimental resolution level is governed by quality control on inputs such as boundary conditions in the experiments, and on instrumentation/diagnostic uncertainties on measured outputs. Ref. [7] explored the paradigm that a model is validated if its deterministic nominal prediction is biased from the mean of the data by an amount encompassed by the experimental uncertainty. Recognizing this uncertainty floor presented by the experiments—below which any model bias cannot be established or resolved—presents one key to the validation puzzle. However the final interpretation in [7], even for models that do qualify by their criterion, was not satisfactory as the first few paragraphs of the next section indicate by analogy.

Another issue is that the more precise the validation experiment, the more difficult it is to find a model valid by such a test; and the more noisy and uncertain the experiment, the easier it is to validate the model. Seen another way, the same model pronounced valid (by such a criterion) in one validation activity could be pronounced not valid by a much better (cleaner) experiment. In the limit, models will always fail this type of vali-

dation test, at every sensor, as the experimental resolution of the experiments becomes infinitely precise (then any non-zero model bias will fail the test). This simply reflects the truth that no model perfectly represents reality. Even so, some models are still very useful. Hence, ultimately the dynamics here are not what one would want from a validation formalism.

One possible way out is to allow an acceptably small amount of bias error between model and data (that is larger than the experimental resolution floor), and then test the hypothesis that the differences are likely less than this allowed error. But, how is a meaningful non-arbitrary accuracy limit established? Even if one could be established, there are substantial practical and theoretical problems with this approach, as will be discussed later.

Now consider use of the fire model at the engineering applications stage. The goal is to predict heat loads on some object of interest in the fire. If the fire model has been successfully validated in isolation (by any scheme, including the author's), is it validated for use in the engineering application? The difficulties in establishing an affirmative answer are numerous and daunting. The logical implication of carrying the pronouncement of model validity from the isolated-fire situation to the situation with the object in the fire is that the accuracy level established in the former situation is also the equivalent accuracy level in the new situation. But how is this rigorously established? How does the validation metric transfer mathematically/quantitatively from the old to the new situation such that the accuracy conclusions can be said to remain the same? Even if not remaining the same, how might the accuracy characterization at the validation setting mathematically/quantitatively export to the new setting such that accuracy estimates can be made there?

In the next section a methodology is proposed by which the validation accuracy characterization can be mapped from the validation setting to an application setting. However, even so, the author does not consider that the model is valid or to be considered validated at the new application point. Instead, the author considers the validation information mapped to the new application point to contribute to what might be called a Best Estimate Plus Uncertainty (BEPU) prediction with an **accredited** model (as discussed in the closing section). This is intended to be a much weaker and less misleading statement than "We're using a validated model in this application" when the model was actually validated at different conditions.

The following examples explain why the author does not advocate claiming that a model validated under one set of conditions in the modeling space can be considered validated at a significantly different set of conditions. Consider the fire/object modeling application. Putting the object in the fire adds new physics and significant interaction effects between the object and fire. For instance, the presence of the object adds surface-shear-driven vorticity to the flow field at the object surface. Since the fire in isolation has only buoyancy-driven shear vorticity generation, the shear, vorticity, and perhaps turbulence generation models might not suitably handle the new type of surface-driven phenomena, even though they might perform fine in the isolated fire conditions where the model was validated. Hence, the surface-shear-driven turbulent mixing of fuel and air may not be adequately modeled, and since mixing strongly affects combustion, the local combustion and consequent heating of the object may not be represented well.

Whenever physical conditions change, new physics modes inevitably come into play. How large the effect will be on the validation quantities of interest, and how well the existing (validated) model will pick up and handle these new modes, is something that is very difficult to predict. As another example, it is generally regarded that structural dynamics models do not extrapolate particularly well. If a model is validated at one or several points in the parameter space, this does not strongly imply that it will be successful at other untested parts of the parameter space. Even though good correlation might exist at the tested points, it is unknown and unknowable whether it will hold up sufficiently in extrapolation. (Sounds a lot like the obligatory caveat with financial markets and investing, that past history is no indicator of future performance). Another example that the author is intimately familiar with are foam thermal pyrolysis and charring/ablation/vaporization. The agreement between present day state-of-the-art models and actual foam behavior can vary quite sensitively at different foam densities, venting conditions, and applied heat fluxes and heating rates. This is why the author insists that validation conclusions can only be drawn at specific validation points in the parameter space, and these conclusions cannot be expected to reliably extrapolate to significantly different conditions. It is unknowable how much is "significantly different" until actual testing determines this.

Hence, it is essential to design validation experiments as close as possible (in the parameter space) to the actual conditions of the application for which the model will be used. (This is of course also the best philosophy for calibration of models.) This would appear to maximize the applicability and relevance of the validation results to the intended application.

With this mind, we return to the problem above where the objective is to predict heat loads on an object of interest in the fire. Let us assume that several objects of different size and/or thermal mass and/or surface roughness and/or rotational orientation are to be analyzed in model calculations after validation to one specific case. Since the physical regime is such that the object's presence substantially affects the fire, through the flow field changes it causes and because it's thermal mass has a back-coupling effect that moderates fire intensity, it is best to design a validation experiment that has a representative object in it.

In this modeling regime, the validation context is one where the object is a sensor that provides an effect measure (of the fire field) with connectivity to a concrete engineering issue. Perhaps it is then possible that the fire model can be validated for this effect (imparted heat to the object) to have the required accuracy to resolve the engineering issue of importance. This case would therefore appear to have a different validation contextualization than the modeling regime described above, which is devoid of a specific engineering object and accuracy requirement, and different also from the modeling regime discussed at the end of this section, which is also generically non-specific to a particular engineering application. However, on closer examination below, this differentiation is typically only cosmetic as regards model validation.

On the surface, it would appear that an engineering application with stated performance requirements would yield commensurate accuracy (validation) requirements on selected effects monitored in the model validation activity. However, this is not usually the case. Accuracy allowables on top-level engineering requirement(s) such as performance, tolerance, or safety requirements cannot be uniquely mapped down to the physics effects level. Assume for the sake of argument that performance, tolerance, or safety "requirements"[4] at the system level can be transformed into modeling accuracy allowables on performance predictions at the system level. At best this is a difficult type of mapping to perform, but the author does have affirmative experience that it is possible, with sufficient conventions and constraints applied to get a unique mapping. Applied to our example problem, say that the mapping yields a requirement that the predicted temperature response of components in a fire-engulfed weapon not differ by more than 10% from actual component response in a validation experiment. In trying to map this error budget at system level into modeling error budgets in the various elements of the fire/object model, however, an essentially infinite number of non-unique combinations will satisfy the top-level error budget.

For example, in a hierarchical model validation approach the requirement could be semi-arbitrarily parsed into an equivalent top-level effect of 2% for thermal modeling errors within the weapon plus an equivalent top-level effect of 8% for fire-modeling errors. Then, in a fractal manner, within the thermal model itself it is possible for an infinite number of different combinations of error budgets in, e.g., the foam pyrolysis and contact-resistance submodels, to meet a top-level error of 2%.

[4] Such "requirements" at the system level are unavoidably subject to some degree of arbitrariness from, e.g., decision-maker knowledge limitations and subjective variabilities in risk perception and tolerance based on individual predispositions and experiences; uncertainty and conflict in program objectives and resource constraints; and non-uniqueness of any tradeoffs involved, etc.

Say that one such combination happens to be a relative error in the foam model of 18%, and in the contact resistance model of 30%, and that the actual modeling errors qualify to these levels. The only way to determine the 18%/30% mix, unfortunately, is to propose an 18% error in the foam model, and then inversely solve the thermal model to determine the maximum error in the contact-resistance model that would yield a top-level error of 2%. If one instead started with a proposal of 25% error for the foam model, then this by itself might lead to a larger than 2% error at the system level, so a lower foam modeling error would have to be proposed, but how much lower? Only trial and error could tell. Say that a 20% error for the foam model qualified, resulting in an allowable of 10% error for the contact-resistance model. If the contact-resistance model has a lower attainable limit of 25% error because of uncertainties in its inputs, does it make sense to reject it? No, because the system error requirement can still be made with this model. Up to a 30% error is allowable with this model, when coupled with an 18% or less error in the foam model. Thus, if the system-level error budget is met, all submodels are valid by definition, no matter what their individually assigned (non-unique) error budgets are, and conversely, if the system requirement is not met then all submodels fail together, irrespective of any non-unique error budgets assigned.

Thus, trial and error, inverse calculations, and arbitrariness would be involved pervasively. Imagine the perplexity when error budgets for all the other submodels are brought into the mix—since none can be required to have zero error!

Now consider the fire model with its (non-unique, substantially arbitrary) budget of 8% top-level equivalent error. The 8% allowable error at top level could be decomposed into an infinite number of spatial-temporal error fields of incident flux at the weapon surface. An infinite number of spatial-temporal error fields also don't meet this requirement. The only way to know is to propagate the errors up to the system level and see whether they meet the 8% "requirement". Say that such error fields are somehow proposed, and the errors are propagated to system level. Then the first one that results in error less than but hopefully close to 8% at system level would be logically picked. Say that an error field yielding 7% equivalent error was found. Does this error field make a suitable standard against which actual fire model error field (determined from comparison with data) should be judged in hopes of validating the model? Consider that the model could be rejected by this standard if the actual spatial-temporal field of model errors was not everywhere lower than the reference error field, even if the actual error field produces less than 7% error at system level.

What does all this suggest? Evaluate conformance with accuracy requirements at the top level instead of trying to map accuracy requirements down. At the lower levels, bound the actual modeling error and propagate this upward to the system level. To the author, a validated

model at any modeling level is one that properly bounds the actual error in a form that can be transmitted or propagated effectively in downstream use of the model. The next section concentrates on this perspective.

The impropriety of validation paradigms that potentially reject models based on arbitrary accuracy "requirements" is now clear. Beyond this, hypothesis tests themselves have subjectivity in their level-of-significance criteria. Small perturbations in the level-of-significance threshold can variously switch between rejecting and accepting a model. Another problem is hypothesis testing when multiple accuracy requirements are involved. "To be validated, the model must be y% accurate in this region of the fluid domain and w% accurate in that region." The error field in space and time discussed above is an example of this. The multiplicity might in combination or alternatively be with respect to different physics fields: "The model must calculate total body drag to within u% error and total heat transfer gain to within v% error." If some of the multiple accuracy requirements are met and some are not, is the model considered validated or not? Is the model of no use or value if it does not meet all the prescribed accuracy requirements to the stipulated degree of significance, even if the ones not met were just barely not met? How would one formulate and interpret a weighted hypothesis test for multiple accuracy objectives? In short, hypothesis tests are very subjective and volatile measures of model quality/value/usefulness, even when they can be applied.

Furthermore, a practical problem exists with the concept of model rejection itself. In most modeling initiatives, resources are constrained and we have to ultimately use the model when the development resources run out, but simply want its error to be characterized so we can account for this in the predictions. Even if the model does not achieve the hoped-for accuracy, we can still use the model with lowered expectations as determined from the validation activity. Since the model will have to suffice in this case, it is "good enough" by default. The other alternative is to truly reject the model. However, this would normally represent a rejection of presumably the best-available codified knowledge of the physics application at hand. Since model development and validation activities are relatively expensive and time-consuming, rarely is the luxury afforded to build multiple models and apply the validation process. Rather, project resources are concentrated on developing what available experience and tools allow to be the best model possible under the resource constraints. If the model is rejected, then to avoid "analysis paralysis" at the modeling stage of the program, another (what would likely be inferior) model would have to be quickly thrown together and used—likely without having gone through the model validation process to characterize its accuracy.

Yet another set of problems normally exists with the hypothesis testing approach. Validation experiments are usually conducted under at least slightly different conditions than the intended application(s) of the validated model. This occurs often because of the need to control input conditions in the validation experiments in order to maximize resolution power to isolate modeling error in the validation activity. Cost and technical practicality usually also drive validation experiments to be simpler than the real applications. In many cases, such as nuclear weapons testing and nuclear power plant accidents, tests in the intended application space are not possible or feasible. A technical issue then exists in mapping requirements from the application space to the validation space, and in mapping validation results back from the validation space to the application space. This mapping cannot even in principle be accomplished 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 the degrees of freedom and model forms in the model must span both domains. In other words, the "same" model must exist 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, physics modes, and boundary conditions that normally exist between validation and application domains. The lack of a parametric relationship between the validation and application domains poses a discontinuous mapping between the two spaces, preventing even the possibility of a rigorous mapping of requirements (which would be non-unique anyway).

A discontinuous mapping also prevents a rigorous propagation of information the other way, from the validation setting to the application setting. Therefore, any potential statistical confidence statements that model error is less than some tolerance in the validation domain cannot be transferred to the application domain. This eliminates the hope of being able to perform extrapolative predictions with rigorous statistical confidence assignable to the predictions. Nonetheless, very useful information gained from the model validation activity can be captured and parameterized in a form that allows propagation from the validation domain to the application domain, as explained in the next section.

Hence, the author concludes that hypothesis testing is not a viable way of validating models in even the modeling regime where engineering performance requirements exist at the system level. A different approach to establishing validated models in this regime will be presented in the next section.

To finish up this section, we consider one last modeling regime. Examples in this regime are material property and constitutive models, turbulence models, convection correlations, thermal contact-resistance correlations, etc. This modeling regime exists, like the first one, in relative isolation from analysis, design, and decision-making associated with a specific engineered system. Here as well, there are no evident accuracy requirements against which the model could be judged to be "sufficiently accu-

rate to be validated". Nonetheless, model quality/accuracy can still be quantitatively expressed and transmitted to downstream uses of the model. The traditional manner for doing this will be cited in the next section. It will then be argued that this same philosophy also applies to models in the other regimes of modeling described above.

A HALLMARK FEATURE OF VALIDATED MODELS (EXTENDED TO NEW MODELING REGIMES)

What are the hallmarks of a model and its usage that allow one to defensibly make the claim that they are using a validated model in an analysis? Here a *hallmark criterion* for validated models is presented that applies in all three modeling regimes discussed in the previous section. It is explained how existing precedent from one modeling regime can be extended to the other two.

First consider the modeling regime at the constitutive level. What constitutes a validated model in this modeling regime? Lacking an external accuracy criterion for validation, one could look for some type of natural or intrinsic criterion. One possibility is to require that the model results lie within "the error bars"[5] of the experimental data. This philosophy could also be applied in the other model regimes discussed above. Indeed, many validation activities reported in the present-day literature advocate this philosophy. However, there are a few serious problems with it. First, the larger the uncertainty in the experiments, the larger the experimental error bars are. So in this conception of model validation, the less precise the experiments, the easier it is to accept or validate a model. The other problem is that a contemplation of the downstream implications of model usage leads the author to the opposite as a rational criterion: *the error bars of the experimental data should lie within error bars associated with the model*. That is, the model predictions with error bars should be validated to encompass reality, rather than the other way around. A few simple examples bring this point home.

Figure 1 presents some hypothetical data for an illustrative example of a measured material property as a function of temperature. The solid straight line shown is determined by a Least-Squares regression of the measured data points (i.e., the bias-corrected experimental readings shown at the midpoints of the affiliated uncertainty bars). The large error bar and associated intervals (dashed lines) shown about the straight line are reflective of the total experimental uncertainty. Their construction will be addressed later, but for now suffice it to say that these represent the best estimate within which the material property values are reasonably expected to lie—in all but the most unforeseen of instances. That is, at any given temperature, we expect that actual surface property values will fall within the large interval.

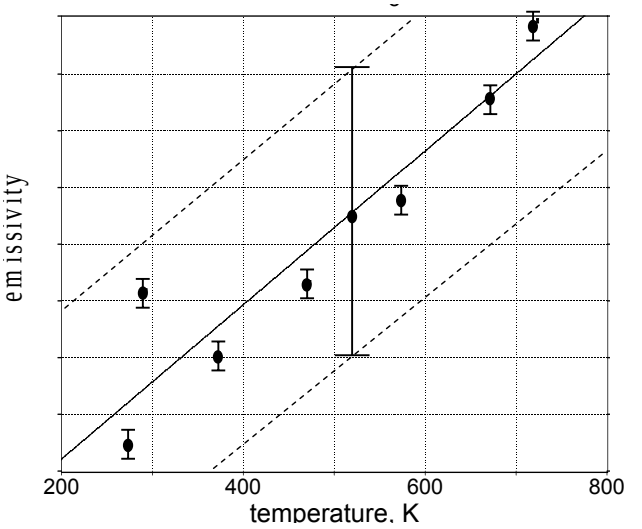

Figure 1: Material property measurements and associated uncertainty intervals, with regression line through the data.

The regression line can be recognized as a deterministic model of the surface-property behavior over the tested temperature range. Certainly, the model results lie within the uncertainty of the experimental data. Some conceptions of model validation would therefore deem the model 'validated for use'. However, the actual material-property data can deviate significantly from the regression line, i.e., from the deterministic linear model $Y(x) = a \cdot x + b$. Thus, the deterministic model by itself does not suffice to represent material-property expectations at some given value of the temperature state variable. If the deterministic model alone (now claimed 'validated') was utilized in downstream simulations at component/subsystem/system level, the material property realizations would be misrepresented as being much more precisely known than is actually the case. This would lead to under-representation of the actual uncertainty associated with use of this property model. Therefore, to claim a model validated because it lies within the uncertainty of the experimental data appears to the author to be wholly inappropriate, and inviting of trouble.

Rather, a legitimately validated model should include a reasonable formula for representing the uncertainty associated with the model's predictions. This would seem to be an essential element of what would constitute "best estimate plus uncertainty" (BEPU) predictions.

[5] Many complications, choices, and caveats are involved in forming such error bars on experimental data (as will be discussed later in this paper), so they are always somewhat subjective. Nonetheless, reasonable results are obtainable in many engineering validation activities with readily available uncertainty quantification and propagation technologies (e.g., [6] – linearized propagation for first-order estimates of mean and variance of experimental uncertainty, [26] – Monte Carlo propagation of uncertainties for higher-order estimates of mean and variance).

A convenient and perhaps optimal way to model the prediction uncertainty is via an uncertainty representation parameterized into the model itself, producing a "**combined** or **augmented model**". Examples will be provided shortly. Thus, the validation uncertainty, parameterized into the model at the validation point(s) in the parameter space, would be carried in the augmented model to prediction points in the parameter space. Evaluation there would provide an estimate of the uncertainty associated with the prediction.

As an example, the deterministic model $Y(x) = a \cdot x + b$ can be conveniently augmented with an appropriate parameterization of the uncertainty of the experimental material-property values. The uncertainty represented by the large intervals in Figure 1 can be parameterized into the deterministic model $Y(x) = a \cdot x + b$ though its ordinate-intercept parameter b. For instance, this parameter can be made to vary from $(b - h)$ to $(b + h)$ with a uniform density function, where h is the vertical distance from the reference regression line to the lower or upper large interval. Alternatively, if these intervals signify the approximate 0.025 and 0.975 percentiles (2σ intervals) of what is being modeled as a normally distributed uncertainty from the experiment, then replacing b by a normal distribution of variance σ^2 centered about b would parameterize the experimental uncertainty into the material-property model.

Note that the effect of parameterizing the experimental uncertainty into the model is to associate an uncertainty intrinsically with the model, such that its representation is no longer the infinitely thin line $Y(x) = a \cdot x + b$ in Figure 1, but a thick *prediction band* with extents the same as, and driven by, the experimental uncertainty. It is this augmented model that is the appropriate one for modeling the material in downstream component/subsystem/system simulations where prediction uncertainty associated with this model is to be accounted for. This is the essence of the author's assertion above that a legitimately validated model should have error bars that contain or encompass the error bars of the experimental data, instead of vice-versa as some would define the criterion for a validated model. As another example, the author does not deem a carefully developed convection correlation to be a validated model unless appropriate uncertainty intervals are cited along with it. Indeed, such correlations are usually presented in the literature with admonishments that the resulting point estimates can have typically ±20-30% error, and should be used with this kept in mind. Accordingly, the author deems use of such models to be improper in applications of substantial importance unless this uncertainty affiliated with the validated models is propagated to the predictions.

Note that in this conception of model validation, there is no "free lunch" where noisier or more uncertain experiments allow less accurate models to be validated than if the experiments were of higher quality. Instead, greater uncertainty in the validation experiments is reflected in greater uncertainty in the augmented (validated) model, which is carried forward into predictions.

These considerations for what is required of a validated model at the material-property modeling level can be extended to other modeling regimes, including those discussed in the previous section involving very complex systems (this will be elaborated later). This requirement remains invariant over all modeling regimes: *the error bars of a validated model should encompass the total uncertainty of the experimental data for pragmatically important measures of the system*. (The physics field *effects* discussed earlier are included here as measures of the system.) This stated criterion is necessary for validated models, but perhaps not sufficient in all modeling regimes and for all modeling purposes. Other criteria may exist in various circumstances.

Construction of the large uncertainty band in Figure 1 will now be addressed briefly. In general, total experimental uncertainty may include contributions from: a) stochastic variability of system response and associated confidence intervals due to finite numbers of repeat experiments; b) bias uncertainties associated with any models used to correct and/or interpret the measured data; c) measurement uncertainty on system outputs and contributed uncertainty from experimental inputs to the system, including apparatus/setup, test conditions, and boundary conditions.

For example, in Figure 1 let the small uncertainty bars about each data point represent the measurement uncertainty on the ***output*** of the experiment (emissivity value) due to sensor/diagnostic variability and bias uncertainties. To make things easier for illustrative purposes, assume that the measurement uncertainties are not significantly affected by temperature of the material, even though the property value itself is. Thus, the error bars are the same size about each data point.

Experimental uncertainty also exists with respect to the inability to exactly control and measure critical ***input*** factors such as boundary conditions in the experiment. Here, this means the temperature of the material sample, as well as its surface condition. Uncertainties in the inability to exactly control and measure material temperature can be transformed into implied uncertainty on the output of the experiment (emissivity as a function of temperature) through standard techniques.[6] This implied

[6]For many engineering purposes it is adequate to treat the system in the validation experiment as having locally linear behavior over the uncertainty of the experimental inputs to the system. Then the following relationship can be used:

Variance of [system response or output]$_{\text{due to input factor}}$ = Var[input factor]$*$[∂response/∂input_factor]$**2$. The partial derivative above would usually be approximated by one-sided or central differencing using the model of the system. Since the model is actually the subject of the validation activity, this is somewhat suspect. However, the model is only being used for relative trend information here, so does not need to be accurate in an absolute sense, but only in a much weaker relative sense.

uncertainty on the experimental output is convolved with the measurement uncertainty (error bars) on the output.

The particulars of propagating the uncertainties associated with control and measurement of the temperature in the experiment will be avoided for simplicity in the following discussions by presuming these uncertainties small enough to have negligible mapped impact on the uncertainty of the output measured quantity, emissivity. Nonetheless, it is important to mention the generic presence of this type of experimental uncertainty, and its mapping to output uncertainty. In many validation experiments, the output uncertainty due to e.g. uncertain input boundary conditions is not negligible, and can in fact be the dominant contributor to the aggregate uncertainty on the output(s) of the experiment. This was the case with the heating boundary conditions in a recent validation of a thermal model of a complex electromechanical device ([29]).

The other critical input factor in our example is the surface condition of the material. This implies surface finish as well as surface preparation in the experiments (such as cleaning and polishing). In our example, the surface condition of the material as employed in the field varies substantially. Our experiment is therefore designed to employ different material samples covering a representative variety of surface conditions, in a randomized order as temperature is increased in the experiment. Use of a different surface sample for each temperature experiment is presumed for the sake of illustration here to cause the oscillations (from linear) seen in the seven data points as the temperature state-variable increases.

If many more experiments could be afforded, it would be better to take each of the seven surface samples and test each of them through the temperature range. This might, for instance, give seven experimental data points at each of the seven temperature stations in the figure. To form the total experimental uncertainty (analogous to the large error bar in the figure), one of three routes would be taken. One route addresses the foreseeable case where the same sensor/instrumentation is systematically used for all seven emissivity measurements at a given temperature. In this case, the associated measurement uncertainty (small error bars in the figure) would be convolved with a density function estimated to contain the population from which the seven data values are likely to have come. This density function would account for confidence intervals on the mean and variance due to the limited number of data samples.

If instead, a different sensor/instrumentation unit was used for each of the seven emissivity measurements, then options 1 or 2 below are taken.

1) If the seven measurement packages used have random bias errors (random accuracy errors) reasonably represented by the manufacturer-published uncertainty or other characterization activity, then the associated variance may be used to decrement the variance (width) of the density function mentioned above. The result is called the **reduced experimental variance**. Because some of the original variance in the data would be due to sensor variability, and this is independently characterized, this knowledge can be used to decrease the realized variance in the experimental data, thereby reducing the measurement-uncertainty penalty from the experiments that must be carried with the augmented model.

2) If a separate characterization of sensor bias uncertainty is not available, or is not deemed to be representative for the current experiments, then reduction cannot be applied to the raw experimental variance. Hence, any variability in the different measurement units adds to the perceived data variability in the experiments, and comprises an experimental-resolution uncertainty penalty that must be carried along with the validated model.

Returning to the example in Figure 1 where only seven data samples exist over the entire temperature range, the following is a simplistic way to construct the larger error bar applicable at any temperature in the range. (More sophisticated and correct approaches take into account confidence intervals from limited sampling, but the added complexity detracts in making the essential point here.) The Least-Squares regression of emissivity rise as a function of temperature serves as a reference mean trend line about which the random deviations of the data can be parameterized by a nominal variance value σ^2. The associated uncertainty is increased, decremented, or unchanged by the measurement uncertainty (small error bars) to arrive at the uncertainty represented by the larger error bar, according to whether the measurement uncertainty is systematic, perfectly random, or somewhere in-between, over the set of experiments.

Note that this makes any numerical nonconvergence bias or uncertainty of the model immaterial to final results except for any non-constant bias under the small perturbations of the input factors. Because such perturbations frequently do precipitate small-scale bias non-uniformity or "noise" in the computed response (see [25]), care should be taken to assure derivative accuracy in view of interaction effects between finite-difference step size, model noise, and model solver tolerances. Whether system behavior is linear enough over the uncertainty ranges of the inputs for the above approximation to be effective can be tested by probing the model over the uncertainty ranges. If nonlinearity is found to be too great, then a Monte Carlo approach to the uncertainty mapping can be taken. This approach avoids the expense of any model runs to ascertain whether the behavior is linear enough, and the expense of forming the derivatives and assuring their numerical accuracy. For validation activities involving more than 3 or 4 input factors, the author finds the Latin-Hypercube Monte Carlo approach to be less expensive regardless of linearity.

A much more involved example is presented in reference [26] that involves several systematic and random experimental uncertainty contributors to the total uncertainty in the failure level of a device. Failure is characterized as a function of two state variables of heating rate and surface area being heated of the device, for use in downstream risk calculations for a fire-heated weapon containing the device. The total uncertainty from the failure characterization experiments (analogous to the large error bars in Figure 1) was mapped cleanly into two parameters μ and σ of the failure threshold model, categorically similar to how the experimental uncertainty depicted by the large error bars in Figure 1 is mapped one-to-one into an uncertainty in the parameter *b* of the linear-regression emissivity model.

The mappings of total experimental uncertainty into parameters of the emissivity and failure models are relatively straightforward because the models have purposefully simple and convenient mathematical forms for accomplishing such mappings. Even the more complex models in this modeling regime, such as turbulence models, have convenient free parameters whose values are determined by what is effectively statistical regression to best fit the relevant data. Hence, it would appear feasible to map (to these regression parameters of the model) the total experimental uncertainty from physical stochastic variance and instrumentation uncertainty in the experiments. In fact, a practical formalism for mapping the total experimental uncertainty into the regression parameters of the model may be derivable from existing Bayesian model calibration techniques (e.g. [18]).

Now consider a different type of model for emissivity as a function of temperature. Instead of a purely calibrated model having free parameters determined from regression to the data and not constrained by physics principles such as conservation laws, consider a model developed from physics principles.[7] To convey a subtle distinction regarding a new term to be introduced in connection with a complete model validation process to be proposed below, presume that the physics-based model (with a nominal set of values for the model inputs) predicts emissivity behavior exactly the same as the regression line in Figure 1. Also assume here that the model results are completely converged numerically.

Unlike regression models—with their malleable structures which facilitate molding to the data, physics-based models have parameters that are estimated *a priori*, from independent experimental characterization (such as discussed in Footnote 7), or from analysis (e.g. molecular dynamics simulations for material properties). These parameters may have significant associated uncertainty. For example, input electromagnetic and/or surface molecular properties in our emissivity model may be quite uncertain. If these uncertainties, which are **intrinsic** to the model, when propagated through the model encompass the large experimental uncertainty interval in Figure 1, then this {model + uncertainty representation} as a set meets the conditions for BEPU prediction, and therefore meets the author's criterion for a validated model. (See [4] for a validation case in this realm that the author recently served as an advisor for.) Of course, these intrinsic uncertainties must be present in downstream use of the model, and are propagated to the associated predictions, along with any other uncertainties specific to the particular application setting.

If the intrinsic uncertainties when propagated through the model do not completely encompass the total experimental uncertainty[8], then additional action is re-

[7] Even models based on physics principles usually involve some degree of calibration, where parameters of the model are empirically set through characterization experiments. For example, a finite-element model of heat diffusion (based on thermal energy conservation and transport principles) through a geometrically complex 3-D stainless steel plate relies partly on the thermal conductivity parameter in the partial-differential-equation diffusion model, Fourier's Law. The stainless-steel conductivity is obtained empirically by iterating for a conductivity value that makes predictions (using a 1-D version of Fourier's Law for the heat transport through a rod specimen of the material) best match 1-D experimental results. Hence, a self-consistent matched set of the partial-differential-equation (PDE) model and its conductivity parameter are arrived at that best replicates the experimental data. The presence of state-variable-dependent material properties, such as temperature-dependent thermal conductivity, signifies that the extent of the PDE model's predictive capability is limited. Since the model does not explicitly include mechanisms for the increased molecular vibration modes and activity at elevated material temperatures, this lack of explicit representation is compensated by recalibrating the value of conductivity that produces the best match at elevated temperatures. The outward appearance is a temperature-dependent material property, which when coupled with the PDE model outwardly appears to be predictive over a range of temperatures. However, this is only true in a post-dictive sense, where a moving calibration was developed over the temperature range to prevent empirical divergence of the model predictions otherwise. Even so, models that explicitly incorporate the important physics principles and constraints in an application would appear to have the best opportunity for predicting well in extrapolation, so they are preferable for such purposes—though this is not necessarily true for interpolation (see e.g. [29]).

[8] This is often the case because the independently characterized or estimated intrinsic uncertainties of the separate elements of the system or subsystem have no recognition of any potential nonlinear interactions and/or stochastic variability present in the assembled sub/system. Nor do the intrinsic uncertainties reflect the experimental uncertainties present in the validation activity at the assembly level. Hence, the so-called "Top-Down" ([30]) assessment at the assembly level is necessary for ascertaining whether the "Bottom-Up" propa-

quired before it can be deemed validated. One possible course of action consists of identifying one or more suitable parameters through which the additional uncertainty in the validation experiment can be mapped into the model—although such mapping is typically not as straightforward as for purely calibrated models.[9] Thus, the mapped uncertainty in the selected parameters, plus the intrinsic parameter uncertainty, propagate through the model together to produce an uncertainty interval which bounds the total uncertainty interval from the validation experiments. Hence, the augmented uncertainty (intrinsic + mapped) is transported to the application setting via uncertainties in various parameters of the model, and is propagated to the associated predictions along with any other uncertainties specific to the particular application setting.

Although illustrated at the material modeling level, the validation reasoning here applies fractally to more complex modeling endeavors.

The author refers to the mapping of experimental uncertainty to the model as "**model conditioning**" with respect to the validation experiments. Note that this is distinguished from model calibration in that the purpose and result of model conditioning is not to bias-correct the model to the data.[10] Nonetheless, the conditioning operation also unavoidably perturbs the model being validated (just as calibration does). Hence, the author observes that *the model validation process often involves model conditioning based on the validation experiments* in order to validate the model for BEPU predictions.

Even when a model's intrinsic uncertainty by itself bounds the total experimental uncertainty (so that no model conditioning is needed), it stands that *a {deterministic model + associated uncertainty representation} are validated together as a complementary set*. Therefore, "model validation" is really validation of the complementary set {model + uncertainty representation} that comprises the augmented model. Previous thinking on model validation does not seem to explicitly promote this idea.

Another method of model conditioning involves transporting the uncertainty in the validation experiments as a separate "uncertainty layer" to the model. This can be as simple as e.g. the blanket prescription that a ±20% uncertainty should be placed on a calculated convection coefficient from a given correlation. Much more involved parameterizations of uncertainty layers for various types of models can be imagined. For example, [13] and [30] appear to provide frameworks for generating a bias-correcting layer to the model, with residual uncertainty included.

In some cases, transportation of the experimental uncertainty is only accomplishable (or is best accomplished) through such a layer or some combination of a separate layer and inverse mapping to inputs of the model. In [25] the magnitude of the experimental uncertainty for a subsystem model was so large due to uncertainty of experimental boundary conditions that the uncertainty could not be sufficiently captured through uncertainties on model inputs applicable at the system modeling level.

CONCLUDING REMARKS

The author sees much reason and value in extending to other modeling regimes what he sees as existing precedent for validated models for BEPU predictions in certain modeling regimes—at least as a necessary although perhaps not sufficient condition for validated models in these other regimes. That is, *a **validated model** is one that (at least) has—mapped into selected parameters of the model and/or carried as a separate "layer" to the model—an uncertainty representation that reasonably bounds the experimentally established uncertainty of system measures pragmatically important at that point in the parameter space.* In this new conception of validation, the {model + uncertainty representation} is validated as a set, rather than the model alone being thought of as the object of validation.

One of the criteria for claiming "use of a validated model in an analysis" is that the model is being used for essentially interpolatory predictions, such as use of a material model or failure model within the state-variable range over which it was properly established. This would also include, for example, more aggregate models such as the 1-D PDE model in Footnote 7 where it's accuracy has been established over the temperature range of calibration, as discussed earlier. However, consider a 3-D version of the PDE model applied to the complex steel plate, for temperatures within the said range. The author does not consider the validated 1-D diffusion model, extrapolated to the 3-D application, to be validated for the 3-D application (even assuming isotropic material properties and accurate geometry modeling in the 3-D setting). Unmodeled phenomena could be masked in the 1-

---

gation of the intrinsic uncertainties encompasses the experimental characterization at the assembly level.

[9] For physics-principles models like finite-element PDE models, the model forms usually imply nonlinear mappings between input parameters and model outputs, and various interactions between parameters (specific to the different model outputs). Therefore, inverse techniques are required to map output uncertainties (i.e., total experimental uncertainty) to uncertainty/ies on model input/s. Hence, for such models it is much more difficult to apply the mapping step. Many considerations are involved in selecting the optimal input parameter or subset of parameters to map the experimental uncertainty to, and in selecting the most appropriate procedure for accomplishing the inverse mapping. In fact, this is a wide-open area of research. More discussion of the issues is given in [29].

[10] Calibration of physics-based models has the potential to be either beneficial or detrimental depending on the choices of calibration parameter(s) and procedure, and the particular extrapolation or interpolation involved, so is not necessarily a model improvement. This is discussed further in Ref. [29].

D calibration setting that could show up as modeling error in the 3-D setting.

Hence, if making extrapolatory predictions, the author does not think it justifiable to claim that a validated model is being used for the predictions. Rigorously, all bets are off in extrapolation. What *is* reasonable is to imply a non-trivial degree of quality control and risk assessment by claiming an '**accredited model'** for the predictions, if it has been validated at proximal points in the parameter space, and in the analyst's experienced opinion, is anticipated to give a reliable or trustworthy result for the issue resolution purposes of the analysis. Thus, validation implies hard direct evidence, but accreditation is an expertise-based belief (a leap of faith) that a model properly validated at proximal point(s) will perform adequately enough in extrapolation to support effective resolution of the issues of interest.

If a model is not validated at point(s) in the parameter space sufficiently near to the extrapolation conditions, it is difficult to defensibly argue that it is accredited for making predictions there. However, since accreditation is a subjective value judgment—an assertion by the modeler/analyst and an acceptance by the modeling customer—the determination depends on how convincing the arguments are. The author does accept that an adequately diverse and credentialed group of experts, adequately peer-reviewed, as in the case of nuclear power plant accident modeling, can deem a model accredited for their modeling purposes—with stated constraints on what the modeling can be used for. For instance, "To be used only for ascertaining which of several accident perturbations appears to be worse, or to determine which model parameters most affect outcomes." Such ordinal ranking purposes are fairly forgiving of model inaccuracies. This is crucial because without validation which enables approximate compensation for model-form error in the extrapolated result, it is not apparent how this error is otherwise estimated and compensated for. Ref. [8] offers detailed considerations and formal procedures concerning model accreditation when actual validation cases are rare. The accreditation process is one of assessing risk and weighing benefits of model use, with associated statements of what the models can reasonably be used for.

As another example, the discontinuous modeling extrapolation from the 1-D rod to the 3-D plate is a case where the author's experience strongly suggests that such an extrapolation is likely to be effective. Hence, the author would be amenable to accrediting the model for the 3-D problem, based on the evidence from the 1-D problem. Although there is undeniable risk here, the author considers this extrapolation risk to be a minimal, and in fact this is the standard practice in most engineering modeling.

Another aspect is that the author might, for instance, accredit a physics-based model for a small extrapolation, while not granting this for a neural network model, even if both match the data reasonably well at the validation points. This might be because the author's experience with the physics-based model suggests that it is likely to extrapolate acceptably to the given conditions, whereas he is personally inexperienced with neural network models, or his experience indicates that they usually don't extrapolate well.

Importantly, even if a model cannot be defensibly claimed to be 'accredited' for a given extrapolatory prediction, this does not necessarily mean that the model should not be used for the prediction. This just means that the risk involved is relatively unmanaged. Nonetheless, empirical indications are that such usage has been successful on balance. The author has seen countless instances where models have been developed in certain settings, and very effectively used in extrapolatory settings where there is a significant discontinuity in the parameter mapping between the two settings. In fact, this is most often the case in the real world, where the empirical evidence is that, although industry has not typically used strictly validated and/or accredited models, the modeling benefits seem to have outweighed the costs and risks, as the use of modeling in industry is currently very popular and continues to grow.

In closing, although we cannot guarantee accuracy of predictions or accompanying uncertainty bands, we can still set about the objective of contextualizing and improving our estimates as well as possible through appropriate quality control procedures. That is, we can attempt to maximize accuracy potential through optimized design of validation experiments and optimized model development, validation, and extrapolation procedures for a given prediction task. Additionally, we can assess modeling risk through InfoGap [3] type analyses like that discussed in [27] to determine the degree to which the models can be incorrect before changing the conclusion obtained with the model. This having been said, quality assessment and control in modeling and simulation is an engineering science still in the very early stages of development, and much needs to be done to bring this young science to a mature state.

BIBLIOGRAPHY

[1] AIAA, 1998, Guide for the Verification and Validation of Computational Fluid Dynamics Simulations, AIAA G-077-1998, American Institute of Aeronautics and Astonautics, Reston, Va.

[2] ASME, V&V 10 - 2006 Guide for Verification and Validation in Computational Solid Mechanics, available at http://catalog.asme.org/Codes/PrintBook/VV_10_2006_Guide_Verification.cfm

[3] Ben-Haim, Y., *Information Gap Decision Theory-Decisions under severe uncertainty*, Academic Press, London, 2001.

[4] Black, A.R., M.L. Hobbs, K.J. Dowding, T.K. Blanchat, "Uncertainty Quantification and Model Validation of Fire/Thermal Response Predictions," to be presented at the 18th AIAA Computational

Fluid Dynamics Conference, 25-28 June 2007, Miami, FL.

[5] Carter, J.N., P.J. Ballester, Z. Tavassoli, P.R. King, "Our Calibrated Model has No Predictive Value: An Example from the Petroleum Industry," presented at the 4th International Conference on Sensitivity Analysis of Model Output (SAMO) Conference, Santa Fe, NM, March 8-11, 2004. paper available at http://library.lanl.gov/ccw/samo2004/.

[6] Coleman, H. W., and Steele, Jr., W. G., 1989, *Experimentation and Uncertainty Analysis for Engineers*, John Wiley & Sons, New York.

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

[8] Department of Defense (2006), Key Concepts of Verification, Verification, & Accreditation, VV&A RPG document dated Sept. 15, 2006.

[9] Dowding, K.J., R.G. Hills, I.H. Leslie, M. Pilch, B.M. Rutherford, M.L. Hobbs, "Case Study for Model Validation: Assessing a Model for Thermal Decomposition of Polyurethane Foam," Sandia National Laboratories report SAND2004-3632, printed September 2004.

[10] Easterling, R. G., "Measuring the Predictive Capability of Computational Models: Principles and Methods, Issues and Illustrations," Sandia National Laboratores report, SAND2001-0243, Unlimited Release, February 2001.

[11] Easterling, R. G., "Measuring Predictive Capability of Computational Models: Foam Degradation Case Study," Foundations for V&V (Computer Model Verification and Validation) in the 21st Century Workshop, October 22-23, 2002.

[12] Easterling, R. G. and Berger, J. O., "Statistical Foundations for the Validation of Computer Models," Foundations for V&V (Computer Model Verification and Validation) in the 21st Century Workshop, October 22-23, 2002.

[13] Hasselman, T., Yap, K., Wathugala, G., Anderson, M.C., "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.

[14] Hazelrigg, G. A., "Thoughts on Model Validation for Engineering Design," Proceedings DETC'03, ASME 2003 Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Chicago, Illinois, September 2-6, 2003.

[15] Hills, R. G. and Trucano, T. G., Statistical Validation of Engineering and Scientific Models: Background, Sandia National Laboratories report, SAND99-1256, Unlimited Release, May, 1999.

[16] Hills, R.G., "Model Validation: Model Parameter and Measurement Uncertainty," *Journal of Heat Transfer*, April 2006, Vol. 128, pp. 339 – 351.

[17] Hodges, J. S., "Six (or so) Things You Can Do with a Bad Model," *Operations Research*, Vol. 39, No. 3, May – June, 1991.

[18] Kennedy, M.C., and O'Hagan, A., "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society Series B - Statistical Methodology*, Vol. 63, No. 3, 2001, pp. 425-464.

[19] Logan, R.W., Nitta, C.K., "Comparing 10 Methods for Solution Verification, and Linking to Model Validation," Journal of Aerospace Computing, Information and Communication, Vol. 3, 2006, pp. 354-373.

[20] Logan, R.W., Nitta, C.K., and Chidester, S. K., "Uncertainty Quantification During Integral Validation: Estimating Parametric, Model Form, and Solution Contributions," 47th AIAA/ASME/ASCE/AHS/ ASC Structures, Structural Dynamics, and Materials Conference, May 1-4, 2006, Newport, RI.

[21] Mayes, R.L, "Validation of a Blast Pressure Loading Model for a Shell-Payload Shock Response Model," paper #251 of the 2006 International Modal Analysis Conference (IMAC XXIV), Jan. 30 - Feb.2, 2006, St. Louis, MO.

[22] Oberkampf, W. L., and Trucano, T. G., 2002, "Verification and Validation in Computational Fluid Dynamics," *Progress in Aerospace Sciences*, 38(3), pp. 209-272.

[23] Oberkampf, W. L., and Barone, M. F., 2004, "Measures of Agreement between Computation and Experiment: Validation Metrics," AIAA 34th Fluid Dynamics Conference, Portland, OR, June 2004.

[24] Roache, P.J., *Verification and Validation in Computational Science and Engineering*, Hermosa Publishers, Albuquerque, NM, 1998.

[25] 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, Seattle, WA, April 16-19, 2001. Updated and extended version available from the author.

[26] 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.

[27] Romero, V.J., "On Model Validation and Extrapolation for Best-Estimate-Plus-Uncertainty Predictions," Sandia National Laboratories document SAND2005-7678C, November 2005.

[28] Romero, V.J., "Some Issues in Quantification of Margins and Uncertainty (QMU) for Phenomenologically Complex Coupled Systems," paper AIAA-2006-1989 presented at the 8th AIAA Non-Deterministic Approaches Conference, May 1-4, 2006, Newport, RI.

[29] Romero, V.J., "Validated Model? Not So Fast. The Need for Model "Conditioning" as an Essential Ad-

dendum to Model Validation," to be submitted for the 9th AIAA Non-Deterministic Approaches Conference, Honolulu, Hawaii, April 23-26, 2007.

[30] Rutherford, B.M., and K.J. Dowding, "An Approach to Model Validation and Model-Based Prediction: Polyurethane Foam Case Study," Sandia National Laboratories report SAND2003-2336, printed July 2003.

[31] Tieszen, S.R., S.P. Domino, A.R. Black, "Validation of a Simple Turbulence Model Suitable for Closure of Temporally-Filtered Navier-Stokes Equations Using a Helium Plume," Sandia National Laboratories report SAND2005-3210, printed June 2005.

[32] Trucano, T.G., M. Pilch, W.L. Oberkampf, "General Concepts for Experimental Validation of ASCI Code Applications," Sandia National Laboratories report SAND2002-0341, printed March 2002.

[33] Trucano, T.G., L.P. Swiler, T. Igusa, W.L. Oberkampf, M. Pilch, "Calibration, validation, and sensitivity analysis: What's what," *Reliability Engrng. and System Safety*,. Vol. 91, No. 11, pp. 1331-1357.

[34] Urbina, A., T.L. Paez, D.O. Smallwood, "A Hierarchy of Validation Measures for Structural Dynamics," paper #181 of the 2006 International Modal Analysis Conference (IMAC XXIV), January 30 – Feb. 2, 2006, St. Louis, MO.