

SANDIA REPORT

SAND2011-7342

Unlimited Release

Printed November 2011

Elements of a Pragmatic Approach for dealing with Bias and Uncertainty in Experiments through Predictions:

- Experiment Design and Data Conditioning
- “Real Space” Model Validation and Conditioning
- Hierarchical Modeling and Extrapolative Prediction

Vicente J. Romero

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.



Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

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

Printed in the United States of America. This report has been reproduced directly from the best available copy.



Elements of a Pragmatic Approach for dealing with Bias and Uncertainty in Experiments through Predictions:

- Experiment Design and Data Conditioning**
- “Real Space” Model Validation and Conditioning**
- Hierarchical Modeling and Extrapolative Prediction**

Vicente Romero
Model Validation and Uncertainty Quantification Dept. 1544
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87195-0828

Abstract

This report explores some important considerations in devising a practical and consistent framework and methodology for utilizing experiments and experimental data to support modeling and prediction. A pragmatic and versatile “Real Space” approach is outlined for confronting experimental and modeling bias and uncertainty to mitigate risk in modeling and prediction. The elements of experiment design and data analysis, data conditioning, model conditioning, model validation, hierarchical modeling, and extrapolative prediction under uncertainty are examined. An appreciation can be gained for the constraints and difficulties at play in devising a viable end-to-end methodology. Rationale is given for the various choices underlying the Real Space end-to-end approach. The approach adopts and refines some elements and constructs from the literature and adds pivotal new elements and constructs. Crucially, the approach reflects a pragmatism and versatility derived from working many industrial-scale problems involving complex physics and constitutive models, steady-state and time-varying nonlinear behavior and boundary conditions, and various types of uncertainty in experiments and models. The framework benefits from a broad exposure to integrated experimental and modeling activities in the areas of heat transfer, solid and structural mechanics, irradiated electronics, and combustion in fluids and solids.

Acknowledgements

The author thanks James Kamm and Tim Trucano of Sandia for their extraordinary reviews of an early version of this report. Three anonymous reviews of Reference [\[50\]](#) are also greatly appreciated. Insights gained from the reviews led to further thought and evaluation in several areas, and reconfiguration of the report to improve the presentation and content. Any remaining shortcomings are solely the responsibility of the author.

Table of Contents

1	Introduction	9
2	Model Validation—What is it?	
	What does it Entail and Imply?	12
2.1	Terminology, Ambiguity, and Various Interpretations of ‘Model’, ‘Model Validation’, and ‘Validated Model’	12
2.2	Definitions of ‘Model’ in Validation and Prediction Contexts	13
2.2.1	Experiment Model and Traveling Model	13
2.2.2	Strongly and Weakly Defined Models	15
2.2.3	Cases where the Discretization Grid is part of the Traveling Model	18
2.3	Accuracy and Adequacy Aspects of Model Validation	19
2.4	A Restatement of the Definition of Model Validation	25
3	Several Fronts of Current Procedural Difficulty in Model Validation, and a Workable Real-Space Approach	29
3.1	Deciding what is an adequate validation basis of comparisons — Scalar vs. Field comparisons, State Variables vs. Resultant Effect, etc.	29
3.2	Deciding on the Formulation or Metric of Discrepancy Characterization	31
3.3	Deciding the Threshold for Model Adequacy	36
3.3.1	Two Potentially Different Validation Settings regarding Accuracy Requirements for Model Adequacy Determination	36
3.3.2	Hierarchically Coupled Adequacy Determination vs. Stand-alone Adequacy Determination	44
4	Projecting Results and Outcomes from the Validation Setting to the Model Intended-Use Setting—Dealing with Extrapolation	45
4.1	Type X Validation Error and Lack of Consequence when Bias and Uncertainty travels “consistently” from Validation to Extrapolation Settings	45
4.2	Absence of Traveling Consistency between Validation and Prediction Settings, associated Extrapolation Risk from Type X Validation Error, and Data Conditioning to Mitigate the Risk	50
5	Design of Model Validation (and Conditioning) Experiments and Scope of the Experiment Model	53
6	Model Acceptance, Endorsement, Accreditation, etc.	55
7	Model Conditioning	57
8	Closing	59
	References	60

Appendices

**Appendix A: Summary Comparison of Real Space Validation Approach vs.
Two Other Established Frameworks69**

List of Figures

Figure 1.	Generic categories of comparison outcomes in the framework's model validation and model conditioning real-space comparisons of uncertainty intervals of experimental versus simulation results for a scalar output quantity.	33
Figure 2.	Material property measurements as a function of temperature, with vertical and horizontal uncertainty bars associated with the measurements, and net uncertainty bounds (dashed lines) for the set of data. At any given temperature it is expected that actual property values will fall within the experimental uncertainty bounds depicted at that temperature. The solid line shows model predictions of the material property value as a function of temperature.	39
Figure 3.	Beam deflection results from the physical beam and from a biased model. (Curves only illustrative.) Low and high limits of possible experimental and model results are shown for an uncertainty range [<i>Lo-Low</i> , <i>Lo-High</i>] in beam length. A deception “risk zone” is shown (see also Figure 4.) where a realization of experimental input and corresponding system response would not lie outside the uncertainty of model response computed with the known input uncertainty range [<i>Lo-Low</i> , <i>Lo-High</i>] in beam length.	47
Figure 4.	High and low deflection values defining ranges of possible experimental and model output consistent with the uncertainty [<i>Lo-Low</i> , <i>Lo-High</i>] in beam length. The high and low response values here are those in Figure 3. A deception risk zone for Type X validation error is shown, where an experimental realization could occur, and because it lies within the uncertainty intervals of the associated model prediction, would provide no indication or warning that the model is biased, even though it actually is.	47
Figure A.1.	Real Space vs. Transform Space Representations of Model Discrepancy	69
Figure A.2.	Real Space vs. Transform Space Discrepancy Metric Support for Extrapolation.....	70
Figure A.3.	ASME V&V-20 Subtractive-Difference Metric limitation for Some Types of Random Variability in Repeated Experiments.....	71
Figure A.4.	Roy & Oberkampf skew toward Model User’s Risk arising from SystematicUncertainties in Experimental Input Conditions	72
Figure A.5.	Summary Table of demonstrated Capabilities and Features of the Compared Model Validation Frameworks.....	73

This page intentionally left blank

1 Introduction

Methodologies for modeling and prediction in the presence of bias and uncertainty are being actively researched and formulated by the modeling and simulation (M&S) community. Comprehensive and detailed frameworks are still elusive. References [1]–[82] highlight various lines of thinking and progress in these areas. Refs. [42]–[52] develop key aspects of the author’s particular paradigm of modeling and prediction summarized in this report. The paradigm and procedures were developed within a larger context of integrated experimental-modeling-analysis programs with end-to-end scope involving experiment design and analysis, data conditioning, model conditioning¹, model validation, hierarchical modeling, and extrapolative prediction under uncertainty.

The majority of this report centers on model validation, although the other elements of the end-to-end process are also addressed. The author’s “Real Space” model validation approach (outlined in sections 3.2 and 3.3 of this report) was arrived at by working backwards from an end objective of “best estimate with uncertainty” (BEWU) modeling and prediction. The approach reflects a pragmatism and versatility derived from working many industrial scale validation problems in the following application areas:

- device thermal response and initiated failure [43], [46]
- device thermal-structural response and failure [51]
- foam thermal pyrolysis and vaporization [28], [33], [53]
- fire and object-response modeling [40], [49]
- radiation-damaged electronic device response [56].

Working from this broad basis, the Real Space model validation (and conditioning) approach has evolved to address uncertainties of random and systematic; correlated and uncorrelated; interval and distributional; and aleatory, epistemic, and combined natures that commonly arise in modeling and experiments, including:

- experimental variability in repeated experiments;
- associated epistemic uncertainty from limited numbers of repeated experiments;
- measurement uncertainties in experimental inputs and outputs;
- uncertainties that arise in data processing and inference from raw experimental and simulation outputs;
- parametric and functional-form uncertainties associated with the model;
- numerical solution uncertainty from model discretization effects.

1. **Model conditioning** is a term signifying a superset of model initialization and corrective adjustment techniques in the presence of uncertainty. Model conditioning includes but extends beyond procedures otherwise variously known as parameter estimation, model calibration, model updating, etc. Model conditioning is the subject of section 7 of this report.

Among the many model validation paradigms, frameworks, and methodologies surveyed in this report, the Real Space formulation has been found by the author to be uniquely capable (adaptable, robust, and practical) to handle the diverse set of challenging validation application problems and attributes cited above. The challenges emanate from:

- analysis and processing of the experimental data and associated interpretation of results;
- model accuracy characterization—deciding on the formulation or metric for characterizing discrepancy between model and experiment results;
- model adequacy characterization—deciding the threshold or criterion for model adequacy (acceptable agreement with reality);
- extrapolation of validation information/results/products—deciding how to project results and outcomes from validation settings to other prediction settings subsequent to the model validation activity. (This item is not part of model validation, but is an essential consideration in the formulation of a useful validation approach.)

The Real Space validation approach differs from others established frameworks in the literature in that model accuracy and adequacy relative to experimental data are not posed in terms of transform measures and acceptance criteria in a discrepancy *transform space*. A large variety of mathematical transforms exist in the literature to characterize discrepancy between experiment and simulation results (e.g., the subtractive difference transform in [6] and [17] and the distribution function ‘area’ validation metric in [20], [54]). The transform measures in the literature can be relatively involved, with varying transparency and interpretability of the physical and decision-making significance of the numerical values yielded by the measures. The transforms can also constrain what forms and types of uncertainty can be handled, as discussed in section 3.2 and Appendix A of this report. Furthermore, as explained in section 3.3.1, workable criteria to demarcate adequacy of model-experiment agreement in transform space remain elusive, whereas a simple and useful criterion can be applied in Real Space.

Besides greater transparency, intuitiveness, and interpretability, the Real Space discrepancy measure better reveals characteristic differences between model and experimental results that affect prediction risk. All transform metrics that the author is familiar with have non-exclusive mappings between real space and transform space; the same transform-space value can accompany different conditions in real space (see Appendix A). Therefore it can be risky to use transform space metrics to make validation judgments on model performance and adequacy, and to guide model conditioning and extrapolation.

These considerations will be central items of discussion in this report. Before presenting the Real Space paradigm in section 3, section 2 considers some fundamental issues and conceptions regarding models and model validation. Attention is given to the appropriate definition of ‘model’ in ‘model validation’. The concepts of *strong* and *weak* models ([65]) are discussed. These are found to be consistent with, but not as useful as, the newer

concepts of *experiment models* and *traveling models* ([45]-[52]), essential concepts which previous model validation literature appears to lack. A traveling model travels to use beyond the validation or conditioning activity, and is a subset of the larger model of the experiments. The traveling model and experiment model signify the two different modeling scopes that must be distinguished and treated appropriately in model validation and conditioning activities; uncertainties in the traveling model are handled differently than non-traveling uncertainties in the Real Space methodology.

Section 4 of this report describes consistent (traveling) and non-consistent (non-traveling) bias and uncertainty in relation to extrapolation risk in going from model validation or conditioning settings to model application settings. The affiliated concept of *Type X* error in model validation or conditioning is discussed. *Data conditioning* procedures ([47]-[49]) that can effectively mitigate prediction risk associated with Type X error are also discussed. Connections are drawn between these new concepts to the literature and prior concepts of Type I and II errors and Model Builder's risk and Model User's risk. Section 5 considers how optimized design of model validation and conditioning experiments can reduce bias uncertainty and risk in extrapolation. Section 6 on model Endorsement, Accreditation, etc. considers the more subjective realm of deciding on the acceptability of extrapolation risk for particular model uses away from the validation or conditioning points in the modeling space. Section 7 summarizes important distinctions between various objectives, approaches, and methodologies that comprise the model initialization and corrective adjustment techniques of model conditioning. Some closing remarks are made in Section 8.

2 Model Validation—What is it? What does it Entail and Imply?

2.1 Terminology, Ambiguity, and Various Interpretations of ‘Model’, ‘Model Validation’, and ‘Validated Model’

There seems to be fairly uniform agreement in the validation community on what model validation implies at a vague conceptual level, i.e., at the level of the following brief contemporary definition: *Model Validation is the process of determining the degree to which a computer model is an accurate representation of the real world from the perspective of an intended use of the model.* This statement and close variants of it are the accepted definition in the recognizing organizations [1] – [6] (there may be others).

This statement is a concise expression of some fundamental notional aspects of model validation. However, despite broad agreement on this notional definition, at a detailed interpretational and implementational level there is room within the definition for considerable debate among validation methodology developers and practitioners concerning the specific procedures, steps, and products of model validation. As a result, phrases like “the model has been validated” can mean very different things to different people even within the model validation community.

Various conceptions of model validation are discussed below. A large variety of paradigms produce profound differences in the validation outlook, approach, procedures, and end products—including validation criteria, results interpretation and usage, and ultimately how the validation activity supports the objective of BEWU extrapolative prediction beyond the validation setting. For example, the definition statement above is originally from the DoD document [1], arising largely from a tradition of operations research and systems simulation. The statement was adopted consecutively in [2]-[6] essentially without change. However, the nature of the models and decision scopes pertinent in [1] are largely different from those in [2]-[4], [6], and perhaps [5]. The models addressed in [1] vary from discrete-event and agent type models, e.g. for supply chain logistics, battlefield simulations, and theatre defense simulations, to computational physics type models for simulation of fluid mechanics, heat transfer, and structural mechanics phenomena, etc. The type of models addressed in [2]-[4], [6], and to large measure in [5] are computational physics type models. Interpretation of the “standard” definition statement above is also substantially different in the DoD validation community versus the others cited. According to personal discussions with several validation practitioners in DoD, validation includes not only the process of assessing model performance vs. empirical data (if suitable data exists), but also the subjective process of adjudicating whether it *appears* that the model will perform acceptably for specified *intended uses* outside the conditions where the model has been calibrated or tested against data. In contrast, the author perceives that [2]-[6] do not consider validation to involve a process for judging whether the model will extrapolate acceptably; this is seen as a largely subjective and situation-dependent endeavor that is a different element of the M&S process (e.g., BEWU extrapolative prediction, model accreditation, etc.).

Other ambiguities in the standard definition are discussed in the following. Other validation definitions from the literature are considered. These are leveraged to offer a new definition intended to clear up ambiguities in relation to validation of computational physics models (according to the author’s paradigm). Other terminology refinements are introduced to capture what is meant in this report by ‘model’, ‘model validation’, and ‘validated model’.

2.2 Definitions of ‘Model’ in Validation and Prediction Contexts

It is important to identify what model or set of models is being validated in a validation activity. Often, validation frameworks and activities are ambiguous in this respect, which can lead to confused interpretation and improper usage of validation results, as well as improper accounting steps regarding uncertainties and their propagation to predictions.

Consider a 1-D heat conduction experiment involving a heated rod. If a finite-element model is built to simulate the experiment, how does one target or differentiate whether the 1-D heat diffusion equation (partial differential equation, PDE) alone is being validated, or the validation applies to the larger set of models (equations and parameter values/ranges) consisting of the PDE and the geometry, material property, and boundary condition descriptions? The ‘model’ to be validated could also potentially include the affiliated discretization scheme and solution algorithm (see section 2.2.3 for a discussion).

Accordingly, it is necessary to be specific about which model or set of models is the focus of a validation exercise. This is important for planning and performing the validation activity and interpreting and using the results. The author’s mechanism for defining the boundary of the ‘model’ to be validated is explained next as the “*traveling model*” ([47]).

2.2.1 “Experiment Model” and “Traveling Model”

Here we provide a more precise definition of ‘model’ by recognizing two different modeling scopes in model validation, conditioning, and extrapolation.

The notion of a *traveling model* is employed to delineate the set of models in a validation or conditioning activity that will be used (as a set) in subsequent predictions. This is a subset of the larger set of models employed in the validation activity, referred to as the *experiment model* or *e-model*. This larger set of models cannot be avoided. No matter what the model of traveling interest is—whether a set of PDEs, a material behavior model, a model of a hardware device, a process model, etc.—the experiment will have aspects that need to be modeled that are *auxiliary* to the traveling model of interest.

For example, consider an electromechanical device with specific behaviors of interest for which a model is being validated. The physical device is part of a larger hardware system. A finite-element (FE) model of the device is to be validated under certain loading conditions in validation experiments. The model is then to be used in a hierarchical system-level model to make predictions of device response under *other* loads that the device may experience within the assembled system subjected to loading at the system

level. Here, the hardware device in the validation experiment is the traveling physical system, the FE model is the traveling model, and the applied loads in the validation experiment are not part of the traveling system or model. Here the traveling model includes the PDE equation sets and the mathematical descriptions of the device geometry and material properties.

In contrast to the concrete image that a traveling hardware device and FE model would have, the traveling physical system and model might not have a definite image. This occurs when the physical system of traveling interest is a phenomenological behavior such as material property behavior, turbulence, etc. Here the traveling model is a phenomenological equation set with associated parameter values/ranges, but without a specific traveling geometrical-object form.

Such *amorphous* phenomenological models can only be validated through some embodiment or instantiation of a particular physical problem. For example, to develop or validate a constitutive equation set for elastic-plastic behavior of a certain material, any associated experiments will involve a particular specimen of material, shape/geometry, initial state, and loading. A corresponding e-model is built of the physical instantiation. Any e-model aspects such as applied loading and specimen shape/geometry that are not part of a physical setting for which the constitutive model is later used are outside the traveling model of interest.

Although only a certain portion of an experiment model travels to application/prediction settings away from the validation experiments, the uncertainty that issues with the traveling model is often accumulated from various uncertainties in the experiments, including those from the non-traveling aspects. Uncertainties in the non-traveling aspects depend on the design and implementation of the experiments; the diagnostic instrumentation used; and the chosen scope of the e-model, all of which should be optimized to reduce the non-traveling uncertainty as much as achievable within project constraints (see section 5 for elaboration). It is also important to note that uncertainties in the experiments are treated differently in the Real Space framework if the uncertainties are affiliated with traveling aspects versus non-traveling aspects ([49], [51], [52]).

The traveling model can contain uncertainties that are inherently affiliated with the traveling model, such as an uncertainty range on parameter values in a turbulence or material model. These uncertainties are defined prior to the current experiment (they are not determined by or in the experiment) and come to the experiment model as *a priori* uncertainties in model form and/or parameter values. See [28], [40], [46], [49], [51], [52], and [56] for illustration of how the framework handles these traveling “*model-intrinsic*” uncertainties in model validation and conditioning.

The model that travels from a validation or conditioning activity could even be the representation of the loads or boundary conditions in an experiment. One may want to develop or validate a model of an experimental facility’s applied loading or excitation, e.g. the spatial/temporal radiation intensity profile in a testing chamber. The model would be

of continuing interest for use in predicting device response to simulated radiation shots in the test chamber.

The following is a poignant example where the traveling model concept would have been helpful in an eventful exchange ([64] - [66]) in the groundwater modeling literature regarding the boundaries of the ‘model’ being validated. Konikow & Bredehoeft [64] give an example where a calibrated groundwater flow model under 10 years of precipitation conditions was used to predict the next 10 years of ground-water flow, assuming that precipitation in the calibration period would be representative of the future. The actual precipitation (stream recharge) happened to be substantially different in the following 10 years, contributing to disagreement with the actual ground-water flow measured in that period. De Marsily et al. [65] point out that Konikow & Bedeloft are “...mixing apples and oranges...In no way has any groundwater flow model claimed that it can predict climate variability [precipitation and stream recharge] in the future...the ten years following the model calibration are more humid than those used for the calibration, why should this in any way invalidate the groundwater flow model?”

In the vernacular of the present report, Konikow & Bredehoeft include the model inputs of stream recharge rates as being in their traveling model; the recharge rates are included as part of the model from the 10-year calibration period that travels to the new prediction setting and is used to predict behavior in the next 10 years. De Marsily et al. argue that recharge rates should not be part of the model on trial; that recharge should not be assumed to remain consistent over the next 10 years. Instead, a variety of recharge scenarios should be run in the predictions to reflect the uncertainty (historic climatic variability) of future recharge in the next 10 years. Whether it was reasonable for Konikow & Bredehoeft to include the recent 10-yr. history of stream recharge rates in their traveling model is for the groundwater community to discuss and decide. Such a discussion would be focused and facilitated by the concept and terminology of the ‘traveling model’. In general, a precisely defined traveling model should be a routine aspect of the thinking in model conditioning, validation, and prediction.

It is the traveling model that is the subject model of a validation or conditioning activity. The traveling model carries an implicit connotation that the physical attributes and phenomena it represents are proposed to carry-over consistently to the new prediction settings, such that the traveling model applies there as well. Hence, Konikow & Bredehoeft were implicitly proposing that the stream recharge rates measured in the previous 10 years would remain stable over the next 10 years. This didn’t turn out to be the case, although the geology was probably stable over the time period and hence that part of the model would carry over well to the new prediction setting.

2.2.2 Strongly and Weakly Defined Models

A related contribution of interest was put forward by Leijnse & Hassanizadeh [66]. They observed that “there is no unanimity within the groundwater modeling community on what constitutes a model.” They further observed that much of the difference of opinion between De Marsily et al. and Konikow & Bredehoeft was that ‘model’ and ‘validation/

validated' were defined differently in the two view points. The present report recognizes their different definitions of 'model' as a difference in their scopes for the traveling model, i.e., where the boundaries of the traveling model were drawn.

Leijnse & Hassanizadeh go on to introduce the concept of '*weakly defined*' models like Darcy's law and the Navier-Stokes equations. Such governing equations are considered weakly defined in that they do not involve set parameter values or boundary conditions, geometry, etc. When an instantiation of the equations of a weakly defined model occurs for a particular application, specific parameter values, boundary conditions, geometry, etc. become defined and the collection comprises a '*strongly defined*' model in the Leijnse & Hassanizadeh terminology. They observe that models can exist anywhere in the spectrum between weak and strong extremes, with most models in practice involving both strongly defined elements (deterministic parameter values and model forms) and weakly defined elements (where parameter values and model forms are varied to express uncertainty).

They argue that the rift between the view of Konikow & Bredehoeft that groundwater models cannot be validated, and the opposite view of De Marsily et al., is explainable by the fact that the two camps were talking about validation of models at very different places on the weak/strong spectrum. Leijnse & Hassanizadeh contend that it is easier for De Marsily et al. to validate their more weakly defined model (using measured precipitation for input to the model) than for Konikow & Bredehoeft to validate their more strongly defined model (involving stream recharge rates modeled as those from the previous 10 year period).

Leijnse & Hassanizadeh generalize that it is easier to validate weakly defined models than strongly defined ones. A different view is held in the current report. The difficulty that Konikow & Bredehoeft experienced was not necessarily that they defined their future recharge rates strongly, it was that the proposed recharge rates were inaccurate. If the strongly defined recharge rates were accurate, the model's prediction accuracy would have been much better. Conversely, it is not necessarily true that a more weakly defined model would have been easier to validate. Consider if the model was more weakly defined by modeling the future recharge rates as uncertain, say with a range $\pm x\%$ about the strongly defined rate $R(\text{time})$. Then at one end of the uncertainty range, say $(1 + x\%)R(\text{time})$, the model output results would be closer to the actual measured 10-yr. data. But at the opposite extreme of the uncertainty range, $(1 - x\%)R(\text{time})$, the model predictions would presumably be *further* from the data. How does one deal with this new complication presented by the weaker model? A detailed discussion is given in sections 3.2 and 3.3 in connection with Figure 1. Anyhow, it is not necessarily easier to validate weaker models than stronger ones.

More broadly, the following views are taken in this report:

1. Completely strong models with no uncertainty in their parameter values and/or model form cannot be validated to capture reality.
2. Weaker models that explicitly incorporate uncertainty in their parameter values and/or model form are eligible to possibly capture reality, and whether they do can

be reasonably objectively determined and the model thereby substantiated or refuted as described in section 3.3.

3. Substantiating models defined in the weak limit (generic governing equation sets like Darcy's law) is vastly more difficult than substantiating a model in circumstance 2 above. This is because substantiation of a generic equation set proposed to govern over a large range of physical instantiations requires non-rejection over an immense number of particular instantiations of the weak model under diverse parameter values, geometries, boundary conditions, etc.² In fact, the testing never stops with regard to laws and theories. Darcy's law is said to hold, period (not "to hold within $\pm 10\%$ " as might be the relaxed standard of acceptance for a more strongly defined application model).

For reasons associated with item 3, it is not realistic to speak of *code* validation because that would imply validation over all possible (infinite) particular instantiations of the code's equations. Rather, we speak of instantiation-specific *model* validation.

In the spirit of what Leijnse & Hassanizadeh may intend, the present author, as a long-time modeler and non-purist, finds value in loose or "*anecdotal*" corroboration of weak models by way of a broad empirical basis of loosely quantitative affirming experiences with the model. References [41] and [6, Appendix C-9] recognize consensus community acceptance as a legitimate means of substantiating a model. Still, all seek a more rigorous approach to model substantiation for application-specific models not at the weak limit.

The assessment in this paper is that the rift between De Marsily et al. and Konikow & Bredehoeft viewpoints is less an issue of where their respective models lie on the weak-strong spectrum, and more aligned with what De Marsily et al. cite: A) a difference in their strictness of definitions of Validation; and B) their already-discussed difference on viable scope of the traveling model—whether the past 10 yr. record of recharge rates should have been proposed to represent the future 10 years. Regarding their definitions of validation, in the vernacular of the present report Konikow & Bredehoeft use a strict internal-correctness criterion for model validation, while De Marsily et al. use a more relaxed

-
2. Popper ([39]) argues that a theory cannot be proven true because at some point evidence can come along to overturn the theory. This has happened repeatedly in history. The best that can be done is to continually subject the theory to as many diverse tests as possible and if/while it survives the tests, not reject the theory. A related issue is that non-rejection of a hypothesis in individual tests does not imply trueness of the hypothesis because noisy and uncertain experimental data can leave much room for bias error to go undetected (see Section 3.3.1-Validation Setting I). However, the odds that the theory is valid would appear to increase as non-rejection evidence mounts over a large number of diverse tests; non-rejection in individual cases is not convincing, but taken together the assembly of evidence becomes convincing—though not absolute proof. It is conceivable that statistical arguments can be made that non-rejection in a large number of diverse cases statistically averages out the bias "parallax" in the individual tests such that a probability can be estimated that the law/theory/model is valid. However, the author is aware of no such methodology applied to positively substantiate (within estimated statistical uncertainty) weak models such as Darcy's law or the Navier-Stokes equations. Their broad acceptance is assumed to be based on less rigorous anecdotal evidence, although a lot of it.

definition that concerns model effectiveness—whether the model, though unavoidably erred to some degree, is good enough to be useful for particular analysis and forecasting needs. These two viewpoints will be discussed in Section 2.3-Part I.

The current report largely takes the De Marsily et al. view, a widely held view in the validation community. However this report takes a different view than expressed in the last paragraph of De Marsily et al., “...we do not validate our models...we try to show that they are not invalidated by the data.” There is a nuanced problem with this view as discussed at the end of section 2.3-Part I and illustrated in section 3.3.1.

2.2.3 Cases where the Discretization Grid is part of the Traveling Model

As pointed out by Roache in the section “Does *Model* Include the Grid?” in [41], mesh discretization dependence is built into the modeling scheme for many types of geophysics models, such as groundwater flow, climate, ocean, and weather models. Physics subgrid models are tuned to the particular grid spacing being used, and solution results are produced on that grid. There is no notion similar to the one in engineering computational mechanics (e.g. [2], [4], [6]) of refining grid size to obtain solutions with negligible error relative to the exact solution of the governing PDEs, or of generating several successively refined solutions from which convergence behavior can be used to estimate solution error by various “calculation verification” or “solution verification” methods (see e.g. [38], [41]).

For geophysics models where the grid and grid-dependent subgrid models travel as a coupled set from the tuning setting to validation or other prediction settings, the grid is considered in this report to be an integral, and traveling, aspect of these models. For engineering computational mechanics models, in contrast, current modeling strategies the author is aware of do not involve explicit preservation of mesh spacing in going from validation or conditioning settings to new prediction settings. Grid related solution error is assumed to be appreciably different in both settings and is to be estimated within quantified uncertainty by calculation verification procedures. Associated grid refinement studies are conducted to levels separately feasible in each settings, so a particular grid does not usually travel between the two settings (grid is not part of the traveling model). For industrial scale models, grid refinement procedures are often difficult and computationally expensive, frequently prohibitively so. Even if estimates of solution error and associated uncertainty can be obtained, they are often suspect because solutions that can be afforded rarely exhibit the convergence smoothness or other properties that current calculation verification techniques require for accuracy. For more background see ([13], [58], [89]-[99]).

Even in theory, convergence to exact solution of the governing PDEs is undermined or precluded in engineering models that involve grid-dependent submodels like turbulence models and material behavior and failure models. Such models may only enable solution accuracy for grid sizes (and even discretization element types and stress states) that the submodel was developed at or tuned to; spatial filtering scales in turbulence models can interact with grid discretization scales when discretization in application settings becomes too coarse or fine. Changes in the problem geometry itself can also occur with grid

refinement (e.g. the fractal geometry problem in [15]). For the many reasons cited it seems worthwhile to further investigate paradigms where mesh dependence of solutions is inherent to the modeling and prediction schemes, i.e. where mesh dependency is a part of the traveling model. In this vein, strategies from the geophysics modeling communities may be helpful for dealing with discretization dependency (solution bias) in computational mechanics modeling and prediction.

2.3 Accuracy and Adequacy Aspects of Model Validation

Part 1: ‘accuracy’ pertains to the model’s mapping of inputs to output results, not to accuracy/correctness of the model’s internal representations of reality

In the standard definition statement the following words “...*degree to which the model is an accurate representation of the real world...*” can engender a misconception of what model validation is capable of ascertaining. Without going too deeply into scientific philosophy and logic in the following discussion, it is necessary to be aware of the fundamental constraints on what model validation can actually establish.

An important aspect of the present framework is the recognition of the basic nature of models and modeling. A concise and popular expression that contrasts model validation from code verification is the following: model validation is concerned with “*solving the right governing equations*” whereas code verification is concerned with “*solving the governing equations right*.” This phraseology was first used in [13], cited as a computational fluid dynamics adaptation of the software engineering phraseology in [83]: “*Validation is doing the right job and verification is doing the job right*.” At one level the succinct statement concerning model validation is useful, but it requires further refinement when thinking at a deeper level about models, modeling, and model validation. Because models are mathematical abstractions of physical properties, processes, and behaviors of real systems, it is not proper or productive to approach model validation as an activity in assessing or judging how “right” or “correct” in a strict sense particular modeling equations and parameters are.

For example, a homogeneous material property treatment of mass density in a continuum description of a steel plate might be very effective in modeling certain thermal or mechanical behaviors. Nonetheless, this does not correspond to equation/parameter correctness; the steel material is in reality mostly vacant space with mass concentrated at numerous discrete locations (atoms). Although clearly not correct, the model representation is effective in this case. Would turbulence modelers claim that their equations model the rich phenomena in turbulence in a mechanistically correct or right manner? What about crack-propagation models, or thermal contact resistance models, or convection correlation models? Hence, when equations/models seem to work well in particular circumstances, ‘model effectiveness’ seems the better way to think about and term things than ‘model correctness’ does. Zeigler ([72]) frames this as ‘predictively valid’ models versus ‘structurally valid’ models. In the latter the model structure agrees with the internal workings of the real system. Predictively valid models predict well even though they may not faithfully represent all the internal workings of the system.

Continuum mechanics partial differential equations (PDEs), as the foundation of the engineering models of interest in this report, while very effective for certain modeling tasks and purposes, are not correct in a strict sense. These simplifications of reality are only superficially representative of the underlying systems and phenomenology. They do not and cannot represent the full rich physics in real systems.

The words ‘right’ and ‘correct’ also convey a notion of model uniqueness. The implication is that a uniquely right or correct mathematical model/basis exists that describes the underlying phenomena in some system for which a model is being validated. However, even if the model exactly matches the output data of the system it cannot logically be concluded that the present model is the only one capable of this. Other models that are perturbations of the said model, or even dramatically different from it, might also be able to produce exact matches to the experimental data. For example, in parameter estimation and model calibration one must deal with the usual circumstance that non-unique combinations of parameter values give the same output results of the model. There are even numerous non-unique model paradigms and structures that yield effectively the same solutions in regimes of overlap, e.g. Lagrangian vs. Eulerian formulations of transport phenomena, Newtonian Mechanics vs. Einstein Relativistic Mechanics, and Navier-Stokes vs. Euler vs. Potential Flow.

Therefore, even if a model exactly matches the output data of a validation experiment, it cannot be concluded that the model is uniquely the right one; that the equations/parameter values are the right ones; or that the model is correct in a strict sense.

For even our most revered and fundamental continuum-level behavioral models (PDEs) and simple geometries and boundary conditions, matching the behavior of real systems requires adjustable parameter values to make up for the limited causal representation in the models.³ For more complex models with many governing equations, materials/properties, complex geometry, multiple parts and pieces with boundary interactions like thermal contact resistance and frictional contact, and coupled nonlinear phenomenological interactions, it is expected that *a priori* parameter values input to the

3. Parameters like thermal conductivity, modulus of elasticity, and viscosity in the governing equations are state-variable-dependent (e.g., temperature-dependent conductivity). This signifies that the extent of the modeling scheme's predictive capability is limited. For example, Fourier's Law of heat diffusion does not explicitly include terms for the increased molecular vibration at elevated material temperatures, which normally increases the effective thermal conductivity. This lack of explicit representation in the model is compensated by inverse-calculating the conductivity values that produce the best match to experimental results at various temperatures. The outward appearance is a temperature-dependent material property (thermal conductivity) which, when coupled with Fourier's Law outwardly appears to be predictive over a range of temperatures. However, this is only true in a post-dictive sense, where a temperature dependence relationship for conductivity is developed over the temperature range to prevent empirical divergence of the model predictions otherwise. Even so, models that explicitly incorporate the essential physics mechanisms and principles would appear to have the best opportunity for predicting well in extrapolation, so would be preferable for extrapolative predictions purposes—although this is not necessarily true for interpolative prediction purposes, where physics-free statistical models may do better if sufficient data is available for these models.

model cannot possibly make up for the various model-form deficiencies. This reflects back to point 1 in section 2.2.2. An important implication is that it seems therefore inappropriate to propose validation as a determination of whether model results are different from experimental results. The expectation is that they will be different. The relevant issues are how much different and whether the difference is acceptable for an envisioned use purpose of the model, such as scoping and sensitivity analysis, design and decision-making support, and assessment of system performance and/or safety margins relative to stated requirements.

If a validation assessment finds that model output results are very close to experimental results (for now assuming exact solution of the equations and experiments with no error or uncertainty), the hope is that this is the result of negligible errors in all the aspects of the model. Unfortunately this cannot be strongly concluded. Numerous possibilities exist where relatively large errors occur in the various modeling aspects but the errors offset each other to a large degree. In fact, the probability of “right for the wrong reasons” would appear to be substantially greater than the probability of “right for the right reasons” in situations where complex models have not been assembled incrementally with careful experimental validation as each new significant element of the model is added.

However, such incremental validation cannot occur with more granularity than physical and experimental factors allow. This limits the granularity with which “right for the right reasons” can be established. It can be difficult or impossible to separate strongly comingled physics for individual validation of the contributing elements. For example it is not possible experimentally to build up to a fire (as the full system of interest) by incrementally adding combustion, then turbulence, then soot generation/agglomeration, then radiative participation, etc. Here one cannot claim that agreement of model and experimental results at the full system level is “right for the right reasons”. Nonetheless, model credibility in this respect can be alternatively built in the spirit of Footnote 2. by investigating a diverse set of conditions/scenarios in a validation matrix and showing that the model performs well (robustly) over the diverse tests.

Another type of granularity constraint in engineering projects is that time and resource constraints present practical restrictions which almost always limit hierarchical decomposition to something much coarser than what a detailed in-depth understanding of the system would require. As a practical matter it is often not necessary or advantageous to pursue understanding of system internal workings beyond a certain level of granularity. Beyond that level, characterizing what’s “under the hood” becomes counter-productive from cost and unnecessary-complexity standpoints.

In general, a validation activity at a given level of modeling can only be used to make assertions on the predictivity of the tested model at that level, but not to make statements on the accuracy or correctness of the individual elements of the model. Any statements on the accuracy of the underlying elements must come from probing and characterizing them individually.

The validation concerns brought up in this section are prevalent in the natural sciences modeling community and are cogently discussed in e.g. [64] – [71]. It is not possible with natural systems (geological, biological, ecological, climatological, astrophysical, etc.) to pursue incremental system construction and characterization accompanied by incremental hierarchical model validation. Much of the thinking from those modeling realms is directly applicable to engineering validation where incremental model validation is not or cannot be pursued (e.g. modeling of as-built and in-place structures and systems).

The author agrees with the sentiment in [64] – [71] that model uniqueness and internal causal correctness are not definitively establishable via validation or otherwise. The author also agrees with the majority sentiment in [64] – [71] that ‘absolute validation’ of internal causal correctness is not what model validation is intended to establish anyway (contrary to some views in that community, e.g. [64] and [68]). While it might be that absolute causal correctness of a model (if even a theoretical possibility) could ensure accurate extrapolative prediction beyond the validation conditions, this does not require that absolute validation of model causal correctness is the goal. A more practical and productive goal is to assess whether the model, even though admittedly imperfect, predicts acceptably well in significant extrapolative tests relevant to specific use purposes of the model.

Hence the wording of the standard definition statement is somewhat troublesome in that the following interpretation in brackets [~] could easily be taken: “...*determining the degree to which a computer model is an accurate [internal causal] representation of the real world...*” The following early (1979) consensus definition of model validation seems less troublesome in this regard, although it has other weaknesses as will be discussed. From the SCS Technical Committee⁴ on Model Credibility of the Society for Modeling and Simulation International [57]:

“Model Validation is the substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model.”

Close variants of this definition have been used for over three decades in the significant body of V&V work by Sargent and Balci (e.g. [75]-[82]). References [75]-[78] have valuable surveys of pre-1980 concepts and definitions of verification and validation, many of which have survived the test of time and are relevant today. These references provide a good compliment to the current report’s survey of more recent validation literature.

4. It is noteworthy that the SCS Committee was “...composed of members from diverse disciplines and backgrounds, with the intent that it [the committee’s set of definitions] could be employed in all types of simulation applications. Great care was taken to develop the definition which would be equally applicable to simulations of physical systems that embody readily measurable phenomena, and social and biological systems for which data may be ill-defined...” ([57]). Representation on the committee came from Europe, Canada, and the US, spanning engineering (aerospace, electrical, and mechanical), maritime studies, natural resources studies, Dept. of Agriculture, Dept. of Defense, and the US Air Force Human Resources Lab.

Some important points from the SCS definition are inferred below for discussion.

1. ‘*range of accuracy*’ is read to imply that model predictions are regarded to have some error/uncertainty.
2. ‘*satisfactory range of accuracy*’ connotes that discernment of the adequacy or acceptability of the model’s accuracy is a key aspect of model validation.
3. ‘*Model Validation is the substantiation that*’ connotes that model validation entails more than a process of assessing model accuracy and adequacy. It entails making an evidence-based statement that the model’s accuracy has been found to be satisfactory. Thus, the term ‘validated model’ indicates satisfactory model quality; the model has been substantiated to be adequate, not merely assessed for accuracy and adequacy by applying a model validation assessment process (the outcome of which could be that the model is found inadequate). In contrast, the standard definition statement in [1]-[6] only mentions “...a process of [assessing]...”

Perhaps because one cannot establish that a model has internal causal correctness even if its results match those of the real system, some in the V&V community contend that the most positive statement that can be made regarding substantiation of a model is that “The model was not invalidated.” In other words, “We cannot definitively establish that the model is correct, but neither did the validation results show definitively that the model was flawed.” However, using this as a basis for model corroboration will accept arbitrarily biased models within the uncertainty in the validation exercise (see section 3.3.1-Validation Setting 1, the ‘not inconsistent’ criterion). The not-invalidated paradigm for model validation is conceivably applicable if many non-refuting results are obtained over a diversity of validation instantiations (see Footnote 2.), but the norm with engineering application models is that validation assessments are very limited in the number of model instantiations investigated.

Besides this technical argument, the not-invalidated paradigm is philosophically inconsistent with how validation of application models is more reasonably viewed and posed. The concept that one can only establish that a model is not invalidated derives ([68]) from Popper’s arguments ([39]) that a theory cannot be proven true by evidence that supports the theory, but can only be proven false (if/when any evidence comes along to overturn the theory). Thus, the terminology ‘model not invalidated’ has a strong natural association with the adjudication system for scientific theories. The natural connotation is therefore that the model is being proposed as something capable of being causally correct and that a non-refuting validation result is something that does not prove the model correct, but also does not disprove it. While it is true that a non-refuting result does not disprove the model, it is also true that for the vast majority of engineering models that would be the subject of validation in practice, these would not be proposed to be unflawed representations of reality in the first place. Rather, the hope would be to show that the model, as an approximate representation of the real system, is effective in predicting results that match experimental results to within acceptable error tolerances for particular engineering needs. One would not similarly talk about acceptable error tolerances of $\pm X\%$

regarding the trueness of a theory, in the realm where the concept and language of ‘not invalidated’ comes from.

Indeed, *the applicable validation question is whether the model is “good enough” for a particular use purpose.* The objective in validation is to ascertain the answer with reasonable definitiveness and evidentiary basis.

Part 2: Accuracy and Adequacy Measures and Objectives in Model Validation

In the standard definition statement of model validation there is substantial ambiguity regarding accuracy and adequacy determination of the model’s predictive performance. One could reasonably conclude from the definition statement that model validation involves just the quantification of model *accuracy* versus reality, for various experiment and model output quantities. From the definition statement it is fair to say, “We compared the model results against experimental data and found a discrepancy of x (in some particular error measure). Therefore we have validated the model. The model has been validated.”

This accuracy-quantification-only viewpoint is not at odds with the wording of the definition statement, and some validation frameworks stop with accuracy quantification only (e.g. [6], [54]). Nonetheless it seems evident that pronouncements like “the model has been validated” should connote that some reasonable standard of model quality/usefulness/acceptability has been met with respect to anticipated application needs. This cannot be established by simply quantifying model discrepancy. Hence it seems that a complete validation methodology should address model adequacy as well.

Indeed the question of model adequacy is so central to what is of concern to users and customers of model validation work that determination of model adequacy (not just quantification of model accuracy) has been recognized to be a central element of model validation in the ASME standard [4] (work is underway on a supplemental document [7] that proposes some approaches to model adequacy determination). As a practical matter, it is the author’s experience that project funding for model validation work usually comes with expectations beyond just quantification of model discrepancy. The real deliverables sought are quantitative evidence and analysis supporting a decision regarding model adequacy for an intended use purpose. Certainly this is the perspective of DoD ([1]) and many in the DoE ([3]) and perhaps NASA as well ([5]).

The practical difficulties of adjudicating model adequacy for extrapolatory uses of the model are daunting. The difficulties are reflected in a dearth of methodology in the literature. Nonetheless, some progress has been made. This is the subject of section 3.3 of this report.

Besides the recognizing bodies [1], [3], [4] that view adequacy assessment as being an essential element of model validation, this is also central to the view of the geohydrological community ([71]). Older works like the SCS standard [57] and the notable 1988 work of Miser & Quade [34] specifically emphasize model adequacy. From Miser & Quade:

“Validation is the process by which the analyst assures himself and others that a model is a representation of the phenomena being modeled that is adequate for the purposes of the study of which it is a part.”

2.4 A Restatement of the Definition of Model Validation intended to clarify things

The early definition statements of model validation by SCS and Miser & Quade have the common-language connotation that model validation implies a positive, successful, or desirable outcome concerning model adequacy: a substantiation of model predictive quality. The more recent standard definition statement does not specifically make this tie. Thus, varying interpretations of what model validation entails are found among [1] – [6] and others subscribing to the standard definition statement. The newer and older definitions convey some validation elements in common, but also have some disjoint elements. None clearly convey all essential elements, and all have some problematic wording that can lead to troublesome interpretations.

For instance consider the following terminology problem. The standard definition statement and the Miser & Quade definition can be construed to define that validation is just the *process* of assessment because they respectively state ‘...the *process* of...’ and ‘...the *process* by which...’. However, if model validation is just an assessment *process* it is fair to say, “We applied a model validation process of accuracy and adequacy assessment. We validated the model by this process.” This can be said whether the model is found to be adequate or not. “We validated the model to be inadequate” is even fair to say. Certainly this usage is not desirable or transparent from a common-language standpoint or from a standpoint that promotes reliable communication about model quality. Clearly, model validation should entail more than the said processes—it also involves making meaningful characterizations and judgments about the model itself, i.e. about its predictive performance. Validation processes must produce meaningful products, ultimately model accuracy and adequacy characterizations, in order to yield value (and perceived value) in most engineering pursuits—thereby justifying funding and resource allocations.

The emphasis on *process* may have to do with the popular notion expressed by Miser & Quade and [1] – [6] and much validation literature that, in the spirit of Footnote 2., model validation entails a continuing process of testing a model and hopefully building “confidence” in it by amassing evidence of model corroboration in various use settings. The complementary notion is that by noting where the model fails, a map can be effectively constructed to delineate regions in the parameter space where the model is and is not valid for specific use purposes. These are appealing sentiments. However, it is not evident how the currently adhoc processes of confidence building and mapping out regions of model adequacy and inadequacy are quantitatively formalizable or accomplishable beyond the anecdotal stage—whether for fully weak models or for more strongly defined models. Furthermore, the paradigm of on-going testing toward these objectives is not practical in most model validation activities, where time and resources for model validation testing are usually very limited.

A related point is that ‘confidence’ in a model or its predictions does not seem to be quantifiable on any measurable absolute scale. It is reasonable to say that confidence increases as a model passes various validation tests, but how is this confidence quantified and measured? How is it used as a quantitative caveat or qualifier for new prediction results beyond the validation setting? Viable answers do not appear to be currently available so the author prefers to frame things in terms of *uncertainty* in models and predictions rather than *confidence* in models and predictions. While any uncertainty estimates won’t be perfect, an absolute scale is involved and a workable framework and methodology presently exists to address the validation problem and objectives in terms of quantitative *uncertainty*.

Because *confidence* also has a technical meaning in the field of statistics, use of this word in the context of model validation can lead to confusion and misunderstanding, so the author avoids associating this word with model quality. Indeed, in the statistical context increased confidence is associated with *larger* uncertainty intervals. In a connected matter, some may argue that uncertainty bars corresponding to statistical confidence levels can be determined for extrapolative predictions (e.g. [11]), but this presumes that computational physics model accuracy does not change in extrapolation. This is generally not a reliable assumption, especially under geometry and boundary condition changes. Model ‘credibility’ (e.g. [57], [73], [75]) seems to be a better term to express built-up belief in a model’s trustworthiness.

Although it is not evident how to pursue (practically or technically) approaches of establishing model confidence and mapping out regions of model adequacy and inadequacy, a concrete conception of validation is presented in sections 3.2 and 3.3 that is practical, self contained (not ongoing or open ended), and based on quantitative uncertainty. The methodology has been successfully demonstrated on the variety of challenging validation projects cited in Section 1.

Finally, a potential conflict comes into focus between all the quoted validation definitions and the fact that an affirmative validation conclusion (model is adequately accurate) at validation points in the modeling space does not necessarily imply that the model will retain acceptable accuracy in extrapolative prediction settings. If tested at new prediction points the model could be found not adequately accurate for the intended purposes of the predictions. The phrases ‘*from the perspective of the intended use*’ and ‘*intended domain of applicability*’ and ‘*purposes of the study of which it is a part*’ in the quoted definitions signify extrapolative predictions where the model is anticipated to be used. (A resolution of this logical conflict is proposed at the end of this section.) Besides degenerate cases of limited practical interest it appears otherwise impossible to determine a model’s accuracy for prediction settings different from the points at which the actual validation assessments are conducted. Establishing model adequacy in extrapolations (per the zeroth order criterion explained in section 3.3) is only slightly less tenuous. Accordingly the author tries to avoid the phraseology ‘validate a model for an intended use’ in favor of ‘validate a model with respect to an intended use’. (It is found that [67] also uses the latter phraseology, but does not cite any particular reasons.)

By now it is clear that substantial variation has historically existed and presently exists regarding philosophy, terminology, and definitions of model validation and validated models. ‘Model validation’ and ‘validated model’ can mean any number of different things to different people and in different contexts. The terms are treacherously ambiguous when communicating among modelers, analysts, decision makers, and customers. This calls for appropriate refinement of the concepts and terminology.

In this report the terms *model substantiation* or *corroboration*, *model refutation*, *substantiated model*, *corroborated model*, *refuted model* are used to signify the particular conclusion from a validation adequacy assessment. Saying “the model has been substantiated according to stated adequacy criteria” or “the model was refuted according to stated adequacy criteria” are much less ambiguous than saying “the model has been validated according to stated adequacy criteria.” Furthermore, an operational definition of model validation as interpreted and applied in this report follows.

Operational definition of model validation—*the compilation of useful indicators regarding the accuracy and adequacy of a model’s predictive capability for output quantities (possibly filtered and transformed) that are important to predict for an identified purpose, where meaningful comparisons of experiment and simulation results are conducted at points in the modeling space that present significant prediction tests for the model use purpose.*

The definition entails not only the act or process of assessing via appropriate validation experiments and procedures, but also the insights gained from the process as reflected in the validation products of model accuracy and adequacy results and conclusions. *Significant* prediction tests go as far toward the intended model-use settings and input conditions as can be accomplished in the validation project.

This operational definition reflects that we can rigorously determine model accuracy and adequacy only where experimental data exists. The *idealized objective* of model validation is to establish that a model will be adequate when used beyond validation points in the modeling space. This is the desired objective, but unfortunately is not wholly possible to attain. Because this idealized objective is present in the SCS, standard, and Miser & Quade definition statements, the author considers these not to be definitions of model validation *per se* (that can be put into practice), but instead to be *objective statements* for model validation. It is fine to be optimistic when stating objectives, as the SCS, standard, and Miser & Quade statements are, but operationally a different definition is needed.

Toward the idealized objective of model validation, if model adequacy per the methodology in sections 3.2 and 3.3 can be established at the validation points, then this constitutes a ***zeroth order satisfaction*** of the idealized objective of model validation. Nonetheless, other analysis procedures should also be brought to bear in judging whether a model will perform satisfactorily in specific uses subsequent to the validation tests, see section 6.

This page intentionally left blank

3 Several Fronts of Current Procedural Difficulty in Model Validation, and a Workable Real-Space Approach

Discerning whether the model is “good enough” is difficult on at least the following fronts.

- Deciding what is adequate in terms of an abbreviated basis of raw and processed output quantities for validation comparisons.
- Deciding on the formulation or metric of discrepancy (accuracy) characterization.
- Deciding the threshold or criterion for model adequacy (acceptable agreement with reality).
- Deciding how to project results and outcomes from the validation setting/s to other prediction settings—dealing with extrapolation of validation information/results/products. (This item is not part of model validation, but is an essential consideration in the formulation of a workable validation paradigm.)

Many modeling frameworks consider the above items (or some of them) to be separable, and devise sub-frameworks to address them separately. However, the present framework considers these items to be strongly interdependent. It is reasoned that only by considering them together can a comprehensive and workable end-to-end framework be formulated.

This interconnectivity bears not only on the technical procedures of model validation, but also on the interpretability and usability of the information and products from the model validation activity, which is paramount in importance. It is not worthwhile to go through the expense and rigors of a validation exercise unless the results can be interpreted in a useful manner, where salient and meaningful conclusions can be drawn and what is learned can be used to quantitatively characterize and perhaps 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 funding and resources. Therefore, to be relevant beyond just qualitative insights, there must be a plan for how the information gained will be packaged and interpreted to provide a useful quantitative characterization of model accuracy and adequacy in a larger context of downstream model use for predictions beyond the model validation setting. These considerations pertain similarly to model conditioning and are addressed by the framework similarly.

3.1 Deciding what is an adequate validation basis of comparisons — Scalar vs. Field comparisons, State Variables vs. “Resultant Effect” quantities, etc.

One important issue in model validation assessments is the “representativeness” of raw or processed experiment and model output quantities that are compared in a validation assessment. For example, we might be interested in how well a computational fluid dynamics (CFD) fire simulation represents a hydrocarbon fire in calm wind conditions. We could contemplate performing validation comparison of the simulation’s field

variables (e.g., pressure, three components of velocity, species concentrations, etc., all as a function of time and space) against corresponding quantities measurable in the experiment (e.g., with laser sheets and particle image velocimetry). Then there are potentially as many scalar or individual validation assessments to be made as there are points of data (in time, space, and category = pressure, x -direction velocity, etc.).

The problem would necessarily be reduced to comparisons for a manageable number of data points, say on a regular grid on the time-space domain, or using a randomly selected statistical sample of time-space data points for comparison. Unfortunately, current validation theory does not appear to exist for how much thinning of the data can be done and the remaining comparison points still be considered to be representative of the larger set of time-space information.

A possible simplification is that the time dimension might be managed by comparing whole time histories of the various quantities at a manageable number of spatial locations in the domain. Additional reduction could be gained by comparing time-averaged or steady-state values, etc. Further reduction could be gained by employing suitable spatial integration or averaging. Another type of reduction would consider only a subset of the categories of output (velocity but not pressure, etc.).

All of these potential reductions from the full-field validation problem make things more manageable from a validation practice standpoint, but have the unavoidable consequence of reducing the veracity of the validation test and the strength of the conclusions that can be drawn about model accuracy and adequacy. The power for asserting the model is “right for the right reasons” is commensurately diminished. Indeed, a *sufficiency* condition for such claims is probably only reachable in concept. One might alternatively appeal to statistical arguments to establish validation sufficiency in practical terms (e.g., odds associated with the hypothesis “right for the right reasons”), but this is beyond the scope of current model validation theory. At present it appears that only *necessary* conditions for “right for the right reasons” are approachable.

A different type of reduction in the difficulty of the validation problem, but also with weakening of the power to assert the model is “right for the right reasons”, comes from comparing derived resultant quantities of application-specific interest that are computed from (some of) the primitive-variable field data. For instance, the prediction of total drag on a vehicle is often the focus of interest instead of point details or statistical quantities of the pressure and velocity fields; or the focus of interest may be the time between when two parts in a heated device reach their failure temperatures, rather than the temporal-spatial temperature field that occurs in the device (the said field would probably not even be determinable with current experimental techniques).

In the fire example, instead of comparing field quantities the *resultant effect* that the fire has on a number of sensors or transducers of the field would normally be compared in the validation activity. Such sensors could be flux gauges distributed throughout the fire, or suitably located and oriented calorimetric plates, objects, and walls outfitted with thermocouples. Because the fire is stochastically fluctuating in time and space, the sensor

data would have to be processed to provide sufficient time and space averaging to get well-behaved quantities for comparison. By evaluating such quantities, the computed *effects* of the fire can be compared to measured *effects*.

Again, 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. 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, even if the modeled physics is not precisely correct at a detailed level, but the aggregate *effect* is sufficiently accurate, the model can be quite useful—the empirical evidence is that use of imperfect models and modeling has been quite productive in science and engineering.

Even so, in model validation one would like to produce and compile as much evidence as practical that the full physics fields themselves are not drastically dissimilar. The more spatially and temporally diverse (well-distributed) the sensors are, and the more categories of quantities monitored (velocity and pressure, etc.), the greater the chances of identifying any significant dissimilarities. This drives toward using as many diverse sensors as affordable in validation experiments.

Although a best effort is made, ultimately it is only practical to compare experimental and simulation results for a small number of possible output indicators, with some or all of the compared indicators often being indirect: transformed integral, averaged, or resultant-effect measures. An inevitable incompleteness of model substantiation evidence always exists (usually to a great degree) and we are left with acknowledging that it is only possible at best to achieve incomplete circumstantial corroboration of model predictiveness. Nonetheless, carefully designed and executed model validation assessments usually add greatly to the modelers' and the model users' knowledge about the performance of the model. This is an essential element of due diligence and applied risk management in developing and using models.

3.2 Deciding on the Formulation or Metric of Discrepancy Characterization

Meaningful and relevant accuracy characterization of a model's mapping of inputs to outputs is also a difficult issue. It is not as simple as quantifying discrepancy or non-equivalence (in whatever norm one chooses) between output results of the experiment and of the model. This would be the case only if the true experimental conditions (e.g. boundary conditions) applied to the real system are also input to the model. When the applied loads and conditions in the experiment like temperature, species concentration, humidity, flow inlet turbulence intensity, etc. are uncertain to some degree, then this complicates matters.

The author’s experience is that inexact control of experimental input conditions, and uncertainty in their values due to measurement uncertainty, is typically a leading source of uncertainty concerning the mapping of experimental system inputs to outputs. Such uncertainty is often quite substantial, and can lead to “Type X” error in model validation and conditioning if it is ignored (see [47] and Section 4 of this report). Type X error is a systematic-uncertainty analogue to the random-uncertainty Type II hypothesis testing error associated in [77] with “model user’s risk”. Experiment output data must be properly conditioned to reflect important sources of systematic uncertainty in experimental inputs (e.g. [47]-[52]). This consideration appears to be overlooked by other model validation and conditioning frameworks in the literature.

The real-space validation framework is also different from most others in the literature in that model accuracy and adequacy relative to experimental data are not posed in terms of transform measures and acceptance criteria in a discrepancy “*transform space*”. A large variety of mathematical transforms exist in the literature to characterize discrepancy between experiment and simulation results (e.g. the linear subtraction transform in [6] originated in [17], and the “Area CDF” discrepancy metric in [7] and [54] originated in [20]). The transform measures can get relatively sophisticated and involved, with varying degrees of transparency or interpretability of the physical and decision-making significance of the numerical values yielded by the discrepancy measures. The transforms can also put constraints on what forms and types of uncertainty can be handled (see [52]). Indeed, workable criteria to demarcate adequacy of model-experiment agreement in the transform space of the various mathematical measures are found to be rather elusive (see next section).

Instead, by considering the problems of model accuracy characterization and adequacy determination to be coupled (in anticipating the need for transparency and interpretability in model adequacy determination), the real-space approach was developed ([42]-[52]). Project requirements of pragmatism, versatility, and demonstrated value have led to the approach, which is sketched next.

Figure 1 concisely expresses the real-space comparison paradigm for model accuracy and adequacy assessment. Analogues exist for assessment of time-varying and space-varying (vector) output (see e.g. [46], [49], [56]), as opposed to just the single (scalar) output cases shown in Figure 1. Three different categories of comparison outcomes are shown, where uncertainty bars of a model prediction are compared to uncertainty bars of experimental data. The construction of the uncertainty bars is briefly outlined below. The three categories of comparison outcome regarding model adequacy are interpreted in the next section.

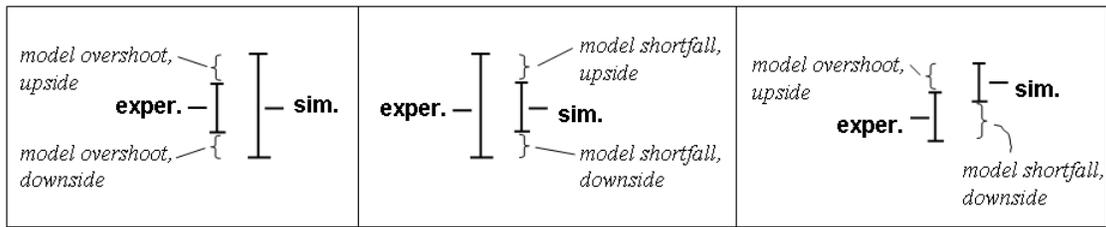


Figure 1. Generic categories of comparison outcomes in the framework's model validation and model conditioning real-space comparisons of uncertainty intervals of experimental versus simulation results for a scalar output quantity.

The uncertainty bars on the experimental output data in [Figure 1](#) depict the range of uncertainty associated with experimental output response. This uncertainty can come from various aleatory and epistemic sources.⁵ In general the net experimental uncertainty may include contributions from:

- A. test-to-test stochastic variability, including:
 - i. the tested systems and/or phenomena (e.g. part geometries, material properties, stochastic behavior, etc.);
 - ii. measurement and processing errors on system outputs;
 - iii. measurement and processing errors on system inputs (environmental conditions, applied loading/excitation, etc.);
- B. epistemic uncertainty on applicable response statistics or quantities determined from limited sampling of the above stochastic factors in a limited number of repeated experiments (see e.g. [\[101\]](#));
- C. systematic bias uncertainties associated with measurement of system outputs and any procedures and models used to process, correct, and/or interpret the measured data;
- D. systematic measurement and processing of uncertainty on system inputs and experimental factors, including apparatus/setup, test conditions, and boundary conditions.

The uncertainty bars on the model results depict the net uncertainty associated with the predictions. These issue from uncertainties in numerical discretization effects and results processing, from input uncertainties in the model simulations (e.g. uncertainty in the experimental conditions), and from model-intrinsic uncertainties (different plausible

5. Many complications, choices, and caveats are involved in forming uncertainty bars on experimental data, so they are always somewhat subjective. Nonetheless, reasonably objective results are obtainable in engineering validation activities, with readily accessible uncertainty quantification and propagation tools and procedures.

model forms and/or parametric model-form uncertainty). Note that in the framework's bookkeeping scheme, uncertainties from numerical discretization and processing of model output results normally *decrease* the size of the simulation uncertainty bar relative to the experimental uncertainty bar. This is due to the way the framework uses the relative sizes of the experimental and simulation uncertainty bars to gauge model adequacy. See [47], [49], [52] for further explanation and examples.

The framework's expression of experimental and simulation uncertainties ultimately in terms of uncertainty intervals embraces the non-probabilistic nature of uncertainty that is unavoidably present in experimental and modeling projects and is usually dominant relative to probabilistic distributional uncertainty. Indeed, for a framework to be fully general and viable it must suitably handle non-probabilistic types of uncertainty like interval uncertainty and discrete model-form uncertainty among alternate plausible models. See for [47] - [49] and [51] - [52] for examples of how the Real Space framework propagates and aggregates distributional, interval, and discrete uncertainties into a representative uncertainty intervals of experimental and simulation uncertainties for validation comparison (*c.f.* Figure 1).

In contrast, most model validation and conditioning frameworks in the literature are based on a probability distribution model of uncertainty. The thinking might be that the framework should be fashioned around this type of description because it is the best one can do in terms of specificity of uncertainty. However, it is perhaps too optimistic to base a framework on this ideal limit of what can be achieved in terms of knowledge of uncertainty.

The present framework derives from a standpoint at the other extreme: almost never in real experimental and modeling programs are probabilistic uncertainty distributions available or achievable. In fact, very often *no* direct probabilistic or statistical information is available for many of the driving uncertainties. Rather, much of this uncertainty is estimated by extrapolation from other experiments; by expert judgment; or by modeling and analysis rather than actual empirical data. The current framework proceeds from *this* perspective as the problem description.

The author is aware of no validation framework founded on a probabilistic basis that has been demonstrated to adequately handle problems dominated by interval and/or discrete model-form uncertainty. However, this is not to say that such frameworks do not or cannot exist.

Much of the framework's uncertainty representation and comparison machinery is the same whether the purpose is model conditioning or model validation (e.g. [46], [56]). In contrast, many of the discrepancy metrics and formulations in the literature do not quantify model error in real space, even though real-space characterization is often necessary to affect a correction to the model before proceeding to prediction (see e.g. [46] and Appendix A).

Another important consideration in devising or choosing a formulation or metric to characterize discrepancy concerns discrimination ability. For example, let the center and leftmost cases in [Figure 1](#) represent 5th and 95th percentiles of probability distribution functions (PDFs) of independent uncertainties associated with experimental and simulation results. Let the experimental uncertainty be from measurement uncertainty on the experiment output results, and let the simulation uncertainty be from model-affiliated intrinsic traveling uncertainty (e.g. a material property uncertainty). If, for instance, the methodology of the ASME V&V standard [\[6\]](#) is applied, it yields a transformed discrepancy measure that does not differ between the left and center cases in [Figure 1](#) (see [Appendix A](#)). The “Area” discrepancy measure from [\[20\]](#), [\[38\]](#), [\[54\]](#) also gives the same value for the left and center cases in [Figure 1](#). [Appendix A](#) shows that both metrics mask another fundamental mode of discrepancy that can exist between experimental and simulation results. Indeed, all such discrepancy transform measures that the author is familiar with have non-exclusive mappings of real-space results to transform space. This can undermine model adequacy determination in validation assessments; the next section describes how the left and center cases in [Figure 1](#) are crucially different in terms of model adequacy with regard to prediction risk in downstream use of the model for design, analysis, or decision-making purposes. [Appendix A](#) also shows the shortcomings of using transform quantification of model discrepancy to extrapolate model error or corrections.

Hence, if the scope of use of a discrepancy measure includes model adequacy determination, extrapolation support, and prediction risk in downstream use of the model, the Real Space approach has significant advantages over transform discrepancy characterizations like the subtraction and Area transform metrics, for example.

From a limited scope of just accuracy assessment it could be argued that the real-space comparison finds no difference between the [Figure 1](#) center and left cases in terms of amount of mismatch error between experiment and simulation: the shortfall errors in the central case are the same magnitude as the overshoot errors in the leftmost case. However, a broader scope of consideration anticipates the adequacy question. Then directionality (not just magnitude) of the error matters and the Real Space metric also represents this important aspect.

Furthermore the Real Space characterization does not constrain model error to be posed in terms of a *singular bias* of the model, either downward or upward by some value. That is, *the real-space metric allows model results to be **simultaneously biased upwards and downwards** relative to the experimental data*, as the center and leftmost cases in [Figure 1](#) indicate. This can be contrasted to e.g. the formulation in [\[6\]](#) which poses model error/bias as a single but unknown value.⁶

Another advantage of the Real Space portrayal of the relative sizes and positions of the experimental and simulation uncertainty bars is a transparent and easily interpretable presentation of the discrepancy. The discrepancy characterizations from transform approaches are less intuitive. This impedes adjudication of model adequacy. Workable

criteria for model adequacy are found to be more accessible for the Real Space approach than for the transform approaches. This is elaborated in the next section.

In summary, many transform discrepancy metrics exist, but none appear to have practical advantage over the simple and versatile Real Space comparison approach. Transform metrics add complexity (and sometimes constraints) in the analysis, and present difficulty in deriving useful meaning that supports analysis insight and decision making. In particular, it is not evident how transform-space criteria are determined for deciding whether models are adequate.

3.3 Deciding the Threshold for Model Adequacy

3.3.1 Two Potentially Different Validation Settings regarding Accuracy Requirements for Model Adequacy Determination

Validation Setting 1. Consider an example where a CFD fire model is being developed, say at a university. Let the model development context be devoid of a particular engineering application for which the model is to help resolve certain issues, answer specific questions, or make decisions. The goal at this stage is simply that the model replicate an actual fire “well.” What criteria can be applied in characterizing the quality and usefulness of the model? In more stark validation terms, if one wants to assert that the model adequately matches experimental data, then what standard or threshold must the model pass?

Other examples of models in this type of validation setting include material property and constitutive models, turbulence models, convection correlations, thermal contact-resistance correlations, etc. These types of models are usually developed in relative isolation from analysis, design, and decision making associated with an engineering project having accuracy requirements specified for the modeling.

Lacking an external accuracy criterion, one could look for some type of natural or intrinsic criterion. Since the objective in these settings is simply that the model replicate the target phenomena “well”, one approach might be to apply a statistical hypothesis test for whether the model results can pass as being the “same” as the data with respect to particular measures like similar experimental and predicted probability distributions of response, or similar means of the results. The point-null hypothesis is that the model and experiment agree, and the evidence is weighed for rejecting this hypothesis. This is in accord with the

6. The framework [6] concentrates on systematic uncertainty categories C and D listed earlier. The formulation holds that one value exists for each uncertain quantity in the experimental and modeling uncertainty sources, but the value is unknown to within the prescribed uncertainty. However, for modeling and validating stochastic systems (most systems are stochastic to some degree), uncertainties in categories A and B arise concerning populations of many results, rather than just one value but unknown to within a specified range or probability distribution. Figure A.3 in Appendix A presents a case where the subtractive difference metric does not accommodate stochastic uncertainty in repeated experiments. Conversely, an early draft of the ASME V&V-10 document [7] considered methodology for only stochastic uncertainty (category A). Ref. [7] is currently being revised. All categories A – D are treated in the Real Space framework, see [49], [51], [52]. Appendix A and reference [112] compare several validation frameworks.

paradigm that a theory or model cannot be proven to be valid or correct, but is subject to being unseated by contrary evidence. The hypothesis test sorts through the statistical uncertainty associated with random variability in repeat experiments and quantifies the level of statistical significance at which the hypothesis of “sameness” of model and experimental results can be rejected.

Since relatively strong (statistically significant) evidence of inconsistency between model and experimental results is required before the model is rejected, the approach has a drawback in that the greater the stochastic variability in the tested systems and/or experimental conditions, and/or the smaller the number of repeat experiments, the relatively smaller the chances of detecting and therefore rejecting a biased model. Barring strong evidence of inconsistency, a significantly biased model could enjoy a “free lunch” in that non-rejection can be misinterpreted as implying model goodness. However non-rejection does not necessarily imply model goodness because noisy and uncertain data can leave much room for model bias error to go undetected. This is the classical Type II error in statistical hypothesis testing, designated *model user’s risk* in [77]. In fact, the greater the probabilistic certainty sought that one is not incorrectly rejecting an unbiased model (where incorrect rejection is called Type I error), the greater the chance of committing Type II error of not rejecting a biased model. Ref. [47] and section 4 of this report demonstrate an analogue of Type II error for *systematic* uncertainty as opposed to the *random* uncertainty in statistical hypothesis testing. Hence, that the model is “not inconsistent” with the data is not a reliable indicator of model accuracy or adequacy.

The converse problem is that the more precise the experiments are, and the greater their number, the more likely the hypothesis test is to find that the model is biased, as one would normally expect. Indeed, it appears to be an improper formulation that the classical point-null hypothesis test is posed to support the less reasonable hypothesis (that the model has zero error) unless overturned by strong evidence. These shortcomings underlie the problem with the anti-invalidation conception of model validation discussed at the end of section 2.3-Part I.

The more proper question is whether the model has acceptably small bias. Accordingly, “*interval null*” hypothesis tests (see e.g. [74]) judge whether model bias is smaller than some specified level. But as pointed out in [74] and [85], the problem remains that the hypothesis test formulation is skewed toward a finding of “not inconsistent” with the allowed discrepancy tolerance. This does not provide positive evidence that a model is unbiased to within the allowed tolerance.

Several validation approaches are proposed in [74] that get away from the not-inconsistent type of testing and toward more direct indicators of model accuracy/quality. Bayesian interval-hypothesis testing is introduced, along with a “model reliability metric as the probability of data falling within model predictions”, but these approaches focus on the accuracy of predicted vs. experimental mean results, as do the point-null and interval-null hypothesis tests. Unfortunately, testing whether the model matches mean experimental results (within some specified allowable bias error) is usually insufficient to validate model

predictivity for design and decision-making purposes (see Figure 2 and related text below).

References [17] and [58] introduced what could be viewed as an analogue of the means hypothesis testing approach—for systematic uncertainties in modeling and experiments instead of the random uncertainty in multiple experimental trials that traditional hypothesis testing is affiliated with. They observed that the combined experimental and model parameter uncertainty provide an uncertainty “noise” level or resolution floor below which any model bias cannot be detected. This is analogous to the principle that statistical hypothesis tests work on. They held that the combined experimental and modeling uncertainty, the “validation uncertainty” of magnitude scale u_V , is an uncertainty measure with special significance in model validation as a natural or intrinsic criterion that bears on model acceptability. They proposed that validation is achieved at a u_V level of uncertainty if the mean indication of model bias (i.e. the bias E between the means of the experimental and simulated response distributions) is less than the combined experimental and modeling uncertainty u_V . (However, in actuality the model could be biased by as much as $2u_V$ and still satisfy their criterion for validation at the proclaimed u_V level.) Note that all three cases in Figure 1 would pass their test for model validity at the u_V level, but the center and rightmost cases possess major risk liabilities (in post-validation downstream predictions) that the leftmost case does not.

The u_V -level criterion for validation of the model at that accuracy level was dropped in the subsequent ASME effort [6], but most of the uncertainty characterization and propagation framework was kept for quantifying model accuracy in terms of bias from experimental data. Indeed, the standard [6] is limited to model accuracy characterization and says that validation only provides an uncertainty statement on the model: the model is ‘validated’ to have a bias within the range $E \pm xu_V$, where x is a proposed “coverage factor” set by the modeler based on various considerations. The determination of whether the model is adequate with respect to a particular use is not addressed: “...validation does not include acceptability criteria, which are relegated to certification or accreditation...[model] intended use is very general (with specific intended use being tied to acceptability criteria embedded in project-specific certification rather than validation).” Somewhat inconsistently, the same Appendix C-9 also states that “full validation” can be considered to include pass/fail determination of whether the bias is acceptable or adequate for a particular application. However, no guidance is given for how acceptability criteria might be rationally determined for project-specific model adequacy determination. A particular approach is introduced with the notion of *coupled hierarchical adequacy determination* under *Validation Setting 2*, below. Substantial practical difficulties are involved, although not necessarily insurmountable. It is presently more practical to use the non-project-specific (*stand-alone* or *uncoupled*) nearly universal criterion for model adequacy described next.

Consider the simple illustrative example in Figure 2. One possible validation criterion in a real-space approach is that the model results lie within the net uncertainty bounds of the experimental data, as in the figure. Many validation approaches in the literature advocate

this paradigm. However, this conception has the undesirable “free lunch” property where the larger the uncertainty in the experiments, the easier it is to find the model valid. So less precise experiments are rewarded with easier findings of model adequacy by this criterion—a bad property to have in a model validation scheme.

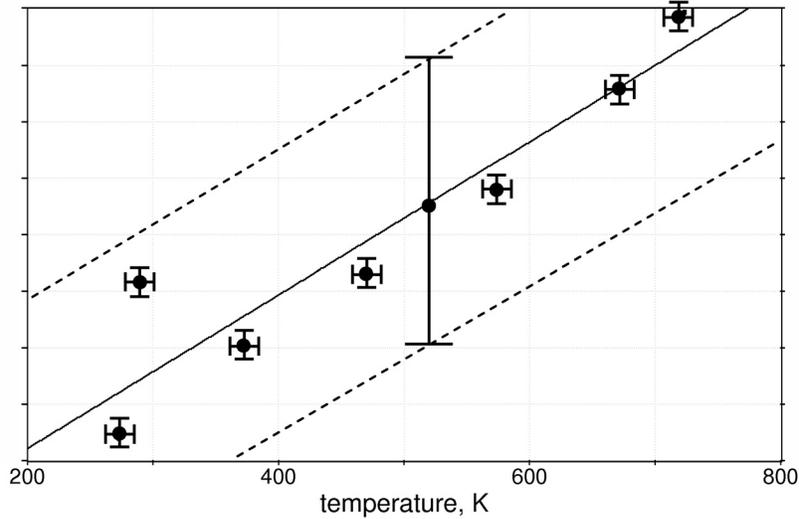


Figure 2. Material property measurements as a function of temperature, with vertical and horizontal uncertainty bars associated with the measurements, and net uncertainty bounds (dashed lines) for the set of data. At any given temperature it is expected that actual property values will fall within the experimental uncertainty bounds depicted at that temperature. The solid line shows model predictions of the material property value as a function of temperature.

A contemplation of model use after the validation assessment condemns the criterion even more strongly. Although model predictions lie within the uncertainty of the experimental data, actual property values deviate significantly from the prediction curve. Thus, the deterministic model does not suffice to fully represent material-property expectations at a given value of temperature. If the model (claimed ‘validated’ according to the said criterion) is utilized in downstream simulations and evaluated over the same temperature range, the model would significantly misrepresent the material property as being much more precisely known than it actually is. This would under-represent the actual material property uncertainty in downstream model use. Therefore, to claim a model validated because its results lie within the uncertainty of the experimental data in validation settings appears to be inappropriate and inviting of trouble in model-use settings.

Consideration of model use downstream from the validation activity is essential in devising or choosing viable model adequacy criteria to apply in the validation setting. If the modeled phenomenological mechanisms extrapolate *consistently* (see Section 2.2.1) from the validation to the prediction setting, then for the middle and rightmost cases in Figure 1 a similar relationship will hold in new prediction settings—much of reality will lie outside of model predictions in new settings. Clearly, this is **not** what most designers, analysts, and decision makers would want from model predictions. Rather, they would like

to be reasonably assured that reality lies *within* the uncertainty of the model predictions—indicating the leftmost case in [Figure 1](#) as being the most desirable of the three cases. To the extent that extrapolation behavior can be anticipated, risk mitigation in prediction is best served by a validation criterion that *the uncertainty bounds of the prediction in the validation setting should encompass reality*. Then if the modeled phenomena extrapolates consistently a similar relationship will hold in new prediction settings—reality will lie within the model predictions. This supports the objective of BEWU modeling and prediction and satisfies to **zeroth order** the idealized validation objective discussed at the end of [Section 2.4](#). Of course, other analysis should also be sought in judging whether a model will perform satisfactorily in specific uses subsequent to the validation tests. More research needs to be conducted in this area. This could lead to higher-order satisfaction of the validation objective.⁷

Model adequacy also requires that the upside and downside model overshoot errors in the leftmost case are small enough that the model will be useful in downstream predictions. For instance, infinite uncertainty bars on the simulation results (say due to infinite model-intrinsic uncertainties in the model) would contain the experimental results in the figure, but the model with its infinite uncertainty would not be useful for prediction purposes. Therefore it could not be considered to be an adequate (acceptably accurate) model. The broader issue of acceptable magnitude of model overshoot and undershoot errors in the three cases of [Figure 1](#) is discussed in the following.

Validation Setting 2. Recall the setting discussed in [Section 3.1](#) where the engineering or programmatic objective is to predict heat loads on an object of interest in a fire. Perhaps for this situation, given concrete system-level prediction accuracy targets, it is possible that commensurate accuracy targets could be set for validation of the fire CFD model as a submodel of the larger hierarchical system-level model. This case has a different validation contextualization than Validation Setting 1, which is devoid of a specific engineering

7. For example, zeroth-order satisfaction of the idealized validation objective is presently said to occur whether the leftmost case in [Figure 1](#) exists, or as an extreme alternate possibility, the uncertainty bars of the simulation just barely encompass the experimental uncertainty (essentially equaling the range of the experimental uncertainty). From the standpoint of the validation objective the latter case leaves less room (effectively no room) for model accuracy loss in extrapolative predictions. The leftmost case in [Figure 1](#) has larger margin for model accuracy loss in extrapolation, and therefore has a tangibly larger chance that the validation objective will be met. Thus it is foreseen that refinements along these lines can be made to the present zeroth-order model validation criterion. It is also emphasized that models with lower embedded uncertainty are not necessarily better for prediction; they are necessarily less robust to model-form error (e.g. [\[86\]](#)). The validation methodology should reflect this. However, many validation conceptions in the literature intrinsically favor a validated/adequate model with less internal uncertainty to a validated/adequate model with more internal uncertainty (however adequacy is defined). Such automatic preference lacks an end-to-end view that sufficiently includes downstream prediction in the scope of the thinking about model validation. These considerations also affect the amount of uncertainty that should be mapped to models for model conditioning. Perhaps appropriate factors of safety should be used in model validation and conditioning to buffer against model accuracy degradation in extrapolative prediction. These are important subjects for research.

application and associated accuracy requirement. The difference is significant as regards model adequacy assessment.

It is first necessary to discuss the feasibility of mapping top-level program requirements for performance, safety, etc. of an engineered system to accuracy requirements on the various system-model elements (submodels) in a hierarchical model validation conception. Assume that performance, tolerance, or safety “requirements”⁸ at the system level could be transformed into modeling accuracy allowables on performance predictions at the system level. Say that a requirement is stated that the predicted temperature-induced failures of components in a fire-engulfed weapon not differ by more than 10% from actual component failure temperatures in a validation experiment. Then, in trying to map this error budget at system level into error budgets for the various elements of the system model, an infinite number of combinations will satisfy the top-level error budget.

For example, let the 10% accuracy requirement be parsed (perhaps arbitrarily) into equivalent top-level effects of no more than 6% error contributed by the fire-heating submodel, plus up to 2% error from modeling weapon/component thermal response, plus up to 2% error from modeling the temperature thresholds at which the components fail. Then, fractally downward, within the thermal modeling for instance, it is possible for an infinite number of different combinations of error budgets for its modeling elements (foam vaporization, material properties, geometry, thermal contact resistance, etc.) to meet an aggregate error of 2%. How is the 2% error budget best, or even non-arbitrarily, parsed among the submodels?

References [45] and [46] elaborate on the non-uniqueness issue and the associated technical problems and trial-and-error difficulty involved with trying to parse a system-level modeling accuracy budget to corresponding budgets of the individual submodels. Besides the difficulty of devising a defensible parsing scheme for the reverse-mapping problem, significant computational difficulty and expense would exist in solving the reverse problem. Even if successful, the submodel error budgets may not be mappable to individual validation experiments for the submodels. In many cases an explicit parametric map does not exist between a submodel’s instantiation in the system setting and its instantiation in the validation setting, particularly when changes in device geometry or spatial/temporal loading are necessitated for practicality in the validation experiment. Without a parametric map the error budget cannot be mapped to the equivalent accuracy requirements at the validation setting. Ultimately, the paradigm of downward-mapping to obtain submodel accuracy requirements for validation is severely impractical, if not impossible; not a viable paradigm.

8. Such “requirements” at the system level are themselves unavoidably subject to some degree of arbitrariness from: uncertainty in program objectives and constraints; multiple program objectives (some even conflicting) that disallow a unique project optimum (create a Pareto front) among the tradeoffs involved; flexibility or negotiability of objectives and constraints; decision-maker knowledge limitations and subjective variabilities in risk perception and tolerance based on individual predispositions and experiences; etc.

However, an approach is now outlined for deciding whether submodel accuracy meets top-level accuracy requirements specified for a project. Submodel adequacy is cast in integral terms such that all submodels together meet top-level accuracy needs, or fail as a group. The discussion assumes a real-space characterization of discrepancy between experimental and model results, but analogues may be devisable for transform metrics of discrepancy.

In the general case where multiple submodels of a hierarchical model are being assessed, a *coupled* validation problem exists where all the submodels' overshoot and shortfall errors (potentially from all three generic classes shown in [Figure 1](#)) can be combined via forward propagation through the system model. The resultant can be compared against the system level error budget. All of the assessed submodels together, with their overshoot and/or undershoot errors cancelling and adding as they do, are found to be acceptably accurate as a set, or are found to be unacceptably inaccurate as a set. Thus, any submodels in the center and rightmost classes in [Figure 1](#) may be found to be adequately accurate for their use in the system model at the validation conditions, even though individually they do not meet the stand-alone adequacy criterion of capturing reality within their prediction uncertainty. Although these individual submodels push the system model toward under-prediction of uncertainty, their shortfall errors may be small enough to result in acceptably accurate predictions at the system level. Another possibility is that all submodels individually qualify to the stand-alone adequacy standard (leftmost class in [Figure 1](#)), but together the propagated overshoot errors result in unacceptably large overshoot error at system level. Then the submodels when used in combination are not accurate enough.

This scheme can be recognized as a propagation of errors associated with submodels, e.g. in contrast to propagation of errors associated with model parameters. In cases of system-model inadequacy, the individual submodels that contribute most error to the system-level results would be the ones to consider improving first to reduce error in system-level predictions.

Coupling and involvement exists that is not mentioned with the principle commonly advocated in the literature of *hierarchical model validation* where submodels are to be individually validated before assembly into the system model. This strives to establish to the degree practical that a system model if substantiated is right for the right reasons because the submodels have been substantiated individually (“bottom-up validation” of the system model, [\[22\]](#))—and not just the integral result from the combined system level model has been validated (“top-down validation”, [\[22\]](#)). The principle is certainly appealing.

The principle appears to be advocated (e.g. in the ASME works [\[4\]](#) and [\[7\]](#)) under the assumption that rational *a-priori* accuracy thresholds for submodel acceptance can be issued *before* the validation experiments and simulations are run and processed. This would be ideal for implementation of the principle.

However, as already said, a downward (reverse) mapping of system-level model accuracy requirements to submodel accuracy validation requirements is generally not feasible.

Then the vision of validating models to *a priori* adequacy criteria (specified before the validation experiments) would seem to depend on some other way of arriving at accuracy requirements. There is little in the literature to indicate how this might be done. The author has seen a few instances where validation accuracy requirements have been stipulated *a priori*, but no technical basis or methodology for arriving at these objectively was provided. Without providing such a basis the outward appearance is that the accuracy “requirements” are substantially subjective and arbitrary—not quantitatively grounded. One must be prepared for internal and external scrutiny and critical review of a validation project, where the question may need to be answered, “What is the scientific or technical basis behind your model adequacy criteria?”

This is critical because model validation exercises usually take significant time, planning, and resources. Why go through the rigors and expense of a model validation exercise just to validate against arbitrary accuracy requirements? This does not seem to be a reasonable way for a project to use precious resources. Moreover, posing arbitrary accuracy requirements is liable to be counter-productive. Such requirements can have profound implications on model acceptance and reliance—either through overconfidence or unreasonable aspersion ascribed to the model, depending on whether it passes the arbitrary criterion or not. This can lead to unnecessary, unjustified, and counter-productive courses of action regarding project strategies and resource allocations.

Furthermore, a lower-bound constraint on validation accuracy requirements is set by the uncertainty in the validation experiments—one cannot substantiate a model to better accuracy than the accuracy to which reality is known. Thus, any *a priori* accuracy criteria, however obtained, must meet this lower-bound constraint. Since this lower bound is not known until after the experiments are run and processed, validation accuracy requirements arrived at *a priori* are subject to being found infeasible after the experiments are run. What happens with infeasible “requirements”? To avoid derailing the validation project the validation accuracy “requirements” would in all practicality be altered to meet the experimental uncertainty lower-bound constraint. If so, then they really weren’t requirements.

It is suggested in the literature that validation experiments be designed and executed with the necessary experimental techniques, instrumentation, data processing procedures, etc. to achieve accuracy levels well below any prescribed validation accuracy requirements. This would avoid the requirements infeasibility problem. However, this might be like “the tail wagging the dog.” An extremely difficult inverse experiment-design and execution problem arises, with ominous cost, time, and resource implications. It is exceedingly difficult to get a project to commit money and resources to the cost-unconstrained problem of meeting experimental accuracy requirements when the requirements cannot be substantiated to be rigorous or unique. Instead, the cost-constrained circumstance usually predominates: given a negotiated budget for validation experiments, the experimentalists work to minimize total experimental uncertainty within their resource constraints. A certain realizable level of accuracy is obtained in the experiments, irrespective of what the accuracy goals were. Importantly, the experimental uncertainty usually has contributions from stochastic variabilities inherent to the tested system of

interest. This aleatory uncertainty is usually not known *a priori* and cannot be limited or reduced in trying to get more accurate experiments per some *a priori* posed uncertainty requirement.

A priori specification of suitable accuracy requirements for transform-space discrepancy measures seems even less plausible. Thus, paradigms and frameworks that pursue validation procedures and judgments based on specification of *a priori* accuracy requirements seem implausible. Hierarchical model validation has been traditionally advocated under the questionable assumption of readily attainable rational *a priori* requirements. Nonetheless the principle is important and should be pursued. One foreseeable approach is *a posteriori* coupled (simultaneous) validation assessment of the submodels in the validation hierarchy. The forward propagation is cumbersome but tractable, and does not suffer from the non-uniqueness or technical difficulties of downward propagation of accuracy requirements, nor from any arbitrariness of submodel accuracy requirements somehow conjured up.

3.3.2 Hierarchically Coupled Adequacy Determination vs. Stand-alone Adequacy Determination

The demands of the coupled validation scheme may not be affordable or be considered necessary or “worth it” to the project. Indeed, for validation at the submodel level, the author has yet to experience all the required elements in place to pursue anything but uncoupled validation determinations. Such validation is much easier to accomplish. Additionally, it can be done early in hierarchical modeling efforts where no system level model exists, or when error tolerances at the system modeling level have not been specified (the usual situation).

In stand-alone validation the integrity of a conclusion of model adequacy rests on the correctness of the assumption that the model overshoot errors are not so large that this becomes troublesome in downstream predictions. Often the satisfaction of this assumption is clear. Conversely, it may be obvious that the overshoot errors are too large for the model to be useful, so it is therefore inadequate. In non-obvious cases it should be made clear that a conclusion of ‘adequate model’ rests on the said assumption.

Establishing that the model is validated to encompass experimental reality, with reasonable indication that its uncertainty is not too large to undermine the model’s applicability in downstream use, is a relatively strong statement of indicated model fitness for prediction (in as much as trend consistency can be anticipated in extrapolation). This is far beyond what is currently done in most modeling efforts.

4 Projecting Results and Outcomes from the Validation Setting to the Model Intended-Use Setting—Dealing with Extrapolation

4.1 “Type X” Validation Error and Lack of Consequence when Bias and Uncertainty travels “consistently” from Validation to Extrapolation Settings

Here we consider any aspects of a physical system (geometry, materials, and phenomenology aspects) that differ *inconsequentially* between the validation and prediction settings.

Consider a hardware device employed in a validation experiment. The device is eventually to be used in somewhat different deployment conditions, and predictions of its performance under those conditions are to be made. If the geometry and materials of the device are the same in the validation and deployment settings, then these modeled aspects in the validation setting “flow through” to the prediction setting.

If such traveling aspects of the system are mismodeled in the validation setting but the model as a whole is substantiated nonetheless, then the mismodeling error can, under conditions to be explained next, be inconsequential as regards model performance in downstream predictions. Consider the following cantilever beam example.

Let D signify the cantilever beam’s deflection at the free end of the beam and in the direction of an applied load P there. Assume the typical model-problem conditions ([102]) of zero deflection ($D=0$) and zero slope ($dD/dx = 0$) of the beam where it horizontally protrudes from a rigid unyielding vertical wall. Let D be the vertical deflection pointing in the direction of the vertical load, which is perpendicular to horizontal coordinate x that starts at the wall ($x=0$) and runs along the length of the beam to its free end at $x=L$.

The deflection equation for this classical problem, which involves other assumptions ([102]) is:

$$\frac{d^2}{dx^2} \left(IE \frac{d^2}{dx^2} D(x) \right) = w(x) = P\delta(x-L) \quad \text{EQ1}$$

where L is the length of the beam, E is its modulus of elasticity, I is its cross-sectional moment of inertia with respect to the loading direction, and $w(x)$ represents a generalized distributed load on the beam wherein the point load P in our example is represented by a delta function of x that recovers the point load P at $x=L$ (see [102]).

The solution of this problem is

$$D_m = PL^3/(3EI) \quad \text{EQ2}$$

where the subscript m refers to a result from the model.

Now consider a particular problem with the following parameter values and uncertainties. Let E , I , and the applied load P_o in the validation experiment be known with certainty, but let uncertainty exist in the measured length L_o of the beam; its length lies within two ruler tick marks with values L_{o-Low} and L_{o-High} that differ by 1%:

$$L_{o-High} = 1.01L_{o-Low}$$

Let the beam response model give a result that is biased 2% high relative to the actual (experimental) beam response. Let the physical and modeled beam deflections at the end of a beam of length L be respectively given by:

$$D_e = 0.98P_oL^3/(3EI) \quad \text{EQ3}$$

$$D_m = P_oL^3/(3EI). \quad \text{EQ4}$$

To simplify the problem, let $P_o/(3EI)$ have a magnitude of unity. Then the experimental and model responses are respectively given by:

$$D_e = 0.98L^3;$$

$$D_m = L^3 \quad .$$

[Figure 3](#) shows experimental and modeled cantilever beam end-deflections as functions of hypothetical values of beam length L . (Note that the abscissa in [Figure 3](#) corresponds to total beam length L and **not** to the running value x of distance along the beam.)

Now consider a validation assessment. The framework dictates that the prediction uncertainty for system response (here beam deflection) must reflect that the experimental quantity of beam length is not precisely known (it is uncertain to within the range $[L_{o-Low}, L_{o-High}]$). This length uncertainty when propagated through the model yields uncertainty in predicted beam deflection shown by the uncertainty bar in [Figure 4](#). (as obtained from [Figure 3](#)). Given this same range of length uncertainty, the experimental result could come in anywhere within the experiment uncertainty range shown in [Figure 4](#). (as obtained from [Figure 3](#)).

If the actual beam length in the experiment is between L_{o-Low} and $1.0067L_{o-Low}$, then the experimental deflection would be outside the model-predicted deflection range in [Figure 4](#). The modeler would be warned that bias exists in the model. Appropriate caveats, model correction, etc. could then be pursued if desired.

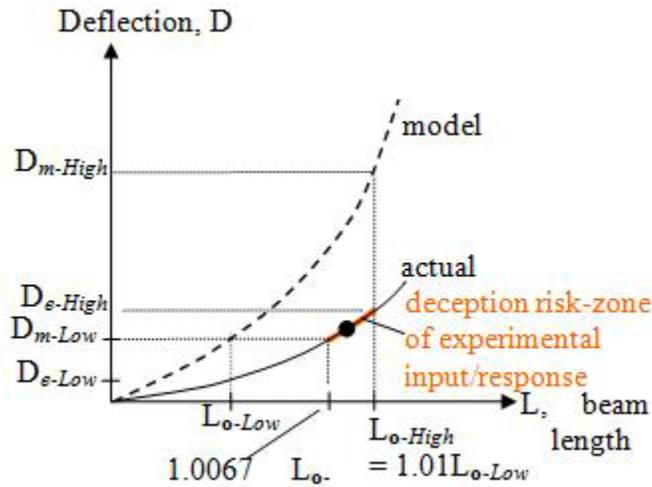


Figure 3. Beam deflection results from the physical beam and from a biased model. (Curves only illustrative.) Low and high limits of possible experimental and model results are shown for an uncertainty range $[L_{o-Low}, L_{o-High}]$ in beam length. A deception “risk zone” is shown (see also Figure 4.) where a realization of experimental input and corresponding system response would not lie outside the uncertainty of model response computed with the known input uncertainty range $[L_{o-Low}, L_{o-High}]$ in beam length.

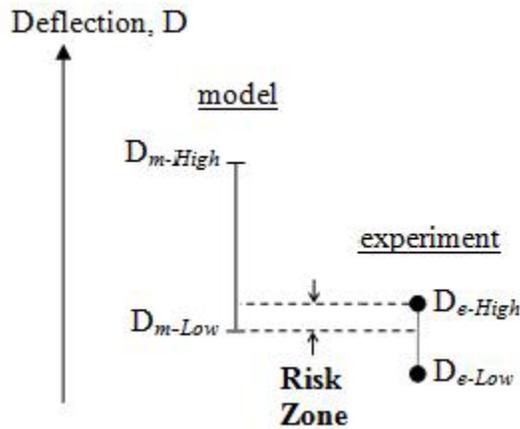


Figure 4. High and low deflection values defining ranges of possible experimental and model output consistent with the uncertainty $[L_{o-Low}, L_{o-High}]$ in beam length. The high and low response values here are those in Figure 3. A deception risk zone for Type X validation error is shown, where an experimental realization could occur, and because it lies within the uncertainty intervals of the associated model prediction, would provide no indication or warning that the model is biased, even though it actually is.

However, if the actual beam length is between $1.0067L_{o-Low}$ and L_{o-High} , then the experimental deflection will lie within the indicated deception “risk zone” in Figure 4.. This is called the risk zone because any experimental result in this zone lies within the model-predicted deflection uncertainty. Then no explicit indication of bias exists in the validation assessment even though the model is biased. Non-indication of model bias in this circumstance is termed “Type X” validation error ([47]). Type X error occurs when model bias is obscured by **systematic** uncertainty in one or more *inputs* to the experiment. (This differs from classical Type II error of incorrectly accepting a biased model or hypothesis in statistical hypothesis testing. Type II error occurs when a biased hypothesis/model is accepted, where imprecision error from sampling of **random** uncertainty of *output data* of a system obscures the fact that the model/hypothesis is biased. See [47] for more discussion.)

What about the unrevealed model bias associated with Type X validation error? Will this have downstream prediction consequences? This depends on the particulars of the physical system and model that travel to a new application setting. One set of particulars is investigated next, while a converse set of circumstances is investigated in the next section.

Let the actual beam length be $L_{o-actual}$ and let the traveling physical system be the cantilever beam used in the validation experiments. Let the applied load in the new prediction setting change to P' from the non-traveling value P_o in the validation experiment. All the other physical conditions are the same (they travel) between validation and prediction settings because the same physical beam is to be employed in the new setting, at the same temperature and wall attachment conditions as in the validation setting. Then the traveling model includes the beam deflection equation EQ1 and its solution EQ2, both specialized to the said parameter values. In the new prediction setting, accounting for the known uncertainty in beam length (same beam, so same length uncertainty [L_{o-Low} , L_{o-High}] in the validation and prediction settings) yields the following prediction bounds for deflection:

$$D_{m-Low} = P' L_{o-Low}^3 / (3EI) \quad \text{EQ5}$$

$$D_{m-High} = P' L_{o-High}^3 / (3EI) = P' (1.01L_{o-Low})^3 / (3EI). \quad \text{EQ6}$$

It can be shown by substitution that any value of actual beam length $L_{o-actual}$ within the risk range $1.0067L_{o-Low} \leq L_{o-actual} \leq L_{o-High}$ yields prediction bounds EQ5 and EQ6 that contain the experimental result in the new setting:

$$D_e = 0.98P' L_{o-actual}^3 / (3EI). \quad \text{EQ7}$$

In fact this is true for P' of any load magnitude in the new (prediction) setting—as long as the load is not high enough to cause a change in the applicable physics in the new setting such that the governing behavior equation (EQ1) no longer applies. Here the system internal state and output values change because the excitation load in the problem changes between the validation and new application settings. A linear rescaling of the results in

Figures 3 and 4 occurs as the load value changes, but the system's physical mechanisms that convert inputs to outputs (as described by the governing PDEs and material models) do not change.

Even though the model is biased here, if the equations and parameter values/uncertainties as a set bound reality in the validation setting, they also bound reality in the new prediction setting. This is because the model in the validation setting travels to the new prediction setting with a consistency that reflects the physical consistency between the two settings. This will not be the case in the next section, where the physical consistency in beam length does not travel, so related modeling aspects must be removed from the traveling model.

The demonstration here shows that biased models can have non-adverse consequences in downstream predictions under conditions where traveling consistency exists between validation and new prediction settings. That is, imperfect models (as they all are) can extrapolate reliably under certain conditions. This is fortunate because models of complex devices or systems can have multitudes of parts and materials modeled with "strong" deterministic property values and geometry descriptions that are incorrect to some degree (assuming that not all can be modeled as uncertain where actual values lie within the stipulated uncertainties). These input errors combine with physics modeling errors, such as lack of modeled contact resistance or contact friction at the part interfaces, to yield biased models to some degree. Nonetheless, such models can be productively utilized if the necessary consistency exists between validation (or conditioning) and extrapolative prediction settings.

Another type of consistency between validation (or conditioning) and prediction settings is the following. Consider a material property that varies as a function of temperature. Let the temperature range experienced by the material in the validation experiment be different from the temperature range in the new setting. If both temperature ranges are spanned by a consistent enveloping database⁹ of property characterization, then the property is deemed to be "mappably associated" between the validation and prediction settings, where a consistent representation for the phenomena spans and includes the validation and application settings.

Another type of consistency between validation (or conditioning) and application settings concerns any common-mode error in postprocessing of model and experiment results ([43], [47]). Other types of potentially consistent traveling quantities are material or component failure thresholds that are keyed e.g. to consistently biased models or discretization-dependent solution results in both settings. Section 2.2.3 cites some types of sub-grid models that rely on consistency of grid-dependent solution bias between validation and prediction settings. Approximate traveling consistency of grid related solution bias (relative to exact solution results) can exist over a modeling space and can be

9. Here a consistent enveloping database would span the temperature range exercised in the validation and the application settings, and any error in the measured property value would be contained within the published uncertainty of the data (e.g., as a \pm percentage of a nominal property value) at any given temperature.

taken advantage of to decrease computational cost of parameter space sampling in optimization and uncertainty quantification. See related discussion and empirical examples in [95], [103], [104].

4.2 Absence of Traveling Consistency between Validation and Prediction Settings, associated Extrapolation Risk from Type X Validation Error, and “Data Conditioning” to Mitigate the Risk

In this subsection prediction risk associated with model bias is shown to be consequential when traveling consistency is not present. In the example below this occurs when continuity in a physical aspect of the cantilever beam is not maintained between the validation and prediction settings.

Consider a case where the beam in the validation experiment (Beam 1) has an actual length somewhere in the risk range $1.0067L_{o-Low} \leq L_{o-actual} \leq L_{o-High}$ as portrayed in Figure 4. No model bias is indicated in the validation assessment although the model is actually biased.

Let a nominally identical second physical beam (Beam 2) be used in the new application setting. Beam 2 has a different length, but also within the same measurement and modeling uncertainty range as the validation Beam 1 (Beam 2 length also lies within the two ruler tick marks L_{o-Low} and L_{o-High}). Assume that everything else is the same between the validation and prediction settings: E , I , P_o , P' , and phenomenological behavior of the beams. Let the actual length of Beam 2 lie within the range $L_{o-Low} \leq L_{o-actual} \leq 1.0067L_{o-Low}$. Then the actual deflection in the new setting (7) is outside the predicted bounds given by equations 5 and 6. Of course, this would also be the case in the validation setting if Beam 2 was used there. Then the validation assessment would reveal an obvious bias where the experimental result lies below the bottom extent of the prediction uncertainty bar in Figure 4. This was not caught because different beams were used in the validation and application settings.

We ran into trouble here because beam length was modeled as though it flows through to the prediction setting as part of the traveling model, when really beam length changes between the two settings in changing from Beam 1 to Beam 2. Physical beam length in this example is not consistent between the validation and prediction settings even though both beam lengths lie within the modeled uncertainty range $[L_{o-Low}, L_{o-High}]$. A different example with similar implications (when a physical factor is not consistent between validation and application settings) is illustrated in [47].

Actual beam deflection in the new setting would lie within the predicted bounds (equations 5 and 6) if beam length (Beam2) happens to be between $1.0067L_{o-Low}$ and $1.01L_{o-Low}$. To avoid dependence on any such serendipity the Real Space framework takes a defensive measure for the most significant systematic uncertainties in the non-traveling aspects of the validation experiment. If it is known in advance that the beams in the

validation and prediction settings are nominally identical but not exactly the same (i.e., have characteristics only within the same range of uncertainty), then the framework would treat beam length and its other geometric and material characteristics as non-traveling. Uncertainties in these beam characteristics would be treated according to “*data conditioning*” procedures demonstrated in [47]–[52]. The procedures makes appropriate accommodations to avoid Type X error of undetected model bias in model validation and conditioning activities. The precautionary measures help avert associated adverse consequences in downstream predictions.

The data conditioning procedures address risk from model bias that can be masked by uncertainties in aspects of the validation experiments that do not travel. Those aspects designated as traveling are considered or assumed to be consistent between the validation and new prediction settings. There is always risk that some aspects designated for the traveling model will not be consistent in the new prediction setting. Beyond aspects like geometry and material properties, whether the same “physics” applies in both situations is often very open to question. Will the behavioral equations and parameter values/uncertainties that were corroborated in the validation setting apply in the new setting? In the new temperature regime, will the equation for specific heat as a function of temperature still work with the rest of the traveling model to capture reality in the new physical regime? Will the different submodels that bound model-form uncertainty for laminar-to-turbulent transition length and for turbulence behavior still adequately bound these in the new physical setting? This is the implicit hope going into prediction if the traveling aspects have been corroborated at the validation conditions. Whether this hope is reasonable depends on many things that only a diversity of experience and viewpoints may help assess and divine (see Section 6, accreditation).

Rigorously, all bets are off in extrapolation. Even if good agreement exists at the tested points, it is unknown and unknowable whether a model will hold up well in extrapolation. Whenever physical conditions change, physics modalities change. How large the effect will be on the validation quantities of interest, and how well the validation-corroborated model will account for these new modes, is very difficult to reliably predict.

Consider the potential consistency of a fire dynamics model traveling between a setting where an object does not exist in a fire and one where an object does exist within the fire. Putting the object in the fire produces significant interaction effects which add new physics to the situation. 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 turbulence generation models might not suitably handle (travel to) the new type of surface-driven phenomena, even though they might perform well in the isolated fire conditions where the fire dynamics model was initially developed and 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.

As another example, it is generally regarded that models of complex structures do not extrapolate well from the conditions at which they are calibrated. Another example familiar to the author is foam thermal pyrolysis and charring/ablation/vaporization models ([18], [28], [53] [55], [105]-[107]). The agreement between present day state-of-the-art models and actual foam behavior can vary quite sensitively at different foam densities, venting conditions, applied heat fluxes, heating rates, and spatial application of heating. These models are generally considered unreliable in extrapolation.

Thus, 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 what “significantly different” is until actual testing determines this.

Therefore it is not appropriate to claim that a model substantiated in validation settings will be valid at, or is to be considered validated for, new application settings. Instead, the framework speaks only of BEWU prediction at the new settings. This is intended to be a less misleading statement than “We’re using a validated [validation-substantiated] model for these extrapolative predictions...,” when the model was actually substantiated at different conditions.

5 Design of Model Validation (and Conditioning) Experiments and Scope of the Experiment Model

Because of the vagaries that extrapolation presents, it is essential to plan validation experiments as close as possible to the actual conditions of the applications for which the model will be used. This is of course also the best philosophy for conditioning of models. This maximizes the applicability and relevance of the validation or conditioning results to the intended applications.

In the prior example of fire modeling, where the prediction problem involves heat loads on an object in the fire, the physical regime is one where the object's presence substantially affects the fire, through its impact on the flow field and because its thermal mass has a coupling affect that moderates fire intensity. Therefore it is best to design one or more fire-model validation experiments that involve a representative object. Validation exercises would ideally be performed with objects of different size, thermal mass, and surface roughness, and locations within and apart from the fire.

Countervailing objectives and constraints often cause validation experiments to be relatively simple and far (in modeling parameter space) from the intended model application conditions. One driver is the need to control conditions in the validation experiments in order to maximize resolution power to isolate model bias. Cost and technical practicality also often drive validation experiments to be simpler than the eventual applications. In many cases, such as nuclear weapons testing and nuclear power plant accidents, tests in the intended application space are not feasible.

Balanced judgment over all the considerations involved must be applied to derive the most benefit from validation experiments and assessments. Other aspects of good experiment planning to be discussed next include:

- i. the use of modeling to help design the validation or conditioning experiments and diagnostic instrumentation;
- ii. optimizing the scope/boundaries of the experiment model (*e-model*) and of the traveling model to reduce uncertainty and increase resolution in the validation assessment.

The validation experiment does not usually dictate a unique scope for the experiment model. For example, a bank of heat lamps in an experiment warms a heat-spreader plate, which then radiates to a test object of interest. The heating conditions on the test object (traveling model) are needed. The scope of the *e-model* could stop at the heat spreader plate that radiates to the test object, or could extend further to include the banks of heat lamps in the model. It is much easier and less uncertain to paint the heat-spreader plate with an emissivity-controlling paint and instrument it with thermocouples (providing a reasonably accurate radiative heating condition applied to the test object) than to try to model the complex geometries of the heating lamps and accurately determine and specify their effective temperatures and emissivities.

As another illustration, consider a fuel-pool fire being modeled with a fire CFD code. If the fuel-level regression rate of the pool of fuel is measured in the fire experiment, then the fuel vaporization rate that feeds the fire can be calculated from fundamental principles and other experimental measurements of ambient pressure and liquid fuel pool temperature. The fuel vaporization rate could then be used as input data to the fire CFD model. Alternatively the fire CFD code may offer a modeling option to calculate fuel vaporization from inputs to the model of measured ambient pressure and initial fuel pool temperature. From these and the fire simulation's calculated heat flux to the pool surface and consequent evolution of fuel surface temperature, the fuel vaporization rate that feeds the fire can be calculated internally. Thus, various utilizations of the measured data, traded off with internal modeling, affect the scope and form of the e-model and its uncertainty.

Thus, when one sets out to “model a validation experiment”, there are often multiple options regarding the model's scope: what are the model boundaries and what is to be phenomenologically modeled versus what is to be input to the model from experimental measurement or published values or empirical relationships? Although a unique option does not exist with regard to scope of the experiment model, a best choice can be made. This is the choice that most reduces uncertainty in the validation activity, given the available freedoms in the experiments and modeling. That is, the best choice is the one that allows the smallest uncertainty in the traveling model. (This includes any uncertainty mapped to the traveling model due to uncertainty in non-traveling aspects of the experiments as demonstrated in [46], [47], [49], [56] and discussed in section 7 of this report.)

Whatever the case, the uncertainty of all the options should be assessed up front with a goal of uncertainty minimization, which may impact:

- design of the experiment, and scope of the e-model;
- measurement and sensor types, locations, and spatial and temporal resolution;
- procurement of more and/or better equipment for measurement of outputs, measurement of inputs, and/or control of inputs such as applied excitations;
- procurement of sufficient information regarding material properties, geometry, boundary conditions, equipment accuracy, calibration accuracy, etc.

Optimized experiment design in these respects is crucial. Optimized design usually benefits immensely from computational modeling and simulation. With ***model-assisted experiment design*** the experiments and planned measurements are modeled before the experiments are finalized, with the goal of identifying to the greatest extent possible the best experiment parameter settings, measurements, instrumentation, and e-model that minimize traveling-model uncertainty. Model-assisted experiment design examines the uncertainties in various experiment and modeling facets and options to help arrive at the most revealing experiments and model-experiment comparisons possible under the constraints in the model validation or conditioning project. Model-assisted experiment design includes but goes beyond the concepts of traditional statistical design of experiments (e.g. [108]). This traditionally deals with planning trials or runs to sample a parameter space of system operation for efficient discernment and analysis of process or product variation over noise variables and exploratory/control/operational variables.

6 Model Acceptance, Endorsement, Accreditation, etc.

Although validation conclusions cannot be assured to accurately project to extrapolation conditions, quantitative substantiation at validation points in the space does provide some corroborating evidence that lends credibility to the model. Such corroboration, combined with other evidence of substantive quality control and risk management in the modeling process, can be used to support a decision to tentatively accept or even formally *endorse* ([5], e.g. formally Accredited, Certified, or Qualified) the model for certain design, analysis, and decision-support uses. Acceptance or endorsement is ultimately a subjective decision based on circumstantial evidence and human judgment regarding the model's anticipated trustworthiness in resolving a given issue. Hence, technical expertise in the particular modeling realm is essential in the decision-making process, both to make a reasonable decision and to project to others the credibility of the decision. The deliberation should always weigh potential risks against benefits of model use.

Thus, model acceptance or endorsement is based on a subjective belief (leap of faith) that a model will perform sufficiently well in extrapolation to support effective resolution of issues of interest. In contrast, model validation involves direct quantitative discernment.

If a model is refuted in validation assessments at points in the parameter space near the extrapolation conditions, it is difficult to defensibly argue that it should be endorsed for making predictions there. However, since endorsement is a subjective value judgment, one cannot unequivocally say that the model should not be endorsed. If, for instance, only minimal model conditioning involving a small isolated model fix is necessary to bring the model into line with the experimental data at the tested points in the parameter space, then it might be reasonable to endorse the conditioned model for certain use purposes.

Indeed, reference [1] offers detailed considerations and formal procedures concerning model accreditation when actual validation substantiations are unavailable because experiments and data are unavailable. This is for modeling endeavors like battlefield simulations, but might also apply to things like weapon accident and nuclear power plant accident modeling. The accreditation process weighs potential risks and benefits of model use, and forms statements of what the models can reasonably be used for, e.g. "To be used only for ascertaining which of several accident perturbations appears to be worst, or to determine which model parameters affect outcomes most." Such ordinal ranking purposes tend to be fairly forgiving of model inaccuracies. For such endorsement deliberations it is crucially important to assemble appropriately diverse and credentialed experts and utilize independent peer review of their deliberation processes and conclusions.

The latter might be more aptly called 'model acceptance' rather than 'model endorsement' because of the greater reassurance of prediction quality implied by 'endorsement' (or 'accreditation' or 'certification'). As an engineering example, consider the task of modeling 3-D heat diffusion through an object. Let the thermal conductivity to be used (assumed to be isotropic) come from a material property characterization using a 1-D version of the heat equation and a heated rod of the same material as the object. Even if the thermal conductivity is characterized in the 1-D activity over the relevant temperature range for

the 3-D application conditions, there could be significant inconsistencies between the 1-D and 3-D settings that would cause the thermal conductivity properties to not travel appropriately. Lack of isotropy of the material property is one possible issue; another is that mismodeled phenomena like convective and radiative boundary conditions in the 1-D parameter estimation activity could yield incorrect property values that would manifest as modeling error in 3-D settings. Nonetheless, the author's experience is that extrapolations of this nature between the 1-D and 3-D settings are generally low-risk. Hence, the material property model might be accepted for cautionary 3-D use, but endorsement might involve much greater formalism—closer examination of the 1-D characterization activity; testing for isotropy; etc.

7 Model Conditioning

What if the model does not meet the validation accuracy criteria? It is presumed that generally the model is the best one that can be afforded and obtained, given that model validation activities are fairly complex and expensive endeavors—not applicable at will to just any model. Hence, the model will normally not be rejected and abandoned if it does not meet the validation accuracy criteria. Rather, as the best available codification of knowledge of the physics and solution approach for the problem, efforts may be made to better reconcile the model with reality. Generally the model or some modification of it will be leveraged for prediction needs, even if with lowered expectations and reduced prediction duties and domain of application.

Reconciliation can come from an investigation of experimental and modeling factors perceived to contribute most to the failure to meet the acceptance criteria, with remediation of the factors as feasible. This can take the form of more accurate and precise measurement and control of experimental inputs and conditions in the experiments; more experiments if experimental uncertainty is being driven by a small number of trials of stochastically varying phenomena; improved data processing procedures; reducing the discretization-related uncertainties in the model solutions, etc.

If the remediation actions do not fully reconcile things, then it may be decided that the best chance for success in upcoming uses of the model is to *condition* it to match the validation data as well as possible. It is illustrated in [46] that this can reduce error in extrapolative prediction; the conditioned model will be more accurate in at least a local neighborhood of extrapolative prediction, and this advantage may extend to larger extrapolations.

Hence, even if the model is found inadequate, model conditioning can be used to extract value from the validation characterization before going forward with predictions. Model conditioning applied to the center and rightmost categories in Figure 1 would bias correct and/or add uncertainty to the model to yield results that match the uncertainty range of the experimental data as closely as can be achieved. The left-most category of results in Figure 1 may also require or benefit from model conditioning, especially if the uncertainty carried in the model is excessive such that the model predictions are very suboptimal or effectively non-useful.

What is the best procedure for conditioning a model? There is no single approach that works best in all circumstances. Approaches in at least the following two categories exist. A combination of these approaches can also be used.

Approach 1: correction or uncertainty “layer function” superposed-on or scaled-to the prediction results of the unaltered model (also called “add factoring” [70]). Reference [46] summarizes a project app. where Approach 1 worked and Approach 2 was not suitable.

Approach 2: adjust or ‘*calibrate*’ model parameter values to correct the output results. A simple approach is demonstrated in [47] and [56], where *interval* uncertainty is mapped

to the calibration parameters. The approach is simpler and less expensive than approaches like Bayesian calibration and Maximum Likelihood estimation, which provide uncertainty *distributions* on the calibrated parameter values ([14], [63]). (The term ‘calibration’ also applies to methods that yield point values for the calibration parameters such that model response matches e.g. the mean of the experimental data, or yields the best fit to the data in a Least Squares sense. For purposes of model conditioning as defined in this report, the calibration problem involves addressing the full span of the experimental uncertainty.)

The best model parameter(s) to manipulate in a given model conditioning circumstance can depend on the particular model extrapolations in mind and are a matter of expert judgment in the model and the physics of the problem. Extrapolation in one direction in the parameter space might be best achieved by manipulating one set of parameters, while extrapolation in another direction might be best accomplished with a different set. Alternatively, a correction layer (Approach 1) might work best in the given circumstance.

If the model will be used for extrapolations in a number of different directions in parameter space, then one desires conditioning that is robust over the various extrapolations. Selection of the optimal parameters and method to best accomplish a given model conditioning purpose is presently more an art than a science.

A constraint on the allowable set of conditioning parameters is that they must be parameters of the traveling portion of the E-model. Often, the most versatile traveling parameters for model conditioning are material property and constitutive model parameters and other physics submodel parameters.

It is somewhat common in the literature to blur distinctions between model validation (which tests predictive ability of the model in an extrapolative setting where the validation data has not been used to condition the model) and model conditioning or “updating” where the model is adjusted to match the validation data. There are many papers in the literature that affix the label ‘model validation’ to procedures like Bayesian and Maximum Likelihood model calibration and parameter estimation¹⁰. As discussed in [46], this distinction can be somewhat arbitrary under certain conditions such as interpolative use of the model within its calibration data base (e.g. the application in [43]). However, it is necessary to make and emphasize the distinction when the intended use of the model is extrapolatory, in which case validation implies testing the model’s predictive ability in meaningful extrapolations away from the calibration conditions.

10. Parameter estimation is algorithmically similar to model calibration. However, parameter estimation is considered by the author to occur at the modeling stage where a parameter's value is initially characterized and represents a physically related parameter in the modeling equations. If the parameter value is subsequently manipulated to make model predictions better match experimental data in point-of-use settings where model deficiency shows up, then such readjustment or updating is viewed as model calibration instead of parameter estimation. When a parameter does not represent a physically related parameter in the modeling scheme, but instead exists as one or more non-physical degrees of freedom in the model that are adjusted to attain model agreement with data, then this is considered to be model calibration (not parameter estimation). However, these distinctions are somewhat subtle and are not universally agreed upon.

8 Closing

Modeling and simulation are increasingly being utilized and depended upon for analysis, design, and decisions of consequence. Therefore it is increasingly important to develop and implement quality control procedures and risk management in modeling and prediction. Associated methodologies are presently in a vigorous state of research and development within and outside Sandia. A large variety of viewpoints and precedents were surveyed in this report, as were constraints and difficulties in devising a viable methodology that accommodates the end-to-end experiments-to-prediction problem. New concepts, definitions, and terminology were presented to address perceived gaps.

A novel Real Space approach was presented and discussed. Rationale was given for the various choices taken in developing the Real Space approach. The approach is driven by pragmatic requirements of reasonable simplicity and straightforwardness; affordability of the methodology and procedures; limited data availability and versatility of the types of uncertainty that must be accommodated; and readily interpretable results for decision makers who are non-specialists in modeling and uncertainty quantification.

An important contribution of the approach is its end-to-end project perspective. Workable process elements are required. The Real-Space framework's pragmatic discrepancy measures and acceptance criteria and associated uncertainty representation and processing machinery are the same whether the purpose is model conditioning or model validation. Other methodologies in the literature appear to address only isolated portions of the end-to-end problem.

Although validation assessments can only yield limited circumstantial corroboration of model predictiveness, carefully designed and executed assessments can add greatly to the modelers' and the model users' knowledge about the performance of the model. This is an essential element of due diligence and applied risk management in developing and using models. Likewise, although accuracy of predictions cannot be guaranteed, contextualizing and improving the predictions as much as possible through appropriate methodology should be pursued. We can seek to maximize accuracy potential through careful design and analysis of experiments and model development, validation, and extrapolation procedures optimized toward the desired prediction tasks. Additionally, modeling risk can be assessed through InfoGap type analyses [111] to determine the degree to which a model can be incorrect before changing the conclusion obtained with it (see e.g. [44]). This having been said, quality assessment and control in modeling and simulation are in the very early stages of development. Much still needs to be done to bring these young engineering sciences to maturity.

References

- [1] Department of Defense Instruction 5000.61: *Modeling and Simulation Verification, Validation, & Accreditation (VV&A)*, Defense Modeling and Simulation Office, Office of the Director of Defense Research and Engineering, dated April 29, 1996, www.dmsso.mil/docslib.
- [2] American Institute of Aeronautics and Astronautics, *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, AIAA G-077-1998, AIAA, Reston, Va.
- [3] Department of Energy – Defense Programs, *Accelerated Strategic Computing Initiative Program Plan*, 2000, DOE/DP-99-000010592.
- [4] American Society of Mechanical Engineers, V&V 10 – 2006 *Guide for Verification and Validation in Computational Solid Mechanics*, available from ASME Codes & Standards website.
- [5] NASA, *Technical Standard for Models and Simulations*, report NASA-STD-7009, released July 11, 2008.
- [6] American Society of Mechanical Engineers, V&V 20 – 2009 *Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer*, available from ASME Codes & Standards website.
- [7] American Society of Mechanical Engineers, V&V 10 – draft document in preparation illustrating the concepts of V&V in an end-to-end computational solid mechanics example
- [8] Babuska, I., F. Nobile, R. Tempone, “Reliability of Computational Science,” *Num. Methods in Partial Differential Equations*, (2007) Vol. 23, pp. 753-784.
- [9] Babuska, I., J.T. Oden, M. Papadrakakis “Verification and Validation in Computational Engineering and Science: Basic Concepts,” *Comput. Methods in Applied Mechanics and Engrng.*, Vol. 193, 2004, pp. 4057-4066.
- [10] Balci, O., “Verification, Validation and Testing,” *Handbook of Simulation*, ed. J. Banks, Engineering and Management Press (1998), pp. 335-393.
- [11] Bayarri, M., J. Berger, P. Rui, J. Sacks, J. Cafeo, J. Cavendish, C-H. Lin, J. Tu, “A Framework for Validation of Computer Models,” *Technometrics*, May 2007, Vol. 49, no. 2, pp. 138 - 154.
- [12] Black, A. R., Hobbs, M. L., Dowding, K. J., Blanchat, T. K., 2007, “Uncertainty Quantification and Model Validation of Fire/Thermal Response Predictions”, 18th AIAA Computational Fluid Dynamics Conference, Miami, FL, June 25-28, paper AIAA-2007-4204.
- [13] Blottner, F.G., “Accurate Navier-Stokes Results for the Hypersonic Flow over a Spherical Nosedip,” *AIAA J. Spacecraft and Rockets*, Vol. 27, No.2, March-April 1990, pp. 113-122.

- [14] Campbell, K., "A Brief Survey of Statistical Model Calibration Ideas," International Conference on Sensitivity Analysis of Model Output, March 8 – 11, 2004, Santa Fe, NM.
- [15] Calder, A., B. Fryxell, T. Plewa, R. Rosner, L. Dursi, V. Weirs, T. Dupont, H. Robey, J. Kane, B. Remington, R. Drake, G. Dimonte, M. Zingale, F. Timmes, K. Olson, P. Ricker, P. MacNeice, H. Tufo, "On Validating an Astrophysical Simulation Code," submitted to *Astrophysical Journal*.
- [16] Chen, W., Y. Xiong, K-L. Tsui, S. Wang, "Some Metrics and a Bayesian Procedure for Validating Predictive Models in Engineering Design," ASME Design Technical Conference, Design Automation Conf., Philadelphia, PA, Sept. 10-13, 2006.
- [17] Coleman, H.W., and Stern, F., "Uncertainties in CFD Code Validation," *Journal of Fluids Engineering*, Dec. 1997, vol. 119, pp. 795-803.
- [18] 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.
- [19] Easterling, R. G., "Measuring the Predictive Capability of Computational Models: Principles and Methods, Issues and Illustrations," Sandia National Laboratories report, SAND2001-0243, Unlimited Release, February 2001.
- [20] Ferson, S., W.L. Oberkampf, L. Ginzburg, "Model Validation and Predictive Capability for the Thermal Challenge Problem," *Comput. Methods in Applied Mechanics and Engrng.*, Vol. 197, 2009, pp. 2408 – 2430.
- [21] Hanson, K.M., and Hemez, F.M., 2001, "A Framework for Assessing Confidence in Computational Predictions," *Experimental Techniques*, Vol. 25.
- [22] Hasselman, T., G. Wathugala, T. Paez, A. Urbina, "Comparison of Top-Down vs. Bottom-Up Approaches for Uncertainty Quantification and Predictive Accuracy Assessment of Computational Mechanics Models," Proceedings of the 9th International Conference on Structural Safety and Reliability (ICOSSAR), Rome, Italy, June 19-23, 2005.
- [23] Hasselman, T., G. Lloyd, T. Hinnerichs, C. O’Gorman, A. Urbina, "Validation of Uncertainty Quantification For Structures with Encapsulated Epoxy Foam," 8th AIAA Non-Deterministic Analysis Conference, Newport, RI, May 1-4, 2006.
- [24] 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.
- [25] Hills, R.H., and T.G. Trucano, "Statistical Validation of Engineering and Scientific Models: Background," Sandia National Laboratories Report SAND99-1256, printed May 1999.

- [26] Hills, R.G., and Dowding, K.J., "Statistical Validation of Engineering and Scientific Models: Bounds, Calibration, and Extrapolation," Sandia National Laboratories document SAND2005-1826, April 2005.
- [27] Hills, R.G., "Model Validation: Model Parameter and Measurement Uncertainty," *ASME J. Heat Transfer*, April 2006, Vol. 128, pp. 339 - 351.
- [28] Hobbs, M.L., and Romero, V.J., "Uncertainty Analysis of Decomposing Polyurethane Foam," *Thermochemica Acta*, 384 (2002), 393-401
- [29] Horta, L.G., S.P. Kenny, L.G. Crespo, K.B. Elliott, "NASA Langley's approach to the Sandia's structural dynamics challenge problem," *Comput. Methods in Applied Mechanics and Engrng.*, Vol. 197, 2009, pp. 2607-2620.
- [30] Kleindorfer, G.B., L. O'Niell, R. Ganeshan, "Validation in Simulation: Various Positions in the Philosophy of Science," *Management Science*, Vol. 44, No. 8, August 1998, pp. 1087-1099.
- [31] Lew, J-S., "Model Validation Using Interval Modeling with Performance Sensitivity," paper AIAA-2007-2348, 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 23 - 26 April, 2007, Honolulu, HI.
- [32] Logan, R.W., Nitta, C.K., and Chidester, S. K., "Uncertainty Quantification During Integral Validation: Estimating Parametric, Model Form, and Solution Contributions," 8th AIAA Non-Deterministic Analysis Conference, May 1-4, 2006, Newport, RI.
- [33] McFarland, J., S. Mahadevan, V. Romero, L. Swiler, "Calibration and Uncertainty Analysis for Expensive Computer Simulations with Multivariate Output," *AIAA Journal*, vol. 46, no. 5, May 2008, pp. 1253-65.
- [34] Miser, H.J., and E.S. Quade, "Validation," Chapter 13 in *Handbook of Systems Analysis: Craft Issues and Procedural Choices*, North-Holland, New York, 1988, pp. 527-565.
- [35] Oberkampf, W.L., and Trucano, T.G., 2002, "Verification and Validation in Computational Fluid Dynamics," *Progress in Aerospace Sciences*, 38(3), pp. 209-272.
- [36] Oberkampf, W.L., T.G. Trucano, C. Hirsch, "Verification, Validation, and Predictive Capability in Computational Engineering and Physics," Sandia National Laboratories Report SAND2003-3769, printed Feb. 2003.
- [37] 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.
- [38] Oberkampf, W.L., and Roy, C.J., *Verification and Validation in Scientific Computing*, Cambridge University Press, 2010.
- [39] Popper, K., *Conjectures and Refutations: The Growth of Scientific Knowledge*, Basic Books, New York, 1963.

- [40] Ricks, A., V. Nicolette, V. Romero, W. Erickson, J. Hewson, "Fuego Solid-Propellant Fire Model Verification and Validation," Sandia National Laboratories report in preparation.
- [41] Roache, P.J., *Fundamentals of Verification and Validation*, Hermosa Publishers, Socorro, NM, 2009.
- [42] Romero, V.J., "On Model Validation and Extrapolation for Best-Estimate-Plus-Uncertainty Predictions," Sandia National Laboratories document SAND2005-7678C, November 2005.
- [43] 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, 45th AIAA/ASME/ ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, April 18-21, 2005, Austin, TX.
- [44] Romero V.J., "Issues and Needs in Quantification of Margins and Uncertainty (QMU) for Phenomenologically Complex Coupled Systems," paper AIAA-2006-1989, 8th AIAA Non-Deterministic Analysis Conference, Newport, RI, May 1-4, 2006.
- [45] Romero, V.J., "A Paradigm of Model Validation and Validated Models for Best-Estimate-Plus-Uncertainty Predictions in Systems Engineering," paper 2007-01-1746 for Soc. Automotive Engineers 2007 World Congress April 16–20, Detroit, Michigan.
- [46] Romero, V.J., "Validated Model? Not So Fast. The Need for Model 'Conditioning' as an Essential Addendum to Model Validation," paper AIAA-2007-1953, 9th Non-Deterministic Approaches Conference, Honolulu, HI, April 23-26, 2007.
- [47] Romero, V.J., "Type X and Y Errors and Data & Model Conditioning for Systematic Uncertainty in Model Calibration, Validation, and Extrapolation," SAE paper 2008-01-1368 for Society of Automotive Engineers 2008 World Congress, April 14-17, 2008, Detroit, MI.
- [48] Romero, V.J., "Data & Model Conditioning for Multivariate Systematic Uncertainty in Model Calibration, Validation, and Extrapolation," paper AIAA-2010-2511, 12th AIAA Non-Deterministic Approaches Conference, April 12-15, 2010, Orlando, FL.
- [49] Romero, V.J., A. Luketa, M. Sherman, "Application of a Versatile "Real Space" Validation Methodology to a Fire Model" *AIAA J. of Thermophysics and Heat Transfer*, Vol. 24, No. 4, Oct. – Dec. 2010, pp. 730-744.
- [50] Romero, V.J., "Comparison of Several Model Validation Conceptions against a "Real Space" End-to-End Approach," Soc. Automotive Engrs. *Intn'l. J. of Materials and Manufacturing*, June 2011.
- [51] Romero, V.J., J.F. Dempsey, G. Wellman, B. Antoun, M. Sherman, "Model Validation and UQ Techniques applied to a Temperature Dependent Stainless-

- Steel Constitutive Model tested on Heated Pipes Pressurized to Failure,” Sandia National Laboratories report in preparation.
- [52] Romero, V.J., “An Applied Real-Space Framework for Characterizing and Aggregating Uncertainty in Experiments and Model Validation Assessments,” Sandia National Laboratories report in preparation. See reference [112] for summary presentation available.
- [53] Romero, V.J., Shelton, J.W., and Sherman, M.P., “Modeling Boundary Conditions and Thermocouple Response in a Thermal Experiment,” 2006 ASME Int’l. Mechanical Engineering Congress and Exposition, Nov. 5-10, 2006, Chicago, IL.
- [54] Roy, C.J., and Oberkampf, W.L., “A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing,” *Computer Methods in Applied Mechanics and Engineering*, 200 (2011).
- [55] 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.
- [56] Rutherford, R., V. Romero, J. Castro, R. Hoekstra, “Methods of Uncertainty Quantification, Model Calibration, and Prediction for the QASPR Silicon Device Prototype Demonstration,” Sandia National Laboratories report SAND2011-7940 printed November 2011.
- [57] Schlesinger, S., R. Crosbie, R. Gagne, G. Innis, C. Lalwani, J. Loch, R. Sylvester, R. Wright, N. Kheir, D. Bartos, “Terminology for Model Credibility,” report of the Technical Committee on Model Credibility to the general membership of the Society for Modeling and Simulation International (SCS), *Simulation*, 1979, Vol. 32, pp. 103-104.
- [58] Stern, F., Coleman, H.W., Wilson, R.V., Paterson, E.G., “Verification and Validation of CFD Simulations,” Proceedings of the 3rd ASME/JSME Joint Fluids Engineering Conference, July 18-23, 1999, San Francisco, CA.
- [59] Trucano, T.G., M. Pilch, W.L. Oberkampf, “General Concepts for Experimental Validation of ASCII Code Applications,” Sandia National Laboratories Report SAND2002-0341, printed March 2002.
- [60] 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 (2006), No. 11, pp. 1331-1357.
- [61] 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.
- [62] Swiler, L.P., B.M. Adams, M.S. Eldred, “Model Calibration under Uncertainty: Matching Distribution Information,” paper AIAA-2009-5944, 11th AIAA Non-Deterministic Approaches Conference, 4-7 May 2009, Palm Springs, CA.

- [63] Xiong, W., W. Chen, K-L. Tsui, D. Apley, "A better understanding of model updating strategies in validating engineering models," *CMAME* 198 (2009) 1327-1337.
- [64] Konikow, L.F., and Bredehoeft, J.D., "Groundwater models cannot be validated," *Advances in Water Resources*, Vol. 15, 1992, pp. 75-83.
- [65] De Marsily, G., P. Combes, P. Goblet, "Comment on 'Ground-water models cannot be validated' by L.F. Konikow & J.D. Bredehoeft (*Advances in Water Resources*, Vol. 15, 1992, pp. 75-83)," *Advances in Water Resources*, Vol. 15, 1992, pp. 367-369.
- [66] Leijnse, A., and Hassanizadeh, S.M. "Model definition and model validation," Short Communication in *Advances in Water Resources*, Vol. 17, 1994, pp. 197-200.
- [67] Tsang, Chin-Fu, "The Modeling Process and Model Validation," *Ground Water*, Vol. 29, No. 6, Nov.-Dec. 1991, pp. 825-831.
- [68] Oreskes, N., K. Shrader Frechette, K. Belitz, "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences," *Science*, Vol. 263, Feb. 1994, pp. 641-646.
- [69] Refsgaard, J.C., and Henriksen, H.J., "Modeling Guidelines-Terminology and Guiding Principles," *Advances in Water Resources*, Vol. 27, 2004, pp. 4057-4066.
- [70] Stermann, J. D., Letter to the Editor, in *Science* Vol. 264, April 1994, re: "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences," by Oreskes et al., *Science*, Vol. 263, Feb. 1994, pp. 641-646.
- [71] Anderson, M.G., and Bates, P.D., editors, *Model Validation—Perspectives in Hydrological Science*, Wiley, 2001.
- [72] Zeigler, B.P., *Theory of Modeling and Simulation*, John Wiley & Sons, 1976.
- [73] Mehta, U.B., "Guide to Credible Computer Simulations of Fluid Flows," *Journal of Propulsion and Power*, vol. 12, no. 5, Sept.-Oct. 1996, pp. 940-948.
- [74] Rebba, R, Mahadevan, S., "Computational methods for model reliability assessment," *Reliability Engineering and System Safety*, Vol. 93 (2008), pp. 1197-1207.
- [75] Sargent, R.G., "Validation of simulation models," Proc. Winter Simulation Conf., San Diego, CA, December 1979, 497-503.
- [76] Sargent, R.G., "Verification and validation of simulation models," Chapt. IX in *Progress in Modelling and Simulation*, ed. F.E. Cellier, Academic Press, London, 1982.
- [77] Balci, O., and Sargent, R.G., "A methodology for cost-risk analysis in the statistical validation of simulation models," *Commun. of the ACM*, 24 (6) April 1981, 190-197.
- [78] Balci, O., and Sargent, R.G., "Bibliography on validation of simulation models," Newsletter -TIMS College on Simulation and Gaming, 4 (2), Spring 1980, 11-15.

- [79] Balci, O., "Verification, Validation, and Accreditation of Simulation Models," Proc. 1997 Winter Simulation Conf., ed. S. Andradottir, K.J. Healy, D.H. Withers, B.L. Nelson, 135-141.
- [80] Balci, O., "Verification, Validation, and Testing," Chapt. 10 in *The Handbook of Simulation*, ed. J. Banks, 1997, John Wiley & Sons.
- [81] Balci, O., "Quality assessment, verification, and validation of modeling and simulation applications," Proc. 2004 Winter Simulation Conf., ed. R.G. Ingalls, M.D. Rossetti, J.S. Smith, H.A. Peters, IEEE Piscataway, NJ, 122-129.
- [82] Sargent, R.G., "Verification and Validation of Simulation Models," Proc. 2005 Winter Simulation Conf., ed. M.E. Kuhl, N.M. Steiger, F.B. Armstrong, J.A. Joines, 130-143.
- [83] Boehm, B.W., *Software Engineering Economics*, Prentice-Hall, 1981.
- [84] Ferson, S., and L. Ginzburg, (1996) "Different Methods are Needed to Propagate Ignorance and Variability," *Reliability Engineering and System Safety*, Vol. 54, No. 11, pp. 133-144.
- [85] Liu, Y., W. Chen, P. Arendt, H-Z Huang, "Towards a Better Understanding of Model Validation Metrics," paper AIAA-2010-9240, *13th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Sept. 12-15, 2010, Fort Worth, TX.
- [86] Ben-Haim, Y., F.M. Hemez, "Robustness, Fidelity and Prediction-Looseness of Models," paper ESDA 2004-58008, Proc. of ESDA04, 7th Biennial Conf. on Engr. Sys. Design and Analysis, July 19-22, 2004, Manchester, United Kingdom. Revised and extended version submitted Jan. 2011 to *Royal Society Proceedings A*, as Los Alamos National Laboratories document LA-UR-11-0497.
- [87] Coleman, H.W., and Steele, Jr., W.G., *Experimentation and Uncertainty Analysis for Engineers ? 2nd Edition*, John Wiley & Sons, New York, NY, 1999.
- [88] Kline, S.J., and McClintock, F.A., "Describing Uncertainties in Single-Sample Experiments," *Mechanical Engineering*, Jan. 1953, pp. 3 - 8.
- [89] Domino, S.P., G. Wagner, A. Luketa, A. Black, J. Sutherland, "Verification for Multi-Mechanics Applications," 9th AIAA Non-Deterministic Methods Conference, April 23 - 26, 2007, Honolulu, Hawaii.
- [90] Eça, L., Hoekstra, M., "An Evaluation of Verification Procedures for CFD Applications," 24th Symposium on Naval Hydrodynamics, Fukuoka, Japan, 8-13 July 2002.
- [91] Hemez, F., and Marcilhac, M., "A Fresh Look at Mesh Refinement in Computational Physics and Engineering," paper AIAA-2008-2153, 10th AIAA Non-Deterministic Methods Conference, 7 - 10 April 2008, Schaumburg, IL.
- [92] Knupp, P.M., and Salari, K., *Verification of Computer Codes in Computational Science and Engineering*, Chapman & Hall/CRC Press, 2003.

- [93] Maupin, R., "Calculation Verification for Threaded Assembly," IMAC-XXIV Conference on Structural Dynamics, Jan. 30 – Feb. 2, 2006, St. Louis, MO.
- [94] Roache, P.J., "Verification of Codes and Calculations," paper AIAA-95-2224, 26th AIAA Fluid Dynamics Conference, San Diego, California, June 19-23, 1995.
- [95] Romero, V.J., "Model-Discretization Sizing and Calculation Verification for Multipoint Simulations over Large Parameter Spaces," paper AIAA2007-1953, 9th AIAA Non-Deterministic Methods Conference, April 23 - 26, 2007, Honolulu, HA.
- [96] Roy, C.J., Raju, A., Hopkins, M.H., "Estimation of Discretization Errors Using the Method of Nearby Problems," *AIAA Journal*, Vol. 45, No. 6, June 2007.
- [97] Salas, M.D., "Some Observations on Grid Convergence," *Computers & Fluids*, Jan. 2006.
- [98] Salas, M.D., Atkins, H.L., "On Problems Associated with Grid Convergence of Functionals," submitted to *Computers & Fluids*.
- [99] Stewart, J.R., Gullerud, A.S., Heinsteins, M.W., "Solution Verification for Explicit Transient Dynamics Problems in the presence of Hourglass and Contact Forces," *Comput. Meths. Appl. Mech. Engrg.* 195 (2006) 1499-1516.
- [100] Dept. of Energy, *Predictive Science Academic Alliance Program-II (PSAAP-II) Verification, Validation, and Uncertainty Quantification Whitepaper*, jointly issued by Lawrence Livermore, Los Alamos, and Sandia National Laboratories as Lawrence Livermore National Laboratory publication LLNL-MI-481471, May, 2011.
- [101] Romero, V., L. Swiler, A. Urbina, "An Initial Comparison of Methods for Representing and Aggregating Experimental Uncertainties involving Sparse Data," paper AIAA-2011-1705, 13th AIAA Non-Deterministic Approaches Conference, April 4-7, 2011, Denver, CO.
- [102] Byars, E.F., Snyder, R.D., and Plants, H.L., 1983, *Engineering Mechanics of Deformable Bodies*, Harper & Row Publishers, NY.
- [103] Romero, V. J., "Efficient Global Optimization Under Conditions of Noise and Uncertainty - A Multi-Model Multi-Grid Windowing Approach," in the Proceedings of the 3rd WCSMO (World Congress of Structural and Multidisciplinary Optimization) Conference, Amhearst, NY, May 17-21, 1999.
- [104] Romero, V.J., "Characterization, Costing, and Selection of Uncertainty Propagation Methods for Use with Large Computational Physics Models," paper AIAA-2001-1653 presented at the 42nd Structures, Structural Dynamics, and Materials (SDM) Conference, Seattle, WA, April 16-19, 2001.
- [105] Chu, T.Y., M.L. Hobbs, K.L. Erickson, T.A. Ulibarri, A.M. Renlund, W. Gill, L.L. Humphries, T.T. Borek, "Fire-induced Response in Foam Encapsulants," in Proceedings SAMPE 1999-44th Intn'l. SAMPE Symposium & Exhibition (1999).

- [106] K.L. Erickson, S.M. Trujillo, K.R. Thompson, A.C. Sun, M.L. Hobbs, K.J. Dowding, “Liquefaction and flow behavior of a thermally decomposing removable epoxy foam,” *Computational Methods in Materials Characterization*, Southampton, Boston: WIT Press, 2004.
- [107] Hobbs, M.L., “Modeling epoxy foams exposed to fire-like heat fluxes,” *Polymer Degradation and Stability*, Vol. 89 (2005), pp. 353 – 372.
- [108] Box, G.E.P., Hunter, W.G., Hunter, J.S., *Statistics for Experimenters—An Introduction to Design, Data Analysis, and Model Building*, Wiley & Sons, 1978.
- [109] Du, L., K.K. Choi, B.D. Youn, “Inverse Possibility Analysis Method for Possibility-Based Design Optimization,” *AIAA Journal*, Vol. 44, No. 11, Nov. 2006, pp. 2682 – 2690.
- [110] Eldred, M.S., and Swiler, L.P., “Efficient Algorithms for Mixed Aleatory-Epistemic Uncertainty Quantification with Application to Radiation-Hardened Electronics, Part 1: Algorithms and Benchmark Results,” Sandia National Laboratories report SAND2009-5805, printed Sept. 2009.
- [111] Ben-Haim, Y., *Information Gap Decision Theory—Decisions under severe uncertainty*, Academic Press, London, 2001.
- [112] Romero, V.J., “Comparison of a Pragmatic and Versatile Real-Space Model Validation Framework against Several Other Frameworks,” Sandia National Laboratories document SAND2011-7613C presented at the Intn’l. Workshop on Verification and Validation in Computational Science,” Notre Dame University, South Bend, IN, Oct. 17-19, 2011.

Appendix A: Summary Comparison of Real Space Validation Approach vs. Two Other Established Frameworks

The following figures are extracted from a slide presentation ([112]) that compares the Real Space validation approach against established validation frameworks [6] and [54]. A number of advantages of the Real Space approach are shown. See [112] for more details.

Real Space vs. Transform Space Representations of Model Discrepancy

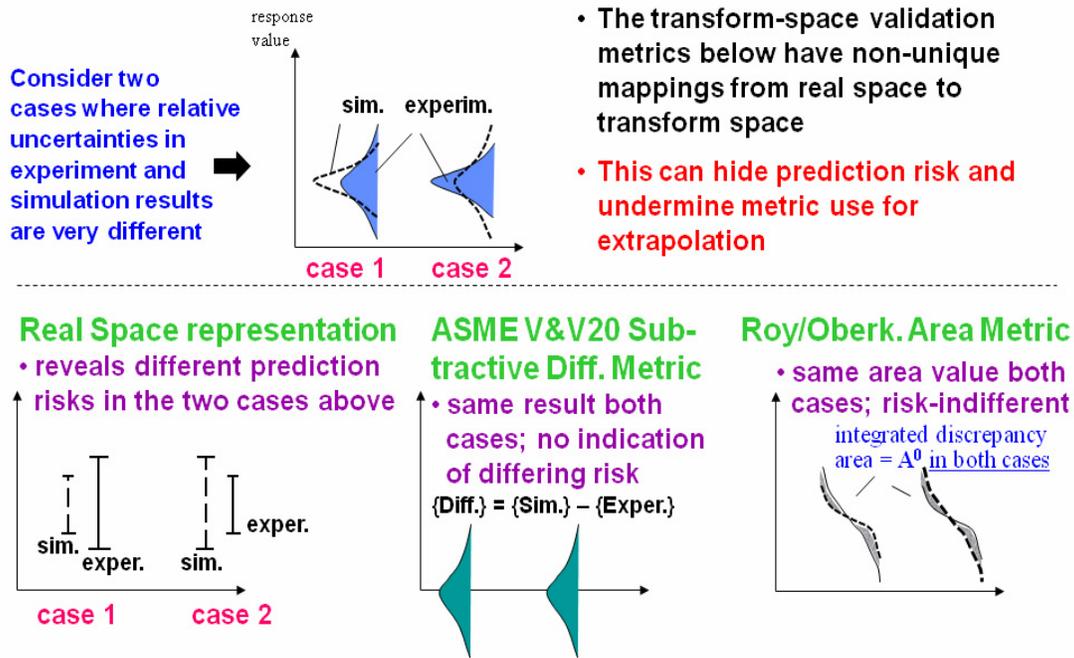


Figure A.1. Real Space vs. Transform Space Representations of Model Discrepancy

Real Space vs. Transform Space —Support for Extrapolation

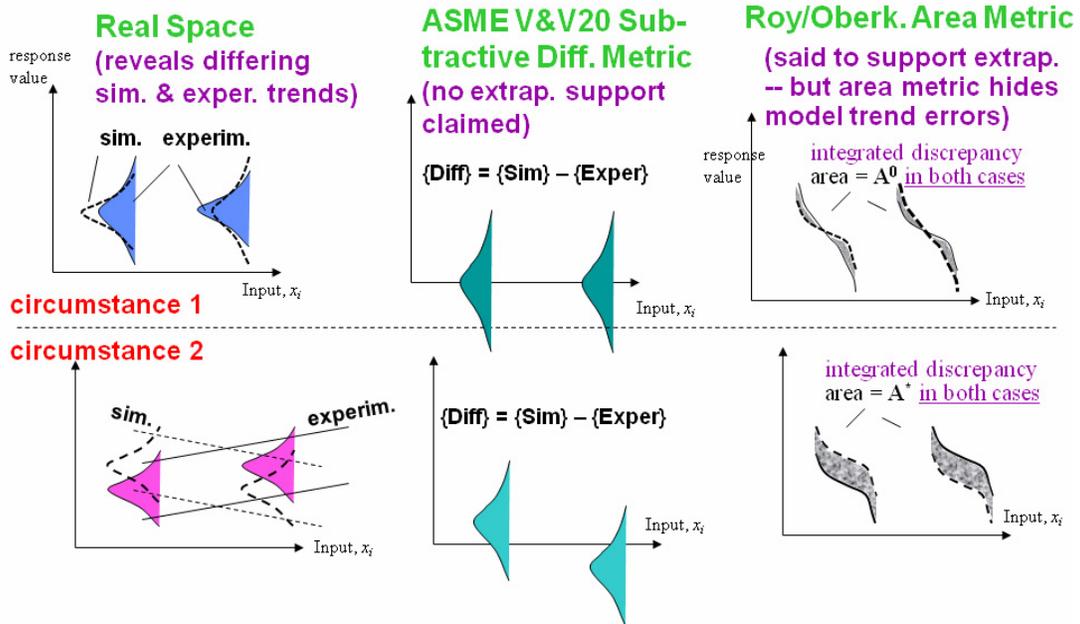
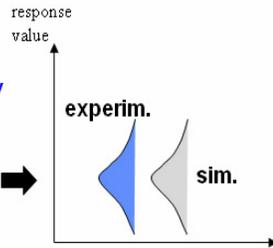


Figure A.2. Real Space vs. Transform Space Discrepancy Metric Support for Extrapolation

Subtraction Metric prevents proper handling of some types of Random Variability in a population of repeated experiments



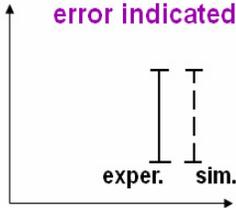
E.g., let simulated stochastic variability of system exactly equal variability of real system tested many times



- **Conditions:** no measurement errors in the experiments; and “large” # of tests
- Observed response variability is due to unit-to-unit stochastic variability of the tested systems
- and/or due to variability of experimental input conditions
- variability sources independently characterized for simulations, and play out consistently in experims.

Real Space approach

✓ works; no model error indicated



ASME V&V20 Subtractive Diff. Metric

• exaggerates uncertainty re. model bias

$$\{Diff\} = \{Sim\} - \{Exper\}$$



Roy/Oberk. Area Metric

✓ works; no model error indicated

Figure A.3. ASME V&V-20 Subtractive-Difference Metric limitation for Some Types of Random Variability in Repeated Experiments

Model Builder's Risk vs. Model User's Risk with respect to systematic uncertainty in experiment conditions



- Where place burden of proof?
 - **Optimistic** stance: Assume model is unbiased if don't have positive proof that it is biased (outside expr. uncer.)
 - **Roy & Oberkampf**: “When the simulation is a P-Box due to insufficient information provided by the validation experiment, the model is given more leeway in comparing with the experiment, *as is appropriate.*”
 - “**Free lunch**”— the more experimental uncertainty, the more model-bias leeway allowed
 - eliminates Model Bldr.’s risk but increases Model User risk
 - **Conservative**: Treat model as potentially biased up to magnitude allowed by resolution uncertainty in exers.
 - **Real Space** – reduces Model User’s risk but increases Model Builder’s risk
 - **ASME V&V20** – similar

Figure A.4. Roy & Oberkampf skew toward Model User’s Risk arising from Systematic Uncertainties in Experimental Input Conditions

Frameworks Comparison Summary



Demonstrated Capability/Feature	Real Space	ASME V&V-20	Roy & Oberkampf
Discrimination of Prediction Risk and assoc. Extrapolation Support	✓	✓ ⁻	✓ ⁻
Model adequacy criterion	✓	✓ ⁻	—
Traveling Model vs. Experiment Model recognition & extrapolation strategy	✓	—	—
Probabilistic uncertainty	✓	✓	✓
Epistemic uncertainty due to small # of data samples	✓	—	—
Interval uncertainty	✓	—	✓
Discrete Function Non-Parametric uncertainty	✓	—	—
Random Variability of test units and conditions in repeated experiments	✓	✗	✓
Controls Type X Model User's risk from systematic uncertainty in experiment conditions	✓	✓	—

Figure A.5. Summary Table of demonstrated Capabilities and Features of the Compared Model Validation Frameworks

DISTRIBUTION

1	MS0829	431	B. M. Rutherford
1	MS1318	1441	J. R. Stewart
1	MS0378	1441	J. R. Kamm
1	MS1318	1441	P. Knupp
1	MS1318	1441	L. P. Swiler
1	MS1320	1441	V. G. Weirs
1	MS1323	1443	W. J. Rider
1	MS0316	1445	J. P. Castro
1	MS0828	1514	M. Pilch
1	MS0380	1540	D. E. Womble
1	MS0828	1544	K. F. Alvin
1	MS0828	1544	A. R. Black
1	MS0828	1544	B. Carnes
1	MS0897	1544	K. D. Copps
1	MS0828	1544	K. J. Dowding
1	MS0828	1544	R. G. Hills
1	MS0828	1544	K. Hu
1	MS0828	1544	J. R. Red-Horse
3	MS0828	1544	V. J. Romero
1	MS0828	1544	A. Urbina
1	MS0828	1544	W. R. Witkowski
1	MS1138	6923	B. S. Paskaleva
1	MS9042	8250	M. E. Gonzalez
1	MS9042	8259	J. A. Crowell
1	MS9159	8954	G. A. Gray
1	MS9159	8954	P. D. Hough
1	MS0899	RIM-Reports Management, 9532 (electronic copy)	

