SAND2014-19727C

A Perspective on the Integration of Verification and Validation into the Decision Making Process

Ken Hu, Angel Urbina and Joshua Mullins
Sandia National Laboratories
PO Box 5800, MS 0828
Albuquerque, NM 87185

Nomenclature

V&V: Verification and validation
UQ: Uncertainty quantification
M&S: Modeling and simulation

ABSTRACT

As more and more high-consequence applications such as aerospace systems leverage computational models to support decisions, the importance of assessing the credibility of these models becomes a high priority. Two elements in the credibility assessment are verification and validation. The former focuses on convergence of the solution (i.e. solution verification) and the “pedigree” of the codes used to evaluate the model. The latter assess the agreement of the model prediction to real data. The outcome of these elements should map to a statement of credibility on the predictions. As such this credibility should be integrated into the decision making process. In this paper, we present a perspective as to how to integrate these element into a decision making process. The key challenge is to span the gap between physics-based codes, quantitative capability assessments (V&V/UQ), and qualitative risk-mitigation concepts.

Keywords: Verification, validation, credibility, decision making

Introduction

Verification and validation (V&V) of physics based models has progressed quickly in the past few decades. Modeling and simulation (M&S) is now an everyday activity, and is routinely used to influence decisions. Engineering decisions are made based on both testing and M&S information. In our engineering context, the decision often relates to whether a system or components meet a requirement. In addition to the best action (accept vs. reject), sometimes the decision itself must be changed when there is not enough information (redesign/add margin, or gather additional information - i.e. renegotiate the intended uses). Decision theory tells us that the best action must be selected based on the available information.

Engineering decisions are based on proxy indicators. “Will this system work as intended?” is translated to “will the system work under a representative scenario?” What is the appropriate level of abstraction to request M&S information? In the waterfall concept, the decision maker is supposed to request a prediction or statistic about the system that is accurate to within a threshold. At some point the proxies are no longer representative of the actual question of interest. In addition, the capabilities to meet those requirements are unlikely to exist, at least with any degree of credibility. This is true for both M&S and experimental work.

The proxy indicators are the intended uses of M&S, or the M&S information that will be provided to the decision maker. It is not useful to set intended uses that exceed the predictive capability (the ability to make predictions with any confidence or credibility). Instead the intended uses must be negotiated so that they are both useful to decision makers, and achievable by the M&S/V&V analysts.

If we accept that this is a negotiation of capability vs. intended use, we still need a framework for comparing the two. We propose that decision theory is a suitable approach – the predictive capability (prediction+uncertainty and credibility) must be high enough for the customer to feel they can make a high quality decision. A high quality decision is one where you are confident that the action taken was the best available, based on the available information. Often times, decisions are not high quality because the available information is so uncertain that a proper analysis is not possible. We propose that high quality, engineering decisions are possible based on M&S information. This requires a combination of predictions, uncertainty analysis, and also credibility assessments. The argument is that the ability to make a decision is related to the available information (predictions, tests, uncertainty analysis) and the quality of that information (credibility).

This paper will dig deeper into this argument with the goal of exposing a gap between the stated purpose of V&V and the practical aspects of decision support. The authors' aim for this paper is not a position statement as to how V&V should support credibility but to start the conversation around this topic.

We start with some crucial terminology, but we stress that these are working definitions that we apply to the contents of this paper and are not assumed to be universally accepted.

- Verification – the process of confirming that a governing equation has been properly implemented into a mathematical model and solved correctly
- Validation – the process of confirming that a governing equation and the associated model are indicative of reality for a target application
- Uncertainty quantification – the process of mathematically characterizing unknown/random features and variables in a target application and predicting their influence on a quantity of interest
- Value – the benefit achieved as a result of a particular event
- Risk – the expected loss associated with a particular event or set of events
- Risk management – the identification, assessment, and prioritization of risks followed by coordinated and economical application of resources to minimize, monitor, and control the probability and/or impact of unfortunate events [1]
- Credibility – the believability of a message or conclusion, as determined by objective and subjective components including the expertise of the source and the quality and completeness of the available information

As a starting point for this paper, we conducted a cursory review of published work on this topic [2-7]. This is not an exhaustive literature search but provides a means to identify established processes in other organizations, and understand the concepts and principles and cultural issues to explain why they work. One formal framework to establish confidence in an extrapolated system level model response within the context of nuclear weapons was proposed at Sandia National Laboratories (referred to as Sandia hereafter) and it is referred to as Risk Informed Decision Analysis (RIDA). A central statement of RIDA which provides the main motivation for this research states (Pilch et al., 2006):

"Whatever mathematical form an application of Risk Informed Decision Analysis (RIDA) to a stockpile lifecycle decision might take, it requires that all uncertainties be identified and characterized. This includes the separate quantification of both variability (i.e., aleatoric uncertainty) and lack-of-knowledge uncertainty (i.e., epistemic uncertainty), as well as definitions of "other factors" and quantified characterizations of their individual contributions to uncertainty. RIDA also requires attention to uncertainties in requirements and decision criteria, such as definitions of performance thresholds that are fundamental to the decision making. In addition, RIDA requires complete transparency of all the information to make the decision process understandable, traceable, and reproducible (documented)."

A key feature in the statement above is the last sentence which implies a formal process leading to the decision making effort, in order to achieve the required level of transparency. Addressing this fact in a systematic way is what motivates this paper and thus, we organize this paper by posing questions and providing commentary – not answers. The "best" way to support and influence engineering decisions, based on M&S, V&V, and UQ evidence, is very much open to debate. The questions are posed at three levels

First level:

1. What is the relationship between V&V, credibility, and uncertainty?
2. How does V&V/UQ evidence and communication impact risk perception?
3. What is and how to characterize risk?
4. What is the "decision rule" that takes into account all the available information to make an engineering decision? What principles underlie this rule? Risk management? Value maximization?

Second level

1. In the presence of risk, how to allocate resources towards testing, M&S, V&V, UQ, etc.?

2. What is the best process to determine whether a model is valid?
3. What is the role of the Subject Matter Expert and peer review?

Third level:

1. What validation metric should I use?
2. How many Monte Carlo samples do I need?
3. How good is good enough? Is my model valid?

The questions at the third level are representative of much of the published work in the V&V and UQ fields. Indeed the focus of V&V and UQ research at Sandia has been on developing methods, metrics, heuristics, etc. The second level is more general, asking questions about the culture and processes of an engineering organization. The first level is the most abstract – dealing with the concepts and principles that govern how decisions are made. In this work, we focus on first level – the most abstract. The hope is that providing clarity on the principles will highlight the cultural and procedural changes that are required to connect M&S, V&V, and UQ work and decision making. Then the detailed, practical questions of methods and metrics will become more meaningful.

How is M&S, V&V/UQ and Credibility Connected to Decision Making and Risk?

The answer to this question has everything to do with the engineering culture. A fundamental question is: what is the effect of a certain activity – either testing, M&S, V&V, UQ in the total uncertainty vs. credibility space. For example, doing UQ improves the understanding of uncertainty in predictions but the particular method used to quantify the uncertainty can introduce epistemic uncertainty; in addition, the boundary conditions might not be well characterized (garbage in, garbage out). So in this case, perhaps the fidelity of the uncertainty increases, but the credibility could potentially go down. Credibility is a subjective thing, and institutional knowledge and experience is the key.

To begin addressing the topic of this section, we have arbitrarily divided the problem into two broad categories: 1) where V&V/UQ is not part of the process and 2) where V&V/UQ is integrated into the workflow. These categories are illustrated below and some commentary on each one is made. To reiterate, the purpose of this paper is not to critique any particular approach or to suggest one is superior to another but merely to contrast between them and highlight the positive aspects and shortcomings of each one.

Case 1: A deterministic workflow

The first case, shown in Figure 1, represents a workflow that is pervasive in many industries. It involves a deterministic M&S single prediction with no formal understanding and treatment of uncertainty. This leads to a decision which can be characterized as uninformed (i.e. sources of uncertainty not identified and/or incorporated into the analysis) and thus yielding little understanding of value and risk. In this case, the decision maker relies on expert opinion and "tribal knowledge" to prescribe a level of risk associated with the M&S predictions that are presented. In such a scenario, the concept of "I trust this predictions because analyst XYZ made them and has years of experience in this field" are commonplace and serve as the basis for assigning credibility to an M&S body of work. One issue with the previous statements is that it is a subjective and possibly bias way of assigning credibility of M&S prediction. We are not discounting the value of experience in the process of decision making but it is possible that a less experienced analyst could provide similar or even technically superior predictions yet his/her results will be highly questioned and at worse, dismissed mainly based on perception. We feel this is one of the biggest risks in not having a systematic approach to determining credibility in an M&S prediction. On the positive side, this approach has been in place for many years and it is used by industry to inform and guide decisions on a regular basis. Both analysts and decision makers are comfortable with this mainly due to familiarity with this.

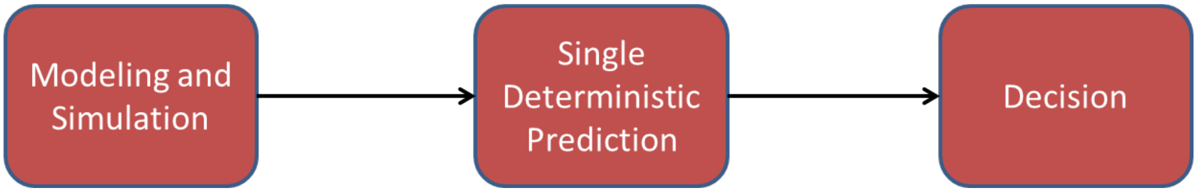


Figure 1: Deterministic, black box M&S single prediction with little to no understanding of uncertainty.

Case 2: An M&S and V&V/UQ integrated workflow

In contrast to Case 1, the second case, shown schematically in Figure 2, represents an attempt to integrate the elements of V&V/UQ into the decision making process. It is noted from the figure that the process is not linear or serial but a recursive process where feedback loops help to enhance various aspect of the process. Modeling and simulation, where elements of V&V and UQ are integrated into the process, will tend to yield predictions plus an estimate of uncertainty and an assessment of credibility and risks given the activities that were and were not performed. This, we posit, leads to enhanced informed decisions with a formal assessment of the risks associated with predictions used in the process. This knowledge leads naturally to risk management/mitigation opportunities. Figure 2 includes the pieces that we believe are necessary, but not sufficient, for supporting decisions. Decisions must be made with some purpose or context beyond model predictions. In the engineering work performed at Sandia, the purpose is to manage technical risks. Those risks must drive the development, verification, validation, and eventual use of models. The loop is completed when modeling results are communicated to decision makers. We believe that distinct concepts of uncertainty and credibility are useful in this setting. We also envision that this identification of uncertainties and risks will naturally lead to a formulation in which the optimal allocation of resources, to reduce risk and improve decisions, can be achieved. This is further explored later in this paper.

We note that a decision can be made in both cases presented above. However, the first case must rely entirely on "intangibles" like assumptions, experience, credibility of the person(s) informing the process, etc., while the second provides the necessary evidence to make a defensible, high-quality, informed decision. Since risk assessment is an intrinsic element in this workflow, it is possible to introduce risk management and mitigation processes into the workflow. This is a fundamentally different way of thinking; one that emphasizes a system's engineering approach to the problem. In other words by looking at the "big picture", it can enable to make informed decisions at the analysis level.

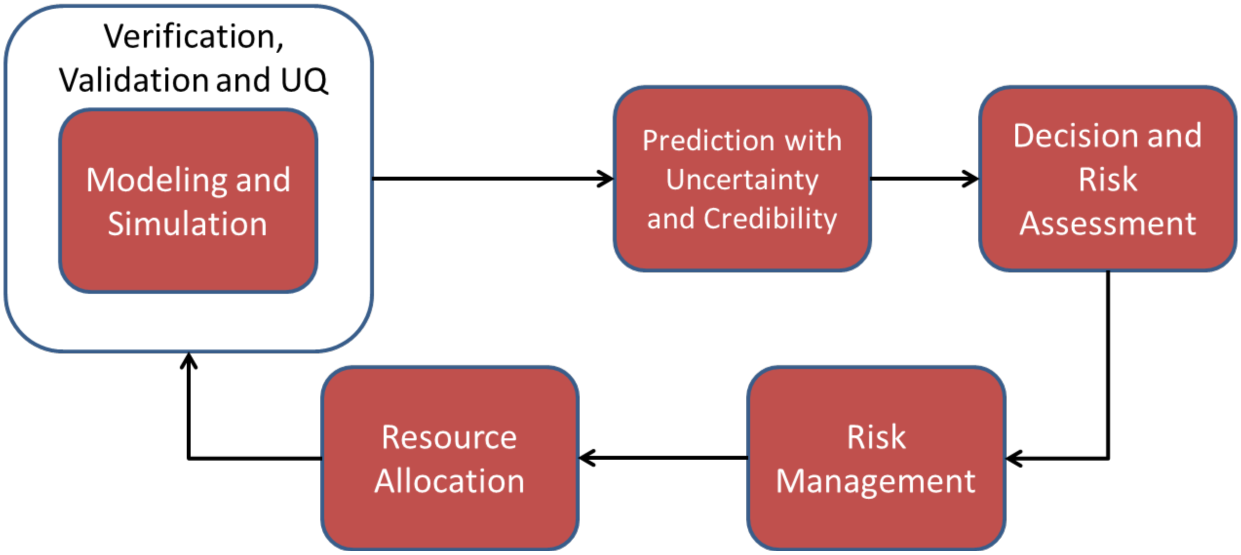

Verification, Validation and UQ
Modeling and Simulation
Prediction with Uncertainty and Credibility
Decision and Risk Assessment
Risk Management
Resource Allocation

Figure 2: M&S with elements of V&V and UQ incorporated into the process.

To enable the upper left box in Figure 2 (i.e. the integration of M&S and V&V/UQ, we have developed an integrated M&S/V&V/UQ workflow and it is shown in Figure 3. The value proposition of this integrated workflow is the identification of an end-to-end path which includes elements of V&V/UQ leading to a customer driven deliverable.

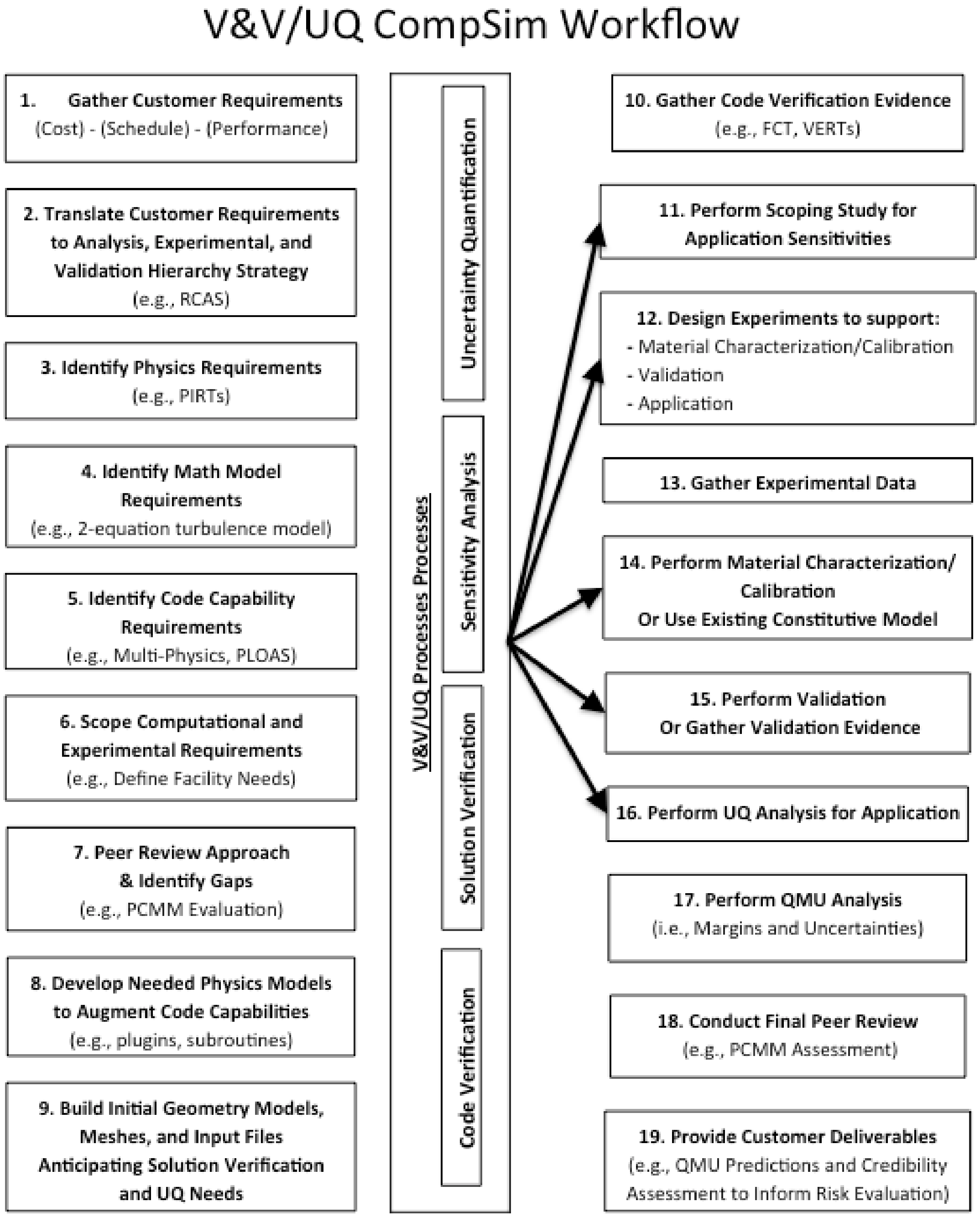
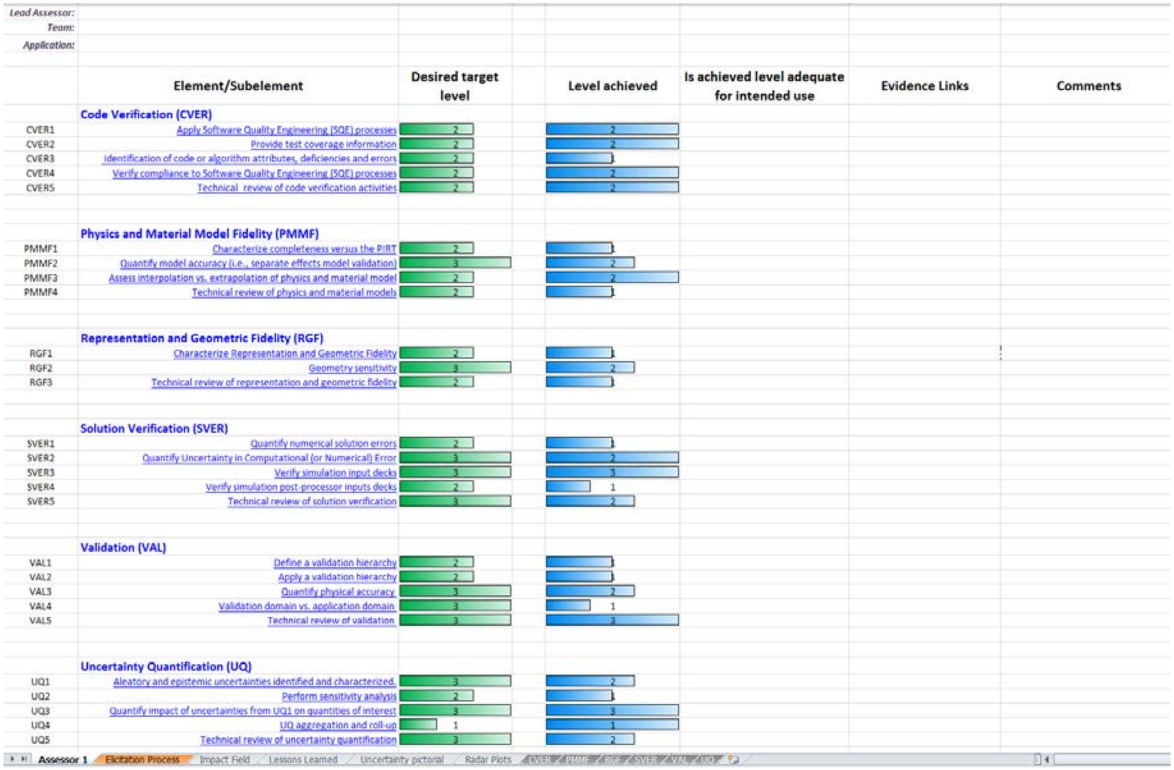

V&V/UQ CompSim Workflow
1. Gather Customer Requirements
(Cost) - (Schedule) - (Performance)
2. Translate Customer Requirements to Analysis, Experimental, and Validation Hierarchy Strategy
(e.g., RCAS)
3. Identify Physics Requirements
(e.g., PIRTs)
4. Identify Math Model Requirements
(e.g., 2-equation turbulence model)
5. Identify Code Capability Requirements
(e.g., Multi-Physics, PLOAS)
6. Scope Computational and Experimental Requirements
(e.g., Define Facility Needs)
7. Peer Review Approach & Identify Gaps
(e.g., PCMM Evaluation)
8. Develop Needed Physics Models to Augment Code Capabilities
(e.g., plugins, subroutines)
9. Build Initial Geometry Models, Meshes, and Input Files Anticipating Solution Verification and UQ Needs
V&V/UQ Processes Processes
Uncertainty Quantification
Sensitivity Analysis
Solution Verification
Code Verification
10. Gather Code Verification Evidence
(e.g., FCT, VERTs)
11. Perform Scoping Study for Application Sensitivities
12. Design Experiments to support:
- Material Characterization/Calibration
- Validation
- Application
13. Gather Experimental Data
14. Perform Material Characterization/ Calibration
Or Use Existing Constitutive Model
15. Perform Validation
Or Gather Validation Evidence
16. Perform UQ Analysis for Application
17. Perform QMU Analysis
(i.e., Margins and Uncertainties)
18. Conduct Final Peer Review
(e.g., PCMM Assessment)
19. Provide Customer Deliverables
(e.g., QMU Predictions and Credibility Assessment to Inform Risk Evaluation)

Figure 3: Integrated M&S/V&V/UQ Workflow

In the context of Figure 2 which includes elements of credibility and risk assessment, the integrated workflow is envisioned to aid in the following ways. It clearly outlines the steps that are needed to comprehensibly perform an end-to-end M&S activity with the proper context in mind. It provides an opportunity to adjust the scope of the effort by "pruning" those elements (i.e. the boxes in the workflow) that are unattainable due to funding, schedule and/or technical impediments. It is this "pruning" of elements that introduce risk in to the process. For example, if the Code Verification element, which addresses how good is the software that is used to run an analysis, is "pruned", the net effect is felt downstream of this element. Thus there is a significant risk involved (and a corresponding reduction in credibility) due to this step being eliminated from the workflow. At this point, there is no formal quantification of the amount of risk that is introduced into the process by "pruning" any particular box or boxes in the workflow. This is a qualitative assessment. Future research will be needed to determine if there is any practical way to quantify this. Also the amount of risk is highly dependent on the context (or intended use) of the M&S activity. For example, if the analysis is mainly to understand a relative behavior of a quantity of interest, then the risk is fairly low. Conversely, if the analysis is supporting a major decision say, certification of an airplane, then the risk associated with the “pruning” of the Code Verification element is fairly high. It is thus imperative that both the context and the workflow that leads from requirements to deliverables are well defined.

Communication of credibility as a function of M&S, V&V and UQ

The credibility of a computational simulation analysis has historically been based largely on the experience of the analyst using a tool and judgment by that analyst on the suitability of the results produced for a particular application. At Sandia, we developed the Predictive Capability Maturity Matrix (PCMM) [8] to provide more structure and formality in assessing the credibility of a M&S analysis for a target application, to reduce the ambiguity in such assessments, and to provide specificity as to what should be assessed and communicated to the analyst’s customer. PCMM evaluations have the potential to provide information for effective planning as well as for communication. The PCMM is based on six evaluation dimensions, or elements, that are deemed fundamentally important to the quality of as M&S analysis. These elements address 1) the fidelity in representing physics, including the material models; the 2) the geometric fidelity in representing the system or subsystem element being modeled; 3) the completeness in addressing whether the computational simulation code has been verified from a software assurance point of view and 4) verified from a solution convergence point of view; 5) assessments against experimental data; and 6) the evaluation of uncertainty in the results due to uncertainties such as the input information used to characterize the specific geometry, environment, and material properties for the application. The current tool used at Sandia to perform the PCMM assessment is shown in Figure 4.



Lead Assessor:
Team:
Application:

	Element/Subelement	Desired target level	Level achieved	Is achieved level adequate for intended use	Evidence Links	Comments
	Code Verification (CVER)					
CVER1	Apply Software Quality Engineering (SQE) processes	2	2			
CVER2	Provide test coverage information	2	2			
CVER3	Identification of code or algorithm attributes, deficiencies and errors	2	1			
CVER4	Verify compliance to Software Quality Engineering (SQE) processes	2	2			
CVER5	Technical review of code verification activities	2	2			
	Physics and Material Model Fidelity (PMMF)					
PMMF1	Characterize completeness versus the PIRT	2	1			
PMMF2	Quantify model accuracy (i.e., separate effects model validation)	3	2			
PMMF3	Assess interpolation vs. extrapolation of physics and material model	2	2			
PMMF4	Technical review of physics and material models	2	1			
	Representation and Geometric Fidelity (RGF)					
RGF1	Characterize Representation and Geometric Fidelity	2	1			
RGF2	Geometry sensitivity	3	2			
RGF3	Technical review of representation and geometric fidelity	2	1			
	Solution Verification (SVER)					
SVER1	Quantify numerical solution errors	2	1			
SVER2	Quantify Uncertainty in Computational (or Numerical) Error	3	2			
SVER3	Verify simulation input decks	3	3			
SVER4	Verify simulation post-processor inputs decks	2	1			
SVER5	Technical review of solution verification	3	2			
	Validation (VAL)					
VAL1	Define a validation hierarchy	2	1			
VAL2	Apply a validation hierarchy	2	1			
VAL3	Quantify physical accuracy	3	2			
VAL4	Validation domain vs. application domain	3	1			
VAL5	Technical review of validation	3	3			
	Uncertainty Quantification (UQ)					
UQ1	Aleatory and epistemic uncertainties identified and characterized.	3	2			
UQ2	Perform sensitivity analysis	2	1			
UQ3	Quantify impact of uncertainties from UQ1 on quantities of interest	3	3			
UQ4	UQ aggregation and roll-up	1	1			
UQ5	Technical review of uncertainty quantification	3	2			

Figure 4: PCMM Tool’s Main Assessment Sheet

Resource allocation

When the result of the risk assessment is unacceptable, a mitigation strategy must be undertaken. Here we refer to this process as risk management. When the risk assessment is made deterministically, risk management decisions must be made based largely on expert judgment. While domain experts frequently understand which aspects of their models are inadequate (w.r.t. physics fidelity or numerical resolution), they are not likely to be able to gauge the relative contributions of these aspects in the overall system prediction. By including UQ and credibility assessment in the decision process, the improvement decisions can instead be made in a quantitative way that takes a comprehensive view of the system and the contributions of error and uncertainty sources to the various aspects of the system prediction. Since this risk management process is subject to budget constraints, an important issue is the proper allocation of resources to achieve the maximum benefit.

We explore a quantitative risk reduction-based strategy for resource allocation [9, 10]. Classically, the risk of an event (e.g. system failure) has two key components: (1) the probability of the event and (2) the consequences of the event. These two components have a simple, logical relationship, in which the risk S is a product of the consequence of an event L and the probability of the event $P(L)$.

$$\boldsymbol{S = L * P(L)}$$

Within this context, risk can be viewed as the expected value of the cost of a particular failure scenario. For some systems, a relatively large failure probability does not pose a great risk because the failure event will not result in any particularly severe consequences. Therefore, the events of greatest concern are those that have both high probability and extreme consequence (e.g. human life loss and/or major property destruction). In many applications, there are many different potential risk events, and the overall system risk S_T is the summation of all of m discrete risk scenarios.

$$\boldsymbol{S_T = \sum_{i=1}^{m} L_i P(L_i)}$$

Designers and decision-makers have little control over the consequences of an event, so risk minimization is achieved by reducing the probability of the negative events. This reduction is significantly enabled by minimizing prediction uncertainty while maintaining prediction accuracy (i.e. low bias).

From an economic perspective, it is not logical to spend large resources on UQ without also considering the total benefit of the analysis. Since risk can be directly interpreted as a cost, it provides a convenient space to analyze design and management decisions. The total cost of the UQ activities can then be viewed as a “risk” event with 100% probability since the cost is always incurred once the spending decision is made. Then, a reasonable overall goal for the decision maker is to minimize the total risk coming from these two components (the failure risk S_f and the UQ/model development cost S_d). This view leads to the following formulation for determining how much UQ spending is enough:

$$\boldsymbol{S_d = 1 * L_d}$$
$$\boldsymbol{S_f = p_f * L_f}$$
$$\boldsymbol{S_T = S_d + S_f}$$

Here, L_d is the total cost of UQ/development activities, p_f is the system failure probability, and L_f is the consequence of system failure. It is assumed that L_f is a constant that the decision maker cannot control, and L_d is the primary decision variable during the resource allocation phase of the process. With the overall goal of minimizing S_T the spending budget L_d must now be selected based on its impact on p_f, which is a dependent variable in the analysis that is affected by UQ and credibility assessment. The primary challenge is that the functional relationship between spending activities and failure probability reduction is never explicitly known. However, this framework emphasizes the importance of looking at these decisions at the level of the system prediction rather than making refinement decisions locally without specific regard for their overall impact.

Potential spending decisions that could be made include model selection, model improvement, and test selection. Since these are topics unto themselves in the literature and the intent of this paper is to provide a high-level framework for decision-making, we only mention them briefly here. Uncertainty propagation can be prohibitively expensive due to the large number of model evaluations that are needed. As a result, models of lower fidelity/complexity including reduced-order, reduced-physics, and mathematical surrogate models are often used for UQ activities. There is a tradeoff decision of accuracy vs. efficiency when selecting among these candidate models. Any of them may potentially be improved/refined, or they may be selected adaptively for different predictions. Additionally, these candidate models must be calibrated and validated, which couples the modeling decisions with test selection decisions. The number and type of calibration and validation tests that are performed impacts the quality of the models and therefore the quality of the prediction and associated UQ. For each of these decisions, the important consideration is the impact on the overall prediction uncertainty since it is directly tied to the system failure probability component of the total risk. This context for the decisions is critical to efficient resource allocation.

Summary

In this paper, we present a perspective of how elements of V&V/UQ and Credibility could be incorporated into the decision process. We have developed an integrated workflow that is useful in understanding sources of risks in the end-to-end analysis. To properly establish the importance of these sources of risk, it is imperative that both the context and the workflow that leads from requirements to deliverables are well defined. At present, risks arising from "pruning" elements in the workflow are mainly qualitatively defined. A quantitative solution is a future research direction which once established, will enable a more formal and mathematical treatment of the larger issue of resource allocation.

Acknowledgments

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

References

1. Hubbard, D. W., The Failure of Risk Management: Why It's Broken and How to Fix It, Wiley, 1st Ed., New York, 2009.

2. Elele, J.N. Assessing risk levels of verification, validation, and accreditation of models and simulations. 2009.

3. Elele, J.N. and J. Smith. Risk-based verification, validation, and accreditation process. 2010.

4. Elele, J.N. and N. Gould. Methodology for designing M&S that integrates V&V processes and documentation. 2012.

5. Youngblood, S.M., et al., Risk Based Methodology for Verification, Validation, and Accreditation (VV&A) M&S Use Risk Methodology (MURM), 2011, Johns Hopkins University Applied Physics Laboratory.

6. Nitta, C.K. and R.W. Logan, Qualitative and Quantitative Linkages from V&V to Adequacy, Certification, Risk, and Benefit / Cost Ratio, 2004, Lawrence Livermore National Laboratory.

7. Blattnig, S.R., et al., Towards a Credibility Assessment of Models and Simulations. American Institute of Aeronautics and Astronautics, 2009002E

8. Oberkampf, W.L., et. al., Predictive Capability Maturity Model for Computational Modeling and Simulation, Sandia National Laboratories , SAND2007-5948, 2007

9. Mullins, J. and Mahadevan, S., "Variable-fidelity model selection for stochastic simulation," Reliability Engineering & System Safety, Vol. 131, pp. 40-52, 2014. doi: 10.1016/j.ress.2014.06.011

10. Mullins, J., Li, C., Mahadevan, S., and Urbina, A., "Optimal Selection of Calibration and Validation Test Samples Under Uncertainty," Model Validation and Uncertainty Quantification, Vol. 3, Conference Proceedings of the Society for Experimental Mechanics, pp. 391-401, 2014.