16

Development and responsibilities for verification, validation and uncertainty quantification

This chapter discusses some perspectives on the responsibilities for the development, conduct, delivery, and management of verification, validation, and uncertainty quantification (VV&UQ). The topics considered here deal with both technical and nontechnical issues, but all are focused on the strategic goals of improved credibility and proper use of simulations. Our experience, and the experience of others, has convinced us that while technical issues and computing resources are important, they are not the limiting factor in improving the credibility and usefulness of scientific computing used in a decision-making environment. We believe that nontechnical issues have significantly constrained the improvements in credibility that VV&UQ can provide.

Examples of the issues discussed in this chapter are (a) suggestions for needed technical developments in VV&UQ, (b) responsibilities for the various technical activities that encompass VV&UQ, (c) recommendations for management responsibilities and leadership in deploying VV&UQ practices, (d) development of V&V databases for general use, and (e) development of industrial and engineering standards for V&V.

There are diverse perspectives on many of these topics from different groups, such as companies that produce hardware products or commercial software, government organizations, organizations that sell simulation services, and special interest groups. We do not claim to have all of the answers, or even the correct answers, to some of the questions raised. We are simply attempting to provide perspectives on some important technical and nontechnical themes of responsibility and development of VV&UQ. As scientific computing continues to have an increasing impact on everyday life, economic competitiveness, environmental safety, global warming analyses, and national security, we believe these topics must involve a wider discussion and debate.

16.1 Needed technical developments

In most chapters we discuss various technical weaknesses and challenges related to the methods and techniques discussed in the chapter. Here we present a summary of some of the key research topics that we feel are needed to advance VV&UQ in scientific computing. This list is not meant to be comprehensive, but only suggestive. These topics are presented in the order that the topics appeared in the book:

- automated (in-situ) code verification, where a manufactured solution/source term is automatically generated and an order of accuracy test is automatically run based on chosen code options one desires to verify;
- construction of manufactured solutions for much wider range of technical fields in scientific computing;
- improved methods to measure coverage of code verification testing based on features and capabilities options in the code;
- reliable, i.e., demonstrably asymptotic, discretization error estimation for complex scientific computing problems, including hyperbolic partial differential equations and problems with discontinuities;
- reliable mesh adaptation for complex scientific computing problems, especially for anisotropic problems such a boundary layers and shock waves;
- further development and use of the validation hierarchy as it is focused on the system of interest;
- further development and use of the phenomena identification and ranking table (PIRT) and the gap analysis table;
- development of validation metrics for various types of system response quantities, particularly time dependent responses with a wide range of frequencies, and SRQs that depend on a large number of input parameters;
- development of methods for incorporating validation metric results into uncertainty estimation of predicted responses, e.g., extrapolation of model uncertainties as a function of input parameters and inclusion of epistemic uncertainty in predictions;
- improved understanding of the positive and negative effects of model calibration on model predictive capability.
- uncertainty estimation of SRQs at higher levels in the validation hierarchy, based on validation metric results where available, and multiple mathematical models at all levels in the validation hierarchy.

16.2 Staff responsibilities

By *staff* we mean individuals who are technically trained to conduct any part of the computational analysis related to VV&UQ. This includes engineers, scientists, mathematicians, statisticians, and computer scientists involved in the theoretical, computational, or experimental activities of the overall effort. Management responsibilities will be discussed in Section 16.3.

16.2.1 Software quality assurance and code verification

16.2.1.1 Who should conduct SQA and code verification?

Software quality assurance (SQA) and code verification are primarily the responsibility of code developers and commercial software suppliers. Most commercial software companies are subjected to demanding quality and reliability requirements from their customers. Yet, it is impossible to obtain public information concerning the percentage of resources devoted to SQA and code verification compared to total resources devoted to developing and selling

commercial software. Factors that should be important for these companies as part of their business activities include the following.

1 The software company should put in place and practice a high level of SQA for every stage of the software life cycle. These stages range from a new research module, to preliminary release of modules, to bug-fixing patches of released software products. Each of these modules will require separate SQA tracking and different levels of code verification testing for its quality assurance.
2 The software company should educate its customers to understand that it is impossible to cover all possible combinations of physical, boundary, material, and geometrical conditions available in the code during the verification testing process. The software company should (a) carefully prioritize what combinations are important for its range of customers, and (b) inform the interested customers which combinations have been tested and which ones have not. If the customer is not satisfied with extent of the testing for their applications, they should express those views to the software company.
3 The software product should be compiled and ported to a wide variety of computer platforms (serial, multi-core, and parallel processor machines) that may be used by their customers. They should also be rigorously tested under different operating systems and a range of versions of operating systems.

The SQA and code verification procedures within a software company's development environment constitute an ongoing process. With every software release, the company should have SQA test procedures that are set up under the supervision of a software quality group. These procedures should include automatic reporting features for the results of the testing. In particular, the activities related to fixing, tracking, and managing software bugs should be a critical part of the global SQA procedures of the software company.

Code development within organizations, separate from software suppliers, usually falls into two groups: (a) groups within a large organization or a large project that develop either proprietary software or special purpose software, and (b) research groups within corporations, universities, or institutes. Within many research groups, there is considerable resistance to the implementation of formal SQA practices. The degree of formality, software development constraints, user restricted access, and costs of SQA are key factors in their resistance. For many research groups, there is also a well-established value system that new knowledge, as well as publishable results, is the crucial product, not reliable software, algorithms, reproducibility, and SQA and code verification documentation.

The notable exception to the generally low level of SQA and code verification procedures is a computational project involved in high-consequence systems. The two most important examples are nuclear power reactor safety and underground storage of nuclear wastes. These projects have intense government and public scrutiny not only in results of the analyses, but also in all of the steps leading up to the results. In the US, primary oversight of nuclear power reactor analyses resides with the Nuclear Regulatory Commission. Primary oversight of underground storage of nuclear wastes resides with the Department of Energy, the Nuclear Regulatory Commission, and the Environmental Protection Agency. These projects have extraordinarily rigorous SQA procedures, including detailed documentation of all of the software components. In fact, there is credence to the argument that the procedures

are so demanding that they severely constrain the introduction of improved mathematical models and numerical algorithms. A proper balance must be struck between appropriate SQA procedures and improved numerical techniques.

16.2.1.2 Who should require SQA and code verification?

Code users are primarily responsible for requiring SQA and code verification from code developers and commercial software companies. Code users may also contribute to conducting SQA and code verification, but when doing so they are serving more in the role of code developers. When we refer to *code*, we are referring to the entire software package or any subset of the software used in a computational analysis. We believe the appropriate conceptual framework for understanding responsibilities in all of the VV&UQ activities discussed in Section 16.2 is that of supplier and customer. *Suppliers* are individuals or organizations that provide products or information used in the analysis. *Customers* are individuals and organizations that receive a product (such as a code) or information produced by the code or the analysis. An individual or an organization can serve as both a supplier and a customer during different activities or phases of the analysis.

The code developers and commercial software companies are clearly the suppliers of the code as a product. The final responsibility for the quality of the code used in the analysis, however, lies with the customer. The customer is the user of the product or the information produced by the product and he/she must provide the requirements, due diligence, and proper use of the product or information. As a result, we believe the value placed on SQA and code verification by the customer is the *driving factor* in quality. If the customer simply *assumes* that SQA and code verification are adequately done and does not require any documentation, details, or evidence of these activities, then the supplier appropriately considers the requirement void of content or bogus. For *any* item of importance to a customer, the customer should require demonstration of capability, proof of performance, inspection, or documentation of quality and reliability.

If the customer is dealing with a supplier that essentially has a monopoly on the software market, then this situation is very different and more complex. If the market is relatively small and the customer is important to the supplier, then the customer has some leverage with respect to specifying requirements. If on the other hand, the market is large, individual customers are at a distinct disadvantage and have few options to require improvements. They can either tolerate the product quality provided by the supplier, or they can pursue more strategic solutions. For example, the customer may seek to promote new competitors in the market, or if the situation warrants it, the customer may seek governmental or legal actions to break the monopoly. These solutions, or course, are very slow and costly.

Commercial software companies have long recognized the importance of documentation for their product's use and capabilities. Their documentation has mainly described the use of the graphical user interface (GUI), the capabilities of the code, mesh generation capabilities, numerical algorithm techniques, post-processing capabilities, and output graphics. Documentation of code verification procedures and results by commercial software

companies, however, has generally been very limited, or nonexistent. Customers presume that the software company has satisfactorily conducted SQA and code verification, but there is a great deal of evidence to the contrary. We believe the old adage of quality processes applies here: if it isn't documented, it didn't happen.

Some large commercial code companies specializing in solid mechanics have developed an extensive set of verification benchmarks that can be exercised by licensed users of their codes (see, for example, (ANSYS, 2005; ABAQUS, 2006)). These documented benchmarks are intended to demonstrate the performance of the code on relatively simple benchmark problems, as well as serve as a training aid for users of the code. The primary weakness of essentially all of the documented verification testing is that it only demonstrates "engineering accuracy" of the code; but does not carefully quantify the order of convergence of the numerical methods (Oberkampf and Trucano, 2008). As stated in one set of documentation, "In some cases, an exact comparison with a finite-element solution would require an infinite number of elements and/or an infinite number of iterations separated by an infinitely small step size. Such a comparison is neither practical nor desirable" (ANSYS, 2005). This argument is without merit. As discussed by many authors, and at length in this book, the discretization order of convergence and iterative convergence characteristics of the numerical methods can be practically and rigorously tested. It is of great value to a customer to demonstrate what the code actually produces in terms of convergence rate characteristics on a variety of relevant test problems, versus what is claimed in glossy marketing material.

Computational fluid dynamics (CFD) software has, in general, not provided the same level of documentation of code verification as that provided by commercial solid mechanics codes. As an indication of the poor state of maturity of CFD software, a recent paper by Abanto *et al.* (2005) tested three unnamed commercial CFD codes on relatively simple verification test problems. The poor results of the codes were shocking to some people, but not to the authors of the paper, nor to us.

16.2.2 Solution verification

16.2.2.1 Who should conduct solution verification?

Solution verification is primarily the responsibility of the analyst conducting the simulation. If it is a simulation involving multiple analysts and software packages, there should be close communication between the analysts to ensure adequate monitoring of numerical solution error. As discussed in Chapter 7, Solution verification, and Chapter 8, Discretization error, the recommended procedure is to quantitatively estimate both the iterative and mesh discretization error for each SRQ of interest. If a sequence of codes is used where the output of one code provides the input to the next in order to produce a final set of SRQs, then the estimation procedure for solution error becomes convolved with the uncertainty estimation procedure.

The level of effort needed to ensure high quality solution verification can vary tremendously, depending on the complexity of the analysis. For large-scale analyses, the effort

needed to carefully estimate and document numerical solution error is major. For example, with a hundred uncertain input quantities resulting in tens of uncertain SRQs, the solution error would need to be estimated for each SRQ over the 100-dimensional input space. For those SRQs where the estimated error is not negligible, then the error would need to be combined with other uncertainties estimated in the prediction. A new method for combining these uncertainties was presented in Chapter 13, Predictive capability. This is a time-consuming task in a large-scale analysis, especially if a sequence of codes is used where the output of one becomes the uncertain input for the next code. Our experience, and observations from the journal literature, indicates there is careful attention paid to solution verification only for high-consequence, high public visibility projects.

For a deterministic analysis with relatively few SRQs, the task of numerical error estimation is relatively straightforward. For those SRQs where the estimated error is not negligible, the SRQ should be represented as an interval in order to reflect the uncertainty in the SRQ. Even for this type of problem, it is still common to see little or no information concerning solution verification in conference papers or journal articles. For analyses documented in internal corporate or project reports, it is our experience that solution verification is addressed infrequently.

Probably the simplest technique for quantifying discretization error is to use the grid convergence index (GCI) method (Roache, 1998). To use the method, the analyst (a) must have a solution on two different mesh and/or temporal resolutions; (b) should have some type of evidence that the solution for the particular SRQ of interest is in the asymptotic region for each of the independent variables in the mathematical model; (c) should compute the observed order of convergence, or alternatively assume the formal order of convergence, of the discretization method; and (d) should pick an appropriate factor of safety for the error estimate.

Sometimes even the GCI is not used and the analyst resorts to qualitative methods. For example, the solution is computed on two different mesh resolutions and the difference between the solution values are examined for each of the SRQs of interest. If the differences are all considered "small," the mesh resolution is deemed acceptable and the analysis proceeds. This latter approach is unreliable and can be very misleading with respect to the inferred accuracy of the results of the analysis.

16.2.2.2 Who should require solution verification?

Users of the analysis results from the simulation are primarily responsible for requiring quantitative information concerning solution verification. We again apply the supplier–customer model discussed above. Here, the analyst is the supplier of information to the customer who uses the analysis results. The customer can be anyone who directly uses the results of the analyses; e.g., a designer, a project manager, or decision maker. Stakeholders, i.e., those who have an appropriate and legitimate interest in the analysis results, may also share some of the responsibility of solution verification, but this would normally be minimal. Stakeholders are typically secondary users of the information provided by the

analysis, such as those who benefit or profit from the analysis, or those who must plan their future activities based on the analysis.

Some have argued that the complete responsibility for solution verification should rest with the analyst conducting the work, instead of a shared responsibility with the customer. They argue that "The analysts are the experts in error estimation, so they should be responsible." Or, "We are not trained in these matters, why should we be responsible?" This argument has little merit and we reject it. We argue that it would be delinquent, if not negligent, to *assume* that solution verification was adequately done in an analysis. The argument for shared responsibility for solution verification between the analyst and the customer is analogous to the previous discussion of shared responsibility for SQA and code verification between code developers and code users.

Commonly the situation arises where the customer may be inexperienced or completely ignorant of solution verification techniques. For example, the customer could be a project leader or decision maker who is experienced in other technical fields, such as systems engineering and testing, but *not* in aspects of scientific computing. A more problematic situation would be if the project leader or decision maker is not technically trained, for example, their background is in business, marketing, or law. For these situations, it would be useful to have advisory staff or consultants experienced in scientific computing.

16.2.3 Validation

16.2.3.1 Who should conduct validation?

Validation is the joint responsibility of the analysts conducting the simulation and the experimentalists conducting the experiment. As discussed in Chapter 11, Design and execution of validation experiments, validation experiments should be designed, executed, analyzed, and documented as a team effort. Various members of the team must be familiar with and decide on a wide range of technical details for the validation activity, such as (a) what predictive capabilities of the model are of primary importance to the application of interest? (b) can the available experimental facilities provide adequate conditions to assess the model's predictive capabilities? (c) what SRQs are of interest and can they be measured with adequate experimental accuracy? (d) what are the key input quantities to the model and how can these quantities, along with their uncertainties, be measured in the experiment? (e) can the numerical solution errors in the SRQs of interest be estimated and adequately reduced? and (f) does an appropriate validation metric exist that is relevant to the application of interest, or does one need to be developed? Assembly of a competent validation team is challenging, and joint funding and coordination can also be formidable.

If poor agreement between simulations and experiments occur, some may feel that the validation effort has been wasted. We believe this response reflects a misunderstanding of the primary goal of validation. Validation is conducted for the purpose of critically assessing the predictive capability of a mathematical model; *not* for the purpose of showing good agreement of model and experiment. Consider two very beneficial outcomes that can occur

when agreement between simulation and experiment is found to be unexpectedly poor. The usual situation for these outcomes is when it is strongly believed that the model should be very accurate for the range of input conditions tested. First, it can be found that the cause of the disagreement with experiment is due to a previously undetected programming error in the software or an unexpected numerical algorithm deficiency. Although there is a definite penalty to be paid for this lack of sufficient code verification, it is a very valuable piece of information not only for the present analysis, but also for past and future analyses. Two examples of the penalty that may result are (a) past simulations may be affected by this code bug or algorithm deficiency and these will need to be revisited, and (b) all model parameters that have been calibrated and are affected by this code bug will need to be recalibrated. These kinds of repercussion can be very expensive, humiliating, and damaging to the confidence in the organization responsible for the simulations.

Second, it can be found that the cause of the disagreement is due to an unexpected systematic (bias) error in the experimental measurements. As just discussed, this can have negative repercussions to the experimentalist who conducted the experiment or the facility in which the experiment was conducted. On the positive side, it can also be beneficial in convincing staff and management to improve measurement techniques and experimental facilities in the future.

Code developers or commercial software providers are only secondarily responsible for validation. They commonly conduct various unit level validation activities, but these are usually generic in nature. For example, commercial software providers typically compare their simulations with classic experiments using publicly available experimental data. These cases are usually documented as part of marketing material for the software or user training cases. One procedure that can be very helpful to a potential new customer for a commercial software license is to request that the software company compute a blind prediction as a validation test case. Most potential customers have proprietary or restricted distribution experimental data on systems, subsystems, or components that they design, market, or use. All of the needed input data from the experiment can be given to the software supplier and then the software company computes the SRQs of interest to the software customer. The comparisons of the simulations with the data may, or may not, be shared with the software supplier, depending on the business or security sensitivity of the situation. In this type of blind prediction, the software customer must be extremely careful to provide to the software supplier all of the needed information that is appropriate for their modeling approach. If some input quantities are not available, then the software supplier will need to conduct a nondeterministic analysis for comparison with the experimentally measured SRQs.

16.2.3.2 Who should require validation?

The analysis results user, i.e., the customers for the information produced, is primarily responsible for requiring detailed information concerning validation activities. Customers of computational analyses have traditionally recognized this responsibility with regard to

validation activities. We believe this tradition exists because most customers of computational analyses have intuitively expected critical comparisons of predictions and experimental measurements that are closely related to their system of interest.

Given the validation framework discussed in this book, there are four aspects of validation particularly important to a customer. (a) At what levels of the validation hierarchy were the experiments conducted, compared to the top-level system of interest in the hierarchy? (b) What is the relationship of the validation domain where the experiments were conducted, compared to the application domain of the system? (c) Were the validation metrics used in model accuracy assessment relevant to the application of interest? and (d) How accurate were the model predictions over the validation domain? Here, we comment on the first two areas of importance.

Occasionally, it has been observed that the validation conditions are inappropriately influenced by the needs of the model developers as opposed to the needs of the customer. For example, model developers (and sometimes analysts) can invent reasons why certain validation conditions that would test the model more critically are not possible, practical, affordable, or useful. This type of thinking reflects a cognitive bias and is usually referred to as *confirmation bias* (Pohl, 2004). That is, most people have a strong tendency to process and search for information in a way that confirms their preconceptions or existing viewpoints, and avoid information that might contradict, challenge, or threaten their present beliefs. This bias is the primary reason for lack of enthusiasm for blind-test experiments among many model developers. This bias, of course, is not restricted to model builders, but is even more dangerous when exhibited by the customers of computational analyses.

16.2.4 Nondeterministic predictions

16.2.4.1 Who should conduct nondeterministic predictions?

Analysts are primarily responsible for conducting nondeterministic predictions. The nondeterministic predictions referred to here are those focused on predictions of SRQs at the application conditions for the system of interest. These predictions are the culmination of all of the efforts devoted to V&V. This activity is commonly the most demanding in terms of time and resources because it brings together (a) the V&V efforts directed to the system of interest, (b) the modeling effort devoted to identifying and characterizing the environments and scenarios of the system of interest, and (c) the UQ for the system of interest. The level of effort devoted to the nondeterministic predictions can vary widely, depending on the needs of the analysis. As a result, the burden of responsibility on the analyst can vary considerably. We briefly describe two extremes of nondeterministic predictions to stress the range of responsibility.

For the first extreme, consider the emergency cooling of the core of a nuclear power reactor. This type of analysis involves a large team of people, commonly spread over several organizations, because it deals with a high-consequence system in an abnormal environment. For this environment, many possible scenarios are identified and several of

the highest risk scenarios are analyzed in detail. The nondeterministic analysis includes many different types of mathematical models, codes that have gone through years of verification testing, experimental data from a wide variety of experiments, and an extensive range of uncertainties that are carefully characterized. As is appropriate, much is expected of the analysis in order to support risk-informed decision making related to the system. This type of nondeterministic analysis is highly demanding in terms of time, resources, and the variety of technical expertise needed on the team, but it is appropriate for high-consequence systems.

For the second extreme, consider the solid mechanics analysis of a non-safety-critical component of a gas turbine engine. Here, we presume the analysis is directed toward an incremental design change of the component, but the change is expected to have minimal impact on the performance or reliability of the system. As a result, one individual conducts the nondeterministic analysis. Although the analysis could take into account different environments, scenarios, and a wide range of uncertainties, the analyst decides that because of schedule and resource constraints, he is only able to conduct a meager nondeterministic analysis. This situation is very common in industrial settings where competition between manufacturers is intense and time scales are very short. This type of industrial environment for scientific computing was well summarized by Hutton and Casey (2001). Risks due to approximations and short cuts in the non-deterministic analysis are greatly mitigated because relatively little is required from the analysis. For this type of situation, the primary basis for decision-making is the extensive operating experience and developmental testing of the system.

16.2.4.2 Who should require nondeterministic predictions?

Consistent with our previous argument, the customer for the results should require nondeterministic predictions, if these are needed for the decisions at hand. It is now standard practice in the design of most engineered systems that some portions of the analysis of the performance, reliability, and safety of the system are nondeterministic. Some fields of engineering still rely on a factor of safety design procedure or some even declare the reliability of their system is so high that the reliability is unity. In preliminary design studies and physical phenomena research, however, nondeterministic analyses are rarely conducted. For these types of situations a deterministic analysis is appropriate, except for the following cases: (a) when there are large uncertainties in important input quantities to the analysis, and (b) when the SRQs of interest are highly sensitive to uncertainties in any input quantities. For risk-informed decision making on final designs that rely heavily on M&S, the appropriate level of effort that is devoted to nondeterministic analyses should depend on a wide range of technical, liability, and resource factors.

A summary of our views on primary responsibilities for the conduct of an activity and who should require confirmation of the activity is shown in Table 16.1. The table shows how the responsibilities can change between the various participants depending on how the participant is involved in a particular activity. The *supplier* designation in the table

Table 16.1 *Primary role of participants in VV&UQ activities.*

Activity	Code developer	Analyst (code user)	Analysis result user	Experimentalist
SQA and code verification	Supplier	Customer	Customer	–
Solution Verification	–	Supplier	Customer	–
Validation	–	Supplier	Customer	Supplier
Nondeterministic predictions	–	Supplier	Customer	–

indicates the participant who is primarily responsible for the conduct of the activity, and the *customer* designation indicates who is primarily responsible for confirmation that the activity is satisfactorily completed.

16.3 Management actions and responsibilities

This section discusses various actions that management can take to help or hinder the development, implementation, and effectiveness of good VV&UQ practices in an organization. Most of the discussion is directed toward private businesses and government organizations, while some discussion is also appropriate for research institutes and universities. The topics are discussed in the context of responsibilities of management of scientific computing projects whose results are either used in their organization or are provided as a service to other organizations. Some of the recommendations for management are for the purpose of leading and motivating staff toward more effective and efficient VV&UQ practices.

When we refer to *management*, we are primarily referring to line management of an organization. Line managers are those appointed to the relatively permanent structure needed for the coordination and functioning of an organization. In certain instances, the discussion of actions and responsibilities will also refer to project management. Project managers are those whose responsibilities are directly tied to the execution and completion of a project. Project managers in business are commonly under severe constraints related to scope, schedule, and cost of the project. We recognize and respect their perspective and project responsibilities.

16.3.1 Implementation issues

Many of the responsibilities discussed in Section 16.2, with regard to who should require a certain activity, are directed at staff and project management. Often the situation arises where the manager is not familiar with, or technically trained in, all of the VV&UQ activities in a project. As a result, the manager is placed in a difficult position with regard to assessing the adequacy of all of the activities needed to achieve the goals of the computational analysis. For example, there may be inadequate solution verification that could place

the analysis results into question, or worse, the results are more influenced by solution error than physics modeling. The manager may also perceive a risk with regard to large uncertainties in predicted system performance and he/she may need to move resources from one activity to fund another activity that appears to be lacking. The predictive capability maturity model (PCMM), discussed in Chapter 15, Maturity assessment of modeling and simulation, can be helpful in this regard. The PCMM is a framework for assessing the maturity of each of the four activities discussed above, in addition to the activities of representation of geometric fidelity and physics and material model fidelity. The PCMM is focused on assessing M&S maturity for an application-focused project. The manager can require that the analysts involved in the six M&S activities complete the PCMM table and present their maturity assessment at progress reviews of the project. If appropriate, the manager may also wish to add an activity referred to as *reproducibility, traceability, and documentation* of the six previously mentioned activities. Even if the manager is not familiar with certain activities being assessed, he/she can learn a great deal not only from the maturity assessment presented, but also how the analyst justifies the assessment results. After the PCMM assessment is complete, the manager can then determine if the assessed levels of maturity are adequate for the requirements of the project.

It is our observation that in many organizations there is either a competitive relationship between computational analysts and experimentalists, or there is a major disjunction between the two. This type of relationship can be found at the individual level as well as between computational and experimental groups. It could be due to (a) competition over organizational resources, (b) a perception of unjustified recognition of one group over the other, (c) experimentalists feeling that increasing capability in scientific computing will reduce or eliminate the need for their expertise, or (d) simply due to the divisional structure of the organization. Management often does not recognize the problem, or if they do, they may consciously or subconsciously ignore it. Even without competitive or adversarial pressures, there is commonly a notable difference in technical and personal cultures between computational analysts and experimentalists. For validation activities to contribute to the greatest degree possible, it is imperative that management assesses the state of the relationship between analysts and experimentalists in their organizations. In addition, it is critical that management *creates opportunities* for bringing the different individuals, groups, and cultures together in cooperative and mutually beneficial efforts. For example, management must make it clear that the success of a validation team effort will benefit both groups equally, and that failure will be the responsibility of both groups. By *success* we mean high quality computational simulations, experimental measurements, model accuracy assessment, and timeliness of the effort; *not* whether the agreement between computation and experiment was good or bad.

Implementation of most of the approaches and procedures recommended in this book will be neither inexpensive nor easy. Furthermore, some of these approaches may even be technically or economically impractical in particular situations. In addition, some of the approaches and procedures have not been developed satisfactorily for implementation in a design-engineering environment. With each included step, however, the quality of the

VV&UQ processes will be improved, resulting in increased confidence in the simulation results. We firmly believe that VV&UQ is a process, *not* a product. In addition, VV&UQ are not processes that can be *inspected in* the simulation results. That is, the technical complexities of M&S and VV&UQ do not allow for a comprehensive set of rules to be laid down, and then simply followed. Good VV&UQ processes are akin to best practices in engineering and business. The development of good VV&UQ processes requires a change in the culture of scientific computing; a culture of healthy skepticism and disbelief. Even though technological change can occur at a stunning rate, changes in habits and traditions are exceedingly slow and painful.

Some individuals and managers will overtly or covertly thwart the implementation of VV&UQ activities because they believe that: (a) the costs of the activities exceed their value added, and (b) it unnecessarily delays the completion of a computational analysis. If the resistance to VV&UQ is done openly, then the management team must have frank discussions concerning the value added by VV&UQ, and the costs and schedule implications of implementing VV&UQ activities. If resistance is hidden or surreptitious, then it is much more difficult for management to address.

From a business perspective, the results from a computational analysis is considered as a *service*, because it involves the production of intangible goods; specifically, the generation of knowledge. As has been stressed throughout this book, VV&UQ add quality to the computational analysis, i.e., to the service. Business will evaluate the quality improvement due to VV&UQ in the same way as the introduction of any new technology or business process; what is the value added compared to the resources and time expended? Although many business sectors have devised innovative methods to measure quality of the service provided, we contend that the quality added by VV&UQ to the knowledge generated in a computational analysis is significantly more difficult to measure.

The following is a suggested framework useful for considering the business issue of cost versus benefit: what is the cost and time required for VV&UQ compared to the risk of incorrect or improper decisions made based on simulation. In probabilistic risk assessment, the risk is typically defined as the product of the probability of the occurrence of an adverse event and the consequence of the event. As the level of effort devoted to VV&UQ increases, it is reasonable to expect that the probability of incorrect or improper decisions based on simulation decreases. The decrease in the probability, however, is difficult to measure. One recommended guideline is that one must weigh how much of the decision is based on computational analysis versus other traditional factors, such as experimental testing and operational experience with the system. Stated differently, how far beyond our base of experience and experimental data are we relying on the predictions from the analysis? The consequence of an incorrect or improper decision can also be difficult to measure, primarily because there is such a variety of consequences that can be considered. The only guideline we can recommend is that the assessment of consequences should be broadly examined, as discussed above, not only in the short term, but also over the long term. For example, suppose an erroneous conclusion is made concerning the physics of some process simulated in an article in a research journal. The erroneous result would rightly be viewed

as a low-consequence risk, if noticed at all. On the other hand, if erroneous conclusions based on simulation are made on important aspects of a system, decision makers could place at risk their corporation, their customers, the public, or the environment.

16.3.2 Personnel training

Individuals experienced and trained in V&V are relatively rare at the present time because V&V, as recognized fields of research and application, are actually quite new. There is much wider experience and training, however, in the foundational fields of V&V: SQA, numerical error estimation, experimental uncertainty estimation, and probability and statistics. University graduate courses in the foundational topics have been in existence for at least four decades. Some major universities in the US are now beginning to teach graduate courses specifically in the field of V&V. Consequently, it will be at least another decade or two before there will be a sizeable cadre of highly qualified individuals in the field of V&V.

We suggest that if there are at least several experienced individuals in the field of VV&UQ within an organization, then management should consider forming a group of these individuals. This group can provide training and mentoring for other staff members, as well as leading VV&UQ activities within a project. This option will be discussed more in Section 16.3.4.

During the near term, the best method for training staff and managers in V&V is through professional development short courses. These continuing education courses typically range from one to five days and are offered by professional societies, universities, and institutes. They are also taught at professional society conferences or, upon request, at the site of the interested organization. At the present time there are roughly five short courses taught in the field of V&V by different organizations.

16.3.3 Incorporation into business goals

The incorporation of computational analyses into the goals of a business, or any organization, is a broad topic. Here, we will focus on how VV&UQ contribute to the quality of the information generated in an analysis. In Chapter 15, we discussed the results of the comprehensive study by Wang and Strong (1996) concerning the most important qualities of information. They categorized the key attributes into four aspects:

- *intrinsic information quality*: believability, accuracy, objectivity, and reputation;
- *contextual information quality*: value added, relevancy, timeliness, completeness, and amount of information;
- *representational information quality*: interpretability, ease of understanding, consistent representation, and concise representation;
- *accessibility information quality*: accessibility and security aspects.

In the following sections, we will briefly discuss how VV&UQ contribute to the first three attributes of information quality.

16.3.3.1 Intrinsic information quality

Throughout this book, we have dealt with four intertwined issues of VV&UQ: (a) quality of simulation software, (b) accuracy of the numerical calculations, (c) uncertainty in the validation experiments, and (d) uncertainty in predictions. When we consider the intrinsic information quality of simulation in terms of how it is used, we must broaden our perspective away from the details of the analysis. The proper perspective should be: how is the information from an analysis properly used by management in a decision-making environment? We argue that the proper use of the information is, in large part, dependent on the accuracy and objectivity aspects of intrinsic information quality. To make this point, we contrast the broader and more comprehensive meaning of the term *accuracy*. Here, we stress the meaning of accuracy in sense of faithfulness, objectiveness, and completeness, as opposed to the implication of precision of the results. We make this point in two different contexts.

First, predictions for any real engineering system must deal with uncertainty in the response of the system. The uncertainty can be due to many sources, for example: (a) the environment in the sense of normal, abnormal, or hostile; (b) the possible scenarios that can occur in the system, (c) the uncertainty inherent in the system; (d) the uncertainty in the influence of the surroundings on the system; and (e) the uncertainty due to the mathematical model used in the analysis. The simulation result, therefore, must be characterized as an uncertain number; *not simply a number*. The uncertain number can be expressed in several ways, such as an interval, a precise probability distribution, or a p-box. Several fields of engineering and systems analysis moved to this paradigm decades ago; e.g., nuclear reactor safety, structural dynamics, and the broad field of risk assessment. Many fields of engineering and science, however, have not moved from the tradition of deterministic predictions. More importantly, many decision makers in business are not comfortable with the concept of simulation results presented as an uncertain number. It can be disconcerting to let go of the apparent precision of a deterministic prediction, but the precision is only a ruse for accuracy. Quoting a Chinese proverb, "To be uncertain is uncomfortable, but to be certain is ridiculous."

Second, when uncertainty is incorporated in the analysis, through whatever source, one must commonly deal with the issue of aleatory and epistemic uncertainty. As discussed in several chapters, including epistemic uncertainty in the analysis can result in a large increase in uncertainty in the response of the system, as compared to characterization of the uncertainty as a random variable. For example, if an input uncertainty is initially characterized as a uniform distribution over some interval, but then changed to an interval-valued quantity over the same interval, there can be a large increase in the uncertainty of a SRQ. The characterization of the SRQ changes from a single probability distribution associated with the uniform distribution to a p-box, i.e., an infinite ensemble of distributions. Some experienced risk analysts have commented that simply changing an input probability distribution to an interval will not yield a significant change in the characterization of the output. If the uncertain input quantity were a significant contributor to uncertainty in the output, then they would be stunned at the increase in uncertainty. With regard to simulation

accuracy, the point can be made in the following way. If certain inputs to the simulation are known so poorly that the knowledge can only be represented as an interval, then the p-box result of the system responses is the *least uncertain result* that can be claimed. An assumption that characterizes the input as a random variable, such as assuming a uniform distribution over the interval, will actually *under-represent* the uncertainty of the responses to the decision maker. The result showing less uncertainty in the system response can be viewed as being more precise, but it is indeed less accurate and, at best, misleading to the decision maker.

16.3.3.2 Contextual information quality

The two attributes of contextual information quality that we will stress here, with regard to decision making, are value added and timeliness. Several chapters in this book have dealt with the added value of computational analyses in the sense of detailed knowledge generated for the decision maker. To better appreciate our stress on added value, consider the wisdom that can be learned from examining past failures. Here, we specifically refer to knowledge that can be learned by examining the root causes of past engineering system failures (Dorner, 1989; Petroski, 1994; Vaughan, 1996; Reason, 1997; Chiles, 2001; Gehman *et al.*, 2003; Lawson, 2005; Mosey, 2006). Most of these authors stress the importance of errors in judgment of the responsible decision makers. Sometimes the error is made by an individual decision maker, but in catastrophic failures of large systems the error is more commonly made by a group of decision makers, i.e., an organizational failure. Petroski (1994) bluntly states:

> Human error in anticipating failure continues to be the single most important factor in keeping the reliability of engineering designs from achieving the theoretically high levels made possible by modern methods of analysis and materials. This is due in part to a de-emphasis on engineering experience and judgment in the light of increasingly sophisticated numerical and analytical techniques. (pp. 7–8)

His view, and that of many others who help us to learn from our past mistakes, is primarily directed at the lack of vigor that some project managers have toward investigating how their system can fail or could cause other associated systems to fail. Many managers tend to view VV&UQ with the same mind set. Instead of grasping how VV&UQ adds value to the quality of the information provided in the analysis, they see it as a potential risk to their project or personal agenda, or simply as a drain on resources. This is especially true if the VV&UQ activity is controlled and funded by a separate project or line manager who does not report to them, because then they have little or no control over the activities. This type of systemic failure in the attitude of a project toward a safety organization was scathingly criticized by Adm. Harold Gehman as part of his testimony to a US Senate Committee concerning the loss of the Space Shuttle Columbia and her crew: "The [NASA] safety organization sits right beside the person making the decisions. But behind the safety organization, there's nothing back there. There's no people, money, engineering expertise, analysis." (AWST, 2003).

A widespread difficulty encountered when using simulation results in a design-engineering environment is the lack of appreciation of timeliness among simulation analysts. The time scales required for making the multitude of design decisions on a system are generally very short. Some analysts, especially applied research analysts, are completely unaccustomed to these schedule requirements. For example, it is not uncommon for the design engineer to tell the analyst "If you get the analysis results to me by next week, I will use the results to make a design decision. If you don't, I will make the decision without your help." VV&UQ activities take time to complete, possibly resulting in a degradation of the prompt response times required. However, as discussed above, there must be an appropriate tradeoff between the time required to produce an adequately reliable result and the consequences of improper decisions based on an inaccurate or incomplete result.

16.3.3.3 Representational information quality

As part of the attribute of representational information quality, we will stress the importance of interpretability (and ease of understanding) and consistent representation of results and VV&UQ activities. Most people think about this attribute of information quality with regard to written documentation of a simulation. Although written documentation is certainly important for a permanent record of the details of an analysis, we believe representational information quality is a more important quality with regard to oral presentations to management. It is our view that much of the discussion and debate concerning significant decisions occurs when presentations to management are made. By the time detailed documentation is prepared and published, usually at the completion of the project, most decisions are cast in stone.

In any summary presentation of a mathematical modeling approach and computational results to management, the importance of interpretability and ease of understanding cannot be overstated. This is in no way a criticism of management's abilities or backgrounds. As part of human nature, technical staff who prepare the summary presentations tend to think that others have similar backgrounds; technically, experientially, and culturally. This is, in general, a gross misjudgment that can be devastating to the clarity and effectiveness of a presentation. Summary presentations must greatly condense the amount of information generated from developing, testing, analyzing, and understanding the models and the simulations, obtaining experimental data, and interpreting the results. For large team efforts, the condensation factor is even greater. Similarly, large condensations of information also occur when presentations are given to external review panels or governmental regulatory agencies. As a result, management must stress to their staff the crucial importance of interpretability and ease of understanding of an analysis in their presentations.

In this same regard, we urge that the presentations have a proper balance, explaining the strengths *and* weaknesses of not only the M&S effort, but also the VV&UQ activities. Presentations tend to overly stress the strengths of an analysis and minimize the weaknesses, if they are mentioned at all. A presentation or detailed documentation of an M&S effort that does not discuss the effects of important assumptions should be highly suspect by managers

and external reviewers. Ignoring the justification and effects of important assumptions is, in our view, one of the single most damaging indicators of the quality of an M&S effort. Analyses that have not identified, or do not discuss, their primary weaknesses can expose decision makers and stakeholders to significant risks.

The importance of significant assumptions also carries over to the second aspect of representational information quality: consistent representation. It is common practice that summary presentations of a major M&S effort are repeated to different audiences, usually with a slightly different emphasis, depending on the nature and interest of the audience. Although it is appropriate to stress different aspects of the analysis in each presentation, it is important that the primary issues receive consistent representation. For example, consider the situation where presentations to lower level management are made by the staff who conducted the analysis. At this level, time is usually available for significant detail to be presented, along with the major assumptions made in the analysis. It is common in many organizations, especially in organizations that place heavy emphasis on a hierarchical management structure, that managers of the staff who conducted the analysis brief higher levels of management on the analysis. The presentations to upper management are usually accompanied by additional condensation of the information. We recommend that VV&UQ activities be included in these higher level presentations, and not eliminated from the presentations with the argument "Of course we conducted adequate VV&UQ measures." This can jeopardize the consistency of the representation of the information and it can put higher levels of management at increased risk of making poor decisions.

16.3.4 Organizational structures

There are two basic organizational approaches to the deployment of V&V practices in a business or government agency. One involves the formation of a small organization within the parent organization that is composed of experienced V&V staff. This organization is responsible for developing and deploying V&V practices and training of staff and management in other organizations within the parent company. The second is a dispersed approach where the experienced staff are placed in various organizations that are conducting computational analyses and experimental validation activities. Since there are a relatively small number of individuals trained and experienced in V&V at the present time, and since V&V serves in a support role to simulation, it is an open question as to which approach is best suited to the deployment of V&V. In fact, it is certainly reasonable that one approach may be optimum for one organization, while the other approach would be better in a different organization. There are many site-specific factors that managers should consider in deciding which approach would be best for their organization. Here, we will briefly discuss some of the features of each approach, as well as some of the advantages and disadvantages.

There are a number of advantages in forming a V&V group. With close interactions in the group, it can serve as a critical mass of ideas and expertise so that additional energy is generated around the topic. This approach was used at Sandia National Laboratories

beginning in 1999 with a large degree of success. Included in this group were staff members who, although not familiar with V&V, were experienced in UQ. This combination of both activities in the same group proved to be an excellent approach. As V&V methodologies developed, the strong connection between validation and UQ, particularly model uncertainty, became better understood. Another benefit in the initial formation of the group was the inclusion of experienced individuals in nuclear reactor safety analyses. These individuals were experienced not only in each topic of VV&UQ, but also in risk analysis techniques, e.g., fault tree analyses and the phenomena identification and ranking table (PIRT).

The existence of a VV&UQ group in an organization is, even today, an anomaly. Since there is essentially no formal university training devoted exclusively to VV&UQ at the present time, the mixture of technical disciplines of the staff in a VV&UQ group is expected to be diverse. Although the principles of VV&UQ are the same regardless of the application area, there are noticeable differences in some of the techniques depending on the application area. For example, in fluid dynamics, heat transfer, and solid mechanics there are many applications that are steady state or involve slowly varying responses as a function of time. In structural dynamics, shock wave dynamics, and electrodynamics, however, the time series nature of the responses is a dominant factor that significantly complicates VV&UQ. Although the group would probably be focused on simulation, there should be some expertise included with a background in experimental methods. It has been our experience that individuals who have experience in both computational and experimental methods are the most adept at grasping the philosophy of validation and the principles for the design and execution of validation experiments. An individual who was entirely focused on experiments, however, would probably not fit well in a group such as this.

The responsibilities of the VV&UQ group should include three areas. First, the group should develop VV&UQ techniques that are useful not only for the needs within the group, but also for other organizations within the parent organization. Some examples of development areas are (a) manufactured solutions, (b) numerical solution error estimators, (c) validation metrics, and (d) propagation methods applicable to both aleatory and epistemic uncertainties. Second, the group should deploy VV&UQ practices to other computational and experimental groups. Deployment of practices can take several forms, such as (a) explain and promote the concepts, procedures, and benefits of VV&UQ; (b) apply VV&UQ techniques to various computational projects within the parent organization, i.e., become a practitioner instead of just a theorist; and (c) write and document software packages that are useful in various VV&UQ tasks. Concerning this last suggestion, it responds to the criticism that is sometimes posited by other organizations: "We don't have the expertise or the time to write the needed software packages to do what you are promoting." Third, the group should train staff and managers in other organizations concerning VV&UQ practices. Ideas for different types of training are (a) offer short courses or formal training on topics, (b) serve as consultants for computational and experimental projects in other organizations, and (c) serve as mentors to train and advise staff in VV&UQ. We should

stress that significant training is usually needed with regard to UQ concepts and methods. Very few engineering and science educational disciplines involve courses in statistics and the analysis of uncertainty in systems.

The second approach to the deployment of V&V practices is to disperse experienced V&V staff into various computational and experimental project groups. For a small organization, this is a more appropriate method. The individuals could accomplish the suggestions given in the previous paragraph, but their focus would be more on tactical, near term issues of their parent project group. The success of this approach would depend in part on how much time the project manager would allow the staff member to devote to V&V activities. If the manager were to allow a significant portion of time to be devoted to both near-term and long-term V&V needs, then this arrangement could be effective. If project tasks, for example those directed toward building new mathematical models for the computational analysis, were to always take priority over V&V needs, then little would be accomplished in moving the group toward better V&V practices.

Whichever deployment method is used, one key factor is crucial to the success of V&V deployment in an organization: the strong and genuine support of management for VV&UQ practices throughout the organizations involved in computational analyses and validation experiments. If management, in all organizations involved and at multiples levels, does not genuinely support VV&UQ, then deployment will be spotty at best; or at worst, a failed effort. As has been discussed in this and several other chapters, the key to long-term success of VV&UQ is the understanding of the value added to the information quality of computational analyses. To comprehend this, management must embrace a commitment to change the culture of M&S in their organizations.

16.4 Development of databases

This section discusses the major V&V databases and gives recommendations for improvements in the quality and usability of databases in the future. We only discuss major databases that are either publicly available or available on a membership basis. There are, as one would suspect, major proprietary databases built and maintained by for-profit organizations. Some examples of these are the large commercial software companies and essentially every large manufacturer that uses scientific computing in the design, optimization, and evaluation of their products. These proprietary databases are not addressed here.

The suggested recommendations for future development of publicly available or membership-based databases are made because we believe that the influence of scientific computing in the world will continue to grow in depth and breadth. We also believe that V&V databases built around individual technical fields can be a very positive influence on the quality and reliability of scientific computing in these fields. For those industries where the competitive pressure between manufacturers is intense, we argue that simplified test cases, i.e., lower levels in the validation hierarchy, can be found where competitors can safely share data. For these simplified cases, it can be more cost effective and efficient to jointly build and share V&V data than using proprietary databases. Some corporations

feel that their proprietary data is the lifeblood to their continued competitiveness. We agree that some of these data are, but not all. Unless these data are well documented, effectively catalogued, easily found, and efficiently retrieved, the data are of little value. The business model of having key senior staff be the keepers of the data jewels is no longer realistic and it can easily fail for various reasons.

16.4.1 Existing databases

During the last two decades, the National Agency for Finite Element Methods and Standards (NAFEMS) has developed some of the most widely known V&V benchmarks (NAFEMS, 2006). Roughly 30 verification benchmarks have been constructed by NAFEMS. The majority of these benchmarks have targeted solid mechanics simulations, though some of the more recent benchmarks have been in fluid dynamics. Most of the NAFEMS verification benchmarks consist of an analytical solution or an accurate numerical solution to a simplified physical process described by a partial differential equation. The NAFEMS benchmark set is carefully defined, numerically demanding, and well documented. However, these benchmarks are currently very restricted in their coverage of various mathematical and/or numerical challenges (such as discontinuities) and in their coverage of physical phenomena. Further, the performance of a given code on the benchmark is subject to interpretation by the user of the code. It is also likely that the performance of a code on the benchmark is dependent on the experience and skill of the user.

In the field of nuclear reactor engineering, the Nuclear Energy Agency, Committee on the Safety of Nuclear Installations (CSNI) devoted significant resources toward developing validation benchmarks, which they refer to as International Standard Problems (ISPs). This effort began in 1977 with recommendations for the design, construction, and use of ISPs for loss-of-coolant accidents (LOCAs) (NEA, 1977). The CSNI recognized the importance of issues such as (a) providing a detailed description of the actual operational conditions in the experimental facility, not simply those conditions that were requested or desired; (b) preparing careful estimates of the uncertainty in experimental measurements and informing the analyst of the estimates; (c) reporting the initial and boundary conditions that were realized in the experiment, not those conditions that were simply desired; and (d) conducting a sensitivity analysis to determine the most important factors that affect the predicted system responses of interest. The CSNI has continually refined the guidance for ISPs, such that the most recent recommendations for the ISPs address any type of experimental benchmark, not just benchmarks for LOCA accidents (CSNI, 2004). Thus, the primary goal of the ISPs remains the same for all types of benchmark: "to contribute to a better understanding of postulated and actual events" that could affect the safety of nuclear power plants.

A number of efforts have been undertaken in the development of validation databases that could mature into well-founded benchmarks. In the United States, the NPARC Alliance has developed a validation database that has roughly 20 different flows (NPARC, 2000). In Europe, starting in the early 1990s, there has been a much more organized effort to develop

validation databases. These databases have primarily focused on aerospace applications. ERCOFTAC (the European Research Community on Flow, Turbulence and Combustion) has collected a number of experimental datasets for validation applications (ERCOFTAC, 2000). QNET-CFD is a thematic network on quality and trust for the industrial applications of CFD (QNET-CFD, 2001). This network has more than 40 participants from several countries who represent research establishments in many sectors of the industry, including commercial CFD software companies. For a history and review of the various efforts, see Rizzi and Vos (1998) and Vos *et al.* (2002).

We note that the validation databases described by Rizzi and Vos (1998) and Vos *et al.* (2002) contain many cases that are for complex flows, which are sometimes referred to as *industrial applications*. We have observed, both through our own experience and in the scientific literature, that attempts to validate models on complex physical processes are commonly unsuccessful for two reasons. First, inadequate information concerning detailed system features, boundary conditions, or initial conditions are provided by the experimentalists. Second, the computational results compare very poorly with the experimental measurements for difficult-to-predict SRQs. Then, the computational analysts often do one of the following: (1) they engage in a model calibration activity, adjusting both physical and numerical parameters in the model, to obtain better agreement; (2) they reformulate the assumptions in their model to obtain better agreement, thereby changing the model; or (3) they start pointing accusatory fingers at the experimentalists about either what is wrong with the experimental data or what the experimentalists should have measured to make the data more effective for validation. Our view of these responses by the analysts has been discussed in several chapters dealing with validation and prediction.

16.4.2 Recent activities

Oberkampf *et al.* (2004) introduced the concept of *strong-sense benchmarks* (SSBs) in V&V. They argued that SSBs should be of sufficiently high quality that they could be viewed as *engineering reference standards*. They stated that SSBs are test problems that have the following four characteristics: (1) the purpose of the benchmark is clearly understood, (2) the definition and description of the benchmark is precisely stated, (3) specific requirements are stated for how comparisons are to be made with the results of the benchmark, and (4) acceptance criteria for comparison with the benchmark are defined. In addition, they required that information on each of these characteristics be promulgated, i.e., the information is well documented and publicly available. Although a number of benchmarks are available, a few of which were discussed above, these authors asserted that SSBs do not presently exist in science or engineering. They suggested that professional societies, academic institutions, governmental or international organizations, and newly formed nonprofit organizations would be the most likely to construct SSBs.

Oberkampf and Trucano (2008) present an in-depth discussion of how SSBs should be constructed and describe the key features that are needed to qualify as an SSB for both verification and validation. Concerning verification benchmarks, the following elements

are discussed that should be contained in the documentation of a verification benchmark: (a) conceptual description, (b) mathematical description, (c) accuracy assessment, and (d) additional user information. Examples are provided for applying these elements to the four types of benchmarks, namely, manufactured solutions, analytical solutions, numerical solutions to ordinary differential equations, and numerical solutions to PDE models. Oberkampf and Trucano (2008) recommend that when a candidate code is compared with a verification benchmark, the results of the comparisons with benchmarks should *not* be included in the benchmark documentation *per se*. They also discuss how formal comparison results could be used and identify the types of information that should be included in the comparisons.

Concerning validation benchmarks, Oberkampf and Trucano (2008) present four elements that should be contained in the documentation of a validation benchmark: (a) conceptual description; (b) experimental description; (c) uncertainty quantification of benchmark measurements; and (d) additional user information. They also discuss how candidate code results could be compared with the benchmark results, paying particular attention to issues related to the computation of nondeterministic results to determine the uncertainty of SRQs due to uncertainties in input quantities, the computation of validation metrics to quantitatively measure the difference between experimental and computational results, the minimization of model calibration in comparisons with validation benchmarks, and the constructive role of global sensitivity analyses in validation experiments.

They also discuss why validation benchmarks are much more difficult to construct and use than verification benchmarks. The primary difficulty in constructing validation benchmarks is that experimental measurements in the past have rarely been designed to provide true validation benchmark data. The validation benchmarks that have been compiled and documented by organized efforts, some of which were discussed above, are indeed instructive and useful to users of the codes and to developers of physics models. However, they argue that much more needs to be incorporated into the validation benchmarks, both experimentally and computationally, to achieve the next level of usefulness and critical assessment.

16.4.3 Implementation issues of Databases

If V&V SSBs and a database to house them were to become a reality, a number of complex and difficult implementation and organizational issues would need to be addressed. Some of these issues would be, for example,

- agreement on the primary and secondary goals of the database,
- initial construction of the database,
- review and approval procedures for entries into the database,
- open versus restricted use of the database,
- structure of the software framework for searching and retrieving information on SSBs in the database,

- organizational control of the database,
- relationship of the controlling organization to existing private and governmental organizations and engineering societies,
- initial and long-term funding of the database.

These issues are of major importance to the joint community of individuals, corporations, commercial software companies, nonprofit organizations, engineering societies, universities, and governmental organizations with serious interest in improving scientific computing.

Initial construction of the database would be technically and organizationally complex, as well as costly. Populating the database with relevant, high-quality benchmarks would require a wide-ranging effort that cuts across major communities of applied mathematics, model building, experimentation, computation, engineering applications, and business decision making. Putting this kind of collaborative effort together hinges on a careful plan that takes the long-term view for the database. The construction of SSBs is not feasible as a short-term task. Much of what we recommend clearly aims at a sustainable and long-term use of the database, with an implication that the quality and breadth of the database improves over a long period of time. The long-term success of the database requires a sound starting point, with broad consensus from all interested parties about goals, use, access, and funding over the long term.

Broad organizational issues must be addressed very early in the planning stage. For example, will a single organization (nonprofit, academic, or governmental) have responsibility for database maintenance, configuration management, and day-to-day operation? Will the database have a role beyond its immediate community? *Broad impact* then implies that there is the goal of open access to the database for the good of the simulation community, specifically the world community in each of the traditional scientific and engineering disciplines. But how is this goal compatible with the significant expense needed to create, maintain, and improve the database? Potential financial supporters and users of the database would need to be convinced of the value returned to them for their investment. This returned value could be in many forms, such as improvements in their software products, the ability to attract new customers to their software products, and use of the database as a quality assessment tool for organizations or government agencies to allow contractors to bid on new projects. If proprietary information is used in the database, we believe it would greatly diminish or possibly eliminate the ability to create and sustain the database. Some have argued that the database could be constructed so that proprietary information could be segregated from generally available information. We believe that private corporations would not be convinced such segregation could be accomplished with high confidence, and that the database manager would not be able to adequately protect the proprietary information.

It seems that V&V databases of the type we have discussed should be constructed along the lines of traditional engineering and science disciplines, e.g., fluid dynamics, solid dynamics, electrodynamics, neutron transport, plasma dynamics, and molecular dynamics.

How each of these disciplines might begin to construct databases certainly depends on the traditions, applications, and funding sources in each of these fields. The nuclear power industry, for example, has a deeply embedded, long-term tradition of international cooperation. On the other hand, the aerospace industry, both aircraft and spacecraft builders, has a fierce competitive nature. We envision that different implementations and database structures would be chosen in various communities.

We also suggest that a secondary purpose for the establishment and use of SSBs is for the development of best practices in scientific computing. As recognized by NAFEMS (NAFEMS, 2006) and ERCOFTAC (Casey and Wintergerste, 2000), there is a compelling need for improvements in the professional practice of scientific computing. In our opinion, a convincing argument could be made that the most common failures in industrial applications of scientific computing result from mistakes made by practitioners using the code. Corporate and governmental management, of course, shoulders the ultimate responsibility for mentoring and training these practitioners, as well as for monitoring their work products. Given the qualities of SSBs discussed previously, these benchmarks could be viewed as very carefully documented step-by-step sample problems from which practitioners, new and experienced, could learn a great deal.

Rizzi and Vos (1998) and Vos *et al.* (2002) discuss how validation databases could be built and used by a wide range of individuals and organizations. They stress the importance of close collaboration between corporations and universities in the construction, use, and refinement of a validation database. In this regard, they also stress the value of workshops that are focused on specialty topics to improve the modeling efforts and simulations that are compared to experimental data. They discuss a number of workshops and initiatives in Europe, primarily funded by the European Union. Often, these workshops provide dramatic evidence of the power of carefully defined and applied V&V benchmarks. One such effort organized in the United States, but with participants from around the world, is the series of Drag Prediction Workshops (Levy *et al.*, 2003; Hemsch, 2004; Hemsch and Morrison, 2004; Laflin *et al.*, 2005; Rumsey *et al.*, 2005). These workshops have been extraordinarily enlightening from two perspectives: (a) there was great variability in the drag predictions from computational analysts for a relatively simple aircraft geometry, and (b) there were surprisingly large differences between the computational results and the experimental measurements. The key factor in this exercise that resulted in a "surprising large range of results" is that this was a blind comparison. Results from these types of workshop could form the basis for initial submittals of new V&V benchmarks into the database.

We believe an Internet-based system would provide the best vehicle for the deployment of V&V databases for three reasons. First, the ability to build, quickly share, and collaborate with an Internet-based system is now blatantly obvious. A paper-based system would be completely unworkable, as well as decades behind the current state of information technology.

Second, descriptive terms for a particular application of interest could be input to a search engine that could find all of the benchmarks that would contain those terms. The search

engine could operate much like that found in Google or Wikipedia. Functionality could be expanded to include a relevancy-ranking feature that would further improve the search-and-retrieval capability. The overall system design would include configuration-, document-, and content-management elements. Then the benchmarks that were retrieved could be sorted according to their relevance to the words input to the search. One could then select the hyperlinks embedded within any of the benchmarks found. When a particular benchmark is displayed, it could have links from important words in the benchmark description to more detailed information in the benchmark.

Third, the computer-based system could instantly provide much more detail about each benchmark. We recommend that the documentation of V&V benchmarks be produced in an electronic format that is widely usable and robust across many computer operating systems. Of the electronic formats available, Adobe Portable Document Format (PDF) is the most commonly used and has many desirable characteristics; however, we also recommend that this format be supplemented with additional file formats for specialized information. For example, tabular data could be stored in ASCI text files or in Microsoft Excel files; high-resolution digital photographs should be stored in easily usable formats such as TIFF, PDF, and JPEG; digital video files should be stored in formats such as QuickTime, MPEG, or AVI; and computer software should be written in common languages such as C++, Fortran, or Java. The computer software would be necessary for documenting the source terms in database entries submitted for the method of manufactured solutions.

In the long term, new validation experiments should be funded either by the organization controlling the database or by for-profit private, nonprofit, university, or governmental organizations. The organization controlling the database could receive its funding from subscribing members to the organization, and possibly from governmental funding. The funding could be directed to both operation and maintenance of the database and to constructing new V&V benchmarks. When new validation results are entered into the database, there would be a unique opportunity for blind comparisons. As we have stressed several times, blind comparisons are the real test of predictive-capability prowess. We believe that identification of new validation experiments should be the responsibility of both the application community and the database organization. The organizational role and facilitation of discussions regarding which experiments should be conducted is best served by the database organization. For example, the database organization could serve as an unbiased referee between for-profit corporations desiring more application-relevant experiments and model builders who are more knowledgeable of the weaknesses of modeling for complex systems.

16.5 Development of standards

The efficient maturation of V&V depends on concerted efforts by individuals and organizations throughout the world to develop international standards with regard to terminology, basic philosophy, and proven procedures. In the US, the American National Standards

Institute (ANSI) oversees the development and approval of voluntary consensus standards in products and services. Although ANSI does not develop standards itself, it sets the rules and procedures by which standards are developed by member organizations, e.g., engineering societies. ANSI is the official US representative to the two major international standards organizations, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). In a new field, such as V&V, it is critically important to develop standards that (a) have broad participation from industry, government, and universities, (b) are thoroughly vetted both within the committee and to interested groups, and (c) are on sound technical and practical grounds. In a new field, there must be extraordinary caution in developing standards because they are much more than research ideas or seemingly good practices. Standards are more than guides for engineering practices or recommended practices; they are expected to establish the highest levels of reliability and permanence to industry.

Concerning terminology, the efforts of the US Department of Defense, the American Institute of Aeronautics and Astronautics, and the American Society of Mechanical Engineers have been extremely productive in providing understandable, useful, and workable definitions for critical terms in scientific computing. Their definitions are on solid rational and pragmatic grounds. There is still some debate and confusion, however, concerning the encompassing view and the restricted view of validation, as discussed in Chapter 2, Fundamental concepts and terminology. However, we do not feel this is a critical issue, as long as people make it clear which definition they are using. Our primary concern is with the significant difference in V&V terminology between the scientific computing community (the focus of this book) and the ISO/IEEE communities, as discussed in Chapter 2. We are adamant that a rejection and dismissal of the scientific computing definitions by the ISO/IEEE communities would be a *major step backwards* in the development and use of scientific computing around the world. How this dichotomy in definitions will be resolved, and if it will be resolved, is unknown.

In our view, the most appropriate organizations for developing new standards are the professional engineering societies that have ANSI-approved standards writing committees and procedures. At the present time, only ASME has standards committees that are actively producing V&V standards. The AIAA and the American Nuclear Society have standards committees, but they have not recently produced new V&V standards. When these two societies, and possibly others, become more active in developing new V&V standards there must be a concerted effort to promote consistent concepts and procedures. One possibility would be for ANSI to take a more active role in coordinating these activities.

A complementary approach, one that is appropriate at the worldwide level, is the involvement of ISO in writing V&V standards. If this were to occur, it would require close coordination between ISO and ANSI. We believe there is also a contribution to be made by organizations such as the US National Institute of Standards and Technology (NIST), and by similar organizations in the European Union, such as NAFEMS and ERCOFTAC. As more organizations become interested in V&V standards, however, there must be significantly improved coordination among the interested parties.

16.6 References

Abanto, J., D. Pelletier, A. Garon, J.-Y. Trepanier, and M. Reggio (2005). Verification of some commercial CFD codes on atypical CFD problems. *43rd AIAA Aerospace Sciences Meeting and Exhibit*, Paper 2005–0682, Reno, NV, AIAA.

ABAQUS (2006). *ABAQUS Benchmarks Manual*. Version 6.6, Providence, RI, ABAQUS Inc.

ANSYS (2005). *ANSYS Verification Manual*. Release 10.0, Canonsburg, PA, ANSYS, Inc.

AWST (2003). Slamming shuttle safety. *Aviation Week & Space Technology*, May 2003, 23.

Casey, M. and T. Wintergerste, eds. (2000). *ERCOFTAC Special Interest Group on Quality and Trust in Industrial CFD: Best Practices Guidelines*, Lausanne Switzerland, European Research Community on Flow, Turbulence, and Combustion.

Chiles, J. (2001). *Inviting Disaster-Lessons from the Edge of Technology*. 1st edn., New York, HarperCollins.

CSNI (2004). *CSNI International Standard Problem Procedures, CSNI Report No. 17 – Revision 4*. NEA/CSNI/R(2004)5, Paris, France, Nuclear Energy Agency, Committee on the Safety of Nuclear Installations.

Dorner, D. (1989). *The Logic of Failure, Recognizing and Avoiding Error in Complex Situations*, Cambridge, MA, Perseus Books.

ERCOFTAC (2000). *Portal to Fluid Dynamics Database Resources*. ercoftac.mech.surrey.ac.uk.

Gehman, H. W., J. L. Barry, D. W. Deal, J. N. Hallock, K. W. Hess, G. S. Hubbard, J. M. Logsdon, D. D. Osheroff, S. K. Ride, R. E. Tetrault, S. A. Turcotte, S. B. Wallace, and S. E. Widnall (2003). *Columbia Accident Investigation Board Report Volume I*. Washington, DC, National Aeronautics and Space Administration, Government Printing Office.

Hemsch, M. (2004). Statistical analysis of computational fluid dynamic solutions from the Drag Prediction Workshop. *Journal of Aircraft*. **41**(1), 95–103.

Hemsch, M. and J. H. Morrison (2004). Statistical analysis of CFD solutions from 2nd Drag Prediction Workshop. *42nd AIAA Aerospace Sciences Meeting and Exhibit*, Reno, NV, American Institute of Aeronautics and Astronautics, 4951–4981.

Hutton, A. G. and M. V. Casey (2001). Quality and trust in industrial CFD – a European perspective. *39th AIAA Aerospace Sciences Meeting*, AIAA Paper 2001–0656, Reno, NV, American Institute of Aeronautics and Astronautics.

Laflin, K. R., S. M. Klausmeyer, T. Zickuhr, J. C. Vassberg, R. A. Wahls, J. H. Morrison, O. P. Brodersen, M. E. Rakowitz, E. N. Tinoco, and J.-L. Godard (2005). Data summary from the second AIAA Computational Fluid Dynamics Drag Prediction Workshop. *Journal of Aircraft*. **42**(5), 1165–1178.

Lawson, D. (2005). *Engineering Disasters – Lessons to be Learned*, New York, ASME Press.

Levy, D. W., T. Zickuhr, R. A. Wahls, S. Pirzadeh, and M. J. Hemsch (2003). Data summary from the first AIAA Computational Fluid Dynamics Drag Prediction Workshop. *Journal of Aircraft*. **40**(5), 875–882.

Mosey, D. (2006). *Reactor Accidents: Institutional Failure in the Nuclear Industry*. 2nd edn., Sidcup, Kent, UK, Nuclear Engineering International.

NAFEMS (2006). NAFEMS Website. www.NAFEMS.org.

NEA (1977). *Loss of Coolant Accident Standard Problems*. Committee on the Safety of Nuclear Installations, Report No. 17, Paris, France, Nuclear Energy Agency.

NPARC (2000). *CFD Verification and Validation: NPARC Alliance*. www.grc.nasa.gov/WWW/wind/valid/homepage.html.
Oberkampf, W. L. and T. G. Trucano (2008). Verification and validation benchmarks. *Nuclear Engineering and Design*. **238**(3), 716–743.
Oberkampf, W. L., T. G. Trucano, and C. Hirsch (2004). Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*. **57**(5), 345–384.
Petroski, H. (1994). *Design Paradigms: Case Histories of Error and Judgment in Engineering*, Cambridge, UK, Cambridge University Press.
Pohl, R., ed. (2004). *Cognitive Illusion: a Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. New York, Psychology Press.
QNET-CFD (2001). *Thematic Network on Quality and Trust for the Industrial Applications of CFD*. www.qnet-cfd.net.
Reason, J. (1997). *Managing the Risks of Organizational Accidents*, Burlington, VT, Ashgate Publishing Limited.
Rizzi, A. and J. Vos (1998). Toward establishing credibility in computational fluid dynamics simulations. *AIAA Journal*. **36**(5), 668–675.
Roache, P. J. (1998). *Verification and Validation in Computational Science and Engineering*, Albuquerque, NM, Hermosa Publishers.
Rumsey, C. L., S. M. Rivers, and J. H. Morrison (2005). Study of CFD variation on transport configurations for the second Drag-Prediction Workshop. *Computers & Fluids*. **34**(7), 785–816.
Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*, Chicago, IL, The University of Chicago Press.
Vos, J. B., A. Rizzi, D. Darracq, and E. H. Hirschel (2002). Navier–Stokes solvers in European aircraft design. *Progress in Aerospace Sciences*. **38**(8), 601–697.
Wang, R. Y. and D. M. Strong (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*. **12**(4), 5–34.