# 13

# Predictive capability

This chapter will synthesize the key results from the previous chapters and incorporate them into modern predictive capability in scientific computing. This chapter, in contrast to all other chapters, does not stress the theme of assessment. Here we discuss the fundamental steps in conducting a nondeterministic analysis of a system of interest. With this discussion we show how verification and validation (V&V) can directly contribute to predictive capability.

The previously covered material and the new material are organized into six procedural steps to make a prediction:

1 identify all relevant sources of uncertainty,
2 characterize each source of uncertainty,
3 estimate numerical solution error in the system response quantities of interest,
4 estimate uncertainty in the system response quantities of interest,
5 conduct model updating,
6 conduct sensitivity analysis.

All of these steps, except step 3, are widely practiced in nondeterministic simulations and risk analysis. Step 3 is not commonly addressed for three reasons. First, in many simulations the numerical solution error is assumed to be small compared to other contributors to uncertainty. Sometimes this assumption is quantitatively justified, and sometimes it is simply posited with little or no evidence. Second, in some computationally intensive simulations, it is understood that the numerical solution error is important, and possibly even dominant, but it is argued that various modeling parameters can be adjusted to compensate for the numerical error. If the application of interest is sufficiently similar to the conditions for which experimental data are available, then it is claimed that the adjustable parameters can be used to match the existing data and thereby make reasonable predictions. Third, even if the numerical error is estimated, and it is not small relative to other uncertainties, there are no generally accepted procedures for including its effect on the system response quantities (SRQs) of interest.

It is the beyond the scope of this chapter to deal in depth with each of the steps. Many techniques in predictive capability are well summarized in the following texts (Cullen and Frey, 1999; Melchers, 1999; Modarres *et al.*, 1999; Haldar and Mahadevan, 2000a; Bedford

and Cooke, 2001; Bardossy and Fodor, 2004; Nikolaidis *et al.*, 2005; Ayyub and Klir, 2006; Singpurwalla, 2006; Ang and Tang, 2007; Choi *et al.*, 2007; Kumamoto, 2007; Singh *et al.*, 2007; Suter, 2007; Vinnem, 2007; Vose, 2008; Haimes, 2009; EPA, 2009). Some of these texts should be consulted for a more in-depth understanding of uncertainty quantification (UQ) and risk assessment. Even though the classic text of Morgan and Henrion (1990) is rather dated, we believe it is still one of the most comprehensive discussions of the myriad aspects of UQ and risk assessment. It is highly recommended reading, not only for people new to the field, but also for experienced UQ analysts.

The three dominant approaches to UQ and risk assessment are: traditional probabilistic methods, Bayesian inference, and probability bounds analysis (PBA). As discussed in several earlier chapters, this text concentrates on PBA. Some of the key references in the development and use of PBA are Ferson (1996, 2002); Ferson and Ginzburg (1996); Ferson *et al.* (2003, 2004); Ferson and Hajagos (2004); Kriegler and Held (2005); Aughenbaugh and Paredis (2006); Baudrit and Dubois (2006) and Bernardini and Tonon (2010). PBA is closely related to two other approaches: (a) two-dimensional Monte Carlo sampling, also called nested Monte Carlo, and second-order Monte Carlo (Bogen and Spear, 1987; Helton, 1994, 1997; Hoffman and Hammonds, 1994; Cullen and Frey, 1999; NASA, 2002; Kriegler and Held, 2005; Suter, 2007; Vose, 2008; NRC, 2009), and (b) evidence theory, also called Dempster–Shafer theory (Krause and Clark, 1993; Almond, 1995; Kohlas and Monney, 1995; Klir and Wierman, 1998; Fetz *et al.*, 2000; Kyburg and Teng, 2001; Helton *et al.*, 2004, 2005a; Oberkampf and Helton, 2005; Bae *et al.*, 2006). The PBA approach stresses the following aspects in an analysis: (a) keep aleatory and epistemic uncertainties segregated throughout each step of the analysis; (b) mathematically characterize aleatory uncertainty as probability distributions; (c) characterize epistemic uncertainty as interval-valued quantities, i.e., all values over the range of the interval are possible, but no likelihood is associated with any value; (d) if independence between uncertainty quantities cannot be justified, then dependence should be considered as an epistemic uncertainty; (e) map all input uncertainties through the model; and (f) display SRQs as bounds of probability distributions, i.e., a p-box. A p-box is special type of cumulative distribution function that represents the set of all possible CDFs that fall within the prescribed bounds. As a result, probabilities can be interval-valued quantities as opposed to a single probability. A p-box expresses both epistemic and aleatory uncertainty in a way that does not confound the two.

Returning to the topic of the six steps for prediction, the six steps discussed here are similar to the six phases of computational simulation discussed in Chapter 3, Modeling and computational simulation. The phases in Chapter 3, however, stressed the modeling and computational aspects of simulation. In the six steps discussed here, more emphasis is given to the UQ aspects because we believe that V&V are supporting elements in nondeterministic predictions. It should also be stressed that we assume that before the six steps discussed here are initiated, the goals of the simulation analysis have been clearly identified and agreed upon by those conducting the analysis, as well as those who will use the results of the analysis. As discussed in Chapter 14, Planning and prioritization in modeling and simulation, this is a critical, but difficult, task.

In addition, the following aspects of modeling should be considered and specified *before* the analysis is begun:

- systems and surroundings,
- environments,
- scenarios,
- application domain of the system.

These aspects were discussed in Chapter 2, Fundamental concepts and terminology, Chapter 3, and Chapter 14.

For complex system analyses, there may be multiple possibilities considered for each of these aspects, resulting in multiple sets of simulations, each addressing a particular aspect of the system. For example, in an abnormal environment there may be many failure scenarios identified; however, only some may be analyzed in detail. If multiple environments and scenarios are identified, each one may have an estimated probability of occurrence associated with it. If some type of consequence can be quantified for each possibility identified, then one may only choose to analyze the highest risk possibilities. In the discussion that follows, however, we will not address these probabilities or consequences. For simplicity, we will only consider one set of conditions at a time, i.e., system, surroundings, environments, scenarios, and application domain. In most engineering analyses, multiple sets of conditions must be analyzed.

## 13.1 Step 1: identify all relevant sources of uncertainty

When the above mentioned aspects of formulating a model have been completed, then a process is conducted to identify all the aspects of the model that will be considered as uncertain and those that will be considered as deterministic. For example, in the analysis of the performance of an electrical automatic control system, the electrical properties of many of the components will be considered as uncertain due to manufacturing variability and assembly. Properties such as Planck's constant, the speed of light in a vacuum, and the elementary charge would normally be considered as deterministic. The *goals of the analysis* should be the primary determinant for what is considered as fixed versus what is considered as uncertain. The general philosophy that should be used is: consider an aspect as uncertain unless there is a strong and convincing argument that the uncertainty in the aspect will result in minimal uncertainty in *all* of the system response quantities of interest in the analysis. It should be convincing to the project leader, as well as all of the members of the team conducting the analysis. If the sensitivity analysis conducted in step 6 shows that the contribution to uncertainty from certain aspects is small, then these aspects can be considered as deterministic. If, however, an aspect is considered as deterministic, the model cannot provide any indication of how sensitive the results are to that assumed aspect.

In large-scale analyses, a *screening analysis* or *scoping study* is commonly conducted to obtain a better indication of what aspects may be important and what may be unimportant. A screening analysis is a preliminary modeling and UQ analysis, done with simplified

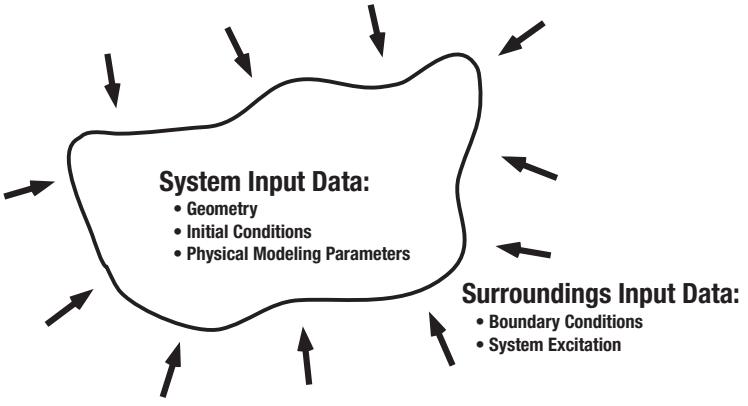


Figure 13.1 System and surroundings input data.

models, to assist in identifying the most important and the least important contributors to uncertainty. The screening analysis is intentionally biased toward being conservative toward the important outcomes of the analysis. That is, a screening analysis attempts to err on the side of identifying all possible changes in modeling and uncertainties that may result in detrimental outcomes of interest. For example, in a complex system there are many subsystems and components that can interact in several ways. Additionally, there would commonly be many different types of physics occurring, along with different types of interactions. A screening analysis attempts to identify what aspects of the system, subsystems, components, physics, and interacting physics should be included in the full analysis and what could safely be excluded. A proper screening analysis can aid analysts and decision makers in directing limited resources toward the more important aspects of the work. This also includes resources devoted to obtaining experimental measurements on needed model inputs.

### *13.1.1 Model inputs*

Model inputs can be divided into two general groups: system input data and surroundings input data. Figure 13.1 depicts these groups, along with the subgroups of each. Quantities in each of the subgroups can be deterministic or uncertain, depending on the needs of the UQ analysis. Each subgroup will be briefly discussed in order to point out what type of uncertainty might need to be considered, and what kinds of difficulty commonly arise.

System geometry data can be specified in a number of ways. For example, it could be specified in detail in a computer aided design (CAD) software package. In addition, computer aided manufacturing (CAM) software may be used for more detail on the actual manufacturing and assembly process, such as final surface preparation, methods and specifications for installing rivets and bolts, welding procedures, installation, and assembly of hydraulic lines, and electrical sensor and wiring installation. If CAD/CAM software is used,

however, the ability to consider uncertainties in the geometry in an automated way is rather very limited. By *automated*, we mean that the user of the CAD/CAM package can specify a subset of the geometry features as uncertain, assign values to that subset, and then have all of the remaining geometry features re-computed. Because of the many ways in which CAD/CAM packages are structured, and because of the multitude of ways the geometry within a package can be built, the ability to automate uncertainties in a system design can be quite problematic. As a result, one should be very cautious in choosing CAD/CAM packages so that they have the flexibility that is needed in specific UQ analyses.

A similar situation occurs when a user constructs a simplified geometry within a commercial software package that is designed for specific types of analysis, e.g., solid mechanics or fluid dynamics. The user usually specifies many of the geometry features as parameters. Then, if they want to consider some of them as uncertain, they must individually input many of the uncertain parameters in the geometry. However, one must be careful so as not to over-specify the geometry or cause inconsistencies in the geometry when certain features are considered as uncertain. As a simple example, suppose one is interested in computing the deflection of a triangular-shaped plate due to a load distribution over the surface of the plate. Because of manufacturing variability, the three internal angles of the plate are considered as continuous random variables. One can only choose two of the angles, because choosing three would over-specify the geometry. This example also points out that there is a correlation structure between the three angles. Correlation of input information will be discussed shortly.

Initial conditions (ICs) provide required information for a model of a system that is formulated as an initial value problem. ICs provide required information concerning: (a) the initial state of all of the dependent variables in the partial differential equations; and (b) the initial state of all other physical modeling parameters, including geometric parameters, that could be dependent on time. As a result, the IC data could be a function of the remaining independent variables in the PDEs. Typically, the most important aspect of the ICs is the state of all of the dependent variables over the domain of the PDE. In addition, the initial state of all of the dependent variables in all of the submodels, e.g., auxiliary PDEs, must be given. If the ICs are considered as uncertain, the uncertainty structure is clearly more complicated than input geometry data because one must deal with functions of one or more of the independent variables.

The most common inputs that are considered uncertain in analyses are data for parameters that appear in the model. There are a number of types of parameters that can occur in models. The following is a useful classification:

- geometry parameters,
- parameters that characterize features of the ICs,
- physical modeling parameters that characterize features of the system,
- parameters that characterize features of the boundary conditions,
- parameters that characterize the excitation of the system due to the surroundings,
- parameters occurring in the mathematical characterization of uncertainties,
- numerical solution parameters associated with the numerical algorithms used.

Depending on its role in the model, a parameter can be a scalar, a scalar field, a vector, or a tensor field. Although we will primarily discuss physical modeling parameters dealing with the system, ICs, and BCs, many of the concepts will apply to the other types of parameters listed.

Surroundings input data consists of two subgroups: boundary conditions and excitation of the system. BCs can be dependent on one or more of the independent variables of the PDEs. These independent variables are typically other spatial dimensions and time, if the problem is formulated as an initial-boundary value problem. For example, in a fluid-structure interaction problem, the boundary condition between the structure and the fluid is a compatibility condition. For the boundary condition of the structure there is a distributed pressure and shear stress loading imposed by the fluid. For the boundary conditions of the fluid, there is no flow through the surface, and the fluid on the boundary must be equal to the local velocity of the boundary. Examples of different types of BCs are Dirichlet, Neumann, Robin, mixed, periodic, and Cauchy. If the BCs are considered as uncertain, the effect on the solution procedure can range from minimal to a situation where the solution procedure must be completely changed. For example, if the uncertainty in the BCs does not cause a change in the coupling of the BCs to the solution procedure for the PDEs, then the uncertainty can usually be treated similarly to parametric uncertainty. For example, one could use a sampling procedure to propagate the effect of the uncertainty in the BCs onto the SRQs. If, on the other hand, the uncertainty in the BCs causes a change in the way the BCs must be coupled to the solution to the PDEs, then more sophisticated procedures must be used. For example, if the uncertainty in the BC deforms the boundary to such a degree that the deformation cannot be considered as small, then one must significantly change the numerical solution procedure, and possibly even the mathematical model, to deal with the uncertainty.

System excitation refers to how the surroundings affect the system, *other than* through the BCs. System excitation always results in a change in the PDEs that are being solved. Sometimes system excitation is referred to as a change in the right hand side of the PDEs to represent the effect of the surroundings on the system. Common examples of system excitation are (a) a force field acting on the system, such as that due to gravity or an electric or magnetic field; and (b) energy deposition distributed through the system, such as by electrical heating, chemical reactions, and ionizing or nonionizing radiation. System excitation uncertainties are usually treated as an uncertain parameter that is a scalar or tensor field. Similar to large uncertainties in BCs, however, if large uncertainties occur in system excitation, then the mathematical model and/or the numerical solution procedure may need to be changed.

### *13.1.2 Model uncertainty*

By *model uncertainty*, we specifically mean uncertainty that is caused by the assumptions embedded in the formulation of the model, as opposed to uncertainty in inputs to the model. As discussed in Chapter 3, formulation of the model occurs in both the conceptual

modeling and mathematical modeling phases. Sometimes model uncertainty is referred to as *model form uncertainty*, and we will use that term when the context is not clear as to what uncertainty we mean. It must be emphasized that model uncertainty is the uncertainty due to the *entire aggregation of all components of the formulation of the structure of the model*, exclusive of model input uncertainty. For example, this would include (a) the specification of the environment of interest, (b) the scenario of interest, (c) physical interactions or couplings that are included or ignored, (d) the PDEs of the primary model, and (e) the PDEs of all submodels that complete to the primary model. Stated differently, model uncertainty includes all assumptions, conceptualizations, abstractions, and mathematical formulations on which the model relies.

Model uncertainty is rarely analyzed in texts on UQ and risk analysis because it is difficult to deal with. It is much more difficult to deal with than input uncertainty, for two reasons. First, model uncertainty is *totally* an epistemic uncertainty, i.e., it is completely due to lack of knowledge as opposed to the inability to know the precise outcome of a random process. Recall from Chapter 2 that epistemic uncertainty was divided into two types: (a) *recognized uncertainty*, an epistemic uncertainty for which a conscious decision has been made to either characterize or deal with it in some way, or to ignore it for practical reasons; and (b) *blind uncertainty*, an epistemic uncertainty for which it is not recognized that the knowledge is incomplete and that the knowledge is relevant to modeling the system of interest. Model uncertainty can be either a recognized or blind epistemic uncertainty. Second, estimating any type of useful bound on model uncertainty is very difficult. These difficulties are, of course, rooted in the fact that model uncertainty necessarily deals with a property of, or choices made by, the modeler or observer. In a UQ analysis, one should not ignore model uncertainty simply because it is difficult to deal with and conceptualize. That would be equivalent to the idiom of ignoring the elephant in the room. In order to achieve reliable predictive capability, model uncertainty must commonly be dealt with, even though it is messy, controversial, and causes a great deal of discomfort. Section 13.2 will discuss some methods for addressing and characterizing model uncertainty.

### 13.1.3 Example problem: heat transfer through a plate

Consider the heat transfer analysis of a system that is coupled to a larger system solely through the boundary conditions. We are interested in simulating the heat transfer through a solid metal plate of size $1 \times 1$ m and thickness 1 cm (Figure 13.2). The SRQ of interest is the total heat flux through the west face of the plate. The key assumptions for modeling the heat transfer through the plate are the following:

- the plate is homogeneous and isotropic;
- the plate is in steady-state condition;
- thermal conductivity of the plate is not a function of temperature;
- heat transfer only occurs in the x-y plane, i.e., there is no heat loss or gain over the surface of the plate in the z-direction.

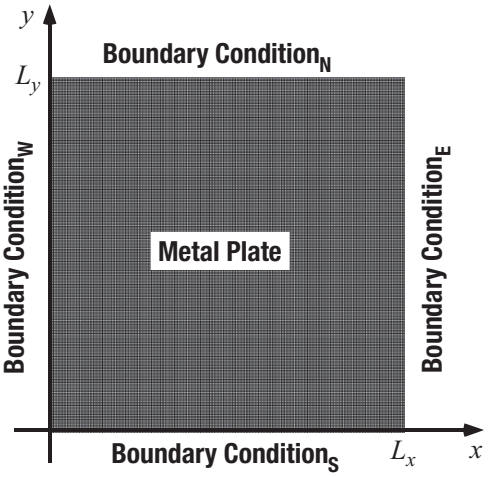


Figure 13.2 System geometry for heat transfer through a metal plate.

The PDE for the temperature distribution through the plate is given by Laplace's equation:

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0. \tag{13.1}$$

As shown in Figure 13.2, boundary conditions are given on the north, south, east, and west boundaries.

The dimensions of the plate, $L_x$, $L_y$, and the thickness, $\tau$, are assumed to be well controlled in manufacturing, so that they are characterized as deterministic quantities. The plate is made of aluminum and the thermal conductivity, $k$, is considered to be uncertain due to manufacturing variability. That is, due to metal composition, forming, and rolling processes there is variability in $k$ from one manufactured plate to the next, i.e., inter-individual variability. The BCs for the east and west faces are $T_{\mathrm{E}} = 450$ K and $T_{\mathrm{W}} = 300$ K, respectively, and they are considered as deterministic. The north face is exposed to air that freely circulates above the top edge of the plate. As a result, the BC on the north face is given as

$$q_{\mathrm{N}}(x) = -k\left(\frac{\partial T}{\partial y}\right)_{y=L_y} = h(T_{y=L_y} - T_{\mathrm{a}}), \tag{13.2}$$

where $h$ is the convective heat transfer coefficient at the surface and $T_{\mathrm{a}} = 300$ K is the ambient air temperature above the plate. $h$ is an empirical coefficient that depends on several factors, such as the air pressure, the speed of air currents above the surface, and whether the plate may possibly have a small amount of moisture on its surface. These conditions are poorly known for the operational conditions of the system, so $h$ is characterized as an epistemic uncertainty and represented as an interval.

The south face of the plate is well insulated so that the BC is given by

$$q_S(x) = -k\left(\frac{\partial T}{\partial y}\right)_{y=0} = 0. \tag{13.3}$$

The model will be used to predict the total heat flux through the west face of the plate, which is given by

$$(q_W)_{\text{total}} = \tau \int_0^{L_y} q_W(y)\,dy, \tag{13.4}$$

where

$$q_W(y) = -k\left(\frac{\partial T}{\partial x}\right)_{x=0}. \tag{13.5}$$

$(q_W)_{\text{total}}$ is of interest because the adjacent system to the west of the system of interest could be damaged due to high heating levels.

To develop confidence in the model, a validation experiment is designed and conducted so that the measurements from the experiment can be compared to predictions from the model. As commonly occurs, the system cannot be tested in available experimental facilities because of its size. So the model will be evaluated using predictions on a scale-model; a plate of size 0.1 m × 0.1 m, but the same thickness, $\tau = 1$ cm. The validation experiment uses the same plate material as the system, and the facility is able to replicate two of the four BCs of the actual system. The BCs on the south and west faces can be duplicated, but the BCs on the east and north faces are modified from the system of interest. Because of facility limitations on heating capability, the maximum temperature that can be achieved in the facility is 390 K. To evaluate the model over a range of temperatures, experiments are conducted at east face temperatures of 330, 360, and 390 K. For each of these $T_E$ conditions, multiple measurements of the SRQ, $(q_W)_{\text{total}}$, are measured in the validation experiment.

For the north face, a different kind of situation exists. The experimentalist, being familiar with the design of validation experiments, realizes that the model accuracy cannot be precisely assessed if significant epistemic uncertainty in the convective heat transfer coefficient, $h$, is allowed to occur in the validation experiment. The experimentalist recommends that in the validation experiment the north face of the plate be provided with a well controlled and carefully measured value of $h$. In consultation with the computational analyst, they agree to set the value of $h$ at the middle of the interval range of $h$ for the system of interest. Table 13.1 summarizes the system and surroundings input data for both the system of interest and the scale model used in the validation experiments.

In addition to the model input data uncertainties, one should also try to identify the potential weaknesses in the modeling, i.e., possible sources of uncertainty in the formulation of the model. As discussed in Chapter 3 and Chapter 12, Model accuracy assessment, identifying and quantifying model form uncertainty is always difficult. The task is made more challenging if the analyst does not have an open mind concerning the various sources of model uncertainty. One procedure is to try and identify some of the assumptions that may be questionable in the modeling. This aids in improved understanding of modeling

Table 13.1 *Model input data for the system of interest and the validation experiment for the heat transfer example.*

| Model input data | System of interest | Validation experiment |
|---|---|---|
| **System input data** | | |
| Geometry, $L_x$ and $L_y$ | $L_x = L_y = 1$ m, deterministic | $L_x = L_y = 0.1$ m, deterministic |
| Geometry, $\tau$ | $\tau = 1$ cm, deterministic | $\tau = 1$ cm, deterministic |
| Thermal conductivity, $k$ | $k$, aleatory uncertainty | $k$, aleatory uncertainty |
| **Surroundings input data** | | |
| BC east face | $T_E = 450$ K, deterministic | $T_E = 330, 360, 390$ K, deterministic |
| BC west face | $T_W = 300$ K, deterministic | $T_W = 300$ K, deterministic |
| BC north face | $h$, epistemic uncertainty | $h$, deterministic |
| | $T_a = 300$ K, deterministic | $T_a = 300$ K, deterministic |
| BC south face | $q_S = 0$, deterministic | $q_S = 0$, deterministic |

uncertainties, not only in the actual system, but also in the validation experiment. The following describes some concerns with modeling assumptions, listed in order of decreasing concern.

- The assumption that thermal conductivity is independent of temperature is fairly well justified for the temperature range and metal considered here. However, it is believed to be the weakest assumption of those listed above in formulating the analysis.
- The assumption of no heat loss or gain over the front and back surfaces of the plate is well justified in the validation experiment because it is a well-controlled and well-characterized environment. In the actual system, however, it is a questionable assumption because of the design, manufacturing, and assembly of the complete system, i.e., the larger system in which the present system operates.
- The assumption of a homogeneous plate, i.e., $k$ equal to a constant throughout the plate, was made for both the full-size system plates and the scale-model plates used in the validation experiments. In the validation experiment, multiple plates are cut from the actual production plates of the system. However, since the validation plates are 100th the size of the system plates, the homogeneity of $k$ in the validation plates may be higher than in the system plates. Stated differently, the validation experiments may not fully test the assumption of homogeneity in the system plates.
- The steady-state heat transfer assumption is very good after the system has been operating for a period of time. During startup of the system, however, the assumption is erroneous. The purpose of the simulation discussed here is to predict the heat flux through the west face of the plate because of possible damage the heating might have on the adjacent system. During startup of the system most of the thermal energy goes into heating the plate, as opposed to being transferred to the adjacent system. As a result, the assumption of steady-state heat transfer will tend to produce a higher heating value into the adjacent system, requiring a design of the adjacent system that is more tolerant of higher temperatures.

In the following sections we will discuss this example in the context of the steps of prediction.

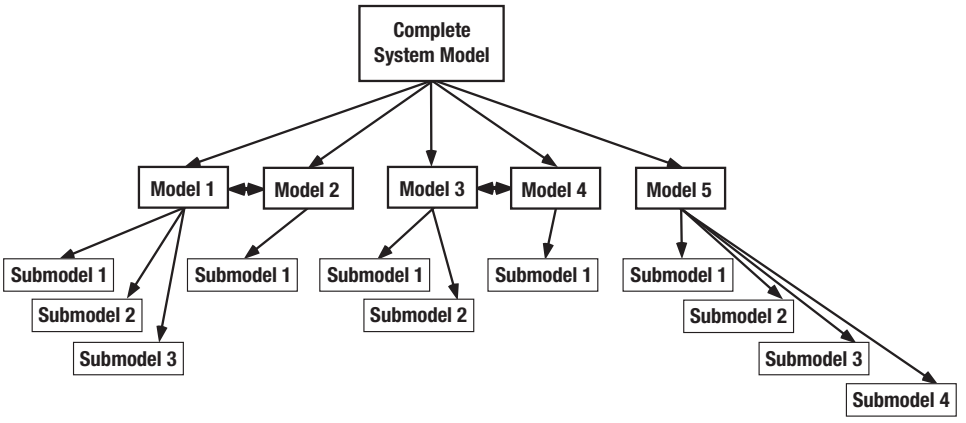


Figure 13.3 Tree structure for models and submodels.

### *13.1.4 Final comments on step 1*

As part of the completion of step 1, a list should be compiled of all of the sources of uncertainty that will be considered in the model. For complex analyses, this list of uncertainty sources could number over a hundred. Some type of logical structure should be developed to help understand where all the sources of uncertainty appear in the analysis. This structure will not only aid the analysts involved in the project, but also the project managers and stakeholders in the analysis. If the project is exposed to an external review panel, it is critically important to devise a method for clearly and quickly displaying what is considered as uncertain, and what is considered as deterministic.

One method for summarizing the model input uncertainties and deterministic quantities is to begin with a tree-structured diagram of models and submodels that make up the complete system model. Figure 13.3 gives an example of a complete system model that is composed of five models, and each model has anywhere from one to four submodels. All five models interact in the complete system model, but only models 1 and 2, and models 3 and 4 directly interact. That is, models 1 and 2, and models 3 and 4 are strongly coupled, whereas the remaining models are only coupled in the complete system model. Once the tree-structured diagram of models is created, then the model input uncertainties and deterministic quantities can be summarized in a table for each model and each submodel. A simplified version of Table 13.1 for the heat transfer example could be used for each model and submodel. This summary information is time consuming to compile, but it is of great value not only to project managers, stakeholders, and external reviewers, but also to analysts because it can uncover inconsistencies and contradictions in a complex analysis.

## 13.2 Step 2: characterize each source of uncertainty

By *characterizing a source of uncertainty* we mean (a) assigning a mathematical structure to the uncertainty and (b) determining the numerical values of all of the needed elements of

the structure. Stated differently, characterizing the uncertainty requires that a mathematical structure is given to the uncertainty and all parameters of the structure are numerically specified, such that the structure represents the state of knowledge of every uncertainty considered. The primary decision to be made concerning the mathematical structure for each source is: should it be represented as a purely aleatory uncertainty, a purely epistemic uncertainty, or a mixture of the two? As discussed in Chapter 2, a purely aleatory uncertainty is one that is completely characterized by inherent randomness, i.e., purely chance. The classic examples are the roll of a die and Brownian motion. Purely epistemic uncertainty is one that is completely characterized by lack of knowledge. Stated differently, if knowledge is added to the characterization of the uncertainty, the uncertainty will decrease. If sufficient knowledge is added, it is conceptually possible that the source will become deterministic, i.e., a number.

At first glance, it may seem rather easy to segregate uncertainties into aleatory or epistemic. In reality, it can be difficult. The difficulty commonly arises because of very different, yet pragmatic reasons. First, the risk assessment community has had a long tradition of *not* separating aleatory and epistemic uncertainties. Only during the last decade or so have a number of leading risk analysts begun to stress the importance of different mathematical representations for aleatory and epistemic uncertainty. See, for example, Morgan and Henrion (1990); Ayyub (1994); Helton (1994); Hoffman and Hammonds (1994); Rowe (1994); Ferson (1996); Ferson and Ginzburg (1996); Frey and Rhodes (1996); Hora (1996); Parry (1996); Paté-Cornell (1996); Rai *et al.* (1996); Helton (1997); Cullen and Frey (1999); and Frank (1999). Second, essentially all of the commercial risk assessment, UQ, and SA software available are completely focused on purely aleatory uncertainty. To deal with this, most of the large risk assessment projects build their own software to separate aleatory and epistemic uncertainty. Medium and small risk assessment projects, however, usually do not have the resources to develop the software tools. And third, a slight change in the perspective or the question that is asked concerning an input uncertainty can change its mathematical structure. For example, if one question were asked concerning the source, it could be characterized as an aleatory uncertainty. If a slightly different question were asked, it could be characterized as an epistemic uncertainty or as a mixed uncertainty. As a result, careful planning must go into what questions should be asked so that they are aligned with the goals of the system analysis. In addition, UQ analysts must be very careful and clear in explaining the question to experts providing opinions, or experimentalists providing empirical data.

Three different examples of an epistemic and aleatory uncertainty will be discussed. First, consider the case of guessing the number of marbles inside of a jar. Suppose that the jar is transparent so that a person could see a relatively large number of the marbles inside the jar. The number of marbles inside the jar is a pure epistemic uncertainty. It is *not* a random number, but a unique number that is simply unknown to the viewer. Depending on the motivation for guessing the right number of marbles, for example some type of wager, we may guess a single number. However, this type of situation is not what engineering is about: *adequacy is the issue, not perfection*. We may guess an interval in which we believe

the actual number may lie, or we may guess an interval with some type of personal belief structure on the interval. For example, we may give a triangular belief structure over the range of the interval. As we study the jar, we may start estimating the number of marbles and possibly make some measurements of the marbles and the jar. With this time and effort, we are improving our knowledge. As a result, we may revise our interval estimate of the true value of the number of marbles. If we spend a significant amount of time, and maybe even modeling of the marbles in the jar, we may significantly reduce the size of our interval estimate. If we empty the jar and count the number of marbles, we have added sufficient knowledge so that the number is exactly known. In engineering, seldom do we know the exact value; we have to make decisions based on imprecise knowledge.

Second, consider the roll of a fair die. Before the die is rolled, the uncertainty in purely aleatory and the probability of each face of the die is 1/6. After the die is rolled, but *before* the die is observed, the uncertainty is purely epistemic. That is, after the die is rolled, there is a fixed outcome, whether we know it or not. In this example it is seen that whether we consider the uncertainty as aleatory or epistemic depends on what question is asked. Are we asking an uncertainty question before the die is rolled, or after the die is rolled? A similar example occurs in risk assessment. Suppose the safety of a certain design nuclear power plant is being analyzed. The question could be: what is the estimated safety of the set of plants of similar design, based on our knowledge as of today? Or, *after* an accident has occurred at one of the plants: what is our estimate of the safety after we have investigated the accident and studied related issues at the other plants? If our estimate of the safety has decreased after a plant accident, then we have either underestimated or underrepresented the safety of the plants before the accident.

Third, consider the case of pseudo-random number generation. Suppose a person observes a long sequence of numbers and asks the question: is this a random sequence of numbers? They may conduct various statistical tests and conclude that the sequence is indeed random, and that the next number is not knowable. Suppose now that the person was provided the algorithm that generated the sequence and the seed that started the sequence. With this knowledge, the person can determine, with perfect confidence, what the next number in the sequence will be. Without this knowledge, the sequence would be characterized as aleatory. With this knowledge, it would become completely deterministic.

Now consider the case of an uncertainty that is a mixture of aleatory and epistemic. This case will be discussed by way of two examples. First, consider the situation of a stranger approaching you in a casino and asking if you would like to place a wager on the roll of a die. You're feeling lucky, so you say Yes. He reaches in his pocket and pulls out a die. He says he will pick a number between 1 and 6, and then you will pick another number between 1 and 6. He will throw the die and whoever's number comes up first, wins the wager. How much do you want to bet? Before you answer, you start considering various ways to estimate the probability of winning or losing. If you were a Bayesian, you would assume a noninformative prior distribution, i.e., assume a uniform probability distribution that any number between 1 and 6 is equally likely.

Being cautious and skeptical, you note that you have essentially *no basis* for assuming a uniform distribution. You have not seen the die, and you have never seen this person before in your life. So you ask if you can see the die. You look at the die, it indeed has six unique faces, and it looks *normal*. With this step, you have added significant knowledge to the decision process. There is now *some* evidence that the uniform distribution is reasonable.

Being really cautious, and gambling with your own money, you seek to add more knowledge before you characterize the uncertainty. You ask if you can roll the die a number of times to see if it appears to be fair. He agrees, and you start rolling the die. Each roll of the die adds knowledge concerning the probability of each number of the die appearing. You continue to roll the die a large number of times and, finally, conclude that the die *is* fair. About this time, the stranger shakes his head in frustration, takes the die, and walks off.

This example shows that when the stranger initially asks you to bet, you can *only* defend a characterization of pure epistemic uncertainty; everything else is *presumption*. As you gather information, the uncertainty becomes a mixture of aleatory and epistemic uncertainty, and at the final stage, it becomes purely aleatory uncertainty. Without this knowledge, you cannot be assured that the stranger is not making a Dutch book against you. (See Leonard and Hsu, 1999; Kyburg and Teng, 2001; and Halpern, 2003 for a discussion of a Dutch book in statistics.)

The second example of mixed aleatory and epistemic uncertainty deals with characterizing an uncertainty based on samples from a population. Suppose you are interested in the variability of the mass of individual manufactured parts. Suppose you have just received the first shipment of parts from a new supplier. The contract with the supplier specifies the metal from which the parts are to be machined, the dimensional tolerance requirements for the part, as well as the material properties, but nothing specifically dealing with mass variability of the part. You have very little knowledge of their manufacturing process, their quality control processes, or their reputation for quality manufacturing. Suppose that all of the parts from the new supplier were dimensionally inspected and they were all found to satisfy the dimensional tolerances in the contract. Before any mass measurements were made of the parts, one could compute the maximum and minimum volume of the part based on the tolerances given for each dimension. A reasonably defensible characterization of the maximum and minimum mass of the part would be to assume a maximum and minimum density of the metal and use these values multiplied by the maximum and minimum volume, respectively. To assign a variability of mass over this range, it would be reasonable to assign a uniform probability distribution over this range. One could argue that there should be less variability of the mass than a uniform distribution, but there is little evidence to support that view.

If some of the parts were also measured for their mass, then one has a good deal more information concerning the variability. To decide what theoretical family of distributions might be used to represent the variability, one could use the PUMA technique, i.e., Pulled oUt of MidAir; then one would compute a best fit for the parameters of the chosen distribution. Or, one could conduct various statistical tests to determine which distribution appears

reasonable to characterize the variability. (A wide variety of commercial software exists for analyses such as this; for example, JMP and STAT from SAS Inc., BestFit from Palisade Inc., Risk Solver from Frontline Systems, Inc., STATISTICA from StatSoft, Inc., and the statistical toolbox in MATLAB from The MathWorks, Inc.). Suppose that a two-parameter log-normal distribution looked sensible, so it is chosen to characterize the variability. Using the samples available, one could estimate the two parameters in the distribution using various methods. If this were done, one would be characterizing the variability as a purely aleatory uncertainty. Although this is common practice, it actually under-represents the true state of knowledge. If the number of mass measurements made is rather small, or the choice of the distribution is not all that convincing, then the strength of the argument for the mass as a purely aleatory uncertainty becomes embarrassing.

A more defensible approach would be to characterize each of the parameters of the log-normal distribution (i.e., the parent distribution) as having probability distributions themselves. A mathematical structure such as this is usually referred to as a *second-order distribution*. The parameters of the second-order distributions are usually referred to as second-order parameters. This mathematical structure directly displays the uncertainty due to sampling, or as it is sometimes referred to, the epistemic uncertainty of the variability. A detailed discussion of how the second order distributions can be calculated is beyond the scope of this book. See Vose (2008) for a more detailed discussion. The second-order distribution is actually a special type of p-box, referred to as a *statistical p-box*. The statistical p-box could be computed by sampling the second-order parameters. For each sample, a cumulative distribution function (CDF) of the parent distribution can be computed. After a number of samples are computed, one generates an ensemble of CDFs. For all samples within the outer envelope of the p-box constructed from sampling, there is a statistical structure within the p-box. One could contrast this structure with the p-box where the parameters are intervals. For the case of interval-valued parameters, there would be *no structure* within the p-box. Both types of p-box have epistemic uncertainty, but the statistical p-box contains structure because of the knowledge of sampling uncertainty, whereas the p-box resulting from intervals contains no knowledge of the inner structure.

In the discussion that follows concerning model input uncertainty, aleatory, epistemic, and mixed structures can be used, depending on whether we are dealing with a random variable or not and the amount of information available. For model uncertainty, only an epistemic uncertainty structure should be used.

### *13.2.1 Model input uncertainty*

Characterizing the uncertainty in model inputs is commonly a major effort in any UQ analysis. For large-scale analyses, uncertainty characterization can take as much time and financial resources as the development of the model and the analysis of the results. Information obtained for the characterization of input quantities can come from one or more of the following three sources:

- experimentally measured data for quantities taken from the actual system or similar systems under relevant conditions;
- theoretically generated data for quantities appearing in the model of the system, but the data come from separate models that provide information to the larger analysis;
- opinions expressed by experts familiar with the system of interest and the models used in the analysis.

In using each of these sources, one is attempting to characterize the uncertainty in an input quantity. However, when any one of these is used, the uncertainty due to the source itself is convolved with the uncertainty in the quantity. Different procedures should be used to minimize the effect of the source uncertainty. For example, it is well known that in experimental measurements there are random measurement uncertainties and there are systematic, or bias, uncertainties. Within a UQ analysis, one usually employs a mixture of the above listed sources. These sources of information will be briefly discussed in this section.

For small-scale analyses of a relatively simple system, the analyst may be able to estimate the uncertainties in all of the input quantities. For most analyses, however, a wide range of expertise is required to gather the needed information. This expertise may have no association with the larger UQ analysis or the organization conducting the analysis. If systems similar to the one of interest have been operational and tested in the past, then significant information can be obtained from these sources. However, this route usually requires searching through old records, digging through data, and finding individuals who are familiar with the data in order to fill in gaps in the information and provide the proper interpretation. In many cases, separate laboratories or organizations are contracted by the organization conducting the larger UQ analysis so that needed data can be generated. The data can be either experimental measurements or theoretical studies using models that are specifically constructed so that their SRQs are the input quantities for the larger UQ analysis of interest.

Eliciting and analyzing expert opinion has received a great deal of attention recently (Cullen and Frey, 1999; Ayyub, 2001; Meyer and Booker, 2001; Vose, 2008). This is due to the recognition of how important it is in UQ analyses, as well as how often it must be done. The references cited list a number of procedures for eliciting, analyzing, and characterizing expert opinion. It is important to recognize that these elicited experts should have two kinds of expertise: substantive expertise on the issue at hand, i.e., in-depth technical knowledge of the issue; and normative expertise, i.e., understanding of the methods of quantification of the uncertainty in the elicited information. The references given discuss a number of pitfalls that can occur in expert elicitation, as well as methods for reducing or eliminating their impact. Two of the primary pitfalls are misinterpretation and misrepresentation of expert data. By *misinterpretation*, we mean that either the expert misinterpreted the question being asked by the elicitor or the elicitor misinterpreted the information provided by the expert. By *misrepresentation*, we mean that the elicitor misrepresents, unintentionally or intentionally, the information from the expert. This most commonly occurs when the elicitor converts the expert information into a mathematical structure that is used as input to the model.

In this regard, risk analysts using PBA have found that the most common difficulty is the lack of understanding of aleatory, epistemic, and mixed uncertainties by the experts. It is the responsibility of the *elicitor* to clearly explain and give a number of examples to the expert before the elicitation process is initiated. After it appears that the expert understands each type of uncertainty, then specific questions dealing with the model inputs can be queried. After the expert provides answers to the questions, it is highly advisable that the elicitor gives back to the expert his/her interpretation of what the expert seemed to have said. Often one finds that there is a miscommunication, primarily because of the subtle differences between aleatory, epistemic, and mixed uncertainties that are not fully grasped by the expert.

Cullen and Frey (1999) stress the importance of both the expert and elicitor understanding what kind of aleatory uncertainty is being captured in the elicitation. Here, we will refer to aleatory uncertainty as simply variability. Cullen and Frey (1999) point out that there are three types of variability: (a) temporal variability, (b) spatial variability, and (c) inter-individual variability. Temporal variability deals with the question of how a quantity varies as a function of the time scale of interest in the analysis. For example, suppose the variability of wind speed at a given location is needed in the analysis. Suppose the analysis requires, because of the assumptions in the model, the distribution of wind speed averaged over the period of a month, individually for each month of the year. All of these aspects should be made clear to the expert so that no confusion or miscommunication occurs. For example, the expert may have never dealt with the wind speed variability over such a long time period.

Spatial variability refers to how a quantity varies in space. If a quantity varies over space, ignoring time dependence for the present, then one must clarify for the expert what type of spatial averaging one is interested in for the model. For example, suppose the model is dealing with the dispersal of a contaminant in the atmosphere. Suppose the finest scale for the discretization of space in the computational model is 1 $m^3$ of air and this occurs near the surface of the Earth. As a result, the finest spatial scale of concentration of the mass of the contaminant is 1 $m^3$ of air. If an expert is questioned about his/her opinion of spatial scales of fluid dynamic turbulence, then it must be clarified that the smallest spatial scale of turbulence that exists in the model is 1 $m^3$.

Inter-individual variability refers to how an outcome can vary over the sample space of all possible outcomes, i.e., the population. Outcomes can be the result of physical measurements, a theoretical model, or a sequence of observations. When a model for the UQ analysis is constructed, a specific definition of a population is defined. For example, one may be interested in the population of all parts manufactured by a supplier during a particular month or a particular year. If this information is not experimentally available, it may be elicited from an expert knowledgeable about similar manufactured parts, but not necessarily from the same supplier. The elicitation process must be very clear to the expert concerning exactly what population is of interest for the model in the UQ analysis.

Depending on the perspective of the elicitor, they may require that the expert provide information in terms of very rigid mathematical structures, e.g., "We will not let the expert

out of the room until they give us a probability distribution for the input." We, of course, do not subscribe to that type of interrogation technique. The expert should not be pressured into providing more information than they feel they can support. The following are examples of the types of mathematical structure that can be provided by an expert for an input quantity. The list is ordered in terms of the least information provided to the most information.

- A single (deterministic) value for the quantity is presumed to exist. The expert only claims that they know this value to within an interval.
- A single (deterministic) value for the quantity is presumed to exist. The expert claims that they know this value over an interval, but they have a higher level of confidence over certain regions of the interval than others. As a result, they can provide a belief structure over the interval.
- The quantity is a continuous random variable. The expert claims that the sample space cannot be less than a certain value and it cannot be greater than a certain value.
- The quantity is a continuous random variable. The expert claims that the probability distribution is a specific theoretical distribution, the sample space cannot be less than a certain value and it cannot be greater than a certain value.
- The quantity is a continuous random variable. The expert claims that the probability distribution is a specific theoretical distribution and that all the parameters in the distribution are known to be within specified intervals.
- The quantity is a continuous random variable. The expert claims that the probability distribution is a specific theoretical distribution and all the parameters are precisely known.

Mathematical characterization of certain types of information can be constructed by some of the software packages mentioned above. Another package that can deal with characterizing information with epistemic uncertainties is CONSTRUCTOR (Ferson *et al.*, 2005).

Whenever there is uncertainty, either aleatory or epistemic, in more than one input quantity, correlation or dependencies commonly exist between the quantities. There are two basic types of dependency, aleatory dependency and epistemic dependency. How to deal with aleatory dependence is fairly well understood. See, for example, Cullen and Frey (1999); Devore (2007); and Vose (2008). How to deal with epistemic or mixed dependence, however, is still is still a research topic. See, for example, Couso *et al.* (2000); Cozman and Walley (2001); and Ferson *et al.* (2004). Deducing dependency between model inputs can be aided by experimental data, theoretical modeling information, or expert opinion. The task of determining dependency increases rapidly as the number of uncertain inputs in the UQ analysis increases. As a result, the most common approach is to assume independence between all inputs and proceed with the UQ analysis. This assumption, although very expedient, can greatly underestimate the uncertainty in the outcomes of the UQ analysis. It is beyond the scope of this summary to deal with issues of characterization of dependency. For a detailed discussion, see the references just mentioned.

### *13.2.2 Model uncertainty*

Here we are interested in characterizing the model form uncertainty, as discussed in Section 13.1.2, by estimating the model uncertainty over the domain where the model

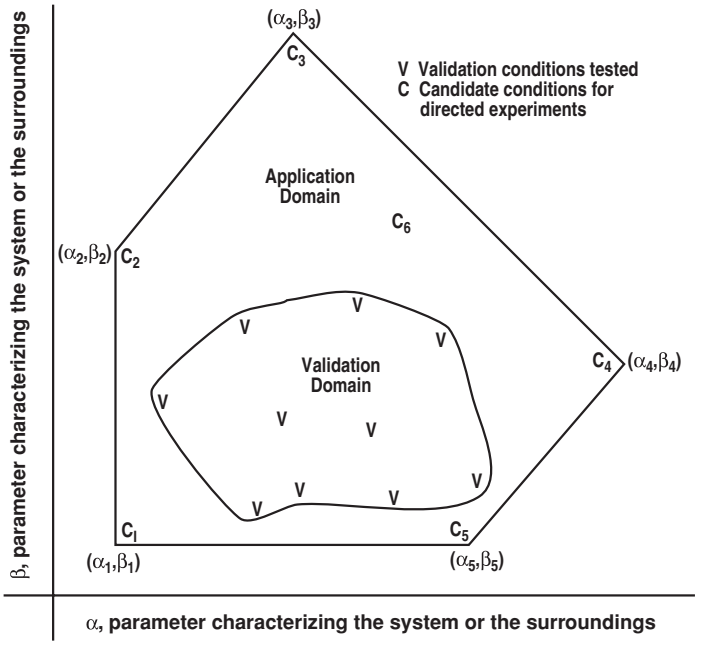


Figure 13.4 Validation domain and application domain in two dimensions (adapted from Trucano *et al.*, 2002).

will be used, i.e., the application domain. We have also referred to model uncertainty as model bias uncertainty, calling attention to the analogy with experimental bias, or systematic, uncertainty. If the application domain is completely enclosed in the validation domain, then model uncertainty can generally be well estimated based on validation metric results, as discussed in Chapter 12. When we compute a validation metric we ask two questions. First, how well do the predictions match the actual measurements that are available for the system? And second, what does the model uncertainty in the predictions tell us about what we should infer about other predictions? That is, when we make a new prediction, it is based on the physics in the model, how well the model has performed in the past, and the conditions for the new prediction. Model uncertainty is based directly on what has been observed in (preferably blind) *prediction performance* of the model. If any portion of the application domain is outside the validation domain, then some type of extrapolation procedure must be used for the estimation of model uncertainty. In practical engineering applications, some degree of extrapolation is commonly required.

Figure 13.4, from Chapter 10, Model validation fundamentals, captures the essence of the concept of interpolation and extrapolation of the model in two dimensions. Recall that $\alpha$ and $\beta$ are parameters characterizing conditions of the system or the surroundings. The Vs denotes conditions where experimental data have been obtained and $(\alpha_i, \beta_i)$, $i = 1, 2, \ldots, 5$, denote the corners of the application domain for the engineering system, sometimes called the operating envelope of the system. A validation metric result can

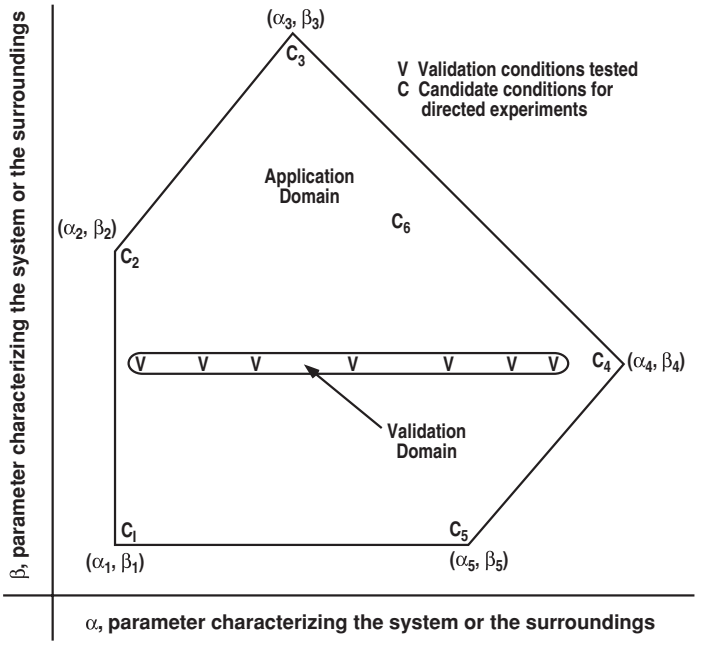


Figure 13.5 Example of a validation domain in only one dimension of the parameter space.

be computed at each of the Vs in Figure 13.4. One can imagine a surface above the $\alpha - \beta$ plane representing the estimated uncertainty in the model over the validation domain.

In Chapter 12, two approaches were discussed in detail for computing validation metrics: the confidence interval approach, and the method of comparing CDFs from the model and the experiment. The first approach estimated the model uncertainty in the mean of the SRQ of interest, and the CDF approach estimated the evidence for model mismatch of the SRQ. To estimate the model uncertainty over the validation domain, either an interpolation function or regression fit of the estimated uncertainty at each of the Vs can be computed. Whether one uses an interpolation function or a regression fit, one must include in the representation the scatter in the experimental data and the aleatory and epistemic uncertainty that may exist in the model prediction. If one uses a regression fit, one must also include the uncertainty due to the lack of fit because of the choice of the regression function.

More commonly, one of the following situations occurs: (a) the data are sparse over the validation domain, (b) the dimensionality of the parameter space characterizing the system or surroundings is very large and data are available only for a few of these dimensions, and (c) all of the data are for a fixed value of one of the dimensions of the parameter space. An example of this last case is shown in Figure 13.5. For the case where there is sparse data, one should use a low order polynomial regression fit of the model uncertainty estimates (the Vs) in the dimensions where data are available. A regression fit using a first- or second-degree polynomial would probably not capture all of the features of the model uncertainty over the validation domain, but it would be a much more robust and reliable estimate

than the vagaries of an interpolation function. Robustness of the estimation of the model uncertainty is especially important if extrapolation of the uncertainty is required outside of the validation domain. For the case where data are available only along certain dimensions of the parameter space, one is forced to extrapolate the model uncertainty in all of the remaining dimensions. For example, in Figure 13.5, model uncertainty in the $\beta$ dimension must be estimated either by extrapolation or by the use of alternative plausible models. For the case shown in the figure, the extrapolation is so weak that it would only support a model uncertainty function that does not change in the $\beta$ direction. Both the extrapolation approach and the alternative models approach will be discussed in more detail in Section 13.4.2.

When the application domain is outside the validation domain, extrapolation must deal with two issues. First, there is extrapolation of the model itself, in the sense that the model is used to make a prediction in terms of the input data and parameters characterizing the system or surroundings. For physics-based models, this extrapolation can be viewed as constrained by the equations for conservation of mass, momentum, and energy, and any other physics-based principles embedded in the model or submodels. For nonphysics-based models, e.g., purely regression fits of data, extrapolation would be foolhardy. Second, one must also extrapolate the model uncertainty that has been observed over the validation domain. Extrapolating model uncertainty is a complex theoretical issue because it is extrapolating the error structure of a model, combined with the uncertainty in the experimental data, in a high dimensional space. However, it is *not as risky*, in our view, as extrapolating a regression fit of the measured SRQs themselves without the benefit of physics. For a system exposed to abnormal or hostile environments, the concept of interpolation or extrapolation of a validation metric result is questionable because of the complexity of the environment. These environments are usually not well characterized by parameters defining a validation domain because there are commonly strong interactions between subsystems, poorly known system geometries or surroundings, and strongly coupled physics.

### *13.2.3 Example problem: heat transfer through a solid*

Here, we continue with the development of the heat transfer example begun in Section 13.1.3.

#### *13.2.3.1 Model input uncertainty*

Recalling Table 13.1, there are two uncertain model input parameters in the heat transfer example, $k$ and $h$. The characterization of the uncertainty in $k$ is based on experimental measurements conducted on small samples of aluminum cut from the actual plates used in the system. Samples were cut from multiple locations on multiple plates so that the measured variability in $k$ is representative of both causal factors. The location of the samples was drawn randomly over the area of the plate, and the plates were drawn randomly from multiple production lots of plates. A total of 20 samples were cut from the various plates. Since there was a concern about the dependence of $k$ on temperature, $k$ was measured for

Table 13.2 *Experimentally measured values of k for the heat transfer example (W/m-K).*

| Sample no. | $T = 300$ K | $T = 400$ K | $T = 500$ K |
|---|---|---|---|
| 1 | 159.3 | 164.8 | 187.8 |
| 2 | 145.0 | 168.0 | 180.1 |
| 3 | 164.2 | 170.3 | 196.1 |
| 4 | 169.6 | 183.5 | 182.1 |
| 5 | 150.8 | 165.2 | 186.4 |
| 6 | 170.2 | 183.6 | 200.4 |
| 7 | 172.2 | 182.0 | 199.6 |
| 8 | 151.8 | 170.2 | 192.7 |
| 9 | 154.4 | 165.8 | 191.8 |
| 10 | 163.7 | 175.6 | 194.3 |
| 11 | 157.1 | 169.0 | 191.9 |
| 12 | 167.1 | 181.7 | 192.4 |
| 13 | 161.1 | 174.6 | 185.2 |
| 14 | 174.5 | 194.4 | 199.4 |
| 15 | 165.8 | 177.3 | 181.9 |
| 16 | 163.9 | 172.4 | 189.4 |
| 17 | 171.3 | 182.2 | 195.9 |
| 18 | 154.3 | 170.3 | 191.9 |
| 19 | 159.4 | 174.4 | 196.7 |
| 20 | 155.1 | 170.6 | 184.1 |

each sample at three temperatures; 300, 400, and 500 K. All of the measured values of $k$ are shown in Table 13.2.

Figure 13.6 plots the measured $k$ values for each of the 20 samples as a function of temperature. Scatter in the measurements is due not only to manufacturing variability, but also experimental measurement uncertainty. Here, we do not attempt to separate the two sources of uncertainty, although this could be done statistically using design of experiments (DOE) techniques (Montgomery, 2000; Box *et al.*, 2005) discussed in Chapter 11, Design and execution of validation experiments. Also shown in Figure 13.6 is a linear regression fit of the data using the method of least squares. Although the assumption has been made that $k$ is independent of temperature, the data show an 18% increase in the mean value of $k$ over the range of measured temperatures. The effect of this temperature dependence on the SRQ of interest should be detected in the validation experiments. The regression fit of $k$ and the residual scatter is given by

$$k \sim 116.8 + 0.1473T + N(0, 7.36) \quad \text{W/m-K}. \tag{13.6}$$

The normal distribution indicates the residual standard error from the regression analysis results in a mean of zero and a standard deviation $\sigma = 7.36$. $\sigma$ quantifies the vertical scatter

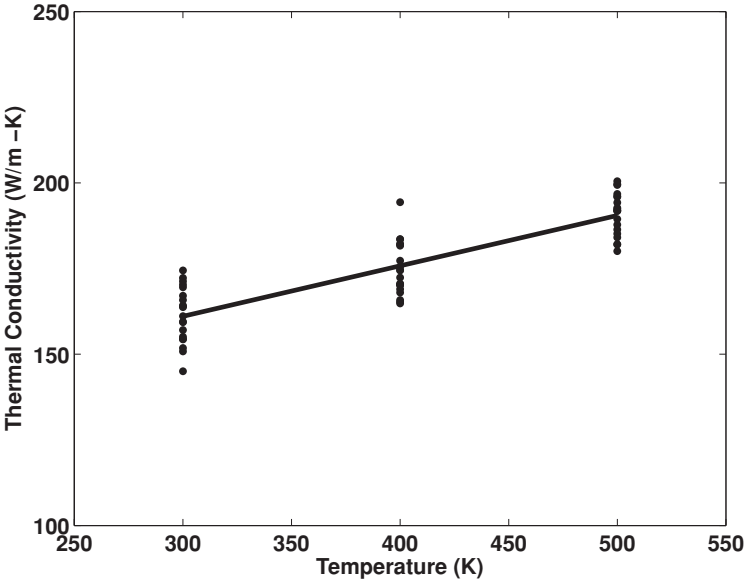


Figure 13.6 Measurements of $k$ for 20 plate samples as a function of temperature.

of $k$ at a given value of temperature. The $R^2$ value of the regression fit is 0.735. $R^2$ is referred to as the square of the multiple correlation coefficient, or the coefficient of multiple determination (Draper and Smith, 1998). $R^2$ is interpreted as the proportion of the observed variation in $k$ that can be represented by the regression model.

Concerning various sources of uncertainty in $k$, one could ask the question: for a given temperature, how much of the variability in $k$ is due to the variability from plate to plate, versus how much is due to location on the plate? If the information on each of the samples was recorded as to which plate it was cut from and where on each plate it was cut, then the question could be answered using design of experiment (DOE). As is common, however, the analyst may have only thought about this question after the experiments were conducted and after he starts to think about what could be the sources of uncertainty. In any type of experimental measurement program, it is not uncommon for the experimentalist to tell the analyst who requests measurements: "If you had only told me to measure it, I could have easily done it." We make this comment to stress Guideline 1 discussed in Section 12.1.1: *A validation experiment should be jointly designed by experimentalists, model developers, code developers, and code users working closely together throughout the program, from inception to documentation, with complete candor about the strengths and weaknesses of each approach.*

Figure 13.7 shows the empirical distribution function (EDF) (solid line) for all of the measurements of $k$. To characterize the possibility of more extreme values that were not seen among the limited samples, and to obtain a continuous approximation of the EDF for a large number of samples, it is common practice to fit a distribution to the data. We

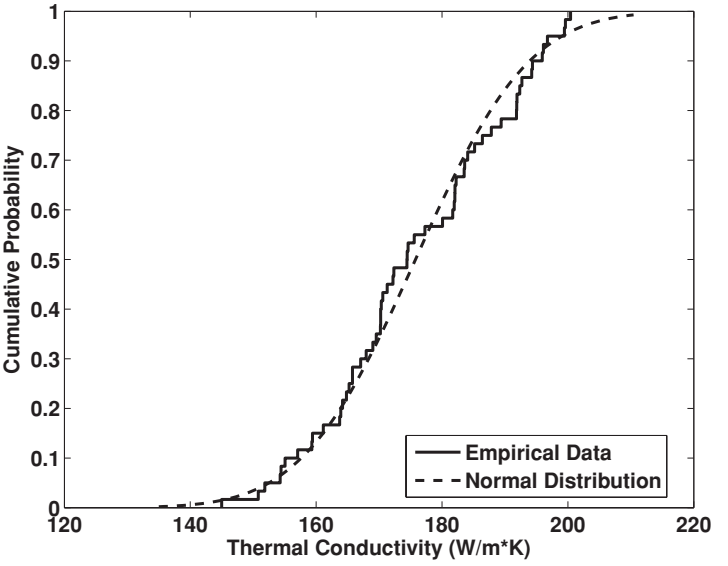


Figure 13.7 Empirical distribution function for $k$ and the normal distribution obtained by the method of matching moments.

used a normal distribution to characterize the variability of $k$, such that the distribution had the same mean and standard deviation as the data set, according to the method of matching moments (Morgan and Henrion, 1990). It is easily computed that the sample mean, $\bar{k}$, is 175.8 W/m-K and the sample deviation is 14.15. Therefore, the computed normal distribution is given by

$$k \sim N(175.8, 14.15) \quad \text{W/m-K}. \tag{13.7}$$

The normal distribution is shown in Figure 13.7 as the dashed line.

A discrete representation of the PDE describing the heat transfer, the boundary conditions, and the heat flux through the west face will now be considered. Using a second order, central difference scheme for the PDE, Eq. (13.1), one obtains

$$\frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{\Delta x^2} + \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{\Delta y^2} = 0, \quad i = 1, 2, \ldots, i_{\max} \quad \text{and} \quad j = 1, 2, \ldots, j_{\max}, \tag{13.8}$$

where $i_{\max}$ and $j_{\max}$ are the number of mesh points in the $x$ and $y$ directions, respectively.

Using a second-order, one-sided, finite-difference approximation for the BC on the north face, Eq. (13.2), one obtains

$$-k\left(\frac{3T_{i,j_{\max}} - 4T_{i,j_{\max}-1} + T_{i,j_{\max}-2}}{2\Delta y}\right) = h\left(T_{i,j_{\max}} - T_{\mathrm{a}}\right). \tag{13.9}$$

Solving for the boundary temperature, $T_{i,j_{\max}}$, one obtains the BC for the north face:

$$T_{i,j_{\max}} = \frac{(2\Delta y h/k)T_a + 4T_{i,j_{\max}-1} - T_{i,j_{\max}-2}}{3 + (2\Delta y h/k)}. \tag{13.10}$$

$h$ is considered a fixed but unknown quantity over the north face of the plate for a given operating condition. However, the system can operate under many different conditions that can alter the value of $h$. Because of the variety of poorly known conditions, $h$ is characterized as an interval-valued quantity. Both fluid dynamics simulations and expert opinion based on operating experience of similar systems are used to determine the interval

$$h = [150, 250] \text{ W/m}^2\text{-K}. \tag{13.11}$$

For the validation experiments, the airflow over the north face is adjusted and calibrated to yield a value of $h = 200$ W/m$^2$-K.

Using a second-order, one-sided, finite difference approximation for the BC on the south face, Eq. (13.3), we have

$$-k\left(\frac{-3T_{i,1} + 4T_{i,2} - T_{i,3}}{2\Delta y}\right) = 0. \tag{13.12}$$

Solving this equation for the boundary temperature, $T_{i,1}$, we have the BC for the south face

$$T_{i,1} = \frac{4}{3}T_{i,2} - \frac{1}{3}T_{i,3}. \tag{13.13}$$

Using a second-order, one-sided, finite-difference approximation for the local heat flux through the west face, Eq. (13.5), we have

$$q_W(y) = -k\left(\frac{-3T_{1,j} + 4T_{2,j} - T_{3,j}}{2\Delta x}\right) + O(\Delta x^2). \tag{13.14}$$

$q_W(y)$ can be directly evaluated once the iterative solution for $T_{i,j}$ has converged. The mid-point rule of integration can be used for the total heat flux through the west face, Eq. (13.4), resulting in

$$(q_W)_{\text{total}} = \frac{\tau}{2}\sum_{j=1}^{j_{\max}-1}\left[(q_W)_j + (q_W)_{j+1}\right]\Delta y. \tag{13.15}$$

The appropriate BCs for each of the four surfaces are coupled to the solution for the finite difference equation for the interior mesh points, Eq. (13.8). Solving Eq. (13.8) for the interior mesh point $T_{i,j}$, one has

$$T_{i,j} = \frac{T_{i+1,j} + T_{i-1,j} + (\Delta x^2/\Delta y^2)T_{i,j+1} + (\Delta x^2/\Delta y^2)T_{i,j-1}}{2 + (\Delta x^2/\Delta y^2)}. \tag{13.16}$$

This equation, coupled with the appropriate finite difference equations for the BCs, is solved iteratively. We use the Gauss–Seidel method discussed in Chapter 7, Solution verification. In Section 13.3 we discuss a method for estimating the discretization and iterative errors.

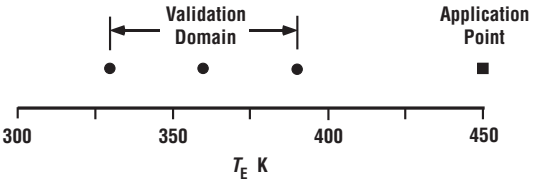


Figure 13.8 Validation domain and application point for the heat transfer example.

### *13.2.3.2 Model uncertainty*

Examining Table 13.1, we can characterize the space of system and surroundings parameters as an eight-dimensional space; $L_x$, $L_y$, $\tau$, $k$, $T_E$, $T_W$, $h$, and $q_S$. For this example, there is only one parameter, $T_E$, which is actually varied. All of the other parameters are either deterministic or an uncertain parameter. As a result, we have a 1-dimensional application and validation domain. This 1-dimensional space is depicted in Figure 13.8. Validation data are obtained at the three conditions shown in Figure 13.8, 330, 360, and 390 K, and a validation metric result is calculated for these conditions. The validation metric result, however, will need to be extrapolated to the application point of 450 K. Since we observed that $k$ increases slightly with temperature, there will clearly be increased uncertainty in the prediction due to the extrapolation of the validation metric results.

As discussed in Section 13.1.3, the validation experiment used aluminum plates of size $0.1 \times 0.1$ m that were cut from the system plates. Similar to the material samples for measuring $k$, validation plates were cut from multiple locations on system plates, from multiple plates, and from multiple production lots. All of these samples were drawn randomly from their respective populations. As a result, the variability of $k$ in the validation experiment plates should be similar to the variability in $k$ shown above in Figure 13.6 and Figure 13.7.

Four sets of independent validation experiments were conducted. One experimental setup was used for each of the three temperatures tested, $T_E = 330$, 360, and 390 K. That is, for a given experimental setup, one measurement of $q_W$ was made at each of the three temperatures. Then the experimental setup was disassembled, old diagnostic sensors were removed, new sensors installed, and the experimental setup was reassembled with a new plate. This procedure takes advantage of DOE principles to reduce systematic (bias) uncertainties in the experimental measurements. For all of the validation experiments the convective heat transfer coefficient on the north face was controlled so that $h = 200$ W/m$^2$-K.

Table 13.3 gives the experimental data for $q_W$ from the four sets of validation experiments. The negative sign for the heat flux indicates that the heat transfer is in the $-x$ direction, i.e., out of the west face of the plate. Just as with measurements of $k$, the experimental uncertainty in $q_W$ includes both the experimental measurement uncertainty and the variability in $k$.

To characterize model uncertainty, we use the confidence interval approach that is based on comparing the mean response of the model and the mean of the measurements, as discussed in Sections 13.4 through 13.7. Using the experimental data in Table 13.3 and the appropriate equations from Section 13.6.1, we obtain

Table 13.3 *Experimental measurements of $q_W$ from the validation experiments (W).*

| Experimental setup | $T_E = 330$ K | $T_E = 360$ K | $T_E = 390$ K |
|---|---|---|---|
| 1 | −41.59 | −100.35 | −149.71 |
| 2 | −49.65 | −95.85 | −159.96 |
| 3 | −55.06 | −99.45 | −153.68 |
| 4 | −49.36 | −104.54 | −155.44 |

number of samples $= n = 12$,
$y$ intercept from the linear curve fit $= \theta_1 = 533.5$,
slope from the linear curve fit $= \theta_2 = -1.763$,
standard deviation of the residuals of the curve fit $= s = 4.391$,
coefficient of determination $= R^2 = 0.991$,
0.9 quantile of the $F$ probability distribution for $\nu_1 = 2$ and $\nu_2 = 10$, yields $F(2,10,0.9) = 2.925$,
mean of the $x$ input values $= \bar{x} = 360$,
variance of the $x$ input values $= s_x^2 = 654.5$.

To keep this analysis simpler, we have made an approximation that should be pointed out. This confidence interval analysis assumes that the 12 experimental samples are independent samples. However, as described above in the description of the validation experiments, there were four independent experiments and each experiment made three measurements of $q_W$. As a result, this analysis could under-represent the uncertainty in the experimental measurements.

The linear regression equation for the experimental data is

$$[\bar{q}_W(T_E)]_{exp} = 533.5 - 1.763T_E, \tag{13.17}$$

and the Scheffeé confidence intervals are

$$\text{SCI}\ (x) = \pm 3.066\sqrt{1 + 0.001667(T_E\ - 360)^2}. \tag{13.18}$$

This validation metric approach only uses the mean of the model result to compare with the estimated mean of the experimental data. We could take the common approach of simply computing one solution for each of the three temperatures used in the experiment. Each of these solutions would use the sample mean of $k$ determined from the material characterization experiments, $\bar{k} = 175\ .8$ W/m-K, to compute the three values of $q_W$. As discussed in Section 13.4.1, this approach is not recommended because the mean of an uncertain input quantity maps to the mean of the SRQ *only* for the case when the model is *linear* in the uncertain quantity. This situation rarely occurs, even for linear PDEs. If the coefficient of variation (COV) of the important input random variables is small and the model is not extremely nonlinear with respect to these random variables, then the

Table 13.4 *Values of $\bar{q}_W$ computed from Monte Carlo sampling for the validation temperatures.*

| $T_E$, (K) | $\bar{q}_W$, (W/m$^2$) |
|---|---|
| 330 | −51.59 |
| 360 | −103.93 |
| 390 | −155.54 |

approximation of using the mean value of all of the uncertain input variables can give reasonably accurate results. The COV is defined as $\sigma/\mu$, where $\sigma$ and $\mu$ are the standard deviation and mean of the random variable, respectively.

Even though for the present example the $(\text{COV})_k = 0.08$, we will still use Monte Carlo sampling to propagate the distribution of $k$ through the model to obtain the mean of $q_W$, $\bar{q}_W$, for each value of $T_E$. Each distribution of $q_W$ was calculated by computing 1000 Monte Carlo simulations using the variability of $k$ given by $N(175.8, 14.15)$ W/m-K. (A more detailed discussion of Monte Carlo sampling will be given in Section 13.4.1.) With these three distributions for the SRQ, we have a great deal more information about the predicted uncertain response of the system. As a result, we could also use the validation metric approach based on comparing CDFs, but we reserve using this metric for the example problem discussed later in this chapter. Table 13.4 gives $\bar{q}_W$ computed from each set of 1000 Monte Carlo simulations for the three temperatures used in the validation experiment.

Figure 13.9 shows the linear regression of the experimental data, the Scheffé confidence intervals, and the $\bar{q}_W$ resulting from the model. The quadratic interpolation of the model is given by

$$(\bar{q}_W)_{\text{model}} = 572.3 - 2.025\ T_E + 0.0004056\ T_E^2. \tag{13.19}$$

Over this range in temperature the model is nearly linear. It is seen that the model slightly over-predicts (in absolute value) the measured heat flux over the temperature range measured. One may not expect this over-prediction because Figure 13.6 shows that the thermal conductivity increases with temperature. The explanation for the over-prediction is that the CDF of $k$ (Figure 13.7) is derived from the *entire* set of material characterization data, resulting in a shift of $\bar{k}$ to a higher value that would be appropriate for higher temperatures. As the temperature increases, it can be seen in Figure 13.9 that the model gives a more accurate prediction of the data.

Figure 13.10 shows the estimated model error and the Scheffé confidence intervals of the experimental data as a function of temperature. This plot is much clearer than Figure 13.9 because we concentrate on the *difference* between the model and the mean of the experimental data; not on the magnitude of the SRQ itself. Since the mean of the

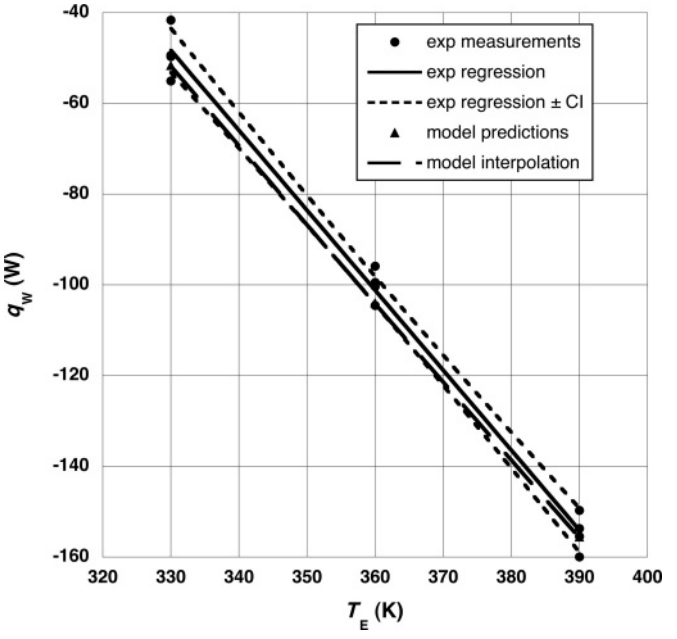


Figure 13.9 Comparison of experimental measurements and model predictions for heat flux over the range of validation data.

experimental data is expressed as a linear function, and the model prediction is nearly linear, it can be seen in Figure 13.10 that the estimated model error is nearly linear. Using a linear regression function to represent the experimental data, the Scheffé confidence intervals, Eq. (13.18), are seen to be symmetric hyperbolic functions. The confidence intervals shown represent the extent of the true mean of the experimental data for a confidence level of 90%, given the experimental data that has been observed. This shows that the estimated model error falls within the ±90% confidence intervals of the data.

Even though we are primarily interested in the model error, we must also include the uncertainty in the estimate. For the confidence interval approach to validation metrics, the uncertainty is *only* representative of the experimental uncertainty, i.e., it does not take into account any uncertainty in the model prediction. The validation metric function $d$ is the characterization of the estimated model uncertainty and is written as

$$d(T_E) = [\bar{q}_W(T_E)]_{model} - \left\{[\bar{q}_W(T_E)]_{exp} \pm SCI(T_E)\right\}. \tag{13.20}$$

Into this equation we substitute the linear fit of the experimental data, Eq. (13.17); the Scheffé confidence interval, Eq. (13.18); and the quadratic interpolation of the model prediction, Eq. (13.19), to obtain the final expression

$$d(T_E) = 38.8 - 0.262\,T_E + 0.0004056\,T_E^2 \mp 3.0066\sqrt{1 + 0.001667(T_E - 360)^2}. \tag{13.21}$$

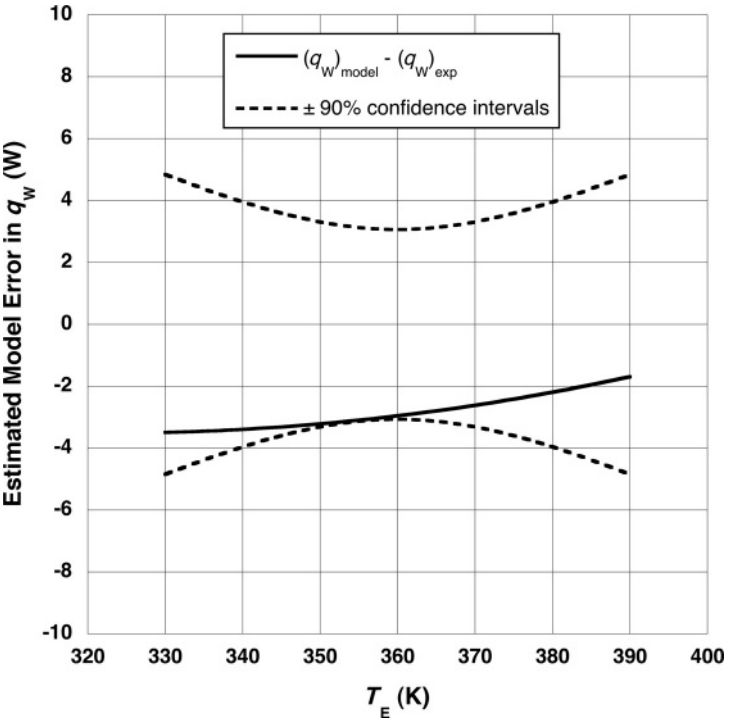


Figure 13.10 Estimated model error and the Scheffé confidence intervals for a 90% level of confidence over the validation domain.

Figure 13.11 shows a graph of Eq. (13.21) along with the estimated model error over the validation domain. This figure contains essentially the same information as Figure 13.10, but with Figure 13.11 certain interpretations are clearer. First, even though the model error is estimated to be small relative to the magnitude of the $q_W$, the uncertainty in the estimate is noticeably larger due to the uncertainty in the experimental measurements. For example, at $T_E = 330$ K, the estimated model error is only 3.5 W, but it may be as small as zero or as large as 8.3 W with a confidence level of 90%. Second, because the confidence intervals are a hyperbola, the uncertainty in the estimated model prediction will grow rapidly when the model is extrapolated beyond the validation domain.

## 13.3 Step 3: estimate numerical solution error

This step involves the estimation of two key sources of error in the numerical solution of PDEs: iterative and discretization error. As discussed in several earlier chapters, there are a number of other error sources that can occur in the numerical solution of PDEs, for example: (a) defective numerical solution algorithms, (b) data input

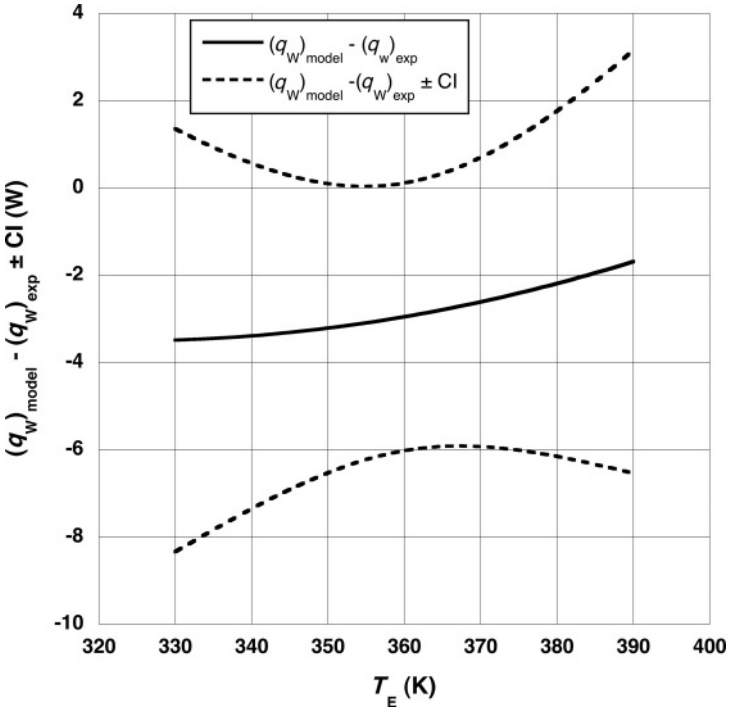


Figure 13.11 Validation metric function for $q_W$ over the validation domain.

mistakes, (c) computer programming errors, (d) computer round-off errors, and (e) data post-processing mistakes. In this step, however, we will not deal with these sources.

It is stressed here that the estimation of iterative and discretization error must be conducted *on the SRQs of interest* in the analysis. It has been pointed out in various contexts that the sensitivity of SRQs to iterative and discretization error can vary drastically from one SRQ to another. Stated differently, for a fixed number of iterations and mesh points, some SRQs can be numerically converged orders of magnitude more (or less) than other SRQs. As discussed in Chapter 10, the discretization convergence rate is commonly related to the order of the derivatives and integrals of the dependent variables in the PDE. That is, the higher the order of the spatial or temporal derivative of one SRQ as compared to another SRQ, the slower the convergence rate. For example, in fluid dynamics the local shear stress at a point on a surface converges more slowly than the total shear stress integrated over an entire surface. As a result, the convergence of all of the SRQs of interest should be evaluated or checked in some way. If certain SRQs are known to be less sensitive to iterative and discretization error than others, and if the low sensitivity is observed over the entire domain of all conditions of interest in the analysis, then one can be more confident in the higher level of convergence of these SRQs.

### 13.3.1 Iterative error

#### 13.3.1.1 Iterative methods

As discussed in Chapter 7, Solution verification, iterative convergence error arises due to incomplete convergence of the discrete equations. Iterative methods are usually required for the solution of nonlinear discrete equations (except for explicit marching methods), as well as for large linear systems. Various types of iterative error estimation technique were discussed for both stationary iterative methods and nonstationary (Krylov subspace) methods. Stationary methods tend to be more robust in applications compared to nonstationary methods. However, stationary methods tend to produce lower convergence rates than nonstationary methods.

Iterative methods are commonly applied in two different types of numerical solution: initial value problems (IVPs) and boundary value problems (BVPs). For IVPs where implicit methods are used, very few iterations are typically required for each step in the marching direction. This is because the initial guess for the iterative solution is based on the previously converged solution, and that the solution does not change much in the marching direction. For BVPs, typically a larger number of iterations are required to obtain a converged solution. As a result, for BVPs convergence characteristics should be closely monitored to determine if one has monotone, oscillatory, or mixed convergence characteristics. The most convenient method of tracking the iterative convergence is to compute the $L_2$ norm of the residuals of each of the PDEs as a function of iteration number. The most reliable method, however, is to estimate the iterative error as a function of the number of iterations. Depending on whether one has monotone, oscillatory, or mixed convergence, this can be cumbersome.

As discussed in Chapter 7, the iterative error should be driven at least as small as 1% of the discretization error when extrapolation-based discretization error estimators are used. However, it is advisable to reduce the error even smaller, e.g., 0.1%, because complex interactions can occur between iterative and discretization error. For example, if the iterative error is not driven small enough, it can result in a misleading or confusing trend in observed order of accuracy as the mesh resolution is changed. An alternative to estimating the iterative error is to monitor the discrete residuals, i.e., the nonzero remainder that occurs when the current iterative solution is substituted into the discrete equations. However, relating residual convergence to iterative error must be done heuristically for a given class of problems. As pointed out above, different SRQs typically converge at widely differing rates, so one should monitor the most slowly converging SRQs of interest. Monitoring the change between successive iterates is *not* reliable and should always be avoided.

#### 13.3.1.2 Practical difficulties

To stress the variety of SRQs that can exist in an analysis, let the set of all SRQs of interest be written as a vector array $\vec{y}$. Let $n$ be the number of elements in the array, each one being

an SRQ of interest, so that one has

$$\vec{y} = \{y_1, y_2, \ldots, y_n\}. \tag{13.22}$$

Some of the elements of the array are typically the dependent variables in the PDEs. However, some of the array elements may be very different kinds of quantities, for example: (a) derivatives of dependent variables with respect to the independent variables, or derivatives with respect to input parameters to the model; (b) functionals such as integrals, time-averages, or min and max operators on the dependent variables; (c) cumulative distribution functions or a specific value of a CDF; and (d) indicator functions, for example, if the system is in a certain state, then a SRQ has the binary value of zero. If the system is in another state, the SRQ has a binary value of one.

Iterative convergence is normally thought of as dealing with the convergence of dependent variables of the PDEs. For example, one commonly monitors the convergence of a norm of the dependent variables over the domain of the PDE. However, one can also think of the mapping of the dependent variables to the elements of $\vec{y}$. Then one could ask the question: how do we quantify the iterative convergence of each of the elements of $\vec{y}$? If the quantity can be computed using the present solution that is being calculated, then one can monitor the iterative convergence of the quantity. For example, suppose one element of $\vec{y}$ is an integral of some dependent variable over the domain of the PDE. Then this quantity can be computed at each iteration and its convergence monitored. However, if the quantity *cannot* be computed from simply knowing the present solution, then iterative convergence of the quantity can be a difficult, or impossible, to monitor. For example, suppose one of the elements of $\vec{y}$ is a CDF. As will be discussed in Section 13.4, there are a number of methods for propagating input uncertainties to obtain uncertainties in $\vec{y}$. The most common method of propagating the uncertainties is to use sampling. When this is done, there will be an ensemble of hundreds or thousands of $\vec{y}$s, one for each sample of the uncertain input. As a result, the CDF cannot be computed until *after* an ensemble of solutions to the PDEs has been computed. For this latter case, one usually relies on stringent iterative convergence criteria for the dependent variables that are being monitored.

It should be stressed that the iterative convergence characteristics, as well as the mesh convergence characteristics to be discussed shortly, of the elements of $\vec{y}$ can change drastically over the sample space of the uncertain input quantities. Since the number of uncertain input quantities can be quite large for a complex analysis, this presents an additional complexity for monitoring iterative convergence. In addition, for unsteady simulations and hyperbolic PDEs, convergence rates of SRQs, particularly local SRQs, can change significantly over time and space. A good understanding of the dominant physical effects on the SRQs can greatly help in identifying the most troublesome convergence situations. If one is dealing with multi-physics simulations with large differences in temporal and spatial scales, one must be extremely cautious.

To try to deal with this, one should attempt to determine (a) what input quantities cause the most difficulties with regard to iterative convergence, (b) what range of values of these input quantities are the most troublesome, and (c) what SRQs converge at the

slowest rate. Note that if one is dealing with an unsteady simulation or one has mixed or oscillatory convergence, this can be difficult. Even if one identifies the troublesome parameters and the problematic ranges, monitoring the convergence characteristics is still quite time consuming, and is highly prone to oversights. As a result, it is important to automate the monitoring of convergence as much as possible. Automatic tests should be programmed into the important iterative solvers so that if the iterative convergence of an element of $\vec{y}$ is suspect, then a special output warning flag is included in the output of the results. These *red flag* warning indictors should be checked on every sample that is computed for the ensemble.

### *13.3.2 Discretization error*

#### *13.3.2.1 Temporal discretization error*

When temporal discretization errors are present in a problem, there are two basic approaches for controlling it: (a) the error is estimated at each time step, compared to some error criterion, and then adjustments are possibly made to the step size; and (b) an entire solution is computed with a fixed time step and then recomputed with either a larger or smaller fixed time step. The former approach is usually referred to as a variable or adaptive time step method and the latter as a fixed time step method. Note that for a variable time step method, the error criterion must be satisfied for *all* dependent variables and *all* spatial points in the domain of the PDE. In addition, since variable time step methods only estimate the per step temporal discretization error, they are susceptible to error accumulation when a large number of steps are taken. Both methods can be effective, but the former is typically more accurate, reliable, and efficient. It should be noted that although we refer to integration in time, the integration could be in any independent variable of a hyperbolic system. For example, in fluid dynamics, the boundary layer equations are integrated in a wall-tangent spatial direction (the downstream flow direction).

Practitioners of finite element methods typically use Runge-Kutta methods for time integration. Most Runge-Kutta methods are explicit methods of order 2, 3, or 4, and they are able to estimate the discretization error at each time step. Variable time step Runge-Kutta methods are known to be very reliable and robust, resulting in their widespread use. The primary shortcomings of Runge-Kutta methods, and all explicit methods, are that they require relatively small time steps because they are conditionally stable schemes and that they can suffer from error accumulation when a large number of steps is required.

Implicit time integration methods provide a significant increase in numerical stability. This allows very large time steps to be taken, while damping both temporal and spatial modes of instability. Of course, this trade-off is at the expense of temporal discretization accuracy. Implicit methods provide significant advantages in the solution of stiff differential equations. Since most implicit methods are only second order, there is a rapid increase in discretization error as the time step is increased. Implicit Runge-Kutta methods have recently been developed and these have advantageous properties. See Cellier and Kofman

(2006) and Butcher (2008) for a detailed discussion of numerical methods for ordinary differential equations. The choice of implicit versus explicit methods should depend on both the stability restrictions and the accuracy needed to resolve the time scales in the problem.

#### 13.3.2.2 Finite-element-based methods for mesh convergence

The Zienkiewicz–Zhu (ZZ) error estimator combined with the super–convergent patch recovery (SPR) method is probably the most widely used discretization error estimator (Zienkiewicz and Zhu, 1992). The ZZ-SPR recovery method can obtain error estimates in terms of the global energy norm or in terms of local gradients, given two conditions: (a) finite element types are used that have the superconvergent property, and (b) the mesh is adequately resolved. The method has recently been extended to finite-difference and finite-volume schemes. However, accurate error estimates can only be made in the global energy norm, which is a significant disadvantage for addressing many SRQs of interest.

Residual-based methods also provide error estimates in terms of the global energy norm. Since the global energy norm is rarely an SRQ, the adjoint system for PDEs must be solved along with the primal error PDE. As discussed in Chapter 7, an extension of the residual methods, referred to here as adjoint methods, has recently been applied to various SRQs as well as to other discretization schemes (e.g., finite-volume and finite-difference methods). These approaches are very promising and are currently under investigation by a number of researchers.

#### 13.3.2.3 Richardson extrapolation error estimators for mesh convergence

Richardson extrapolation based error estimators are quite popular, particularly in fluid dynamics, because they are extremely general in their applicability. These methods do not depend on the numerical algorithm used, whether it is an IVP or BVP, the nonlinearity of the PDEs, the submodels that may be used, or whether the mesh is structured or unstructured. One disadvantage is that they require multiple solutions of the PDEs using different mesh resolutions. Standard Richardson extrapolation can be used if the numerical algorithm is second-order accurate and if the mesh is refined by a factor of two in each coordinate direction. Generalized Richardson extrapolation can be used for any order accuracy algorithm and with arbitrary refinement factors. Richardson extrapolation requires uniform mesh refinement or coarsening as one changes from one mesh solution to another. That is, the refinement or coarsening ratio must be nearly constant over the entire mesh from one mesh to the other. This type of refinement or coarsening requires significant capability of the mesh generator, particularly if an unstructured mesh is used.

The primary difficulty with using Richardson extrapolation is that the meshes must be sufficiently resolved so that all numerical solutions are within the asymptotic convergence region for the SRQ of interest. The spatial resolution required to attain the asymptotic region depends on the following factors: (a) the nonlinearity of the PDEs; (b) the range of

spatial scales across the domain of the PDEs, i.e., the stiffness of the equations; (c) how rapidly the spatial mesh resolution changes over the mesh, i.e., how nonuniform the mesh is; (d) the presence of singularities or discontinuities in the solution e.g., shock waves, flame fronts, or discontinuities in the first or second derivatives of the boundary geometry; and (e) lack of iterative convergence.

If only two mesh solutions are computed, the grid convergence index (GCI) method (Roache, 1998) can be used with a safety factor of three to indicate spatial discretization error. However, if one has no empirical evidence that the meshes are in the asymptotic region, then the GCI estimate is not reliable. Computational analysts, even experienced analysts, commonly misjudge what spatial resolution is required to attain the asymptotic region. A more reliable procedure is to obtain three mesh solutions and then compute the observed order of convergence based on the three solutions. If the observed order is relatively near the formal order of the numerical method, then one has direct empirical evidence that the three meshes are in, or near, the asymptotic region for the SRQ of interest. A factor of safety of 1.25 or 1.5 would then be justified to yield a reliable uncertainty indication for the discretization error.

#### *13.3.2.4 Practical difficulties*

Three practical difficulties were mentioned with regard to iterative convergence: (a) the wide range of types of SRQ in $\vec{y}$, (b) dealing with hundreds or thousands of samples of $\vec{y}$, and (c) the sensitivity of certain elements of $\vec{y}$ to certain ranges of input quantities that are sampled. Temporal and mesh convergence methods suffer from these same difficulties. In addition, there is a difference in the mathematical structure between iterative convergence and temporal/spatial convergence. Iterative convergence characteristics are typically monitored by examining the magnitude of the residuals during the iterations of each solution to the discrete equations. This, of course, is monitored by tracking the norm of various quantities computed over the domain of the PDEs. With temporal/spatial convergence characteristics, however, one must monitor the change in local quantities in the domain of the PDEs as the mesh is resolved. This presumes that at least one of the elements of $\vec{y}$ is a local quantity over the domain. Local quantities could include not only dependent variables, but also spatial or temporal derivatives of dependent variables. As discussed in Chapter 10, derivatives, particularly higher order derivatives, converge much more slowly with respect to discretization error than dependent variables.

If one retreats to dealing with the $L_2$ norm of the error for the SRQs, then the risk is identical to those methods discussed earlier that only yield error norms of the spatial discretization error. Monitoring the $L_\infty$ norm provides a much more sensitive indicator of temporal/spatial convergence error for each solution. The disadvantage, however, is that a great deal more *noise* will exist in the $L_\infty$ norm. The noise results from this norm's ability to jump from any point in the domain for one level of spatial or temporal resolution to any other point in the domain for another level of spatial or temporal resolution. Stated differently, noise in the $L_\infty$ norm is expected because of the nature of the mathematical

operator, whereas noise in the $L_2$ norm is a clear sign of lack of convergence. Even though more noise exists with the $L_\infty$ norm, it is a recommended quantity to monitor because it keeps a direct indication of the magnitude of the largest local error in a quantity over the domain of the PDEs.

A final recommendation is made to deal with some of the practical difficulties of temporal/spatial convergence, as well as iterative convergence. In an analysis that must deal with a wide range of values of the input quantities, it is advisable to try to identify what combinations of input quantities produce the slowest temporal, spatial, and iterative convergence rates for the most sensitive elements of $\vec{y}$. Sometimes the troublesome combinations of inputs can be deduced based on the physics occurring within the domain of the PDEs. Sometimes the troublesome combinations simply need to be found by experimentation with multiple solutions with different time steps, spatial discretizations, and iterative convergence criteria. Whatever method is used, it can greatly improve the efficiency and reliability of monitoring the temporal, spatial, and iterative error if the troublesome combinations of input are identified. That is, if the solution errors are estimated for the troublesome combinations, then these estimates can be used as bounds for the solution errors over the entire range of input quantities. These bounds may be extremely large for certain combinations of inputs, but this limited number of bounds is much easier to keep track of than estimates over a high-dimensional space of inputs.

### *13.3.3 Estimate of total numerical solution error*

The risk assessment community has, in general, taken the view that the numerical solution error should be reduced to the point that its contribution to uncertainty is much less than the aleatory and epistemic uncertainties in the analysis. This is a very prudent approach because then one can be certain that the predicted outcomes are truly a result of the assumptions and physics in the model and the characterized uncertainties, as opposed to some unknown numerical distortion of the two. There are a number of science and engineering communities, however, that continue to develop models of such complexity that available computer resources do not allow the numerical solution error to be demonstrably neglected. Faced with this situation, one has two options for proceeding. First, calibrate the adjustable parameters so that computed results agree, or nearly agree, with available experimental measurements. Sometimes researchers are unaware they have even chosen this option because they have not quantified the magnitude of the numerical solution error on the SRQs of interest. Second, explicitly quantify the numerical solution error and characterize it in some way as an uncertainty in the prediction. Essentially all researchers choose the first option.

To our knowledge, the only researchers who have made an attempt to characterize numerical solution error and then explicitly include it in the uncertainty analysis were Coleman and Stern (1997); Stern *et al.* (2001); and Wilson *et al.* (2001). They define the uncertainty due to numerical solution error, $U'_{\mathrm{N}}$, to be the square root of the sum of the squares of the following contributors: $U_{\mathrm{I}}$ is the estimated iterative solution error; $U_{\mathrm{S}}$

is the estimated spatial discretization error; $U_T$ is the estimated time discretization error; and $U_P$ is the estimated solution error caused by adjustable parameters in the numerical algorithms. $U'_N$ can be written as

$$U'_N = \sqrt{U_I^2 + U_S^2 + U_T^2 + U_P^2}. \tag{13.23}$$

Although they represent the solution error as an interval, it is clear from Eq. (13.23) that they characterize the solution error as a random variable. That is, if $U_I$, $U_S$, $U_T$, and $U_P$ are independent random variables, then the sum of their variances is equal to the variance of $U'_N$. It is clear from the previous chapters that none of these quantities is a random variable. In addition, there is a dependency structure between these four quantities of which little is usually known or quantified. One could argue that far from the region of smooth iterative convergence, or far from the asymptotic range of spatial or temporal convergence, these quantities would display a random character. However, in this random region none of the methods for error estimation would be applicable, specifically Richardson extrapolation.

Our approach for explicitly including the estimated numerical solution error is to consider each contributor as an epistemic uncertainty for each of the SRQs of interest. Even though some numerical error estimators will provide a sign for the error, we will not take advantage of this knowledge because it may not be reliable. For example, if Richardson extrapolation is used to estimate the spatial discretization error, but we are not confident we are in the asymptotic region, the estimate could be significantly in error and even the sign may be incorrect. As a result, we will always take the absolute value of the estimate of each contributing quantity. Without assuming any dependence structure between the contributors, we can write

$$(U_N)_{y_i} = |U_I|_{y_i} + |U_S|_{y_i} + |U_T|_{y_i} \quad \text{for } i = 1, 2, \ldots, n. \tag{13.24}$$

We stress again, as discussed above, that the combination of input quantities that produce a maximum value of $U_N$ for one $y_i$ is commonly different than the combination of input quantities that produce a maximum for a different $y_i$.

We do not include the uncertainty due to the adjustable parameters in the numerical algorithms because this contribution is already included in the three terms shown in Eq. (13.24). That is, adjustable parameters such as relaxation factors in iterative algorithms, numerical damping parameters, and limiters in algorithms are already reflected in the terms shown in Eq. (13.24). It can be easily shown that

$$U'_N \leq U_N \quad \text{for all } U_I, U_S, \quad \text{and} \quad U_T. \tag{13.25}$$

It is clear from Eq. (13.24) that $U'_N = U_N$ only when two of the three contributors to uncertainty are negligible relative to the third.

### 13.3.4 Example problem: heat transfer through a solid

Here we continue with the development of the example problem discussed in Sections 13.1.3 and 13.2.3. We will discuss the iterative and discretization error related to the numerical solution of Eq. (13.16) in combination because they are always intertwined in computational analyses.

#### 13.3.4.1 Iterative and discretization error estimation

As mentioned above, an analyst should attempt to identify what range of input parameters cause the SRQ of interest to converge the slowest. From a physical understanding of the heat transfer example, the slowest iterative and discretization convergence rates will occur when: (a) $T_E$ is the highest; and (b) the thermal conductivity of the plate, $k$, is the highest. Referring to Table 13.1, the highest value of $T_E$, 450 K, occurs for the system of interest, so we will only address that case. Since $k$ is given as a normal distribution without any stated bounds, i.e., the distribution has infinite support, some reasonable cumulative probability from the distribution must be chosen. We choose a cumulative probability of 0.99. Using Eq. (13.7), which is graphed in Figure 13.7, we find that $k = 209.7$ W/m-K for $P = 0.99$.

The largest temperature gradients in the plate will occur along the north face of the plate because heat is being lost along the north face and it is adjacent to the highest temperature surface, the east face. It can be seen from Eq. (13.9) that the highest heat flux through the north face will occur when $h$ is a maximum. Since the highest value of $h$ occurs in the system of interest, as opposed to the validation experiment, we use the interval-valued characterization of $h$, given by Eq. (13.11), to find that $h_{max} = 250$ W/m-K.

Chapter 7, Solution verification, and Chapter 8, Discretization error, discuss a number of different methods for estimating iterative and discretization error, respectively. For the heat transfer example, we will use the iterative error estimation technique developed by Duggirala *et al.* (2008) and we will use Richardson extrapolation to estimate the discretization error. Since these errors commonly interact in a simulation, the proper procedure is to evaluate each error component in a stepwise fashion and then conduct tests to determine if the interaction has been eliminated. Figure 13.12 depicts the 11 steps in the form of a flowchart for the estimation of iterative and discretization error.

1 Pick a sequence of three mesh resolutions such that the finest mesh resolution is believed to be adequate to satisfy the discretization error criterion. During either the preparation of a V&V plan or the conceptual modeling phase, one should decide on the maximum allowed discretization error. A discretization error criterion should be picked for each of the SRQs of interest. For the heat transfer example, we only have one SRQ, $q_W$. It is usually best to pick a relative error criterion, since this automatically scales the absolute error with the magnitude of the quantity. One should choose a demanding error criterion relative to the accuracy needs of the analysis because one should be certain that any disagreement between the model predictions and the experimental measurements is due to the physics assumptions in the model and *not* due to numerical errors. Here we pick a relative discretization error criterion for $q_W$ of 0.1%.

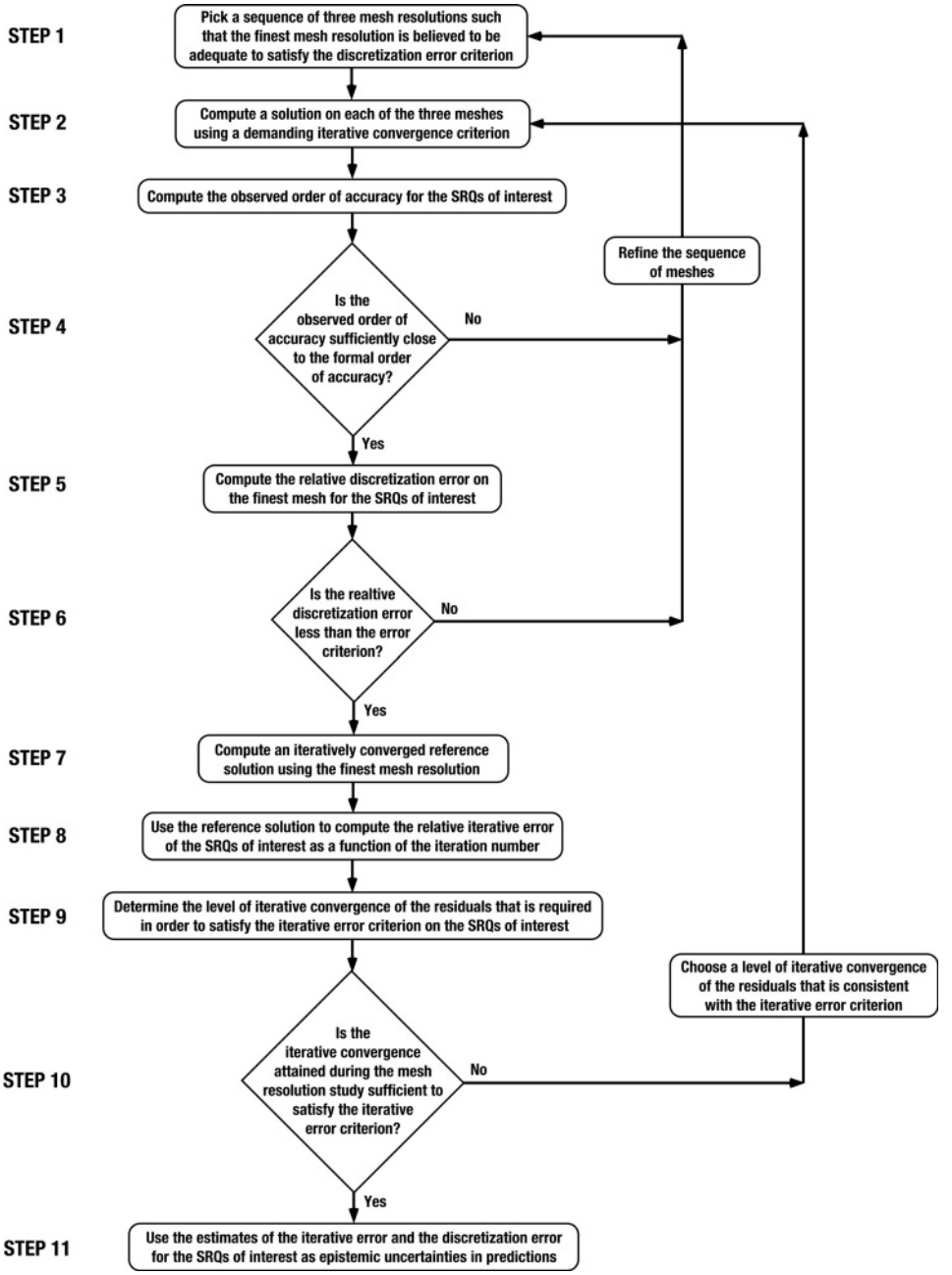


Figure 13.12 Flow chart for determining satisfactory iterative and discretization convergence.

The estimation of what mesh resolution will satisfy the error criterion is commonly based on experience with the numerical solution of similar problems with similar numerical methods. The mesh resolution required also depends on the effectiveness of the mesh clustering structure. Our experience, along with many others, is that numerically accurate solutions almost always require finer meshes than one would expect. Here we pick the three mesh resolutions, $21\times21$, $31\times31$, and $46\times46$, yielding a constant mesh refinement factor of 1.5 for both refinements. As discussed in Chapter 8, when one chooses a noninteger refinement factor and one of the SRQs of interest is a local quantity over the domain of the PDE, one must resort to an interpolation procedure to make the needed calculations for Richardson extrapolation. Since $q_W$ is not a local quantity, we avoid this difficulty.

2 Compute a solution on each of the three meshes using a demanding iterative convergence criterion. Eq. (13.16) is solved for the interior mesh points and the BCs are given by $T_E = 450$ K, $T_W = 300$ K, Eq. (13.10) for the north face with $k = 209.7$ W/m-K and $h = 250$ W/m-K, and Eq. (13.13) for the south face. The initial guess for temperature over the solution domain is simply taken as the average temperature between the east and west faces, $T_{initial} = 375$ K. The iterative method used is the well-known Gauss–Seidel relaxation method. The iterative convergence criterion chosen here is based on the $L_2$ norm of the residuals of the dependent variable in the PDE, i.e., $T$, being solved. The convergence criterion chosen at this point is rather arbitrary because the definitive criterion will be set later in Step 9. One attempts to pick a criterion that is sufficiently demanding so that it is expected to satisfy the criterion evaluated in Step 10. Here we require that the norm of the residuals decrease by nine orders of magnitude compared to the first $L_2$ norm computed.

3 Compute the observed order of accuracy for the SRQs of interest. As discussed in Chapter 8, the observed order of discretization accuracy, $p_O$, can be computed using Richardson's method and three solutions. We have

$$p_O = \frac{\ln\left(\frac{f_3 - f_2}{f_2 - f_1}\right)}{\ln(r)}, \tag{13.26}$$

where $f_1$, $f_2$, and $f_3$ are the solutions for the SRQ on the fine, medium, and coarse meshes, respectively, and $r$ in the mesh refinement factor.

4 Test if the observed order of accuracy of the SRQs of interest is sufficiently close to the formal order of accuracy. To use Richardson's method, one must have some evidence that the three mesh solutions are in or near the asymptotic region of spatial convergence. Computing $p_O$ and comparing it with the formal order of accuracy, $p_F$, which is 2 for the present case, can provide the evidence. There is no strict requirement on how close $p_O$ must be to $p_F$, but a typical requirement is

$$|p_O - p_F| < 0.5. \tag{13.27}$$

If $p_O$ is not sufficiently close to $p_F$, then the mesh resolutions are not sufficiently in the asymptotic region and we must refine the sequence of meshes, and return to Step 1. If $p_O$ and $p_F$ are sufficiently close, then we can proceed to Step 5.

5 Compute the relative discretization error on the finest mesh for the SRQs of interest. The relative discretization error, normalized by the extrapolated estimate of the exact solution, $\bar{f}$, is given by

$$\frac{f_1 - \bar{f}}{\bar{f}} = \frac{f_2 - f_1}{f_1 r^{p_O} - f_2}. \tag{13.28}$$

This equation can also be solved for $\bar{f}$ to give

$$\bar{f} = f_1 + \frac{f_1 - f_2}{r^{p_o} - 1}. \tag{13.29}$$

6 Test if the relative discretization error is less than the error criterion. Using Eq. (13.28), we can compute the relative discretization error in $q_W$ and determine if it is less than the relative error criterion of 0.1%. If the computed error is not less than the error criterion, then we must refine the sequence of meshes and return to Step 1. If the computed error is less than the error criterion, then we can proceed to Step 7.

7 Compute an iteratively converged reference solution using the finest mesh resolution. The technique we use for estimating the iterative error is based on generating a reference solution and then developing a mapping between the reference solution and the $L_2$ norm of the residuals for the PDE being solved. The norm of the residuals for the reference solution should be converged at least ten orders of magnitude; possibly even to machine precision. This solution will be used as the fully converged or reference solution to the discrete equations. When this reference solution is being iteratively converged, the $L_2$ norm of the residuals as well as the SRQs of interest should be saved every 50 or 100 iterations. The saved results will be used to construct the mapping between the reference solution and the $L_2$ norm of the residuals for the PDE. If any of the SRQs are quantities defined over the domain of the PDEs, e.g., dependent variables, then the $L_2$ norm of the quantity should be used.

8 Use the reference solution, $f_{\text{reference}}$, to compute the relative iterative error as a function of the iteration number for the SRQs of interest. Using the reference solution as the exact solution, one can compute the relative iterative error as a function of the iteration number using

$$\%\ \text{error in}\ f_{i\text{th iteration}} = \left| \frac{f_{\text{reference}} - f_{i\text{th iteration}}}{f_{\text{reference}}} \right| \times 100\%, \tag{13.30}$$

where $f_{i\text{th iteration}}$ is the value of the SRQ, or the $L_2$ norm of the quantity if it is a dependent variable in the PDE, at the $i$th iteration.

9 Determine the level of iterative convergence of the residuals that is required in order to satisfy the iterative convergence criterion on the SRQs of interest. The iterative convergence criterion should be specified during the preparation of the V&V plan or during the conceptual modeling phase. It is recommended to be 1/100th of the relative discretization error criterion specified for any SRQs of interest. This much smaller value for the relative iterative error criterion is chosen so that we can be certain that there is little or no interaction between iterative convergence and discretization convergence. Here we pick a relative iterative convergence criterion of 0.001% for $q_W$.

Using the quantities saved in Step 7, one can plot the iterative convergence history of the residuals along with the relative iterative error in the SRQs as a function of the iteration number. Using the relative iterative error criterion as the requirement, one can then determine the level of iterative convergence of the residuals that is needed in order to satisfy the iterative error criterion. That is, the combined plot provides a mapping between the desired iterative error in the SRQs and the iterative convergence of the residuals. When other numerical solutions are computed, such as varying the input parameters in a design study or in Monte Carlo sampling of the uncertain inputs, the iterative convergence rate of most SRQs will change. However, as long as the mapping relationship between the iterative error in the SRQs and the iterative convergence of the residuals remains the same, then this iterative error estimation procedure will be reliable.

10 Test if the iterative convergence attained in the $L_2$ residuals during the mesh resolution study is sufficient to satisfy the iterative error criterion in the SRQs of interest. The purpose of this test is to determine if the level of convergence of the $L_2$ residuals used in Step 2 is smaller than the level of convergence of the $L_2$ residuals determined in Step 9. If the convergence of the residuals in Step 2 was inadequate, then we must use the residual convergence level determined in Step 9 and return to Step 2. If the convergence of the residuals in Step 2 is satisfactory, we can proceed to Step 11.

11 Use the estimates of the iterative error and the discretization error for the SRQs of interest as epistemic uncertainties in predictions. The values computed for the relative iterative error and the relative discretization error for the SRQs of interest are converted to absolute errors so that they can be substituted into Eq. (13.24). Since these errors are, hopefully, for the case(s) of slowest iterative and discretization convergence, they should be conservative bounds on the errors for all other conditions computed in the analysis.

#### 13.3.4.2 Iterative and discretization error results

Now, we give the key results for the heat transfer example from various steps discussed above.

From Step 2: The $L_2$ norm of $T$ on the first iteration was computed to be $1.469 \times 10^4$ K. Using the preliminary relative iterative convergence criterion of nine orders of magnitude decrease in the $L_2$ norm, we have $10^{-9} \bullet 1.469 \times 10^4 = 1.469 \times 10^{-5}$ K as the preliminary $L_2$ norm criterion.

From Step 3: The observed order of accuracy in $q_W$ is computed from the solutions on the three meshes $21 \times 21$, $31 \times 31$, and $46 \times 46$. The values of $q_W$ from these three solutions are $-270.260\,295\,374$, $-270.477\,983\,330$, $-270.575\,754\,310$ W, respectively. Using these values and Eq. (13.26), the result is $p_O = 1.974$.

From Step 4: Since $p_O = 1.974$ satisfies Eq. (13.27), the test in Step 4 is satisfied.

From Step 5: The relative discretization error in $q_W$ is computed on the $46 \times 46$ mesh using Eq. (13.28), the solutions on the $31 \times 31$ and $46 \times 46$ meshes, and the observed order of accuracy of 1.974. The relative discretization error is computed to be $-3.1214 \times 10^{-4}$. Using Eq. (13.29), the estimate of the converged solution obtained from Richardson extrapolation is computed to be $-270.660\,237\,711$ W. This results in an estimated discretization error in the $46 \times 46$ mesh solution of 0.08448 W.

From Step 6: Since this relative discretization error is less than the error criterion of $1 \times 10^{-3}$, the test in Step 6 is satisfied.

From Steps 7–10: Figure 13.13 shows the iterative convergence of the $L_2$ residuals of temperature on the left axis and the relative iterative error in $q_W$ on the right axis. The reference solution was converged 12 orders of magnitude compared to the magnitude of the $L_2$ norm computed on the first iteration. The $L_2$ norm on the first iteration is $1.469 \times 10^4$ K. The relative iterative error in $q_W$ was computed using the reference solution and Eq. (13.30). Using Figure 13.13 one can map from right to left the relative iterative error criterion of $1 \times 10^{-5}$ to the $L_2$ norm of the residuals resulting in a value of $2.2 \times 10^{-5}$ K. Since this value of the $L_2$ norm is larger than the criterion imposed on the $L_2$ norm in Step 2 ($1.469 \times 10^{-5}$ K), the test in Step 10 is satisfied. The value of $1.469 \times 10^{-5}$ K in the $L_2$ norm can then be mapped back to $0.64 \times 10^{-5}$ for the relative iterative error in $q_W$.

Table 13.5 *Levels of iterative convergence of the $q_W$ for the 46 × 46 mesh*

| Order of magnitude drop in the $L_2$ norm | $L_2$ norm at convergence (K) | Number of iterations at convergence | Calculated relative discretization error | Observed order of accuracy, $p_O$ |
|---|---|---|---|---|
| 4 | $1.469 \times 10^{0}$ | 1250 | $1.266 \times 10^{-2}$ | −0.592 |
| 5 | $1.469 \times 10^{-1}$ | 2135 | $-1.049 \times 10^{-3}$ | 1.136 |
| 6 | $1.469 \times 10^{-2}$ | 3021 | $-3.595 \times 10^{-3}$ | 1.864 |
| 7 | $1.469 \times 10^{-3}$ | 3907 | $-3.166 \times 10^{-3}$ | 1.963 |
| 8 | $1.469 \times 10^{-4}$ | 4793 | $-3.125 \times 10^{-3}$ | 1.973 |
| 9 | $1.469 \times 10^{-5}$ | 5679 | $-3.121 \times 10^{-3}$ | 1.974 |
| 10 | $1.469 \times 10^{-6}$ | 6565 | $-3.121 \times 10^{-3}$ | 1.974 |

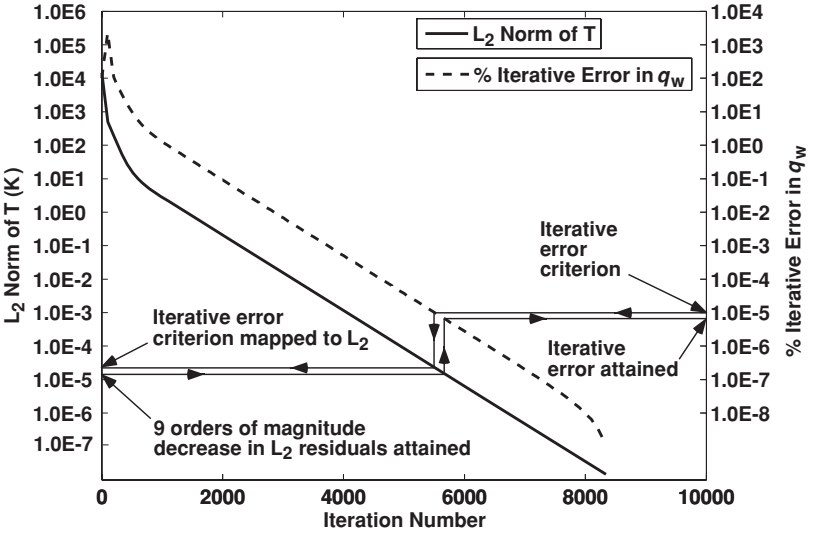


Figure 13.13 Iterative convergence history of the $L_2$ norm of temperature and the relative iterative error in $q_W$.

As an independent check on the adequacy of the iterative convergence, the following procedure is commonly done. When Step 7 is satisfactorily attained, one can compute a series of solutions that are iteratively converged to increasingly higher degrees. Table 13.5 shows varying levels of iterative convergence for $q_W$ on the 46 × 46 mesh. The solutions are converged by four orders of magnitude up to ten orders of magnitude compared to the initial value of the $L_2$ norm. It can be seen from the fourth and fifth columns of the table that the calculated relative discretization error and the observed order of accuracy do not stabilize at the correct value until the solution has been converged at least eight orders

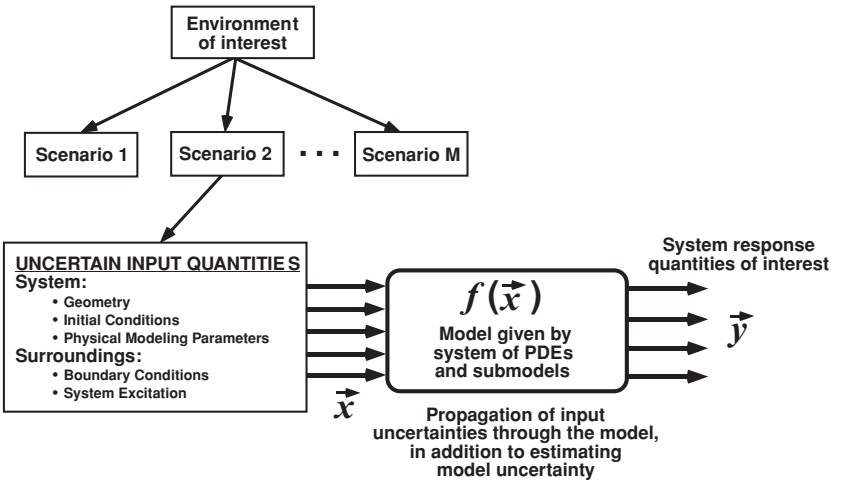


Figure 13.14 Example of sources of uncertainties that yield uncertain system response quantities.

of magnitude. Stated differently, until the solution is converged at least eight orders of magnitude there is an interaction of the iterative error and the discretization error, which results in erroneous results for the computed discretization error and the observed order of accuracy. This type of pollution of observed order of accuracy has certainly occurred in the published literature.

Referring back to the stepwise procedure described above and the results shown in Figure 13.13, it is seen that the procedure yields results that are consistent with the convergence results show in Table 13.5. For example, the requirement that the relative iterative error criterion should be set at 1/100th of the relative discretization error criterion is justified by noting the results in Table 13.5. An iterative error criterion of 1/10th of the relative discretization error criterion could have been used for this example, but it would have been right at the edge of noticeable interaction of the iterative and discretization errors.

From Step 11: Summarizing, from Step 5, we use the estimated discretization error on the $46 \times 46$ mesh to give $U_S = 0.\ 08448$ W. From Step 10 we use the value of $0.8 \times 10^{-5}$ for the relative iterative error in $q_W$ to compute $U_I = -0.00173$ W. Although these values are very small compared to the estimated model uncertainty discussed in Section 13.2.3, they will be included in the next section to estimate the total predictive uncertainty of the simulation.

## 13.4 Step 4: estimate output uncertainty

Figure 13.14 shows an overview of the procedure for estimating output uncertainty. Here we start with the specification of the environment of interest, identify the scenarios of interest, characterize the uncertain inputs, propagate these through the model, and produce a vector of SRQs of interest, $\vec{y}$. Although there may be probabilities associated with a

particular environment and a particular scenario, for our present purposes we will ignore those probabilities. Here we are focused on how the input, model, and numerical solution uncertainties combine to affect $\vec{y}$. Let $\vec{x}$ be the vector of all uncertain input quantities. Let $m$ be the number of uncertain elements in the vector, so that we have

$$\vec{x} = \{x_1, x_2, \ldots, x_m\}. \tag{13.31}$$

Let $f(\vec{x})$ represent the dependence of the model on the uncertain input quantities. $f(\vec{x})$ can also be thought of as the function that maps input uncertainties to output uncertainties. If there are multiple models for $f(\vec{x})$, then there will be multiple mappings of input to output uncertainties. Here we assume, for generality, that all of the input quantities of the model are uncertain. Because of the different way in which input, model, and numerical uncertainties occur in the model of the system and the surroundings, they are normally treated separately. Determining the affect of the input uncertainties, $\vec{x}$, on the response vector $\vec{y}$ is termed *propagation of input uncertainties*.

There is a wide range of methods to propagate input to output uncertainties. It is beyond the scope of this book to discuss all of them and when each is appropriate in a UQ analysis. For a detailed discussion of various methods, see Morgan and Henrion (1990); Cullen and Frey (1999); Melchers (1999); Haldar and Mahadevan (2000a); Ang and Tang (2007); Choi *et al.* (2007); Suter (2007); Rubinstein and Kroese (2008); and Vose (2008). As discussed earlier, probability bounds analysis (PBA) is used in the characterization of the input uncertainties, $\vec{x}$, the model uncertainty, and the characterization of the uncertainties in the SRQs, $\vec{y}$. To our knowledge, the only methods that are able to propagate aleatory and epistemic uncertainties through an arbitrary, i.e., black box, model are Monte Carlo sampling methods. For very simple models that are not black box, one could program each of the arithmetic operations in the model and propagate the input uncertainties to obtain the output uncertainties (Ferson, 2002). Because this approach is very limited in applicability, we will only discuss Monte Carlo sampling methods.

### *13.4.1 Monte Carlo sampling of input uncertainties*

Monte Carlo methods were first used about a century ago, but they have only become popular during the last half-century. They are used in a wide range of calculations in mathematics and physics. There are a number of variants of Monte Carlo sampling (MCS), each serving particular goals with improved efficiency (Cullen and Frey, 1999; Ross, 2006; Ang and Tang, 2007; Dimov, 2008; Rubinstein and Kroese, 2008; Vose, 2008). The key feature of all Monte Carlo methods is that the mathematical operator $f$ of interest is evaluated repeatedly using some type of random sampling of the input. As indicated in Figure 13.14 above, we write

$$\vec{y} = f(\vec{x}). \tag{13.32}$$

Let $\vec{x}^k$ denote a random sample drawn from all of the components of the input vector $\vec{x}$. Let $\vec{y}^k$ be the response vector after evaluation of the model using the random sample $\vec{x}^k$. Then

Eq. (13.32) can be written for the number of samples $N$ as

$$\vec{y}^k = f(\vec{x}^k), \quad k = 1, 2, 3, \ldots, N. \tag{13.33}$$

The key underlying assumption of *simple* or *basic* MCS can be stated as: given a set of $N$ random samples drawn from all of the aleatory and epistemic input uncertainties, one can make strong statistical statements concerning the nondeterministic response of the model. There are *no assumptions* concerning: (a) the characteristics of the uncertainty structure of the input to the model, e.g., parametric versus nonparametric, epistemic versus aleatory, correlated versus uncorrelated; or (b) the characteristics of the model, e.g., whether there is any required smoothness or regularity in the model. The only critical issue in MCS is that the strength of the statistical statements about the system response depends on the number of samples obtained. If $f(\vec{x}^k)$ is computationally expensive to evaluate, then one may have to deal with less precise statistical conclusions concerning the response of the system. Alternatively, one may have to simplify the model or neglect unimportant submodels in order to afford the needed number of samples. However, with the power of highly parallel computing, this computational disadvantage of MCS is mitigated. This is also true with the latest desktop computer systems being designed with multiple compute cores on a single chip. Each of the function evaluations, $f(\vec{x}^k)$, can be done in parallel.

Some academic researchers developing new techniques for propagating input to output uncertainties are harshly critical of any form of MCS because of the computational expense involved in evaluating $f(\vec{x}^k)$. Other methods have been proposed, such as the use of stochastic expansions, i.e., polynomial chaos expansions and the Karhunen–Loeve transform (Haldar and Mahadevan, 2000b; Ghanem and Spanos, 2003; Choi *et al.*, 2007). These methods converge rapidly and are most promising. However, many of these methods require that the computer code that solves for $f(\vec{x}^k)$ be substantially modified before they can be used, i.e., they are intrusive to the code. For academic exercises or specialized applications, this is quite doable. However, for complex analyses of real systems, where the codes can have hundreds of thousands of lines, have possibly been used for decades, and where multiple codes are executed in sequence, this route is completely impractical. In addition, and just as important, the new methods have not been able to deal with epistemic uncertainty. Generality with regard to models chosen, robustness of the method, and ability to deal with both aleatory and epistemic uncertainty has sustained the use of MCS for decades in UQ analyses.

#### *13.4.1.1 Monte Carlo sampling for aleatory uncertainties*

The initial discussion of MCS will only deal with uncertain inputs that are aleatory, i.e., they are given by precise probability density functions or CDFs. Figure 13.15 depicts the basic concepts in MCS for a system of three uncertain inputs $(x_1, x_2, x_3)$ resulting in one uncertain output, $y$. The first step is to draw uniformly sampled random deviates between zero and one. Figure 13.15 shows how these samples applied to the probability axis will generate the random deviates $(x_1, x_2, x_3)^k$, based on the particular CDF characterizing each input quantity. Each of these random deviates is used to evaluate the function $f$ to compute

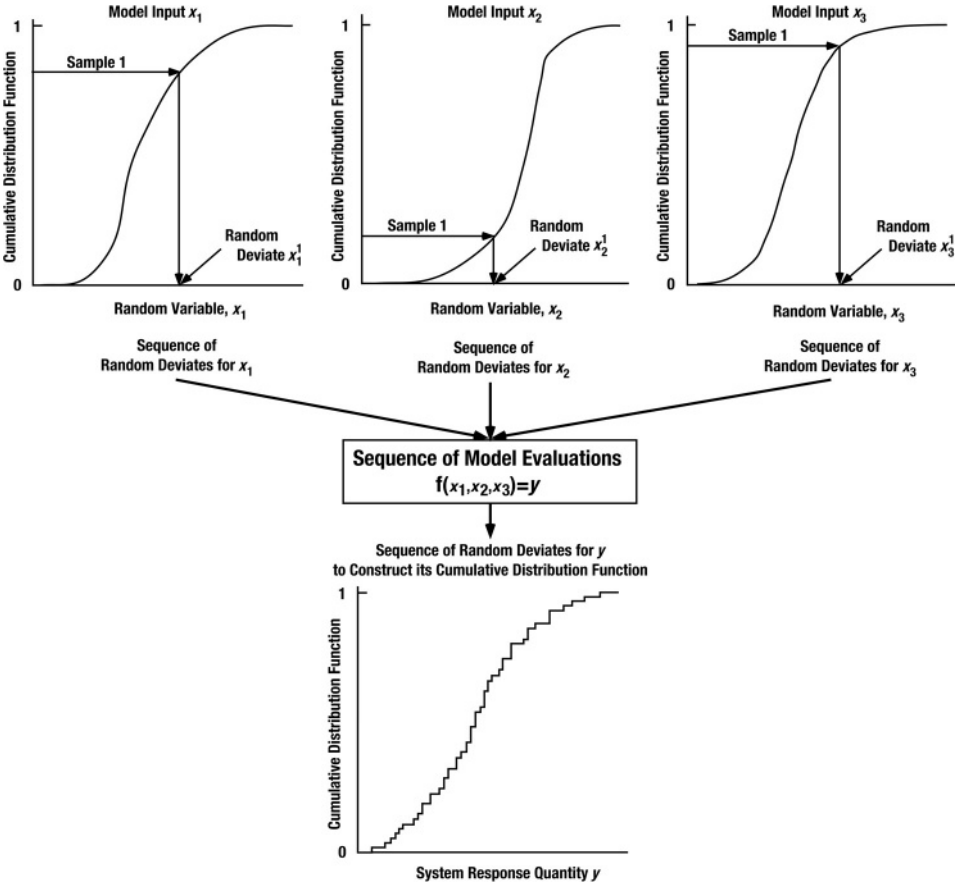


Figure 13.15 Schematic of MCS for three uncertain inputs and one output.

one $y$. After many function evaluations using many random deviates, $k = 1, 2, 3, \ldots, N$, one can construct an empirical CDF as shown in the bottom panel of Figure 13.15.

Figure 13.16 shows a flowchart and gives more of the details for the activities in the various steps in the process for uncorrelated uncertain inputs. Let the number of uncertain aleatory inputs be $\alpha$ and the remainder of the uncertain inputs, $m-\alpha$, be epistemic. Sampling of the epistemic uncertainties will be discussed in the next section. The following explains each of the steps for basic MCS.

1 Generate $\alpha$ sequences of $N$ pseudo-random numbers, one for each of the uncertain aleatory inputs. Since each sequence must be independent of the other, each sequence must use a different seed for the pseudo-random number (PRN) generator or be a continuation of a sequence. As is common with most PRN generators, the numbers range between zero and unity.
2 Picking an individual number from each of the sequences of PRNs, create an array of numbers of length $\alpha$. That is, take one number from the first sequence of PRNs, take one number from the

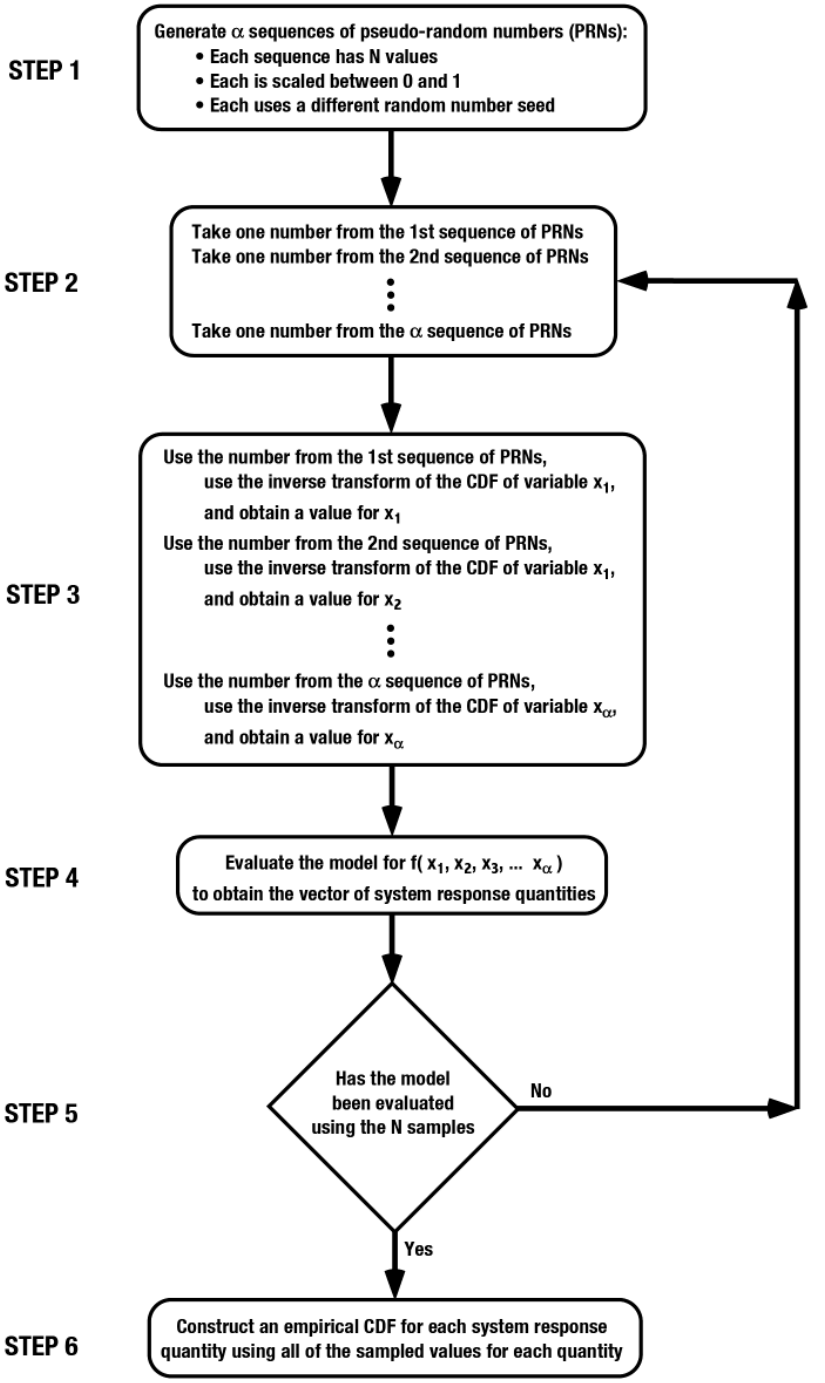


Figure 13.16 Flow chart for simple MCS with only aleatory uncertainties and no correlated input quantities.

second sequence, etc., until the array of length $\alpha$ is created. Once a number from the sequence has been used, it is not used again.

3 Each number in the array generated in Step 2 can be viewed as drawn from a uniform distribution over the range of (0,1). The probability integral transform theorem in statistics is used to map the uniform distribution, using the CDF for each uncertain aleatory input, to obtain a value of the uncertain input (Angus, 1994; Ang and Tang, 2007; Dimov, 2008; Rubinstein and Kroese, 2008). That is: given a distribution $F$ and a uniform random value $u$ between zero and one, the value $F^{-1}(u)$ will be a random variable distributed according to $F$. This is what it means for a random variable to be "distributed according" to a distribution. The result of this step is an array of $\alpha$ model input quantities, $(x_1, x_2, \ldots x_\alpha)$, that are distributed according to the given distribution for input 1, 2, 3, $\ldots$, $\alpha$, respectively. The process in this step is what is depicted in the top panel of Figure 13.15.

4 Use the array $(x_1, x_2, \ldots, x_\alpha)$ as input to evaluate the mathematical model. By solving the PDEs, along with all of the ICs, BCs, and system excitation, compute the array, $(y_1, y_2, \ldots, y_n)$, of SRQs.

5 Test if all $N$ PRNs have been used to evaluate the mathematical model. If No, return to Step 2. If Yes, go to Step 6.

6 Construct an empirical CDF for each element of the SRQ array, $(y_1, y_2, \ldots, y_n)$. The empirical CDF is constructed using each of the samples from the Monte Carlo simulation. The abscissa of the empirical CDF is the observed (sampled) values of an SRQ, $y_i$, and the ordinate of the empirical CDF is the observed cumulative probability for all of the sampled values less than or equal to $y_i$. The $y_i$ are ordered from the smallest to the largest. The empirical CDF is constructed as a nondecreasing step function with a constant vertical step size of $1/N$, where $N$ is the sample size from the MCS. The locations of the steps correspond to the observed values of $y_i$. Such a distribution for the samples $y^k, k = 1, 2, 3 \ldots, N$ is

$$S_N(y) = \frac{1}{N} \sum_{k=1}^{N} I(y^k, y), \tag{13.34}$$

where

$$I(y^k, y) = \begin{cases} 1, & y^k \leq y, \\ 0, & y^k > y. \end{cases} \tag{13.35}$$

$S_n(y)$ is simply the fraction of data values in the data set that are at or below each $y^k$. From Eq. (13.34) it is clear that the total probability mass accumulated from the total of $N$ empirical samples is unity. An example of an empirical CDF is shown in the bottom panel of Figure 13.15.

There are a number of additional topics concerning MCS that are important in practical analyses and are not addressed in Figure 13.15 and Figure 13.16. Two of these will be briefly mentioned here, but the reader should consult the following references for details (Cullen and Frey, 1999; Ross, 2006; Ang and Tang, 2007; Dimov, 2008; Rubinstein and Kroese, 2008; Vose, 2008). First, some type of correlation or dependence structure may exist between various uncertain inputs. If a correlation structure exists between two (or more) inputs, then it means that one is statistically related to the other(s). For example, suppose one considers uncertainty in both thermal and electrical conductivity. For most materials there is a strong correlation between thermal and electrical conductivity. If a

correlation structure exists between two (or more) inputs then one is claiming that one has a *causal* relationship with the other(s), e.g., there is a high dependence between how much rest time an equipment operator has and the likelihood of errors made during the operation of the equipment. Determining a correlation or dependence structure is usually an involved task that is based primarily on large amounts of experimental data, as well as an understanding of the physical relationship between the quantities. If correlation or dependence structures are quantified, it is accounted for in Step 3.

Second, how many Monte Carlo samples are needed for calculating various statistical quantities of the response depends on different factors. The most important of these is the probability value of interest. Ang and Tang (2007) give an estimate of the sampling error in MCS as

$$\%\ \text{error in}\ P = 200\sqrt{\frac{1-P}{NP}}, \tag{13.36}$$

where $P$ is the probability value of interest. It can be seen that the mean converges most rapidly, while low probability events converge very slowly. For example, if a probability of 0.01 is needed, then 100 000 samples are typically required to assure an error of less than 6%. It should be stressed, however, that one of the important advantages of a Monte Carlo sample is that the convergence rate does *not* depend on the number of uncertain quantities or their variance.

To improve the convergence rate of MCS, variance reduction techniques can be used (Cullen and Frey, 1999; Helton and Davis, 2003; Ross, 2006; Dimov, 2008; Rubinstein and Kroese, 2008). These techniques adjust the random samples of the inputs so that certain features of the output distribution converge more rapidly. The most well known is stratified Monte Carlo sampling or Latin Hypercube sampling (LHS). LHS stratifies the uniform distribution that is used with the inverse transform procedure into equal intervals of probability. The most common procedure is to let the number of intervals be equal to the number of samples that will be computed. Across all of these intervals a PRN generator is used to obtain a separate random number for each interval so that a random sample is obtained from each interval. LHS almost always converges faster than simple MCS, so it is the most commonly used sampling technique. For a small number of uncertain quantities, e.g., less than five, LHS converges much faster than simple MCS. For one uncertain quantity, Dimov (2008) shows that

$$\%\ \text{error in}\ P \sim N^{-3/2}. \tag{13.37}$$

All of the modern risk assessment software packages contain a number of sophisticated features, such as dealing with correlations, dependencies, and different methods for variance reduction in the sampling. Examples of some of the packages are JMP and STAT from SAS Inc., Risk Solver from Frontline Systems, Inc., STATISTICA from StatSoft, Inc., @Risk from Palisade Software, and Crystal Ball from Oracle, Inc.

*13.4.1.2 Monte Carlo sampling for combined aleatory and epistemic uncertainties*

Modifying the simple Monte Carlo procedure to deal with the epistemic uncertainties is rather straightforward. Recall that the epistemically uncertain inputs are $(x_{\alpha+1}, x_{\alpha+2}, \ldots, x_m)$. We presume that each of these quantities is given by an interval, i.e., there is no belief or knowledge structure within the interval. There are generally two types of epistemic uncertainty occurring in the modeling process. First, there are those that occur in the modeling of the physical characteristics of the system and the surroundings. Some examples are geometry characteristics, physical modeling parameters (such as material properties), and boundary conditions (such as pressure loading on a system). Second, there are those that occur in the uncertainty characterization of aleatory uncertainties. The most common example is specifying the parameters of a family of distributions as intervals. For example, suppose one is modeling the manufacturing variability of a material property using a three-parameter gamma distribution, where each parameter is specified as an interval. Then each of the three parameters of the distribution would be an element of the array $(x_{\alpha+1}, x_{\alpha+2}, \ldots, x_m)$. The uncertainty structure of this gamma distribution would be the set of all possible gamma distributions whose parameters lie within each of the specified intervals of the parameters.

Since the epistemic uncertainties have no structure within the interval, one could use basic MCS over the range of each interval-valued quantity. This would be the same procedure as described above for aleatory uncertainties. However, it must be stressed that the sampling of the intervals is *fundamentally different* in terms how these samples are processed and interpreted. Each of these samples represents *possible realizations* that could occur in these epistemically uncertain quantities. There is *no likelihood or probability associated with these samples*, which is in contrast to the samples taken from the aleatory uncertainties. As a result, the key to sampling when epistemic and aleatory uncertainties are present is to separate the sampling for the epistemic uncertainties from the sampling of the aleatory uncertainties. This procedure is the essence of a PBA.

To accomplish this, one constructs a double-loop sampling structure. The outer loop is for sampling of the epistemic uncertainties, and the inner loop is for sampling the aleatory uncertainties. It has been found that LHS is more efficient for sampling the epistemic uncertainties (outer loop) than simple MCS because LHS forces the samples into partitions across the interval-valued quantities (Helton and Davis, 2003; Helton *et al.*, 2005b, 2006; Helton and Sallaberry, 2007). Figure 13.17 shows a flow chart for the double-loop sampling strategy. The flow chart is described as if there is only one SRQ, but the same procedure would be used for any number of SRQs. The following gives a brief explanation of each step of the procedure.

1 Choose the number of samples, $M$, that will be used for sampling of the epistemic uncertainties. Because of the interval nature of the epistemic uncertainties, an appropriate structure for sampling would be a combinatorial design. If LHS is used, $M$ must be sufficiently large to insure satisfactory coverage of the combinations of all of the epistemic uncertainties in the mapping to the SRQs. Based on the work of Ferson and Tucker (2006); Kreinovich *et al.* (2007); and Kleb and Johnston

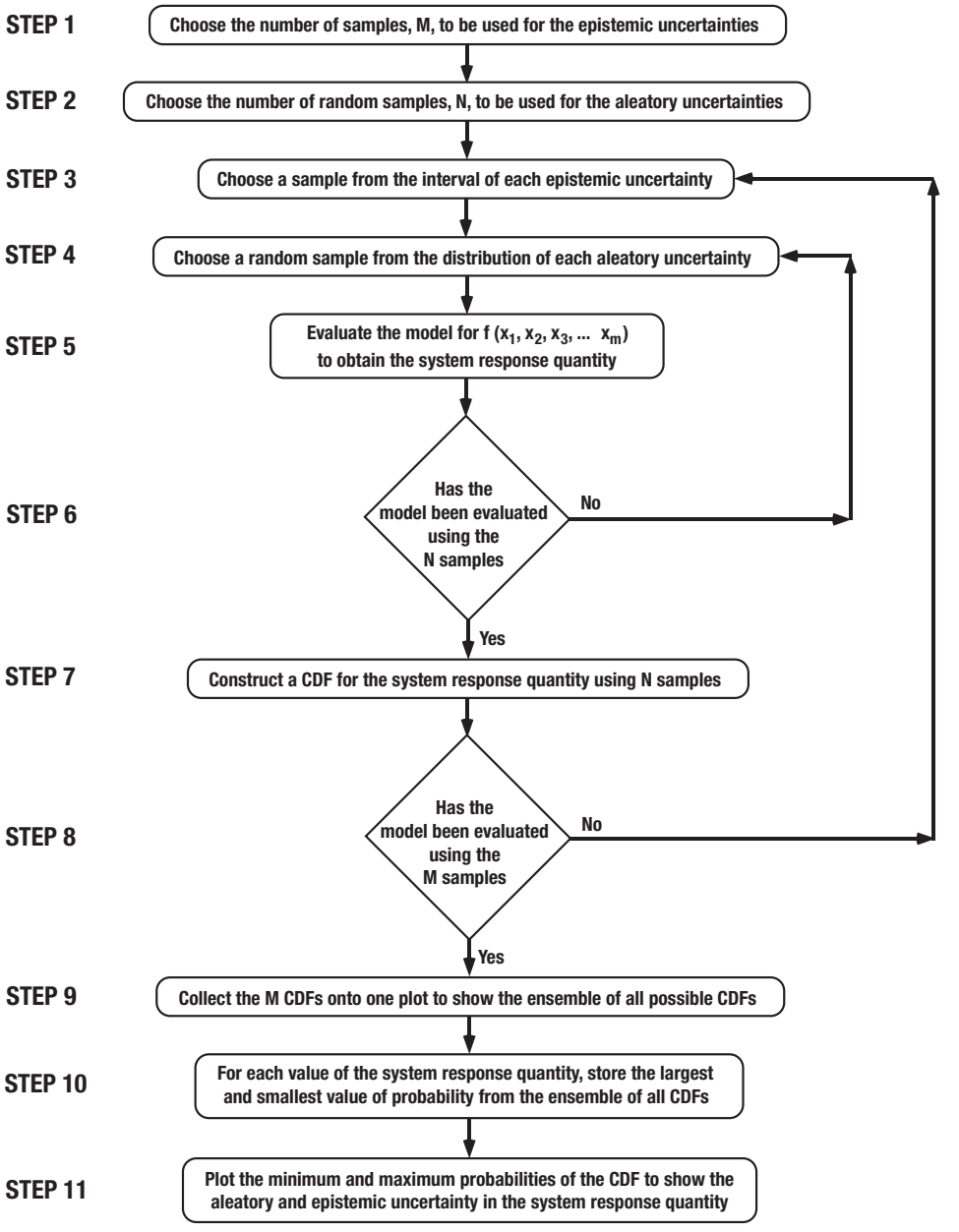


Figure 13.17 Flow chart for MCS with both aleatory and epistemic uncertainties and no correlated input quantities.

(2008), we suggests that a minimum of three LHS samples be taken for each epistemic uncertainty, in combination with all of the remaining epistemic uncertainties. An approximation to this suggestion for the minimum number of samples is given by $(m - \alpha)^3 + 2$, where $(m - \alpha)$ is the number of epistemic uncertainties. If $(m - \alpha)$ is not very large, the suggested minimum number of samples should be computationally affordable. Sallaberry *et al.* (2008) suggest that replicated LHS sampling be used in order to detect if there is sensitivity of the results to the number of samples. Replicated sampling is the procedure where multiple sets of samples are computed, each set having a different seed for the pseudo-random number (PRN) generator. Instead of using LHS sampling, Kreinovich *et al.* (2007) suggest a sampling method based on Cauchy deviates. Although LHS sampling is shown in Figure 13.17, the methods of Sallaberry *et al.* (2008) and Kreinovich *et al.* (2007) could also be incorporated.

2 Choose the number of samples, $N$, that will be used for the random sampling of the aleatory uncertainties. Depending on what quantile is of interest for the SRQ, the number of samples may need to be very large, as discussed earlier.

3 Choose a sample from the interval of each epistemic uncertainty. If LHS sampling is used, the number of strata for each epistemic uncertainty should be set equal to $M$. In addition, it is recommended to use a uniform distribution for each of the strata to map the random samples on the probability axis to obtain the random deviates of the epistemic uncertainties. An additional technique that has proven effective is to require that the end points of each epistemic uncertainty be sampled. That is, for each of the strata on the ends of the interval, one ignores the uniform sampling and chooses the end point of the interval. This technique ensures that the full range of the interval for each epistemic uncertainty is sampled, regardless of the number of LHS samples.

4 Choose a random sample from each distribution of each aleatory uncertainty. LHS is also recommended for sampling the aleatory uncertainties, particularly if there are a small number of aleatory uncertainties. The number of strata of each aleatory uncertainty should be set equal to $N$.

5 Use the complete array of sampled values $(x_1, x_2, \ldots, x_m)$ to evaluate the mathematical model and compute the SRQ.

6 Test if all $N$ samples of the aleatory uncertainties have been used to evaluate the mathematical model. If No, return to Step 4. If Yes, go to Step 7.

7 Construct a CDF based on the $N$ observed (sampled) values of the aleatory uncertainties.

8 Test if all $M$ samples of the epistemic uncertainties have been used to evaluate the mathematical model. If No, return to Step 3. If Yes, go to Step 9.

9 Collect the $M$ CDFs onto one plot to show the ensemble of all CDFs. Each CDF shows a possible distribution of realizations of the SRQ.

10 For each observed value of the SRQ, store the largest and smallest value of probability from the ensemble of all CDFs.

11 Plot the minimum and maximum CDFs over the range of the observed SRQs. This plot shows the possible range in probabilities for all of the observed SRQs. This is referred to as a p-box because it shows interval-valued probabilities for the SRQ. That is, given the information characterized in the uncertain input quantities, no tighter range of probabilities can be claimed for the response of the system.

Figure 13.18 is a sample of a p-box with large epistemic uncertainty in the system response quantity. For any value of the SRQ, only an interval-valued probability can be determined. Likewise, for any value of probability, only an interval-valued response can be determined.

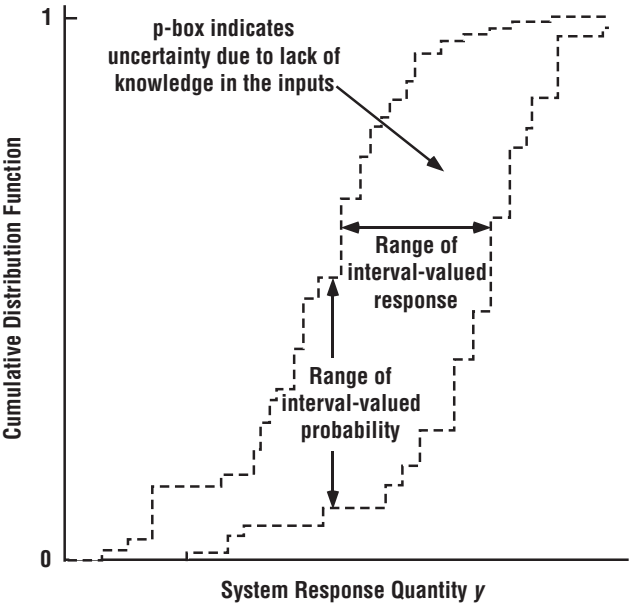


Figure 13.18 Example of p-box obtained for a system response quantity with large epistemic uncertainty.

As a result, this type of mathematical structure is sometimes called an *imprecise probability function*. However, that term suggests the probabilities are vague or *fuzzy* in some sense; this would give the wrong interpretation. The probabilities shown in a p-box are *as precise, and the bounds are as small, as can be stated given the information that is claimed for the input quantities*.

When a decision maker is presented with information that segregates aleatory and epistemic uncertainty, as in Figure 13.18, more enlightened and better decisions or actions can be taken. For example, if the system response is dominated by epistemic uncertainty, then the decision maker must add more knowledge, or make restrictions on the bounds of epistemic uncertainties, in order to reduce the response uncertainty. If one had conducted the same Monte Carlo analysis, but assumed each of the epistemic uncertainties were represented as a uniform distribution, then one would have obtained a plot that would have *one* CDF near the center of the p-box shown in Figure 13.18. This would have presented a *very* different picture of uncertainty to the decision maker. The representation of the interval as a uniform distribution would have had two unjustified changes in a statement of knowledge: (a) the quantity is a random variable instead of a quantity that has a unique, but unknown value; and (b) all possible values of the unknown quantity are *equally likely*.

In Figure 13.18 one should note the relatively distinct changes in trends that can occur in both the upper and lower probability curves. These trend changes are common in the boundaries of the p-boxes because the boundaries represent the minimum and maximum of the ensemble of *all possible CDFs* for the system. For example, in one region of the system

response a particular CDF could represent the maximum, but then at a slightly different response, a different CDF could become the maximum. Stated differently, the p-box boundaries typically have several trade-offs between individually realizable, or possible, CDFs. If one or two epistemically uncertain input quantities dominate the epistemic uncertainty in the response, then there is less chance for these types of trade-off.

A final comment should be made concerning an unintended benefit of MCS or LHS. When random samples are drawn over a wide range of each $\{x_1, x_2, \ldots, x_m\}$, there will be a number of unusual combinations of the $\{x_1, x_2, \ldots, x_m\}$. That is, there will be combinations of $\{x_1, x_2, \ldots, x_m\}$ that no one would ever think about using in a simulation of a system because they would not normally occur together, or they may even be physically impossible. What a number of UQ analysts have found is that when these unusual combinations on inputs are attempted in the code, the code will "crash." When these crashes are investigated, it is commonly found that they were caused by bugs in the code that had not been found in any of the SQE testing. This, of course, is similar to many code developers' experience: "If you want to find bugs in your code, let a new user run it."

### *13.4.2 Combination of input, model, and numerical uncertainty*

How to combine input, model, and numerical solution uncertainties is an open research topic subject to considerable debate. In fact, many researchers and risk analysts either: (a) ignore the quantification of model and numerical uncertainties; (b) ignore the issue of how to combine input, model, and numerical uncertainties because it is such a difficult and controversial issue; or (c) avoid directly dealing with the issue because they use model updating of input and model parameters to attain good agreement between the model and measurements, regardless of model and numerical uncertainties. We, on the other hand, have continually stressed the importance of directly dealing with each of these uncertainties in the prediction of a system response.

Although model and numerical uncertainty are very different kinds of beast, they are both frustrating because we always believe that with more sophisticated physics models and bigger computers we can eliminate them. Often, project leaders and decision makers do not have the luxury of waiting for model and numerical improvements to be made, but must make decisions and move on. The fact of the matter is that in the simulation of complex systems model uncertainty is commonly the *dominant* uncertainty in risk-informed decision-making.

We present two methods for combining model and input uncertainty. The first method uses validation metrics, discussed in Chapter 12, to estimate model uncertainty. The estimate of model uncertainty is then combined with input uncertainty using a method based on recent work by Oberkampf and Ferson (2007) and Ferson *et al.* (2008). The second method is based on using alternative plausible models for the system of interest to try to quantify the combination of model and input uncertainty (Morgan and Henrion, 1990; Cullen and Frey, 1999; Helton *et al.*, 2000; Helton, 2003; NRC, 2009). This method has, of course, been

around for decades, but risk analysts rarely use it because it is expensive and time consuming to develop multiple models of the system of interest, as well as computing the additional simulations from these models. Only in large-scale risk assessments of high-consequence systems are alternative plausible models seriously investigated.

A method for including numerical solution uncertainty will be discussed in the final subsection. It can be applied to either of the methods for combining model and input uncertainty.

#### *13.4.2.1 Combination of input and model uncertainty*

Model uncertainty is fundamentally due to lack of predictive knowledge in the model, so it should be represented as an epistemic uncertainty. If the validation metric is characterized in terms of the physical units of the SRQ of interest, we argue that the most defensible way to combine model and input uncertainty is to add the model uncertainty to the p-box representing the output uncertainty. By *add*, we mean increase the lateral extent of the output p-box (resulting from mapping the aleatory and epistemic input uncertainties to the output) by the amount of the estimated model uncertainty. That is, for every value of cumulative probability, the model uncertainty would be subtracted from the left side of the p-box, and/or added to the right side of the p-box, depending on the sign of the estimated model uncertainty. If the input uncertainty is only aleatory, then the output uncertainty will be a single CDF, i.e., the degenerate case of a p-box. The subtraction from the left side and addition to the right side can only be made if the validation metric is given in terms of the SRQ.

By expanding the p-box on the left and right, we are treating model uncertainty as an interval-valued quantity. The approach is directly equivalent to the addition of simple intervals, and *no assumption* is made concerning dependence between the various sources producing the intervals. Both validation metric approaches developed in Chapter 12, the confidence interval approach and the area metric, can be used in this way because they are both error measures in terms of the dimensional units of the SRQ. If hypothesis testing is used for validation, then one does not have an error measure in terms of the SRQs, but a probability measure indicating a likelihood of agreement between computation and experiment. If Bayesian updating is used for validation, the model form uncertainty is either assumed to be zero, or it is estimated in combination with updating the distributions for the input and model parameters. Even if the latter option is used, updating parameter distributions becomes inextricably convolved with estimating model uncertainty because they are computed jointly.

To apply either the confidence interval approach or the area metric, we need to quantify the model uncertainty at the conditions where the model is applied, i.e., the application condition of interest. That condition, or set of conditions, may be inside or outside the validation domain. If the application condition is inside the validation domain, one can think of the validation metric function as an interpolating function. If the application condition is outside the validation domain, one is extrapolating the validation metric function to the

condition of interest. In this section, we will only consider the combination of input and model uncertainty where the model uncertainty is estimated using the confidence interval approach. In Section 13.7.3.1, we will consider the approach for combining input and model uncertainty where model uncertainty is estimated by the area metric.

The confidence interval approach is rather simple for three reasons. First, it only assesses model uncertainty over a single input or control parameter. If there are additional inputs to the model, the model uncertainty is assumed constant with respect to all other inputs. Second, the model uncertainty is based on taking the difference between the mean value of the experimental measurements and the simulation results. And third, it automatically constructs a validation metric function as part of the approach. These simplifications, particularly the first and second, also restrict the applicability of the approach.

Referring to Sections 12.4 through 12.7 of Chapter 12, we have

$$d(x) = \tilde{E}(x) \pm \mathrm{CI}(x). \tag{13.38}$$

$d(x)$ is the validation metric function, $\tilde{E}(x) = \bar{y}_{\mathrm{m}}(x) - \bar{y}_{\mathrm{e}}(x)$ is the estimated mean of the model error, $\bar{y}_{\mathrm{m}}(x)$ is the mean of the model prediction, $\bar{y}_{\mathrm{e}}(x)$ is the mean of the experimental measurements, and $\mathrm{CI}(x)$ is the confidence interval of the experimental data. $\mathrm{CI}(x)$ is defined with respect to the mean of the experimental data, $\bar{y}_{\mathrm{e}}(x)$. Equation (13.38) can written as an interval at the application point of interest $x^*$.

$$\left(\tilde{E}(x^*) - \mathrm{CI}(x^*),\, \tilde{E}(x^*) + \mathrm{CI}(x^*)\right). \tag{13.39}$$

Note that even if the model matches the experimental data perfectly, that is, $\bar{y}_{\mathrm{m}}(x^*) = \bar{y}_{\mathrm{e}}(x^*)$, $d(x)$ is given by the interval $(-\mathrm{CI}(x^*), +\mathrm{CI}(x^*))$.

For the nonlinear regression case, Section 12.7, it was found that the confidence intervals are not symmetric with respect to $\bar{y}_{\mathrm{m}}(x)$. For this case we would simply average the upper and lower confidence intervals to obtain an average value so that it could be used in Eq. (13.39). The confidence level used in computing the confidence interval would be chosen at the discretion of the analysts and/or the needs of the customer using the simulation results. Typically, 90% or 95% confidence levels are chosen. Note that the magnitude of $\mathrm{CI}(x)$ grows rapidly as the confidence level increases beyond 90%.

The confidence interval approach has a great deal of flexibility concerning the regression function one wishes to use to represent the mean of the experimental data, $\bar{y}_{\mathrm{e}}(x)$. The model prediction, $\bar{y}_{\mathrm{m}}(x^*)$, can either be interpolated, if sufficient data are available, or computed from the model. Let $x_{\mathrm{l}}$ be the lowest value of the validation data and let $x_{\mathrm{u}}$ be the upper value of the data. If $x_{\mathrm{l}} \leq x^* \leq x_{\mathrm{u}}$, we can interpolate the experimental data using the regression function for $\bar{y}_{\mathrm{e}}(x^*)$ and compute $\mathrm{CI}(x^*)$. We will consider this case first.

Consider how Eq. (13.39) is combined with the p-box representing the uncertainty in the predicted SRQ, where the uncertainty is only due to uncertain inputs. To explain the concept, we will simplify the p-box to a continuous CDF for the SRQ. The concept, however, applies equally well to (a) a p-box that is due to epistemic uncertainty in the input,

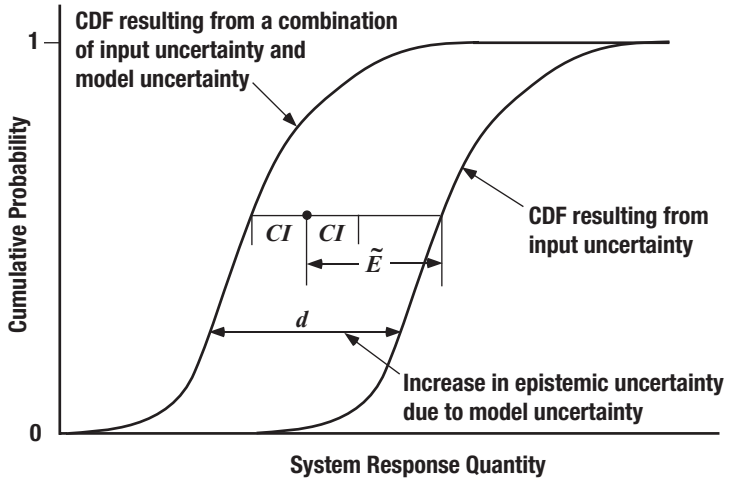


Figure 13.19 Method of increasing the uncertainty on the left of the SRQ distribution due to model uncertainty.

and (b) an empirical distribution function (EDF) that is constructed from a limited number of model evaluation samples.

Figure 13.19 shows how Eq. (13.39) is used to expand the uncertainty to the left of the CDF for the case of $\tilde{E}(x^*) > 0$. The increase due to model uncertainty is not only due to estimated model error, but also due to uncertainty in the experimental data. As a result, the total left displacement of the CDF is

$$\tilde{E}(x^*) + \mathrm{CI}(x^*). \tag{13.40}$$

If

$$\tilde{E}(x^*) + \mathrm{CI}(x) < 0, \tag{13.41}$$

then the left displacement is zero because epistemic uncertainty cannot be negative. The magnitude of the displacement is a constant over the complete range of the CDF, i.e., for all responses of the system at $x^*$. As can be seen from Figure 13.19, even if the response of the system is purely aleatory, due to purely aleatory input uncertainties, the response from the combined input and model uncertainty is a p-box. The p-box can be correctly interpreted in two different ways. First, for a fixed system response anywhere along the distribution, the combined uncertainty is now an interval-valued probability. Second, for a fixed cumulative probability, the combined uncertainty is now an interval-valued response.

A similar development can be shown for the right displacement of the CDF. The equations for the left and right displacement of the CDF, or a p-box if the input uncertainty contains

both aleatory and epistemic uncertainties, can be shown to be

$$d_{\text{left}} = \begin{cases} [\bar{y}_{\text{m}}(x^*)]_{\text{left}} - \bar{y}_{\text{e}}(x^*) + \text{CI}(x^*) & \text{if} \quad [\bar{y}_{\text{m}}(x^*)]_{\text{left}} - \bar{y}_{\text{e}}(x^*) + \text{CI}(x^*) \geq 0, \\ 0 & \text{if} \quad [\bar{y}_{\text{m}}(x^*)]_{\text{left}} - \bar{y}_{\text{e}}(x^*) + \text{CI}(x^*) < 0, \end{cases} \tag{13.42}$$

$$d_{\text{right}} = \begin{cases} \left|[\bar{y}_{\text{m}}(x^*)]_{\text{right}} - \bar{y}_{\text{e}}(x^*) - \text{CI}(x^*)\right| & \text{if} \quad [\bar{y}_{\text{m}}(x^*)]_{\text{right}} - \bar{y}_{\text{e}}(x^*) - \text{CI}(x^*) \leq 0, \\ 0 & \text{if} \quad [\bar{y}_{\text{m}}(x^*)]_{\text{right}} - \bar{y}_{\text{e}}(x^*) - \text{CI}(x^*) > 0. \end{cases} \tag{13.43}$$

$[\bar{y}_{\text{m}}(x^*)]_{\text{left}}$ is the mean of the predicted SRQ from the left boundary of the p-box and $[\bar{y}_{\text{m}}(x^*)]_{\text{right}}$ is the mean from the right boundary of the p-box. If the model is over-predicting the experiment, then the p-box is increased more on the left than the right. If the model is under-predicting the experiment, then the p-box is increased more on the right than the left. It can be seen from these equations that the increase in the lateral extent of the system response p-box is only symmetric left to right when both of the following equations are true:

$$\begin{aligned} & \left[\bar{y}_{\text{m}}\left(x^*\right)\right]_{\text{left}} = \left[\bar{y}_{\text{m}}\left(x^*\right)\right]_{\text{right}} \\ \text{and} \quad & \left[\bar{y}_{\text{m}}\left(x^*\right)\right]_{\text{left}} = \bar{y}_{\text{e}}\left(x^*\right). \end{aligned} \tag{13.44}$$

In this case, the increase on the left and right are equal to the magnitude of the confidence interval of the experimental data.

If one must extrapolate $d(x)$ outside of the validation domain to attain the application condition of interest, one should generally not extrapolate the function $d(x)$ itself. It is not recommended because $d(x)$ will display more complex features than the three individual functions $\bar{y}_{\text{m}}(x)$, $\bar{y}_{\text{e}}(x)$, and $\text{CI}(x)$. It is recommended that a first or second degree polynomial be used in a least squares fit for $\bar{y}_{\text{e}}(x)$ and $\text{CI}(x)$. Low-degree polynomials may not capture the detailed features of $\bar{y}_{\text{e}}(x)$ and $\text{CI}(x)$, but they should be more reliable in extrapolation because they would, in general, have fewer degrees of freedom than $\bar{y}_{\text{e}}(x)$ and $\text{CI}(x)$. Note that the extrapolation of $\text{CI}(x)$ only involves the extrapolation of one function, since the confidence intervals are symmetric around $\bar{y}_{\text{e}}(x)$. One method of improving the least squares fit of the low degree polynomials and capturing a trend that is important for extrapolation is to fit only a portion of the range of the experimental data. One should not use a regression function for extrapolating the model prediction $\bar{y}_{\text{m}}(x^*)$, but simply calculate the value using the model. After the low-order polynomials are computed for extrapolating $\bar{y}_{\text{e}}(x)$ and $\text{CI}(x)$, and $\bar{y}_{\text{m}}(x^*)$ is evaluated, Eqs. (13.42) and (13.43) can be used to calculate $d_{\text{left}}$ and $d_{\text{right}}$.

It should be pointed out that when extrapolation of the model is required, as is commonly the case, the estimate for the *model uncertainty* as presented here is a regression-based extrapolation. That is, the accuracy of the extrapolation of the model uncertainty does *not* depend on the accuracy of the model, but on the accuracy of the extrapolation of the observed uncertainty in the model. The accuracy of the prediction from the model, however,

*is* based on the fidelity of the physics and the soundness of the assumptions incorporated into the model, i.e., its predictive capability.

An uncertainty extrapolation procedure that is not regression based has recently been proposed by Rutherford (2008). The procedure uses the concept of a non-Euclidean space for the input quantities in order to predict the uncertainty in the system responses. The method is intriguing because it does not rely on the concept that the input quantities are simply parameters, i.e., continuous variables, in the extrapolation of the uncertainty structure of the model.

#### *13.4.2.2 Estimation of model uncertainty using alternative plausible models*

When experimental data is sparse over the validation domain, or no data exists for closely related systems under similar conditions, large extrapolations of the model are required. The second approach for estimating model uncertainty is to compare predictions from alternative plausible models. This method is also referred to as the *method of competing models*. The approach is simple, but it is not commonly used because of the time and expense of developing multiple models for a system. There are two important cases where large extrapolations of the model are required, and this method should be used instead of the method for extrapolating the validation metric result. First, models that must predict complex processes far into the future must deal with extraordinary extrapolations. Two timely examples are underground storage of nuclear waste and global climate-change modeling. Models for these systems attempt to make predictions for hundreds and thousands of years into the future for physical processes that are very poorly understood, and the surrounds (e.g., BCs) are even less known. Second, there are model predictions that are needed for systems, particularly failure modes of systems or event trees, which cannot be tested. Some examples are large-scale failure scenarios of nuclear power plants, aging and failure of a full-scale dam, explosive eruption and surrounding damage from a volcano, and a chemical or radiological terrorist attack. Extrapolation of models for these situations can be thought of in terms of the validation hierarchy, but it is of limited quantitative value. For example, certain relevant experiments can be conducted on subsystems, scale models of systems, or surrogate systems in the validation hierarchy. However, the complete system cannot be tested at all, or cannot be tested under relevant conditions. As a result, it is inappropriate and misleading to use the mental model of extrapolation in terms of parameters describing the system or the surroundings.

To use the approach of alternative plausible models, one should have at least two independently developed models of the system and then a comparison is made of the prediction of the same SRQs of interest from each model. In this approach, *none* of the models are presumed to be correct; each one is simply a postulated representation of reality. The only presumption is that each of the models is reasonable and scientifically defensible for the task at hand. Some models may present strong arguments that they are more reliable than others. For example, there has been recent work in the area referred to as hierarchical, or multiscale, physics modeling (Berendsen, 2007; Bucalem and Bathe, 2008; Steinhauser,

2008). If strongly convincing arguments can be made for having higher confidence in the results of some models over others, then the higher confidence models might be used to calibrate the lower level models over some range of conditions. However, if such arguments cannot be decisively made for each aspect of the higher confidence model (physics fidelity, system input data, surroundings input data, and VV&UQ), then each of the models should be treated as having equal confidence. Predictive accuracy assessment of each model may have been done with different sets of experimental data, or possibly even the same set of data. Each model may have used parameter estimation or calibration of model parameters. These details are not critically important.

The most important issue in using alternate models is the *independence* of the teams and the formulations of the models, including assumptions dealing with the surroundings, scenarios, event trees, fault trees, and possible human interaction with the system. In comparing results from each team, it is very often found that each team will have thought of important aspects affecting the modeling and the results that the other team did not think of. Then analyses can be reformulated, improved, and possibly corrected for further comparisons. For example, investigating why two models do not predict similar results for a simplified scenario can help identify blind epistemic uncertainties such as user input and output errors, and coding errors.

This approach does not actually provide an estimate of model uncertainty. It only provides an indication of the similar or dissimilar nature of each model prediction. In the prediction of hurricane or typhoon paths it is a very welcome sight for the news media to show the multiple paths predicted by the various hurricane models. Sometimes these are call *spaghetti plots* of hurricane paths. The results of each of the models should be considered by the decision maker; not averaged or combined in any way. Some argue that the results from each of the models should be considered as equally likely, and treated as a random variable. This would be treating model uncertainty as an aleatory uncertainty, as if physics modeling error is a random process. This forces the square peg of physics modeling into the round hole of probability theory.

Comparing alternative models should be thought of as a sensitivity analysis with respect to model uncertainty, as opposed to an estimation of model uncertainty. Our experience, and the experience of others who have used it, have found it is extremely revealing; often distressing. When results from alternative models are shown together for the first time, particularly from analysts who have been working independently, there is usually a significant *surprise index* (Morgan and Henrion, 1990; Cullen and Frey, 1999). This always generates a great deal of discussion and debate concerning model assumptions, experimental data used for comparisons, and model calibration procedures. This interaction is beneficial and constructive to improving all of the models. On the second iteration of comparing alternative models, there is typically more agreement between models, but not always. Regardless of the level of agreement, we argue that all the results should be presented to the decision makers for their consideration. Given the uncertainty from all sources, the decision makers and managers are responsible for deciding future activities. These activities could be, for example: (a) making system design changes so that the system can successfully

tolerate large uncertainties, (b) restricting the operating envelope of the application domain so that the system will not be exposed to unacceptable uncertainties, or (c) deciding to obtain more experimental data or improve the physics modeling fidelity to reduce the predicted uncertainty.

#### 13.4.2.3 Inclusion of numerical solution uncertainty

As discussed earlier in this chapter, we will represent the estimated numerical solution error as an epistemic uncertainty, $U_N$. Repeating Eq. (13.24) here,

$$(U_N)_{y_i} = |U_I|_{y_i} + |U_S|_{y_i} + |U_T|_{y_i} \quad \text{for } i = 1, 2, \ldots, n, \tag{13.45}$$

where $U_I$ is the iterative solution uncertainty, $U_S$ is the spatial discretization uncertainty, and $U_T$ is the temporal discretization uncertainty. Analogous to the method for increasing the p-box due to model uncertainty, we use each $(U_N)_{y_i}$ computed to expand the width of the p-box for the particular $y_i$ under consideration. Since there is no reference condition, such as experimental measurements in model validation, $U_N$ equally expands the left and right boundaries of the SRQ p-box. That is, the left boundary is shifted to the left by $U_N$, and the right boundary is shifted to the right by $U_N$. This procedure for including numerical solution uncertainty can be applied to both procedures for estimating model uncertainty: the method using validation metrics and the method using alternative plausible models.

Some would argue that the increase in uncertainty due to $U_N$ is generally so small compared to model uncertainty that it should be neglected. If one properly quantifies $U_N$, and it is indeed much smaller than $d$, then this can certainly be done. Our experience is that $U_N$ is not quantified in most simulations, but it is dismissed with claims like "We only use high order accuracy methods, so the numerical error is small." Or "We have computed this solution with many more finite elements than is usually done, so this is much more accurate than past simulations." Or "The agreement with experimental data is excellent, why are you being difficult?" For a simulation to be scientifically defensible, the numerical solution error must be quantified for a number of representative solutions of all the SRQs of interest; preferably solutions that are judged to be the most numerically challenging.

### 13.4.3 Example problem: heat transfer through a solid

Here, we continue with the development of the example problem discussed in Sections 13.1.3, 13.2.3, and 13.3.4. We are now interested in predicting the heat transfer through the west face for the system of interest. As described in Section 13.1.3, the two differences between the validation experiment and the system are (a) in the validation experiment $T_E$ never exceeded 390 K, while in the system of interest attains an east face temperature of 450 K, and (b) in the validation experiment the north face was cooled such that the convective heat transfer coefficient $h$ was set at the middle of the possible range that could exist for the system of interest. The prediction discussed here will include the increase in uncertainty due to these two differences.

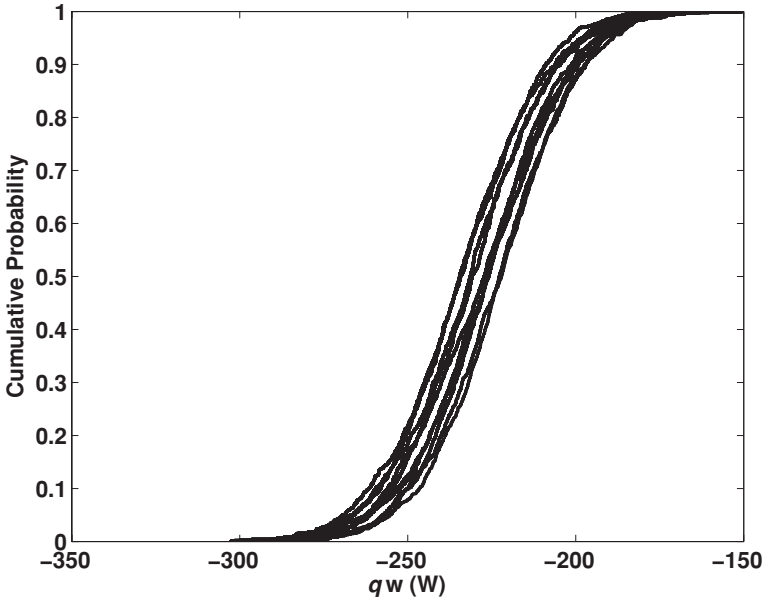


Figure 13.20 Input uncertainty in $q_W$ as represented by ten CDFs from sampling $h$, where each CDF is composed of 1000 MC samples from $k$.

### *13.4.3.1 Input uncertainties*

The two input uncertainties are the thermal conductivity in the plate $k$ and the convective heat transfer coefficient on the north face of the plate $h$. The uncertainty in $k$ is a pure aleatory uncertainty and is given by Eq. (13.7), and the uncertainty in $h$ is a pure epistemic uncertainty and is given by Eq. (13.11). We propagate the uncertainty in $k$ using MCS and the uncertainty in $h$ using LHS, according to the method described in Section 13.4.1. We use ten samples and ten subintervals for $h$, and 1000 samples of $k$ for each sample of $h$. Although the resulting 10 000 samples may be excessive for many analyses, we use this number here to give well-converged results. According to Table 13.1, the remaining BCs for the system are $T_E = 450$ K, $T_W = 300$ K, $q_S = 0$.

Figure 13.20 shows the ten individual CDFs for $q_W$ from the outer loop sampling of $h$. Each CDF results from 1000 MC samples and represents the aleatory uncertainty in $k$, given a sampled realization of the epistemically uncertain $h$. It should be stressed that each one of the CDFs represent the variability in $q_W$ that could occur, given the poor state of knowledge of the epistemic uncertainty $h$. Stated differently, admitting that we only can give an interval for $h$, there is *no more precise statement of uncertainty* that can be made concerning $q_W$ than is represented by the ensemble of all of the CDFs. All of the heat flux predictions shown in Figure 13.20 are negative, meaning that the heat flux is out of the west face of the system being analyzed and into the adjacent system. Note that some of the CDFs cross one another. The likelihood of CDFs crossing depends on the nonlinearity of the SRQ as a function of the epistemic uncertainties in the input.

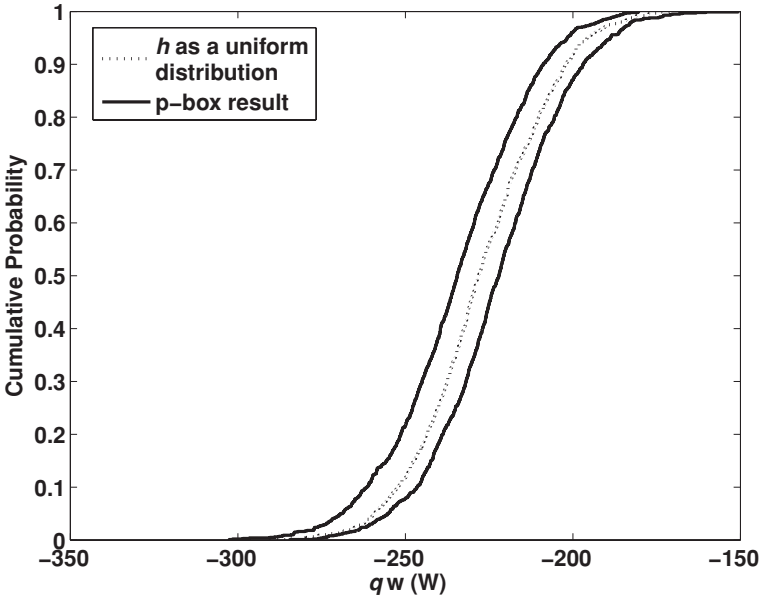


Figure 13.21 p-box for $q_W$ due to input uncertainty.

The p-box for $q_W$ is found by computing the envelope of all of the CDFs shown in Figure 13.20. That is, for each value of $q_W$ computed from both inner and outer sampling, the minimum and maximum of all of the CDFs is found. These minima and maxima form the p-box of the SRQ. Figure 13.21 shows the p-box for $q_W$, due to input uncertainty, as the solid lines. From a system design perspective, the area of interest in the figure is the region of large negative values of heat flux because this flux could possibly damage the system adjacent to the west face of the system of interest. As expected, these heat fluxes have low values of probability. By noting the width of the p-box, one can assess the magnitude of the uncertain response that is due to the epistemic uncertainty in $h$. For example, if one were interested in the range of heat flux that could occur for the cumulative probability of 0.1, one would have the interval $[-247, -262]$ W. Similarly, if one were interested in the range of probabilities that could occur if $q_W = -250$ W, one would have an uncertainty given by the probability interval [0.08, 0.22].

Also plotted in Figure 13.21 is the CDF that would be obtained if $h$ were treated as an aleatory uncertainty instead of an epistemic uncertainty. That is, if the interval-valued characterization of $h$ were replaced by a uniform distribution over the interval, one would obtain the dotted CDF shown inside the p-box. It is clear that the characterization of the uncertainty in $q_W$ would be greatly reduced as compared to the p-box representation. For example, the dotted curve shows that the probability of attaining a heat flux of $-250$ W (or greater in the absolute sense) is 0.12. This interpretation underestimates the uncertainty in $q_W$ due to $h$ and would be misleading to project managers, decision makers, and any other customers of the analysis.

*13.4.3.2 Combination of input, model, and numerical uncertainties*

We first compute the increase in the width of the input uncertainty p-box that is due to model uncertainty. Because the condition of interest, $T_E = 450$ K, is beyond the highest temperature in the validation experiments, 390 K, extrapolation of $\bar{y}_e(x)$, CI($x$), and $\bar{y}_m(x)$ is required. As discussed in Section 13.4.2.1, we use low-order polynomials for the extrapolation of $\bar{y}_e(x)$ and CI($x$). We use a linear regression for $\bar{y}_e(x)$, as given by Eq. (13.17). We could directly use the Scheffé confidence interval as given by Eq. (13.18) for the extrapolation. However, here we use the more general procedure of computing a low degree polynomial fit of the confidence interval. Using Eq. (13.18) to generate the data for the regression fit over the range of $330 \leq T_E \leq 390$ K, we compute the following quadratic fit:

$$\mathrm{CI}(x) = 264.6 - 1.452T_E + 0.002017T_E^2. \tag{13.46}$$

The sampling results discussed above yield $[\bar{y}_m\,(450)]_{\text{left}} = -234.7$ W and $[\bar{y}_m\,(450)]_{\text{right}} = -222.1$ W for the left and right median values, respectively, of the p-box shown in Figure 13.21. Substituting the appropriate values into Eqs. (13.42) and (13.43), we have

$$\begin{aligned} &\text{since} && -234.7 - (-259.9) + 19.5 = 25.2 + 19.5 = 44.7 \geq 0, \\ &\text{we obtain} && d_{\text{left}}(450) = 44.7 \text{ W}, \end{aligned} \tag{13.47}$$

and

$$\begin{aligned} &\text{since} && -222.1 - (-259.9) - 19.5 = 37.8 - 19.5 = 18.3 > 0, \\ &\text{we obtain} && d_{\text{right}}(450) = 0 \text{ W}, \end{aligned} \tag{13.48}$$

respectively. From these equations we see that the model is over-predicting the extrapolated experimental mean (although the model result is smaller in absolute value) by 25.2 and 37.8 W for the left and right bound of the p-box, respectively. For the left bound, the model bias error combines with the extrapolated experimental uncertainty of 19.5 W to yield a total model uncertainty shift to the left of 44.7 W. It can be seen that 44% of the leftward shift of the CDF is due to uncertainty in the experimental measurements. If one required a model prediction for a higher temperature, say $T_E = 500$ K, the percentage of the model uncertainty due to measurement uncertainty would increase to roughly 63%. For the right bound, it is seen that the over prediction of the model is larger than the extrapolated experimental uncertainty, resulting in no rightward shift of the CDF due to model uncertainty.

Now consider the additional increase of width of the p-box due to numerical solution error. As discussed earlier, we treat these estimates of error as epistemic uncertainties. In Section 13.3.4.2, we computed $U_I = -0.00173$ W and $U_S = 0.08448$ W for the iterative and discretization uncertainty, respectively. Substituting these results into Eq. (13.45), we have

$$U_N = |-0.00173| + |0.08448| = 0.08621 \text{ W}. \tag{13.49}$$

It is seen that $U_N$ is two orders of magnitude smaller than $d_{\text{left}}$ (450). As a result, it will be neglected in the final estimate of predictive uncertainty for the heat transfer simulation.

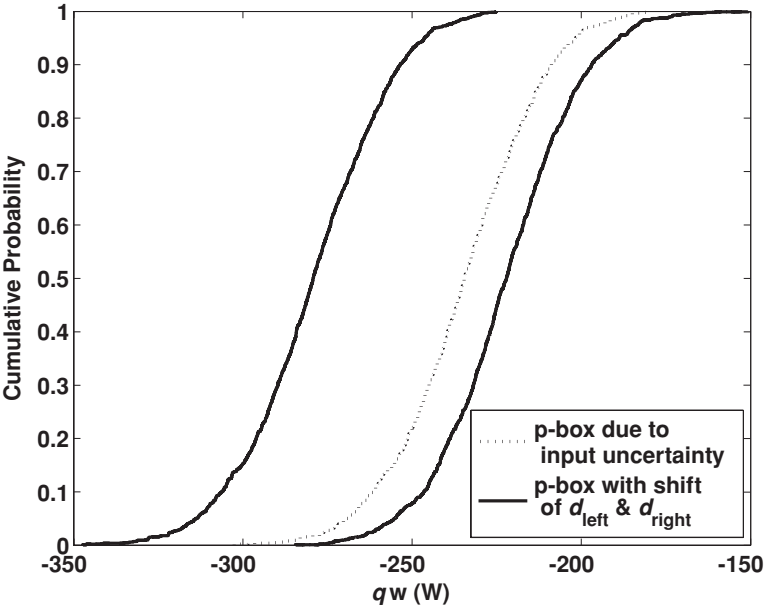


Figure 13.22 p-box for $q_W$ due to the combined input and model uncertainty.

Figure 13.22 shows the final p-box for the combination of input and model uncertainty for $q_W$. The total uncertainty was obtained by combining the input uncertainty shown in Figure 13.21 and the model uncertainty given by $d_{left}(450) = 44.7$ W and $d_{right}(450) = 0$ W. It is obvious that the estimated model uncertainty contributes significant additional uncertainty to the predicted response. This increase is due solely to the leftward shift of the input p-box and it is considered a constant over the entire distribution. The magnitude of the increase is due to a nearly equal combination of model bias error and experimental measurement uncertainty. The question could be raised whether the experimental uncertainty should be separately accounted for in the estimate of model bias error. We strongly argue that it should be included because ignoring it would deny the stochastic nature of experimental measurements, as well as the epistemic uncertainty due to obtaining limited experimental samples. In addition, the experimental uncertainty contribution should increase as a function of the magnitude of the extrapolation.

The impact of including model uncertainty can be viewed in two ways. First, for any value of cumulative probability the uncertainty in the system response is increased by a constant amount; $d_{left}(450) + d_{right}(450)$. Second, the predicted probability interval for a fixed system response can significantly increase depending on what level of system response is of interest. The following two examples capture this impact. First, consider the increase in uncertainty for a heat flux in the mid-range of the predicted response. From Figure 13.21 it can be seen that for $q_W = -230$ W the interval-valued probability is [0.34, 0.59]. However, when model uncertainty is included (see Figure 13.22) the interval-valued probability increases to [0.34, 0.995]. Second, consider the increase in uncertainty for a

large absolute value heat flux that is of concern to over-all system safety, reliability, and performance, say $q_W = -300$ W. The probability interval has increased to [∼0., 0.153] when model uncertainty is included, as compared to [∼0., 0.0025] for input uncertainty alone. Both of these examples are a striking demonstration of the increase in uncertainty when model uncertainty is included.

## 13.5 Step 5: conduct model updating

Model updating can take many different forms. For example, model updating could involve significant reformulation in scenarios, event trees, failure modes, changes in how the system interacts with the surroundings, modeling assumptions, replacement of certain submodels that are performing poorly, and updating of model parameters. Although it is common practice to conduct different types of model updating, here we will only deal with updating of parameters that occur *within* models. That is, the model form or mathematical model structure does not change, but the parameters that occur in the model are altered based on various types of information. These parameters could be part of the mathematical model for the surroundings (such as boundary conditions and system excitation) or the system (such as geometry, initial conditions, and material properties). The parameter can be either a deterministic value, such as a scalar determined by some optimization process, or a nondeterministic value, such as a quantity given by a probability distribution. In addition, the parameter can be, in some circumstances, a scalar field, a vector, or a tensor field.

Updating of parameters in models, and submodels, is often a necessary activity in essentially any simulation. Some readers may feel we have been unfairly harsh on the activity of parameter updating in Chapter 12. Our criticisms have been primarily directed at the detrimental, and dangerous, concept of thinking that updating (or estimating or calibrating) parameters in models is fundamentally the same as model validation. Similarly, we are concerned with the erroneous belief that all types of parameter updating have a similar effect on predictive capability. In this section, we fully recognize the importance and necessity of updating parameters in models. However, we will argue that it is important to keep the concepts of parameter updating and model validation separate. As discussed in Section 13.2, the goal of parameter updating is to improve the agreement between simulation and experiment, whereas the goal of model validation is model accuracy assessment. As will be pointed out in this section, we agree there are shades of gray, or overlap, between parameter updating and validation of a model. We will discuss the extremes of updating and validation and the areas in between to try and clarify the varying effects of updating on predictive capability.

### *13.5.1 Types of model parameter*

To better understand the various approaches to parameter updating, it is appropriate to first discuss the various classes of parameters that occur in simulation activities, a topic that has not been well studied. One would think that theorists in modeling and simulation would

have more carefully examined the issue of parameter classes, but, to our knowledge, this is not the case. Morgan and Henrion (1990) discuss several classes of parameter that are quite helpful for risk assessment and uncertainty quantification. We use their basic classification and we segregate certain important classes into additional classes. For simulation activities, we divide model parameters into the following six classes:

- measurable properties of the system or the surroundings,
- physical modeling parameters,
- ad hoc model parameters,
- numerical algorithm parameters,
- decision parameters,
- uncertainty modeling parameters.

Each of these will be briefly discussed, along with the type of information that is commonly available for updating each class of parameters.

*Measurable properties* of the system or surroundings are physical quantities or characteristics that are directly measurable, at least in principle, either now or at some time in the past or future. Specifically, these are quantities that are measured separately from complex models associated with the system of interest. Metrologists point out that empirical quantities always have some type of conceptual or mathematical model associated with the quantity to be measured (Rabinovich, 2005). Measurable properties of the system or surrounding generally rely on either (a) a simple, well accepted, model of the physical nature of the quantity; or (b) some type of well-understood performance or reliability characteristic of the system or the surroundings. Several examples of system properties in this class are: Young's modulus, tensile strength, mass density, electrical conductivity, dielectric constant, thermal conductivity, specific heat, melting point, surface energy, chemical reaction rate, thrust of an engine, thermal efficiency of a power plant, the failure rate of a valve, and age distribution of subsystems within a system. Examples of properties of surroundings that are in this class are: wind loading on a structure, external heating on a flight vehicle, atmospheric characteristics surrounding a nuclear power plant that could exist after a serious accident, electrical and magnetic characteristics of a lightning strike on a system, and physical or electronic attack on a system.

*Physical modeling parameters* are those that are *not* measurable outside of the context of the mathematical model of the system under consideration. These quantities are physically meaningful in the sense that they are associated with some type of well-accepted physical process occurring in the system. However, because of the complexity of the process, no separate mathematical model exists for the process. Stated differently, some type of basic physical process is known to occur in the system, but because of the coupling of the process, the physical effect of the process is combined into a simple model that only exists within the framework of the complex model. Quantifying these parameters relies on inference within the context of the assumptions and formulation of the complex model. Examples of these types of parameter are (a) internal dynamic damping in a material, (b) aerodynamic damping of a structure, (c) damping and stiffness of assembled joints in a multi-element structure, (d) effective chemical reaction rate in turbulent reacting flow, (e) effective surface

roughness in turbulent flow along a surface, and (f) thermal and electrical contact resistance between material interfaces.

*Ad hoc model parameters* are those that are introduced in models, or submodels, simply to provide a method for adjusting the results of a model to obtain better agreement with experimental data. These parameters have little or no physical justification in the sense that they do not characterize some feature of a physical process. They exist solely within the context of the model or submodel in which they are used. Some examples are (a) most parameters in fluid dynamic turbulence models, (b) most parameters in models for strain hardening in plastic deformation of materials, (c) most parameters in models for material fatigue, and (d) parameters inserted into a complex model that are used solely to adjust model predictions to obtain agreement with experimental data.

*Numerical algorithm parameters* are those that exist within the context of a numerical solution method that can be adjusted to meet the needs of a particular solution. These parameters typically have a recommended range of values, but they can be changed to give better performance or reliability of the numerical method. Here, we *do not mean* discretization levels or iterative convergence levels, but quantities that typically appear in the formulation or execution of a numerical algorithm. Some examples are (a) the relaxation factor in the successive over-relaxation method for the iterative solution of algebraic equations, (b) the number of iterations used on each level of a multi-grid method, (c) artificial damping or dissipation parameters in numerical methods, and (d) parameters to control the hour-glassing phenomena in finite element models for solids.

*Decision parameters* are those that exist within an analysis whose value is determined or controlled by the analyst, designer, or decision maker. Sometime these are referred to as design variables, control variables, or policy decision parameters. The question of whether a quantity is considered as a decision parameter or, say, a measurable property of the system, depends on the context and intent of the analysis. Decision parameters are commonly used in design to optimize the performance, safety, or reliability of a system. In the analysis of abnormal or hostile environments, these parameters can be varied to determine the most vulnerable conditions that could affect the safety or security of the system. These parameters are also used, for example, in models constructed for aiding in public policy decisions, public health regulations, or environmental impact analyses.

*Uncertainty modeling parameters* are those that are defined only in the context of characterizing the uncertainty of a quantity or a model. Uncertainty modeling parameters, simply referred to as *uncertainty parameters* from here on, are typically parameters of a family of distributions, e.g., the four parameters defined as part of a beta distribution. Uncertainty parameters can be a point value, an interval, or a random variable. As a result, an uncertainty parameter can be either a number or an aleatory or epistemic uncertainty itself. Uncertainty parameters can be used to characterize the uncertainty of any of the foregoing parameters.

Of the six classes of parameters just discussed, numerical algorithm and decision parameters are not normally viewed as being updated. They can certainly can be changed and optimized during an analysis so as to better achieve the goals of the analysis. For example, numerical algorithm parameters are usually adjusted to obtain better numerical

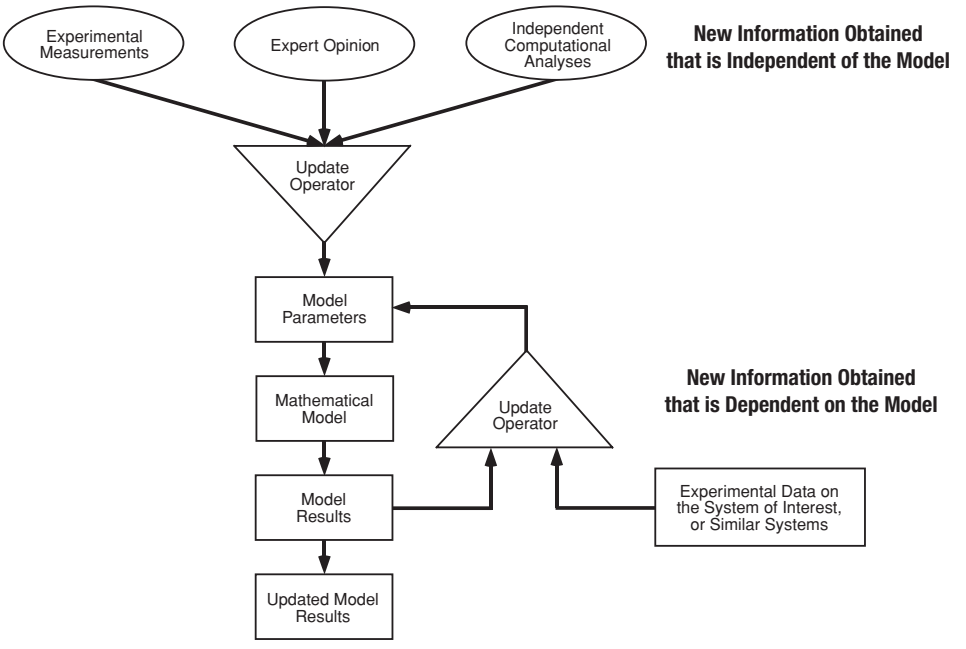


Figure 13.23 Two sources of information for parameter updating.

performance or stability for the various types of physics occurring within the system of interest. However, this type of adjustment of parameters is not normally considered as updating because it does not fundamentally deal with the mathematical model itself. The remaining four classes of parameters are fundamental to the model and they should be updated as new information becomes available.

### *13.5.2 Sources of new information*

New information upon which to update a parameter can come in various forms. Some of the most common sources of new information are (a) newly available experimental measurements of parameters, (b) additions or improvements in expert opinion concerning parameters, (c) separate computational analyses or theoretical studies that add information concerning parameters used in the present analysis, and (d) new experimental results for the system of interest or from systems similar to the present system. From our perspective of calibration and validation, it is clear that the type of information obtained from the system of interest provides fundamentally different inferential information than the other three sources listed. That is, the first three sources of information provide independent information for updating model parameters. The last source, however, provides information that can *only* be interpreted within the context of the mathematical model of the complete system.

Figure 13.23 shows how information for updating parameters falls into two categories. First, information obtained directly on parameters, either from experimental measurements, expert opinion, or independent computational analyses. This information is fundamentally

*independent* from the model of interest. Second, information obtained from comparing model and experimental results on the system or similar systems. This information is fundamentally *dependent* on the model. That is, the updated scalar value or probability distribution for a given parameter is *conditional* on the model of the system. Statisticians refer to any type of updating of parameters from any type of information as *statistical inference*. They do not make a distinction between the two fundamentally different types of information. To them, data are data and all data should be used to update the model. Their traditions, however, are deeply rooted in updating statistical models, i.e., models built on regression fits of data with no causality between input and output. In these models, the parameters have *no physical significance*; they are just *knobs* to be adjusted to obtain a best fit of the model to the experimental data. In physics-based models, however, these knobs are quite often physical modeling parameters with a *clear meaning independent from the present model*. Herein lies the conflict between viewpoints on the scientific justification of updating parameters in models.

With this perspective in mind, some of the key questions in parameter updating are:

- Should the source of new information affect the choice of the method used in updating?
- Should the type of new information, e.g., whether it contains aleatory and/or epistemic uncertainty, affect the choice of the method used in updating?
- How does the source of the new information affect the estimation of uncertainty in the updated model predictions, i.e., how does the source of the information affect the predictive capability of the model?

Some of the issues surrounding these questions will be addressed in the following sections.

### *13.5.3 Approaches to parameter updating*

Approaches to parameter updating are generally divided into methods for estimating scalar quantities and methods for estimating probability distributions for a parameter that is considered a random variable (Crassidis and Junkins, 2004; Raol *et al.*, 2004; Aster *et al.*, 2005; van den Bos, 2007). Here, we will focus on methods for estimating parameters that are given by probability distributions. This type of estimation can be considered as part of the broad field of statistical inference. In fact, most of statistics is devoted to statistical inference. The general problem of statistical inference is, given some information on uncertain quantities, how can the information be used to characterize uncertainty about the quantity. The vast majority of traditional statistical inference is focused on the characterization of random variables, i.e., aleatory uncertainty. Because of the breadth and depth of the development of statistical inference, we will only touch on some of the approaches and issues involved. For an in-depth discussion of the topic of statistical inference, see, for example, Bernardo and Smith (1994); Gelman *et al.* (1995); Leonard and Hsu (1999); Mukhopadhyay (2000); Casella and Berger (2002); Wasserman (2004); Young and Smith (2005); Cox (2006); Ghosh *et al.* (2006); and Sivia and Skilling (2006).

Unfortunately, statisticians are not in agreement about the way in which statistical inference should be accomplished. There are a wide variety of methods available that give rise to different estimates and different interpretations of the same estimate. In addition, statisticians do not agree on the principles that should be used to judge the quality and accuracy of estimation techniques. The two major camps of statistical inference are Frequentists and Bayesians. Both approaches are regularly used, often with much discussion of successes and little discussion of failure or erroneous inferential interpretations. Failure can be the most instructive and beneficial to the continued development of an approach and the development of improved guidance concerning when one approach is more appropriate than another.

The Frequentist approach is also commonly called the classical approach and it involves a number of methods. Two primary features of the Frequentist approach are particularly important with respect to parameter updating. First, probability is *strictly* interpreted as the relative frequency of an event occurring in the limit of a large number of random trials or observations from an experiment. The set of all possible outcomes of a random trial is called the sample space of the experiment. An event is defined as a particular subset, or realization, of the sample space. Second, updating can *only be done* when the new information is derived from experimental measurements obtained directly on the parameters of interest. Some of the more common classes of methods are: point-estimation of a scalar, interval-estimation of a scalar, hypothesis testing, and regression and correlation analyses. Here we are mainly interested in methods for estimating uncertainty parameters, e.g., parameters of a probability distribution that is chosen to represent the random variable of interest. In the estimation of a point value for an uncertainty parameter, it is assumed that the parameter is constant, but imprecisely known because only a sample of the population is available. Various types of best estimate method are available to estimate the point value. There are also methods available for interval-estimates for the parameters of the distribution. When this is done, the uncertain quantity of interest becomes a mixture of aleatory and epistemic uncertainty. The two most commonly used methods are the method of moments and the method of maximum likelihood. Because of the restriction on the type of information that can be used and the quantity of information that is required, Frequentist methods are generally considered less applicable for the type of parameter updating shown in Figure 13.23. The primary advantage of Frequentist methods is that they are widely viewed as simpler and easier to use than Bayesian methods.

Bayesian methods take a different and broader perspective concerning statistical inference. They consider the distribution characterizing the uncertainty in the physical parameter as a function that is (a) initially unknown, and (b) should be updated as more information becomes available concerning the distribution. The initial postulated distribution for the uncertainty parameter can come from any source or mixture of sources, for example, experimental measurements, expert opinion, or computational analyses. More importantly, the distribution for the uncertainty parameter can be updated when any type of new information becomes available. Before updating, the distribution is referred to as the *prior distribution*,

and after updating it becomes the *posterior distribution*. It is well accepted that the proper method of updating distributions is Bayes' theorem.

When no information is available for the prior distribution, the analyst can simply choose as prior his/her personal probability. The Bayesian paradigm uses the notion of subjective probability, or personal degree of belief, defined in the theory of rational decision making. In this theory, every rational individual is a free agent and is not constrained to choose the same prior as another individual. This, of course, leads to vehement criticisms from the Frequentists to the Bayesian paradigm. The Frequentists commonly raise the question, how can such subjectivity in choosing probability distributions lead to trustworthy inference? In the present context of simulation, how can personal beliefs expressed in the input parameters to a model lead to trustworthy decision making based on the model outputs? This question is especially critical if the resulting output probabilities are interpreted as a frequency of occurrence, for example, in terms of regulatory requirements on safety or reliability of a system. These are serious criticisms if little or no justification for prior distributions is available. The primary defense of Bayesian inference is based on the argument that continually adding new information from all available sources minimizes the impact of the subjectivity in the initial choice of prior distributions. And, as a result, the initial choice of prior distributions will have little affect on the final model outputs needed for decision making.

In this discussion, we have focused on the updating of uncertainty parameters for characterizing aleatory uncertainty in measurable parameters, physical modeling parameters, and ad hoc parameters. A related question is: if the uncertainty parameters are epistemic uncertainties, e.g., intervals, how can they be updated? For example, suppose that an expert gives the uncertainty characterization for a measurable system parameter as an interval. How should the parameter be updated or modified if a new, equally credible, expert provides a non-overlapping interval? Neither Frequentist nor Bayesian approaches recognize epistemic uncertainty as a separate type of uncertainty, and as a result, they cannot address this question. Over the last decade, the topic of updating epistemic uncertainties, usually referred to as information aggregation or data fusion, has received increasing attention. The practical need for this type of updating is obvious, if one accepts the separation of aleatory and epistemic uncertainty. The emphasis in this field has concentrated on dealing with contradictions or conflict in information from various sources because the information content in epistemic uncertainty can be minimal and it can represent different types of epistemic uncertainty. For example, one may also have to deal with differences in linguistic interpretation of knowledge and, as a result, use fuzzy set theory or a combination of fuzzy set theory and classical probability theory. For an in-depth discussion of aggregation methods for various types of uncertainties, see Yager *et al.* (1994); Bouchon-Meunier (1998); Sentz and Ferson (2002); Ferson *et al.* (2003); Beliakov *et al.* (2007); and Torra and Narukawa (2007).

### *13.5.4 Parameter updating, validation, and predictive uncertainty*

Whether one uses the Frequentist or Bayesian approach for updating parameters, from our perspective there are two serious difficulties that are caused by convolving parameter

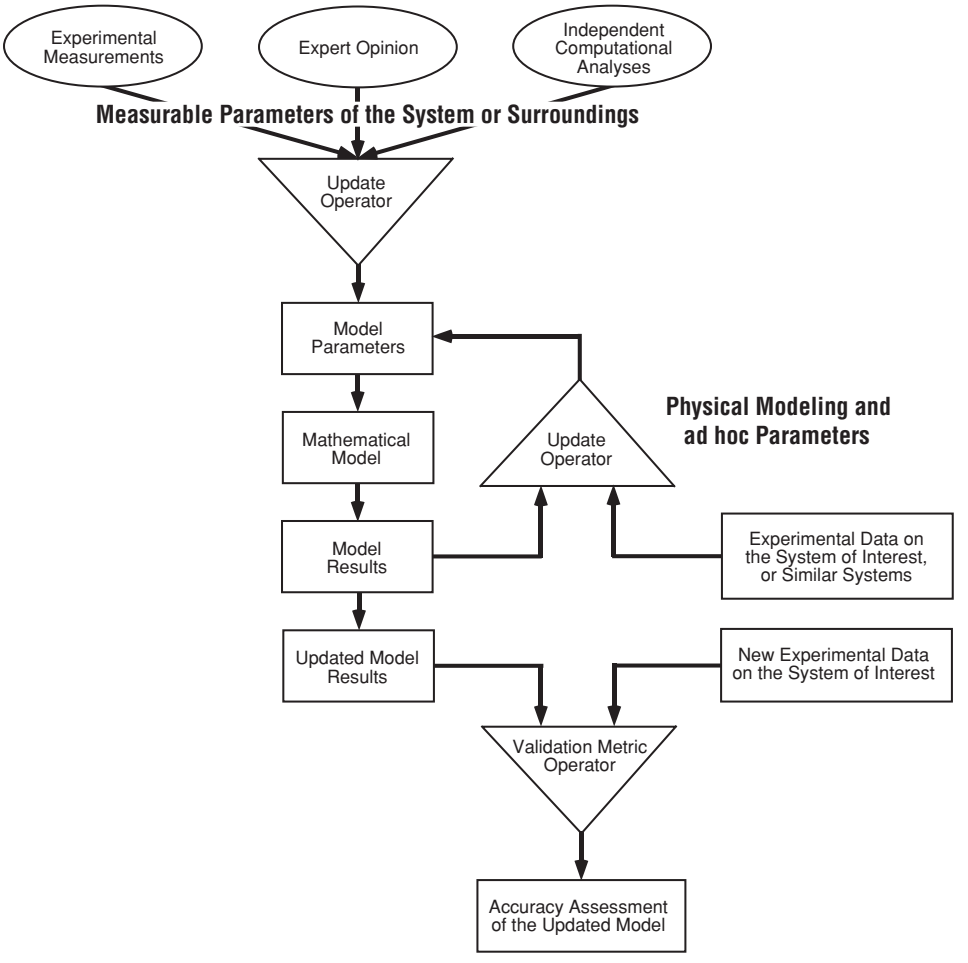


Figure 13.24 Parameter updating, validation, and predictive uncertainty.

updating with the model of the system. We are not the first, of coarse, to raise these concerns (Beck, 1987). These concerns will be discussed in the following two sections.

### *13.5.4.1 Parameter updating*

In Figure 13.23 discussed above, we segregated the various sources of new information into two groups: information obtained that is independent from the model and information obtained that is dependent on the model. Figure 13.24 maps three types of physical parameters discussed in Section 13.5.1 onto the concepts shown in Figure 13.23. Measurable physical parameters of the system or surroundings (simply measurable parameters) can be determined, at least in principle, through measurements that are independent of the system of interest. Physical modeling parameters and ad hoc parameters can *only be determined* by using the model of the system of interest in concert with the experimental data for the

system. As a result, measurable parameters are shown at the top of the figure where they could be independently updated. Physical modeling and ad hoc parameters are shown on the right of the figure because they can only be updated in the feedback loop shown using the mathematical model of the system.

We believe the proper approach to updating the measurable parameters of the system or surroundings is to do so separately from the model for the system of interest. If they are updated with the model of the system, using experimental data for the system, then our concern is the same as with updating physical modeling and ad hoc parameters. That is, our concern is that the updated parameters were determined by convolving parameter updating while using the model of the system. Suppose one used either a Frequentist or a Bayesian approach for model updating and the question was asked: how would you conceptually describe the updating approach to parameters in a physics-based model? There would certainly be a wide range of responses. We suggest that two extremes would cover the range of opinions. Respondent A: The approach attempts to optimally use the information available for the inputs to the model, the physics embodied in the model, and the available experimental results to produce the best prediction possible from the model. Respondent B: The approach is a constrained regression fit of the experimental data in the sense that the physics-based model represents the constraint on the parameters. We would agree more with Respondent B than A. We present the following thought experiment to suggest that B is the more accurate conceptual description.

Recalling the discussion in Chapter 10, the total error in a simulation result can be written as the sum of four sources,

$$E_{\text{sim}} = E_1 + E_2 + E_3 + E_4, \tag{13.50}$$

where the sources are

$$\begin{aligned} E_1 &= (y_{\text{sim}} - y_{\text{Pcomputer}}), \\ E_2 &= (y_{\text{Pcomputer}} - y_{\text{model}}), \\ E_3 &= (y_{\text{model}} - y_{\text{exp}}), \\ E_4 &= (y_{\text{exp}} - y_{\text{nature}}). \end{aligned} \tag{13.51}$$

$E_1$ represents all numerical errors resulting from the difference between the discrete solution, $y_{\text{sim}}$ (obtained using a finite discretization size, finite iterative convergence, and finite precision computer), and the exact solution to the discrete equations obtained on a perfect computer as the discretization size approaches zero, $y_{\text{Pcomputer}}$. $E_1$ is referred to as the solution error and the most important contributor to this error is an inadequately resolved mesh or time step.

$E_2$ represents all errors resulting from the difference between the exact solution to the discrete equations as the discretization size approaches zero, $y_{\text{Pcomputer}}$, and the exact solution of the mathematical model, $y_{\text{model}}$. The most important contributor to this error is coding errors in the software computing the model result.

$E_3$ represents all errors resulting from the difference between the exact solution to the mathematical model, $y_{\text{Pcomputer}}$, and the experimental measurement, $y_{\text{exp}}$. $E_3$ is referred to as the model error or model form error and the most important contributors to this error are model bias error, i.e., $d(x)$, and errors made by the analyst in preparation of the input data.

$E_4$ represents all errors due to the difference between the true, but unknown value of nature, $y_{\text{nature}}$, and the measurement of a physical quantity, $y_{\text{exp}}$. The most important contributors to this error are systematic and random errors in experimentally measured quantities.

Suppose that at a given point in the conduct of an analysis of a system there have been a number of updates on all adjustable parameters using the model and the experimental measurements from a number of systems. Suppose that any one of the following situations occurred that caused a large change in the predicated SRQ of interest: (a) a finer mesh resolution or a smaller time step was used, (b) a coding error was discovered and then corrected in one of the computer codes of the model, (c) an input error was discovered and corrected in one of the computer codes, and (d) an error was discovered and corrected in the data reduction procedure processing the experimental sets of data used in the updating. Each one of these situations would have caused a large increase in the magnitude of at least one of the $E_i$ components discussed above, and a large increase in the newly-calculated total simulation error, $E_{\text{sim}}$.

Whether it is Respondent A or B discussed above, a new effort would be initiated to update all of the adjustable parameters in the model. Given the presumption that there would be a large increase in $E_{\text{sim}}$, there would presumably be large changes in at least some of the three types of updated parameters: measurable, physical, and ad hoc. This is because updating attempts to *achieve error cancellation among the various sources* by adjusting all of the available parameters in the model in order to minimize $E_{\text{sim}}$. There has been, of course, *no change at all* in the physics of the system. Large changes in the measurable parameters are the most embarrassing because there should be *no changes* in these quantities since they are independently measurable quantities. In addition, if large changes occur in the physical modeling parameters, it is also troubling because they have a clear physical meaning associated with well-accepted physical processes. Large changes in both types of parameter demonstrate that updating is conceptually aligned with regression fitting of experimental data, constrained by the physics model structure.

#### *13.5.4.2 Validation after parameter updating*

In the lower portion of Figure 13.24, we show the validation metric operator along with its two inputs and the output of model accuracy assessment of the updated model. Our concern deals with the important question: how should one interpret the validation metric result when an updated model is assessed for accuracy? The answer to this question is critically important to understand because this information is used, for example, by decision makers in assessing the predicted safety, reliability, and performance of systems; and government regulators in determining public safety and potential environmental impact. We give two

situations: the first where the model can be appropriately assessed for accuracy, and the second where the assessment would be erroneous and misleading.

Suppose that new experimental data became available on the system of interest. Presume the data were for input conditions that were substantially different from previous experimental data used in updating the model. For this case, we believe a comparison of model results and the new experimental data would properly constitute an assessment of model accuracy. Consider the following example. Suppose one were interested in several vibrational modes of a built-up structure. The structure is composed of multiple structural elements that are bolted together at a number of connection points of the elements. The experimental data used for updating the model were obtained on the structure with the torsional preload on each bolt specified at a certain value. The stiffness and damping of all of the joints were determined using either Frequentist or Bayesian updating applied to the results from the model and the experimental data from several prototype structures that were tested.

Then, because of some design considerations, the assembly requirements were changed such that the preload on all of the bolts was doubled. The model included a submodel for how the stiffness and damping of the joints of the structure would depend on preload of the bolts. Several of the existing structures were modified and retested as a result of this bolt–preload doubling. We argue that when the new results from the model are compared in a blind-test prediction with the new experimental data, an appropriate assessment of model predictive capability can be made. If so desired, one could compute a validation metric to quantitatively assess model accuracy using the new predictions and measurements. An even more demanding test of the predictive capability of the model would be if the new assembly requirements included changing some of the bolted joints to riveted joints, presuming that a submodel was available for riveted joints.

Now, consider the case where new experimental data became available on the system of interest, and that the model had been updated using previously available data on the same system. For this case, we argue that one could *not* claim that the predictive accuracy of the model was being assessed by comparison with the new experimental data because the new data were obtained for the same system, except for random variability in the system. For example, suppose that in the structural vibration example just described, additional data were obtained on newly assembled structures. However, the structures were assembled using exactly the same specifications for all of the structural elements and the same preload on all of the bolts in the structure as those used earlier to update the model. That is, no new physics was exercised in the model for the new predictions. Even though one would be obtaining new experimental data that has never been used in updating the model, the experimental samples were drawn from the same parent population as the previous samples. The new data would provide additional sampling information concerning manufacturing variability in structural elements and bolt fastening, but it would *not* provide any new test of the physics embedded in the model. Specifically, it would not represent an estimate of the model form error, $d(x)$. As a result, it could *not* be claimed that the accuracy of the model was being assessed by comparison with the new data.

As can be seen from these two examples there is a very wide range of how comparisons with experimental data should be interpreted when model updating is involved. Understanding that the philosophical root of V&V is skepticism, claims of good agreement with data should always be questioned.

## 13.6 Step 6: conduct sensitivity analysis

Morgan and Henrion (1990) provide the broadest definition of sensitivity analysis: the determination of how a change in any aspect of the model changes any predicted response of the model. The following are some examples of elements of the model that could change, thereby producing a change in the response of the model: (a) specification of the system and the surroundings; (b) specification of the normal, abnormal, or hostile environments; (c) assumptions in the formulation of the conceptual or mathematical model; (c) assumptions concerning coupling of various types of physics between submodels; (d) assumptions concerning which model inputs are considered deterministic and which are considered uncertain; (e) a change in variance of an aleatory uncertainty or a change in the magnitude of an epistemic uncertainty of model inputs; and (f) assumptions concerning independence and dependence between uncertain inputs. As can be seen from these examples, changes in the model can be viewed as either what-if analyses, or analyses determining how the outputs change as a function of changes in the inputs, whether the inputs are deterministic, epistemic uncertainty, aleatory uncertainty, or a combination of the two. Uncertainty analysis, on the other hand, is the quantitative determination of the uncertainty in any SRQ as a function of any uncertainty in the model.

Although uncertainty analysis and sensitivity analysis (SA) are closely related, SA is concerned with the characterization of the relative importance of how various changes in the model will change the model predictions. Recalling Eq. (13.32), we have

$$\vec{y} = f(\vec{x}), \tag{13.52}$$

where $\vec{x} = \{x_1, x_2, \ldots, x_m\}$ and $\vec{y} = \{y_1, y_2, \ldots, y_n\}$. SA is a more complex mathematical task than UQ because it explores the mapping of $\vec{x} \rightarrow \vec{y}$ in order to assess the effects of individual elements of $\vec{x}$ on elements of $\vec{y}$. In addition, SA also deals with how changes in the mapping, $f$, changes elements of $\vec{y}$. SA is usually done after a UQ analysis, or at least after a preliminary UQ analysis has been completed. In this way the SA can take advantage of a great deal of information that has been generated as part of a UQ analysis.

Results from an SA are typically used in two ways. First, if an SA is conducted on the elements dealing with the formulation of the model or the choice of the submodels, then one can use the results as a planning tool for improved allocation of resources on a project. For example, suppose that a number of submodels are used and they are coupled through the model for the system. An SA could be conducted to determine the ranking of which submodels have the most impact on each of the elements of $\vec{y}$. For simplicity, suppose that each submodel resulted in one output quantity that was then used in the model of the system.

The output quantity from each submodel could be artificially changed, say by 10%, before the quantity was used as input to the system model. This would be done one submodel at a time. One could then rank in order from largest to smallest, which models produced the largest change on each of the elements of $\vec{y}$. The rank order for each element of $\vec{y}$ would normally be different than another element. If it was found that certain submodels had essentially no effect on any of the elements of $\vec{y}$, then the lead analyst or project manager could alter the allocation of resources within the project. He/she could move funding and resources from the less important submodels to the most important submodels. This can be a difficult adjustment for some involved in a scientific computing project.

Second, it is more common that an SA is used to determine how changes in input uncertainties affect the uncertainty in the elements of $\vec{y}$. Some examples of how this information can be used are (a) reducing the manufacturing variability in key input quantities to improve the reliability of the system, (b) altering the operating envelope of the system in order to improve system performance, and (c) altering the system design to improve system safety when it is discovered that an important input uncertainty cannot be reduced. When used in this way, one usually conducts either a local or global SA. Local and global SAs will be briefly discussed below. For a detailed discussion of SA, see Helton *et al.* (2006) and Saltelli *et al.* (2008).

### *13.6.1 Local sensitivity analysis*

In a local SA, one is interested in determining how outputs locally change as a function of inputs. This is equivalent to computing the partial derivatives of the SRQs with respect to each of the input quantities. Although a local SA can be conducted without the assumption of the existence and continuity of partial derivates of the SRQs, it is a simpler approach to introduce the concepts. Here we will focus on the input quantities that are uncertain, although one could certainly considered all input quantities. A local SA computes $m \times n$ derivatives of the random variable $y_j$ with respect to the random variable $x_i$:

$$\left(\frac{\partial y_j}{\partial x_i}\right)_{\vec{x}=\vec{c}}, \quad i = 1, 2, 3, \ldots m \quad \text{and} \quad j = 1, 2, 3, \ldots, n, \tag{13.53}$$

$x_i = \{x_1, x_2, \ldots, x_m\}$, $y_j = \{y_1, y_2, \ldots, y_n\}$, and $\vec{c}$ is a vector of constants that specifies a statistic of the input quantities at which the derivatives are evaluated. The vector $\vec{c}$ can specify any input condition over which the system can operate. The most common condition of interest is the mean of each of the input parameters, $\bar{x}_i$. In most engineering systems, one can usually consider the input quantities as continuous variables, as opposed to discrete quantities. Since we have focused on sampling techniques, one must compute a sufficient number of samples in order to construct a smooth function so that the partial derivative in Eq. (13.53) can be computed.

As an example, suppose that $\vec{x} = \{x_1, x_2, x_3\}$ and $\vec{y} = \{y_1, y_2\}$. Figure 13.25 depicts $y_1$ and $y_2$ as a function of $x_2$, given that $x_1 = c_1$ and $x_3 = c_3$. Since $\vec{x}$ exists in a three-dimensional space, and $y_1$ and $y_2$ each exist in a four-dimensional space, Figure 13.25 can

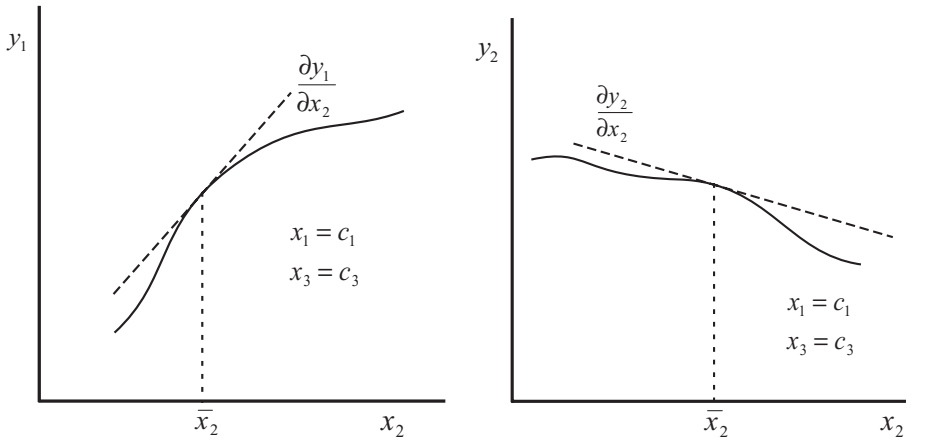


Figure 13.25 Example of system responses and derivatives in a local sensitivity analysis.

be viewed as $y_1$ and $y_2$ as a function of $x_2$, in the plane of $x_1 = c_1$ and $x_3 = c_3$. Also shown in the figure are the derivatives $\left(\frac{\partial y_1}{\partial x_2}\right)_{x_1=c_1, x_3=c_3}$ and $\left(\frac{\partial y_2}{\partial x_2}\right)_{x_1=c_1, x_3=c_3}$ evaluated at the mean of $x_2$, $\bar{x}_2$. Note that $\{y_1, y_2\}$ and the derivatives can be computed without assuming any uncertainty structure of the inputs. $y_1$ and $y_2$ are simply evaluated at a sufficient number of samples over the needed range of each $x_i$. If the mean of any of the input quantities is needed, then a sufficient number of samples of $\{x_1, x_2, x_3\}$ must be computed so that the mean can be satisfactorily estimated.

Results from local SAs are most commonly used in system design and optimization studies. For example, one could consider how all the input parameters affect the performance, reliability or safety of the system. Some of the input parameters the designer has control over; some are fixed by design constraints on the system, such as the operating envelope, size, and weight. For those that can be controlled, a local SA can greatly aid in optimizing the design, including any flexibility in the operating envelope, so that improved performance, reliability, and safety can be obtained.

### *13.6.2 Global sensitivity analysis*

A global SA is conducted when information is needed concerning how the uncertainty structure of all of the inputs maps to the uncertainty structure of each of the outputs. A global SA is appropriate when a project manager and decision maker needs information on what are the most important uncertain inputs driving specific uncertain outputs. This is especially needed, for example, when the outputs fall outside system performance requirements, or exceed governmental regulatory requirements. With the information from a global SA, a project manager can determine what are the most important contributors forcing the outputs of interest into unwanted ranges of response. The information concerning the ordering of

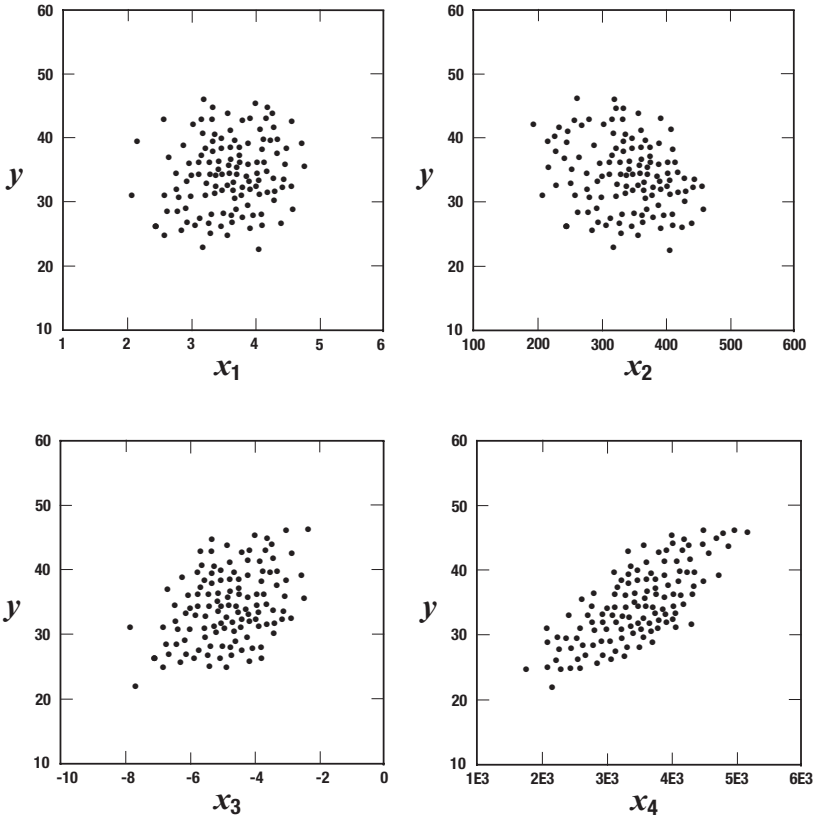


Figure 13.26 Scatter plots for one output quantity as a function of four input quantities.

these input contributors based on their effect on the individual outputs will also include the effects of the uncertainties from all of the other input quantities in the simulation.

A common first step in a global SA is to construct a scatter plot for each of the $y_j$ as a function of each $x_i$. This would result in a total of $m \times n$ plots. The data points in the scatter plots can come from the MCS or LHS that are computed as part of the UQ analysis. The purpose of the scatter plots is to determine if any trend exists in each output quantity $y_j$ as a function of each $x_i$. For example, suppose we have one output quantity, $y$, and four input quantities, $\{x_1, x_2, x_3, x_4\}$. Figure 13.26 shows the four scatter plots that would be generated, given a total of 100 LHS samples. The scatter plots are a projection from a five-dimensional uncertainty space, $\{y, x_1, x_2, x_3, x_4\}$, onto the $y - x_1$ plane, the $y - x_2$ plane, the $y - x_3$ plane, and the $y - x_4$ plane, respectively. The shape of the uncertainty clouds indicate that: (a) there is no discernable trend of $y$ with $x_1$, (b) there is a slightly decreasing trend of $y$ with $x_2$, (c) there is a clear trend of $y$ increasing with $x_3$, and (d) there is a strong trend of $y$ increasing with $x_4$. Note that there could be nonlinear trends buried within the clouds shown in Figure 13.26. One could attempt to rank order the $x_i$ to

determine which one has the strongest influence on $y$ by computing a linear regression of each scatter plot. Using the slope of the linear regression, one could compare the magnitude of

$$\left|\frac{\partial y}{\partial x_1}\right|, \quad \left|\frac{\partial y}{\partial x_2}\right|, \quad \left|\frac{\partial y}{\partial x_3}\right|, \quad \text{and} \quad \left|\frac{\partial y}{\partial x_4}\right|. \tag{13.54}$$

This result, however, would be of limited value in a global SA because the magnitude of each derivative would depend on the dimensional units of each $x_i$. In addition, each derivative has no information concerning the nature of the uncertainty in each $x_i$.

It is standard practice to reformulate the scatter plots so that all the $x_i$ are rescaled by their respective estimated means, $\bar{x}_i$, and $y$ is rescaled by its estimated mean, $\bar{y}$. In this way one can then properly compare the regression slopes given in Eq. (13.54). We then have

$$\left|\frac{\partial(y/\bar{y})}{\partial(x_1/\bar{x}_1)}\right|, \quad \left|\frac{\partial(y/\bar{y})}{\partial(x_2/\bar{x}_2)}\right|, \quad \left|\frac{\partial(y/\bar{y})}{\partial(x_3/\bar{x}_3)}\right|, \quad \text{and} \quad \left|\frac{\partial(y/\bar{y})}{\partial(x_4/\bar{x}_4)}\right|. \tag{13.55}$$

The terms in Eq. (13.55) are either referred to as the sigma-normalized derivatives, or the standardized regression coefficients (SRCs) (Helton *et al.*, 2006; Saltelli *et al.*, 2008). If the response $y$ is nearly linear in all of the $x_i$, and the $x_i$ are independent random variables, then the SRCs listed in Eq. (13.55) can be rank ordered from largest to smallest to express the most important to least important effects of input quantities on the output quantity. In most SAs it is found that even if there are a large number of uncertain inputs $x_i$, there are only several inputs that dominate the effect on a given output quantity. As a final point, it should be clear that the rank ordering of the SRCs for one output quantity need not be the same, or even similar, to the rank ordering for a different output quantity. If there are multiple output quantities of high importance to a system's performance, safety, or reliability, then the list of important input quantities (considering all of the important output quantities) can grow significantly.

The most common method of determining if $y$ is linear with each of the $x_i$, and determining if the $x_i$ are independent, is to examine the sum of the $R^2$ values from each of the linear regression fits. $R^2$ is interpreted as the proportion of the observed $y$ variation with $x_i$ that can be represented by the regression model. For the four input quantities in the present example, these are written as $R_1^2$, $R_2^2$, $R_3^2$, $R_4^2$, respectively. If the sum of the $R^2$ values is near unity, then there is reasonable evidence that a linear regression model can be used to rank order the SRCs. If one finds that the sum of the $R^2$ values is much less than unity, there could be (a) nonlinear trends of $y$ with some of the $x_i$; (b) statistical dependencies or correlations between the $x_i$; or (c) strong interactions between the $x_i$ within the mapping of $x_i \rightarrow y$, e.g., in the physics represented in the model. Applying a linear regression to each of the graphs shown in Figure 13.26, one finds

$$R_1^2 = 0.01, \quad R_2^2 = 0.11, \quad R_1^2 = 0.27, \quad \text{and} \quad R_1^2 = 0.55. \tag{13.56}$$

The sum of the $R^2$ values is 0.94, indicating there is evidence that the SRCs computed from the regression fits can properly estimate the relative importance of each of the input quantities in the global SA. If the sum of the $R^2$ values is significantly less than unity, then more sophisticated techniques such as rank regression, nonparametric regression, and variance decomposition must be used (Helton *et al.*, 2006; Saltelli *et al.*, 2008; Storlie and Helton, 2008a,b).

## 13.7 Example problem: thermal heating of a safety component

This example is taken from a recent workshop dealing with investigating methods for model validation and predictive capability (Hills *et al.*, 2008; Pilch, 2008). The organizers of the workshop constructed three example problems, each one dealing with different physical phenomena: thermal heating of a safety component, static deflection of a frame structure, and dynamic deflection of a structural element. Each mathematical model constructed was purposefully designed with some model weaknesses and vagaries in order to realistically reflect common situations analysts encounter. For example, there are modeling assumptions that seem questionable, including assertions that certain parameters are constants, that interacting variables are mutually independent, and that there are no experimental data on the system of interest at operational conditions.

Each of the three problems constructed for the workshop had a similar formulation: a mathematical model was specified along with an analytical solution, experimental data were given for the system or closely related systems, and a prediction was required concerning the probability that the system satisfy a regulatory safety criterion. The thermal heating challenge problem is described in Dowding *et al.* (2008). For the thermal heating problem, the regulatory criterion specified that the temperature at a specific location and time not exceed a temperature of 900 °C with a probability of larger than 0.01. This criterion can be written as

$$P\{T(x = 0, t = 1000) > 900\} < 0.01, \tag{13.57}$$

where $T$ is the temperature of the component, $x = 0$ is the face of the component, and the time is 1000 s. In this section, we describe an analysis approach to the thermal heating problem that is based on Ferson *et al.* (2008). For a complete description of each challenge problem and the analyses presented by a number of researchers, see Hills *et al.* (2008).

The model for the thermal problem is one-dimensional, unsteady, heat conduction through a slab of material (Dowding *et al.*, 2008). A heat flux, $q_W$, is specified at $x = 0$ and an adiabatic condition ($q_W = 0$) is specified at $x = L$ (see Figure 13.27). The initial temperature of the slab is uniform at a value of $T_i$. The thermal conductivity, $k$, and the volumetric heat capacity, $\rho C_p$, of the material in the slab are considered as independent of temperature. $k$ and $\rho C_p$, however, are uncertain parameters due to manufacturing variability of the slab material. That is, $k$ and $\rho C_p$ can vary from one manufactured slab to the next, but within a given slab, $k$ and $\rho C_p$ are constant. For this simple model, an analytical solution

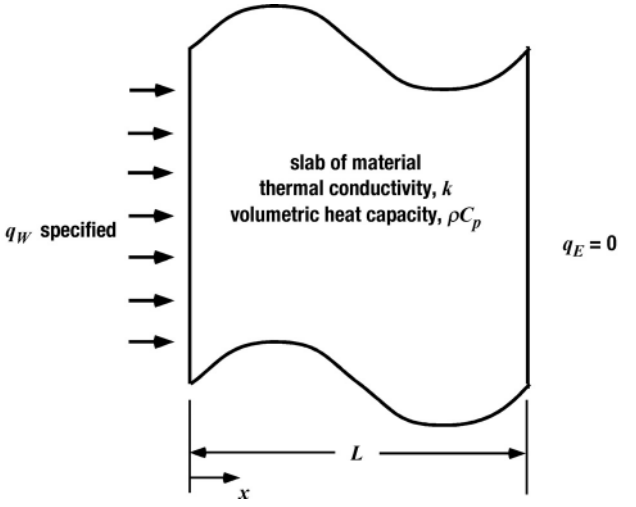


Figure 13.27 Schematic for the thermal heating problem.

to the PDE can be written as

$$T(x,t) = T_i + \frac{q_{\mathrm{W}} L}{k}\left\{\frac{1}{3} - \frac{x}{L} + \frac{1}{2}\left(\frac{x}{L}\right)^2 + \frac{k\,t}{\rho C_{\mathrm{p}} L^2} - \frac{2}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\exp\left[-\frac{n^2\pi^2 k\,t}{\rho C_{\mathrm{p}} L^2}\cos\left(n\pi\frac{x}{L}\right)\right]\right\}. \quad (13.58)$$

$T(x, t)$ is the temperature at any point in the slab, $x$, and at any value of time, $t$.

In comparing this example with the previously discussed example in this chapter it is seen there are some similarities, but there are four significant differences. First, an analytical solution is available to the PDE describing the system response, as opposed to the need to compute a numerical solution. Second, this example deals with a time dependent system response, as opposed to a steady state problem. Third, the validation metric is computed using the mismatch between CDFs of the model prediction and the experimental measurements, as opposed to the confidence interval approach. And fourth, the SRQ of interest depends on four coordinate dimensions $q_{\mathrm{W}}$, $L$, $x$, and $t$, that will be used in computing the validation metric and in the extrapolation of the model, as opposed to one for the earlier example problem.

This example will discuss steps 1, 2, and 4 of the prediction procedure. Step 3 is omitted because an analytical solution is give for the mathematical model, resulting in negligible numerical solution error. Steps 5 and 6 are omitted because they were not part of the analysis of Ferson *et al.* (2008).

Table 13.6 *Model input data for the system of interest and the validation experiments for the thermal heating example.*

| Model input data | System of interest | Validation experiments |
|---|---|---|
| **System input data** | | |
| Slab thickness, $L$ | $L = 1.9$ cm, deterministic | $L = 1.27, 2.54, 1.9$ cm, deterministic |
| Initial temperature, $T_i$ | $T_i = 25\ ^\circ$C, deterministic | $T_i = 25\ ^\circ$C, deterministic |
| Thermal conductivity, $k$ | $k$, aleatory uncertainty | $k$, aleatory uncertainty |
| Volumetric heat capacity, $\rho C_p$ | $\rho C_p$, aleatory uncertainty | $\rho C_p$, aleatory uncertainty |
| **Surroundings input data** | | |
| Heat flux, $q_W$ | $q_W = 3500$ W/m$^2$, deterministic | $q_W = 1000, 2000, 3000$ W/m$^2$, deterministic |
| Heat flux, $q_E$ | $q_E = 0$, deterministic | $q_E = 0$, deterministic |

### *13.7.1 Step 1: identify all relevant sources of uncertainty*

Segregating the model input data into system data and surroundings data, we can construct Table 13.6 for the thermal heating problem. The system of interest, the one for which Eq. (13.57) needs to be evaluated, has the characteristics shown in the middle column of Table 13.6. Various validation experiments were conducted with differing characteristics shown in the right hand column of Table 13.6. $k$ and $\rho C_p$ are considered aleatory uncertainties due to manufacturing variability.

### *13.7.2 Step 2: characterize each source of uncertainty*

#### *13.7.2.1 Model input uncertainty*

The characterization of the uncertainty in $k$ and $\rho C_p$ is based on three sets of experimental data for the material in the slab. The tabular data for each of the sets of data, referred to as low, medium, and high in reference to the quantity of data obtained. Table 13.7 and Table 13.8 give the material characterization data for $k$ and $\rho C_p$, respectively.

Figure 13.28 shows the empirical CDFs for $k$ and $\rho C_p$ for the medium materials characterization data. These observed patterns likely understate the true variability in these parameters because they represent only 20 observations for each one. (For the formulation of the workshop problems, experimental measurement uncertainty was assumed to be zero.) To model this possibility of more extreme values than were seen among the limited samples, it is common to fit a distribution to data to model the variability of the underlying population. We used normal distributions for this purpose, configured so that they had the same mean and standard deviation as the data themselves, according to the method of matching moments (Morgan and Henrion, 1990). The fitted normal distributions are shown in Figure 13.28 as the smooth cumulative distributions. Fitting of distributions for the material characterization is not model calibration, as discussed earlier, because the distributions

Table 13.7 *Thermal conductivity data for low, medium and high data sets (W/m-C) (Dowding* et al.*, 2008).*

| $k$(20 °C) | $k$(250 °C) | $k$(500 °C) | $k$(750 °C) | $k$(1000 °C) |
|---|---|---|---|---|
| Low data set, $n = 6$ | | | | |
| 0.0496 | – | 0.0602 | – | 0.0631 |
| 0.053 | – | 0.0546 | – | 0.0796 |
| Medium data set, $n = 20$ | | | | |
| 0.0496 | 0.0628 | 0.0602 | 0.0657 | 0.0631 |
| 0.053 | 0.062 | 0.0546 | 0.0713 | 0.0796 |
| 0.0493 | 0.0537 | 0.0638 | 0.0694 | 0.0692 |
| 0.0455 | 0.0561 | 0.0614 | 0.0732 | 0.0739 |
| High data set, $n = 30$ | | | | |
| 0.0496 | 0.0628 | 0.0602 | 0.0657 | 0.0631 |
| 0.053 | 0.062 | 0.0546 | 0.0713 | 0.0796 |
| 0.0493 | 0.0537 | 0.0638 | 0.0694 | 0.0692 |
| 0.0455 | 0.0561 | 0.0614 | 0.0732 | 0.0739 |
| 0.0483 | 0.0563 | 0.0643 | 0.0684 | 0.0806 |
| 0.049 | 0.0622 | 0.0714 | 0.0662 | 0.0811 |

Table 13.8 *Volumetric heat capacity for low, medium and high data sets (J/m$^3$-C) (Dowding* et al.*, 2008).*

| $\rho C_p$(20 °C) | $\rho C_p$(250 °C) | $\rho C_p$(500 °C) | $\rho C_p$(750 °C) | $\rho C_p$(1000 °C) |
|---|---|---|---|---|
| Low data set, $n = 6$ | | | | |
| 3.76E + 05 | – | 4.52E + 05 | – | 4.19E + 05 |
| 3.38E + 05 | – | 4.10E + 05 | – | 4.38E + 05 |
| Medium data set, $n = 20$ | | | | |
| 3.76E + 05 | 3.87E + 05 | 4.52E + 05 | 4.68E + 05 | 4.19E + 05 |
| 3.38E + 05 | 4.69E + 05 | 4.10E + 05 | 4.24E + 05 | 4.38E + 05 |
| 3.50E + 05 | 4.19E + 05 | 4.02E + 05 | 3.72E + 05 | 3.45E + 05 |
| 4.13E + 05 | 4.28E + 05 | 3.94E + 05 | 3.46E + 05 | 3.95E + 05 |
| High data set, $n = 30$ | | | | |
| 3.76E + 05 | 3.87E + 05 | 4.52E + 05 | 4.68E + 05 | 4.19E + 05 |
| 3.38E + 05 | 4.69E + 05 | 4.10E + 05 | 4.24E + 05 | 4.38E + 05 |
| 3.50E + 05 | 4.19E + 05 | 4.02E + 05 | 3.72E + 05 | 3.45E + 05 |
| 4.13E + 05 | 4.28E + 05 | 3.94E + 05 | 3.46E + 05 | 3.95E + 05 |
| 4.02E + 05 | 3.37E + 05 | 3.73E + 05 | 4.07E + 05 | 3.78E + 05 |
| 3.53E + 05 | 3.77E + 05 | 3.69E + 05 | 3.99E + 05 | 3.77E + 05 |

Table 13.9 *Characterization of the uncertainty in k and $\rho C_p$ using a normal distribution for the low, medium, and high data sets (Ferson* et al.*, 2008).*

| | Low data set ($n = 6$) | Medium data set ($n = 20$) | High data set ($n = 30$) |
|---|---|---|---|
| Thermal conductivity $k$, (W/m-C) | | | |
| Mean | 0.06002 | 0.06187 | 0.06284 |
| Standard deviation | 0.01077 | 0.00923 | 0.00991 |
| Volumetric heat capacity, $\rho C_p$ (J/m$^3$-C) | | | |
| Mean | 405 500 | 402 250 | 393 900 |
| Standard deviation | 42 065 | 39 511 | 36 251 |

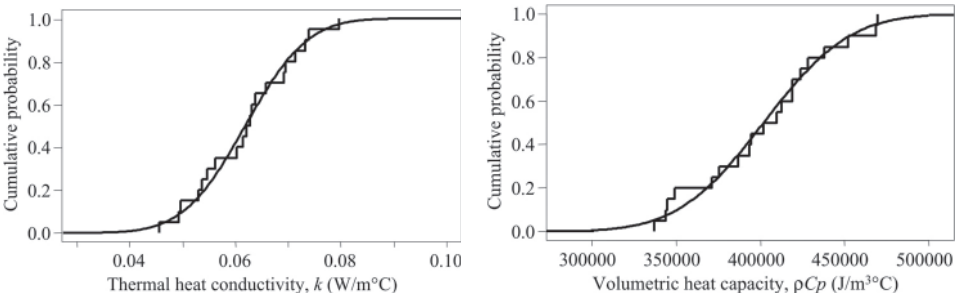


Figure 13.28 Empirical CDF (step functions) and a continuous normal distribution for $k$ and $\rho C_p$ using the medium data for materials characterization (Ferson *et al.*, 2008).

are not selected with reference to the model of the system nor the SRQ of interest. Instead, the distributions merely summarize the material characterization data based on independent measurements of a measurable property of the system.

We fitted normal distributions to the low and high data sets as well. The computed moments for the three data sets for each parameter are given in Table 13.9.

The experimental data were examined for possible dependence between $k$ and $\rho C_p$ in the data collected during materials characterization. Figure 13.29 shows the scatter plot of these two variables for the medium data set, which reveals no apparent trends or evidence of statistical dependence between the two input quantities. For each value of $k$ measured, the corresponding measured value of $\rho C_p$ is plotted, based on the medium set of materials characterization data. When more than two statistical quantities exist, then the scatter plot is shown for two-dimensional planes through the higher-dimensional space. The Pearson correlation coefficient between these twenty points is 0.0595, which is not remotely statistically significant ($P \gg 0.5$, df $= 18$, where df is the number of degrees of freedom). Because there are no physical reasons to expect correlations or other dependencies between these variables, it is reasonable to assume that these quantities are

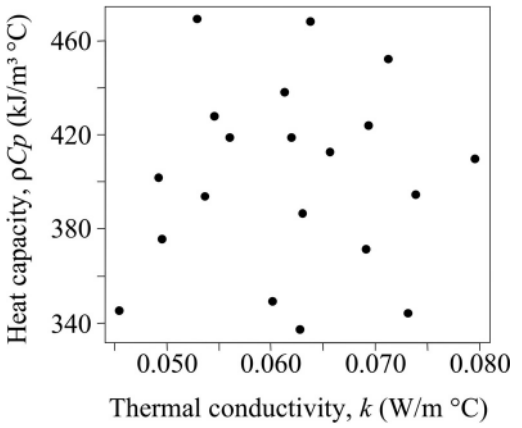


Figure 13.29 Scatter plot of $\rho C_p$ and $k$ obtained from the medium data set for materials characterization (Ferson *et al.*, 2008).

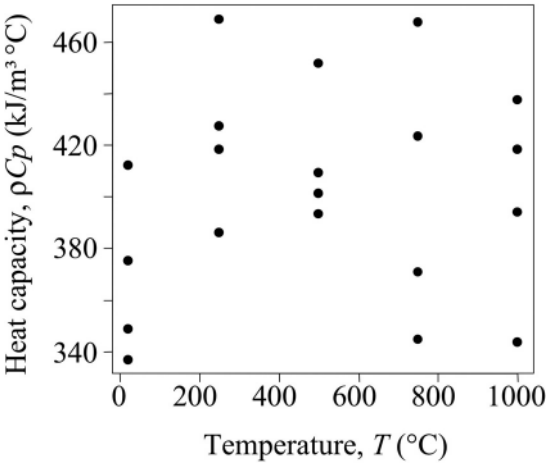


Figure 13.30 Scatter plot of $\rho C_p$ as a function of temperature from the medium data set for materials characterization (Ferson *et al.*, 2008).

statistically independent of one another. Plotting and correlation analysis for the high and low data sets gave qualitatively similar results (Ferson *et al.*, 2008).

### *13.7.2.2 Model uncertainty*

**Possible temperature dependence of material properties**

In the description of the mathematical model for thermal heating $k$ and $\rho C_p$ are assumed to be independent of temperature $T$. It is reasonable to ask whether this assumption is tenable, given the available materials characterization data. Figure 13.30 is the scatter plot for the medium data set for heat capacity, $\rho C_p$, as a function of temperature. Linear and quadratic

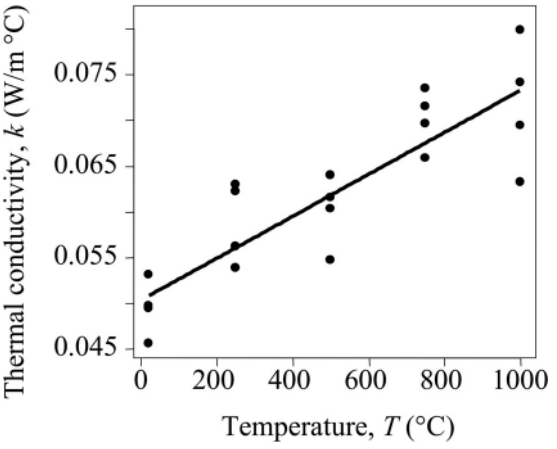


Figure 13.31 Scatter plot of $k$ as a function of temperature from the medium data set for materials characterization (Ferson *et al.*, 2008).

regression analyses reveal no statistically significant trend among these points. The pictures are qualitatively the same for the low and high data sets for $\rho C_p$ in that no trend or other stochastic dependence is evident. Thus, the experimental data for heat capacity support the assumption in the mathematical model.

A similar analysis was conducted for the thermal conductivity data. Figure 13.31 shows the scatter plot of thermal conductivity as a function of temperature for the medium data. The data clearly show a dependence of $k$ on temperature. A linear regression fit was computed using the least squares criterion, yielding

$$k \sim 0.0505 + 2.25 \times 10^{-5} T + N(0, 0.0047) \text{ (W/m-C)}. \tag{13.59}$$

The normal function denotes a normal distribution with mean zero and standard deviation $\sigma$. $\sigma = 0.0047$ is the residual standard error from the regression analysis. This $\sigma$ is the standard deviation of the Gaussian distributions that, under the linear regression model, represent the vertical scatter of $k$ at a given value of the temperature variable. There is no evidence that this trend is other than linear; quadratic regression does not provide a significant improvement in the regression fit.

The medium materials characterization data (Figure 13.31) as well as the low and high data sets, clearly show a dependence of $k$ on $T$. This empirical data show that the model assumption of independence is not appropriate. Weaknesses in models, sometimes severe, are the norm, not the exception, in scientific computing analyses. The important pragmatic question is: what should be done to inform the decision maker of options for possible paths forward? Weaknesses in the modeling assumptions should be clearly explained to the decision maker. He/she may decide to devote time and resources to improve the deficient models before proceeding further with a design or with certain elements of the project. Commonly, however, the decision maker does not have that luxury, so the design and the project must proceed with the uncertainties identified. As a result, the constructive path

forward is to forthrightly deal with the uncertainties and not resort to adjusting the many parameters that are commonly available to the mathematical modeler, the computational analyst, or the UQ analyst.

For the thermal heating problem one option that was explored was to use the dependence of $k$ on $T$ from the materials characterization data directly in the model provided, Eq. (13.58). This, of course, is an ad hoc attempt at repairing the model because Eq. (13.58) would *not* be the solution to the unsteady heat conduction PDE with $k$ dependent on temperature. Even though it is an ad hoc repair of the model, it is *not* a calibration of the model or its parameters to the experimental data for the system, but a use of independent auxiliary data available for a component of the system. A regression fit for the dependence of $k$ on $T$ was computed for each data set, low, medium (Figure 13.31), and high. $k(T)$ from the regression fits and $T(x, t; k)$ in Eq. (13.58) create a system of two equations that can be solved iteratively for each model evaluation as a function of space and time. In this iterative approach, we start with the distribution of $k$ observed in the materials characterization data, and compute from it the resulting temperature distribution through the slab. We then project this distribution of $T$ through the regression function to compute another distribution of $k$. That is, we compute the new distribution $k \sim 0.0505 + 2.25\times10^{-5}T + \mathrm{N}(0, 0.0047)$, where $T$ is the just computed temperature distribution through the slab, and the normal function generates normally distributed random deviates centered at zero with standard deviation 0.0047. As is seen in Figure 13.31, and the low and high materials characterization data sets, the parameters in the normal distribution are independent of the temperature. The resulting distribution of $k$ values conditional on temperature is then used to reseed the iterative process, which is repeated until the distribution of $T$ through the slab converges. We found that only two or three iterations were needed for convergence.

This ad hoc attempt at repair of the model is offered as an alternative model, not as our belief that it is the best approach. We are just exploring this as a simple attempt to see whether it could possibly reduce the model uncertainty. As mentioned above, once the dependence of $k$ on $T$ is found in the materials characterization data, the better physics-based approach is to reformulate the model into a nonlinear PDE with $k$ dependent on $T$, and compute a numerical solution. In the analysis that follows, we present both the results of the specified model, Eq. (13.58), and the ad hoc model for $k$ dependent on $T$. As will be seen next, the weakness in the model due to the assumption of $k$ independent of $T$ will be exposed as model uncertainty.

### Characterization of model uncertainty

The approach used in the present analysis to characterizing model uncertainty is the validation metric operator that is based on estimating the mismatch between the p-box predicted by the model and the p-box from the experimental measurements (see Section 12.8). This metric measures the integral of the absolute value of the difference between the simulation p-box and the experimental p-box. The integral is taken over the entire range of predicted

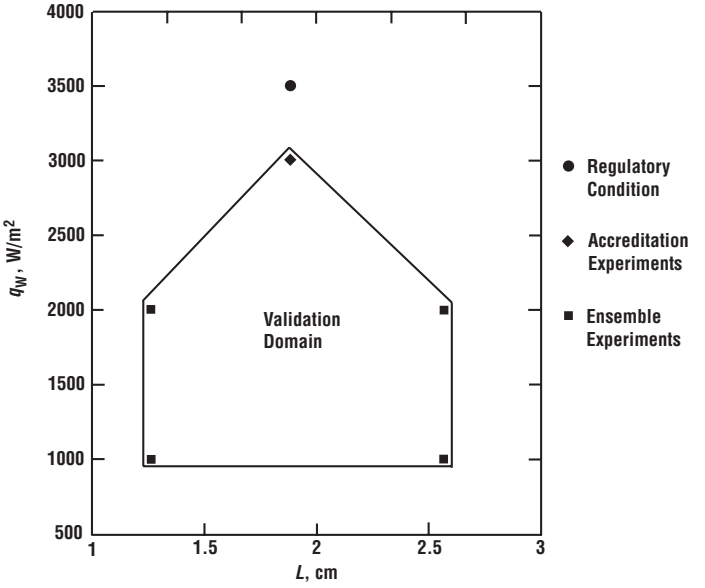


Figure 13.32 Two dimensions of the parameter space for the validation domain and the application condition.

and measured system responses. When no epistemic uncertainty exists in either the simulation or the experiment, such as the thermal heating problem, each p-box reduces to a CDF. Using this metric, we seek the answer to two questions: how well do the predictions match the actual measurements that are available for the system? and what does the mismatch between empirical data and the model's predictions tell us about what we should infer about predictions for which we have no experimental data?

The validation data were divided in the heating problem into ensemble and accreditation data. Figure 13.32 shows these data conditions, along with the regulatory condition where the safety requirement for the system is stated, Eq. (13.57). Only two of the coordinates for the validation domain and the regulatory condition are shown in Figure 13.32, the thickness of the slab, $L$, and the heat flux to the slab, $q_{\mathrm{W}}$. The remaining two coordinates are $x$ and $t$. Validation data for the ensemble and accreditation conditions were taken over the range $0 \leq x \leq L$ and $0 \leq t \leq 1000$ s. The ensemble and accreditation measurements are considered to define the validation domain in the four dimensional space, $q_{\mathrm{W}}$, $L$, $x$, and $t$. As can be seen from Figure 13.32, an extrapolation of the model in the heat flux dimension will be required in order to address the question of the safety requirement at the regulatory condition.

For the ensemble data, temperatures were measured at ten points in time, up to 1000 seconds, and one $x$ position, $x = 0$. The ensemble data are shown in Table 13.10. The conditions for each of the configurations in the ensemble data are shown in Table 13.11.

Table 13.10 *Ensemble data for temperature (°C) at $x = 0$ (Dowding* et al.*, 2008).*

| Time (s) | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 |
|---|---|---|---|---|---|---|---|---|
| | | **Configuration 1** | | | | **Configuration 2** | | |
| 100.0 | 105.5 | 109.0 | 96.3 | 111.2 | 99.5 | 106.6 | 96.2 | 101.3 |
| 200.0 | 139.3 | 143.9 | 126.0 | 146.9 | 130.7 | 140.4 | 126.1 | 133.1 |
| 300.0 | 165.5 | 170.5 | 148.7 | 174.1 | 154.4 | 165.9 | 148.7 | 157.2 |
| 400.0 | 188.7 | 193.5 | 168.5 | 197.5 | 174.3 | 187.2 | 167.7 | 177.2 |
| 500.0 | 210.6 | 214.6 | 186.9 | 219.0 | 191.7 | 205.8 | 184.3 | 194.8 |
| 600.0 | 231.9 | 234.8 | 204.6 | 239.7 | 207.3 | 222.4 | 199.3 | 210.6 |
| 700.0 | 253.0 | 254.6 | 222.0 | 259.9 | 221.7 | 237.6 | 213.0 | 225.0 |
| 800.0 | 273.9 | 274.2 | 239.2 | 279.9 | 235.0 | 251.7 | 225.7 | 238.4 |
| 900.0 | 294.9 | 293.6 | 256.4 | 299.9 | 247.6 | 264.9 | 237.6 | 251.0 |
| 1000.0 | 315.8 | 313.1 | 273.5 | 319.8 | 259.5 | 277.4 | 248.9 | 262.9 |
| | | **Configuration 3** | | | | **Configuration 4** | | |
| 100.0 | 183.1 | 177.8 | 187.2 | 171.3 | 173.4 | 178.9 | 179.3 | 188.2 |
| 200.0 | 247.4 | 240.2 | 254.2 | 231.6 | 234.2 | 241.9 | 242.6 | 254.6 |
| 300.0 | 296.3 | 287.4 | 306.1 | 277.6 | 279.7 | 289.1 | 290.1 | 304.2 |
| 400.0 | 338.7 | 327.8 | 351.9 | 317.4 | 317.4 | 328.4 | 329.6 | 345.4 |
| 500.0 | 378.0 | 364.8 | 395.4 | 354.2 | 350.2 | 362.6 | 363.9 | 381.1 |
| 600.0 | 416.0 | 400.2 | 437.9 | 389.7 | 379.5 | 393.2 | 394.7 | 413.1 |
| 700.0 | 453.5 | 434.8 | 480.0 | 424.5 | 406.3 | 421.1 | 422.8 | 442.1 |
| 800.0 | 490.8 | 469.1 | 522.1 | 459.1 | 430.9 | 446.9 | 448.8 | 469.0 |
| 900.0 | 528.2 | 503.3 | 564.1 | 493.7 | 454.0 | 471.1 | 473.3 | 494.0 |
| 1000.0 | 565.6 | 537.6 | 606.2 | 528.2 | 475.6 | 493.9 | 496.4 | 517.6 |

Table 13.11 *Conditions for the ensemble data (Dowding* et al.*, 2008).*

| Experimental configuration | Data set | Heat flux, $q_W$ (W/m$^2$) | Slab thickness, $L$ (cm) |
|---|---|---|---|
| 1 | low, medium, and high | 1000 | 1.27 |
| 2 | medium and high | 1000 | 2.54 |
| 3 | high | 2000 | 1.27 |
| 4 | high | 2000 | 2.54 |

For the accreditation data, temperatures were measured at 20 points in time, up to 1000 s, at three $x$ positions, $x = 0$, $L/2$, and $L$. The accreditation data are shown in Table 13.12. All of the accreditation data were obtained at $q_W = 3000$ W/m$^2$ and $L = 1.9$ cm. Table 13.13 shows how the accreditation data were segregated into low, medium, and high data sets.

Table 13.12 *Accreditation data for temperature (°C) (Dowding* et al.*, 2008).*

| | Experiment 1 | | | Experiment 2 | | |
|---|---|---|---|---|---|---|
| Time (s) | $x = 0$ | $x = L/2$ | $x = L$ | $x = 0$ | $x = L/2$ | $x = L$ |
| 50.0 | 183.8 | 26.3 | 25.0 | 179.2 | 25.9 | 25.0 |
| 100.0 | 251.3 | 34.0 | 25.1 | 243.9 | 32.2 | 25.1 |
| 150.0 | 302.2 | 47.7 | 26.0 | 292.2 | 44.2 | 25.6 |
| 200.0 | 344.6 | 64.9 | 28.3 | 332.3 | 59.9 | 27.3 |
| 250.0 | 381.7 | 83.9 | 32.7 | 367.1 | 77.5 | 30.6 |
| 300.0 | 414.9 | 103.7 | 39.3 | 398.3 | 96.2 | 35.8 |
| 350.0 | 445.4 | 124.0 | 48.1 | 426.7 | 115.3 | 42.9 |
| 400.0 | 473.6 | 144.4 | 58.7 | 452.9 | 134.7 | 52.0 |
| 450.0 | 500.0 | 164.9 | 71.1 | 477.4 | 154.2 | 62.7 |
| 500.0 | 525.0 | 185.4 | 84.9 | 500.5 | 173.6 | 74.9 |
| 550.0 | 548.8 | 205.9 | 100.0 | 522.4 | 193.1 | 88.4 |
| 600.0 | 571.7 | 226.3 | 116.1 | 543.4 | 212.5 | 103.1 |
| 650.0 | 593.8 | 246.8 | 133.0 | 563.5 | 231.8 | 118.7 |
| 700.0 | 615.2 | 267.2 | 150.7 | 583.0 | 251.1 | 135.0 |
| 750.0 | 636.1 | 287.6 | 169.0 | 602.0 | 270.3 | 152.1 |
| 800.0 | 656.6 | 307.9 | 187.8 | 620.4 | 289.5 | 169.7 |
| 850.0 | 676.7 | 328.3 | 207.0 | 638.5 | 308.7 | 187.8 |
| 900.0 | 696.4 | 348.6 | 226.5 | 656.3 | 327.8 | 206.3 |
| 950.0 | 716.0 | 369.0 | 246.3 | 673.9 | 346.9 | 225.1 |
| 1000.0 | 735.4 | 389.3 | 266.3 | 691.2 | 366.0 | 244.2 |

Table 13.13 *Conditions for the accreditation data (Dowding* et al.*, 2008).*

| Experiment | Data set | Heat flux, $q_W$ (W/m$^2$) | Slab thickness, $L$ (cm) |
|---|---|---|---|
| 1 | low, medium, and high | 3000 | 1.9 |
| 2 | high | 3000 | 1.9 |

The type of data obtained in the ensemble and accreditation experiments is typical of more complex systems in the sense that data are obtained for a variety of conditions and different values of the dependent variables in the PDEs, but there is insufficient data to adequately assess the accuracy of the model. Stated differently, there are usually very few replicated experiments of the system, if any, at the various operating conditions tested so that system variability can be more accurately separated from assessing model accuracy. Project management on real engineering systems tends to place much more emphasis on better understanding system performance than assessing the predictive capability of the model. Using the validation metric approach of comparing CDFs from prediction and

experiment, we could compute a validation metric result, $d$, for every paired comparison of prediction and experiment. If we computed a large number of simulations for each condition of the experiment, we could compute a smooth CDF for the simulation. The difficulty, however, is that for the vast majority of experimental conditions, we have only one experimental measurement so that we only have a single-step function for the empirical CDF. For the ensemble and accreditation data there are a few exceptions where replicate experiments are available. For example, in the ensemble data for the medium data, there are 20 conditions where four replications are available (see Table 13.10 – Table 13.11). These occur for configurations 1 and 2. For the accreditation data, no replications occur for the medium data (see Table 13.12 and Table 13.13). Only using data where several experimental replicates are available would be waste of valuable data.

For the cases where we only have one experimental measurement, we will obtain an inflated estimate of the true value of $d$ (obtained from a large number of experiments) because the metric measures the *evidence for mismatch* between the simulation and the experiment. As discussed in Section 12.8.2, the inflation of the metric is a direct result of the very poor representation of the true CDF (with a large number of experiments) compared to the empirical CDF with only one experimental measurement. This poor representation is due to sampling uncertainty, which is an epistemic uncertainty. To reduce this perpetual problem of conflicting experimental goals of understanding how a system responds to various inputs versus assessing the predictive capability of a model, the concept of $u$-pooling was developed by Ferson *et al.* (2008). This concept was discussed in Section 12.8.3 and it can be used for two different situations: (a) when we have a single measurement of an SRQ at a wide variety of experimental conditions; and (b) when we have a single experimental measurement of *different* SRQs, all of which are predicted by the model. As an example of this last case, suppose the thermal heating model was coupled with a thermal stress model so that both temperature and stress were predicted by the coupled model. If both temperatures and stresses were measured in the slab during the same experiment, then we could combine these different SRQs into a common SRQ so that the accuracy of the models could be better assessed.

The $u$-pooling transforms the experimental measurement and the model prediction for each comparison to probability space, and then uses an inverse transform to return to a transformed measurement coordinate. The inverse transform is based on the model prediction at the application condition of interest so that all of the measurements are transformed in the same way to conditions directly relevant to the application.

We will briefly explain the $u$-pooling concept before applying it to the thermal heating problem. Figure 13.33 shows a comparison of a single experimental measurement, $y_j$, with a distribution for the model prediction for the physical conditions of the experiment and the value of the dependent variables of the PDE appropriate to location of the measurement. The model prediction is shown here as a smooth CDF with no epistemic uncertainty, but it could be a p-box if epistemic uncertainty existed in the model. To obtain a smooth model prediction, one needs to compute a significant number of samples using MCS or LHS. If a set of measurements is made and only the independent variables of the PDE are

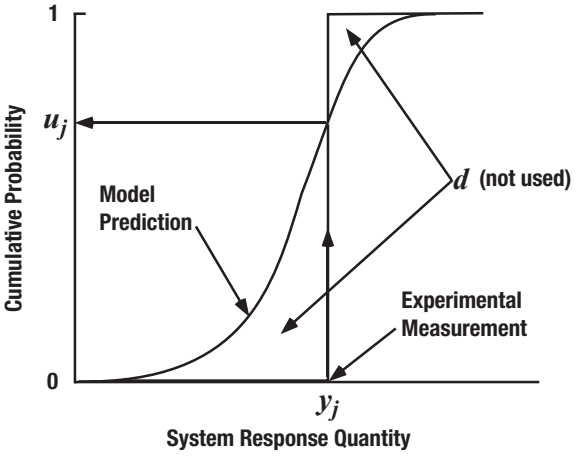


Figure 13.33 Transform of an experimental measurement using the model to obtain a probability.

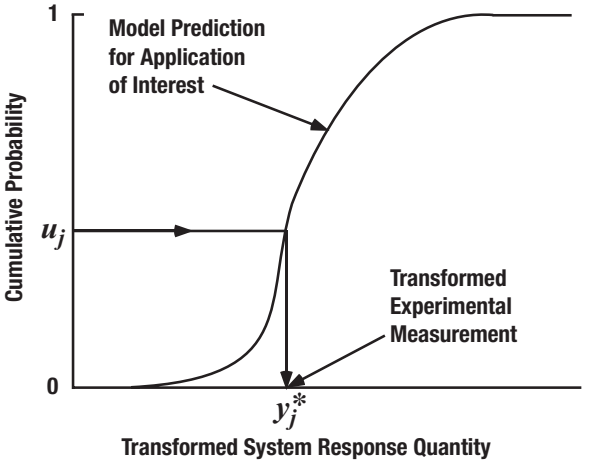


Figure 13.34 Inverse transform using the model for the conditions of the application of interest to obtain a transformed experimental measurement.

different for the individual measurements, then the same simulation can be used for the measurements. If, however, the conditions of the experiment are different, e.g., the input data are different, then each condition requires that a separate CDF be computed for each condition. As shown in the figure, a $d$ can be computed directly from this comparison, but the $d$ value would be inflated, as discussed earlier. Using the concept of $u$-pooling, we transform the physical measurement, $y_j$, into a probability, $u_j$, using the CDF predicted by the model.

Figure 13.34 depicts how the inverse transform takes the probability from Figure 13.33 and maps it into a transformed measurement of the SRQ of interest. For this inverse

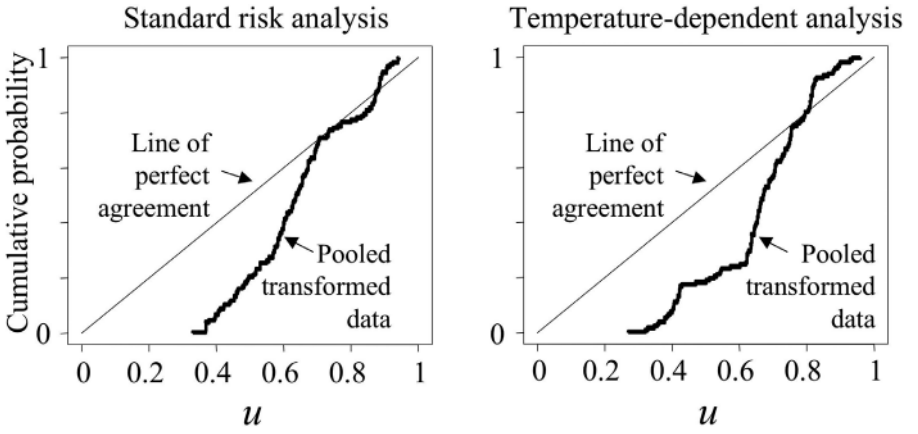


Figure 13.35 Model performance for the medium data set for two analyses as summarized by observed distribution functions of $u$-values (step functions) compared to uniform distributions for perfect agreement (straight lines) (Ferson *et al.*, 2008).

transform, we use the model prediction for the conditions of the application of interest. The same inverse transform is used for all of the probabilities resulting from individual measurements and individual conditions. The inverse transform is simply a consistent monotone rescaling of all of the probabilities. One may criticize the use of the model in the rescaling because the accuracy of the model is what we are trying to assess. This criticism is ineffective because, whether the model is accurate or not, the model is simply used to rescale the measurements in a consistent manner. With all of the original experimental measurements transformed in this way, one can then compute a smoother empirical CDF representing the measurements as if they were mapped to the application condition of interest. One can then compute a $d$ by comparing the CDF from the prediction and the empirical CDF.

We now assess the performance of the thermal heating model over the validation domain by using the method of $u$-pooling. We compute a summary assessment of the overall mismatch between the model and the data by using all the ensemble and accreditation data. Although the ensemble and accreditation data were segregated in the formulation of the validation challenge problem, our analysis dealt with both types of data in the same way. Every observation in the validation domain is associated with values of the control parameters $q_W$, $L$, $x$ and $t$. These four parameters, along with the distributions characterizing the variability of $k$ and $\rho C_p$, were used in 10 000 Monte Carlo samples to compute the prediction distribution of temperatures using Eq. (13.58). Every temperature observation in the validation domain is thereby paired with a prediction distribution of temperature. There are 140 of these pairs in the medium data set. The pairs define the $u$-values $u_j = F_j\,(T_j)$, where $T_j$ is the observed temperature and $F$ denotes the associated prediction distribution and $j = 1,2,\ldots,140$.

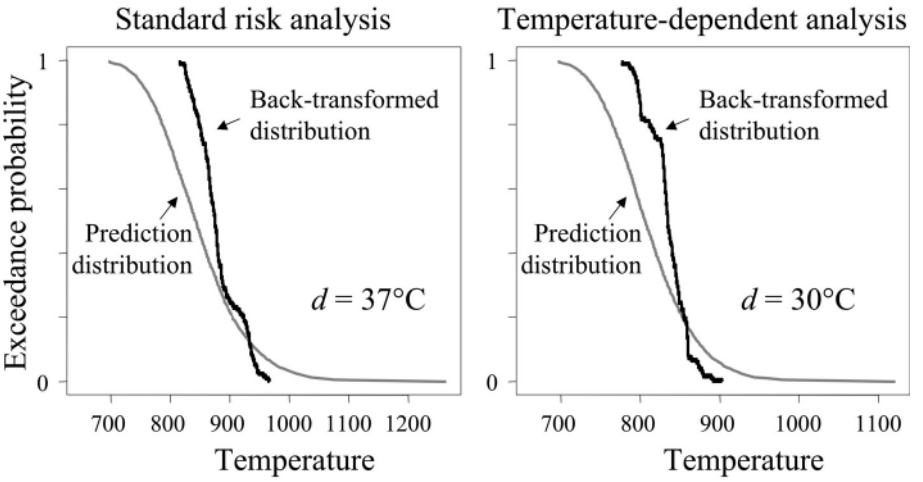


Figure 13.36 Distributions for the medium data set of back-transformed $u$-values (step functions) compared to predicted temperature distributions (smooth lines) for the two analyses (Ferson *et al.*, 2008).

Figure 13.35 shows empirical distributions of these $u$-values using the medium data set. These empirical distributions constitute summaries of the performances of the standard model, Eq. (13.58), and the model for $k$ dependent on temperature. The small step function in each graph represents all 140 observations of temperature in the medium data set compared to their respective prediction distributions generated under that analysis. The step functions are the empirical distributions of the $u$-values produced by these comparisons. These distributions are to be compared against the uniform distribution over [0, 1], whose graph appears in each graph as the 45° line. These 45° lines are lines of perfect agreement. If the observed temperatures were sampled from distributions matching those that were predicted by the model under an analysis, then the step function would match the uniform distribution to within fluctuations due to random chance.

Figure 13.36 shows the back-transformations of these distributions to the physically meaningful scale of temperature. The figure is shown in terms of the exceedance probability, which is the complementary CDF, versus temperature for both $k$ independent and $k$ dependent on temperature. Exceedance probability is commonly used in risk assessment studies to obtain clearer information about the probability of exceeding specified thresholds. The smooth curves in the graphs are the prediction distributions under the two analyses for the conditions of the regulatory requirement. Also displayed in Figure 13.36, as step functions, are the corresponding pooled distributions of $u$-values *back-transformed onto the same temperature scale* via the inverse probability integral transforms specified by the respective prediction distributions. That is, they are the distributions of the quantity $G^{-1}(u_j)$, where $G$ is the prediction distribution on which the regulatory requirement is stated.

The graphs in Figure 13.36 are simply nonlinear rescalings of the graphs in Figure 13.35. Under these rescalings, the straight lines become the smooth curves, and the tails of the

Table 13.14 *Validation metric results over the validation domain using the medium data set (°C) (Ferson* et al.*, 2008).*

| Analysis | $d$, validation metric result | 95% confidence interval |
|---|---|---|
| Model with temperature independent $k$ | 37.2 | [34.0, 42.7] |
| Model with temperature dependent $k$ | 30.4 | [27.4, 33.7] |

step functions are stretched relative to central values of the distributions. For each graph, the transformation of the abscissa is exactly the one that changes the standard uniform distribution into the prediction distribution. The result translates the evidence embodied in all 140 observations in the ensemble and accreditation data sets onto the scale defined by the prediction distribution for the regulatory requirement, Eq. (13.57). These distributions should not be interpreted as though they were themselves actual data collected on this transformed temperature scale. They were, after all, collected for a variety of heat fluxes, various slab thicknesses, different time values, and different locations in the slab. They are pooled for the sake of estimating the predictive accuracy of the model, where data are available. Thus, they do not represent direct evidence about what the temperatures will be under the conditions of the regulatory requirement, but only how well the prediction distributions produced by the model have matched temperatures for all of the data in the medium data set validation domain.

The temperature-dependent analysis, illustrated in the right graph of Figure 13.36, has a somewhat better match with the available empirical data than the standard analysis, illustrated in the left graph. The distribution of back-transformed $u$-values is closer in this analysis to its prediction distribution shown as the smooth curve. The superiority of this match is reflected by the area metric $d$ in Table 13.14.

Recall that the area metric $d$ measures the area between the empirical distribution of back-transformed pooled $u$-values and the prediction distribution, which is our expectation about them. The 95% confidence intervals were computed by a nonparametric bootstrap method based on resampling from the 140 $u$-values. (For a discussion of the bootstrap technique, see Efron and Tibshirani, 1993; Draper and Smith, 1998; and Vose, 2008). That is, we took a random sample of size 140 from the distribution of $u$-values (with replacement) and recomputed the value of $d$ by comparing the distribution of these randomly selected $u$ values to the prediction distribution. We repeated this process 10 000 times and sorted the resulting array of $d$-values. The 95% confidence interval was estimated as $[d_{(2.5N/100)}, d_{(N\text{-}(2.5/100)N)}] = [d_{(250)}, d_{(9750)}]$, i.e., the interval between the 250th and 9750th values from the sorted list of 10 000 $d$ values. Each confidence interval estimates the sampling uncertainty associated with the actual value of $d$ arising from having computed it from only 140 observations.

Table 13.15 *Validation metric results and the 95% confidence interval over the validation domain for the low, medium, and high data sets (°C) (Ferson* et al.*, 2008).*

| Analysis | Low data set ($n = 100$) | Medium data set ($n = 140$) | High data set ($n = 280$) |
|---|---|---|---|
| Model with temperature independent $k$ | 52.6, [49.4, 55.9] | 37.2, [34.0, 42.7] | 18.5, [15.3, 23.9] |
| Model with temperature dependent $k$ | 34.1, [30.0, 38.1] | 30.4, [27.4, 33.7] | 11.6, [9.5, 15.0] |

Both of the graphs in Figure 13.36 suggest that the standard model, Eq. (13.58), is somewhat better at predicting temperatures close to 900 °C than it is at predicting lower temperatures. The temperature-dependent analysis used regression and an iterative convergence scheme to represent the dependence evident in the materials characterization data between temperature and thermal conductivity. The reward for the extra iterative calculation is a modest improvement in the model's assessed performance vis-à-vis the total amount of data. The match for this analysis was quantitatively better than that of the standard risk analysis, Table 13.14.

Calculations were also done with the low and high data sets, which yielded qualitatively similar results. The performances of the model, in terms of the validation metric, under the two analyses for all three data sets are shown in Table 13.15. Also shown in the table is the interval of the metric $d$ using the bootstrap technique and a 95% confidence level. The temperature-dependent analysis was somewhat better than the standard risk analysis for both data sets in the area metric. It should be stressed that validation metrics shown are measures of the mismatch, i.e. inaccuracy, for each of the two models, for different quantities of data, and for *all experimental measurements* made in the validation domain of the indicated data set. This kind of data is valuable for quantitatively comparing the performance of various competing models with the same experimental data.

Figure 13.37 summarizes the validation metric performance results graphically for the low and high materials characterization data. Distributions are shown of the back-transformed $u$-values compared to the respective prediction distributions for the two analyses.

Table 13.15 and Figure 13.37 reveal that the area metric $d$ as used in the validation assessment is strongly sensitive to the sample size of the observations on which it is based. This is not because the model is getting more accurate with more data. Rather, it is the result of there being *less evidence of a mismatch* between the model and the data. Of course, this might not have been true if the model were making very inaccurate predictions. Insofar as assessing model accuracy is concerned, increasing the sample size tends to decrease $d$ to its lower limit of mismatch. The dependence of $d$ on the number of observations means that we should only compare performances of models that are based on the same sample size. In this case, the temperature-dependent analysis has a consistently better (smaller) overall $d$ than the standard analysis at all three sample sizes described in the challenge problem.

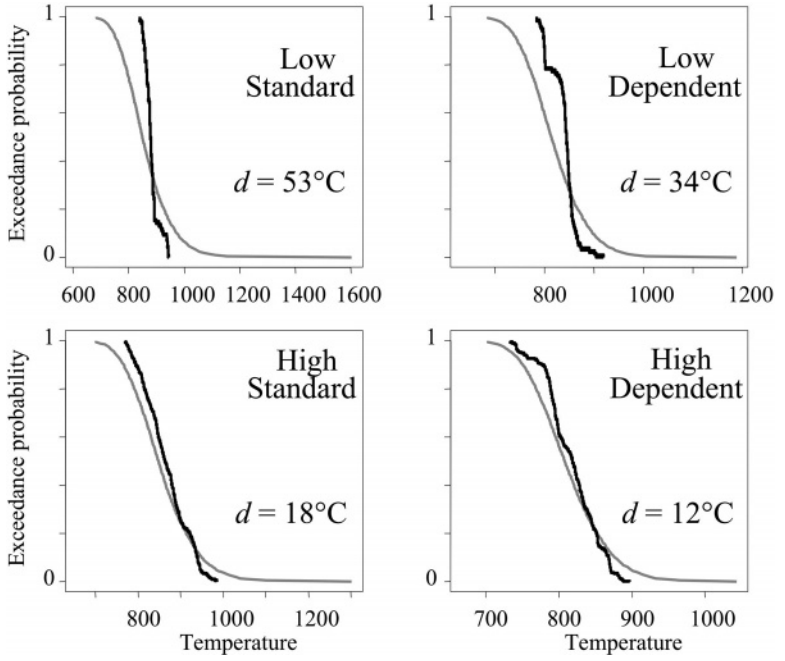


Figure 13.37 Model performances as distributions of back-transformed $u$-values (step functions) compared to prediction distributions (smooth curves) for low and high data sets and the standard and temperature-dependent analyses (Ferson *et al.*, 2008).

### *13.7.3 Step 4: estimate output uncertainty*

Since there is no numerical solution error for the thermal heating problem, output uncertainty need only combine input and model uncertainty. We first discuss the general approach for combining input and model uncertainty using the validation metric based on the mismatch in p-boxes (or CDFs) from prediction and experimental CDFs. Then we apply this approach to the thermal heating problem.

#### *13.7.3.1 General discussion of combining input and model uncertainty*

Since the validation metric approach using the mismatch of p-boxes is directly applicable to a higher dimensional space of input parameters than the confidence interval approach, we begin our discussion in this vein. A function representing the model uncertainty must be constructed over the $m$-dimensional input space of the model. Let $\xi$ be the number of input parameters over which experimental data has been obtained at different conditions during validation activities. Assume the subset of parameters for which experimental data were obtained is listed as the first $\xi$ elements of the array of input quantities, so that we have $(x_1, \ldots, x_\xi, x_{\xi+1}, \ldots, x_m)$. Let $d(x_1, \ldots, x_\xi)$ be the validation metric result that quantifies the model uncertainty over the dimensions where there are experimental data, $(x_1, \ldots, x_\xi)$. Let $\nu$ be the total number of experimental measurements of any of the SRQs made during the validation experiments.

As discussed in Section 12.8, the number of individual $d$ values that can be computed can range from 1 to $\nu$. This is an advantage over the confidence interval approach, but a more important advantage is that when multiple measurements are available, they need not have *any* relationship to one another in the space of $(x_1, \ldots, x_\xi)$. For example, in a particular analysis we could have $(x_1, x_2, x_3, x_4)$ represent the independent variables in the PDE (three spatial coordinates and time), $x_5$ represent a BC, and $(x_6, x_7, x_8, x_9)$ represent four design parameters of the system. We could have one or more experimental measurements from validation experiments in all nine dimensions; none having a commonality or relationship to any of the other measurements.

Since there is no restriction on how many $d$ values can be evaluated, other than that the number is no larger than $\nu$, we consider here the two different ways the available experimental measurements could be used. One could maximize the total number of $d$ values computed by treating each paired experimental measurement and simulation separately. The advantage is that we maximize the number of $d$ values over the space of $(x_1, \ldots, x_\xi)$ with which to construct some type of interpolation or regression function for $d(x_1, \ldots, x_\xi)$. The disadvantage is that by treating each experimental measurement separately, we inflate the validation metric $d$ due to sampling (epistemic) uncertainty. That is, the evidence for mismatch will tend to be greater using fewer experimental measurements compared to using a larger number of measurements. Note that even if the model is *perfect*, in the sense that the model can exactly predict the measured CDFs of the system responses from the experiment, there is still inflation of the validation metric due to limited sampling.

On the other end of the spectrum, if we use $u$-pooling to combine all of the measurements and their respective comparisons with the model, we can end up with one $d$. The advantage of this computed $d$ is that it is a combined or summary $d$ over the entire set of data forming the validation domain. This was the method used in Section 13.7.2.2. This approach is the best representation of the evidence for mismatch between the model and the measurements that can be obtained with the available amount of empirical data. For example, suppose the model is perfect in the sense just mentioned. As a result, when an increasing number of measurements are used to assess the accuracy of the perfect model, we will have $d \to 0$. The disadvantage is that as we combine more data into the evaluation of $d$, we have *less* data with which to construct the function $d(x_1, \ldots, x_\xi)$. That is, an empirical measurement can be used either to better assess the accuracy of the model by contributing to an estimate of $d$, or it can be used in the construction of the hypersurface $d(x_1, \ldots, x_\xi)$ ; but not both. One could investigate the question of the optimum use of the experimental data so as to have the maximum decrease in $d$ with the minimum effect on the inaccuracy in the construction of $d(x_1, \ldots, x_\xi)$. This question, however, has not been investigated.

Now consider the question of how to construct $d(x_1, \ldots, x_\xi)$, whatever balance one chooses between the precision of the model accuracy assessment and the ability to construct $d(x_1, \ldots, x_\xi)$. A reasonable method of constructing this function is to compute a least squares regression fit of the computed $d$ values using a low-degree polynomial. Let $\eta$ be

the number of data points available to compute the regression function, where $\eta \leq \nu$. Note that if we use a first-degree polynomial, we must have at least two data points in each dimension $x_i$. The following situations suggest that a first-degree polynomial should be used instead of a higher-degree polynomial: (a) if $\eta$ is not much greater than $2\xi$, and (b) if a large extrapolation of the regression function beyond the validation domain is needed in order to evaluate $d(x_1, \ldots, x_\xi)$ at the application conditions of interest.

Let the polynomial regression fit for $d(x_1, \ldots, x_\xi)$ be given by

$$d(x_1, x_2, \ldots, x_\xi) \sim \mathcal{P}(x_1, x_2, \ldots, x_\xi) + N(0, \sigma). \tag{13.60}$$

$\mathcal{P}(x_1, x_2, \ldots, x_\xi)$ is the polynomial regression function and $N(0, \sigma)$ is the normally distributed residual error arising from the scatter in the fit to the individual $d$ values. This regression fit, of course, has no knowledge of how the model uncertainty might behave in the remaining $m - \xi$ dimensions of the input space. As a result, $d$ is assumed independent of these remaining parameters. The polynomial portion of Eq. (13.60) will result in a single value of $d$ at each condition of interest. The uncertainty in the value of $d$ should be expressed by using the *statistical prediction interval* based on $N(0, \sigma)$. A prediction interval should be used to represent: (a) the scatter in the experimental data and/or model data, and (b) the scatter of the model uncertainty not captured by the choice of the regression function. A prediction interval should be used instead of a confidence interval because here we are interested in estimating the uncertainty in a *single future value of the uncertainty*. A confidence interval, however, is an estimate of the uncertainty in the mean of a collection of data. Prediction intervals are always larger than confidence intervals, and they do not approach zero as the number of samples becomes larger (Devore, 2007). A reasonable confidence level for the prediction interval would be either 90% or 95%. Because we are trying to capture the epistemic uncertainty in the model as compared to the experimental data, as well as the aleatory uncertainty due to measurement uncertainty, we should always use the upper bound from the prediction interval.

Once the regression function $\mathcal{P}(x_1, x_2, \ldots, x_\xi)$ is computed for the conditions of interest, we can use a procedure similar to that discussed in Section 13.4.2.1 to combine input and model uncertainty. One of the important differences between the metric based on the mismatch of p-boxes and the confidence interval approach for estimating model uncertainty is that the mismatch approach measures the absolute value of the difference between the p-boxes. As a result, there is *no* sign associated with the mismatch measure, leaving no indication whether the model prediction is higher or lower than the empirical measurements. Consequently, when model uncertainty is combined with the p-box due to input uncertainty, the contribution due to model uncertainty expands the left and right sides of the p-box equally. Let $\mathrm{PI}^{\mathrm{u}}(x_1^*, x_2^*, \ldots, x_\xi^*)$ denote the upper bound of the prediction interval evaluated at the conditions of the application of interest, $(x_1^*, x_2^*, \ldots, x_\xi^*)$. We then have for the magnitude of the displacement on the left and right of the input uncertainty p-box

$$d(x_1^* x_1^*, x_2^*, \ldots, x_\xi^*) \sim \mathrm{PI}^{\mathrm{u}}(x_1^*, x_2^*, \ldots, x_\xi^*). \tag{13.61}$$

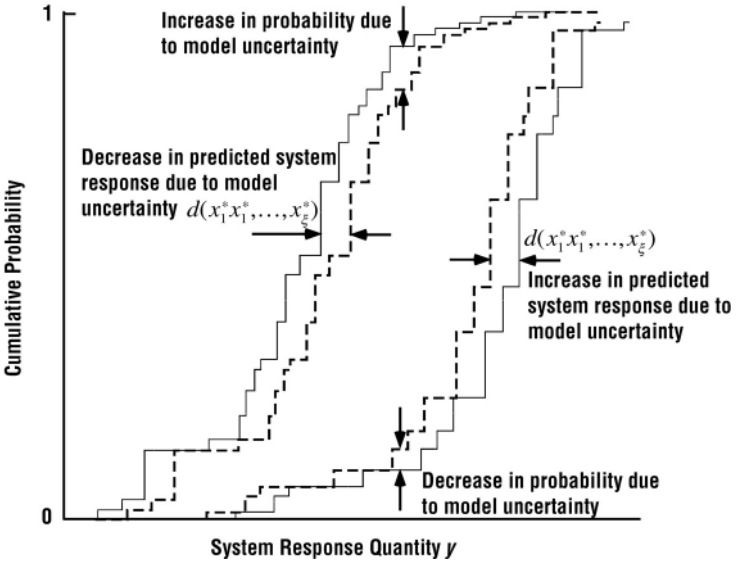


Figure 13.38 Increase in predictive uncertainty due to the addition of model uncertainty.

Figure 13.38 gives an example of how a p-box due to a combination of aleatory and epistemic input uncertainty is expanded due to model uncertainty. As can be seen in the figure, the $d(x_1^*, x_2^*, \ldots, x_\xi^*)$ value that is added to and subtracted from the SRQ p-box is a constant over the entire response. However, the increase and decrease in the probability of various responses will depend on the shape of the p-box. It may appear that the primary effect of accounting for model uncertainty is near the median, $P = 0.5$. This simply occurs because of the probability-constrained shape of CDFs; probability must be bounded by zero and one. Model uncertainty also has a large effect on low and high cumulative probabilities. For example, an undesirable system response can easily change from one in a thousand to one in fifty, as suggested in the left tail of the distribution shown in Figure 13.38.

### *13.7.3.2 Combining input and model uncertainty for the thermal heating problem*

The regulatory condition, Eq. (13.57), is outside the validation domain, defined by the ensemble and accreditation data. As a result, extrapolation of the validation metric, as well as model, is required to determine if the regulatory condition is satisfied. We employed linear regression to extrapolate in the four dimensional space, $q_W$, $L$, $x$, and $t$, to estimate the model uncertainty at the regulatory condition. Here we describe this extrapolation for the temperature-dependent risk analysis under the medium data set (140 measurements). Entirely analogous calculations are possible for the constant conductivity model and the other data sets.

Figure 13.39 shows the model prediction at the regulatory condition for both the temperature-dependent model and the constant conductivity model only accounting for input uncertainty. The figure shows the uncertainty in terms of the exceedance probability,

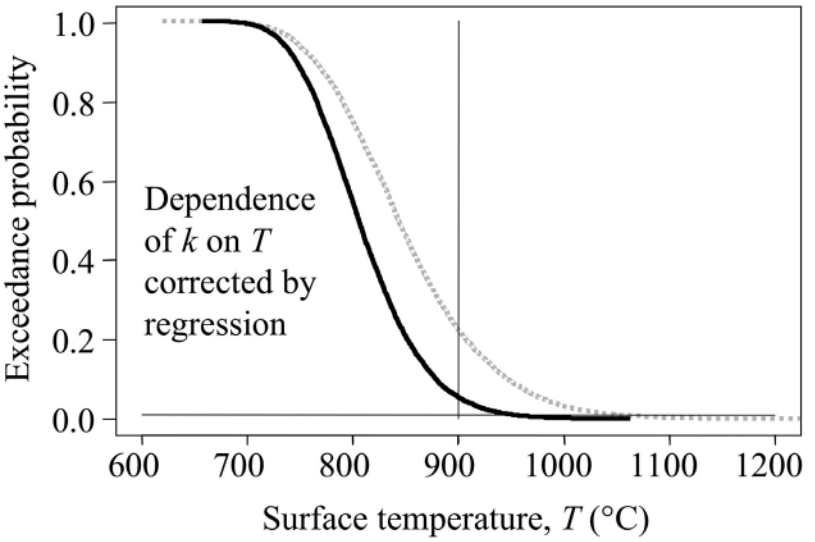


Figure 13.39 Model prediction for exceedance probability at the regulatory condition for the temperature-dependent model (solid line) and the conductivity independent of temperature model (dotted line) (Ferson *et al.*, 2008).

i.e., the complimentary cumulative distribution function (CCDF). The CCDF shows the fraction of the population that would have a response greater than a particular value of the response. Since the regulatory criterion, Eq. (13.57), for the thermal heating analysis is written in terms of an exceedance probability, it is appropriate to show analysis results in this way. Given that the input uncertainties in $k$ and $\rho C_p$ are only aleatory, both model predictions are only aleatory. Note that the predicted exceedance probability for the constant conductivity and the temperature-dependent model, 0.22 and 0.05, respectively, both exceed the safety criterion of 0.01. When the model uncertainty is combined with the input uncertainty, only more uncertainty will be included in the predicted result.

There are various ways to construct the regression function for model uncertainty based on the validation data. As discussed in the previous section, one approach is to compute the metric at the conditions where each of the 140 measures were made. Each of these metrics would be based on single measurements, so they would be inflated estimates of the mismatch of the model vis-à-vis the experiment. With these 140 values spread over the $q_W$, $L$, $x$, and $t$ space, one could compute a linear regression hypersurface. Then one could compute the 95% upper prediction interval at the regulatory condition, $q_W = 3500$ W/m$^2$, $L = 1.9$ cm, $x = 0$, $t = 1000$ s. Using $\mathrm{PI}^u(x_1^*, x_2^*, \ldots, x_\xi^*)$ and the computed CDF resulting from the temperature dependent model, one could estimate the combination of input uncertainty, shown in Figure 13.39, and the model uncertainty using the procedure discussed in Figure 13.38. This would be a reasonable and defensible approach.

Another approach is used because it improves the relevancy of the individual $d$ values to the regulatory condition (Ferson *et al.*, 2008). Instead of pooling these 140 $u$-values, as

we did in the calculation of the summary $d$ in Section 13.7.2.2, we back-transform each $u$-value directly to the temperature scale using the prediction distribution associated with the conditions of the regulatory requirement. That is, we compute the back-transformed temperature $T_j^* = G^{-1}(F_j\ (T_j))$, where $T$ denotes observed temperature, $F$ denotes the associated prediction distribution, $G$ is the distribution from the model with only variability of the input parameters, and $j = 1,2,\ldots,140$ indexes the observations. We then compute the area metric $d(T_j^*, G)$ between the back-transformed temperature and the prediction distribution of regulatory interest for each observation–prediction pair. These comparisons of scalar values and the prediction distribution for temperature at the conditions for the regulatory requirement thus yielded 140 values of the area metric. The mean of these areas was 69, with values ranging between 56 and 129. The primary reason for the large increase in the mean and the scatter, compared to the values listed in Table 13.14, is that this mean is based on individual $d$ values for all of the 140 points, as opposed to pooling all of the $d$ values.

The 140 $d$ values were regressed against the input variables $q_W, L, x, t$ with a linear model for extrapolation. The best fitting linear model for the expected value of the area metric for a single (new) observation as a function of heat flux, thickness, position and time is

$$d \sim 126 - 0.016q_W - 914L + 201x - 0.0124t + N(0, 10.8) \quad ^\circ\text{C}. \tag{13.62}$$

The last term is the normally distributed residual error arising from random scatter in the area values. At the regulatory requirement, this becomes

$$d^* \sim 40.2 + N(0, 10.8) \quad ^\circ\text{C}. \tag{13.63}$$

Because we are extrapolating to conditions specified in the regulatory requirement for which there is no experimental data, it is especially important to account for sampling uncertainty in predicting the magnitude of the area metric under those conditions. The 95% prediction interval for the value of the area metric at the regulatory conditions is computed to be

$$d^* \sim [17, 63] \quad ^\circ\text{C}. \tag{13.64}$$

Because we are trying to capture the epistemic uncertainty of the model's predictions vis-à-vis the data, we use the *upper* bound from the prediction interval as our estimate of the upper bound on the uncertainty of the model, 63 °C.

Figure 13.40 shows the model prediction after combining the input uncertainty and the estimate of model uncertainty. The graph shows a prediction distribution as the inner curve, representing only the input uncertainty, with a parallel distribution on either side of it displaced by 63 °C. The central prediction distribution uses the medium data set for the material characterization and the temperature dependent model. In our view, the p-box prediction represents a reasonable estimate of the combined input and model uncertainty associated with the predictive capability of the model as evidenced by the validation assessment on available data. The expectation is that future data, if it became available, would form a distribution

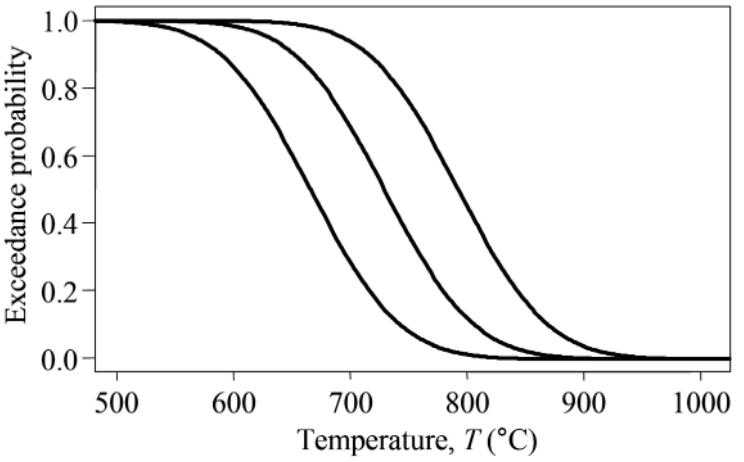


Figure 13.40 P-box prediction after combining input and model uncertainty for the temperature-dependent model using the medium data (Ferson *et al.*, 2008).

within these bounds. In the mid temperature ranges, e.g., 700 to 800 °C, the increase in uncertainty due to model uncertainty is stunning. To many analysts, this method of increasing the predictive uncertainty would appear to be excessive. However, we argue that it is a highly defensible representation of predictive capability that is based directly on the observed accuracy of the model. If data became available at the regulatory condition it could be outside of the p-box shown in Figure 13.40. The p-box *does not guarantee* that the actual result will be inside the p-box. It is a statistical inference based on the model and observations that are available.

An argument can be made for a simpler method for estimating the model uncertainty at the regulatory condition. In Section 13.7.2.2, model uncertainty was estimated using the ensemble and accreditation data and the method of *u*-pooling. In the *u*-pooling process, all of the measurements were transformed to the regulatory condition to compute a summary *d*. One can then argue that the upper bound on the 95% confidence interval can be used as the estimated model uncertainty at the regulatory condition. Referring to Table 13.15, we see that the summary *d* and the 95% confidence interval for the temperature-dependent model at the regulatory condition are 30.4°C and [27.4, 33.7]°C, respectively. One could then treat the 33.7 °C as the model uncertainty at the regulatory condition, $d^*$. This would be about half of the model uncertainty estimated by using the 95% upper bound on the prediction interval from extrapolating the linear regression. 33.7 °C may seem unrealistically low compared to the 63 °C value but we know that this latter value is inflated due to using single experimental measurements to compute *d*. Recall from Table 13.15 that the 95% upper bound on the confidence interval for the temperature-dependent model decreased from 38.1 °C to 15.0 °C as the quantity of data changed from the low data set to the high data set. As a result, the change in the *d* from 63 °C to 33.7 °C between the two different approaches is consistent. Since the concept of *u*-pooling has only recently been published,

Table 13.16 *Probability of failure of the safety component using various heating models and different quantities of experimental data (Ferson* et al.*, 2008).*

| Analysis | Low data set | Medium data set | High data set |
|---|---|---|---|
| Input uncertainty only, temperature independent $k$ | 0.28 | 0.22 | 0.24 |
| Input uncertainty only, temperature dependent $k$ | 0.09 | 0.05 | 0.05 |
| Input and model uncertainty, temperature independent $k$ | [0.13, 0.52] | [0.17, 0.29] | [0.17, 0.32] |
| Input and model uncertainty, temperature dependent $k$ | [0.04, 0.18] | [0.03, 0.09] | [0.03, 0.09] |

additional research is needed to investigate the reliability of using a summary $d$ on a wide range of problems.

#### *13.7.3.3 Predicted probabilities for the regulatory condition*

The regulatory criterion was defined in Eq. (13.57) as

$$P\{T(x = 0, t = 1000) > 900\} < 0.01. \tag{13.65}$$

Table 13.16 gives the summary of results for the failure probabilities, i.e., exceedance probabilities, from the various analyses that were conducted. The first two rows are the results of an analysis that would only consider input uncertainty, i.e., variability in $k$ and $\rho C_p$. This procedure is the standard risk assessment approach that is practiced today. It does not take into account an estimate of the uncertainty in the model that could be inferred from experiments on systems that are similar to the actual system of interest. The standard risk assessment approach predicts that the system safety would be inadequate, but there is a wide difference in results depending on whether $k$ is considered independent or dependent on temperature. If an analyst were to perform both of these analyses and observe the large difference in predicted safety, they would properly suspect that the weakness in the model was due the assumptions concerning $k$. Unless a new nonlinear heat transfer analysis was formulated and computed, however, they would be at a loss to estimate the uncertainty due to model form.

The last two rows of Table 13.16 are the results from the present approach that takes into account input and model uncertainty. The square brackets are the bounds on the failure probabilities from the model when the validation metric results are included in the analysis. The bounds were obtained from the displacement to the left and right of the response distribution due to input variability alone (see Figure 13.40). The predicted interval-valued bounds were assessed by extrapolating the observed uncertainty of the predicted temperature distributions to the regulatory condition. Comparing these probability intervals to

Table 13.17 *Probability of failure of the safety component reported by Liu* et al. *(2009) and Higdon* et al. *(2009).*

| Analysis | Low data set | Medium data set | High data set |
|---|---|---|---|
| Liu *et al.* (2009), temperature independent $k$ | – | 0.02 | 0.04 |
| Higdon *et al.* (2009), temperature independent $k$ | 0.07 | 0.03 | 0.03 |

the standard risk assessment procedures, which do not assess model accuracy; one sees a striking increase in uncertainty. These bounds on the probabilities should be interpreted as the range of failure probabilities that are expected, given the present state of knowledge concerning model uncertainty. Stated differently, given the observed input variability, the observed model uncertainty over the validation domain, and the extrapolation of this uncertainty to the regulatory condition, there is no evidence that the failure probabilities will be outside this range. Noting that all six of the lower failure probabilities are greater than 0.01, one can state that there is *no evidence* that the safety requirement can be met given the present state of knowledge. If any of the lower bounds on the failure probabilities had been *less than* 0.01, then one could conclude that it is possible, i.e., there is some evidence, that the criterion *could* be satisfied. However, this would certainly be insufficient evidence to certify the safety component. If the upper value of the predicted interval is less than 0.01, then one can state that, given the estimated aleatory and epistemic uncertainty, the safety component satisfies the criterion.

The present approach to validation is unlike the more common approaches based on hypothesis testing or Bayesian methods. With its focus on updating, the Bayesian approach convolves calibration and validation, which we believe ought to be rigorously separated. Bayesian methods attempt to use whatever data is available to improve the model as much as possible. Our approach reserves the first use of any validation data for assessing the performance of the model so as to reveal to decision makers and other would-be users of the model an unambiguous characterization of its predictive abilities.

One of the many valuable lessons learned from the Model Validation Challenge Workshop was that many different risk assessment approaches were applied to the three challenge problems. For the thermal challenge problem, two Bayesian analyses were applied by Liu *et al.* (2009) and Higdon *et al.* (2009). Their results are shown in Table 13.17. Comparing their values of failure probability with the present approach shown in Table 13.16, there is a striking difference in the predicted safety. Since the organizers of the Workshop have not released the underlying model that produced the results, it is unknown which analyses more properly represented the "true" failure probability of the safety component.

### 13.8 Bayesian approach as opposed to PBA

The PBA perspective is markedly different from a Bayesian perspective (Bernardo and Smith, 1994; Gelman *et al.*, 1995; Leonard and Hsu, 1999; Ghosh *et al.*, 2006; Sivia and Skilling, 2006). Three central Bayesian principles are (a) all uncertainties are fundamentally dependent on individual subjective belief and they should be interpreted as subjective belief, (b) all uncertainties are represented by subjective probability distributions, and (c) subjective probability distributions are viewed as flexible functions that should be adjusted as new information becomes available. Bayesians reason that the uncertainty results obtained with the PBA approach will commonly yield such large uncertainties that the analysis results are of little value to decision makers. Stated differently, the Bayesians reason that the resources spent on a PBA are wasted because the analysis provides such large ranges of uncertainty with no likelihood information that the decision maker cannot make an informed decision. They further emphasize that the time, resources, and effort spent on a PBA would be better spent on providing a *best estimated prediction* for the system of interest. Since the strategy of PBA is to represent lack of knowledge uncertainty as an interval, as opposed to representing the uncertainty as a probability distribution, it is evident that the uncertainty of the outcomes will be significantly greater with PBA than what a Bayesian approach would yield. We believe the Bayesian criticism of PBA summarized above is without merit and we give the following counter-arguments.

First, PBA lucidly shows the decision maker how poorly the SRQs are predicted because of lack of knowledge. Stated differently, PBA does not bury or disguise any assumptions; if something is not known, it is simply included in the analysis as bounds on possible outcomes. Large uncertainty bounds in PBA usually cause a great deal of angst and frustration in the decision maker. It may even provoke them to the point of saying, "Why did we spend all of this money on modeling and simulation when you can't predict something any better than that?" One might give two responses: "That's the way it is, given our present state of knowledge" Or: "Would you prefer that we make a number of weakly defensible assumptions, mix frequency of occurrence with personal beliefs, and then show much smaller uncertainties?" Sometimes these responses don't seem to help the situation.

Second, because a sensitivity analysis (SA) *must* also be provided with the PBA, the results of the SA tell the decision maker what are the major contributors to the uncertainty in the SRQs of interest. For example, if the p-box of the SRQ is very broad due to epistemic uncertainty, then the SA can identify which contributors to epistemic uncertainty are the culprits. With this list of uncertainty contributors, the decision maker can determine which contributors they have control over, and which ones they do not. For example, they may have significant control over the design of the system, but very little control over the environment or surroundings of the system. For those contributors that are under some control, a well founded cost–benefit analysis can be conducted to determine the time and money needed to reduce certain contributors as compared to other large contributors. A follow-up UQ analysis can then be conducted with estimated reductions in the controllable epistemic uncertainties so that the decision maker can see the impact on reducing the uncertainty in the SRQs of interest.

## 13.9 References


Almond, R. G. (1995). *Graphical Belief Modeling*. 1st edn., London, Chapman & Hall.

Ang, A. H.-S. and W. H. Tang (2007). *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering*. 2nd edn., New York, Wiley.

Angus, J. E. (1994). The probability integral transform and related results. *SIAM Review*. **36**(4), 652–654.

Aster, R., B. Borchers, and C. Thurber (2005). *Parameter Estimation and Inverse Problems*, Burlington, MA, Elsevier Academic Press.

Aughenbaugh, J. M. and C. J. J. Paredis (2006). The value of using imprecise probabilities in engineering design. *Journal of Mechanical Design*. **128**, 969–979.

Ayyub, B. M. (1994). The nature of uncertainty in structural engineering. In *Uncertainty Modelling and Analysis: Theory and Applications*. B. M. Ayyub and M. M. Gupta (eds.). New York, Elsevier: 195–210.

Ayyub, B. M. (2001). *Elicitation of Expert Opinions for Uncertainty and Risks*, Boca Raton, Florida, CRC Press.

Ayyub, B. M. and G. J. Klir (2006). *Uncertainty Modeling and Analysis in Engineering and the Sciences*, Boca Raton, FL, Chapman & Hall.

Bae, H.-R., R. V. Grandhi, and R. A. Canfield (2006). Sensitivity analysis of structural response uncertainty propagation using evidence theory. *Structural and Multidisciplinary Optimization*. **31**(4), 270–279.

Bardossy, G. and J. Fodor (2004). *Evaluation of Uncertainties and Risks in Geology: New Mathematical Approaches for their Handling*, Berlin, Springer.

Baudrit, C. and D. Dubois (2006). Practical representations of incomplete probabilistic knowledge. *Computational Statistics and Data Analysis*. **51**, 86–108.

Beck, M. B. (1987). Water quality modeling: a review of the analysis of uncertainty. *Water Resources Research*. **23**(8), 1393–1442.

Bedford, T. and R. Cooke (2001). *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge, UK, Cambridge University Press.

Beliakov, G., A. Pradera, and T. Calvo (2007). *Aggregation Functions: a Guide for Practitioners*, Berlin, Springer-Verlag.

Berendsen, H. J. C. (2007). *Simulating the Physical World; Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*, Cambridge, UK, Cambridge University Press.

Bernardini, A. and F. Tonon (2010). *Bounding Uncertainty in Civil Engineering*, Berlin, Springer-Verlag.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*, New York, John Wiley.

Bogen, K. T. and R. C. Spear (1987). Integrating uncertainty and interindividual variability in environmental risk assessment. *Risk Analysis*. **7**(4), 427–436.

Bouchon-Meunier, B., ed (1998). *Aggregation and Fusion of Imperfect Information*. Studies in Fuzziness and Soft Computing. New York, Springer-Verlag.

Box, G. E. P., J. S. Hunter, and W. G. Hunter (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd edn., New York, John Wiley.

Bucalem, M. L. and K. J. Bathe (2008). *The Mechanics of Solids and Structures: Hierarchical Modeling*, Berlin, Springer-Verlag.

Butcher, J. C. (2008). *Numerical Methods for Ordinary Differential Equations*. 2nd edn., Hoboken, NJ, Wiley.

Casella, G. and R. L. Berger (2002). *Statistical Inference*. 2nd edn., Pacific Grove, CA, Duxbury.

Cellier, F. E. and E. Kofman (2006). *Continuous System Simulation*, Berlin, Springer-Verlag.
Choi, S.-K., R. V. Grandhi, and R. A. Canfield (2007). *Reliability-based Structural Design*, London, Springer-Verlag.
Coleman, H. W. and F. Stern (1997). Uncertainties and CFD code validation. *Journal of Fluids Engineering*. **119**, 795–803.
Couso, I., S. Moral, and P. Walley (2000). A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy*. **5**, 165–181.
Cox, D. R. (2006). *Principles of Statistical Inference*, Cambridge, UK, Cambridge University Press.
Cozman, F. G. and P. Walley (2001). Graphoid properties of epistemic irrelevance and independence. *Proceedings of the Second International Symposium on Imprecise Probability and Their Applications*, Ithaca, NY, Shaker Publishing.
Crassidis, J. L. and J. L. Junkins (2004). *Optimal Estimation of Dynamics Systems*, Boca Raton, FL, Chapman & Hall/CRC Press.
Cullen, A. C. and H. C. Frey (1999). *Probabilistic Techniques in Exposure Assessment: a Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, New York, Plenum Press.
Devore, J. L. (2007). *Probability and Statistics for Engineers and the Sciences*. 7th edn., Pacific Grove, CA, Duxbury.
Dimov, I. T. (2008). *Monte Carlo Methods for Applied Scientists*. 2nd edn., World Scientific Publishing.
Dowding, K. J., M. Pilch, and R. G. Hills (2008). Formulation of the thermal problem. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2385–2389.
Draper, N. R. and H. Smith (1998). *Applied Regression Analysis*. 3rd edn., New York, John Wiley.
Duggirala, R. K., C. J. Roy, S. M. Saeidi, J. M. Khodadadi, D. R. Cahela, and B. J. Tatarchuk (2008). Pressure drop predictions in microfibrous materials using computational fluid dynamics. *Journal of Fluids Engineering*. **130**(7), 071302–1,071302–13.
Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*, London, Chapman & Hall.
EPA (2009). *Guidance on the Development, Evaluation, and Application of Environmental Models*. EPA/100/K-09/003, Washington, DC, Environmental Protection Agency.
Ferson, S. (1996). What Monte Carlo methods cannot do. *Human and Ecological Risk Assessment*. **2**(4), 990–1007.
Ferson, S. (2002). *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Setauket, NY, Applied Biomathematics.
Ferson, S. and L. R. Ginzburg (1996). Different methods are needed to propagate ignorance and variability. *Reliability Engineering and System Safety*. **54**, 133–144.
Ferson, S. and J. G. Hajagos (2004). Arithmetic with uncertain numbers: rigorous and (often) best possible answers. *Reliability Engineering and System Safety*. **85**(1–3), 135–152.
Ferson, S. and W. T. Tucker (2006). *Sensitivity in Risk Analyses with Uncertain Numbers*. SAND2006–2801, Albuquerque, NM, Sandia National Laboratories.
Ferson, S., V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz (2003). *Constructing Probability Boxes and Dempster-Shafer Structures*. SAND2003–4015, Albuquerque, NM, Sandia National Laboratories.
Ferson, S., R. B. Nelsen, J. Hajagos, D. J. Berleant, J. Zhang, W. T. Tucker, L. R. Ginzburg, and W. L. Oberkampf (2004). *Dependence in Probabilistic Modeling,*

*Dempster-Shafer Theory, and Probability Bounds Analysis*. SAND2004–3072, Albuquerque, NM, Sandia National Laboratories.

Ferson, S., J. Hajagos, D. S. Myers, and W. T. Tucker (2005). *CONSTRUCTOR: Synthesizing Information about Uncertain Variables*. SAND2005–3769, Albuquerque, NM, Sandia National Laboratories.

Ferson, S., W. L. Oberkampf, and L. Ginzburg (2008). Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*. **197**, 2408–2430.

Fetz, T., M. Oberguggenberger, and S. Pittschmann (2000). Applications of possibility and evidence theory in civil engineering. *International Journal of Uncertainty*. **8**(3), 295–309.

Frank, M. V. (1999). Treatment of uncertainties in space nuclear risk assessment with examples from Cassini Mission applications. *Reliability Engineering and System Safety*. **66**, 203–221.

Frey, H. C. and D. S. Rhodes (1996). Characterizing, simulating, and analyzing variability and uncertainty: an illustration of methods using an air toxics emissions example. *Human and Ecological Risk Assessment*. **2**(4), 762–797.

Gelman, A. B., J. S. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*, London, Chapman & Hall.

Ghanem, R. G. and P. D. Spanos (2003). *Stochastic Finite Elements: a Spectral Approach*. Revised edn., Mineola, NY, Dover Publications.

Ghosh, J. K., M. Delampady, and T. Samanta (2006). *An Introduction to Bayesian Analysis: Theory and Methods*, Berlin, Springer-Verlag.

Haimes, Y. Y. (2009). *Risk Modeling, Assessment, and Management*. 3rd edn., New York, John Wiley.

Haldar, A. and S. Mahadevan (2000a). *Probability, Reliability, and Statistical Methods in Engineering Design*, New York, John Wiley.

Haldar, A. and S. Mahadevan (2000b). *Reliability Assessment Using Stochastic Finite Element Analysis*, New York, John Wiley.

Halpern, J. Y. (2003). *Reasoning About Uncertainty*, Cambridge, MA, The MIT Press.

Helton, J. C. (1994). Treatment of uncertainty in performance assessments for complex systems. *Risk Analysis*. **14**(4), 483–511.

Helton, J. C. (1997). Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *Journal of Statistical Computation and Simulation*. **57**, 3–76.

Helton, J. C. (2003). Mathematical and numerical approaches in performance assessment for radioactive waste disposal: dealing with uncertainty. In *Modelling Radioactivity in the Environment*. E. M. Scott (ed.). New York, NY, Elsevier Science Ltd.: 353–389.

Helton, J. C. and F. J. Davis (2003). Latin Hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety*. **81**(1), 23–69.

Helton, J. C. and C. J. Sallaberry (2007). *Illustration of Sampling-Based Approaches to the Calculation of Expected Dose in Performance Assessments for the Proposed High Level Radioactive Waste Repository at Yucca Mountain, Nevada*. SAND2007–1353, Albuquerque, NM, Sandia National Laboratories.

Helton, J. C., D. R. Anderson, G. Basabilvazo, H.-N. Jow, and M. G. Marietta (2000). Conceptual structure of the 1996 performance assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety*. **69**(1–3), 151–165.

Helton, J. C., J. D. Johnson, and W. L. Oberkampf (2004). An exploration of alternative approaches to the representation of uncertainty in model predictions. *Reliability Engineering and System Safety*. **85**(1–3), 39–71.

Helton, J. C., F. J. Davis, and J. D. Johnson (2005b). A comparison of uncertainty and sensitivity analysis results obtained with random and Latin Hypercube sampling. *Reliability Engineering and System Safety*. **89**(3), 305–330.

Helton, J. C., W. L. Oberkampf, and J. D. Johnson (2005a). Competing failure risk analysis using evidence theory. *Risk Analysis*. **25**(4), 973–995.

Helton, J. C., J. D. Johnson, C. J. Sallaberry, and C. B. Storlie (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*. **91**(10–11), 1175–1209.

Higdon, D., C. Nakhleh, J. Battiker, and B. Williams (2009). A Bayesian calibration approach to the thermal problem. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2431–2441.

Hills, R. G., M. Pilch, K. J. Dowding, J. Red-Horse, T. L. Paez, I. Babuska, and R. Tempone (2008). Validation Challenge Workshop. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2375–2380.

Hoffman, F. O. and J. S. Hammonds (1994). Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Analysis*. **14**(5), 707–712.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering and System Safety*. **54**, 217–223.

Kleb, B. and C. O. Johnston (2008). Uncertainty analysis of air radiation for lunar return shock layers. *AIAA Atmospheric Flight Mechanics Conference*, AIAA 2008–6388, Honolulu, HI, American Institute of Aeronautics and Astronautics.

Klir, G. J. and M. J. Wierman (1998). *Uncertainty-Based Information: Elements of Generalized Information Theory*, Heidelberg, Physica-Verlag.

Kohlas, J. and P.-A. Monney (1995). *A Mathematical Theory of Hints – an Approach to the Dempster-Shafer Theory of Evidence*, Berlin, Springer-Verlag.

Krause, P. and D. Clark (1993). *Representing Uncertain Knowledge: an Artificial Intelligence Approach*, Dordrecht, The Netherlands, Kluwer Academic Publishers.

Kreinovich, V., J. Beck, C. Ferregut, A. Sanchez, G. R. Keller, M. Averill, and S. A. Starks (2007). Monte-Carlo-type techniques for processing interval uncertainty, and their potential engineering applications. *Reliable Computing*. **13**, 25–69.

Kriegler, E. and H. Held (2005). Utilizing belief functions for the estimation of future climate change. *International Journal for Approximate Reasoning*. **39**, 185–209.

Kumamoto, H. (2007). *Satisfying Safety Goals by Probabilistic Risk Assessment*, Berlin, Springer-Verlag.

Kyburg, H. E. and C. M. Teng (2001). *Uncertain Inference*, Cambridge, UK, Cambridge University Press.

Leonard, T. and J. S. J. Hsu (1999). *Bayesian Methods: an Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge, UK, Cambridge University Press.

Liu, F., M. J. Bayarri, J. O. Berger, R. Paulo, and J. Sacks (2009). A Bayesian analysis of the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2457–2466.

Melchers, R. E. (1999). *Structural Reliability Analysis and Prediction*. 2nd edn., New York, John Wiley.

Meyer, M. A. and J. M. Booker (2001). *Eliciting and Analyzing Expert Judgment: a Practical Guide*, New York, Academic Press.
Modarres, M., M. Kaminskiy, and V. Krivtsov (1999). *Reliability Engineering and Risk Analysis; a Practical Guide*, Boca Raton, FL, CRC Press.
Montgomery, D. C. (2000). *Design and Analysis of Experiments*. 5th edn., Hoboken, NJ, John Wiley.
Morgan, M. G. and M. Henrion (1990). *Uncertainty: a Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. 1st edn., Cambridge, UK, Cambridge University Press.
Mukhopadhyay, N. (2000). *Probability and Statistical Inference*, Boca Raton, FL, CRC Press.
NASA (2002). *Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners*. Washington, DC, NASA.
Nikolaidis, E., D. M. Ghiocel, and S. Singhal, eds (2005). *Engineering Design Reliability Handbook*. Boca Raton, FL, CRC Press.
NRC (2009). *Guidance on the Treatment of Uncertainties Assoicated with PRAs in Risk-Informed Decision Making*. Washington, DC, Nuclear Regulator Commission.
Oberkampf, W. L. and S. Ferson (2007). Model validation under both aleatory and epistemic uncertainty. *NATO/RTO Symposium on Computational Uncertainty in Military Vehicle Design*, AVT-147/RSY-022, Athens, Greece, NATO.
Oberkampf, W. L. and J. C. Helton (2005). Evidence theory for engineering applications. In *Engineering Design Reliability Handbook*. E. Nikolaidis, D. M. Ghiocel, and S. Singhal (eds.). New York, NY, CRC Press: 29.
Parry, G. W. (1996). The characterization of uncertainty in probabilistic risk assessments of complex systems. *Reliability Engineering and System Safety*. **54**, 119–126.
Paté-Cornell, M. E. (1996). Uncertainties in risk analysis: six levels of treatment. *Reliability Engineering and System Safety*. **54**, 95–111.
Pilch, M. (2008). Preface: Sandia National Laboratories Validation Challenge Workshop. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2373–2374.
Rabinovich, S. G. (2005). *Measurement Errors and Uncertainties: Theory and Practice*. 3rd edn., New York, Springer-Verlag.
Rai, S. N., D. Krewski, and S. Bartlett (1996). A general framework for the analysis of uncertainty and variability in risk assessment. *Human and Ecological Risk Assessment*. **2**(4), 972–989.
Raol, J. R., G. Girija, and J. Singh (2004). *Modelling and Parameter Estimation of Dynamic Systems*, London, UK, Institution of Engineering and Technology.
Roache, P. J. (1998). *Verification and Validation in Computational Science and Engineering*, Albuquerque, NM, Hermosa Publishers.
Ross, S. M. (2006). *Simulation*. 4th edn., Burlington, MA, Academic Press.
Rowe, W. D. (1994). Understanding uncertainty. *Risk Analysis*. **14**(5), 743–750.
Rubinstein, R. Y. and D. P. Kroese (2008). *Simulation and the Monte Carlo Method*. 2nd edn., Hoboken, NJ, John Wiley.
Rutherford, B. M. (2008). Computational modeling issues and methods for the "Regulatory Problem" in engineering – solution to the thermal problem. *Computer Methods in Applied Mechanics and Engineering*. **197**(29–32), 2480–2489.
Sallaberry, C. J., J. C. Helton, and S. C. Hora (2008). Extension of Latin Hypercube samples with correlated variables. *Reliability Engineering and System Safety*. **93**, 1047–1059.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008). *Global Sensitivity Analysis: the Primer*, Hoboken, NJ, Wiley.
Sentz, K. and S. Ferson (2002). *Combination of Evidence in Dempster-Shafer Theory*. SAND2002–0835, Albuquerque, NM, Sandia National Laboratories.
Singh, V. P., S. K. Jain, and A. Tyagi (2007). *Risk and Reliability Analysis: a Handbook for Civil and Environmental Engineers*, New York, American Society of Civil Engineers.
Singpurwalla, N. D. (2006). *Reliability and Risk: a Bayesian Perspective*, New York, NY, Wiley.
Sivia, D. and J. Skilling (2006). *Data Analysis: a Bayesian Tutorial*. 2nd edn., Oxford, Oxford University Press.
Steinhauser, M. O. (2008). *Computational Multiscale Modeling of Fluids and Solids: Theory and Applications*, Berlin, Springer-Verlag.
Stern, F., R. V. Wilson, H. W. Coleman, and E. G. Paterson (2001). Comprehensive approach to verification and validation of CFD simulations – Part 1: Methodology and procedures. *Journal of Fluids Engineering*. **123**(4), 793–802.
Storlie, C. B. and J. C. Helton (2008a). Multiple predictor smoothing methods for sensitivity analysis: description of techniques. *Reliability Engineering and System Safety*. **93**(1), 28–54.
Storlie, C. B. and J. C. Helton (2008b). Multiple predictor smoothing methods for sensitivity analysis: example results. *Reliability Engineering and System Safety*. **93**(1), 55–77.
Suter, G. W. (2007). *Ecological Risk Assessment*. 2nd edn., Boca Raton, FL, CRC Press.
Torra, V. and Y. Narukawa (2007). *Modeling Decision: Information Fusion and Aggregation Operators*, Berlin, Springer-Verlag.
Trucano, T. G., M. Pilch, and W. L. Oberkampf (2002). *General Concepts for Experimental Validation of ASCI Code Applications*. SAND2002–0341, Albuquerque, NM, Sandia National Laboratories.
van den Bos, A (2007). *Parameter Estimation for Scientists and Engineers*, Hoboken, NJ, Wiley-Interscience.
Vinnem, J. E. (2007). *Offshore Risk Assessment: Principles, Modelling and Applications of QRA Studies*, Berlin, Springer-Verlag.
Vose, D. (2008). *Risk Analysis: a Quantitative Guide*. 3rd edn., New York, Wiley.
Wasserman, L. A. (2004). *All of Statistics: a Concise Course in Statistical Inference*, Berlin, Springer-Verlag.
Wilson, R. V., F. Stern, H. W. Coleman, and E. G. Paterson (2001). Comprehensive approach to verification and validation of CFD simulations – Part 2: Application for RANS simulation of a cargo/container ship. *Journal of Fluids Engineering*. **123**(4), 803–810.
Yager, R. R., J. Kacprzyk, and M. Fedrizzi, eds (1994). *Advances in the Dempster-Shafer Theory of Evidence*. New York, John Wiley.
Young, G. A. and R. L. Smith (2005). *Essentials of Statistical Inference*, Cambridge, UK, Cambridge University Press.
Zienkiewicz, O. C. and J. Z. Zhu (1992). The superconvergent patch recovery and a posteriori error estimates. *International Journal for Numerical Methods in Engineering*. **33**, 1365–1382.