RESOURCE ALLOCATION
FOR UNCERTAINTY QUANTIFICATION AND REDUCTION

By

Joshua Mullins

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in

Civil Engineering

December, 2014
Nashville, Tennessee

Approved:

Sankaran Mahadevan, Ph.D.
Prodyot K. Basu, Ph.D.
Caglar Oskay, Ph.D.
Ravindra Duddu, Ph.D.
Alejandro Strachan, Ph.D.
Angel Urbina, Ph.D.

To my mom and dad for their unwavering support

ACKNOWLEDGMENTS

I would like to offer my sincerest gratitude to my research advisor, Professor Sankaran Mahadevan. His wealth of knowledge and experience has provided tremendous insight that has been invaluable to this work. He taught me how to think like a researcher and pushed me to achieve my full potential. His intense dedication to his work and his students is always apparent, and despite his extraordinarily busy schedule, he has always found the time to offer guidance and encouragement when it matters most.

I would like to thank the remaining members of my committee, Professor Prodyot Basu, Professor Caglar Oskay, Professor Ravindra Duddu, Professor Alejandro Strachan, and Dr. Angel Urbina, for their constructive feedback and insight that have helped to make this work a success. In particular, I consider my mentor at Sandia National Laboratories, Dr. Angel Urbina, a close friend and colleague. I am grateful for the resources and insights he has provided to me through Sandia during my internships, and I am pleased to have the opportunity to work with him in the future. I also specifically thank Professor Alejandro Strachan for facilitating many important discussions that have led to the example demonstrations in this work.

At Vanderbilt, I have had the pleasure of working with many faculty members, students, and postdoctoral fellows who have mentored me and provided their expertise. In particular, I thank Dr. Shankar Sankararaman for passing on a wealth of knowledge about validation and uncertainty quantification that has been fundamental to this work. I thank Dr. You Ling for his friendship and close collaboration on many aspects of this work and Chenzhao Li for participating in many discussions and providing feedback that improved the proposed methods. I also thank Dr. Vadiraj Hombal for sharing his programming expertise and passing along software tools that were used in this work. Many other close friends have supported and encouraged me throughout my time at Vanderbilt, and it is greatly appreciated.

I would also like to acknowledge my external research collaborators. In working closely with the Purdue University center for Prediction of Reliability, Integrity, and Survivability of Microsystems (PRISM), many of the faculty and staff members contributed models, data, and expertise that have been critical to the demonstration of this work. In particular, I thank Dr. Lin Sun, Dr. Hojin Kim, Professor Jayathi Murthy and Professor Marisol Koslowski for providing the computational models that were used in this work, Dr. Martin Hunt for providing software assistance, and Professor Muhammad Alam and Professor Dimitrios Peroulis for providing experimental data. I also thank the members of the technical staff at Sandia National Laboratories, Dr. Ken Hu, Dr. Adam Hetzler, Dr. Vicente Romero, and Dr. Laura Swiler for providing helpful feedback in many discussions that has improved this work tremendously.

This dissertation could not have been accomplished without the love and support of my mother and father. I can never repay them for their patience and provision during my academic career. I also thank the rest of my family for their continuous encouragement.

Finally, the funding sources for this work are gratefully acknowledged. The two primary sources were Sandia National Laboratories (Contract No. BG-7732) and the U.S. Department of Energy [National Nuclear Security Administration], award number DE-FC52-08NA28617 to Purdue University (Principal Investigator: Prof. Jayathi Murthy, Director: Prof. Alejandro Strachan), and subaward to Vanderbilt University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

The behavior of a complex engineering system is frequently described by a computational model that is designed to replicate reality as closely as possible. From these models, analysts make decisions about the design and operation of the system, most commonly with respect to a set of decision variables (e.g. material and system configuration/properties in the design phase or the inspection/maintenance interval in the operational phase). However, these decisions are complex because engineering systems are designed and operated under a wide range of uncertainty sources. In the presence of uncertainty, systems are never perfectly reliable, so there is a nonzero probability of system failure for any decision. Each failure event has a corresponding risk that depends on the consequence of the failure. The risk has classically been addressed by applying safety factors based on empirical knowledge, but this practice often makes the mitigation strategy economically inefficient and is difficult to apply to new systems with no experience. The proposed research seeks instead to address risk systematically by properly accounting for all known sources of uncertainty and then reducing them when it is possible and economically feasible to do so.

In the reliability analysis literature, the sources of uncertainty have commonly been grouped into two basic categories: aleatory uncertainty (i.e. natural variability) and epistemic uncertainty (i.e. lack of knowledge) [36, 50, 64, 72, 74]. For aleatory sources, probabilistic methods for characterizing and propagating uncertainty are well-developed [32]. Since this uncertainty source is considered irreducible, engineering system designs must account for it properly, but its contributions cannot be eliminated by any design decision. On the other hand, risk mitigation may be achieved through epistemic uncertainty reduction, by collecting information and improving the understanding of the system. The two main classes of epistemic uncertainty that are considered in this dissertation are data uncertainty and model uncertainty. Data uncertainty arises because economic factors prevent analysts from collecting as much empirical data as is needed (i.e. sparse data) and because human error and instrumentation limitations lead to inaccurate and/or imprecise measurements. Model uncertainty exists because models can always be improved (according to decisions made by the developers), but they are never perfect. The underlying governing equation does not describe the physics of reality completely, and in many cases, the governing equation cannot be solved exactly.

Furthermore, even for a given computational model, many of its inputs are uncertain, and they cannot be measured directly in an experiment. These inputs parameterize the model, and they must be inferred by an inverse problem [3] in which outputs of interest are observed in an experiment. In this dissertation, Bayesian methods [89] are used to handle this inference problem, but the presence of data uncertainty leads to an epistemic probability distribution for these uncertain parameters even when they are deterministic quantities in reality. This parameter uncertainty is critical to system risk assessment and management, especially when the parameters (e.g. material properties) that are calibrated in a simplified domain are common to the usage condition of the model where a prediction is made. Understanding these parameters well can greatly improve the quantification of uncertainty in the system prediction, and the only way to improve understanding is to collect better experimental output data (i.e. larger quantity of data points and/or greater measurement precision). Improving measurement precision may not always be possible, but collecting a larger quantity of data is a feasible option though it will be subject to some economic constraints.

Within this context, the topic of interest in this dissertation is how to perform these activities efficiently by effectively allocating resources to the various uncertainty quantification (UQ) tasks. A comprehensive framework for UQ that includes model calibration, model validation, and uncertainty propagation is proposed. Then, activities such as model selection and test selection are explored in order to improve the accuracy of the computation and minimize the uncertainty in a prediction of interest.

1.2 Research objectives

This work explores resource allocation with the fundamental objective of quantifying and reducing prediction uncertainty in order to enable credible reliability analysis and risk assessment. For each of the two primary epistemic sources that were previously mentioned (data uncertainty and model uncertainty), there is a tradeoff decision of cost vs. value. Data uncertainty reduction requires the tangible expense of performing additional experiments or using more expensive methods and instruments to increase measurement precision. Model uncertainty reduction typically requires additional time and effort for the development of more sophisticated models and/or evaluation of more expensive simulations with higher fidelity and resolution.

This dissertation systematically addresses these tradeoff decisions through several key objectives. To address the evaluation time of expensive simulations, the objective of model selection for uncertainty propagation is considered because efficient uncertainty propagation is needed for both model validation and prediction. Then, to address data uncertainty, test selection for prediction uncertainty reduction is considered. However, to perform this objective for both calibration and validation tests, a more formal understanding of how the model validation results affect the prediction is needed; therefore, the connection of model validation to prediction is considered first. Finally, the overall UQ framework must be connected to risk assessment in order to consider the economic efficiency of the entire approach. These objectives are summarized as follows:

(1) Model selection for uncertainty propagation

(2) Connecting model validation to prediction

(3) Test selection for prediction uncertainty reduction

(4) Risk-based resource allocation

To address the first objective, a methodology to select among available modeling options in order to maximize prediction accuracy within a limited computational budget is proposed. The proposed approach takes advantage of sparse and imprecise information about the prediction quantity to improve the decision-making. The second objective explores the effect of epistemic uncertainty on model validation and examines how different types of validation data impact the prediction of interest. The proposed approach accomplishes this objective by separating the contributions of aleatory and epistemic uncertainty sources and then quantifying the relevance to prediction of different validation tests. The third objective takes advantage of these results to address the test selection problem from the perspective of prediction uncertainty reduction. The proposed method expands test selection methods for model calibration to also include validation experiments in a joint framework. The fourth and final objective explores how the combination of data uncertainty and model uncertainty affects risk assessment. This objective provides insights about how the cost/benefit analysis of the entire resource allocation framework proposed in this dissertation can be used for decision-making.

1.3 Organization of the dissertation

The subsequent chapters of this dissertation are organized to address the research objectives described in Section 1.2. Chapter 2 provides some useful background information about existing UQ frameworks and provides fundamental details of model calibration, model validation, and uncertainty propagation. Chapter 3 proposes a model selection approach for efficient uncertainty propagation in the context of scalar-input systems as well as spatially and temporally varying problems. Chapter 4 explores the separation of uncertainty sources in model validation and proposes an approach to explicitly connect the model validation input conditions to the prediction of interest. Chapter 5 explores the effect of data uncertainty on prediction and proposes an optimization approach to select among available testing options and/or input conditions for calibration and validation. Chapter 6 approaches the resource allocation problem from the perspective of risk and proposes formulations for optimization problems that select an appropriate budget for the UQ problem. Chapter 7 concludes the dissertation and suggests opportunities for future work.

CHAPTER 2

BACKGROUND

This chapter describes some fundamental aspects of a comprehensive UQ framework. Existing frameworks in the UQ literature [39, 92, 95, 98, 107] make predictions on stochastic outputs of interest by performing several key activities: (1) characterization of input uncertainty, (2) model verification, (3) model calibration, (4) model validation, and (5) uncertainty propagation (i.e. prediction). Input uncertainty is typically quantified by repeated tests to explore natural variability, and it can then be characterized by well-established methods of constructing probability distributions. This step provides the input ranges over which existing models should be verified by benchmarking against analytical solutions, and errors pertaining to the numerical solution process can be quantified. Since some additional model parameters cannot be measured directly, they must be inferred from experimental data obtained for measureable output quantities in the calibration process. Since data is sparse and/or imprecise, correct deterministic parameter estimates cannot be obtained confidently, so these model parameters are instead described with uncertainty stemming from lack of knowledge about their values. Calibrated models are then compared with an independent set of experimental data in order to assess the predictive capability of the models. The result of this process, known as model validation, indicates whether the model should be taken forward and used for prediction. If the model is deemed valid, input and parameter uncertainty can be propagated through it to make a prediction for a quantity of interest in the form of a probability distribution.

There are two basic types of model inputs: (1) those which can be measured directly in an experiment, as either a deterministic value or a known aleatory distribution, henceforth denoted by $\boldsymbol{X}$ and referred to simply as inputs and (2) those which are not measureable and must be inferred from observed outputs, henceforth denoted by $\boldsymbol{\Theta}$ and referred to as parameters. For the remainder of this dissertation, note that upper case variables denote random variables while lower case variables represent particular samples from their distributions. Bolded variables are vectors, matrices, or jointly distributed sets of random variables, and variables in plain text are scalar quantities or single random variables.

This chapter explains how these two classes of inputs are treated in model calibration (Section 2.1) and model validation (Section 2.2). Since uncertainty propagation (Section 2.3) is required when performing model validation with stochastic quantities, and model calibration requires solving an inverse problem, both of these activities require a large number of model evaluations. Surrogate models are often needed in order to improve efficiency; therefore, one surrogate modeling approach (Gaussian process modeling) is described in Section 2.4.

2.1 Bayesian model calibration

Bayesian calibration [7, 39, 47, 63, 98] is an approach for inferring unmeasured parameters $\boldsymbol{\Theta}$ by observing particular values of the outputs $\boldsymbol{y_d}$ and corresponding inputs $\boldsymbol{x}$. As opposed to deterministic parameter estimation, which results in only a single value for the parameters, Bayesian calibration results in a posterior probability distribution that represents the subjective probability of each value in the domain. Note that the assumption implicit to this approach is that the parameter values are deterministic in reality, but the values cannot be inferred precisely due to data uncertainty in the observations as well as model errors that may bias the results. Therefore, the posterior distribution represents epistemic uncertainty, not aleatory uncertainty.

The posterior is obtained by applying Bayes' theorem, which states that the posterior probability of the parameters $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{y_d})$ is proportional to the product of the likelihood function $L(\boldsymbol{\theta})$ (i.e. the probability of observing the data $\boldsymbol{y_d}$ given a particular parameter set $\boldsymbol{\theta}$) and the prior density $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$.

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{y_d}) = \frac{L(\boldsymbol{\theta})f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta})f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}} \qquad (2.1)$$

To construct the likelihood function, it is typically assumed that the difference between a particular observation y_{d_i} and the prediction y_m at input $\boldsymbol{x}$ is due to measurement noise in the observation e_{d_i}. This noise is typically assumed to be zero-mean Gaussian white noise, and the standard deviation of the error σ_d may either be computed from the observed data or calibrated along with $\boldsymbol{\Theta}$ when the observation data is sparse.

$$y_m(\boldsymbol{x},\boldsymbol{\theta}) = y_{d_i} + e_{d_i} \qquad (2.2)$$

$$E_d \sim N(0,\sigma_d) \qquad (2.3)$$

The likelihood function is constructed jointly across all observations. It is commonly assumed that the measurement errors associated with the set of observations are independent. In this scenario, the likelihood values for the set of observations can be combined by a product. If there are n_j observations at m different input conditions (each denoted $\boldsymbol{x_j}$**)**, the likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m}\prod_{i=1}^{n_j}\frac{1}{\sigma_d\sqrt{2\pi}}\exp\left\{-\frac{\left[y_{m_j}(\boldsymbol{x_j},\boldsymbol{\theta})-y_{d_{ij}}\right]^2}{2\sigma_d^2}\right\} \qquad (2.4)$$

The likelihood function in Eq. (2.4) includes all of the calibration data from all of the measured input conditions. Since the posterior distribution of $\boldsymbol{\Theta}$ obtained from Eq. (2.1) cannot be normalized and inverted easily, it is difficult to draw samples from the posterior distribution using traditional Monte Carlo Simulation (MCS) [32]. Therefore, samples are typically drawn from the posterior distribution using a function that is proportional to the posterior density. This problem has been solved by applying Markov chain Monte Carlo (MCMC) sampling methods [27, 35, 66, 71], which do not require inversion of the CDF of the posterior distribution.

Note that the relationship given in Eq. (2.2) does not account for model inadequacy. Since model inadequacy is often a leading source of the difference between prediction and observation, many researchers [13, 38, 54] add a stochastic, input-dependent model discrepancy term to the model prediction. The goal of this approach, commonly referred to as the Kennedy-O'Hagan framework [47], is to reduce the bias in the parameter estimates; bias is introduced when parameters are used to fit an inadequate model form to the observed data. However, since the mathematical form of the model inadequacy is always unknown, an additional set of parameters must be introduced to define a stochastic model inadequacy function, and these parameters must be inferred jointly with $\boldsymbol{\Theta}$. This expansion of the calibration problem leads to some additional difficulties, including selection of a proper discrepancy formulation [58] and unique identifiability of the expanded parameter set [7, 58, 88]. Therefore, in this dissertation, no model discrepancy term is included in the proposed methods, and the potential model inadequacy is accounted for through model validation within the prediction framework that will be described in Chapter 5.

2.2 Model validation methods

After the parameters are calibrated, the resulting distributions are propagated through the model, and the output is compared against the validation data in order to assess the predictive capability of the model. The validation data should be independent of the calibration data, and if possible should be data collected in a regime outside the calibration domain. Since this is not practically possible in all cases, data in one regime is sometimes partitioned for calibration and validation. In the presence of both aleatory and epistemic uncertainty, the validation assessment is performed in the probability space by comparing the model prediction (stochastic due to parameter uncertainty) and the observation data (stochastic due to measurement uncertainty). Several methods for performing a stochastic assessment can be found in the literature [57, 59]; available methods include classical hypothesis testing [25, 34, 41], Bayesian hypothesis testing [73, 86, 87, 108], the area metric [22, 23, 95], and the model reliability metric [85, 97]. In particular, the area metric and the model reliability metric are explored in detail in Chapter 4. Brief explanations of these two approaches are provided in Section 2.2.1 and Section 2.2.2 respectively.

2.2.1 Area validation metric

The area metric [22, 23] measures the difference between the cumulative distribution functions (CDF) of model output and experimental data, and is defined as

$$d\left(F_{Y_m}, S_{Y_d}\right) = \int_{-\infty}^{\infty} |F_{Y_m}(y) - S_{Y_d}(y)| dy \qquad (2.5)$$

Here, $F_{Y_m}(y)$ is the CDF of the model output, and $S_{Y_d}(y)$ is the empirical CDF of the experimental data. This metric is inherently designed for a stochastic prediction and observation, but it can also be applied when the model prediction Y_m is deterministic. In this scenario, the model prediction CDF is a step function such that $F_{Y_m}(y) = 0$ for $y < y_m$, and $F_{Y_m}(y) = 1$ for $y > y_m$. One useful feature of the area metric is that the physical unit of d is the same as the unit of Y. Therefore, the area metric value has a direct interpretation that is physically meaningful. The result is nonnegative, but unbounded, since the difference between two cumulative distribution functions can be arbitrarily large.

Since validation tests may be conducted for many different input conditions (i.e. input vectors $\boldsymbol{x_i}$), an important property of a validation metric is how it combines information from different points in the domain. The area metric incorporates different input conditions by applying a "u-pooling" procedure (i.e. a transformation from physical space to probability space). This approach is particularly useful for validating models with sparse data on multiple experimental combinations [59]. For a particular input condition $\boldsymbol{x_i}$, let $F_{Y_{m_i}}$ be the CDF of the model output Y_m, and let y_{d_i} be the corresponding observation. Then, a u-value, $u_i = F_{Y_{m_i}}(y_{d_i})$, can be computed for each input condition. Based on the probability integral transform theorem [6], the u-values would follow the standard uniform distribution, $U[0, 1]$, if the observations y_{d_i} were random samples from the probability distribution of Y_{m_i}. Therefore, if the distributions of the model output and the observation are equal to each other at each input condition, the empirical CDF of the collection of u-values should match the CDF of the standard uniform random variable. Thus, the difference between the two empirical CDF curves can be thought of as the disparity between model outputs and experimental observations across the entire domain of the inputs. Further, the area metric in the transformed space [23] follows similarly from Eq. (2.5) as

$$d(F_u, S_u) = \int_0^1 |F_u - S_u| du \tag{2.6}$$

where F_u is the empirical CDF obtained from the u-values and S_u is the standard uniform CDF. As in Eq. (2.5), small values of d represent good agreement between prediction and observation, and large values represent disagreement. However, in the probability space the metric is no longer unbounded; in fact, it is bounded on the interval $[0, 0.5]$. Therefore, the metric value can no longer be interpreted in terms of the physical unit of the output quantity.

To address this issue, the area metric computed by Eq. (2.6) can be transformed back to physical space to retrieve its physical interpretation. Using the CDF of the model output G_y at some particular input condition, the u-values can be transformed back by inverting the CDF, $y_i = G_y^{-1}(u_i)$. The empirical CDF values y_i can be used to construct an empirical CDF that can then be compared to G_y as in Eq. (2.5). The result of this computation will again have the same physical unit as Y. Thus, transforming back to the physical space makes it easier to set a tolerance threshold for the acceptance of the model. However, it should be noted that the value of the area metric that is obtained after the transformation depends on which value of y is selected for performing the back-transformation.

2.2.2 Model reliability metric

The model reliability metric r [85] is a direct measure of model prediction quality, computed by assessing the distribution of particular values of the difference between a stochastic prediction and observation. It is defined as the probability of the difference (Δ) between observed data (Y_d) and model prediction (Y_m) being less than a given tolerance limit ϵ

$$r = \Pr(-\epsilon < \Delta < \epsilon), \quad \Delta = Y_d - Y_m \tag{2.7}$$

Note that the model reliability is computed separately for each input condition, and Y_d and Y_m are both functions of $\boldsymbol{x}$. This fact will be discussed in detail in Chapter 4, but it is mentioned here to point out that all the stochasticity in Y_m is attributed to uncertainty in $\boldsymbol{\Theta}$ at a particular input condition. Therefore, in Eq. (2.7), experimental observation is treated as a random variable due to measurement error, and the model output is a distribution resulting from the propagation of posterior parameter uncertainty from calibration. Since it is the difference between two random variables, Δ is also a random variable, and the probability distribution of Δ can be obtained from the probability distributions of Y_d and Y_m. Then, the model reliability metric is computed by integration of the distribution of Δ.

$$r = \int_{-\epsilon}^{\epsilon} f_\Delta(\omega) d\omega = F_\Delta(\epsilon) - F_\Delta(-\epsilon) \tag{2.8}$$

For instance, if the model prediction, $Y_m \sim N(\mu_{Y_m}, \sigma_{Y_m}^2)$, and the corresponding observation, $Y_d \sim N(\mu_{Y_d}, \sigma_{Y_d}^2)$, are independent, the distribution of the difference can be computed analytically, $\Delta \sim N(\mu_{Y_d} - \mu_{Y_m}, \sigma_{Y_d}^2 + \sigma_{Y_m}^2)$. For the sake of simplicity, let $\sigma_\Delta = \sqrt{\sigma_{Y_d}^2 + \sigma_{Y_m}^2}$. In this scenario, the model reliability metric r can be computed by evaluating the standard normal CDF Φ as

$$r = \Phi\left[\frac{\epsilon - \left(\mu_{Y_d} - \mu_{Y_m}\right)}{\sigma_\Delta}\right] - \Phi\left[\frac{-\epsilon - \left(\mu_{Y_d} - \mu_{Y_m}\right)}{\sigma_\Delta}\right] \tag{2.9}$$

Since the result of this computation is a probability, the model reliability is considered to be a probabilistic validation metric. Note that Bayesian hypothesis testing generally leads to a single scalar result known as the Bayes factor [46, 73, 76], but the Bayes factor may also be converted to a probability measure. Thus, the methods that are developed in this dissertation for probabilistic validation metrics are also applicable to Bayesian hypothesis testing although they are only illustrated for the model reliability metric.

As mentioned, separate computations of model reliability are performed at each input condition since the distributions of Y_d and Y_m are dependent on where validation experiments are conducted. The set of reliability values at different $\boldsymbol{x_i}$ provides information about the predictive capability of the model as a function of location in the input domain. The suitability of any model for prediction depends on the prediction scenario of interest. Models are often useful in some regions of the domain, but not in others. This fact is used to develop the model selection methodology in Chapter 3, and in Chapter 4, an approach for connecting the validation input conditions to the prediction of interest is proposed.

2.3 Uncertainty propagation techniques

When solving an inverse problem by applying MCMC methods as described in Section 2.1, a large number of function evaluations are needed to solve the parameter estimation problem, and then the resulting posterior distribution must be propagated back through the model for prediction. In addition, stochastic approaches to model validation, as described in Section 2.2, require the propagation of parameter uncertainty through the model at each validation input condition. This propagation is typically performed via MCS, which again requires a large number of model evaluations. When the computational model is expensive, it is often unaffordable to use the computational model for every function evaluation.

There are two basic classes of approaches available to manage computationally intractable UQ problems; either the number of samples required can be reduced in an intelligent way, or the model being evaluated can be simplified so that less time is needed for each sample. Methods of efficient stochastic simulation with respect to the number of samples have been explored in studies on reliability analysis and design optimization. One inexpensive way of propagating input variability and/or parameter uncertainty through a system model is a first-order Taylor series expansion, which requires only $n + 1$ function evaluations for n uncertain variables. This method is referred to as a first-order second moment (FOSM) approach in the reliability analysis literature [32]. Other reliability analysis approaches take advantage of the idea that sometimes only a particular point on the distribution of the output quantity of interest (QoI) is needed for the computation (e.g. probability that stress or deformation exceeds a particular value). This type of analysis typically uses Newton-like optimization methods to search in an equivalent uncorrelated standard normal space for the most probable point (MPP) on a limit state related to the QoI [82, 93]. The failure probability is then approximated via the first-order reliability method (FORM) or the second-order reliability method (SORM) [32].

Alternatively, within the context of MCS for reliability analysis, methods such as importance sampling modify the sampling distribution to ensure that more samples fall within a region of interest, thereby reducing the total number of samples needed for the analysis. For example, Harbitz’s importance sampling approach [33] creates a sampling distribution centered at the MPP; adaptive methods are also available to update the importance sampling distribution after ever few samples [17, 110]. Because each of the aforementioned approaches searches only in a region of interest, they can be restrictive if the goal of the analysis is to determine the entire distribution of the QoI. To calculate the entire distribution, these methods may be applied at several regions of interest and interpolated (note: interpolation introduces additional error and uncertainty), or the analyst must revert to a full MCS.

If a full MCS is to be performed, it may be infeasible to evaluate a high-fidelity physics model (e.g. nonlinear finite element analysis with a very fine mesh) for every Monte Carlo sample, so the class of approaches aimed at reducing computation time per sample is utilized instead. Cheaper models (in terms of CPU time per evaluation) which may be in the form of mathematical surrogate models (also referred to as response surfaces or meta-models), reduced order models, or reduced physics models have been pursued in this regard. Common surrogate models include simple regression models, Gaussian process (GP) or Kriging models [16, 96], polynomial chaos expansion models [111], support vector machines [78], and neural networks [62]. Since additional error is introduced to the system prediction by these surrogates, the uncertainty associated with surrogate modeling is considered in subsequent chapters of this dissertation. In particular, GP surrogate models are used for efficiency throughout the proposed UQ framework. Therefore, the GP modeling approach is described in detail in Section 2.4.

2.4 Gaussian process surrogate modeling

Because of the computational challenges described in Section 2.3, the computational model is commonly replaced by a surrogate model to improve the efficiency of both the calibration and uncertainty propagation activities. GP surrogate models [84] are used in this dissertation because they provide a natural way of quantifying the uncertainty due to the discrepancy between the surrogate and the original computational model. The contribution of this uncertainty can then be incorporated into the validation assessment and the prediction, as will be described in Chapter 4.

Suppose that a GP surrogate model will be used to replace a computational model $y = g(\boldsymbol{x})$. By evaluating the computational model at an arbitrary number of input points $\boldsymbol{x_i}$, a matrix of training points $\boldsymbol{x_T}$ and training values $\boldsymbol{y_T}$ can be generated. Then, the GP model will be used to predict at a new set of input points $\boldsymbol{x_P}$ within the same domain of interest. The GP model has two basic parts: a mean function, which typically isolates a simple polynomial trend relationship between input and output, and a Gaussian process which describes the random variability over the input space. It is assumed that the combination of these two components describes the true response function g [16] as shown in Eq. (2.10). The mean function is represented by $m(*)$ and the GP by $z(*)$.

$$g(\boldsymbol{x}) = m(\boldsymbol{x}) + z(\boldsymbol{x}) \tag{2.10}$$

The mean function can usually be a simple low-order polynomial, and even a constant value over the entire input space may be sufficient [96]. The GP is typically assumed to be stationary with zero mean, which implies that the correlation between prediction point and training point is only a function of the distance between them. The choice of the correlation function may be problem specific, and there are many available options depending on the desired properties of the correlation structure. One form that is frequently chosen is the squared-exponential function, which represents the correlation between two input points in the domain as in Eq. (2.11).

$$r(\boldsymbol{x_1}, \boldsymbol{x_2}) = exp[-\sum_{i=1}^{d} \frac{(X_1^i - X_2^i)^2}{l_i}] \tag{2.11}$$

The dimension of the input space is given by d, and each l_i defines the length scale in the corresponding dimension. Each length scale represents the rate of decay of correlation when moving in the corresponding spatial dimension. The covariance between input points is given by a product of the correlation function and process variance σ_Z^2 as in Eq. (2.12).

$$Cov(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sigma_Z^2 r(\boldsymbol{x_1}, \boldsymbol{x_2}) \tag{2.12}$$

The combination of Eq. (2.11) and (2.12) is used to compute the covariance between each pair of input points in $\boldsymbol{x_T}$ to obtain the covariance matrix $\boldsymbol{\Sigma_{TT}}$. It can also be applied to obtain the covariance matrix between training and prediction points.

The mean function coefficients may be estimated along with the parameters of the covariance function; however, when the mean function is taken as a constant, it is typically chosen to equal the mean of the training values across the available training points. In this situation, there are then $d + 1$ parameters of the GP remaining to estimate: one length scale for each dimension and the process variance. Either Bayesian inference or maximum likelihood estimation (MLE) may be used to compute the parameters. In this dissertation, only single deterministic estimates of the GP parameters are used, as obtained from MLE. To obtain the MLE values, a global optimization problem must be solved, and the shape and smoothness of the likelihood function often make gradient-based approaches ineffective. Therefore, the problem is typically solved using the DIRECT algorithm [24] or the simulated annealing algorithm [49]. Since inversion of the covariance matrix $\boldsymbol{\Sigma_{TT}}$ is required when computing the likelihood, numerical instabilities may arise, and the search algorithms may be costly when the size of $\boldsymbol{\Sigma_{TT}}$ is large (corresponding to a large number of training points). Some improvements to the efficiency and numerical stability of the estimation process can be found in the literature [31, 63, 81].

Once the parameters are determined, the resulting GP model is used to make predictions at a new set of input points within the space. An important property of the model is that it gives an estimate of uncertainty in addition to the mean prediction at particular input point. In particular, the set of prediction values $\boldsymbol{y_P}$ at prediction points $\boldsymbol{x_P}$ are jointly Gaussian distributed according to the following set of equations:

$$\boldsymbol{y_P}|\boldsymbol{x_P}, \boldsymbol{x_T}, \boldsymbol{y_T} \sim N(\boldsymbol{\mu_{y_P}}, \boldsymbol{\Sigma}_{\mathbf{y_P}})$$

$$\boldsymbol{\mu_{y_P}} = m(\boldsymbol{x_P}) + \boldsymbol{\Sigma_{PT}}\boldsymbol{\Sigma_{TT}^{-1}}[\boldsymbol{y_T} - m(\boldsymbol{x_T})]$$

$$\boldsymbol{\Sigma}_{\mathbf{y_P}} = \boldsymbol{\Sigma_{PP}} - \boldsymbol{\Sigma_{PT}}\boldsymbol{\Sigma_{TT}^{-1}}\boldsymbol{\Sigma_{PT}^{T}} \tag{2.13}$$

Here, $\boldsymbol{\mu_{y_P}}$ is the mean vector of prediction values, $\boldsymbol{\Sigma_{y_P}}$ is the covariance matrix of the prediction values, $\boldsymbol{\Sigma_{PP}}$ is the covariance matrix of prediction points, and $\boldsymbol{\Sigma_{PT}}$ is the covariance matrix between training and prediction points.

The uncertainty in the prediction values is zero at the training points, and it increases as the distance from the training points increases. As the variance increases, the GP surrogate model becomes a less suitable replacement for the underlying computational model. The prediction variance can typically be reduced by adding more training points and reconstructing the surrogate. If the underlying function is smooth and well-behaved, the prediction variance is a good predictor of the observed bias, and training points should be added in the regions of the domain with maximum prediction variance. However, when modeling more challenging functional behaviors, an adaptive approach to bias minimization may be implemented [42].

2.5 Summary

This chapter describes some of the fundamental components of a UQ methodology for prediction (Bayesian calibration, probabilistic model validation, and uncertainty propagation) that are widely used across engineering applications. Additional features of a comprehensive UQ framework are developed in subsequent chapters; in particular, the effect of model uncertainty and data uncertainty on these activities is considered. The proposed methods of this dissertation use the UQ framework to perform forward propagation of uncertainty. Efficient uncertainty propagation techniques (Chapter 3) are needed to perform model validation in the context of prediction (Chapter 4). Then, based on the solution approaches for the forward problem, the inverse problem of test selection in Chapter 5 can be developed, and risk-based resource allocation can be explored (Chapter 6).

CHAPTER 3

MODEL SELECTION FOR UNCERTAINTY PROPAGATION

3.1 Introduction

This chapter proposes an efficient approach to uncertainty propagation since it is often prohibitively expensive to evaluate computational models repeatedly. Uncertainty propagation is required in order to obtain the distributions of model output that are needed in both model validation and prediction. This propagation is performed by stochastic simulation that includes both aleatory and epistemic uncertainty in the prediction of an output quantity of interest (QoI). Often, the computational models are hierarchically composed, such that some aspects of the physics are modeled separately from others (leading to individual outputs of each component that are inputs to the prediction of interest). These component models are combined together to make an overall prediction that properly accounts for contributions from the sources of uncertainty that are present in each of the individual component models.

As described in Section 2.3, standard uncertainty propagation techniques, such as MCS, are available to propagate aleatory uncertainty in the model inputs. The presence of epistemic uncertainty sources that are considered in this dissertation makes the simulation procedure more challenging. Data uncertainty arises from sparse, imprecise, missing, subjective, or qualitative data, and also from measurement and data processing errors. Model uncertainty may arise due to model form assumptions, model parameters, and solution approximations. As described in Section 2.1, data uncertainty also affects the estimation of model parameters. If the various types of epistemic uncertainty are represented in a probabilistic format, the model prediction is stochastic at a particular input condition. The additional sources of uncertainty increase the number of model evaluations that are needed for accurate propagation, which likely makes it infeasible to include the full physics fidelity of the computational model in every simulation.

The situation explored in this chapter is one in which a stochastic simulation is performed for UQ and reliability analysis. A comprehensive analysis should accurately predict the full distribution of the output QoI by including all the sources of uncertainty. It is assumed that a high fidelity computational model already exists, but it is too expensive to evaluate at every sample point. Once cheaper models are developed (with respect to spatial resolution and/or physics complexity), the high fidelity model is still available, but the analyst must decide when to use it in order to obtain results of desired accuracy within an allowable amount of time. With this goal in mind, this chapter proposes a multi-fidelity model selection methodology that combines the use of both efficient simulation and surrogate modeling. The proposed framework uses surrogate models to inform the model selection decision at each random sample of the MCS (or each spatial location or time step, depending on the problem) and then executes a single selected model combination at this input. In this way, the framework can account for the possibility that different models may be adequate in different domains (including cheaper vs. expensive models, and even models with competing physical hypotheses). The proposed methodology accommodates different types of information about the QoI (such as actual observations, expert opinion etc.).

To develop a methodology for model selection, it must first be clear whether the ranking of fidelities among candidate models is consistent over the entire domain or whether it may change as a function of the inputs. In some situations, such as the comparison of a mathematical surrogate model with a physics-based computational model, it is obvious that the physics-based model is of higher fidelity. On the other hand, there are also instances in which multiple competing physics-based models are available for the same prediction, but it is not obvious which of them represents reality more accurately for the application of interest. For example, one physical phenomenon may be more dominant in one region of the input space than another. This situation has been addressed by quantifying the discrepancy between the model prediction and some performance benchmark [43]. Since it is not clear which model is providing the better estimate of reality, this benchmark must come from an additional piece of information, most commonly a physical observation, known exact solution, or expert opinion. After a benchmark is selected, the decision is a tradeoff of accuracy vs. computational expense.

Once the appropriate ranking of the fidelities among the candidate models is considered, the goal is to select among available models in an intelligent and efficient manner. Given these various scenarios, the general model selection problem can be posed as a decision based on one or more of the following criteria: (1) parsimony vs. accuracy in regression, (2) discrepancy compared to a benchmark, and (3) computational expense. The problem of selecting among multiple regression models has frequently been addressed by considering the first of these criteria. In several existing metrics based on information theory, accuracy is indicated by the sum of squares of residuals or the maximum likelihood with respect to training data, and parsimony is indicated by the number of terms in the model. Both of these components are included within Mallows' C_p statistic [61], the Akaike information criterion [1] based on information entropy [14], the Bayesian information criterion [102], and the minimum description length [28]. Each of these is addressing the tradeoff between bias and variance in available models, since additional complexity will reduce the residuals (i.e. variance) but also risks “overfitting,” which may increase bias. Typically, the outcome of this problem is the choice of a single model from a set, or possibly a new model which averages a set of available models.

When the models are not statistical regression models, but rather physics-based models, these metrics, based on the accuracy vs. parsimony criterion, can be difficult and inappropriate to employ for a couple of reasons. First, the forms of these models may be complex and in some cases impossible to write in an analytical form, so it will be difficult to define the parsimony of the model. Second, different physical hypotheses may attribute different physical mechanisms as causes for the observed behavior, which makes the associated models difficult to compare with respect to parsimony, and they cannot be combined in a natural way. Therefore, it is more appropriate to look only at model discrepancy and computational expense when addressing this selection scenario.

The tradeoff between accuracy (w.r.t. a benchmark) and computational effort in physics-based models has been addressed in the system design literature. It is possible to develop a more accurate model by introducing additional phenomenological features (i.e. improve the model form) and/or by improving the quality of the numerical approximation to the solution (e.g. discretization refinement). Available methods [65, 83, 90] assign utilities to the candidate models based on expected performance and explore the tradeoff between utility and the associated costs (both model building cost and execution cost). The use of multiple models with varying degrees of fidelity is also studied in the design optimization literature; this is referred to as model management [2]. Lower fidelity models to evaluate the objective and constraints include surrogate models or reduced-order models [4, 80].

Within this context, this chapter develops a model management framework for UQ, based on model discrepancy and computational effort, in the presence of both aleatory and epistemic uncertainty. Model discrepancy is probabilistically quantified for different model choices and traded off against computational effort to develop an optimization-based model selection criterion (instead of information theoretic metrics). Note that model choice is different for different samples of the input (or spatial location or time step), thus taking advantage of all the available models selectively at each point rather than making a single decision for all points in the input domain.

A simple mathematical example is first implemented to demonstrate a situation in which no prior information is available about the appropriate ranking of fidelities among candidate models. In such a case, additional information about the QoI is needed in order to define the relative accuracies in terms of a discrepancy. Otherwise, an informed decision cannot be made based on computational effort alone. Next, a richer engineering example is used to demonstrate the proposed methods for a more complicated simulation where inputs vary both spatially and temporally. Additionally, this second example establishes the model selection approach for a case where the ranking of fidelities among the candidate models is known *a priori*.

3.2 Model selection methodology

Consider a problem of the form given in Figure 3.1. A total of v subsystem models are needed, where each describes a physical phenomenon that produces an output that feeds into a full system model. For each subsystem i, a total of w_i competing models are available; these competing models are denoted g_{ij} (j denotes a model choice; j = 1 to w_i); they take the same inputs $\boldsymbol{X_i}$ but require a different set of parameters $\boldsymbol{\theta_{ij}}$. Each subsystem model produces the same intermediate output quantity of interest Y_i and the system-level QoI Z is a function of these subsystem outputs. These general relationships (for an example case where $v = 2$ and $w_1 = w_2 = 2$) are summarized by Eq. (3.1) to (3.3) below.

$$Y_1 = g_{11}(\boldsymbol{X_1}, \boldsymbol{\theta_{11}}) \text{ or } Y_1 = g_{12}(\boldsymbol{X_1}, \boldsymbol{\theta_{12}}) \tag{3.1}$$

$$Y_2 = g_{21}(\boldsymbol{X_2}, \boldsymbol{\theta_{21}}) \text{ or } Y_2 = g_{22}(\boldsymbol{X_2}, \boldsymbol{\theta_{22}}) \tag{3.2}$$

$$Z = h(Y_1, Y_2) \tag{3.3}$$

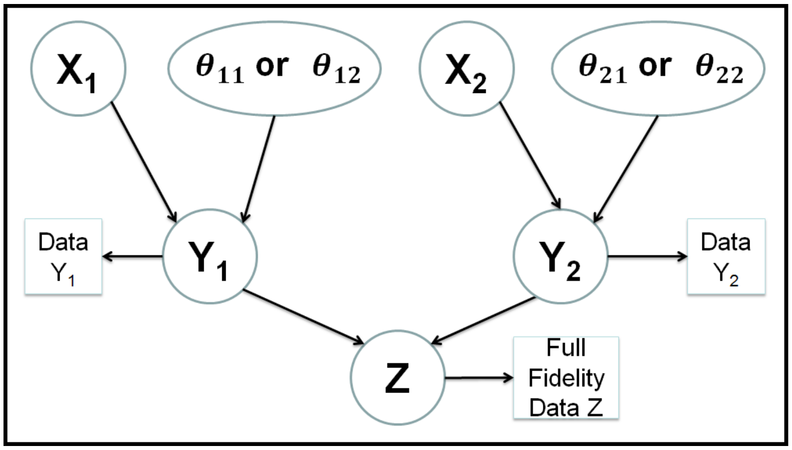


Figure 3.1: Example problem structure given by Eq. (3.1) - (3.3)

The possible model choices in this problem result in four model combinations to be considered, as shown in Figure 3.1:

- g^1 - model g_{11} for Y_1 and model g_{21} for Y_2
- g^2 - model g_{11} for Y_1 and model g_{22} for Y_2
- g^3 - model g_{12} for Y_1 and model g_{21} for Y_2
- g^4 - model g_{12} for Y_1 and model g_{22} for Y_2

In this chapter, a superscript k denotes a particular model combination that propagates a total of v input vectors $\boldsymbol{X_1}, \dots, \boldsymbol{X_v}$ through the corresponding subsystems to obtain $Y_1, \dots, Y_v$. All subsystem outputs are then propagated through the system model h to obtain Z. For w_i possible model choices for the respective subsystems, the total number of possible model combinations is denoted by p, where $p = \prod_{i=1}^{v} w_i$.

3.2.1 Model selection within Monte Carlo simulation

Experimental data may be available at various levels of the system hierarchy. However, one underlying assumption of this work is that data at the subsystem level is cheaper to procure and therefore more abundantly available than at the full system level. As such, data on the subsystem outputs Y_i is treated in a different manner from data on system output Z in this work. Subsystem level data is utilized in Bayesian calibration (described in Section 2.1) to provide updated distributions of each parameter set $\boldsymbol{\theta_{ij}}$. Each competing model within a subsystem can be calibrated from the same subsystem output data Y_i, but a separate calibration must be performed for each model option for each subsystem, requiring at total of, $\sum_{i=1}^{v} w_i$, Bayesian calibrations. Depending on the computational expense of each subsystem model, surrogate models may be necessary for each of them to improve the efficiency of the calibration.

Since the goal of the model selection procedure is to efficiently approximate the distribution of Z as closely as possible without direct regard for accuracy in each Y_i, available data on Z is used to inform system-level surrogate models that predict errors and uncertainties in Z as a function of $\boldsymbol{x}$. These surrogates are then used for online decision making at each sample of the input uncertainty. The surrogate model evaluations represent a trivial increase in the computational expense of the MCS. In particular, the GP surrogate models (described in Section

2.4) can be evaluated in a time on the order of 10^{-4} to 10^{-2} seconds depending on the number of training points and the number of prediction points. This evaluation time is negligible compared to any realistic engineering simulation where high-fidelity MCS is intractable. The system-level data that is used in surrogate model training and decision-making may be sparse, imprecise, or in some cases completely unavailable; these three situations are individually addressed below.

3.2.2 Case 1: Available sparse system-level data

If data can be obtained on the QoI, either by experiment or by some maximum fidelity (reliable) simulation, this data can be used to train error quantification models for the existing model combinations. In this chapter, GP surrogate models are constructed and used for decision making. Since the available output data is assumed to be well-characterized, i.e. measured values of the corresponding inputs are also available, all possible model combinations can be evaluated at these input values and compared with the given output data. However, since the parameters of each model are calibrated using a Bayesian method, posterior PDFs for the model parameters are available, and each model prediction is stochastic for a given set of input values. Therefore, an uncertainty propagation procedure is needed to account for parameter uncertainty in the surrogate model training. Because it requires a small number of model evaluations, a first order second moment (FOSM) approach is selected in this illustration to compute an approximate mean prediction for each model combination at each input value. The FOSM approach utilizes a first-order Taylor series expansion to calculate this mean value, and only one evaluation (at the parameter means) of each model combination is required to perform this calculation as in Eq. (3.4).

$$Z^k = g^k\left(\boldsymbol{x_1}, \dots, \boldsymbol{x_v}, \boldsymbol{\theta^k}\right)$$

$$E(Z^k) = g^k(\boldsymbol{x_1}, \dots, \boldsymbol{x_v}, \boldsymbol{\mu_{\theta^k}}), \text{ for } k = 1, \dots, p \tag{3.4}$$

The calculated mean prediction is subtracted from the experimental value at the corresponding input to give a mean error associated with each model combination. The computed mean errors and corresponding input values are used to train p GP surrogate models, each predicting a mean error in Z as a function of v input vectors $\boldsymbol{x_1}, \dots, \boldsymbol{x_v}$.

This surrogate model structure is important because it provides a direct mapping from the input space to the QoI. Since the decision is based on information predicted at the system level, it implicitly accounts for two important factors: (1) the amount of error associated with each model combination at the inputs of interest and (2) the sensitivity of the QoI to errors made in each subsystem-level prediction. Once training of these models is complete, suppose a full MCS over the input space is to be conducted to approximate the corresponding distribution of the QoI. A model combination is selected at each input sample that minimizes two objectives: cost and mean error. In the context of the model selection problem, the cost is the amount of computer time $\boldsymbol{t}$ required to evaluate the selected model combination at the particular input. The available budget is the amount of time available for the entire MCS. In this illustration, the error and time objectives are combined by a simple product of the two because a product formulation attributes equal weighting to both objectives regardless of the scaling of the quantities. For example, a 10 percent reduction in expected error will have the same impact on the combined objective as a 10 percent reduction in computation time. Other complicated bi-objective formulations can also be explored if there is a reason to attribute more weight to one objective than to the other.

For each iteration of the MCS, a sample of the inputs is taken, and the mean error of each model combination is predicted via an evaluation of the corresponding GP at that input sample. The model combination c with the minimum product of mean error e^k and computation time t^k is selected and executed to calculate a sample of Z. Figure 3.2 gives the pseudo-code for the procedure in Case 1:

```
$i$ = 0
cost = 0
while cost < budget
    $i$ = $i$ + 1
    generate samples $\boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{1},\dots,\boldsymbol{v}}$
    for $k$ = 1 : $p$
            $e^{ik} = GP^{\boldsymbol{k}}(\boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{1}}, \dots, \boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{v}})$
    end
    $c = \text{argmin}_k\, e^{ik} * t^k$
    $z^i = g^c\,(\boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{1}}, \dots, \boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{v}}, \boldsymbol{\theta}^{\boldsymbol{c}})$
    cost = cost + $t^k$
end
```

Figure 3.2: Algorithm 1 for model combination selection

3.2.3 Case 2: Available imprecise system-level data

Frequently, system-level data cannot be collected directly, but some imprecise data may be available in the form of an interval (range of values) for Z, such as from expert opinion. In such a case, it is not possible to build error models for the particular model combinations. Instead, the FOSM procedure is again utilized, but two GP models can be trained for each model combination: one for the mean prediction and one for the variance of the prediction. Since no particular input values are known, they must now be generated in a way that covers the input space in order to effectively train the surrogates. For this purpose, a Latin hypercube sampling technique can be employed. Since the number of samples selected consumes a specified percentage of the total allowable simulation budget, further improvements to the training procedure may be made by utilizing more advanced approaches such as optimal symmetric Latin hypercube sampling, bias-minimizing training techniques [42], and expected improvement functions [11]. The FOSM training procedure now requires $n+1$ evaluations of each model combination at each input point where n is the number of parameters associated with the particular model for which the surrogate is being trained. The additional n evaluations give gradient information at the mean values which is used to calculate the first-order variance in Eq. (3.5) in conjunction with Eq. (3.4).

$$Var(Z^k) = \sum_{i=1}^{n} \left(\frac{\partial g}{\partial \theta_i^k}\right)^2 Var(\theta_i^k) + \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\partial g}{\partial \theta_i^k}\right)\left(\frac{\partial g}{\partial \theta_j^k}\right) Cov(\theta_i^k, \theta_j^k) \text{ for } k = 1, \dots, p \qquad (3.5)$$

Once the mean and variance GP models are trained for each model combination, the procedure is similar to that in Case 1. At each MCS sample, the mean and variance GP models are evaluated for all model combinations. For example, suppose the distribution of the prediction for each model combination is assumed to be normal with mean and variance predicted by the GP. From this distribution, the probabilities of the prediction falling inside and outside the expert opinion interval $[E^L, E^U]$ can be calculated. The procedure continues as in Case 1, except that the "error" to be minimized is now defined by the probability of the prediction falling outside the expert interval, and the objective is again to minimize the product of computation time and "error". With this objective in mind, MCS samples are taken, and the optimal model combination is chosen at each sample until the computation budget is expended as demonstrated in the pseudo-code in Figure 3.3.

```
$i$ = 0
cost = 0
while cost  <  budget
    $i$ = $i$ + 1
    generate samples $\boldsymbol{x}_{\boldsymbol{1}}^{\boldsymbol{i}}, \dots, \boldsymbol{x}_{v}^{\boldsymbol{i}}$
    for $k$ = 1 : $p$
            $\mu_Z^{ik} = GP_{\mu}^{k}(\boldsymbol{x}_{\boldsymbol{1}}^{\boldsymbol{i}}, \dots, \boldsymbol{x}_{\boldsymbol{v}}^{\boldsymbol{i}})$
            $\sigma_Z^{ik} = GP_{\sigma}^{k}(\boldsymbol{x}_{\boldsymbol{1}}^{\boldsymbol{i}}, \dots, \boldsymbol{x}_{\boldsymbol{v}}^{\boldsymbol{i}})$
            $e^{ik} = 1 - \int_{E^L}^{E^U} normpdf(\mu_Z^{ik}, \sigma_Z^{ik})$
    end
    $c = \text{argmin}_k\, e^{ik} * t^k$
    $z^i = g^c\,(\boldsymbol{x}_{\boldsymbol{1}}^{\boldsymbol{i}}, \dots, \boldsymbol{x}_{\boldsymbol{v}}^{\boldsymbol{i}}, \boldsymbol{\theta}^{\boldsymbol{c}})$
    cost = cost + $t^k$
end
```

Figure 3.3: Algorithm 2 for model combination selection

3.2.4 Case 3: No system-level data available

The decisions in Case 3 are the most difficult since no information about the QoI is available. Therefore, there is no available measure of error, and it is difficult to quantify. Furthermore, when no model combination is clearly superior to the others based on physical intuition, there is no obvious benchmark for accuracy. In this case, the proposed procedure begins exactly as it did in Case 2 with the construction of mean and variance GP models for each model combination over a Latin hypercube input sample. The assumption of a normal distribution of the prediction is again made at each MCS sample point. To select among the p possible model combinations, an average distribution is created by taking a simple arithmetic mean of the GP predictions corresponding to each combination. The underlying assumption of the proposed approach for this situation is that the consensus prediction of all possible model combinations is the best indication of the true QoI when no data is directly available. This model averaging approach may not be appropriate in some situations, and it is particularly dangerous when there is substantial difference among the available model predictions. If averaging is not appropriate, the analyst must insist upon additional information on the QoI or some assertion about the ranking of fidelities of the candidate models. Additional information about the QoI would allow admit the proposed methods of Case 1 or Case 2 of this section, or an assertion about the ranking of fidelities would admit the approach presented in Section 3.4.

For cases where an averaging approach is reasonable, an "error" measure can be based on information theory via the Kullback-Leibler (KL) divergence [55] (Eq. (3.6) below), which is calculated between the average distribution and the distribution predicted by each individual model combination. If there is reason to give preference to one or more model combinations over the entire domain, the average distribution can be a weighted average rather than a simple arithmetic average.

$$D(g,h) = \int g(y) \ln[\frac{g(y)}{h(y)}] dy \tag{3.6}$$

The KL divergence is not symmetric, so the distance *from* the average distribution *to* the particular distribution of a given model combination, is used here as the error measure. The objective function for this case is a product of the computation time and this new error measure. The MCS again continues until the budget is reached as shown in the pseudo-code in Figure 3.4.

```
$i$ = 0
cost = 0
while cost < budget
    $i$ = $i$ + 1
    generate samples $\boldsymbol{x_1^i}, \dots, \boldsymbol{x_v^i}$
    for $k$ = 1 : $p$
            $\mu_Z^{ik} = GP_\mu^k(\boldsymbol{x_1^i}, \dots, \boldsymbol{x_v^i})$
            $\sigma_Z^{ik} = GP_\sigma^k(\boldsymbol{x_1^i}, \dots, \boldsymbol{x_v^i})$
    end
    avg_dist = $\frac{\sum_{k=1}^{p} normpdf(\mu_Z^{ik},\sigma_Z^{ik})}{p}$
    for $k$ = 1 : $p$
        $e^{ik}$ = KL distance from avg_dist to pdf $k$ (Eq. (3.6))
    end
    $c = \text{argmin}_k\, e^{ik} * t^k$
    $z^i = g^c(\boldsymbol{x_1^i}, \dots, \boldsymbol{x_v^i}, \boldsymbol{\theta^c})$
    cost = cost + $t^k$
end
```

Figure 3.4: Algorithm 3 for model combination selection

The proposed methods make several simplifying approximations, which are summarized here. First, the mapping from inputs $\boldsymbol{x_1}, \dots, \boldsymbol{x_v}$ to the system-level QoI Z is described by a GP model. Obviously, models of different types of physical phenomena will behave differently, but GP models have been shown to provide a robust and flexible tool for representing a wide range of processes. In most applications, these surrogates will provide a good approximation so that an appropriate model selection can be made. Second, the propagation of parameter uncertainty, which is necessary to train these surrogates, is performed by the FOSM method. The first-order Taylor series approximation may not be sufficient for complex parameter relationships, and a higher order approximation may be necessary. Finally, in the model selection step, the output QoI Z is assumed to have a normal distribution (whose mean and variance are predicted by the corresponding GP models), only in order to compute error measures. This assumption is primarily made for illustration, since in general other distributions could be chosen to suit a given problem if more information about Z is available. Note that this assumption is only for the sake of model selection; the predicted distribution of Z after the simulation could be of any form; only numerical kernel density fits are in fact reported.

3.3 Illustrative example

To demonstrate the proposed model selection methodology, an illustrative problem of the form given in Figure 3.1 with simple analytical models and a known "reality" to generate data is utilized. Both X_1 and X_2 are assumed to follow a uniform distribution over the interval $[-1, 1]$. The "reality" is the cubic model in Eq. (3.7) which connects X_1 to Y_1 and the cubic model in Eq. (3.8) which connects X_2 to Y_2. The outputs Y_1 and Y_2 are used in the system model given by Eq. (3.9) to predict the system-level QoI *Z*.

$$Y_1 = 1 + 2X_1 + 3{X_1}^2 + 4{X_1}^3 \tag{3.7}$$

$$Y_2 = 4 + 3X_2 + 2{X_2}^2 + {X_2}^3 \tag{3.8}$$

$$Z = Y_1 + 5Y_2 \tag{3.9}$$

Now, assume that the actual functions in Eqs. (3.7) and (3.8) are not known. Instead, for each subsystem, a linear and a quadratic model are available. The two models for Y_1 have the forms of Eq. (3.10) and (3.11) respectively and are calibrated to available subsystem data.

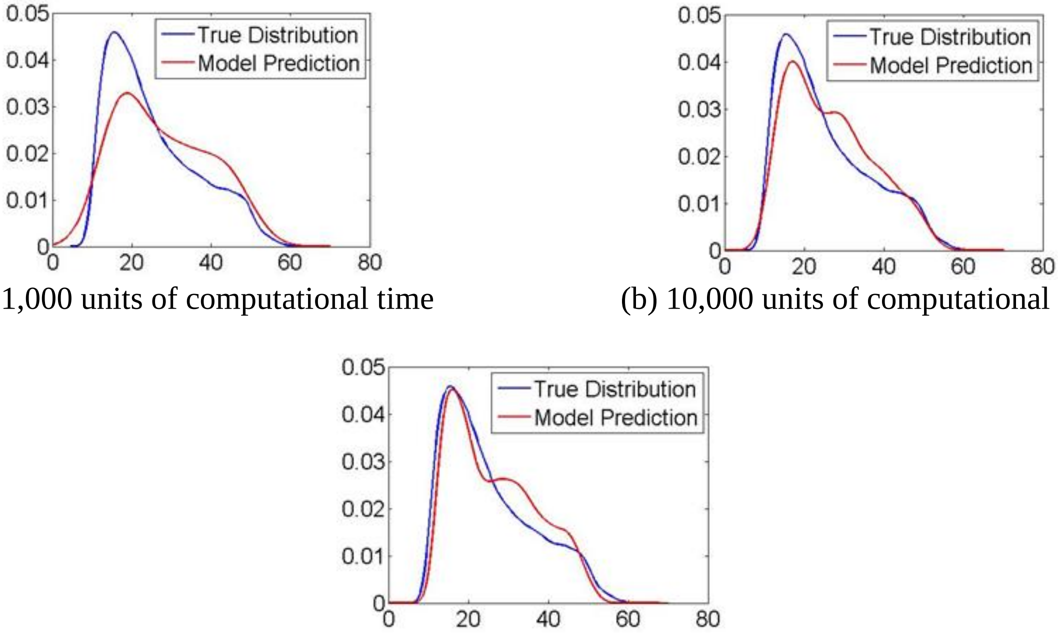
$$g_{11} : \; Y_1 = \theta_{11}^{(1)} + \theta_{11}^{(2)} X_1 \tag{3.10}$$

$$g_{12} : \; Y_1 = \theta_{12}^{(1)} + \theta_{12}^{(2)} X_1 + \theta_{12}^{(3)} {X_1}^2 \qquad (3.11)$$

Similarly, two available computational models for Y_2 are calibrated to a subsystem level data set. For such simplistic analytical models, the computation time needed to evaluate them is obviously negligible on modern machines, but to exercise the methodology the models were assigned costs based on the number of floating point operations required (two for the linear and five for the quadratic models). Therefore, four model combinations are available with computational times 4, 7, 7, and 10 units respectively.

3.3.1 Case 1

Noisy system-level data generated from the "reality" (Eqs. (3.7) and (3.8)) was utilized to construct the GP error models for Case 1. The analysis proceeded in four steps: (1) generate an input sample x_1 and x_2, (2) select a model combination using the surrogate error models; (3) sample a realization of vector $\boldsymbol{\theta^k}$ from the corresponding calibrated joint parameter distribution to account for parameter uncertainty in the selected model combination; and (4) calculate the output z. These four steps are repeated multiple times (as allowed by the computational budget) to construct the predicted distribution of the QoI Z. The "true" distribution of Z (computed from exhaustive sampling of the known "reality") is computed, and shown along with the predicted distribution (based on the proposed model selection strategy) in Figure 3.5. Three results of model selection are shown for budgets of 1000, 10000, and 100000 units of computational time are shown. (Note that the selected model combination is different for each Monte Carlo sample of the input and is chosen using Algorithm 1 in Figure 3.2).



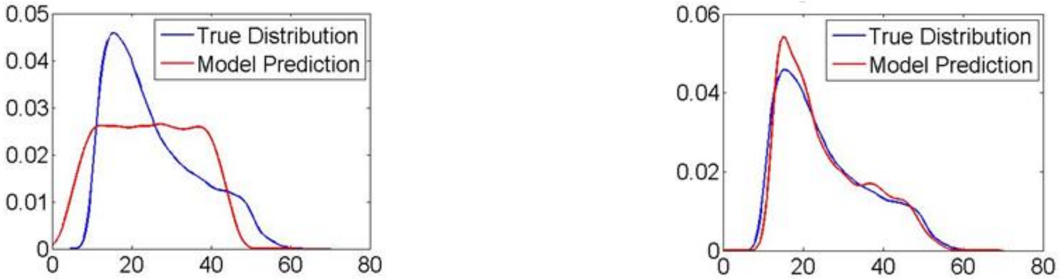
(a) 1,000 units of computational time

(b) 10,000 units of computational time

(c) 100,000 units of computational time

Figure 3.5: Improvement in accuracy with computational budget

As the allowable budget for the computation increases, the predicted distribution converges toward the true distribution. With a budget of 100,000 computational units, the prediction demonstrates good agreement with the unknown truth. When compared to just blindly evaluating the same model combinations everywhere, the method gives a good prediction much more quickly. For example, suppose 50,000 MCS samples are used. If only the linear model was selected for both Y_1 and Y_2 for all of the 50,000 MCS samples (corresponding budget = 200,000 units, the least expensive option), the resulting model prediction is given in Figure 3.6a. If instead the quadratic model was selected for both Y_1 and Y_2 for all of the 50,000 MCS samples (corresponding budget = 500,000 units, the most expensive option), the resulting model prediction is given in Figure 3.6b. The quadratic models are able to describe the population from the cubic model fairly well after a large number of samples, whereas the linear model combination is not sophisticated enough to capture the behavior of the true system for any number of samples since a linear transformation of a uniform input distribution still behaves like a uniform distribution. However, the linear model may be adequate in some regions where the actual system behavior is not too non-linear. The proposed method is able to exploit this property, i.e., linear models are adequate for both subsystems in some regions, quadratic models are necessary for both subsystems in some regions, and linear model for one subsystem and quadratic model for another subsystem are adequate in some regions. Of course, GP surrogate models are used to make this selection; therefore the accuracy of the prediction is also dependent on the accuracy of the GP models.



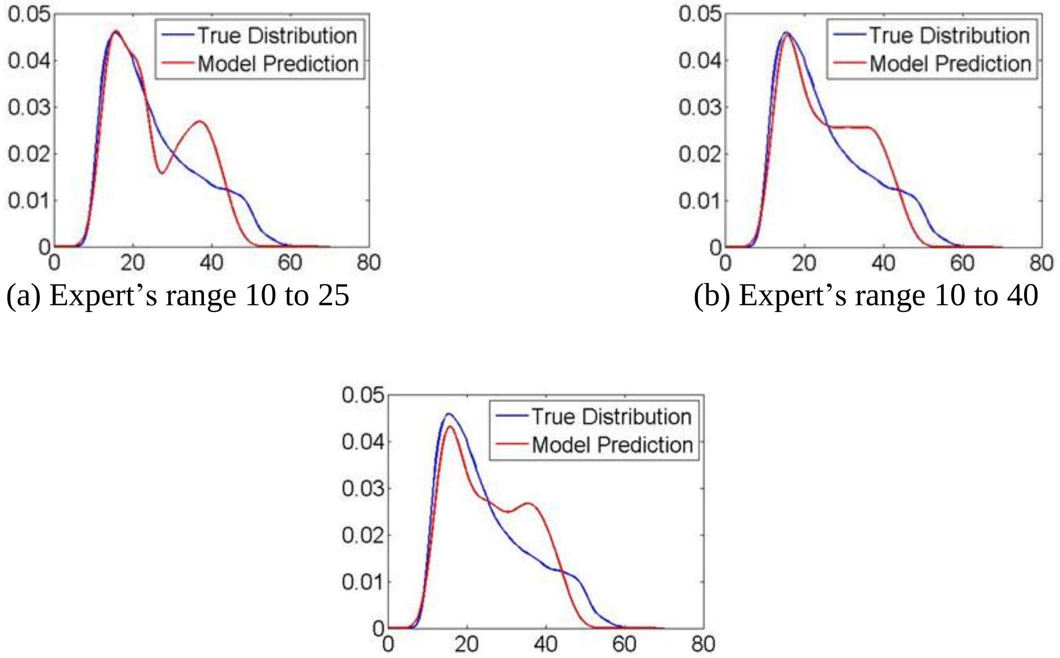
(a) Linear models only

(b) Quadratic models only

Figure 3.6: Effect of fixed model choices for all samples

3.3.2 Case 2

In this case, an expert opinion interval is assumed to be available in order to demonstrate the impact of the quality of the expert opinion given. No data from the reality is assumed to be available to guide the model selection. A Latin hypercube sample is taken over the input space in order to train mean and variance GP models for each model combination. The FOSM procedure is utilized at each sample point to propagate parameter uncertainty and obtain first-order mean and variances, which correspond to GP training values. When the MCS is conducted, the mean and variance are predicted at each sample using the GP models, and the assumption of a normal distribution enables a simple calculation of the probability of falling outside the expert's interval. The model combination that minimizes the product of this probability and the computational time is chosen at each input sample. The results for three different expert intervals for Z are given in Figure 3.7. In each case, a budget of 10,000 units was expended.



(a) Expert's range 10 to 25

(b) Expert's range 10 to 40

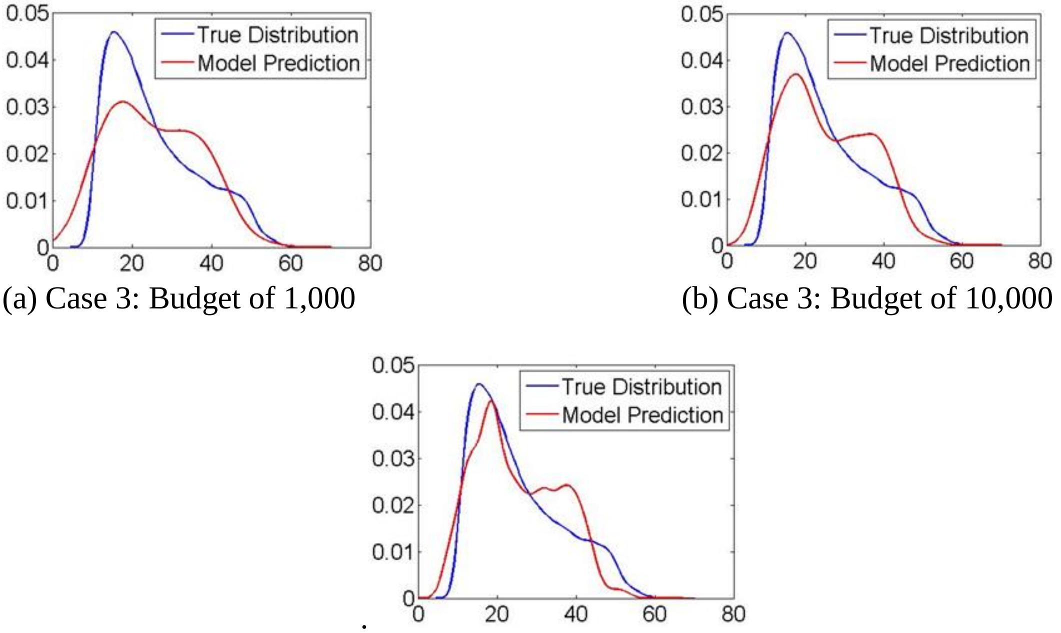
(c) Expert's range 10 to 60

Figure 3.7: Impact of expert opinion quality

The results demonstrate that the range given by an expert will impact the model selection algorithm, and poor information may cause the algorithm to select a model with insufficient fidelity outside the range. As the interval becomes wider, the probability of falling outside of it may be correspondingly smaller for all possible model combinations. If the integrals of the distributions predicted by all the models are close to unity over the range given by the expert, then the cheapest model is always selected. Only on the edges of the interval does the algorithm begin to discriminate between the model combinations well. As shown in Figure 3.7a, the optimized solution describes the true solution well in the small interval that was selected $[10, 25]$ and will choose the cheaper model when both model predictions are likely to fall outside the interval.

3.3.3 Case 3

For Case 3, no observation data or expert opinion is available. Since no *a priori* information about the quality of the model options is available, the model combinations are all given equal weights. For the sake of illustration, it is assumed that the model predictions can be logically combined into an averaged form. The consensus prediction of the four model combinations is treated as the best idea of the true behavior, and the KL distance metric to the average distribution (weighted by computational expense) becomes the selection criterion. If some information about the quality of the models were available upfront, benchmarking off the best available model or assigning unequal weights to the distributions would also be viable alternative methods. Results for the equally weighted case with budgets of 1,000 units, 10,000 units, and 100,000 units are shown in Figure 3.8.



(a) Case 3: Budget of 1,000

(b) Case 3: Budget of 10,000

.

(c) Case 3: Budget of 10,0000

Figure 3.8: Effect of increasing budget with unknown reality

The results for Case 3 converge reasonably well toward the true distribution in some regions of the distribution. The model selection algorithm has no knowledge of the underlying truth at all except via the subsystem data used to calibrate the subsystem model parameters. Treating the linear and quadratic models as equally valid in the weighting process did not skew the result in regions of the domain where the discrepancy was large, but it is clear from Figure 3.6 that choosing both quadratic models is most accurate over the entire domain. Therefore, even when there are small discrepancies between the linear choices and the consensus prediction, it is not optimal to select the linear models, and this selection will cause some prediction errors. Some prior information on the ranking of the fidelities of the models would help to solve this problem by helping to select appropriate weights.

3.4 Simulation over time

The previously described methodology considers a problem where each random input sample of a MCS requires only evaluating a model combination once to predict the QoI. In contrast, many problems vary over space and time and may require repeated calls to a model even for a single input sample. In this case, some input samples may correspond to realizations of random process or random field quantities in the system. For example, a particular input may define a random process cyclic loading on a system, and the output of one cycle becomes an input to the next cycle of the simulation. In such a case, potential frameworks may (1) select a model combination at each cycle of the simulation, (2) perform temporal discretization of the load process and select a model combination for each discrete block load, or (3) select a model combination for the entire load history. The second case is considered here (model selection for each load block). Consider a realization of the input random process $\boldsymbol{X}$ and cyclic output response history $\boldsymbol{Y}$ related at cycle i by

$$Y_i - Y_{i-1} = g(X_i, Y_{i-1}, \boldsymbol{\theta}) \tag{3.12}$$

Note that this will require an initial value Y_0 in order to evaluate the first input. This initial value is itself a random input to the system. If the entire realization $\boldsymbol{x}$ and the initial value y_0 are sampled, $\boldsymbol{x}$ is discretized into blocks of n cycles, and the system can be approximated by Eq. (3.13).

$$Y_{i+n-1} - Y_{i-1} = n * g(X_i, Y_{i-1}, \boldsymbol{\theta}) \tag{3.13}$$

For this situation, the temporal discretization becomes an additional decision variable, and the problem can be posed in two different contexts: (1) optimize cost and discrepancy jointly (Section 3.4.1) or (2) specify an allowable computational budget and minimize the uncertainty in the prediction within that budget (Section 3.4.2).

3.4.1 Minimizing the product of cost and error

Suppose k $(k = 1, \dots, p)$ is a possible model combination that predicts the output for a single cycle of a given input. The FOSM procedure (as in Section 3.2) can be applied to account for the uncertainty in $\boldsymbol{\theta^k}$ by taking a Latin hypercube sample of $\boldsymbol{x_i}$ and $\boldsymbol{y_{i-1}}$ values and propagating the distribution of $\boldsymbol{\theta^k}$ through all possible model combinations at each pair (x_i, y_{i-1}). A GP surrogate model is trained for the mean prediction and the variance of the prediction over the input space for each model combination. One advantage of the GP is that its efficiency allows the model to be evaluated on a cycle-by-cycle basis without discretizing into blocks as is necessary for the higher fidelity models. Thus, starting from cycle i, the mean output after m cycles can be approximated for each of the model combinations as

$$\mu_{Y_{i+m}^{k}} = \sum_{l=i+1}^{i+m}[GP_{\mu}^{k}(x_l^k\,, y_i^k)] \qquad \text{for } k = 1, \dots, p \tag{3.14}$$

The variance for each model combination can also be accumulated under the normality assumption. Therefore, the standard deviation of Y after m cycles have passed starting from cycle i can be approximated for each of the model combinations as

$$\sigma_{Y_{i+m}^{k}} = \sqrt{\sum_{l=i+1}^{i+m}[GP_{\sigma}^{k}(x_l^k\,, y_i^k)]} \qquad \text{for } k = 1, \dots, p \tag{3.15}$$

Frequently, there may be reason to assume that a particular model combination is more accurate than the others if, for example, it uses a finer spatial resolution or a more sophisticated physics model. In Sections 3.2 and 3.3, no such assumption was made (though it could be included by introducing weights as was previously mentioned), and hence all models were given equal weights in constructing the average distribution. A similar approach could be utilized in the time-dependent problem if no information were available about the ranking of fidelities among the candidate models. However, in some cases, it might be obvious that one model should be trusted more than the others because this maximum fidelity model includes all of the physics described by its competitors in addition to incorporating additional physical complexity or providing higher resolution. Even so, it might not be necessary to use the maximum fidelity model for every realization or for every instant and location in order to meet a given accuracy target. Since this scenario poses a tradeoff decision between accuracy and complexity, the methodology that follows here is a technique for selecting the model (among several cheaper models and the highest fidelity model) to evaluate over each block discretization of the input by considering the expense of a model and its discrepancy from the highest fidelity choice. A normal distribution can be constructed for the output of each model combination with the mean and standard deviation estimated by Eq. (3.14) and (3.15). The highest fidelity model combination b (among the possible candidates k) is assumed to be the maximum fidelity model for each subsystem.

Given that the most accurate (and expensive) model is known, the analyst must decide how much deviation from this model is acceptable. From a decision maker’s perspective, it is often possible to establish some acceptable error bars on the prediction (e.g. based on the precision of experimental instrumentation or the width of the maximum fidelity model’s uncertainty).

Therefore, a tolerance ϵ for the discrepancy between the most accurate (and expensive) combination and the other combinations is introduced. Given that the computation time t associated with each model combination is known, the optimal model combination c can then be selected by taking the model combination with the lowest product of computational time and probability of discrepancy greater than the specified tolerance as shown in Eq. (3.16) and (3.17).

$$e^k = P[(Y_i^k - Y_i^b) > \epsilon] \tag{3.16}$$

$$c = \text{argmin}_k\, e^k\, t^k \tag{3.17}$$

The implication of this treatment is that a less expensive model combination will be selected when its mean prediction agrees strongly with the mean prediction of the highest fidelity model and the variance of the prediction is small. Once a model combination is selected, it cannot be evaluated cycle-by-cycle as the GP was, so only one input value x_i can be chosen for the entire duration of the block n_k. Since the GP corresponding to the selected combination has already been evaluated at all $\boldsymbol{x}$ between x_i and x_{i+n_k}, Eq. (3.13) can be applied to guide the selection decision. In particular, the discrete input point x_l^k from that range with mean GP prediction, i.e. $y_{i+n_k}^k = n_k * GP_\mu^k(x_l^k, y_i^k)$, closest to the accumulated mean GP prediction for the maximum fidelity model combination, $y_{i+n_k}^b$ should be selected.

This selection procedure continues until the number of cycles that have been discretized and analyzed is equal to the desired total simulation length N. This procedure can then be repeated for many realizations of the input $\boldsymbol{x}$ and initial output values $\boldsymbol{y_0}$. From these samples, the distribution of interest will describe Y_N, the final value of the output for each realization.

3.4.2 Variance minimization for a fixed simulation time

If instead of simultaneously considering time and discrepancy there is a fixed time to perform a simulation of a given number of realizations, the corresponding time for a single realization of the random process can be specified. The temporal discretizations required to achieve the desired simulation time follow directly from the time required for one cycle of each model combination. Model combinations with more computational expense must be discretized more coarsely in order for them to run within the specified budget. Once all model combinations are forced to take the same amount of time, they can be compared on the basis of error alone. Given a number of cycles n_T to simulate over total time t_T and a vector $\boldsymbol{t}$ of computational times for one evaluation of each model combination k, a vector $\boldsymbol{n}$ of the temporal discretization for each combination can be computed with Eq. (3.18), and m is calculated as the largest value of $\boldsymbol{n}$ as in Eq. (3.19).

$$\boldsymbol{n} = \frac{n_T \boldsymbol{t}}{t_T} \tag{3.18}$$

$$m = \max(\boldsymbol{n}) \tag{3.19}$$

Then, starting from cycle i, the mean output after m cycles can be approximated for each of the model combinations as in Eq. (3.14). However, the variance for each model combination is only accumulated for the number of cycles for the particular temporal discretization required. Therefore, the standard deviation of Y at m cycles after cycle i can be approximated for each of the model combinations as

$$\sigma_{Y}{}_{i+m}^{k} = \sqrt{\sum_{p=i+1}^{i+n_k} [GP_{\sigma}^{k}(x_p^k, y_i^k)]} \qquad \text{for } k = 1, \dots, p \tag{3.20}$$

The result of this treatment is that the standard deviations of the predictions of the faster (cheaper) models are smaller than their more expensive counterparts because of the finer temporal discretization. Since time is no longer a consideration (all models are allotted equal computational time), only the discrepancy needs to be considered in the decision, and the optimal combination c can be selected in a similar manner to Section 3.4.1 as

$$c = \mathrm{argmax}_k \, P[\left(Y_i^k - Y_i^b\right) < \epsilon] \tag{3.21}$$

If none of the alternative model combinations can meet this tolerance criterion with a high confidence (e.g. 95%), then the benchmark model combination itself should be executed. The simulation then proceeds exactly as shown in the previous section.

3.5 Numerical example

To demonstrate the methodology developed in Section 3.4 for time-dependent analysis, an engineering example problem is developed here. The problem under consideration is a cantilever beam with a planar fatigue crack at a small distance from the fixed support. The randomness in the beam's elastic modulus (E) is described by a random field along the length of the beam. A random process cyclic loading P is applied to the free end of the beam for a period n_T equal to 100,000 cycles. Random process and random field variation have been accounted for by several available approaches in the literature such as ARMA methods [67], spectral representation methods [103], Karhunen - Loeve (K-L) expansion [26], and wavelet representations [29]. The K-L expansion approach is utilized here for the sake of illustration, and as a result, the random process P and random field E are represented by a small number of random variables to be sampled within MCS. The beam model and a single realization of the load process are illustrated in Figure 3.9. The structure is analyzed by the commercial finite element method (FEM) solver ANSYS.

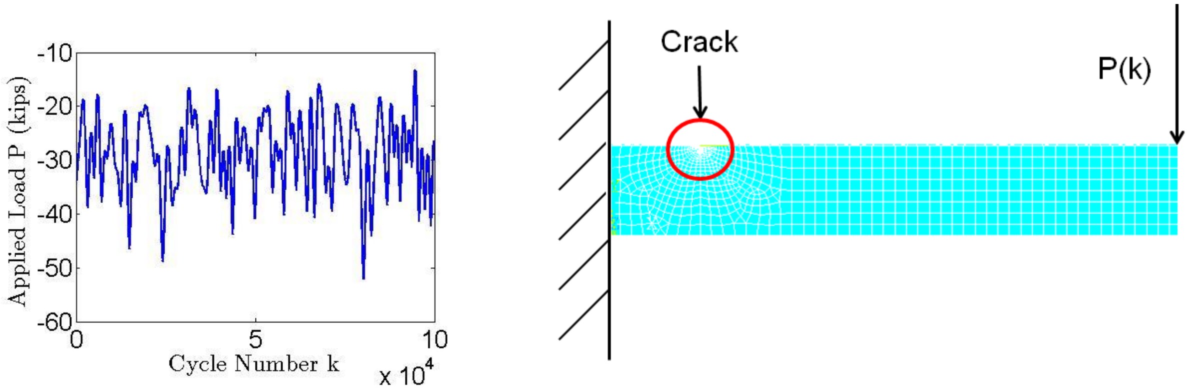


Figure 3.9: Example problem structure

The goal of the problem is to determine the predicted distribution of the final crack size A_f at the end of 100,000 cycles. This distribution could then be utilized within a reliability framework to easily estimate the probability of the beam deflection exceeding an allowable deformation. To solve a problem of this form, the stochastic simulation has to be performed at two levels: (1) an outer loop in which the problem inputs and parameters common to an entire load process are sampled and (2) inner loop cyclic simulations in which the model combinations are selected and the uncertain crack growth parameters needed for each load block are sampled. The main distinction between these two sources of uncertainty is that the outer loop captures aleatory variability in the uncertain inputs to the system while the inner loop captures epistemic uncertainty about the precise value of the crack growth parameters, which are in reality deterministic for a given material specimen. The outer loop variables are the random variables that define the random load process P, the material properties of the beam, and the initial crack size A_I^b. The inner loop samples of the model parameters C and m define the Paris law [75], a simple power law commonly used for fatigue crack growth as shown in Eq. (3.22), based on linear elastic fracture mechanics. The stress intensity factor ΔK is a function of the current crack geometry and applied load, and it is used to predict the rate of crack growth during the cycle ($\boldsymbol{\frac{da}{dN}}$) as

$$\frac{da}{dN} = C(\Delta K)^m \tag{3.22}$$

Data is assumed to be available at two subsystem levels: (1) an axial test used to calibrate the parameters of the random field E and (2) a simple mode I fracture test used to calibrate C and m. Two potential modeling choices are made within the context of this example: (1) linear vs. nonlinear material behavior and (2) coarse vs. fine mesh around the crack tip. The linear material model requires only the random field elastic modulus E. The nonlinear material model assumes bilinear isotropic hardening which requires E in addition to the yield stress σ_y and the tangent modulus T which defines the stress-strain relationship above the yield stress. Two mesh refinements around the crack tip h_1 and h_2 are considered for each of these material models leading to four possible model combinations that may be selected:

- g^1 – linear model with coarse mesh
- g^2 – linear model with fine mesh
- g^3 – nonlinear model with coarse mesh
- g^4 – nonlinear model with fine mesh

A diagram of the test problem structure is provided in Figure 3.10.

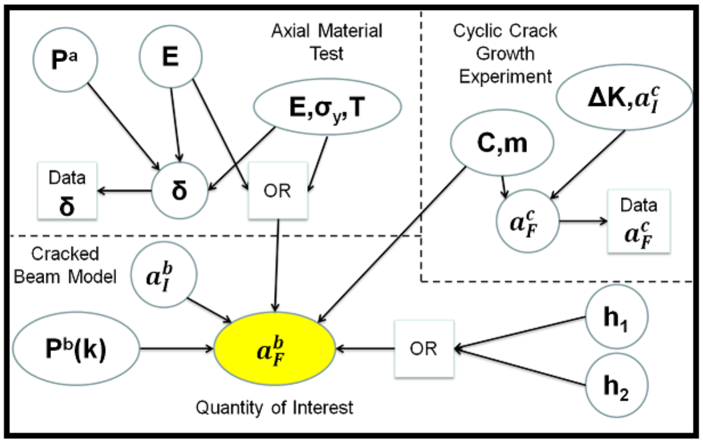


Figure 3.10: Crack growth test and simulation system diagram

Parameter uncertainty in C and m is propagated using FOSM over a Latin hypercube sample of the inputs for all available competing models. Here x_i is equal to p_i and y_{i-1} is equal to a_{i-1} in Eq. (3.12). The first order means and variances of each model combination prediction at each input sample are used to train GP surrogate models that predict crack growth in a single cycle given a current load step and geometry. The MCS begins by sampling a realization of the random field E, random process P, and initial crack size a_I^b (uniform distribution between 0.36 and 0.42 inches).

3.5.1 Stochastic simulation results considering both cost and time simultaneously

A simulation for n_T = 100,000 cycles is to be performed for each of 1,000 input realizations. The computational time vector $\boldsymbol{t}$ (here [0.7703, 0.8251, 1.0804, 1.0720] seconds for the four aforementioned model combinations) is calculated by the average times required for evaluations of each model combination during training. A block size of 4,000 cycles was fixed for this portion of the study, so 25 blocks were required for each realization. Utilizing Eq. (3.14) and

(3.15), the mean and variance is predicted for each model combination for each load block of each realization. Using model combination 4 as the benchmark, the probability of agreement within the tolerance is determined and weighted by the computation time as in (3.16) and (3.17) to select an appropriate model combination for each block. Alternative models will only be selected when they provide a significant time savings and agree well with the benchmark. Since the times for all the model combinations are very close to one another in this example (causing the benchmark itself to be predominantly selected), the effectiveness of the proposed procedure is illustrated by artificially increasing the expense of the benchmark model (combination 4) to five times and ten times its actual duration. A comparison of the results for these two cases and the unscaled case is shown in Figure 3.11.

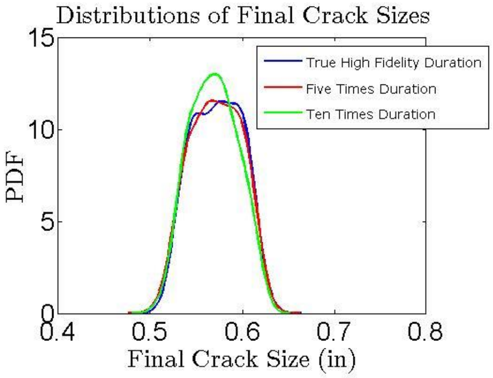


Figure 3.11: Effect of full fidelity model expense

Table 3.1 demonstrates the amount of utilization of each model combination as a function of the relative expense of the high-fidelity model combination. Each simulation requires a total of 25,000 model decisions (25 blocks for each of 1,000 realizations). Table 3.2 compares the total simulation times for each of these three levels of high-fidelity model expense.

Table 3.1: Model combination utilization for three levels of high-fidelity model expense

Model Comb.	Material Model	Refinement	% of Calls Unscaled Time	% of Calls Five Times Scaled	% of Calls Ten Times Scaled
1	Linear	Coarse	0.30	0.88	0.92
2	Linear	Fine	15.51	57.92	69.00
3	Nonlinear	Coarse	0.26	0.50	0.46
4	Nonlinear	Fine	83.92	40.70	29.62

Table 3.2: Total simulation times for three levels of high-fidelity model expense

High-Fidelity Time Per Evaluation (sec)	% of Calls to High Fidelity	Total Simulation Time (hr)
1.07 (unscaled)	83.92	14.28
5.35 (5 times scaled)	40.70	18.47
10.7 (10 times scaled)	29.62	25.97

*Note: The simulation times for the scaled cases were calculated from the assumed run times.

These results demonstrate that the efficiency of the proposed methodology is closely tied to the expense of the high-fidelity model. The time savings is substantially improved when the benchmark model is substantially more costly to evaluate than its alternatives (a common situation in engineering problems). For example, when the high-fidelity model expense increases by a factor of 10 (i.e. 1000%), the total simulation time only increases by about 80%. As shown in Figure 3.11, the effect on the overall accuracy of the distribution of the QoI is minimal since the algorithm does not allow for an alternative model to be selected when it deviates strongly from the benchmark. If the demand on accuracy is even more stringent, the proposed methodology allows the analyst to make a tradeoff decision by adjusting the tolerance in Eq. (3.21). A tighter tolerance will cause the simulation to run slower but with greater accuracy and vice versa.

3.5.2 Stochastic simulation results for fixed time

The desired time t_T for a single realization is selected to be 50 seconds for n_T = 100,000 cycles. The computational time vector $\boldsymbol{t}$ (as in Section 3.5.1) is used in Eq. (3.18) to compute the vector $\boldsymbol{n}$ (here [1541, 1650, 2161, 2144] cycles) of discretizations required for each model combination. For the first temporal discretization block, the mean predicted crack growth is calculated for m (here 2161, given by Eq. (3.19)) cycles. This is done for each model combination using the mean GP models as in Eq. (3.14), and the corresponding variance of the predicted crack growth is calculated for n_k cycles using the variance GP models as in Eq. (3.20). The criterion in Eq. (3.21) is utilized to select the optimal model combination for the load block, and the selected model is evaluated at a_{i-1} with P_i selected to match the highest fidelity model most closely at cycle $i + m$. This process is again repeated until i equal to n_T is obtained, and the final value a_f is determined for each realization. The distribution of A_f shown in Figure 3.12 is again obtained for 1,000 realizations of the inputs. Table 3.3 gives the number of calls to each of the four model combinations during the full simulation (100,000 cycles x 1,000 realizations).

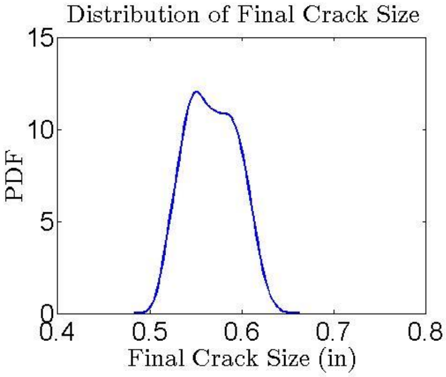


Figure 3.12: PDF of final crack size, A_f

Table 3.3: Comparison of calls for various model combinations

Model Combination	Material Model	Mesh Refinement	Number of Calls	% of Total
1	Linear	Coarse	662	1.20
2	Linear	Fine	43728	79.42
3	Nonlinear	Coarse	237	0.43
4	Nonlinear	Fine	10430	18.94

3.5.3 Discussion

Both of the treatments explored here (Sections 3.5.1 and 3.5.2) show that model combination 2 (linear model, fine mesh) was selected most frequently by the algorithm. The physical interpretation of this is that there was not much nonlinear behavior in a large portion of the input domain being considered, so the mesh refinement was a much more critical factor. Exploiting this type of information is the main objective of the methods in this chapter, and the proposed strategy is seen to satisfy this objective by selecting cheaper simulation options where they are adequate. The discretization error can have a large effect on the numerical calculation of the stress intensity factor (and therefore the crack growth), so this result is reasonable. As the load grows, the nonlinear effect on the result becomes more pronounced, so it is important to use the nonlinear model for some cycles. For such cycles, model combination 4 was typically selected because none of the other three could closely replicate this behavior. The result of the simulation is a synthesis of all the modeling options used selectively throughout the domain. This treatment improves efficiency, and it is protected from deviating significantly from the physics of the highest fidelity model by the tolerance choice in Eq. (3.16). A tighter tolerance can be chosen to ensure a close match with the high-fidelity model at the cost of spending more computation time.

Only 25 function calls (i.e. load blocks) were used for each realization of the load process in Section 3.5.1, and an average of 55 blocks were needed for the analysis in Sec. 3.5.2. The amount of computational time saved through temporal discretization is obviously enormous. The entire simulation of 1,000 realizations ranged from 15 to 35 hours to perform on a single processor of a PC while a cycle by cycle simulation of the highest fidelity combination would require about 70,000 hours (clearly intractable). Both proposed approaches select the model combination that minimizes the prediction variance (when mean predictions agree well), and it may therefore lead to only a negligible error with respect to the cycle by cycle case as is shown in the verification example that follows. It is clear that the proposed approach is most efficient when there is a substantial difference in the runtimes associated with the candidate models (in particular when the highest fidelity model is prohibitively expensive).

The proposed decision-making strategy provides this reduction in the computational expense of the simulation with only a minimal increase in expense coming from the model selection method itself. Some additional evaluations of the computational models are needed to train the GP surrogates; in this example, those calls represent less than 1% of the total evaluations. However, once the surrogates are built, the GP surrogate-based selection process is very fast (less than 0.1% of the expense of an evaluation of the computational models). The only substantial addition to the computational cost comes from the overhead in communicating with a driver program that makes the decisions and calls the computational models. In particular, the example crack growth analysis is performed by ANSYS FEM models that were driven by MATLAB scripts. The overhead associated with these two programs did increase the expense by as much as 50% in some cases. However, that increase seems large because the models in this example are substantially faster than would be expected in most applications; thus, the overhead represents a larger portion of the total expense. The benefit of the proposed approach would be much more dramatic for more expensive simulations since the overhead is only related to communication time between MATLAB and ANSYS and is independent of how long the ANSYS code takes to run. Furthermore, if the amount of overhead is significant, the model can be implemented in a generic programming language where the driver scripts are also implemented, and the overhead expense can be avoided altogether.

The proposed selection approach is intended to be non-invasive (i.e. the models can be treated as black boxes). This method is applicable to the crack growth example presented here because this problem is solved by a series of static analyses. At any cycle, only the output information from the previous cycle is necessary, not the details of the analysis in the previous cycle. The choice of linear vs. nonlinear model, or coarse vs. fine mesh, is based on load value and current crack size in any cycle. The error incurred by the low-fidelity model grows as the stress intensity factor grows larger, and the large stress intensity factor could be due to either a large load value or a large crack size. Since each cycle actually has a separate analysis, the selections can be made independently without causing any physical inconsistencies among the available models.

Since some expert judgment is needed to make decisions about how to define the parameters of the formulation (e.g. ϵ), there is no analytical proof that the proposed approach is “optimal.” However, the proposed method provides a systematic way of dealing with practical simulation constraints. This approach will never be slower than the brute force approach of calling the highest fidelity model every time, and it will be much faster when the lower-fidelity alternatives offer acceptable accuracy (i.e. the expected differences between high fidelity and low fidelity are small). In fact, the frequency with which lower fidelity models are called gives a clear indication of the quality of the lower fidelity options. Furthermore, the tolerance parameter gives an upper bound on how much error could be admitted into the problem by model selection decisions that are forced by computational time constraints.

3.5.4 Verification example

To demonstrate the efficiency of the proposed approach, a cycle-by-cycle simulation of a single set of input realizations was performed using the maximum fidelity model choice (combination 4). The initial crack size was sampled from the given distribution as 0.4122 inch, and crack growth was simulated for 25,000 cycles (because of time constraints). In this period the high fidelity model evaluated cycle-by-cycle predicted a final crack length of 0.4438 inch. Using the same initial crack size as well as the same load random process and material random field realization, the proposed approach predicted a final crack length after 25,000 cycles of 0.4428 inch. Thus, the error produced is only 0.23% while the computation time is reduced from 15 hours to 30 seconds. Note that in this problem, the computation time reduction is primarily due to the load block discretization since all the competing models have similar computational expense.

3.6 Conclusion

In the literature, model selection decisions are typically made only once at the beginning of the simulation and the choice is fixed for the rest of the simulation. This chapter proposes that this practice can be improved by taking advantage of local information about the system. Surrogate models that map the input space to the QoI are very useful as a decision making tool since they can serve to help the analyst understand how errors in subsystem level model predictions impact the system level QoI. Considering these factors serves to reduce the computational expense required to perform large-scale simulations with only a marginal loss in accuracy, and the decision-making method itself represents a very small component of the total simulation expense. In addition, tracking which sample points lead to which model selection decisions may provide useful information to isolate physics-based deficiencies in low-fidelity models.

This work considers two basic situations: (1) the ranking of model fidelities is known for the entire domain because of expert opinion from model developers and (2) competing models (representing different physical hypotheses) may have a different (unknown) ranking of fidelity in different regions of the domain. These two scenarios are handled in two different ways; the second case is demonstrated by the illustrative example in Section 3.3 while the first is investigated in Section 3.5. In the second case, selecting models appropriately during the simulation is difficult if no information on the QoI is available, so whenever possible, independent data should be collected to validate the results and improve the decision-making tools.

Within this framework, model decisions can be fully automated and thus more easily applied to problems with highly sophisticated computational architectures. Further work is needed to integrate this approach with a dynamic computing resource allocation methodology, and with decisions about future model improvements and data collection. A complete orchestration of the UQ process for complicated problems with many component simulations will need algorithms to schedule the selected simulations and take advantage of parallelization in order to further reduce the computational effort while achieving the desired accuracy and precision.

The proposed methods of this chapter are important to a comprehensive framework for UQ. Efficient uncertainty propagation is important to both model validation and prediction activities since both aleatory and epistemic uncertainty are typically accounted for via MCS. Specifically, uncertainty propagation produces the probability distribution of a stochastic output of interest. Such a distribution is needed when performing quantitative model validation by comparing against observed data, and obtaining the distribution itself accurately is the singular goal of the prediction phase. The predicted distribution can then be used to perform reliability analysis and risk assessment by considering failure thresholds and failure consequences respectively. Each of these activities is discussed subsequently in this dissertation.

CHAPTER 4

CONNECTING MODEL VALIDATION TO PREDICTION

4.1 Introduction

This chapter demonstrates how models are validated in the presence of uncertainty that is propagated through computational models as described in Chapter 3. Model validation can be defined as the process of assessing the adequacy of a computational model for an intended prediction application. As described in Section 2.1, computational models are calibrated by updating parameter distributions to match the model output with observation data. However, a calibrated model should not be trusted for prediction without evidence that it is a good representation of reality in other input scenarios, both in terms of the inferred parameters and the underlying form of the model. This evidence should come from additional independent observations, preferably in a different input domain that is closer to the application of interest. The new experimental data that is used for model validation is inherently stochastic in the presence of measurement uncertainty. Since the model prediction is also stochastic once input and parameter uncertainty are propagated through it, quantitative model validation requires the comparison of probability distributions for prediction and observation.

As mentioned in Section 2.2, most recent quantitative validation methods are designed precisely for this purpose. Validation methods that have been developed in the literature include classical hypothesis testing [25, 34, 41], Bayesian hypothesis testing [73, 86, 87, 108], the area metric [22, 23, 95], and the model reliability metric [85, 97]. The connections between these various metrics as well as their strengths and weaknesses have also been explored [57, 59]. Each of these existing approaches assesses the agreement between model prediction and validation observation, but they differ in how they are applied. One view, as is usually taken with the area metric (see Section 2.2.1), is to look at the set of validation observations collectively and compare the distribution of the prediction over the entire input domain against the distribution of the observation data. When the input and corresponding output are measured for each validation experiment (with corresponding stochastic predictions for each input), a synthesis across the domain is accomplished via the “u-pooling" approach defined earlier in Eq. (2.6). An alternate view, as taken with the model reliability metric (see Section 2.2.2), is to perform a series of point comparisons, one for each validation input condition, and assess the predictive capability of the model as a function of the location in the input domain. Both classical and Bayesian hypothesis testing may be cast in a way that is consistent with either of these views by choosing different hypotheses. The proper interpretation depends largely on the type of data that is available to the analyst. This chapter investigates different validation scenarios where one of these two views (ensemble validation vs. point-by-point validation) is more suitable.

A further distinction between these methods is in the interpretation of the results. Conventionally, model validation has resulted in a single positive or negative result that indicates whether the model should be used in prediction or not. By choosing thresholds for the quantitative results, any of the previously mentioned methods could be interpreted in this manner. Alternatively, Bayesian hypothesis testing and the model reliability metric enable the result to be interpreted as a probability of agreement between prediction and observation. Thus, the result is not a single pass/fail decision, but a degree of validity. This dissertation focuses primarily on these probabilistic approaches because they enable other ongoing research efforts that are aimed at including the validation result in the subsequent prediction of a quantity of interest in the usage condition [40, 92, 98].

An important aspect of this discussion is the distinction between aleatory and epistemic uncertainty sources as introduced in Section 1.1. Aleatory uncertainty is unavoidable and must be accounted for in prediction models; however, it is not directly pertinent to decisions about risk and uncertainty reduction because its contribution cannot be eliminated. The primary focus of this dissertation is epistemic uncertainty since resource allocation decisions are aimed at reducing its contributions to the prediction. In the literature, epistemic uncertainty has been modeled in a number of different ways, including Bayesian probability [74], interval analysis [45], evidence theory [101], possibility theory [19], fuzzy logic [94], and generalized information theory [51]. Regardless of the approach to epistemic uncertainty characterization, researchers have become increasingly aware of the importance of separating aleatory and epistemic uncertainty sources [36, 50, 72]. Therefore, the focus of this chapter is the impact of epistemic uncertainty on model validation. The proposed methods demonstrate how to separate the contributions of aleatory and epistemic uncertainty when the available data permits.

Within this context, this chapter aims to address three issues that impact the validation assessment: (1) the type of input-output measurements that are made in validation experiments, (2) the “proximity" of the validation tests to the prediction regime of interest, and (3) the use of surrogate models for uncertainty propagation. The first issue is addressed in Section 4.2 where three different types of validation data scenarios are explored, and appropriate validation approaches are identified. The second issue is addressed in Section 4.3, where a method for weighting validation results by the relevance to the prediction is proposed. The third issue is addressed in Section 4.4, which quantifies the effect of surrogate model uncertainty on the validation result. The proposed methods are demonstrated with a numerical example of a microelectromechanical system (MEMS) device in Section 4.5.

4.2 Aleatory and epistemic uncertainty in model validation

In a probabilistic framework, both prediction and observation are treated as stochastic variables that are described by probability distributions. These distributions, which represent aleatory and/or epistemic uncertainty sources, may be compared by comparing the moments, the shapes, or the samples of the distributions. In the area metric and KL divergence [55] comparison approaches, the shapes of the distributions themselves are compared directly. In the model reliability metric approach, the distance between sampled realizations of prediction and observation is evaluated. Hypothesis testing methods (i.e. classical hypothesis testing and Bayesian hypothesis testing) may be cast in different ways by choosing different hypotheses (e.g. equality of moments or distribution parameters, equality of prediction and observation samples, or allowable distance between prediction and observation samples).

A key factor in the choice of comparison is the stochastic dependence between the prediction and observation. As noted in [22], samples cannot be uniquely generated without some knowledge of the dependence, so it is only possible to compare samples if the dependence information (i.e. the correlation structure) is known. In such a scenario, a comparison of sampled differences can make a stronger statement about the agreement between prediction and observation. For example, positive correlation between prediction and observation may suggest better predictive capability than negative correlation. This section discusses how the separation of uncertainty sources in point-by-point validation enables dependence information to be isolated, such that independent samples can be drawn. However, this separation may not always be possible, and when no such dependence information is known, a shape-based comparison can be performed in order to bypass this requirement. The result can then be bounded for different possible dependence structures [22].

The focus of this chapter is the area metric and the model reliability metric comparison approaches, which were previously described in detail in Section 2.2. The applicability of these approaches depends on the type of information that is available to the analyst.

4.2.1 Validation with fully, partially, or uncharacterized experimental data

Validation observations always include data on output quantities of interest, but the corresponding inputs are not always measured precisely (or at all). Three possible scenarios exist with respect to input measurements: (1) fully characterized (i.e., all the input variables of individual experiments and corresponding outputs are measured and reported as point values), (2) partially characterized (i.e., some inputs and/or outputs of individual experiments are not measured or are reported as intervals), or (3) uncharacterized (i.e., experiments are performed on multiple input combinations, but these input combinations are not measured or are reported as a single interval). In the cases of partially characterized or uncharacterized validation data, the input $\boldsymbol{X}$ is treated as a random vector due to the lack of measurements or the imprecision of the measurements. The reported intervals and expert opinion (if available) are needed to construct a probability distribution of $\boldsymbol{X}$. Note that in the Bayesian approach, the lack of knowledge (epistemic uncertainty) is represented through a probability distribution (subjective probability). This point is critical to the discussion that follows later in this section; the implication is that the “true" output of a single experiment is not a probability distribution, but a single value that cannot be precisely observed. Likewise, the corresponding model prediction would also be a single deterministic value for each experiment if all inputs and parameters were precisely known. The non-probabilistic approaches that were mentioned in Section 4.1 have also been proposed to handle the epistemic uncertainty; this dissertation focuses only on probabilistic methods.

Table 4.1: Input-output data for three different types of validation experiments

Fully Characterized	Input	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	…	$\boldsymbol{x_n}$
	Output	y_{d_1}	y_{d_2}	…	y_{d_n}
Partially Characterized	Input	$f_{X_1}(\boldsymbol{x})$	$f_{X_2}(\boldsymbol{x})$	…	$f_{X_n}(\boldsymbol{x})$
	Output	y_{d_1}	y_{d_2}	…	y_{d_n}
Uncharacterized	Input	$f_X(\boldsymbol{x})$			
	Output	y_{d_1}	y_{d_2}	…	y_{d_n}

For partially characterized validation data, input distributions are assigned to different experiments separately, and these distributions $f_{\boldsymbol{X}_i}(\boldsymbol{x})$ $(i = 1, \dots, n$ for n validation input conditions) represent input data uncertainty in each individual experiment. For example, suppose experiments were conducted at n different nominal load values, but each of the load values is only known up to an interval $[x_i - \epsilon, x_i + \epsilon]$. For uncharacterized validation data, a single distribution is assigned to the variable over multiple experiments, and this distribution $f_{\boldsymbol{X}}(\boldsymbol{x})$ represents the uncertainty due to both natural variability and input data uncertainty. For example, suppose the same n experimental outputs are available; however, there is not a nominal load value for each individual experiment, but rather a single interval that encompasses the load values for all experiments $[x_L, x_U]$. Table 4.1 shows a typical format of input-output data collected from the three types of experiments. Fully characterized data is preferred for the purpose of model validation; however, partially characterized and/or uncharacterized data may still be used when no fully characterized data is available.

4.2.2 Ensemble vs. point-by-point validation

As mentioned in Section 4.1, there are two possible views of validation. The data can be viewed collectively and compared against the overall distribution of the model prediction across the input domain, or the data can be viewed individually and compared against a separate stochastic prediction at each input condition. If the validation assessment is performed only once over the collection of data (i.e. ensemble validation), it is difficult to separate the contributions of aleatory and epistemic uncertainty sources to the validation result. Once the model prediction has been corrected for solution approximation errors and/or calibrated for bias (often referred to as model discrepancy [47]), the distributions of both the prediction and observation are a result of aleatory uncertainty (input variations) and epistemic uncertainty (parameter uncertainty in the prediction and measurement uncertainty in the observation).

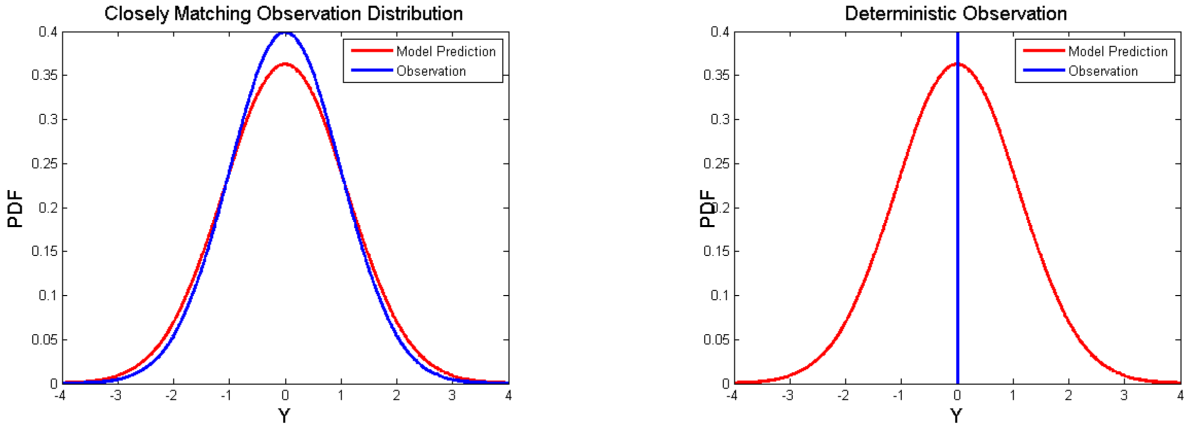
There is no reasonable expectation that the epistemic uncertainty contributions to the total uncertainty in the prediction and observation should be similar to each other because the two sources are independent. In particular, parameter uncertainty is related to the quantity and quality of available calibration data. As more calibration data is collected, parameter uncertainty can be reduced via Bayesian updating. Since the validation data set should be separate from the calibration data in order to make a proper assessment of the model's predictive capability, the measurement uncertainty in the validation data is generally different from the calibration measurement uncertainty. Furthermore, even if the distributions of the measurement errors in the calibration data and the validation data are similar, there is still no reason to expect correlation between particular samples of measurement error. Therefore, the only uncertainty contribution that is common to both the prediction and observation is the aleatory uncertainty in the input.

In the collective view of validation, one option for separating the aleatory and epistemic contributions is the p-box approach [95]. In this treatment, epistemic uncertainty is expressed as an interval while aleatory uncertainty is expressed with probability distributions. Such a treatment is particularly suitable for uncharacterized data because the data quality does not enable point-by-point separation. However, when the dominant effect is epistemic uncertainty, rather than aleatory uncertainty, comparing observations to a p-box may not be very informative since the epistemic uncertainty gives a wide window of acceptance for the model [22, 95].

In many problems, the epistemic contributions are, in fact, large since economic constraints in realistic applications often lead to very sparse/imprecise data. For this reason, the model reliability approach is aimed at epistemic uncertainty in both the observation and the prediction. Note again that parameter uncertainty in this dissertation refers to the subjective probability description of a deterministic parameter value, not aleatory uncertainty. It is possible that parameters may also be affected by aleatory variability across experiments, but this issue is addressed in this chapter by localizing calibration to particular experimental configurations. The parameter uncertainty is expressed by a subjective probability distribution separately for each test, and it is then reduced via Bayesian updating with replicate testing as seen in the example in Section 4.5. Aiming the assessment at epistemic uncertainty leads directly to decisions about what improvements are most necessary (either in the data or the model) in order to improve the predictive quality of the model.

Therefore, when information is available about the particular input condition associated with each data point (either fully or partially characterized data), the use of individual comparisons at each location with the model reliability metric is proposed. The metric is computed for a stochastic prediction and an uncertain observation, but the metric is not maximized when the spreads in the two distributions are the same. This behavior occurs because the metric is not a shape-based comparison; it comes from sampling the distributions to compute the distribution of the difference Δ (see Section 2.2.2).

As mentioned in the opening of this section, the distribution of Δ can only be obtained if the stochastic dependence between prediction and observation is known. However, the correlation between these two variables only occurs through aleatory uncertainty that is common to both, and the epistemic uncertainty sources are independent. Therefore, at a particular input condition, since the stochastic prediction and observation are only sampled over epistemic uncertainty sources, the samples are *conditionally independent*. Since Δ is simply the distribution of bias between deterministic samples of prediction and observation, the maximum reliability metric occurs when the distributions of prediction and observation are unbiased from each other, and each has minimum uncertainty (see Figure 4.1). This behavior agrees with our intuition about how to improve the result if the validation agreement is poor. By reducing measurement uncertainty *or* reducing parameter uncertainty, the validation result at each input can be improved. For the shapes of the distributions to agree, both the measurement uncertainty *and* the parameter uncertainty must be reduced in order to improve the agreement. It is an unnecessary requirement that the shapes agree since they are representing only independent epistemic uncertainty sources. Both collecting more calibration data (to reduce parameter uncertainty) and collecting more precise data (to reduce data uncertainty) should individually improve confidence in the model if the model is actually predicting well.



(a) Measurement uncertainty and parameter uncertainty are similar (Model reliability = 0.86)

(b) Zero measurement uncertainty with the same parameter uncertainty as in 4.1(a) (Model reliability = 0.95)

Figure 4.1: Decreasing measurement uncertainty for the same stochastic prediction improves the confidence in the model if the observation is unbiased

For these reasons, shape-based comparisons are not intended for purely epistemic uncertainty-based comparisons. They should not be used for this purpose because it is possible for the contributions of one or both of the uncertainty sources to increase and improve the comparison. For example, the point comparison shown in Figure 4.1 poses two scenarios, both with the same stochastic prediction. Figure 4.1(b) gives an idealized scenario where the observation data is “perfect" (i.e. no measurement uncertainty). In this scenario, clearly the shapes of the two distributions are not the same, and the distributions will actually match more closely (improving a shape-based measure) by injecting more uncertainty into the observation as in Figure 4.1(a). This result does not occur with the model reliability approach because the metric is lower for larger uncertainty in the observation (i.e. there is less confidence in the assessment because the observation data is not adequate). Since, at a single known input point (fully or partially characterized), the uncertainty sources are completely epistemic, both the prediction and observation would be deterministic values if no epistemic uncertainty existed.

Thus, the scenario shown in Figure 4.1(b) (where ideal quality observation data is available) is actually preferable because there is a higher probability that the deterministic prediction and observation would agree if both were known precisely.

An additional advantage of point-by-point comparison is that it demonstrates the quality of the model as a function of input condition. This information may be very useful in determining whether the model will be appropriate in its intended use, and it may also help isolate potential systematic errors arising from model form inadequacy. For example, if the model is consistently performing poorly for large values of some input (e.g. loading), this may be evidence that the model does not capture some higher order physical behavior (e.g. nonlinearity) that is activated by extreme conditions. Additionally, if different values of model reliability are computed at different inputs, the weighting approach that is presented in Section 4.3 becomes possible, and preferences for particularly important regions of the input domain based on the intended application can be incorporated.

In summary, this section concludes that ensemble validation is best suited for uncharacterized data scenarios, and point-by-point validation is preferable when information is known about the corresponding input conditions (partially or fully characterized validation data scenarios) for the following reasons: 1) distributions of prediction and observation can only be expected to agree when the dominant uncertainty source is aleatory variability that is common to both distributions; 2) point-by-point comparisons with the model reliability metric separate aleatory and epistemic uncertainty and penalize large epistemic uncertainty (from any source) by returning a lower validation result; 3) point-by-point comparisons allow systematic error trends to be isolated in the model; and 4) a set of point-by-point comparisons can be weighted based on relevance to the intended use of the model.

4.3 Integration of model validation results from multiple input conditions

By utilizing the model reliability approach, a value for the validation metric can be computed for each validation input condition. This information is itself useful for decision making about the model adequacy since developers can look at regions of the input domain that perform poorly in validation and investigate potential model improvements. However, the ultimate goal of the validation activity is to assess the current model's prediction capability, and recent research [40, 92, 98] has the additional goal of performing this assessment quantitatively so that it may be included in the prediction. In some applications, including the validation result in a prediction framework may require a single overall measure of the model quality across the entire domain of interest. This measure should be representative of the quality of the model in its intended application condition where the prediction will be made. Thus, the method given by Eq. (4.1) is proposed.

$$v_{overall} = \int v(\boldsymbol{x})\pi(\boldsymbol{x})d\boldsymbol{x} \tag{4.1}$$

Here, $v(\boldsymbol{x})$ is the value of the validation metric at a particular point in the validation domain, represented by the n-dimensional input vector $\boldsymbol{x}$ and $\pi(\boldsymbol{x})$ is the n-dimensional joint probability density of the point $\boldsymbol{x}$ in the prediction domain. This distribution comes from the best available knowledge of the input conditions that will be encountered in the intended application of the model; the distribution may describe both aleatory and epistemic uncertainty. Effectively, the joint density becomes a weighting function for the importance of each validation result according to how likely that input condition is in the prediction scenario. In evaluating the integral in Eq.

(4.1), note that $v(\boldsymbol{x})$ is only available at some discrete values of $\boldsymbol{x}$. Therefore, the integral may be approximated by a weighted sum taken over a set of m validation tests as

$$v_{overall} = \frac{\sum_{i=1}^{m} v(\boldsymbol{x_i})w_i}{\sum_{i=1}^{m} w_i} \tag{4.2}$$

The computation of the weight w_i is straightforward for fully characterized validation data; it is obtained by computing $\pi(\boldsymbol{x_i})$, which is a single value for each validation experiment. When input measurement uncertainty exists, the validation data is considered to be partially characterized, and $\boldsymbol{X_i}$ is not a point value, but rather a random variable. In this scenario, the weighting for the intended application can be obtained for each validation test by taking the expected value over the distribution of the corresponding input measurement uncertainty $f_{\boldsymbol{X_i}}(\boldsymbol{x})$.

$$w_i = \int \pi(\boldsymbol{x}) f_{\boldsymbol{X_i}}(\boldsymbol{x}) d\boldsymbol{x} \tag{4.3}$$

Once all the weights are computed, Eq. (4.2) results in a single deterministic measure of the probabilistic performance of the model over the expected prediction domain. It is an approximation since the set of validation input conditions generally does not cover the full prediction domain of interest; therefore, the summation must be normalized in order to obtain a valid probability. In fact, in some cases the prediction scenario may be for values of $\boldsymbol{x}$ that are not close to the validation domain. In this situation, the validation input conditions fall in the tail of the distribution for the intended application, and $\pi(\boldsymbol{x_i})$ is small for all the validation points. This would imply that none of the validation experiments are in the regime that is most relevant to the intended application, and the prediction represents a significant extrapolation of the model. Such extrapolation scenarios can be dangerous applications of the model, but they are often unavoidable in practice. Additional conservatism is needed for this situation, and the analyst should be especially aware of any trends in the point-by-point validation results that suggest that model inadequacy will be magnified in the prediction regime. Ongoing research efforts are exploring quantitative methods of setting boundaries for the extrapolation of the model and applying additional conservatism to the extrapolation scenarios when they are practically necessary [40, 44, 92, 95, 98].

The proposed integration approach has been described for situations where a single probabilistic value can be obtained from the model reliability metric at each input condition. When additional epistemic uncertainties exist, the validation metric uncertainty can be described by a probability distribution at each validation input, and the overall metric will also be a probability distribution accounting for these additional uncertainties. One example is a stochastic model discrepancy term as in the Kennedy-O'Hagan approach to model calibration [47]. If the discrepancy term is used as a correction for the model prediction, different realizations of the stochastic discrepancy yield different validation results. As another example, when surrogate models are used to generate the distribution of the model output that is used in the validation assessment, different realizations of the surrogate model prediction also lead to different validation results. A final example is model validation in the presence of sparse data, leading to uncertainty about the distribution of Y_d. These additional uncertainty sources should also be accounted for. Thus, the surrogate model scenario is explored in Section 4.4, and the sparse validation data scenario is explored later in Section 5.2. Though it is not explored in this dissertation, note that the mathematics of treating stochastic model discrepancy would follow similarly to the other two examples that are demonstrated.

4.4 Inclusion of surrogate model uncertainty in validation

Probabilistic approaches to model validation, as described in Section 2.2, require the propagation of parameter uncertainty through the model at each validation input condition. This propagation is typically performed via Monte Carlo sampling, which requires a large number of model evaluations. When the computational model is expensive, it is often replaced by a surrogate model to improve the efficiency of the propagation. Ultimately, the goal of the validation assessment is to make a statement about the adequacy of the physics-based, original computational model and not the surrogate, since the former will be used for prediction. Since the surrogate model is not a perfect representation of the original computational model, additional uncertainty is added to the validation result. In this dissertation, GP surrogate models as described in Section 2.4 are used for this purpose because they provide a natural way of quantifying the uncertainty due to the discrepancy between the surrogate and the original computational model. This uncertainty then propagates to uncertainty in the validation assessment.

When GP surrogate models are available, they can be used for affordable uncertainty propagation. The issue with this approach is that it creates an additional source of uncertainty in validation. The validation result must apply to the physics-based computational model (not the surrogate model) since it will be used in the prediction domain. To make this assessment, the additional uncertainty stemming from the uncertain fit of the surrogate to the computational model must be accounted for. Using a GP model, denoted here as $\hat{f}$, to replace the underlying physics model as a function of input $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$ provides a Gaussian distribution at a prediction point arising from surrogate uncertainty as $Y_m = \hat{f}(\boldsymbol{x}, \boldsymbol{\theta}) \sim N(\mu_{Y_m}, \sigma_{Y_m})$. This represents a family of distributions for different values of $\boldsymbol{\theta}$. This family of distributions may be collapsed by employing the auxiliary variable approach [99] in which the dependence on the distribution parameters μ_{Y_m} and σ_{Y_m} can be mapped to a dependence on only the CDF value of the distribution u as in Eq. (4.4).

$$u = F_{Y_m}\left(y_m \middle| \mu_{Y_m}, \sigma_{Y_m}\right) = \int_{-\infty}^{y_m} f_{Y_m}\left(\omega \middle| \mu_{Y_m}, \sigma_{Y_m}\right) d\omega \tag{4.4}$$

Since this auxiliary variable represents a CDF value, $U \sim Uniform[0,1]$, the model reliability R becomes a random variable itself and can be written as a function of the random variables $\boldsymbol{X}$, U, and $\boldsymbol{\Theta}$. As shown in Eq. (4.5), the model reliability metric at input $\boldsymbol{x}$ can be computed for any u by integrating over the distribution of $\boldsymbol{\Theta}$ as in Eq. (2.8). Then, the model reliability at any $\boldsymbol{x}$ is weighted by the pdf of $\boldsymbol{x}$ in the prediction domain (as in Eq. 4.1) to obtain the overall distribution of the metric as a function of the surrogate model uncertainty as shown in Eq. (4.6).

$$r(\boldsymbol{x}, u) = \int_{|y_m - y_d| < \epsilon} [y_m(\boldsymbol{x}, \boldsymbol{\theta}, u) - y_d(\boldsymbol{x})] f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{4.5}$$

$$r_{overall}(u) = \int r(\boldsymbol{x}, u) \pi(\boldsymbol{x}) d\boldsymbol{x} \tag{4.6}$$

The resulting distribution of the validation metric can be computed by sampling the auxiliary variable to demonstrate the contribution of the GP uncertainty to the validation result. The spread in this distribution is the cost of using the surrogate model for propagation. This uncertainty may be reducible by improving the surrogate model by adding additional training points.

The proposed approach formalizes the validation assessment when using surrogate models for uncertainty propagation. When possible, it is preferable to use the original computational model directly, but constraints on computational effort often make such an approach unaffordable. When surrogates are necessary, the additional uncertainty can be included via the method described above. An alternative approach is to apply the uncertainty propagation approach proposed in Chapter 3 to select between the GP and the original computational model across the domain. The need to use the original computational model in some portions of the domain will then depend on the quality of the GP surrogate, which is dependent upon the amount of training data as well as the smoothness of the original computational model's response.

4.5 Numerical example

4.5.1 Validation of MEMS device simulation

To demonstrate the proposed validation methodology, a microelectromechanical system (MEMS) example is introduced. The radio frequency (RF) MEMS switch, shown in the conceptual diagram in Figure 4.2, is subjected to electrostatic loading that causes the membrane to deform. The mechanical properties of the membrane resist the deformation, but at some voltage, known as the pull-in voltage, the electrostatic force pulls the membrane into contact with the substrate. At a voltage level known as the pull-out voltage, the membrane can then be released from contact with the substrate. The pull-in and pull-out voltages are predicted by device simulation, and they are also measured in validation experiments (20 replicate tests on each of six devices).

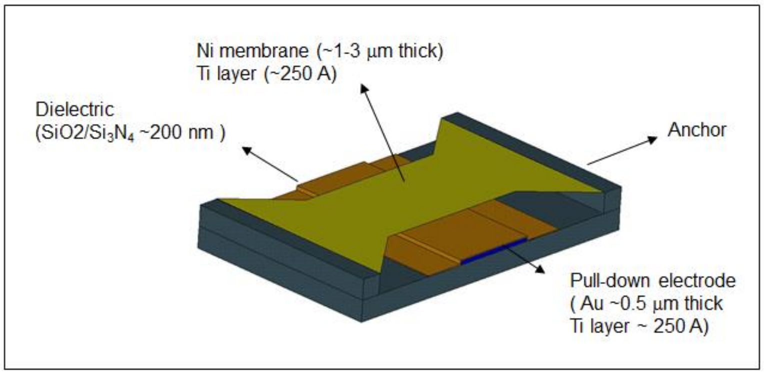
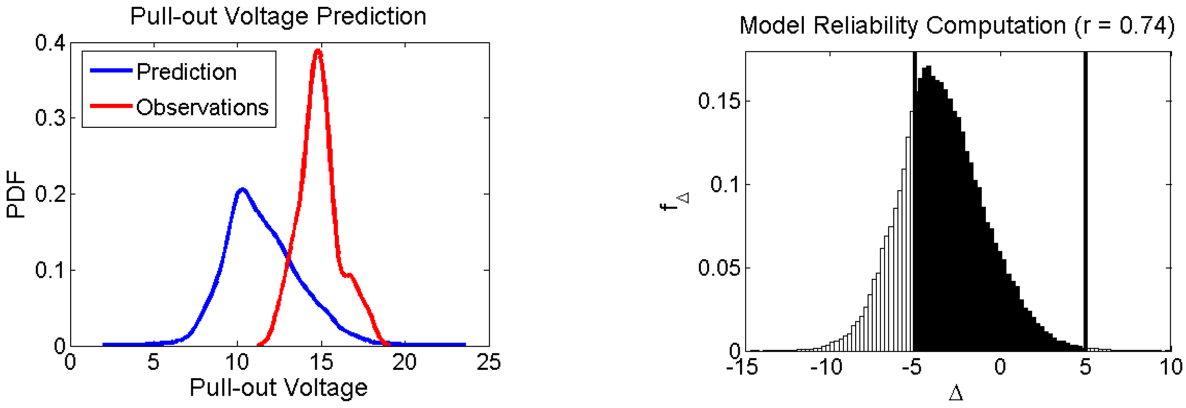


Figure 4.2: RF MEMS switch

Five variables, membrane thickness h, gap between one end of the membrane and the substrate g_1, gap between the other end of the membrane and the substrate g_2, Young's modulus E, and contact height d_c are identified as inputs to the model and experiments. Due to the imprecision of the measurement techniques, the geometry parameters g_1 and g_2 are described by distributions that represent input measurment uncertainty for each of the six devices. Direct measurements of E and d_c are not available, but the ranges of these two variables are obtained via multi-scale simulation [48, 53]. The thickness of the membrane h cannot be measured accurately, so it is treated as a calibration parameter. Using the pull-in voltage measurements, the membrane thickness is estimated separately for each device via Bayesian inference. Then, the predictive simulation is validated using the pull-out voltage measurements. The measurement for each device corresponds to a combination of the input set $[h, g_1, g_2, E, d_c]$, each with associated uncertainty. Thus, the validation measurements are partially characterized.

In a partially characterized data scenario, input measurement uncertainty can be treated in the same manner as parameter uncertainty when performing the validation assessment. For a single device, each of these inputs has a single value in reality, but it cannot be measured precisely. Aleatory uncertainty is only present in the form of device-to-device variation. Therefore, the source of the uncertainty in the prediction for a particular device (i.e. a particular input condition) is completely epistemic. The uncertainty in the observation is attributed to output measurement uncertainty, which is also epistemic. Therefore, a point-by-point comparison for each device using the model reliability metric can be performed.



(a) Comparison of model prediction and observation with associated epistemic uncertainty.

(b) Computation of model reliability using the difference between prediction and observation. For ϵ = 5, the difference Δ is integrated over the interval (-5, 5) as shown

Figure 4.3: Computation of model reliability for partially characterized validation data

For example, Figure 4.3 demonstrates the model reliability metric computation for one of the six devices. The tolerance ϵ is set to 5 volts, and the distribution of the difference between prediction and observation Δ is integrated over the interval (-5, 5) to obtain a model reliability of 0.74. The prediction distribution shown in Figure 4.3(a) is generated by propagating input measurement uncertainty through the prediction model. Since the computational model that predicts the pull-out voltage is expensive (approximately 6 hours per evaluation) and a large number of Monte Carlo samples of the input measurement uncertainty are needed in order to converge the output distribution (10,000 were used in this illustration), using the computational model for propagation is unaffordable. Therefore, GP surrogate models are constructed to improve the efficiency of the computation. For illustration, the surrogate uncertainty is not included in the result shown in Figure 4.3; only the mean prediction from the GP model is used. If the computational model were not expensive, the uncertainty propagation could be performed without constructing a surrogate model, and the computation of the model reliability would proceed exactly as shown, resulting in a single value of the model reliability for each device. However, as mentioned, a surrogate model is needed for this example, and this uncertainty must also be included in the assessment. As a result, the model reliability is instead described by a distribution for each device. This consideration is demonstrated in Section 4.5.2.

4.5.2 Inclusion of surrogate uncertainty

The framework in Section 4.4 is applied to the validation assessment for each of the six devices. For each device, the model prediction is made for a set of samples of the input uncertainty. By sampling the auxiliary variable, many realizations of the GP model are taken; each of these is a candidate prediction of the underlying computational model. The set of realizations produces a family of predictions that represents the possible outcomes for the validation assessment that could be obtained if the computational model were used directly. Note that these realizations are obtained by sampling the auxiliary variable and using the covariance function of the GP model, so the outputs at different samples of the input uncertainty are highly correlated. This correlation may result in a family of predictions with greater uncertainty than the standard deviation at a single prediction point would indicate. For each candidate model prediction, Eq. (4.5) is applied to obtain a value for the model reliability metric. This set of values for the model reliability is used to construct a histogram for the validation result for each device. The histograms are normalized to obtain the frequency diagrams shown in Figure 4.4.

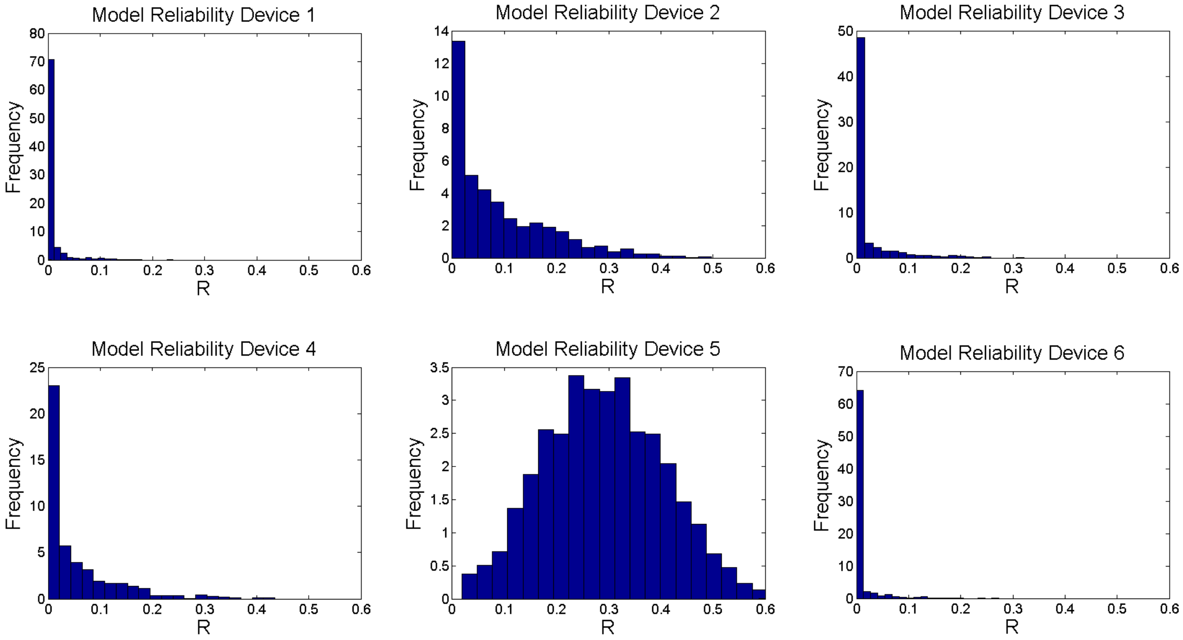


Figure 4.4: Frequency diagrams of model reliability for each of 6 devices

For several of the devices, the mean model reliability is very low because the mean prediction and mean observation were substantially biased from each other. This result may occur due to inadequacies in the model and/or inconsistencies in the observed data. As described in Section 4.2, both input and output measurement uncertainty may also contribute to the poor performance of the model (input measurement uncertainty increases the spread in the prediction while output measurement uncertainty increases the spread in the observation). Additionally, the spread in the potential outcomes of the model reliability indicates that the GP uncertainty is significant. By obtaining more training data, this particular source of epistemic uncertainty can be reduced, and the model reliability would be expected to converge toward the single value that would be obtained by performing the propagation with the computational model directly.

For most applications, the validation results shown in Figure 4.4 would not provide sufficient confidence to use the model going forward in prediction. Either the model form should be improved, or the quality of the observation data should be thoroughly evaluated, and if necessary additional validation data should be collected. However, for illustration, the approach for integrating these results from different devices into a single result is demonstrated in Section 4.5.3 below.

4.5.3 Integration of validation results from multiple devices

Once individual validation results have been obtained for several different devices, it is useful to determine which of the results is most relevant to the prediction of interest. For example, if the beam thickness h is an input of particular interest, it is helpful to assess to the predictive capability of the model as a function of what thickness will be encountered. The validation tests that were conducted for thicknesses similar to those in the prediction scenario are most relevant. The calibrated thickness distributions for each of the six devices are shown in Figure 4.5.

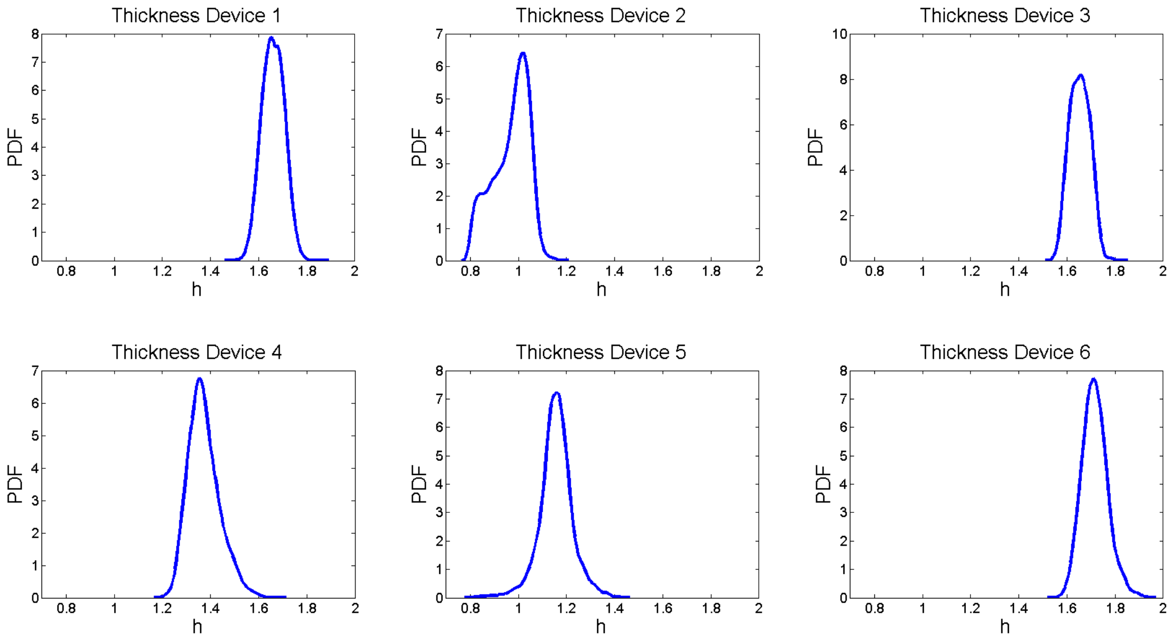
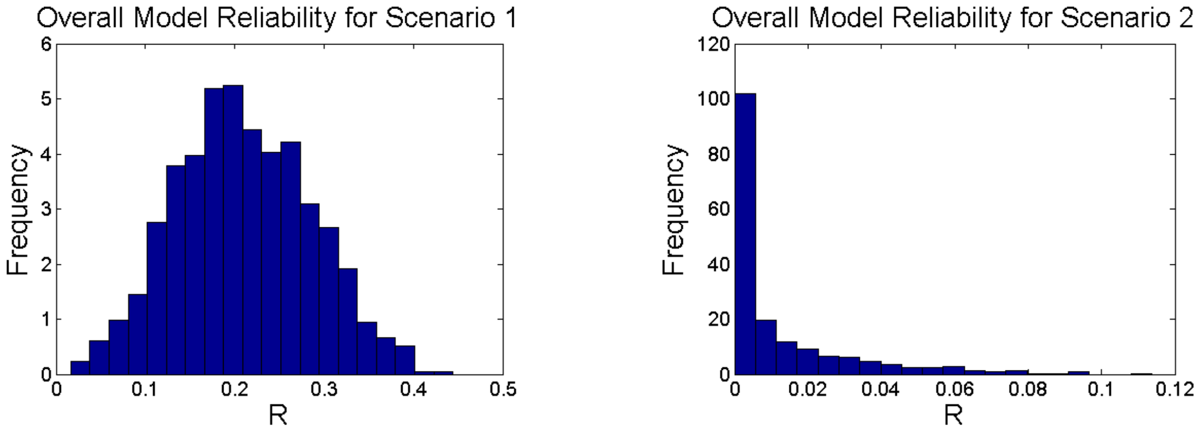


Figure 4.5: Input uncertainty for the thickness of the 6 devices

Table 4.2: Weights for two different prediction scenarios

	Device 1	Device 2	Device 3	Device 4	Device 5	Device 6
$\pi(\boldsymbol{x}) \sim N(1.2, 0.1)$	1.18e-4	0.114	1.11e-4	0.244	0.642	1.57e-5
$\pi(\boldsymbol{x}) \sim N(1.7, 0.1)$	0.328	4.08e-10	0.323	9.56e-3	2.71e-5	0.339

Suppose the model will be used for two different prediction scenarios in which the thicknesses will be $N(1.2, 0.1)$ and $N(1.7, 0.1)$ respectively. By applying Eq. (4.3) and normalizing the weights, the weights for the 6 devices are shown for the two scenarios in Table 4.2. It is clear from this table that device 5 is most relevant to the first prediction scenario while device 4 and device 2 are also somewhat relevant, and the other three devices are not. The second scenario has three device tests that are of nearly equal relevance (devices 1, 3, and 5), and the other three have negligible weight. By using these weights in Eq. (4.2), the integration in Eq. (4.6) can be approximated to produce the distributions for $R_{overall}$ shown in Figure 4.6.



(a) Distribution of $R_{overall}$ for a prediction scenario with $\pi(\boldsymbol{x}) \sim N(1.2, 0.1)$

(b) Distribution of $R_{overall}$ for a prediction scenario with $\pi(\boldsymbol{x}) \sim N(1.7, 0.1)$

Figure 4.6: The model is expected to perform much better for the first scenario since device 5 is the most relevant and also the best performer in the validation assessment.

This validation assessment has shown that there is low confidence in the model in general, but this comparison shows that the model is much more adequate for the first prediction scenario than the second. Since only device 5 gave reasonable prediction quality in the validation assessment, it is reasonable to conclude that if the model is used in prediction at all, it should only be for input scenarios that are similar to the measured inputs of device 5. Therefore, the predictive capability of the model is very limited, which again emphasizes the need to improve both the model and the observation data.

4.6 Conclusion

This chapter presents a model validation methodology for handling different data scenarios. When validation data is uncharacterized (corresponding inputs are not measured for each experiment), an ensemble validation approach is suitable. However, when inputs are also measured in validation tests (either fully or partially characterized data), it is preferable to perform validation individually for each input scenario. This enables aleatory and epistemic uncertainty sources to be separated from one another, which aids in decision making for uncertainty reduction when the model performance is inadequate. Additionally, understanding the reliability of the model as a function of the input may help to identify systematic inadequacies in model form. The individual metric values can be integrated into a single metric by weighting each value with the probability of observing the corresponding input in the prediction domain (i.e. relevance to the intended application of the model). When the computational expense of the model causes uncertainty propagation to be intractable, surrogate models are needed to obtain the distribution of the model prediction. This approach adds additional uncertainty into the assessment that should also be included in the analysis. With a GP surrogate model, the surrogate uncertainty can be readily obtained from the covariance structure of the model, and this uncertainty results in the model reliability metric itself being treated as a random variable with epistemic uncertainty. Once the model reliability metric is obtained (either a single value or a distribution), the metric can be interpreted probabilistically; this allows the validation result to be incorporated into the prediction.

The model validation methodology proposed in this chapter provides the framework for connecting the validation activity to the prediction of interest. The weighting approach demonstrates that there may be large differences in the importance of the various validation experiments for different prediction scenarios. This knowledge is fundamental to the test selection methodology that is proposed in Chapter 5, and it emphasizes the importance of understanding the intended use of the model when performing validation.

CHAPTER 5

TEST SELECTION FOR PREDICTION UNCERTAINTY REDUCTION

5.1 Introduction

It is often possible to collect many different types of data for both model calibration and model validation. Options may include material, component, and subsystem tests, and it may also be possible to conduct some or all of these types of tests at a variety of different input conditions. When many different types of data are available, all of the information must be integrated toward the prediction goal. The previous chapters of this dissertation have shown how information is integrated toward prediction UQ by performing model calibration, model validation, and uncertainty propagation. Then, in Chapter 4, a method for explicitly connecting model validation to the prediction was demonstrated. The proposed UQ framework is used as a foundation for making test selection decisions among many possible options.

Within this context, the experimental data that is collected for calibration and validation is critical to the prediction quality, but not every piece of information has the same impact. For example, in many applications the data that most closely replicates the system usage conditions is the most valuable, but also the most expensive. Furthermore, even if the important types of tests and associated input conditions can be identified, it is typically not sufficient to perform only a single experiment for each test scenario; instead, replicates are needed, due to data uncertainty. The number of replicates that are needed may vary across different test scenarios depending on the relative magnitudes of the sources of uncertainty that are present. Therefore, the goal of the test selection problem posed in this chapter is to determine the number of replicate tests that should be conducted at a discrete set of candidate testing scenarios. Since the decision of what data to collect is closely tied to budget constraints, a constrained optimization approach for addressing the cost vs. value tradeoff is proposed. To formulate the approach, the value of each test is quantified in terms of prediction uncertainty reduction.

Test selection decisions are focused on two different categories of experiments (calibration and validation) and their impact on the prediction. Model calibration is performed via Bayesian methods (see Section 2.1), and model validation is performed with the model reliability metric (see Section 2.2.2). The collection of calibration data is motivated by parameter uncertainty reduction while the collection of validation data is motivated by data uncertainty reduction. Each of these individual epistemic uncertainty reductions results in overall uncertainty reduction for the prediction quantity of interest.

The problem of which tests to perform has been addressed in the literature in terms of information theory [60] and decision theory [10, 68]. Design of experiment has been explored for both classical [9, 20, 79] and Bayesian formulations [12, 79, 104, 106]. Many of these approaches use Kullback-Leibler divergence [55] to compare the support for the various modeling options and make selection decisions, but these decisions are typically made from the perspective of the prior or posterior parameter distributions. Parameter uncertainty alone is not a sufficient indicator of the resulting prediction uncertainty since the sensitivity of the prediction quantity to the parameters must also be considered. Therefore, the proposed formulation instead addresses the selection decision from the perspective of the prediction for the usage condition. In addition, this chapter extends these methods which have primarily focused on calibration tests to a joint formulation for both calibration and validation tests.

In summary, this chapter proposes a test selection methodology that includes the following key features: (1) integration of experimental data from multiple input conditions for both calibration and validation toward prediction, (2) treatment of data uncertainty (sparse/imprecise data), and (3) a joint optimization formulation for prediction uncertainty reduction that accounts for both calibration and validation activities. The framework for the integration of sparse experimental data toward prediction builds upon the UQ framework developed in previous chapters. Some additional considerations that are particularly relevant to test selection are detailed in Section 5.2. Then, Section 5.3 demonstrates how this framework is used to formulate an optimization problem for test selection. In Section 5.4, the methodology is demonstrated for the MEMS numerical example that was introduced in Chapter 4, and the chapter is concluded in Section 5.5.

5.2 Prediction uncertainty quantification

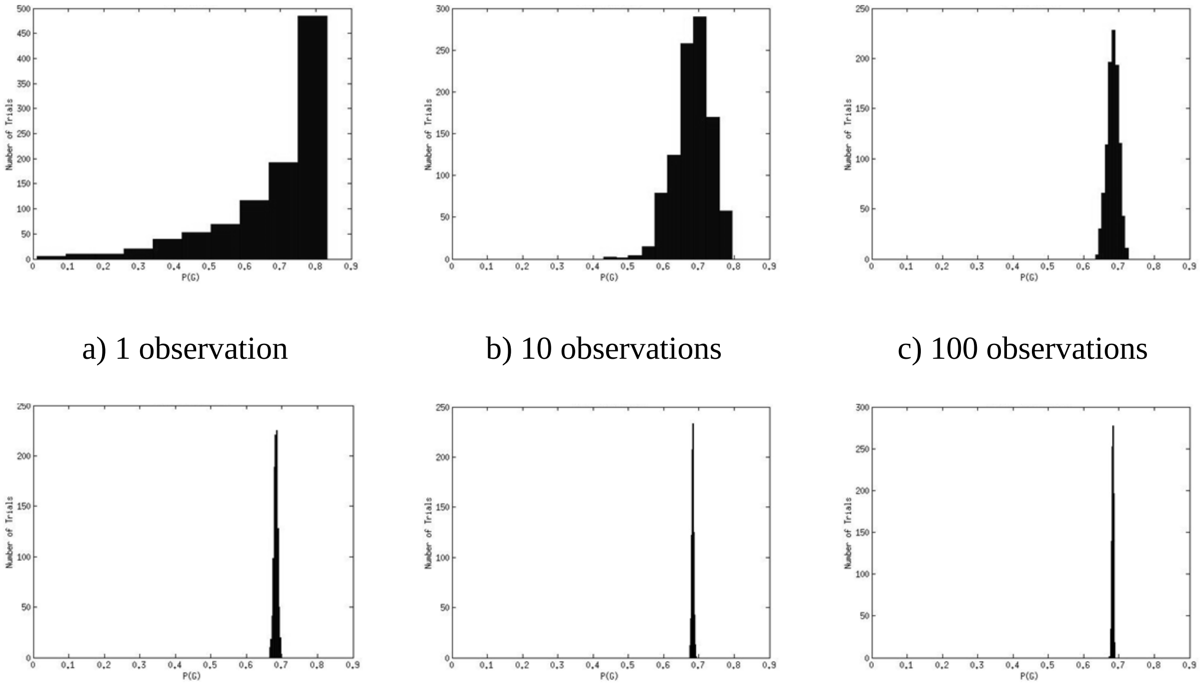
The goal of the prediction methodology is to obtain the distribution of a stochastic output of interest by incorporating both aleatory and epistemic uncertainty sources. Of the three types of experiments described in 4.2.1, only fully characterized and partially characterized tests are considered. Since tests are being selected and have not yet been performed, uncharacterized experiments should be avoided. By restricting to these two types of tests, particular values $\boldsymbol{x_i}$ or test-dependent aleatory distributions $\boldsymbol{X_i}$ for the inputs are known for each test, and the aleatory uncertainty in $\boldsymbol{X}$ across different experiments does not affect calibration or validation at a particular input condition. Therefore, the important uncertain inputs for prediction UQ are the components of $\boldsymbol{\Theta}$ that are common to the calibration and validation experiments and the prediction. Thus, the result of calibration and validation activities is carried through $\boldsymbol{\Theta}$. The proposed methodology consists of three key activities: (1) calibrate $\mathbf{\Theta}$ from the calibration observations , (2) validate the model and the inferred distribution of $\mathbf{\Theta}$ with additional independent observations, and (3) modify the distribution of $\mathbf{\Theta}$ to incorporate the validation result.

5.2.1 Model validation in the presence of data uncertainty

General methods for model calibration (Section 2.1) and model validation (Section 2.2 and Chapter 4) have previously been described in this dissertation. The fundamentals of these methods are not repeated here. Instead, these methods (specifically model validation methods) are expanded to account for the existence of data uncertainty when performing the assessment. Data uncertainty leads to a stochastic validation assessment taken over replicates at each input condition. The separate assessments are then combined with the integration approach across input conditions that was proposed in Section 4.3. The combined treatment is similar to the approach proposed in Section 4.4.

5.2.1.1 Validation uncertainty for sparse observation data

Validation observations may be made at different input conditions, but there should also be replicates at each input condition. These replicates are necessary because there is always measurement error in any experimental observation (e.g. zero mean Gaussian white noise). For a finite number of observations, the distribution of Y_d is approximated empirically. One approach is to construct a discrete probability mass function with equal weights attributed to each observation and then evaluate the model reliability with discrete sampling. With this approach, it will be shown that the expectation of the computed reliability is not sensitive to the number of validation points that are collected. Rather, the impact of the sparseness of the validation data is the uncertainty about the model reliability assessment. To observe this effect, trial computations of the model reliability are conducted for various lengths of the observation vector $\boldsymbol{y_D}$. Assuming that the observations are coming from an unknown true value, polluted by Gaussian white noise, the model reliability is computed for 1,000 trials of six different lengths of the observation vector (1, 10, 100, 1000, 10000, and 100000) to demonstrate the uncertainty in the assessment (see Figure 5.1). In all cases, the mean model reliability taken over the 1,000 trials is equal, but it is clear that the uncertainty in the computation due to the noise in the data is much more severe for sparser sample sets.



a) 1 observation

b) 10 observations

c) 100 observations

d) 1,000 observations

e) 10,000 observations

f) 100,000 observations

Figure 5.1: Uncertainty in model reliability computation for sparse validation data sets

In order for test selection decisions to properly account for the importance of replicate validation tests, the validation assessment must account for the uncertainty arising from data sparseness. If only a single deterministic computation is performed, the computed value may significantly underestimate or overestimate the actual model reliability. In reality, the model reliability is a single deterministic value (i.e. the converged result from many observations as in Figure 5.1f), but this value can only be obtained confidently with a large number of replicate tests. Furthermore, since the expected value of the computation is the same for all sample sizes, a deterministic calculation provides no evidence that the computed result may actually be biased from the true value. Therefore, a stochastic assessment that directly incorporates this uncertainty is needed.

5.2.1.2 Stochastic assessment of model reliability

Since it is obviously not possible to conduct many trials of a fixed number of validation tests (as in the numerical demonstration of Figure 5.1) in realistic problems, a stochastic assessment approach that gives an estimate of the uncertainty in the computation for only a single set of observations is desirable. Note that only the converged deterministic value of the model reliability is of interest for prediction purposes. If there were no measurement error, only a single observation would be needed to obtain this value. In the presence of measurement error, the mean observation corresponds to the desired reliability value as long as the measurement error has zero mean. Therefore, the collection of replicate data can be viewed as a way to estimate the mean observation (i.e. the true observation that is not polluted by noise) accurately.

This goal motivates the use of Student’s t-distribution [105] to describe the mean observation in the presence of sparse data. By definition, the t-distribution is used to describe the uncertainty in the mean of an underlying normally distributed population. Since the measurement error is typically assumed to be zero-mean and normally distributed, the observation mean can often be described by a t-distribution exactly. In order to construct the t-distribution, three pieces of information are needed: 1) sample mean, 2) sample variance, and 3) degrees of freedom. From the set of validation observations, a sample mean $\bar{x}$ and sample standard deviation s can be computed as long as there are at least two observations at a particular input condition. The number of degrees of freedom ν is simply the number of observations minus one. Given these pieces of information, the t-value for the unknown mean μ is obtained as

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \tag{5.1}$$

and the probability density function is given by

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{5.2}$$

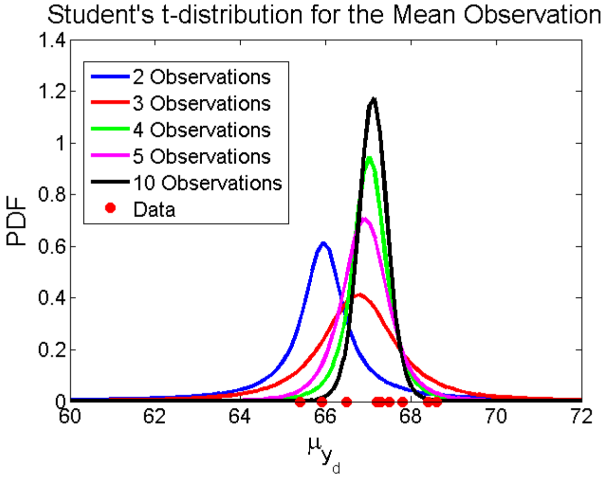


Figure 5.2: Student’s t-distribution of the mean observation for sparse observation sets

The resulting distribution has diminishing uncertainty as the number of validation samples increases. This behavior is demonstrated for different numbers of observations in Figure 5.2. From this distribution, possible values of the mean observation can be sampled, and the model reliability can be computed for each candidate sample. Taken together, these computed reliability values represent the uncertainty in the validation assessment that results from sparse validation observations. The distribution of the model reliability R may have any form depending on the shape of the distribution of Y_m. This t-distribution approach is applied to each validation input condition where replicate samples are available. By sampling the mean observation at each input condition a distribution of the model reliability R_i is obtained for each $\boldsymbol{x_i}$ or $\boldsymbol{X_i}$. Collecting additional replicates at a particular input condition reduces the variance of the t-distribution, which in turn reduces the uncertainty in R_i. The distributions for different input conditions can be combined into an overall distribution by applying the approach in Section 5.2.1.3. This overall distribution is then used to incorporate the result into the prediction as shown in Section 5.2.2.

5.2.1.3 Combination of validation results from different input conditions

By applying the stochastic assessment approach for the model reliability, a distribution for the reliability metric is obtained for each validation input condition. To include the validation result in the prediction framework that will be discussed in Section 5.2.2, a single overall measure of the model quality across the entire domain of interest is necessary. Therefore, the integration approach proposed in Section 4.3 is applied to this situation. Since the model reliability metric is a distribution at each input condition, the approach follows similarly from the method proposed in Section 4.4. By drawing samples from the distributions of each R_i

corresponding to the validation input conditions, sample vectors (henceforth denoted with superscript j) r^j are generated. Then, Eq. (4.2) is applied to obtain

$$r_{overall}^{j} = \frac{\sum_{i=1}^{m} r_i^j w_i}{\sum_{i=1}^{m} w_i} \tag{5.3}$$

The weights are computed from Eq. (4.3), and the set of samples $r_{overall}^{j}$ represent the distribution of $R_{overall}$. This distribution is the overall measure of the probabilistic performance of the model over the expected prediction domain. The overall measure is used to include the result of validation in the prediction by applying weights to the prior and posterior distributions of the parameters that were obtained during model calibration. This methodology is described in the following section.

5.2.2 Inclusion of the validation result in prediction

In the model calibration description of Section 2.1, the model parameters were calibrated jointly with calibration data from multiple input conditions. The posterior parameter distributions were then propagated through the model to validate the calibrated models against some additional data. Since the validation result is a probability, it can be used to modify the posterior parameter distributions to add additional conservatism to the prediction and account for the possibility that the model is not adequate. The underlying assumption of the proposed approach is that parameters calibrated using imperfect models should not be fully trusted when they are propagated forward to the prediction stage. Therefore, the probabilistic validation result is treated as a weight for the posterior distribution, and the remaining weight is given to the prior distribution.

In general, lack of support for the posterior distribution does not imply support for the prior distribution. However, the posterior distribution has added information and generally less uncertainty than the prior distribution. The calibration data effectively reduces the subjective probability of some parameter values that were considered possible in the prior distribution. By adding weight back to the prior when the posterior distribution is invalidated, the possibility that the posterior distribution may have been overconfident and biased is accounted for. Therefore, a wider range of parameter values should be considered, and the method given by Eq. (5.4) is applied to achieve this result.

Using the distribution of $R_{overall}$ from Section 5.2.1.3, a candidate parameter distribution for prediction is obtained for each sample of the overall model reliability as

$$f_{\boldsymbol{\Theta}}^{j}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{C}},\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{V}}\right)=r_{overall}^{j}f_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\middle|\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{C}}\right)+(1-r_{overall}^{j})f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \tag{5.4}$$

Here, $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{C}})$ is the posterior distribution obtained with the full set of calibration observations $\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{c}}$, $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is the prior distribution, and $f_{\boldsymbol{\Theta}}^{j}(\boldsymbol{\theta}|\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{C}},\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{V}})$ is a particular predictive parameter distribution corresponding to a particular sample of the overall model reliability. This predictive parameter distribution is now conditioned on the validation observations $\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{V}}$ in addition to the calibration observations. Each predictive distribution is obtained by discrete sampling of the prior and posterior, weighted by the particular sample of the validation result. Each of them is then propagated through the model at the prediction input conditions to obtain a stochastic prediction on the output of interest. Taken together, they form a family of predictions in which the variance across distributions is the result of the uncertainty in the distribution of

$R_{overall}$, and the spread in each individual prediction is the combined result of aleatory uncertainty in the inputs and the parameter uncertainty in the particular predictive parameter distribution. Thus, the prediction uncertainty is naturally separated into these two components. This information is used to construct the test selection optimization formulation proposed in Section 5.3.

5.3 Test selection optimization methodology

The primary goal of this chapter is to construct a joint optimization formulation for selecting experiments by applying the prediction framework described in Section 5.2. The proposed methods extend previous work [100], which focused only on calibration, to include a combination of validation and calibration testing options. One challenging aspect of the combined calibration/validation test selection problem is that calibration and validation information tend to contribute to the prediction in opposing ways. The calibration information reduces the uncertainty in the posterior distributions of the parameters, which in turn reduces the uncertainty in the prediction. Since all models are imperfect and never perform perfectly in validation, validation information tends to decrease the reliance on the calibrated posterior distributions since they may be overconfident. When applying the framework described in Section 5.2, the validation assessment results in giving more weight to the prior distributions for the parameters, which are independent of model quality. This treatment results in an expansion of the prediction uncertainty because of model inadequacy. Thus, the goal of calibration is to reduce prediction uncertainty, but the goal of validation is to maintain conservatism in the prediction (increased prediction uncertainty).

If viewed in this way, there is not an immediately obvious way to combine these two competing objectives. Therefore, the way the problem is viewed must be altered slightly. Rather than trying to mathematically motivate the idea of validation itself, the proposed formulation aims to mathematically motivate improvements to the quality of the assessment. In other words, the objective function is not constructed to demonstrate the value of performing model validation at all. Rather, it is constructed to demonstrate the value of performing model validation more accurately. As was shown in Section 5.2.1, reducing validation data uncertainty reduces the uncertainty in the assessment, which leads to higher confidence in the prediction. In this context, calibration data can be viewed as a means to reduce the uncertainty *in* the prediction while validation data can be viewed as a means to reduce the uncertainty *about* the prediction.

5.3.1 Objective formulation

Once the two objectives are both viewed in terms of uncertainty reduction, it is more natural to combine them into a single objective function that can be minimized over the feasible set of the number of tests of each type. The set of available testing options typically includes many different possible input conditions for both calibration and validation. The methodology described in Section 5.2 is used to motivate the test selection activities so that the value of each available option can be quantified. By sampling over the uncertainty in the overall model reliability metric, a different predictive parameter distribution is obtained for each sample. Then, each distribution $\boldsymbol{\Theta}^{\mathbf{j}}$ is propagated through the model g along with the aleatory uncertainty in the prediction inputs $\boldsymbol{X_p}$ to obtain a stochastic prediction Z^j.

$$Z^j = g(\boldsymbol{X_p}, \boldsymbol{\Theta^j}) \tag{5.5}$$

The set of distributions for Z collectively represent a family of predictions. An example of this family of distributions in CDF form is given in Figure 5.3. The overall goal is to minimize prediction uncertainty within budget constraints. To make decisions, this goal must be written in the form of an objective function; therefore, variance is used to quantify the prediction uncertainty. The variance of a family of predictions can be expressed using the law of total variance [109], which states

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)] \tag{5.6}$$

for two general random variables X and Y. In the context of the prediction problem, the variance of Z is of interest, and each prediction is conditioned on a particular sample of the overall model reliability. Therefore, Eq. (5.6) can be applied to express the prediction variance in terms of the validation result.

$$Var(Z) = E_{R_{overall}}[Var(Z|R_{overall})] + Var_{R_{overall}}[E(Z|R_{overall})] \tag{5.7}$$

The importance of this variance decomposition is that these two terms correspond to the effects of calibration and validation respectively. In particular, the goal of calibration, expressed by the $E_{R_{overall}}[Var(Z|R_{overall})]$, is to minimize the uncertainty in a single prediction by reducing parameter uncertainty in the posterior distribution that contributes to the predictive parameter distribution. On the other hand, the goal of validation, expressed by the $Var_{R_{overall}}[E(Z|R_{overall})]$, is to reduce the uncertainty about the prediction by driving a family of uncertain predictions toward a single prediction that is not biased by measurement errors. The total variance should be minimized over the set of decision variables (numbers of each type of test).

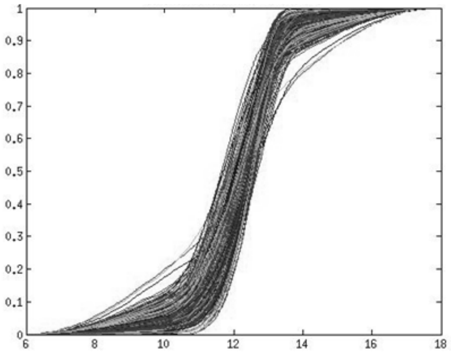


Figure 5.3: Family of CDF predictions

To make the assessment, some synthetic data must be generated to represent expected outcomes of the experiment. In the absence of any prior knowledge about the experiment, the only way that these expected outcomes can be produced is by evaluating the model at the input conditions of the experiment and then adding measurement noise to the data. The distribution of the noise is obtained from the best available information about the instrumentation accuracy. If some historical data is available on closely related experiments, a data-driven model can be created independently of the physics-based model to more accurately estimate the potential outcomes. A data-driven model could also be generated once some tests have been conducted and then improved adaptively. With any of these approaches, the expected outcomes are stochastic, so even at a fixed input condition, many realizations of experimental data can be generated from the model due to the presence of the estimated measurement noise. Therefore, the proposed formulation of the objective function takes an expectation over many realizations of synthetic experimental data.

The decision variables in the optimization problem are the numbers of tests of each type to conduct. In this chapter, a finite set of testing options is considered. Thus, the decision variables are denoted as vectors $\boldsymbol{n_C}$ (length m_C for m_C different calibration input conditions) and $\boldsymbol{n_V}$

(length m_V for m_V different validation input conditions). Once the decision variables are selected, an arbitrary number of data realizations (limited by computational expense) are generated with observation vector lengths equal to the values of the decision variables. Then, for each realization of the data vector, $\boldsymbol{d} = [\boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{C}}, \boldsymbol{y}_{\boldsymbol{d}}^{\boldsymbol{V}}]$, the entire integration procedure described in Section 5.2 is performed, resulting in a family of predictive parameter distributions each denoted as in Eq. (5.5). Within this context, the following formulation for the optimization problem is proposed.

$$\min_{\boldsymbol{n_C},\boldsymbol{n_V}} E_{\boldsymbol{D}}[Var(Z|\boldsymbol{D})]$$

$$s.t.\ \ \boldsymbol{h}_{\boldsymbol{C}}^T\boldsymbol{n_C} + \boldsymbol{h}_{\boldsymbol{V}}^T\boldsymbol{n_V} \leq b \tag{5.8}$$

The constraint function for the decision variables is given by Eq. (5.8) where the total cost is constrained by a total testing budget b. The row vectors $\boldsymbol{h_C}$ and $\boldsymbol{h_V}$ of lengths m_C and m_V respectively contain the costs of the calibration and validation tests at each available input condition, and the superscript T denotes a vector transpose. The formulation of Eq. (5.8) can be further decomposed by applying Eq. (5.7) and taking advantage of the linearity of the expected value operator.

$$E_{\boldsymbol{D}}[Var(Z|\boldsymbol{D})] = E_{\boldsymbol{D}}\{E_{R_{overall}}[Var(Z|R_{overall}, \boldsymbol{D})]\} + E_{\boldsymbol{D}}\{Var_{R_{overall}}[E(Z|R_{overall}, \boldsymbol{D})]\} \tag{5.9}$$

The first term is improved by collecting calibration data since narrowing the posterior distribution of $\boldsymbol{\theta}$ will also tend to reduce the average prediction variance. The second term is improved by collecting validation data since this data will converge the distribution of $R_{overall}$ toward a deterministic value at the limit (i.e. infinite data). A deterministic value of $R_{overall}$

implies that the $Var_{R_{overall}}[E(Z|R_{overall}, \boldsymbol{D})]$ is zero for any $\boldsymbol{d}$. These two terms in Eq. (5.9) can be expanded to

$$E_{\boldsymbol{D}}\left\{E_{R_{overall}}[Var(Z|R_{overall}, \boldsymbol{D})]\right\} = \iint Var_Z(Z|R_{overall}, \boldsymbol{D}) f_{R_{overall}}(r) f_{\boldsymbol{D}}(\boldsymbol{d}) \mathrm{d}r\mathrm{d}\boldsymbol{d} \quad (5.10)$$

$$E_D\left\{Var_{R_{overall}}[E(Z|R_{overall})]\right\} = \int Var_{R_{overall}}\left[\int (z|r) f_{Z|R}(z|r)\mathrm{d}z\right] f_{\boldsymbol{D}}(\boldsymbol{d})\mathrm{d}\boldsymbol{d} \quad (5.11)$$

The goal of the optimization problem is to minimize the sum of the two integrals given in Eq. (5.10) and (5.11). Note that weights could be applied to these two terms if there were reason to preference one over the other. However, the weighted sum would not reflect the overall prediction uncertainty precisely.

Each of these integrals is evaluated by Monte Carlo sampling since the density function for $R_{overall}$ is not known analytically, and the data model $f_{\boldsymbol{D}}(\boldsymbol{d})$ may not have an analytical form either. Since they are evaluated by sampling, the objective function value is inherently stochastic. In addition, the decision variables are discrete quantities, and a relaxation to the continuous space is not possible since fractional tests are meaningless. These two factors (stochasticity and discreteness) significantly limit the available options for solving the optimization problem, which leads to the solution strategy that follows.

5.3.2 Solution approach for the joint optimization problem

As mentioned, the stochasticity and discreteness of the formulation make the selection of an efficient algorithm non-trivial. However, the focus of this section is not optimization methods, but the problem formulation. The use of a simulated annealing algorithm [49] is proposed because it is suited to handle stochastic discrete problems [55] even though it is not a particularly efficient search algorithm. Any other algorithm that is capable of handling discrete, stochastic problems could be substituted. The simulated annealing algorithm starts from an initial guess and then takes random walks in the domain in all dimensions simultaneously. In a discrete problem, these random walks can be made with a continuous proposal density function, but the iterate must be rounded to the nearest discrete value in all dimensions since the objective function cannot be evaluated with numbers of tests that are not integers. Any iterate that improves (i.e. decreases) the objective function value is accepted, and any point that increases the objective function value is accepted with probability p given by

$$p = exp(-\frac{\Delta f}{t}) \tag{5.12}$$

Here, Δf is the change in the objective function from the previous iterate, and t is the current value of the temperature parameter that governs how tight the acceptance criterion should be. The reason for accepting points that do not improve the objective function is to attempt to explore the entire space and reduce the opportunity to stop at a local minimum. As the algorithm proceeds, the threshold for acceptance becomes tighter, so only decreases and very small increases to the objective function can be accepted. This threshold tightening is governed by a reduction to the temperature parameter as

$$t = t_0 \left(1 - \frac{k}{k_{max}}\right)^{\alpha} \tag{5.13}$$

where t_0 is the user-defined starting temperature, k is the current iteration number, k_{max} is the total number of allowable iterations, and α is an exponent that determines the rate of temperature decrease. Once the total number of allowable iterations is expended, the iterate, among all candidate points, with the lowest objective function value is selected. Selecting a point that is not at or near the constraint boundary may be evidence that the search routine should be conducted again locally to ensure that the solution is fully converged.

Applying this method to the objective formulation in Eq. (5.8) produces the optimal number of tests to perform for calibration and validation at each available input condition for a given budget. Once the number of tests of each type is selected, the tests can be conducted. Once some real experimental data is available, it should be used to validate the synthetic data generation models that were used in the optimization problem. If significant bias exists, it may be useful to update the data models and perform the analysis again.

It is noted here that the proposed formulation may be quite computationally expensive to solve. The stochasticity of the objective function value can only be reduced by taking a larger number of Monte Carlo samples when performing the necessary variance and expected value computations. The large number of model evaluations that are required for calibration and validation is likely to make using the physics-based model unaffordable. Therefore, surrogate models should be trained from the physics-based model in order to perform the uncertainty propagation efficiently. In the example demonstration of Section 5.4, GP surrogate models are used.

5.4 Numerical example

To demonstrate the proposed methodology, the RF MEMS example of Section 4.5 is explored further. The pull-in and pull-out voltages are predicted by device simulation, and 6 different devices are available for testing. The pull-in voltage measurements will be used for calibration, and the pull-out voltage measurements will be used for validation. Therefore, the goal of the test selection problem is to determine how many replicate calibration and validation tests to perform on each device. The lengths of the vectors $\boldsymbol{n_C}$ and $\boldsymbol{n_V}$ are each 6, and there are a total of 12 decision variables.

The same 5 input variables $[h, g_1, g_2, E, d_c]$ as in Section 4.5 are considered. In this illustration, the geometry parameters g_1, g_2, and h are described by known aleatory distributions for each of the six devices. Direct measurements of d_c are not available, but an aleatory range for the variable is obtained via multi-scale simulation [48, 53]. These simulations also provide a prior distribution for E, but this variable is treated as the calibration parameter since the material properties are common to all devices for the purposes of calibration, validation and prediction. Therefore, within the framework of Section 5.2, the inputs g_1, g_2, h, and d_c are the components of $\boldsymbol{X_i}$ for each device, and the θ of interest is E.

In the prediction scenario of interest, no information is known about the particular values of g_1, g_2, and d_c. However, suppose the distribution of h in the prediction scenario is expected to be $\pi(\boldsymbol{x}) \sim N(1.2, 0.1)$, and the thickness value is known to be particularly important to the model prediction. Therefore, the suitability of the model for prediction is judged with respect to this expected thickness input condition, and the relevance of the available device validation input conditions is determined according to their proximity to this condition. Note that the weights obtained from the method of Section 4.3 are not dependent on the output observations of the experiments. Therefore, the weights for the validation results can be obtained by applying Eq. (4.2) when only the aleatory distributions for h for each available device are known. The thickness distributions for the experiments are the same as those used in Section 4.5 (see Figure 4.5), and the associated weights are therefore the same as well (see Table 4.2).

Given these scenarios for calibration, validation, and prediction, the optimization problem is formulated as in Eq. (5.8). For this particular problem, some calibration and validation observations had been made prior to this analysis; these observations were used to perform the validation assessment in Section 4.5. Obviously, this scenario would never actually exist when applying the proposed methods because the budget has already been spent, and solving the test selection problem after the fact is not very useful. However, for the sake of illustration, the available measurements are used to compute a sample mean and sample variance to construct the synthetic data generation models that are needed in the formulation. Using only the statistics of the observations (and not the observations themselves) many data realizations can be generated for fixed lengths of the observation vectors.

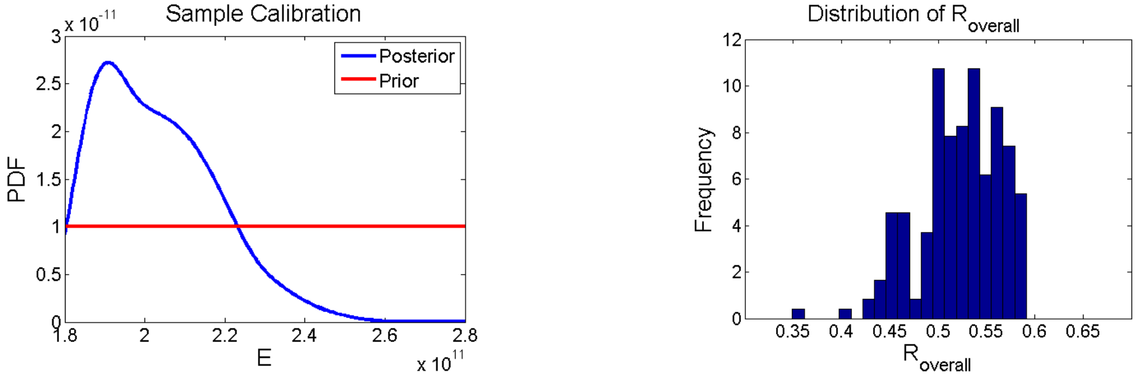
Suppose that a pull-in voltage experiment (i.e. a calibration test) requires 1 cost unit, a pull-out voltage experiment (i.e. a validation test) requires 2 cost units, and a budget of 80 cost units is available for testing. To bound the problem, a maximum of 10 tests of each type is allowed, so it is possible to conduct 0 to 10 calibration tests on each device and 2 to 10 validation tests on each device. (Note that a limitation of the proposed methods is that it is not possible to stochastically assess the model reliability for 0 or 1 validation test.) A full grid search of this 12-dimensional design space would require approximately 900 billion evaluations of the objective function. Even using surrogate models, the objective function in this example still requires about 20 minutes per evaluation (this will vary greatly depending on implementation) due to the expense of the nested Monte Carlo sampling, especially the MCMC routines for calibration. Many of the candidate points do not satisfy the budget constraints, but even the remaining feasible design space is clearly unaffordable to explore fully.

Therefore, the simulated annealing algorithm described in Section 5.3.2 is executed. As a starting point for the optimization algorithm, the budget is divided equally across the 12 available testing options (i.e. 4 calibration and validation tests on each device, expending 72 cost units). The algorithm is allowed to run for a maximum of 200 iterations, and it is constrained such that only feasible points are evaluated. Then, starting from the best point that is found during the initial run, a second run of the algorithm is conducted with reduced temperature (i.e. a stricter acceptance criterion) is allowed to run for 100 iterations. The goal of this approach is to explore the space globally in the first run and then refine the solution locally in the second run. After both runs have been completed, the minimum objective function value that is discovered is assumed to be the optimum. The result for this problem is given in Table 5.1. Since the objective function is stochastic, and the simulated annealing algorithm itself is also stochastic, there is no theoretical guarantee that this result is the global optimum. However, the result provides some very valuable insights about the value of the different testing options.

Table 5.1: Optimal test selection result

Device 1		Device 2		Device 3		Device 4		Device 5		Device 6	
n_C^1	n_V^1	n_C^2	n_V^2	n_C^3	n_V^3	n_C^4	n_V^4	n_C^5	n_V^5	n_C^6	n_V^6
0	2	6	6	0	2	9	9	0	10	3	2

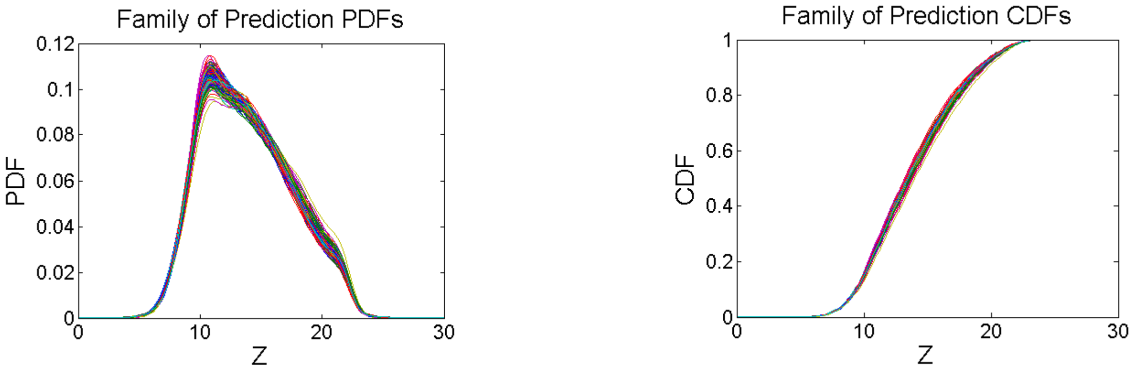
To illustrate the methodology further, the evaluation of the objective function at the optimum is demonstrated. For a particular realization of all the calibration data (i.e. pull-in voltage measurements), the posterior distribution for E, shown in Figure 5.4a, is obtained by updating the uniform prior distribution with Eq. (2.1). This posterior distribution is then propagated through the model at each validation input condition. The particular realization of the validation data (i.e. pull-out voltage measurements) is used to construct a t-distribution for the mean observation for each device. Samples from this t-distribution are taken and compared against the stochastic model output to obtain the distributions R_i for each of the 6 devices. By applying the weights in Table 4.2 in Eq. (5.3), samples from the distributions are combined to obtain the distribution for $R_{overall}$ as shown in Figure 5.4b. Each sample $r_{overall}$ is used as a weight for the posterior distribution in Eq. (5.4), and then the resulting predictive distribution is propagated through Eq. (5.5) to obtain a single prediction distribution. The set of samples produces a family of distributions; the family is shown in PDF form in Figure 5.5a and in CDF form in Figure 5.5b. The first part of the objective function, given by Eq. (5.10), is obtained by taking the variances of the individual distributions and averaging them. The second part, given by Eq. (5.11), is obtained by looking at the variance across the means of the individual distributions. These two together represent the overall prediction uncertainty, which is averaged over the data realizations to obtain the objective function value.



a) Sample calibration of E

b) Distribution of the overall model reliability

Figure 5.4: Parameter uncertainty for a particular data realization



a) Sample PDF family

b) Sample CDF family

Figure 5.5: Family of predictions for a particular data realization $\boldsymbol{d}$

In this particular problem, the optimization result shows that the most valuable testing option is validation tests on device 5. This result is not surprising since device 5 has the largest weight in the overall model reliability. On the other hand, validation tests on devices that are not relevant to the prediction (e.g. devices 1, 3, and 6) provide very little benefit. This fact emphasizes that it is very important to know how a model will be used in prediction in order to validate it efficiently. Since the weights can be computed prior to the test selection analysis if the available validation input conditions are known, validation conditions that are not relevant to the prediction can potentially be ignored in the analysis altogether.

Since all of the calibration tests are taken jointly within the proposed framework, it is more important how many total tests are conducted than on which device they are conducted. The overall magnitude of the first part of the objective function given in Eq. (5.10) is significantly larger than the second part of the objective given in Eq. (5.11). This result is obvious when looking at the family of distributions in Figure 5.5. However, in this problem, the parameter uncertainty can be reduced more rapidly than the uncertainty in the validation result. After a few calibration tests have been performed, the majority of the parameter uncertainty reduction has been achieved, and additional tests give diminishing improvement. The uncertainty in the validation result converges more slowly; therefore, there is more value in doing a larger number of replicate validation tests than calibration tests. This result is evident in the validation result since there are a larger number of total validation tests even though they are twice as expensive.

5.5 Conclusion

This chapter provides a test selection methodology that combines validation and calibration activities. The proposed optimization framework employs a methodology for integrating calibration and validation data probabilistically to make a prediction. By performing Bayesian calibration and a stochastic validation assessment, both calibration and validation data collection are motivated by prediction uncertainty reduction. The prediction uncertainty can be decomposed into two components: one which is improved by adding calibration data and one which is improved by adding validation data. Since model calibration is performed jointly over multiple input conditions, the total number of tests may be more important than which particular test is performed as long as the parameters that are calibrated are common to the different tests. On the other hand, it is very important which validation tests are conducted. The proposed framework weights the validation input conditions according to their relevance to the intended usage condition. Therefore, the tests at the relevant input conditions provide much more value than those at less relevant conditions. For both calibration and validation, uncertainty reduction is fastest with the first few test samples, and then the relative improvement to the prediction decreases as more data is collected. However, in the example shown, this diminishing improvement occurred more rapidly for calibration than validation. The methodology in this chapter is aimed at achieving minimum prediction uncertainty for a fixed budget. Since the value of the tests decreases as more tests are conducted, future work will aim to determine how many tests are enough and what budget is appropriate.

CHAPTER 6

RISK-BASED RESOURCE ALLOCATION

6.1 Introduction

The preceding chapters of this dissertation are aimed at predicting a stochastic output of interest accurately and efficiently. Within the UQ framework that is established, the goal of the proposed resource allocation methods in the previous chapters is to solve the inverse problem of reducing the uncertainty in this prediction by collecting additional data. For a given stochastic prediction, the probability distribution may be compared against a threshold failure criterion (either a deterministic or stochastic value) to compute a failure probability. For example, suppose a failure occurs if the prediction quantity of interest Z exceeds its maximum allowable value z_{max}. Then, the probability of failure p_f is defined as

$$p_f = P(Z > z_{max}) \tag{6.1}$$

Since the prediction for Z is typically obtained via MCS methods, it is only necessary to count the number of failures (i.e. the number of points in a set of samples that do not pass the given threshold) n_f and divide by the total number of samples n to calculate the probability of failure p_f.

$$p_f = \frac{n_f}{n} \tag{6.2}$$

The complement of the failure probability is the system reliability, i.e., the reliability is defined by $r = 1 - p_f$. As mentioned in Section 2.3, MCS and efficient simulation techniques have been developed from the perspective of reliability analysis (i.e. computing or approximating r) [32, 33, 82, 93]. Once the reliability estimate is computed, it is connected to risk by the consequence of the failure event.

Risk assessment is an important extension of reliability analysis. Risk (not reliability) is commonly the motivating factor in design and management decisions because it has a more direct economic interpretation (e.g. expected dollars lost). Classically, the risk of an event (e.g. system failure) has two key components: (1) the probability of the event and (2) the consequences of the event [37]. These two components have a simple, logical relationship, in which the risk S is a product of the consequence of an event L and the probability of the event $P(L)$.

$$S = L * P(L) \tag{6.3}$$

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

$$S_T = \sum_{i=1}^{m} L_i P(L_i) \tag{6.4}$$

In the design and management phases, the goal is to minimize the total system risk, so all of the failure modes should be considered; multiple modes may be affected by a single decision. System risk minimization is directly connected to reliability analysis and stochastic prediction. Designers and decision-makers have little control over the consequences of an event, so risk minimization is significantly enabled by minimizing prediction uncertainty while maintaining prediction accuracy (i.e. low bias). Thus, risk minimization and prediction uncertainty reduction are directly aligned, and the resource allocation framework developed in this dissertation could also be motivated by risk reduction, rather than prediction uncertainty reduction. Although these two formulations are not precisely equivalent, the analyst should reasonably expect that either formulation would lead to a similar conclusion. However, from an economic perspective, it is not logical to spend large resources on UQ without also considering the total benefit of the analysis. Since risk can be directly interpreted as a cost, it provides a convenient space to analyze design and management decisions.

6.2 Failure risk vs. development risk

Two types of risk are considered in this chapter: failure risk and development risk. While it is obvious that system failure carries a cost, and therefore a risk, system development has not classically been viewed as a risk in UQ analysis. In some applications it may be possible to reduce epistemic uncertainty to an arbitrarily small value by collecting large quantities of data (of high quality) and/or dedicating large resources to computational model improvement. In such a scenario, it is possible to perform too much UQ analysis from an economic perspective. Exhaustive UQ techniques may lead to very accurate model predictions, but when failure probabilities are very low, these techniques may be more conservative than is necessary.

Recent research has the goal of determining how far the UQ process should go. Romero [91] refers to "model builder's risk" as the risk associated with rejecting a valid model and compares this risk against the "model user's risk," which is associated with making predictions with an invalid model (i.e. failure risk). Decision-makers must determine how to effectively balance these two types of risk. Development costs and other UQ expenses (i.e. development risk) are spent with 100% probability once improvement decisions are made. On the other hand, system failure typically has a very low probability, but a very high consequence. These types of events have been compared in the literature by using risk matrices [15]. Failure risk is low probability and high consequence while development risk is high probability and low consequence (i.e. the cost of the development activities must in general be much lower than the cost of failure). Typically, decision makers have been biased toward the system failure risk over the development risk, which is ethically correct since the system decisions often have broader impacts that cannot easily be quantified with a monetary value.

6.3 Economic considerations for uncertainty quantification

Within the context of these two types of risk (the development risk S_d and the failure risk S_f), the overall goal is to minimize the total risk S_T, defined in Eq. (6.5).

$$S_d = 1 * L_d$$

$$S_f = p_f * L_f$$

$$S_T = S_d + S_f \tag{6.5}$$

where L_d is the total cost of UQ development activities, p_f is the system failure probability, and L_f is the consequence of system failure. It is assumed that L_f is a constant that the analyst cannot control, and L_d is fully controlled by the analyst. Therefore, the goal of the minimization problem is to decide how much money should be spent on UQ activities. Note that the failure probability is a dependent variable, and it is an unknown function of L_d as

$$p_f = g(L_d) \tag{6.6}$$

This unknown function depends upon how the available resources are allocated to the various UQ activities. Effectively, L_d can be viewed as the budget for UQ.

In Chapter 3, a budget for uncertainty propagation was assumed, and then model selection decisions were made to mximize prediction accuracy subject to that budget. In Chapter 5, a testing budget was assumed, and then tests were selected among the available test scenarios in order to minimize prediction uncertainty subject to that budget. These two sets of activities could be decided jointly subject to total budget L_d.

6.3.1 Combined model selection and test selection

To solve the joint optimization problem of model selection and test selection with a total fixed budget L_d, the first step is to quantify the cost of computational time. For example, suppose the cost of each unit of computational time is α. Then, the total cost of the computational simulations c_m is given by

$$c_m = \alpha t \tag{6.7}$$

Recall from Chapter 5 that

$$c_{test} = \boldsymbol{h}_{\boldsymbol{C}}^T \boldsymbol{n}_{\boldsymbol{C}} + \boldsymbol{h}_{\boldsymbol{V}}^T \boldsymbol{n}_{\boldsymbol{V}} \tag{6.8}$$

Therefore, the joint optimization problem is constrained by the combined total cost of the two sets of activities, $c_m + c_{test} \leq L_d$. Since the goal of the optimization is to minimize risk, that goal can be simplified to a minimization of p_f since the remainder of the variables in the problem are constants. Thus, the combined test selection and model selection problem is formulated as

$$\min_{\boldsymbol{n_C},\boldsymbol{n_V},\boldsymbol{k}} p_f$$

$$s.t.\quad \boldsymbol{h}_{\boldsymbol{C}}^T\boldsymbol{n_C} + \boldsymbol{h}_{\boldsymbol{V}}^T\boldsymbol{n_V} + \boldsymbol{\alpha^T}\boldsymbol{t^k} \leq L_d \tag{6.9}$$

Here, $\boldsymbol{k}$ is the vector of model combination selections that are made, $\boldsymbol{t^k}$ is the column vector of computational times required for the selected models, and $\boldsymbol{\alpha}$ is the column vector of α values repeated to match the length of $\boldsymbol{t^k}$. Recall that $\boldsymbol{n_C}$ and $\boldsymbol{n_V}$ define the number of replicate tests to perform at each of the candidate calibration and validation input conditions. Note that p_f is not deterministic when there are a family of predictions as demonstrated in Chapter 5 (i.e. each prediction yields a single p_f). Therefore, the p_f that is used in Eq. (6.9) may be an expectation taken over a set of predictions. In this context, the variance on the p_f should also be considered.

Solving this problem requires the application of the UQ framework for model calibration, model validation, and uncertainty propagation that has been explored in this dissertation. The problem can then be solved by a nested optimization formulation consisting of the following: (1) an outer loop that takes the total budget L_d and divides it between the test selection and model selection activities and (2) an inner loop which solves two separate and independent optimization formulations that have been addressed in Chapter 3 and Chapter 5 respectively. Since a nested optimization formulation can be quite expensive to solve, the efficiency of the solution approach can be improved by applying a single-loop decoupling strategy [112].

6.3.2 Risk minimization

Solving the joint model selection and test selection problem repeatedly for different total UQ spending L_d provides discrete observations of the functional relationship between L_d and p_f that define the relationship in Eq. (6.6). These pairs of values for L_d and corresponding p_f could be used to train a GP surrogate model for the relationship, which will improve the efficiency of solving the risk minimization problem. The risk minimization problem is then formulated as

$$\min_{L_d} S_T \tag{6.10}$$

Note that this is a continuous unconstrained optimization problem. Any amount of spending from $[0, \infty]$ is possible. Of course, there may be some practical budget constraints on this spending, but this formulation is constructed to determine what the budget itself should be.

Even without supplying any practical constraints, the solution of the problem is still bounded because increasing L_d from any starting point has both a positive and negative effect on S_T. The cost L_d is equal to S_d, which is added to S_T directly, thereby making the objective function value increase. However, increasing L_d is also expected to reduce p_f, which in turn, reduces S_f and decreases the objective function value. Thus, the formulation is a natural way of exploring the economic tradeoff between development spending and reduction of p_f. The optimal solution depends on problem specific variables and relationships, most notably the consequence of failure L_f and the unknown function g in Eq. (6.6) that defines how much p_f decreases with additional spending.

6.4 Conclusion

Earlier chapters of this dissertation were focused entirely on prediction uncertainty quantification and reduction. This chapter introduces the context of risk, which can be used to motivate the general UQ framework. Optimization problems that can be used to guide resource allocation decisions are formulated. The overall goal of any UQ framework is risk reduction, but the framework itself should not cost more than the risk reduction that it achieves. Thus, the risk minimization problem is solved from the perspective of how much UQ spending is economically efficient. In order to solve the risk minimization problem, it must first be clear how the UQ activities are reducing the system failure probability. Answering this question requires solving another joint optimization problem for the combination of model selection and test selection for many different spending budgets.

CHAPTER 7

CONCLUSION

7.1 Summary of accomplishments

This dissertation explores a comprehensive framework for UQ in the context of model calibration, model validation, and uncertainty propagation for prediction. Some of the key features of the forward propagation problem are first discussed, and then the UQ framework is used to solve resource allocation problems. Methods of separating aleatory and epistemic uncertainty sources are proposed, and then epistemic uncertainty reduction strategies are explored and optimized. New techniques that account for data uncertainty in model validation and connect the model validation results to the prediction are proposed. Then, resource allocation strategies for model selection and test selection are proposed from the perspective of prediction uncertainty quantification and reduction. Finally, the concept of risk is discussed and used to motivate UQ and suggest how much is sufficient.

In the model selection framework of Chapter 3, the proposed approach uses GP surrogate models for decision-making and takes advantage of local fidelity preferences by making input-dependent selection decisions. The decision-making methods themselves are very fast to develop, and they can significantly improve the efficiency of the underlying multi-fidelity simulation. The tradeoff decision of accuracy vs. computational expense is considered explicitly by introducing a tolerance on the simulation result. Two different strategies are considered depending on whether the ranking of fidelities is constant across the domain or locally specific.

A model validation methodology for connecting different data scenarios to the prediction of interest is proposed in Chapter 4. Three types of experiments are considered: uncharacterized, partially characterized and fully characterized. The proposed methods enable aleatory and epistemic uncertainty sources to be separated from one another, which aids in decision making for uncertainty reduction when the model performance is inadequate. The individual metric values can be integrated into a single metric by weighting each value with the probability of observing the corresponding input in the prediction domain (i.e. relevance to the intended application of the model). The weighting approach demonstrates that there may be large differences in the importance of the various validation experiments for different prediction scenarios.

The proposed test selection methodology in Chapter 5 combines validation and calibration activities. The proposed optimization framework employs a methodology for integrating calibration and validation data probabilistically to make a prediction. The prediction uncertainty can be decomposed into two components: one which is improved by adding calibration data and one which is improved by adding validation data. The test selection methodology is then aimed at achieving minimum prediction uncertainty for a fixed budget. The validation tests at the input conditions that are relevant to the prediction provide much more value than those at less relevant conditions. The value of both calibration and validation tests decreases as more tests are conducted.

In Chapter 6, the concept of risk is introduced and used to motivate spending in the general UQ framework. Risk minimization problems that can be used to guide resource allocation decisions are formulated. Thus, they are solved from the perspective of how much UQ spending is economically efficient. In order to solve the risk minimization problem, it must first be clear how the UQ activities are reducing the system failure probability. Answering this question requires solving another joint optimization problem for the combination of model selection and test selection for many different spending budgets.

7.2 Future needs

More work is needed to extend and demonstrate the proposed framework. In particular, the impact of many of the parameters that guide the resource allocation decisions needs to be considered carefully. For example, the model selection approach that has been proposed requires an allowable tolerance on the prediction accuracy of lower fidelity (e.g. reduced-order, reduced-physics, or more coarsely refined) models to be chosen. Similarly, the model validation approach requires an acceptable threshold for the difference between prediction and observation to be selected. Future work will explore methods for determining these parameters within the risk reduction formulation by considering how these parameters change the system reliability.

Further work is also needed to integrate the model selection approach with a dynamic computing resource allocation methodology, and with decisions about future model improvements. A complete orchestration of the UQ process for complicated problems with many component simulations will need algorithms to schedule the selected simulations and take advantage of parallelization in order to further reduce the computational effort while achieving the desired accuracy and precision.

In this dissertation, sparse validation data has been incorporated by applying a t-distribution methodology. While this approach is fitting for Gaussian noise (a common scenario), more general forms of this distribution could also be considered. Future work will explore more general formulations such as the Johnson distribution family that can be combined with Bayesian updating methods for the distribution parameters.

The proposed model validation methods have only been demonstrated for the model reliability metric since it can be interpreted probabilistically. More work is needed to demonstrate the compatibility of the proposed methods with other validation metrics including Bayesian hypothesis testing and the area validation metric. The importance of the probabilistic treatment of validation is that it can be used to directly incorporate the validation result into the prediction. In this context, future work will also explore the effect of extrapolation on the system failure risk. The proposed integration approach incorporates the "proximity" of the validation tests to the prediction regime, but the result is then normalized across the available conditions. Therefore, an important issue to address is how far away from the validation regime the tests still retain relevance.

Additional work is also needed to demonstrate the risk minimization and combined test and model selection formulations that are proposed on a realistic example. Important issues to address in this demonstration are robustness and efficiency of the optimization approach. In particular, many of the optimization formulations that are proposed in this dissertation are stochastic because they require nested sampling to propagate uncertainty each time the objective function is evaluated. The effect of this stochasticity on the convergence of the optimization methods should be carefully considered, and for efficiency, methods for determining how many samples can be afforded during each uncertainty propagation step should be explored.

BIBLIOGRAPHY

1 Akaike, H., “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, 1974, pp. 716-723. doi: 10.1109/TAC.1974.1100705

2 Alexandrov, N. M., Lewis, R. M., Gumbert, C. R., Green, L. L., and Newman, P. A., “Approximation and Model Management in Aerodynamic Optimization with Variable-Fidelity Models,” *Journal of Aircraft*, Vol. 38, No. 6, 2001, pp. 1093-1101.

3 Alifanov, O. M., *Inverse Heat Transfer Problems*, Springer-Verlag, London, UK, 1994.

4 Allaire, D. and Willcox, K., "A Bayesian-Based Approach to Multi-Fidelity Multidisciplinary Design Optimization," AIAA-2010-9183, presented at 13th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Fort Worth, TX, September 13-15, 2010.

5 Alrefaei, M.H. and Andradottir, S., “A Simulated Annealing Algorithm with Constant Temperature for Discrete Stochastic Optimization,” *Management Science*, Vol. 45, No. 5, 1999, pp. 748-764. http://dx.doi.org/10.1287/mnsc.45.5.748

6 Angus, J., “The probability integral transform and related results,” *SIAM Review*, Vol. 36, No. 4, 1994, pp. 652-654. doi: 10.1137/1036146

7 Arendt, P. D., Apley, D. W., and Chen, W., “Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability,” *Journal of Mechanical Design*, Vol. 134, No. 10, 2012, pp. 100908. doi:10.1115/1.4007390

8 ASME, Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer, ASME V&V 20-2008, American Society of Mechanical Engineers, 2008.

9 Asprey, S. P. and Macchietto, S., “Designing robust optimal dynamic experiments,” *Journal of Process Control*, Vol. 12, No. 4, 2002, pp. 545-556. doi: 10.1016/S0959-1524(01)00020-8

10 Berger, J. O., *Statistical decision theory and Bayesian analysis*, Springer-Verlag, New York, NY, 1985.

11 Bichon, B. J., McFarland, J. M., and Mahadevan, S., “ Efficient Surrogate Models for Reliability Analysis of Systems with Multiple Failure Modes,” *Reliability Engineering & System Safety*, Vol. 96, No. 10, 2011, pp. 1386-1395.

12 Bingham, D. R. and Chipman, H. A., “Incorporating prior information in optimal design for model selection,” *Technometrics*, Vol. 49, No. 2, 2007, pp. 155-163. doi: 10.1198/004017007000000038

13 Brynjarsdóttir, J. and O’Hagan, A., “Learning about physical parameters: The importance of model discrepancy,” *Inverse Problems*, 2014, Accepted for publication.

14 Burnham, K. P., and Anderson, D. R., *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition, Springer, New York, 2002.

15 Cox, L.A. Jr., “What's Wrong with Risk Matrices?”, *Risk Analysis*, Vol. 28, No. 2, 2008. doi:10.1111/j.1539-6924.2008.01030.x

16 Cressie, N. A. C., *Statistics for Spatial Data*, revised edition, Wiley, New York, 1993.

17 Dey, A., and Mahadevan, S., “Ductile Structural System Reliability Analysis Using Adaptive Importance Sampling,” *Structural Safety*, Vol. 20, No. 2, 1998, pp. 137–154. doi:10.1016/S0167-4730(97)00033-7

18 Du, X., and Chen, W., “Towards a Better Understanding of Modeling Feasibility Robustness in Engineering Design,” *Journal of Mechanical Design*, Vol. 122, No. 4, 2000, pp. 385–394.

19 Dubois, D. and Prade, H., *Possibility Theory: An approach to Computerized Processing of Uncertainty*, Plenum Press, New York, NY, 1986.

20 Fedorov, V. V. and Hackl, P., *Model-oriented design of experiments*. Springer, New York, NY, 1997.

21 Fernández, M., *Models of Computation: An Introduction to Computability Theory*, Springer-Verlag, London, UK, 2009.

22 Ferson, S., and Oberkampf, W., “Validation of imprecise probability models,” *International Journal of Reliability and Safety*, Vol. 3, No. 1, 2009, pp. 3-22. doi:10.1504/IJRS.2009.026832.

23 Ferson, S., Oberkampf, W., and Ginzburg, L., “Model validation and predictive capability for the thermal challenge problem,” *Computer Methods in Applied Mechanics and Engineering*, Vol. 197, No. 29-32, 2008, pp. 2408-2430. doi:10.1016/j.cma.2007.07.030

24 Finkel, D., Kelley, C., “Convergence analysis of the DIRECT algorithm,” *Optimization Online*, 2004, pp. 1-10.

25 Ghanem, R., Doostan, A., and Red-Horse, J., “A probabilistic construction of model validation,” *Computer Methods in Applied Mechanics and Engineering*, Vol. 197, No. 29-32, 2008, pp. 2585-2595. doi: 10.1016/j.cma.2007.08.029

26 Ghanem, R. and Spanos, P., Stochastic finite elements: a spectral approach, Springer, New York, 1991.

27 Gilks, W. and Wild, P., “Adaptive Rejection Sampling for Gibbs Sampling,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 41, No. 2, 1992, pp. 337-348. doi:10.2307/2347565

28 Grunwald, P. D., *The Minimum Description Length Principle*, The MIT Press, Massachusetts, 2007.

29 Gurley, K., and Kareem, A., "On the Analysis and Simulation of Random Process Utilizing Higher Spectra and Wavelet Transforms," *Proceedings of 2nd International Conference on Computational Stochastic Mechanics*, Athens, Greece Balkema, Roterdam, 1994.

30 Gutin, G., Yeo, A., and Zverovich, A., “Traveling salesman should not be greedy: domination analysis of greedy-type heuristics for the TSP,” *Discrete Applied Mathematics*, Vol. 117, 2002, pp. 81–86.

31 Haarhoff, L. J., Kok, S., and Wilke, D. N., “Numerical Strategies to Reduce the Effect of Ill-Conditioned Correlation Matrices and Underflow Errors in Kriging,” *Journal of Mechanical Design*, Vol. 135, No. 4, 2013, p. 044502. doi: 10.1115/1.4023631

32 Haldar, A., and Mahadevan, S., *Probability, Reliability, and Statistical Methods in Engineering Design*, Wiley, New York, 2000.

33 Harbitz, A., "Efficient Sampling Method for Probability of Failure Calculation," *Structural Safety*, Vol. 3, No. 2, 1986, pp. 109–115. doi:10.1016/0167-4730(86)90012-3

34 Hartmann, C., Smeyers-Verbeke, J., Penninckx, W., Vander Heyden, Y., Vankeerberghen, P., and Massart, D., "Reappraisal of hypothesis testing for method validation: detection of systematic error by comparing the means of two methods or of two laboratories," *Analytical Chemistry*, Vol. 67, No. 24, 1995, pp. 4491--4499. doi:10.1021/ac00120a011

35 Hastings, W., "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, Vol. 57, No. 1, 1970, pp. 97-109. doi:10.1093/biomet/57.1.97

36 Helton, J. and Sallaberry, C., "Uncertainty and Sensitvity Analysis: From Regulatory Requirements to Conceptual Structure and Computational Implementation," in: *Uncertainty Quantification in Scientific Computing, IFIP Advances in Information and Communication Technology*, Vol. 377, Springer Berlin Heidelberg, 2012, pp. 60-77. doi:10.1007/978-3-642-32677-6_5

37 Henley, E. J., and Kumamoto, H., *Reliability Engineering and Risk Assessment*, Prentice-Hall, New Jersey, 1981.

38 Higdon, D., Gattiker, J., Williams, B., and Rightley, M., "Computer Model Calibration Using High-Dimensional Output," *Journal of the American Statistical Association*, Vol. 103, No. 482, 2008, pp. 570-583. doi: 10.1198/016214507000000888

39 Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., and Ryne, R., "Combining field data and computer simulations for calibration and prediction," *SIAM Journal on Scientific Computing*, Vol. 26, No. 2, 2005, pp. 448-466. doi:10.1137/S1064827503426693

40 Hills, R. G. and Leslie, I. H., "Statistical validation of engineering and scientific models: validation experiments to application," Sandia technical report (SAND2003-0706).

41 Hills, R. G. and Trucano, "Statistical Validation of Engineering and Scientific Models: Background," Sandia technical report (SAND99-1256).

42 Hombal, V. and Mahadevan, S., "Bias Minimization in Gaussian Process Surrogate Modeling for Uncertainty Quantification," *International Journal for Uncertainty Quantification*, Vol. 1, No. 4, 2011, pp. 321-349.

43 Hombal, V. K., and Mahadevan, S., "Model Selection Among Physics-Based Models," *Journal of Mechanical Design*, Vol. 135, No. 2, 2013, p. 021003. http://dx.doi.org/10.1115/1.4023155

44 Hombal, V. K., Mullins, J., and Mahadevan, S., "Extrapolation confidence assessment for predictions of computational engineering models," Submitted to *Computer Methods in Applied Mechanics and Engineering*.

45 Jaulin, L., Kieffer, M., Didrit, O., and Walter, E., *Applied Interval Analysis*, Springer-Verlag, New York, NY, 2001.

46 Kass, R. and Raftery, A., "Bayes Factors," *Journal of the American Statistical Association*, Vol. 90, No. 430, 1995, pp. 773-795.

47 Kennedy, M. C., and O'Hagan, A., “Bayesian calibration of computer models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 5, 2001, pp. 425-464. doi:10.1111/1467-9868.00294

48 Kim, H., Venturini, G., and Strachan, A., “Molecular dynamics study of dynamical contact between a nanoscale tip and substrate for atomic force microscopy experiments,” *Journal of Applied Physics*, Vol. 112, No. 9, 2012, p. 094325. doi:10.1063/1.4762016

49 Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P., “Optimization by simulated annealing,” *Science*, Vol. 220, No. 4598, 1983, pp. 671-680. doi: 10.1126/science.220.4598.671

50 Kiureghian, A., “Aleatory or epistemic? Does it matter?,” *Structural Safety*, Vol. 31, No. 2, 2009, pp. 105-112. doi:10.1016/j.strusafe.2008.06.020

51 Klir, G. J. and Wierman, M. J., *Uncertainty-Based Information: Elements of Generalized Information Theory*, 2nd Ed., Vol. 15, Physica-Verlag, Heidelberg, DE, 1998.

52 Kokkolaras, M., Mourelatos, Z. P., and Papalambros, P. Y., “Design Optimization of Hierarchically Decomposed Multilevel Systems Under Uncertainty,” *Journal of Mechanical Design*, Vol. 128, No. 2, 2006, pp. 503–508.

53 Koslowski, M. and Strachan, A., “Uncertainty propagation in a multiscale model of nanocrystalline plasticity,” *Reliability Engineering & System Safety*, Vol. 96, No. 9, 2011, pp. 1161-1170. doi:10.1016/j.ress.2010.11.011

54 Koutsourelakis, P., “A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters, *Journal of Computational Physics*, Vol. 228, No. 17, 2009, pp. 6184-6211. doi: 10.1016/j.jcp.2009.05.016

55 Kullback, S. and Leibler, R. A., “On information and sufficiency,” *Annals of Mathematical Statistics*, Vol. 22, No. 1, 1951, pp. 76-86. doi: 10.1214/aoms/1177729694

56 Land, A. H., and Doig, A. G., "An automatic method of solving discrete programming problems," *Econometrica*, Vol. 28, No. 3, 1960, pp. 497–520. doi:10.2307/1910129

57 Ling, Y. and Mahadevan, S., “Quantitative model validation techniques: New insights,” *Reliability Engineering & System Safety*, Vol. 111, 2013, pp. 217-231. doi: 10.1016/j.ress.2012.11.011

58 Ling, Y., Mullins, J., and Mahadevan, S., “Selection of model discrepancy priors in Bayesian calibration,” *Journal of Computational Physics,* 2014, Accepted for publication. doi: 10.1016/j.jcp.2014.08.005

59 Liu, Y., Chen, W., and Arendt, P., “Toward a Better Understanding of Model Validation Metrics,” *Journal of Mechanical Design*, Vol 133, No. 7, 2011, pp. 071005. doi:10.1115/1.4004223

60 MacKay, D.J. C., *Information theory, inference, and learning algorithms*, Cambridge University Press, New York, NY, 2003.

61 Mallows, C. L., “Some Comments on Cp,” *Technometrics*, Vol. 15, No. 4, 1973, pp. 661-675.

62 McCulloch, W., and Pitts, W., “A Logical Calculus of Ideas Immanent in Nervous Activity,” *Bulletin of Mathematical Biophysics*, Vol. 5, 1943, pp. 115-133.

63 McFarland, J. M., “Uncertainty analysis for computer simulations through validation and calibration, Ph.D. thesis, Vanderbilt University, 2008.

64 McFarland, J., and Mahadevan, S., “Multivariate significance testing and model calibration under uncertainty,” *Computer Methods in Applied Mechanics and Engineering*, Vol. 197, No. 29-32, 2008, pp. 2467-2479. doi:10.1016/j.cma.2007.05.030

65 Messer, M., Panchal, J. H., Krishnamurthy, V., Klein, B, Yoder, P. D., Allen, J. K., and Mistree, F., “Model Selection Under Limited Information Using a Value-of-Information-Based Indicator,” *Journal of Mechanical Design*, Vol. 132, No. 12, 2010, pp. 121008-1-121008-13.

66 Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E., “Equation of State Calculations by Fast Computing Machines,” *Journal of Chemical Physics*, Vol. 21, No. 6, 1953, p. 1087. doi: 10.1063/1.1699114

67 Mignolet, M. P., and Spanos, P. D., "Recursive Simulation of Stationary Multivariate Random Processes: Part II," *Journal of Applied Mechanics*, Vol. 54, No. 3, 1987, pp. 681-687.

68 Montgomery, D. C., *Design and Analysis of Experiments*, Wiley, 5th Ed., 2000.

69 Mullins, J., Li, C., Sankararaman, S., Mahadevan, S., and Urbina, A, “Uncertainty Quantification Using Multi-Level Calibration and Validation Data,” *Proceedings of the 54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference,* Boston, MA, AIAA-2013-1872, 2013.

70 NASA, “Standard for Models and Simulations,” NASA-STD-7009, National Aeronautics and Space Administration, 2008.

71 Neal, R., “Slice Sampling,” *Annals of Statistics*, Vol. 31, No. 3, 2003, pp. 705-741. doi:10.1214/aos/1056562461

72 Oberkampf, W. L., Helton, J. C., Joslyn, C. A., Wojtkiewicz, S. F., and Ferson, S., “Challenge problems: uncertainty in system response given uncertain parameters,” *Reliability Engineering & System Safety,* Vol. 85, No. 1-3, 2004, pp. 11-19. doi:10.1016/j.ress.2004.03.002

73 O'Hagan, A., “Fractional Bayes Factors for Model Comparison,” *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 57, No. 1, 1995, pp. 99-138.

74 O’Hagan, A. and Oakley, J. E., “Probability is perfect, but we can’t elicit it perfectly,” *Reliability Engineering & System Safety*, Vol. 85, No. 1-3, 2004, pp. 239-248. doi:10.1016/j.ress.2004.03.014

75 Paris, P.C., Gomez, M. P., and Anderson, W.E., “A rational analytic theory of fatigue,” *The Trend in Engineering*, Vol. 13, 1961, pp. 9-14.

76 Pericchi, L. R., *Handbook of Statistics, Volume 25: Bayesian Thinking, Modeling and Computation,* 1st Ed., North Holland, 2005, Ch. 6, pp. 115-149.

77 Pilch, M., Trucano, T., Moya, J., Froehlich, G., Hodges, A., and Peercy, D., “Guidelines for Sandia ASCI Verification and Validation Plans – Content and Format: Version 2.0,” SAND2000-3101, Sandia National Laboratories, 2001.

78 Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flanner, B. P., *Numerical Recipes: The Art of Scientfic Computing*, 3rd edition, Cambridge University Press, New York, 2007.

79 Pukelsheim, F., *Optimal design of experiments*, Classics in applied mathematics, 50, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2006.

80 Queipo, N. V., Haftka, T. R., Shyy, W., Goel, T., Vaidyanathan, R., and Tucker, P. K., "Surrogate-based analysis and optimization," *Progress in Aerospace Sciences*, Vol. 41, No. 1, 2005, pp. 1-28.

81 Quinonero-Candela, J. and Rasmussen, C. E., "A unifying view of sparse approximate Gaussian process regression," *Journal of Machine Learning Research*, Vol. 6, 2005, pp. 1939-1959.

82 Rackwitz, R., and Fiessler, B., "Structural Reliability Under Combined Random Load Sequences," *Computers and Structures*, Vol. 9, No. 5, 1978, pp. 489–494. doi:10.1016/0045-7949(78)90046-9

83 Radhakrishnan, R., and McAdams, D. A., "A Methodology for Model Selection in Engineering Design," *Journal of Mechanical Design*, Vol. 127, No. 3, 2005, pp. 378-387.

84 Rasmussen, C. E. and Williams, C. K. I., *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, MA, 2006.

85 Rebba, R. and Mahadevan, S., "Computational methods for model reliability assessment," *Reliability Engineering & System Safety*, Vol. 93, No. 8, 2008, pp. 1197-1207. doi:10.1016/j.ress.2007.08.001.

86 Rebba, R. and Mahadevan, S., "Validation of models with multivariate output, *Reliability Engineering & System Safety*, Vol. 91, No. 8, 2006, pp. 861-871, doi: 10.1016/j.ress.2005.09.004

87 Rebba, R., Mahadevan, S., and Huang, S., "Validation and error estimation of computational models," *Reliability Engineering & System Safety*, Vol. 91, No. 10-11, 2006, pp. 1390-1397. doi:10.1016/j.ress.2005.11.035.

88 Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W., "Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors," *Water Resources Research*, Vol. 46, No. 5, 2010, p. W05521. doi: 10.1029/2009WR008328

89 Robert, C. P. and Casella, G., *Monte Carlo statistical methods*, 2nd ed., Springer, New York, 2004.

90 Robinson, T. D., Eldred, M. S., Willcox, K. E., and Haimes, R., "Surrogate-based optimization using multifidelity models with variable parameterization and corrected space mapping," *AIAA journal*, Vol. 46, No. 11, 2008, pp. 2814-2822.

91 Romero, V. J., "Elements of a Pragmatic Approach for dealing with Bias and Uncertainty in Experiments through Predictions," SAND2011-7342, Sandia National Laboratories, 2011.

92 Romero, V., Luketa, A., and Sherman, M., "Application of a Versatile 'Real-Space' Validation Methodology to a Fire Model, *Journal of Thermophysics and Heat Transfer*, Vol. 24, No. 4, 2010, pp. 730-744. doi: 10.2514/1.46358

93 Rosenblatt, M., "Remarks on a Multivariate Transformation," *Annals of Mathematical Statistics*, Vol. 23, No. 3, 1952, pp. 470–472. doi:10.1214/aoms/1177729394

94 Ross, T. J., *Fuzzy Logic with Engineering Applications*, McGraw-Hill, New York, NY, 1995.

95 Roy, C. and Oberkampf, W., "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing," *Computer Methods in Applied Mechanics and Engineering*, Vol. 200, No. 25-28, 2011, pp. 2131-2144. doi: 10.1016/j.cma.2011.03.016

96 Sacks, J., Schiller, S. B., and Welch, W., “Design for Computer Experiments,” *Technometrics*, Vol. 31, No. 1, 1989, pp. 41–47. doi:10.2307/1270363

97 Sankararaman, S. and Mahadevan, S., “Assessing the reliability of computational models under uncertainty,” *Proceedings of the 54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Boston, MA, AIAA-2013-1873, 2013.

98 Sankararaman, S., and Mahadevan, S., “Comprehensive framework for integration of calibration, verification and validation,” *Proceedings of the 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Honolulu, HI, AIAA-2012-1366, 2012.

99 Sankararaman, S. and Mahadevan, S., “Separating the contributions of variability and parameter uncertainty in probability distributions,” *Reliability Engineering & System Safety*, Vol. 112, 2013, pp. 187-199. doi: 10.1016/j.ress.2012.11.024

100 Sankararaman, S., McLemore, K., and Mahadevan, S., “Test Resource Allocation in Hierarchical Systems Using Bayesian Networks,” *AIAA Journal*, Vol. 51, No. 3, 2013, pp. 537-550. doi: 10.2514/1.J051542

101 Shafer, G., *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ, 1976.

102 Schwarz, G. E., “Estimating the dimension of a model,” *Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461-464. doi: 10.1214/aos/1176344136

103 Shinozuka, M., and Deodatis, G., "Simulation of stochastic processes by spectral representation," *Applied Mechanics Review*, Vol. 44, No. 4, 1991, pp. 191–204.

104 Skanda, D. and Lebiedz, D., “An optimal experimental design approach to model discrimination in dynamic biochemical systems,” *Bioinformatics*, Vol. 26, No. 7, 2010, pp. 939-945. doi: 10.1093/bioinformatics/btq074

105 “Student" [Gosset, W.S.], "The probable error of a mean," *Biometrika*, Vol. 6, No. 1, 1908, pp. 1–25. doi:10.1093/biomet/6.1.1

106 Tommasi, C. and Lopez-Fidalgo, J., “Bayesian optimum designs for discriminating between models with any distribution,” *Computational Statistics & Data Analysis*, Vol. 54, No. 1, 2010, pp. 143-150. doi: 10.1016/j.csda.2009.07.022

107 Trucano, T., Swiler L., Igusa, T., Oberkampf, W., and Pilch, M., “Calibration, validation, and sensitivity analysis: What’s what,” *Reliability Engineering & System Safety*, Vol. 91, No. 10-11, 2006, pp. 1331-1357. doi: 10.1016/j.ress.2005.11.031

108 Wang, S., Chen, W., and Tsui, K.-L., “Bayesian Validation of Computer Models,” *Technometrics*, Vol. 51, No. 4, 2009, pp. 439-451. doi: 10.1198/TECH.2009.07011

109 Weiss, N. A., *A Course in Probability*, Addison-Wesley, Boston, MA, 2005.

110 Wu, Y.-T., “An Adaptive Importance Sampling Method for Structural System Reliability Analysis and Design,” Reliability Technology 1992, *Proceedings of ASME Winter Annual* Meetings, Vol. AD-28, 1992, pp. 217–231.

111 Xiu, D., *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, New Jersey, 2010.

112 Zou, T. and Mahadevan, S., “A Direct Decoupling Approach for Efficient Reliability-Based Design Optimization,” *Structural and Multidisciplinary Optimization*, Vol. 31, No. 3, 2006, pp. 190-200.