UNCERTAINTY QUANTIFICATION IN TIME-DEPENDENT RELIABILITY ANALYSIS

By

You Ling

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, 2013
Nashville, Tennessee

Approved:

Professor Sankaran Mahadevan
Professor Prodyot K. Basu
Professor Professor Jayathi Y. Murthy
Professor Professor Caglar Oskay
Professor Ronald D. Schrimpf

To my mother, Yue Zhang

ACKNOWLEDGMENTS

I would like to express my greatest gratitude to my advisor Professor Sankaran Mahadevan. His insight, vision, and vast experience in research have been extremely helpful throughout my graduate career. I can never thank him enough for showing me how to become a researcher whose work can make real impact to both academia and industry. I am indebted to him for his encouragement and enormous support with my postgraduate career plan.

I would like to thank my committee members, Professor Prodyot Basu, Professor Jayathi Murthy, Professor Caglar Oskay, and Professor Ronald Schrimpf, for their kindly support and wise advice for my research and career development.

I would also like to thank our research collaborators, Professor Alejandro Strachan, Dr. Lin Sun, Professor Alina Alexeenko's group, Prof. Muhammad Alams group, Prof. Jayathi Murthys group, Prof. Dimitrios Perouliss group, Prof. Marisol Koslowski, Dr. Martin Hunt, and many others in the Center for Prediction of Reliability, Integrity and Survivability of Microsystem (PRISM) at Purdue University for sharing simulation codes, experimental data, and valuable discussions. I also want to thank Professor Yongming Liu's group at Clarkson University (now at Arizona State University) for discussions on fatigue modeling and testing.

I am grateful for the opportunity to meet so many brilliant people at Vanderbilt University. In particular, I am very fortunate to have Dr. Shankar Sankararaman as a student mentor when I started this research. He was always willing to offer me guidance and even hands-on help. Working closely with him made my rookie year much more productive yet less painful. I would like to thank Dr. Christopher Shantz and Dr. Vadiraj Hombal for their generous sharing of technical knowledge on finite element modeling and Gaussian process modeling. It has also been a joy for me to work with Joshua Mullins, with whom I had many interesting discussions. I also acknowledge the valuable weekly meetings with all the students and scholars in the Reliability and Risk Engineering program.

This dissertation would not have been possible without the unconditional love and support from my mother, Yue Zhang. Also, many friends have offered me warm friendship and kindly support during my five-year stay in Nashville. My deep appreciations go to Dr. Yang Yang, Bin Liang, Tong Hui, David Morse, Chen Liang, Chenzhao Li, and many others.

The research was supported by funding from several sources: (1) the NASA ARMD/AvSP IVHM project under NRA Award NNX09AY54A (Project Monitor: Dr. K. Goebel, NASA AMES Research Center) through subcontract to Clarkson University (No. 375-32531, Principal Investigator: Dr. Y. Liu), (2) the Federal Aviation Administration William J. Hughes Technical Center through RCDT Project No. DTFACT-06-R-BAAVAN1 (Project Monitors: Dr. John Bakuckas, Ms. Traci Stadtmueller), and (3) the U.S. Department of Energy [National Nuclear Security Administration] Predictive Science Academic Alliance Program under Award Number DE-FC52-08NA28617 to Purdue University (Principal Investigator: Prof. Jayathi Murthy, Deputy Director: Prof. Alejandro Strachan) with subaward to Vanderbilt University. The support is gratefully acknowledged. The numerical studies were conducted using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, and the Rosen Center for Advanced Computing (RCAC) at Purdue University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

1.1 Overview

Time-dependent reliability analysis of engineering components/devices is a challenging interdisciplinary problem, which involves many major research areas, such as multi-physics modeling and experiments, numerical analysis, probability theory, and statistical inference. An important part of reliability analysis is to propagate forward uncertainty from different sources to the performance prediction of target components/devices. Several analytical and numerical methods have been developed for uncertainty propagation in reliability analysis, including first-order reliability methods (FORM), second-order reliability methods (SORM), out-crossing methods, surrogate model-based methods, and Monte Carlo simulation-based methods [Melchers, 1999; Mahadevan and Dey, 1997; Kuschel and Rackwitz, 2000]. No matter which method is chosen, the accuracy of reliability prediction is significantly affected by uncertainty quantification (UQ), which is the process of quantifying uncertainty at different levels (e.g., input level, modeling level, and prediction level) of scientific computing.

Two categories of uncertainty, aleatory uncertainty (uncertainty due to natural variability) and epistemic uncertainty (uncertainty due to lack of knowledge/information), have been identified and widely accepted in the UQ community [O'Hagan and Oakley, 2004; McFarland and Mahadevan, 2008; Der Kiureghian, 2009]. In our recent research, epistemic uncertainty is further divided into two types, namely model uncertainty (uncertainty due to the use of approximate models) and data uncertainty (uncertainty due to the use of insufficient or imprecise data) [Sankararaman et al., 2011b,a]. Very limited amount of research have aimed at developing a comprehensive UQ framework [McFarland, 2008; Lucas et al., 2008;

Roy and Oberkampf, 2011; Sankararaman et al., 2011a] that integrates multiple types of uncertainty. Most of the reliability research has only focused on quantifying the natural variability at the input level, and then estimate the uncertainty at the output level via uncertainty propagation [Beran et al., 2006; Maître et al., 2007; Wojtkiewicz et al., 2001; Najm, 2009; Reagan et al., 2003; Guo et al., 2010].

Within the UQ community, three types of activities are pursued to address different aspects of model uncertainty: (1) model calibration [Kennedy and O'Hagan, 2001], which aims to quantify the uncertainty in the estimation of model parameters, (2) model validation [Oberkampf and Trucano, 2004; Rebba et al., 2006; Ling and Mahadevan, 2013b], which is used to quantify the uncertainty and confidence in model prediction, and (3) model verification [Babuska and Oden, 2004; ASME, 2006; Rangavajhala et al., 2011], which focuses on solution approximation error. Probabilistic methods to address data uncertainty in engineering problems are also being actively pursued, and some recent development can be found in [Zaman et al., 2010; Sankararaman and Mahadevan, 2011a, 2012b]. With the rapid development in UQ activities and fundamental methods, the need to obtain more rigorous prediction of reliability accounting for multiple sources of uncertainty has emerged, and it has become more feasible to perform time-dependent reliability analysis with the inclusion of UQ results.

1.2 Research objectives

The overall objective of the proposed research is to develop a rigorous and efficient framework for integrating time-dependent reliability analysis with comprehensive uncertainty quantification that considers natural variability, data uncertainty, and model uncertainty. Four tasks are pursued in order to achieve this overall objective. The first task focuses on Bayesian calibration of computer simulation models that are used in time-dependent reliability analysis using available information, which aims at the quantification of uncertainty at the modeling level (including uncertainty in model parameters and model form), and the uncertainty introduced by the use of insufficient data in model calibration. The second task is to study and develop quantitative model validation methods, that quantify the confidence of models in the domain of intended use (e.g., prediction domain) and guide decision making in model selection. The third task is to explore efficient and accurate methods for uncertainty propagation and the calculation of failure probability in time-dependent problems. The fourth task investigates methods to include data collected on a system in service (e.g., load data, inspection data) in the reliability analysis. Two application examples: (1) long-term life prediction of a radio frequency (RF) micro-electromechanical system (MEMS) switch, and (2) fatigue crack growth analysis of a rotorcraft mast component, are developed to illustrate the overall uncertainty framework proposed in this research.

1.3 Organization of the dissertation

The subsequent chapters of this dissertation will be devoted to the objectives proposed above.

Chapter II provides an introduction to the state-of-the-art tools and methods developed for uncertainty quantification and time-dependent reliability analysis, including: (1) Bayesian networks, (2) Bayesian model calibration, (3) quantitative model validation, (4) PHI2 algorithm for time-dependent reliability estimation, and (5) surrogate modeling techniques.

Chapter III addresses several challenging issues in the application of Bayesian model calibration to realistic problems: (1) appropriate formulation of model discrepancy function to account for model form uncertainty, (2) identifiability of model parameters in an inverse problem, and (3) calibration strategy for multiple physics models with shared parameters. Methods to address these issues are presented in detail in this chapter.

Based on the various quantitative model validation methods illustrated in Chapter II, Chapter IV discusses possible scenarios of model validation given data from three types of experiments. Extensions to Bayesian hypothesis testing-based method are developed, and the relationships and interpretations of the existing methods are investigated.

Chapter V develops a Bayesian network-based uncertainty quantification framework for time-dependent reliability analysis, with application to the life prediction of a MEMS device, which is a multi-physics problem involving elasticity, electrostatics, creep, gas damping, and contact. We construct a Bayesian network in order to integrate multiple physics models and the corresponding experimental data. Various physics models at different levels of the system are also validated using the quantitative model validation techniques discussed in Chapter IV. The results of calibration and validation are then included within time-dependent reliability prediction of the MEMS device.

Chapters VI and VII develop a probabilistic framework to include data regarding a system in service in the reliability analysis, with application to fatigue crack growth prognosis of a rotorcraft component.

Chapter VI explores various modeling choices of fatigue load history, and statistical updating of load models using real-time monitoring data is outlined. The highlight of this chapter is the development of an ARIMA model-based approach, which integrates the ARIMA modeling technique, Bayesian calibration, and Bayesian model averaging in order to quantify various sources of uncertainty in applied loading.

Chapter VII considers the integration of health monitoring data, in particular, crack inspection data into the prediction of the remaining useful life of the rotorcraft component. Various sources of uncertainty are identified using global sensitivity analysis. Gaussian process surrogate modeling technique is used to simulate non-planar crack growth under variable amplitude and multi-axial loading. A Bayesian network is constructed to integrate fatigue-related models and system monitoring data. Bayesian calibration is applied to infer the probability distribution of equivalent initial crack size, and the ARIMA model-based approach is applied to characterize and update the applied load history. Fatigue prognosis of the rotorcraft components under different scenarios of system monitoring data is then performed, by integrating the information from all the models and available data through the Bayesian network.

CHAPTER II

BACKGROUND

This chapter starts with a brief introduction to Bayesian networks, a platform we use throughout this dissertation to integrate various sources of information in a multi-level multi-physics system. Two main research areas of uncertainty quantification, model calibration and validation, are reviewed. The basic methodology of Bayesian model calibration is given in Section 2.2. Section 2.3 presents the currently available model validation techniques. Detailed discussions and further developments are provided in Chapters III and IV. The theory of time-dependent reliability analysis is given in Section 2.4 along with a brief introduction to the PHI2 algorithm. Surrogate modeling techniques that can expedite time-dependent reliability analysis are discussed in Section 2.5.

2.1 Brief introduction to Bayesian networks

A BN is a directed acyclic graph formed by the variables (nodes) together with the directed edges, attached by a table of conditional probabilities of each variable on all its parents [Jensen and Nielsen, 2007]. Therefore, it can be used as a graphical representation of uncertain quantities, and explicitly incorporates the probabilistic causal dependence between the variables. Following certain rules, BN allows information to pass from component level to system level, which makes it a powerful tool in system reliability analysis [Torres-Toledano and Sucar, 1998; Mahadevan et al., 2001; Weber and Jouffe, 2006; Langseth and Portinale, 2007; Straub and Der Kiureghian, 2010; Bensi, 2010].

Based on the chain rule in probability theory, a BN consisting of a set of nodes $U =$

$\{X_1, X_2, ..., X_n\}$ presents the joint probability distribution of these nodes as

$$p(U) = \prod_{i=1}^{n} p(X_i|\mathrm{Pa}_{X_i}) \tag{2.1}$$

where Pa_{X_i} is the set of parent nodes of the node X_i, which can be identified based on the directed edges in the graph. The conditional probability of X_i given its parent nodes, $p(X_i|\mathrm{Pa}_{X_i})$, can be found in the conditional probability table.

If there are observation data D available for a subset (U_{obs}) of the nodes, the joint probability distribution of the other nodes $\bar{U}_{obs}$ in the BN can be updated based on Bayes' theorem

$$p(\bar{U}_{obs}|U_{obs} = D) = \frac{p(\bar{U}_{obs}, U_{obs} = D)}{\int p(\bar{U}_{obs}, U_{obs} = D)\ \mathrm{d}\bar{U}_{obs}} \tag{2.2}$$

Note that the integration in the denominator of Eq. 2.2 is valid if $\bar{U}_{obs}$ contains only continuous random variables and $p(*)$ stands for probability density function. In the case of discrete random variables, the integration should be replaced by the summation of probability mass function. Eq. 2.2 provides the basis for model calibration which will be discussed in Section 2.2.

In addition to updating the unobserved nodes, forward prediction of the quantity of interest (QoI) can also be performed in the BN. For example, if QoI is in the set $\bar{U}_{obs}$, the posterior probability distribution of QoI can be obtained by marginalizing the joint posterior distribution $p(\bar{U}_{obs}|U_{obs} = D)$ in Eq. 2.2.

2.2 Bayesian model calibration

Model calibration is the process of adjusting unknown model parameters in order to improve the agreement between model output and observed data [Campbell, 2006], and it is a widely used technique in the uncertainty analysis of engineering systems [McFarland, 2008]. Kennedy and O'Hagan [Kennedy and O'Hagan, 2001] developed a Bayesian calibration framework (commonly known as the KOH framework), which is capable of including various sources of uncertainty. One of the main features of this framework is the use of a model discrepancy function to explicitly account for the uncertainty due to model inadequacy. It has been shown that by including an appropriate model discrepancy function in calibration, bias and overfitting in the estimation of physical parameters can be mitigated or avoided [Brynjarsdottir and O'Hagan, 2013]. Based on the KOH framework, efficient and rigorous calibration methods have been developed for realistic problems [Higdon et al., 2008; Koutsourelakis, 2009; Arendt et al., 2012; Ling and Mahadevan, 2013a], and practical applications can be found in many areas of science and engineering, such as environmental management [Arhonditsis et al., 2008], heat transfer [McFarland and Mahadevan, 2008], astronomy [Bower et al., 2010], hydrology [Renard et al., 2010], geochemistry [Sarkar et al., 2012], fatigue [Sankararaman et al., 2011b], and aerothermal modeling [DeCarlo et al., 2013].

This dissertation will focus on model calibration with direct measurement data of the model output variable. In the case that the available information are the moments of the probability distributions of the model output variable, some recently developed methodologies based on optimization with constraints on the moments may be considered [Zabaras and Ganapathysubramanian, 2008; Guan et al., 2009]. In addition, a Bayesian approach has been developed to include the information on the moments of unknown model parameters [Berry et al., 2012].

Section 2.2.1 illustrates the basic framework of model calibration. In Sections 2.2.2 and 2.2.3, two types of experimental data are considered, namely interval data and time series data, and the corresponding details of calibration are developed. Some discussions on computing likelihood functions, which can be computationally expensive for complex systems, are provided in Section 2.2.4.

2.2.1 Basic methodology

Consider a computer model, $y_m = G(\boldsymbol{x}; \boldsymbol{\theta})$, with input $\boldsymbol{x}$, parameter $\boldsymbol{\theta}$, and output y_m. This model is constructed to predict a physical quantity of interest y, which is observable through experiments. The model input $\boldsymbol{x}$ is a set of quantities that modelers consider unnecessary to calibrate, since these quantities can be either measured in experiments or computed from other physics models. Therefore, the model input $\boldsymbol{x}$ is set to be $\boldsymbol{x}_D$ in the process of model calibration, where $\boldsymbol{x}_D$ is a set of measured/computed values of $\boldsymbol{x}$. Note that in the presence of measurement error in experiments, or other sources of uncertainty in the models used to compute $\boldsymbol{x}$, some elements of $\boldsymbol{x}_D$ may be treated as random variables with known probability distributions. In contrast to $\boldsymbol{x}$, the model parameter set $\boldsymbol{\theta}$ is considered unknown due to the lack of experimental data or physical knowledge, and the objective of model calibration is to estimate these parameters based on available information.

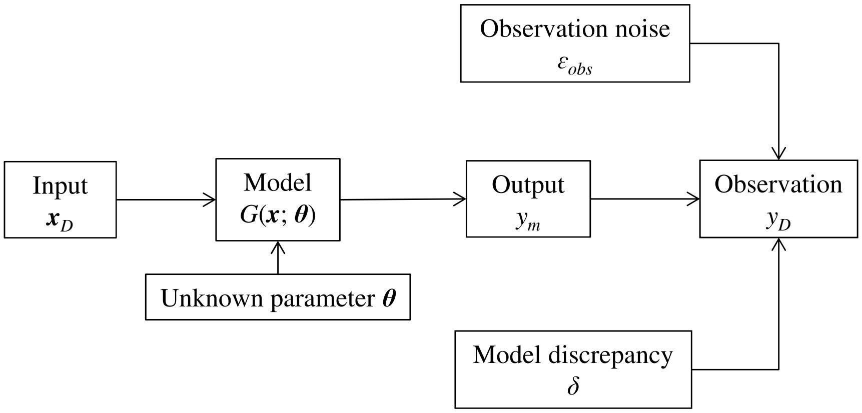


Figure 2.1: Relationship between a computer model and corresponding experimental observation

In the KOH framework, the relationship between experimental observation y_D, true value of the quantity y, and model output y_m is described as (Fig. 2.1)

$$\begin{aligned} y_D &= y + \varepsilon_{obs} \quad (2.3) \\ y &= y_m + \delta = G(\boldsymbol{x}_D, \boldsymbol{\theta}) + \delta(\boldsymbol{x}_D) \quad (2.4) \end{aligned}$$

where ε_{obs} represents measurement uncertainty and is often treated as a zero-mean Guassian random variable with variance σ_{obs}^2. Uncertainty due to model inadequacy is represented by a model discrepancy term δ, which could be a function of model input $\boldsymbol{x}$. Since σ_{obs} and δ are usually unknown, they may also need to be calibrated. Note that y_D is treated as a random variable in Eq. 2.3, and the samples of y_D are the actual observation data of y. Assuming m samples of y_D (denoted as $\boldsymbol{D} = [D_1, D_2, ..., D_m]$) are collected for a single input setting $\boldsymbol{x}_D$, we can calibrate the unknown parameters $\boldsymbol{\theta}$, σ_{obs}, and δ using Bayes' theorem as

$$\begin{aligned} \pi(\boldsymbol{\theta}, \sigma_{obs}, \delta | \boldsymbol{D}) &= \frac{\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta)\ \pi(\boldsymbol{\theta})\ \pi(\sigma_{obs})\ \pi(\delta)}{\int \mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta)\ \pi(\boldsymbol{\theta})\ \pi(\sigma_{obs})\ \pi(\delta)\ \mathrm{d}\boldsymbol{\theta}\ \mathrm{d}\sigma_{obs}\ \mathrm{d}\delta} \\ \mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta) &\propto \prod_{i=1}^{m} \pi(y_D = D_i | \boldsymbol{x_D}, \boldsymbol{\theta}, \sigma_{obs}, \delta) \quad (2.5) \end{aligned}$$

where $\pi(\boldsymbol{\theta})$, $\pi(\sigma_{obs})$, $\pi(\delta)$ are the prior PDFs of $\boldsymbol{\theta}$, σ_{obs}, and δ respectively, representing prior knowledge of these parameters before calibration; $\pi(\boldsymbol{\theta}, \sigma_{obs}, \delta | \boldsymbol{D})$ is the joint posterior (or calibrated) PDF of $\boldsymbol{\theta}$, σ_{obs}, and δ; the joint likelihood function of $\boldsymbol{\theta}$, σ_{obs}, and δ, which is denoted as $\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta)$, is proportional to the conditional probability of observing the data $\boldsymbol{D}$ given these parameters. Note that $\pi(*)$ denotes probability density function in this paper.

In the case where experimental observations are taken for multiple input settings, paired data $\{\boldsymbol{x}_{Di}, y_{Di}\}_{i=1}^{n}$ may be available. Let $\boldsymbol{X}_D = [\boldsymbol{x}_{D1}, \boldsymbol{x}_{D2}, ..., \boldsymbol{x}_{Dn}]$, and the corresponding output observation $\boldsymbol{y}_D = [y_{D1}, y_{D2}, ..., y_{Dn}]$. Note that $\boldsymbol{y}_D$ becomes a random vector. Assuming that m realizations fo $\boldsymbol{y}_D$ (denoted as $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_m]$) are collected in the experiments, the likelihood function in Eq. 2.5 becomes

$$\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta) \propto \prod_{i=1}^{m} \pi(\boldsymbol{y}_D = \boldsymbol{D}_i | \boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta) \tag{2.6}$$

where $\pi(\boldsymbol{y}_D = \boldsymbol{D}_i | \boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta)$ is the joint PDF of the random vector $\boldsymbol{y}_D$ evaluated at $\boldsymbol{D}_i$, and this probability distribution is conditioned on $\boldsymbol{X}_D$, $\boldsymbol{\theta}$, σ_{obs}, and δ.

In this section, the likelihood function is constructed based on measurement data reported as point values. In practical problems, various types of experimental data may be available, including interval data and time series data. For these different types of data, which will be discussed in the subsequent sections.

2.2.2 Bayesian calibration with interval data

Due to the imprecision of measurement techniques and limited experimental resources, the data of many quantities is only available in the form of an interval, which brings in additional data uncertainty (i.e., the actual experimental value lies within an interval). Sankararaman and Mahadevan [Sankararaman and Mahadevan, 2011a] developed a likelihood-based approach to quantify this type of uncertainty. In the example shown in Fig. 2.1, the experimental data D may be reported as an interval, i.e., $D = [D^a, D^b]$, and we can derive the corresponding expression for the likelihood function of unknown parameters based on the method developed in [Sankararaman and Mahadevan, 2011a] as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta) \;\propto\;& \Pr(D^a \le y_D \le D^b | \boldsymbol{\theta}, \sigma_{obs}, \delta) \\ =& \int_{D^a}^{D^b} \pi(y_D | \boldsymbol{x}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta)\ \pi(\boldsymbol{x}_D)\ \mathrm{d}y_D\ \mathrm{d}\boldsymbol{x}_D \end{aligned} \tag{2.7}$$

If data are in the form of half intervals, i.e., $y_D \ge D^a$, the likelihood function can be obtained by letting $D^b = +\infty$ in Eq. 2.7. Similarly, let $D^a = -\infty$ if $y_D \le D^b$.

In some problems, measurements may be available at multiple input settings $\boldsymbol{X}_D =$

$[\boldsymbol{x}_{D1}, \boldsymbol{x}_{D2}, ...\boldsymbol{x}_{Dn}]$, and the data may be in the form of multiple intervals or a mixture of intervals and point values. We can conveniently extend Eq. 2.7 to these two cases. Suppose the available data are now a set of intervals, i.e., $\boldsymbol{D} = \{[D_1^a, D_1^b], [D_2^a, D_2^b], ..., [D_n^a, D_n^b]\}$, which forms a n-dimensional hypercube Ω_n. The probability of observing the data is thus equivalent to the probability of the n-dimensional random vector $\boldsymbol{y}_D = [y_{D1}, y_{D2}, ..., y_{Dn}]$ lying inside the hypercube Ω_n. Hence, the likelihood function of unknown parameters can be derived as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta) &\propto \Pr(\boldsymbol{y}_D \in \Omega_n | \boldsymbol{\theta}, \sigma_{obs}, \delta) \\ &= \int_{\Omega_n} \pi(\boldsymbol{y}_D | \boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta)\ \pi(\boldsymbol{X}_D)\ \mathrm{d}\boldsymbol{y}_D\ \mathrm{d}\boldsymbol{X}_D \end{aligned} \tag{2.8}$$

In the case that the available information is a mixture of k intervals and $(n-k)$ point values, the k intervals form a k-dimensional hypercube Ω_k. Let $\boldsymbol{y}_{D,k}$ represent the elements of the random vector $\boldsymbol{y}_D$ corresponding to interval data, and $\boldsymbol{y}_{D,n-k}$ represent the rest of the elements corresponding to point data $\boldsymbol{D}_{point} = [D_{n-k+1}, D_{n-k+2}, ..., D_n]$. The likelihood function $\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta)$ can be derived as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta) \propto \Pr\left((\boldsymbol{y}_{D,k} \in \Omega_k) \cap (\boldsymbol{y}_{D,n-k} = \boldsymbol{D}_{point}) | \boldsymbol{\theta}, \sigma_{obs}, \delta\right) \\ = \int_{\Omega_k} \pi(\boldsymbol{y}_{D,k} | \boldsymbol{y}_{D,n-k} = \boldsymbol{D}_{point}, \boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta)\ \pi(\boldsymbol{X}_D)\ \mathrm{d}\boldsymbol{y}_{D,k}\ \mathrm{d}\boldsymbol{X}_D \end{aligned} \tag{2.9}$$

where $\pi(\boldsymbol{y}_{D,k} | \boldsymbol{y}_{D,n-k} = \boldsymbol{D}_{point}, \boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta)$ is obtained by substituting $\boldsymbol{y}_{D,n-k} = \boldsymbol{D}_{point}$ into the joint PDF of the random vector $\boldsymbol{y}_D$.

2.2.3 Bayesian calibration with time series data

In dynamic systems, information is commonly available in the form of time series data. This type of information leads to several additional challenges; in particular, the model prediction and the corresponding measurement data are dependent on the states of the system in the previous time steps, and replicate time series observations may be taken with a large number of time points. Both of these characteristics may complicate the computation of the likelihood function. To perform model calibration with time-series data, we again use the KOH framework discussed in Section 2.2.1. In general, a dynamic model can be written as $y_{m,t} = G(\boldsymbol{y}_{m,-t}, \boldsymbol{x}, t; \boldsymbol{\theta})$; $y_{m,t}$ represents the model prediction at time t; $\boldsymbol{y}_{m,-t}$ represents the model predictions for the previous time steps; $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are the same as in Section 2.2.1. Note that $y_{m,t}$ is deterministic for given values of $\boldsymbol{y}_{m,-t}$, $\boldsymbol{x}$, t and $\boldsymbol{\theta}$. The model discrepancy function may also become time dependent as δ_t. To simplify the problem, we assume that δ_t can be parameterized as a function with time-invariant coefficients $\boldsymbol{\phi}$, and thus the problem of calibrating δ_t is converted to the inference of a finite number of unknown coefficients. The measurement uncertainty is still represented as $\epsilon_{obs} \sim \mathcal{N}(0, \sigma_{obs}^2)$, which is time-independent. Similar to Eqs. 2.3 and 2.4, the relationship between experimental observation $y_{D,t}$ and the corresponding model prediction $y_{m,t}$ can be written as

$$y_{D,t} = y_{m,t} + \delta_t + \varepsilon_{obs} \tag{2.10}$$

The fact that $y_{m,t}$ is dependent on $\boldsymbol{y}_{m,-t}$ renders the construction of the likelihood function $\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \boldsymbol{\phi})$ difficult. For example, suppose we have one set of time-series data available $\boldsymbol{D}_t = [D_{t1}, D_{t2}, ..., D_{tn}]$, i.e., the measurements are taken at several time points $\boldsymbol{t}_D = t_1, t_2, ..., t_n$ when the values of the experimental inputs $\boldsymbol{x}_D$ remain the same. Note that again we consider that the actual data $\boldsymbol{D}_t$ are random realizations of the random vector $\boldsymbol{y}_{D,t} = [y_{D,t1}, y_{D,t2}, ..., y_{D,tn}]$. The corresponding likelihood function can be written as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta_t) \propto \pi(\boldsymbol{y}_{D,t} = \boldsymbol{D}_t | \boldsymbol{\theta}, \sigma_{obs}, \boldsymbol{\phi}) \\ = &\int \Big(\pi(\boldsymbol{y}_{D,t} = \boldsymbol{D}_t | \boldsymbol{y}_{m,-t}, \boldsymbol{x}_D, \boldsymbol{\theta}, \sigma_{obs}, \boldsymbol{\phi}) \\ &\pi(\boldsymbol{y}_{m,-t} | \boldsymbol{x}_D, \boldsymbol{\theta}) \; \pi(\boldsymbol{x}_D) \Big) \, \mathrm{d}\boldsymbol{y}_{m,-t} \, \mathrm{d}\boldsymbol{x}_D \end{aligned} \tag{2.11}$$

Note that if strictly written, $\boldsymbol{y}_{m,-t}$ in Eq. 2.11 should be different for different data points. We chose not to write out each "$\boldsymbol{y}_{m,-t}$" to avoid making the equation unnecessarily complex. It can be seen that the PDF of model predictions for the past time points $\pi(\boldsymbol{y}_{m,-t}|\boldsymbol{x}_D, \boldsymbol{\theta})$ and an integration over all the elements of $\boldsymbol{y}_{m,-t}$ are needed, which makes the evaluation of the likelihood function analytically intractable, and numerically expensive methods (such as Monte Carlo simulation) may be needed.

The extension to the cases where multiple sets of time series data are available is straightforward in theory. The difference from the case of single time series data is that $\boldsymbol{y}_{D,t}$, $\boldsymbol{y}_{m,-t}$ and $\boldsymbol{x}_D$ in Eq. 2.11 become matrices instead of being vectors. In the special case that the multiple series are replicates, i.e., we have repeated measurements at the same time points for the same set of inputs, the variation from one series to another can be attributed to the observation noise ε_{obs}, and thus we can directly compute ε_{obs} based on these repeated time series. Assuming that ε_{obs} is a zero-mean Gaussian random variable with variance σ_{obs}^2 independent of time t, the variance σ_{obs}^2 can be estimated as

$$\sigma_{obs}^2 = \frac{1}{n_1(n_2-1)} \sum_{j=1}^{n_1} \sum_{i=1}^{n_2} [y_{D,t_j}^i - \frac{1}{n_2} \sum_{i=1}^{n_2} (y_{D,t_j}^i)]^2 \qquad (2.12)$$

where n_2 is the number of repeated time series, and n_1 is the number of time points in each series. The estimated σ_{obs} and the average time series $(1/n_2)\sum_{i=1}^{n_2}(y_{D,t_j}^i)$ can be further used to compute the likelihood as in the single time series case.

If the measurement uncertainty in the experimental inputs is negligible, $\boldsymbol{x}_D$ can be treated as constant. Note that $\boldsymbol{y}_{m,-t}$ is deterministic for given values of $\boldsymbol{x}_D$ and $\boldsymbol{\theta}$ (initial condition is also considered as input). Then, the calibration with time series data $\boldsymbol{D}_t$ becomes much simpler, as Eq. 2.11 can be simplified as

$$\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \boldsymbol{\phi}) \propto \pi(\boldsymbol{y}_{D,t} = \boldsymbol{D}_t | \boldsymbol{x}_D, \boldsymbol{\theta}, \sigma_{obs}, \boldsymbol{\phi}) \qquad (2.13)$$

where $\pi(\boldsymbol{y}_{D,t} = \boldsymbol{D}_t | \boldsymbol{x}_D, \boldsymbol{\theta}, \sigma_{obs}, \boldsymbol{\phi})$ can be evaluated once the model discrepancy is formulated as shown in Section 3.2.1.

In the case that the time series data set is becoming available in real time during the operation of a system, i.e., the observation is made in real time and the model is used to make predictions for future time points, we can continuously calibrate (or update) the model using algorithms such as Kalman filter (for linear models), extended Kalman filter (for non-linear models), particle filter (sampling implementation of sequential Bayesian calibration) [Thrun et al., 2005], etc.

2.2.4 Computational issues

Bayesian calibration in Eq. 2.5 can be computationally expensive due to two reasons: (1) the likelihood function may be expensive to compute numerically, and (2) the multivariate integration in the denominator of Eq. 2.5 can be time consuming if the number of parameters is large.

Adding a model discrepancy term to the original model can lead to a high dimensional-parameter space, as a set of coefficients which parameterize the model discrepancy needs to be estimated in addition to the actual physics model parameters $\boldsymbol{\theta}$. For the case that the data points are sparse, it may not be feasible to calibrate the model along with the estimation of these parameters of model discrepancy. A compromised solution is to use a simplified model discrepancy function with less flexibility, i.e., a smaller number of parameters. Another possible method is to estimate the model parameters $\boldsymbol{\theta}$ and the coefficients of δ in two sequential steps. First, the model parameters $\boldsymbol{\theta}$ are calibrated without considering model discrepancy. Then, we can estimate the coefficients of δ based on the *a posteriori* estimate of $\boldsymbol{\theta}$ (denoted as $\boldsymbol{\theta}^*$), i.e., we can obtain the posterior PDF of the coefficients of δ which is conditioned on $\boldsymbol{\theta}^*$.

The likelihood function represents the probabilistic relationship between measured data and unknown parameters, and repeated runs of the computer model $G(\boldsymbol{x};\boldsymbol{\theta})$ are required to compute this relationship. Hence, previous studies have mostly focused on approximating the computational model with a surrogate model [Kennedy and O'Hagan, 2001; Sankararaman et al., 2010], i.e., replacing the physics-based model $G(\boldsymbol{x};\boldsymbol{\theta})$ with a faster model without losing much accuracy. Surrogate modeling techniques that have been developed in literature include Kriging or Gaussian Process (GP) interpolation [Rasmussen and Williams, 2006], polynomial chaos expansion [Xiu and Karniadakis, 2002; Ghanem and Spanos, 2003; Marzouk and Najm, 2009], support vector machine (SVM) [Vapnik, 1999], relevance vector machine [Tipping, 2001], adaptive sparse grid collocation [Ma and Zabaras, 2009], etc. Then, the likelihood function of the parameters can be evaluated based on executing the surrogate model a number of times.

If the measurement uncertainty is the only source of uncertainty considered and can be represented using a Gaussian random variable, the likelihood function can be calculated analytically based on the model predictions. However, in the case that various sources of uncertainty exist (e.g., natural variability in the input $\boldsymbol{x}$, data uncertainty in input and output measurement, and model uncertainty), the likelihood function is no longer simple to compute. In that case, sampling methods like Monte Carlo simulation are needed to compute the function for given parameter values. If the number of calibration parameters is relatively large, the evaluation of the likelihood function can become expensive even with a fast surrogate model for $G(\boldsymbol{x};\boldsymbol{\theta})$. In such cases, another surrogate model can be built to directly approximate the joint likelihood function of all the parameters, based on actual evaluations of the likelihood function for selected values of the parameters. Thereafter, we can evaluate this surrogate model, instead of the actual likelihood function, in the calculation of the posterior PDFs, which can speed up Bayesian calibration under multiple sources of uncertainty. For example, Bliznyuk et al. [Bliznyuk et al., 2008] approximated the unnormalized posterior density (the product of likelihood function and prior density) using radial basis functions.

If the number of parameters is relatively small, the integration of the product of the likelihood function and the prior PDFs of parameters can be conducted accurately and efficiently using numerical integration methods, such as Gaussian quadrature or the trapezoidal rule. When the number of parameters becomes large, Markov Chain Monte Carlo (MCMC) methods are widely used due to the relative insensitivity of the computational effort to the number of parameters. MCMC methods do not conduct the integration explicitly, but instead directly generate the random samples from the unnormalized posterior density of the parameters, upon convergence. Several algorithms are available for MCMC sampling, including Metropolis-Hastings [Hastings, 1970; Chib and Greenberg, 1995; Green, 1995; Gelman et al., 1996; Haario et al., 2006; Zuev et al., 2012], Gibbs [Casella and George, 1992], slice sampling [Neal, 2003], etc.

2.3 Quantitative model validation techniques

Model validation is defined as the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended use of the model [AIAA, 1998; ASME, 2006]. Qualitative validation methods such as graphical comparison between model predictions and experimental data are widely used in engineering. However, statistics-based quantitative methods are needed to supplement subjective judgments and to systematically account for errors and uncertainty in both model prediction and experimental observation [Oberkampf and Trucano, 2002].

Previous research efforts include the application of statistical hypothesis testing methods in the context of model validation [Hartmann et al., 1995; Hills and Trucano, 1999, 2002; Hills and Leslie, 2003; Rebba et al., 2006; Rebba and Mahadevan, 2006, 2008], and development of validation metrics as measures of agreement between model prediction and experimental observation [Urbina et al., 2003; Oberkampf and Barone, 2006; Rebba and Mahadevan, 2008;

Ferson et al., 2008; Ferson and Oberkampf, 2009]. Discussions on the pros and cons of these validation methods can be found in [Rebba and Mahadevan, 2008; Liu et al., 2011].

Suppose a computational model is constructed to predict a physical quantity. Quantitative model validation methods involve the comparison between model prediction and experimental observation. In this thesis, we use the following notations

- Y represents the "true value" of the system response
- Y_m is the model prediction of this true response Y
- Y_D is the experimental observation of Y.

The various quantitative validation methods investigated in this thesis are classified into two categories: (1) statistical binary hypothesis testing-based methods, and (2) non-hypothesis testing-based methods.

2.3.1 Hypothesis testing-based methods

Statistical binary hypothesis testing involves deciding between the plausibility of two hypotheses - the null hypothesis (denoted as H_0) and the alternative hypothesis (denoted as H_1). H_0 may be something that one believes could be true, whereas H_1 is a hypothesis opposite to H_0 [Marden, 2000]. For example, given a model for damping coefficient prediction, H_0 can be the hypothesis that the model prediction is equal to the actual damping coefficient, and correspondingly H_1 states that the model prediction is not equal to the actual damping coefficient. The null hypothesis H_0 will be rejected if it fails the test, and will not be rejected if it passes the test. Two types of error can possibly occur from this exercise: rejecting the correct hypothesis (type I error), or failing to reject the incorrect hypothesis (type II error). In the context of model validation, it should be noted that the underlying subject matter domain knowledge is also necessary for the implementation of the hypothesis testing-based methods, especially in the formulation of test hypotheses (H_0 and H_1) and the selection of model acceptance threshold. To formulate appropriate H_0 and H_1 for the validation of a model with stochastic output prediction Y_m, we need to be clear about the physical interpretation of "model being correct". In other words, we need to decide whether or not accurate prediction of statistical moments or the entire PDF of Y_m suggests that the model is correct.

2.3.1.1 Classical hypothesis testing

In classical hypothesis testing, a test statistic is first formulated and the probability distributions of this statistic under the null and alternative hypothesis are derived theoretically or by approximations. Thereafter, one can compute the value of the test statistic based on validation data and thus calculate the corresponding p-value, which is the probability that the test statistic falls outside a range defined by the computed value of the test statistic under the null hypothesis. The p-value can be considered as an indicator of how good the null hypothesis is, since a better H_0 corresponds to a narrower range defined by the computed value of the test statistic and thus a higher probability of the test statistic falling outside the range.

The decisions whether or not to reject the null hypothesis can be made based on the acceptable probabilities of making type I and type II errors (specified by the decision maker). The concept of significance level α defines the maximum probability of making type I error, and the probability of making type II error β can be estimated based on α and the probability distribution of the test statistic under H_1. The null hypothesis will be rejected if the computed p-value is less than α, or the computed β exceeds the acceptable value. Correspondingly, we will reject the model if H_0 is rejected, and accept the model if H_0 is not rejected. An alternative approach to comparing p-value and α is to use confidence intervals. A confidence interval can be constructed based on the confidence level $\gamma = 1 - \alpha$, and the null hypothesis will be rejected if the confidence interval does not include the predicted value from the model.

Various test statistics have been developed corresponding to different types of hypotheses.

The t-test or z-test can be used to test the hypothesis that the mean of a normal random variable is equal to the model prediction; the chi-square test can be used to test the hypothesis that the variance of a normal random variable is equal to the model prediction; and the hypothesis that the observed data come from a specific probability distribution can be tested using methods such as the chi-square test, the Kolmogorov-Smirnov (K-S) test, the Anderson-Darling test and the Cramer test [Lehmann and Romano, 2005]. The discussion in this thesis is limited to the tests on distribution mean, namely the t-test and the z-test, since the tests on variance or probability distribution require relatively large amounts of validation data, which is rarely available in practice.

The t-test is based on Student's t-distribution. Suppose the quantity of interest Y is a normal random variable with mean μ and standard deviation σ, and the measurement error ε_D is a normal random variable with zero mean and standard deviation σ_D. Thus, the experimental observation $Y_D = Y + \varepsilon_D \sim N(\mu, \sigma^2 + \sigma_D^2)$, i.e., Y_D also follows a normal distribution with mean μ and variance $\sigma^2 + \sigma_D^2$. For the sake of simplicity, let $\sigma_{Y_D} = \sqrt{\sigma^2 + \sigma_D^2}$. The validation data are a set of random samples of Y_D with size n (i.e., y_{D1}, y_{D2}, ..., y_{Dn}) and the corresponding sample mean is $\bar{Y}_D$ and sample standard deviation is S_D. The variable $(\bar{Y}_D - \mu)/(S_D/\sqrt{n})$ is modeled with a t-distribution with $(n-1)$ degrees of freedom. Therefore, if one assumes that the model mean prediction μ_m (if model prediction is deterministic, μ_m equals to the prediction value) is the mean of Y, i.e., the null hypothesis is $H_0 : \mu = \mu_m$, then the corresponding test statistic t (follows the same t-distribution) is

$$t = \frac{\bar{Y}_D - \mu_m}{S_D/\sqrt{n}} \tag{2.14}$$

The p-value for the two-tailed t-test (considering both ends of the distribution) can be obtained as

$$p = 2F_{T,n-1}(-|t|) \tag{2.15}$$

where $F_{T,n-1}$ is the cumulative distribution function (CDF) of a t-distribution with $(n-1)$ degrees of freedom. If the chosen significance level is α, then one will reject the null hypothesis H_0 if $p < a$, or fail to reject H_0 if $p > a$.

The t-test requires a sample size $n \geq 2$ in order to estimate the sample standard deviation S_D. If $n = 1$, the z-test can be used instead by assuming that the standard deviation of Y is equal to the standard deviation of model prediction Y_m, i.e., $\sigma = \sigma_m$ and $\sigma_{Y_D} = \sqrt{\sigma_m^2 + \sigma_D^2}$. Thus, the variable $(\bar{Y}_D - \mu)/(\sigma_{Y_D}/\sqrt{n})$ follows the standard normal distribution. Under the null hypothesis $H_0 : \mu = \mu_m$, the test statistic z is

$$z = \frac{\bar{Y}_D - \mu_m}{\sigma_{Y_D}/\sqrt{n}} \tag{2.16}$$

The corresponding p-value for the two-tailed z-test can be computed as

$$p = 2\Phi(-|z|) \tag{2.17}$$

where Φ is the CDF of the standard normal variable. Similar to the t-test, one will reject H_0 if $p < a$, or fail to reject H_0 if $p > a$.

To compute the probability of making type II error β, an alternative hypothesis H_1 is needed and a commonly used formulation is $H_1 : \mu = \mu_m + \epsilon_\mu$. In t-test, under the alternative hypothesis $H_1 : \mu = \mu_m + \epsilon_\mu$, the t statistic follows a non-central t-distribution with noncentrality parameter $\delta = \sqrt{n}\epsilon_\mu/\sigma_{Y_D}$ [Srivastava, 2002; McFarland, 2008], the probability of making type II error β can then be estimated as

$$\beta = 1 - \Pr(|t| > t_{1-\alpha/2,n-1}|\delta) \tag{2.18}$$

where the term $\Pr(|t| > t_{\alpha/2,n-1}|\delta)$ is called the power of the test in rejecting H_0. In this thesis, we use "Pr" in mathematical expressions and equations to represent the probability of a certain event. Similarly, β in the z-test can be estimated as

$$\beta = 1 - \Pr\big(|z - \delta| > \Phi^{-1}(1 - \alpha/2)\big) \tag{2.19}$$

2.3.1.2 Bayesian hypothesis testing

In probability theory, Bayes' theorem reveals the relationship between two conditional probabilities, e.g., the probability of occurrence of an event A given the occurrence of an event E (denoted as $\Pr(A|E)$), and the probability of occurrence of the event E given the occurrence of the event A (denoted as $\Pr(E|A)$). This relationship can be written as [Haldar and Mahadevan, 2000]

$$\Pr(E|A) = \frac{\Pr(A|E)\Pr(E)}{\Pr(A)} \tag{2.20}$$

Suppose event A is the observation of a single validation data point y_D, and event E is hypothesis H_0 being true (or hypothesis H_1 being true). Using Bayes' theorem, we can calculate the ratio between the posterior probabilities of the two hypotheses given validation data y_D as

$$\frac{\Pr(H_0|y_D)}{\Pr(H_1|y_D)} = \frac{\Pr(y_D|H_0)}{\Pr(y_D|H_1)} * \frac{\Pr(H_0)}{\Pr(H_1)} \tag{2.21}$$

where $\Pr(H_0)$ and $\Pr(H_1)$ are the prior probabilities of H_0 and H_1 respectively, representing the prior knowledge one has on the validity of these two hypotheses before collecting experimental data; and $\Pr(H_0|y_D)$ and $\Pr(H_1|y_D)$ are the posterior probabilities of H_0 and H_1 respectively, representing the updated knowledge one has after analyzing the collected experimental data. The likelihood function $\Pr(y_D|H_i)$ in Eq. 2.21 is the conditional probability of observing the data y_D given the hypothesis H_i ($i = 0$ or 1), and the ratio $\Pr(y_D|H_0)/\Pr(y_D|H_1)$ is known as the Bayes factor [O'Hagan, 1995; Kass and Raftery, 1995] and is used as the validation metric.

The original intent of the Bayes factor was to compare the data support for two models [Pericchi, 2005], and thus the two hypotheses become H_0: model M_i is correct and H_1: model M_j is correct. If $\boldsymbol{\theta_i}$ and $\boldsymbol{\theta_j}$ are the parameters of the two competing models respectively, the Bayes factor is calculated as

$$B = \frac{\Pr(y_D|H_0)}{\Pr(y_D|H_1)} = \frac{\int \Pr(y_D|\boldsymbol{\theta_i})\pi(\boldsymbol{\theta_i})d\boldsymbol{\theta_i}}{\int \Pr(y_D|\boldsymbol{\theta_j})\pi(\boldsymbol{\theta_j})d\boldsymbol{\theta_j}} \tag{2.22}$$

where $\pi(\boldsymbol{\theta_i})$ and $\pi(\boldsymbol{\theta_j})$ are the probability density distributions of $\boldsymbol{\theta_i}$ and $\boldsymbol{\theta_j}$ respectively.

In the context of validating a single model, H_0 and H_1 need to be formulated differently. Rebba and Mahadevan [Rebba and Mahadevan, 2008; Rebba et al., 2006] proposed the equality-based formulation ($H_0 : y_m = y_D$, $H_1 : y_m \neq y_D$) and the interval-based formulation ($H_0 : |y_m - y_D| < \epsilon$, $H_1 : |y_m - y_D| > \epsilon$) for Bayesian hypothesis testing, where y_m is the model prediction for a particular input $\boldsymbol{x}$.

2.3.2 Non-hypothesis testing-based methods

Besides the binary hypothesis testing methods, various other validation metrics have been developed to quantify the agreement between models and experimental data from other perspectives, such as the Mahalanobis distance [Srivastava, 2002; Rebba and Mahadevan, 2006], Kullback-Leibler divergence [Seghouane et al., 2005; Jiang and Mahadevan, 2006], probability bounds [Urbina et al., 2003], confidence intervals [Oberkampf and Barone, 2006], reliability-based metric [Rebba and Mahadevan, 2008], and area metric [Ferson and Oberkampf, 2009; Ferson et al., 2008]. Our discussion is restricted to the reliability-based metric and the area metric, since these two metrics have clear probabilistic or physical interpretations regarding the degree of model validity, and both can be applied to validation of a model with multiple input variables using data from discrete test combinations.

2.3.2.1 Reliability-based metric

The reliability metric r proposed by Rebba and Mahadevan [Rebba and Mahadevan, 2008] is a direct measure of model prediction quality and is relatively easy to compute. It is defined as the probability of the difference (d) between observed data (Y_D) and model prediction (Y_m) being less than a given tolerance limit ϵ

$$r = \Pr(-\epsilon < d < \epsilon), \quad d = Y_D - Y_m \tag{2.23}$$

Note that in Eq. 2.23, experimental observation is treated as random variable due to measurement error, and model output is treated as stochastic as well under the combined effect of epistemic and aleatory uncertainty. As the difference between two random variables, d is treated as a random variable, and the probability distribution of d can be obtained from the probability distributions of Y_D and Y_m. For instance, if the model prediction $Y_m \sim N(\mu_m, \sigma_m^2)$, and the corresponding observation $Y_D \sim N(\mu, \sigma_{Y_D}^2)$, the difference $d \sim N(\mu - \mu_m, \sigma_{Y_D}^2 + \sigma_m^2)$. For the sake of simplicity, let $\sigma_d = \sqrt{\sigma_{Y_D}^2 + \sigma_m^2}$. In this case, the reliability-based metric r can be derived as

$$r = \Phi[\frac{\epsilon - (\mu - \mu_m)}{\sigma_d}] - \Phi[\frac{-\epsilon - (\mu - \mu_m)}{\sigma_d}] \tag{2.24}$$

2.3.2.2 Area metric-based method

The area metric proposed by Ferson et al. [Ferson and Oberkampf, 2009; Ferson et al., 2008] measures the difference between the cumulative distribution functions (CDF) of model output and experimental data, and is defined as

$$d(F_{Y_m}, S_{Y_D}) = \int_{-\infty}^{+\infty} |F_{Y_m}(y) - S_{Y_D}(y)| \, \mathrm{d}y$$

where $F_{Y_m}(y)$ is the cumulative distribution function (CDF) of model output, and $S_{Y_D}(y)$ is the empirical CDF of experimental data.

The area metric can incorporate experimental data collected from multiple test combinations [Liu et al., 2011] using the so-called ‘‘u-pooling” procedure (transformation from physical space to probability space). For a single test combination with input $\boldsymbol{x_i}$, suppose $F^m_{\boldsymbol{x_i}}$ is the corresponding CDF of model output Y_m and y_{Di} is the observation, one can compute $u_i = F^m_{\boldsymbol{x_i}}(y_{Di})$ for this experimental combination. Based on the probability integral transform theorem [Angus, 1994], if the observation y_{Di} is a random sample from the probability distribution of model output, the computed u_i will be a random sample from the standard uniform distribution, and thus the empirical CDF of all the u_i's ($i = 1, 2, ..., N$) should match the CDF of the standard uniform random variable. The difference between these two CDF curves is a measure of the disparity between model predictions and experimental observations. Hence, an area metric in the transformed probability space can be developed as [Ferson et al., 2008]

$$d(F_u, S_u) = \int_0^1 |F_u - S_u| \, \mathrm{d}u \tag{2.25}$$

where F_u is the empirical CDF of all the u_i's and S_u is the CDF of the standard uniform random variable. If the value of $d(F_u, S_u)$ is small/large, the model predictions are correspondingly close/not close to experimental observations.

2.4 Time-dependent reliability analysis

In a system with time-varying inputs and degradation mechanisms, probability of failure (P_f) becomes a function of time, and continuous tracking of P_f is necessary for long-term reliability analysis. This section will discuss the theory and computational methods for time-dependent reliability analysis with a single or multiple performance criteria.

2.4.1 Time-dependent reliability analysis with a single performance criterion

For devices with a single performance criterion (denoted as y), the limit state function can be expressed as

$$G(\boldsymbol{x}_t, t) = y - y_{cr} = f(\boldsymbol{x}_t, t) - y_{cr} \tag{2.26}$$

where $\boldsymbol{x}_t$ is the set of input variables that characterize the device and the operational environment. Due to the existence of aleatory and epistemic uncertainty, $\boldsymbol{x}_t$ are either treated as random variables (if the probability distribution of $\boldsymbol{x}_t$ is independent of time), or random processes (if the probability distribution of $\boldsymbol{x}_t$ is time-dependent). The degradation function $f(\boldsymbol{x}, t)$ describes how y changes with time. The device is identified as failed when the performance variable y falls below a certain critical value y_{cr}. The probability of failure at a given time instant can be obtained as

$$P_f(t) = \int_{G(\boldsymbol{x}_t, t) < 0} \pi(\boldsymbol{x}_t) \, \mathrm{d}\boldsymbol{x}_t \tag{2.27}$$

where $\pi(\boldsymbol{x}_t)$ is the joint probability density function of $\boldsymbol{x}_t$.

In some practical problems, the analytical expression of $f(\boldsymbol{x}_t, t)$ may not be readily available but instead a computer simulation model $y = G_M(\boldsymbol{x}_t, t; \boldsymbol{\theta})$ is created to approximate the limit state function. Note that the original set of input variables $\boldsymbol{x}_t$ is divided into two sets: $\boldsymbol{x}_t$ is the set of variables for which the probability distribution can be obtained directly from measurement data; $\boldsymbol{\theta}$ is the set of variables for which the probability distribution can be obtained by model calibration, as presented in Section 2.2.

In the context of long term reliability analysis with a large number of input and parameter dimensions, the calculation of P_f can be hindered by two factors: (1) Integration of the joint PDF $\pi(\boldsymbol{x}_t)$ over the entire input and parameter space is required, and (2) repetitive evaluations of $G_M(\boldsymbol{x}_t; t, \boldsymbol{\theta})$ are needed at each time instant. The first issue may be addressed by using first-order reliability methods (FORM) and second-order reliability methods (SORM) [Haldar and Mahadevan, 2000], or importance sampling-based Monte Carlo integration techniques, which is an effective numerical method for high-dimensional integration [Zwillinger, 1995; Dey and Mahadevan, 2000]. The second issue may be addressed by using response surface (or surrogate model) [Bichon et al., 2011].

In addition to the failure probability at a single time point, the reliability of a device is often connected to the probability of failure during a time period. Compared against brute-force Monte Carlo simulation, the out-crossing approach is relatively efficient and can give the upper bound of the failure probability as [Ditlevsen and Madsen, 1996]

$$P_f(0,T) \leq P_f(0) + E[N^+(0,T)] \tag{2.28}$$

where $N^+(0,T)$ is the number of outcrossings of limit state function out of the safe domain $G(\boldsymbol{x}_t,t) > 0$ during the time period $[0,\ T]$. The expectation of $N^+(0,T)$ can be expressed using outcrossing rate $\nu^+(t)$, as

$$E[N^+(0,T)] = \int_0^T \nu^+(t)\mathrm{d}t \tag{2.29}$$

If the probability of having more than one outcrossing in a small time interval is negligible, the outcrossing rate $\nu^+(t)$ can be defined as [Rackwitz, 2001]

$$\begin{aligned}&\nu^+(t) = \\ &\lim_{\Delta t \to 0} \frac{\Pr\big(G(x_t,t) > 0 \cap G(x_{t+\Delta t}, t+\Delta t) \leq 0\big)}{\Delta t}\end{aligned} \tag{2.30}$$

Eq. 2.30 can be approximated using the PHI2 method with FORM algorithm [Andrieu-Renaud et al., 2004], as

$$\nu^+(t) = \lim_{\Delta t \to 0} \frac{\Phi_2[\beta(t), -\beta(t+\Delta t); \rho_G(t,t+\Delta t)]}{\Delta t} \tag{2.31}$$

where Φ_2 stands for standard binormal cumulative distribution function. $\rho_G(t, t + \Delta t)$ is the linear correlation between two limit states $G(\boldsymbol{x}_t, t)$ and $G(\boldsymbol{x}_{t+\Delta t}, t + \Delta t)$ in the equivalent standard normal space, which can be computed as

$$\rho_G(t, t + \Delta t) = \boldsymbol{\alpha}(t)\boldsymbol{\alpha}(t + \Delta t)$$

where $\boldsymbol{\alpha}(t)$ is the gradient of the limit state function $G(\boldsymbol{x}_t, t)$.

2.4.2 Time-dependent reliability analysis with multiple performance criteria

For devices with multiple dominant failure mechanisms, the performance variable becomes a vector (denoted as $\boldsymbol{y}$). Similar to the case of single failure mechanism, what we actually use in practice may be a computer model $\boldsymbol{y} = G_M(\boldsymbol{x}_t, t; \boldsymbol{\theta})$. In general, the failure region E of a device can be defined by a group of inequalities with the elements of $\boldsymbol{y}$ as variables. An example of these inequalities is $y_i < y_{cr,i}$ $(i = 1, 2, ...)$, i.e., the device fail when any element of the performance variable $\boldsymbol{y}$ falls below its corresponding critical value. The probability of failure at time t can be written as

$$\begin{aligned} P_f(t) &= \Pr\big(\boldsymbol{y} = G_M(\boldsymbol{x}_t, t; \boldsymbol{\theta}) \in E\big) \\ &= \int_E \pi(\boldsymbol{x}_t, \boldsymbol{\theta}) d\boldsymbol{x} d\boldsymbol{\theta} \end{aligned} \tag{2.32}$$

As we can see from Eq. 2.32, the existence of multiple failure mechanisms increases the difficulty of determining the integration domain E. Meanwhile, due to the correlation between multiple performance criteria, it is difficult to apply the PHI2 method to estimate the failure probability over a certain time period $P_f(0, T)$, and Monte Carlo simulation combined with response surface models may be needed.

2.5 Surrogate modeling techniques

As shown in the previous sections, Monte Carlo sampling-based methods require large amounts of limit state function evaluations. Even though the PHI2 method may reduce the number of function evaluations at a single time point, high number of time points can still render the long term reliability prediction for complex systems impossible. Time-dependent problem can be further complicated by the fact that the current system state is often dependent on the system states at k previous time points, e.g.,

$$\boldsymbol{y}_t = G(\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ..., \boldsymbol{y}_{t-k}, \boldsymbol{x}; \boldsymbol{\theta}) \tag{2.33}$$

In such cases, the state transition function G needs to be evaluated recursively to obtain the limit state function at time t. One practical example is the non-planar growth of a 3-D crack under fatigue load, where the current crack surface is obtained by evolving the crack surfaces at previous time steps. Therefore, developing computationally efficient surrogate models for the original limit state function or state transition function may become necessary for time-dependent reliability analysis.

Various surrogate modeling techniques have been developed and applied in reliability analysis, including linear/quadratic polynomial-based response-surface [Faravelli, 1989; Roussouly et al., 2013], artificial neural networks [Hurtado and Alvarez, 2001; Gomes and Awruch, 2004], support vector machine (SVM) [Rocco and Moreno, 2002; Li et al., 2006], polynomial chaos expansion (PCE) [Ghanem and Spanos, 2003; Xiu and Karniadakis, 2002; Huang et al., 2007], and Gaussian process (GP) interpolation (or Kriging) [Romero et al., 2004; Kaymaz, 2005]. Some reviews and discussions of these metamodeling techniques can be found in [Rackwitz, 2001; Gomes and Awruch, 2004; Hurtado, 2004; Blatman, 2009]. In this dissertation, PCE and GP interpolation, especially the latter one, are explored to accelerate time-dependent reliability analysis, as will be shown in the subsequent sections. Note that the discussion is limited to surrogate models with a single output variable. For surrogate modeling of high dimensional limit state functions or state transition functions, principal component analysis-like methods may be needed to reduce the dimension of the problem. One such example can be found in [Hombal et al., 2012; Hombal and Mahadevan, 2013b], where the evolution of a 3D crack surface characterized by 93 coordinate variables is approximated by using a Gaussian process surrogate model in a low-dimension latent space.

2.5.1 Gaussian process interpolation

Let $y = G(\boldsymbol{x})$ be the target function which we intend to describe with a surrogate model. GP interpolation (or kriging) assumes that the output y over the domain of input $\boldsymbol{x}$ is a Gaussian random process with a mean function $m(\boldsymbol{x})$ and a covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$, i.e.,

$$\begin{aligned} \mathrm{E}[y|\boldsymbol{x}] &= m(\boldsymbol{x}) \\ \mathrm{Cov}(y, y') &= k(\boldsymbol{x}, \boldsymbol{x}') \end{aligned} \tag{2.34}$$

Given a set of training data $\{\boldsymbol{X}_T, \boldsymbol{y}_T\}$ and the input $\boldsymbol{X}_p$ where prediction is desired, the conditional probability distribution of the output $\boldsymbol{y}_P$ follows a multivariate Gaussian distribution [Rasmussen and Williams, 2006] as

$$\begin{aligned} &\boldsymbol{y}_P|\boldsymbol{X}_P, \boldsymbol{X}_T, \boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \\ &\boldsymbol{\mu} = m(\boldsymbol{X}_P) + \Sigma_{PT}\Sigma_{TT}^{-1}\big(\boldsymbol{y}_T - m(\boldsymbol{X}_T)\big) \\ &\Sigma = k(\boldsymbol{x}, \boldsymbol{x}')\Sigma_{PP} - \Sigma_{PT}\Sigma_{TT}^{-1}\Sigma_{PT}^T \end{aligned} \tag{2.35}$$

where $\boldsymbol{\mu}$ is the mean vector of the prediction $\boldsymbol{y}_P$ conditioned on the training data, and Σ is the conditional covariance matrix of $\boldsymbol{y}_P$; Σ_{PT} is the covariance matrix between the prediction and the training data; Σ_{TT} is the covariance matrix of the training data; Σ_{PP} is the unconditional covariance matrix of $\boldsymbol{y}_P$.

In order to determine the multivariate Gaussian distribution in Eq. 2.35, we need to formulate the mean function $m(\boldsymbol{x})$ and the covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$. $m(\boldsymbol{x})$ is often formulated as a polynomial function of $\boldsymbol{x}$. The form of $k(\boldsymbol{x}, \boldsymbol{x}')$ may be selected from a number of commonly-used covariance functions based on the desired properties (order of continuity, stationary/nonstationary, isotropic/anisotropic) [Rasmussen and Williams, 2006]. Note that the selected formulations of $m(\boldsymbol{x})$ and $k(\boldsymbol{x}, \boldsymbol{x}')$ contain unknown coefficients. Let $\boldsymbol{\phi}$ denotes the coefficients of $m(\boldsymbol{x})$, and $\boldsymbol{\varphi}$ denotes the coefficients of $k(\boldsymbol{x}, \boldsymbol{x}')$. These coefficients can be estimated by maximizing their likelihood function, which is proportional to the probability density of training data conditioned on $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$, i.e.,

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\varphi}) \propto \pi(\boldsymbol{y}_T | \boldsymbol{X}_T, \boldsymbol{\phi}, \boldsymbol{\varphi}) \tag{2.36}$$

Since $\boldsymbol{y}_T | \boldsymbol{X}_T, \boldsymbol{\phi}, \boldsymbol{\varphi} \sim \mathcal{N}\big(m(\boldsymbol{X}_T), \Sigma_{TT}\big)$, the conditional probability density function $\pi(\boldsymbol{y}_T | \boldsymbol{X}_T, \boldsymbol{\phi}, \boldsymbol{\varphi})$ is a multivariate Gaussian PDF as

$$\begin{aligned} \pi(\boldsymbol{y}_T | \boldsymbol{X}_T, \boldsymbol{\phi}, \boldsymbol{\varphi}) &= (2\pi)^{-\frac{n}{2}} |\Sigma_{TT}|^{-\frac{1}{2}} * \\ & \exp\Big(-\frac{1}{2} [\boldsymbol{y}_T - m(\boldsymbol{X}_T)]^T \, \Sigma_{TT}^{-1} \, [\boldsymbol{y}_T - m(\boldsymbol{X}_T)] \Big) \end{aligned} \tag{2.37}$$

where n is the size of the training data set.

Since the estimation of coefficients depends on finding the maximum of the likelihood function, one needs to be cautious on the choice of optimization algorithms. In a high-dimension coefficient space, the likelihood function is rarely convex, and thus gradient-based local optimization algorithms may not be effective. When global optimization methods such as the DIRECT algorithm [Finkel and Kelley, 2004] and simulated annealing [Kirkpatrick et al., 1983] are used, careful selection of algorithm parameters is suggested. In order to ensure achieving the global maximum, we may need to manually divide the coefficient space into multiple smaller regions, and then search for maximums in these regions separately.

Note that the inverse of the covariance matrix Σ_{TT} is needed in the computation of the likelihood function in Eq. 2.36, as well as the construction of model prediction in Eq. 2.35. Some discussions on the efficient numerical strategies of inversing Σ_{TT} can be found in [McFarland, 2008; Haarhoff et al., 2013]. Also note that the size of Σ_{TT} increases with the size of training data set, which may lead to ill-conditioned matrix and high computational cost. In such cases, sparse Gaussian process approximations may be used, which estimate the inverse of Σ_{TT} via projections from the high-dimension training data space to a low-dimension latent space [Quinonero-Candela and Rasmussen, 2005].

2.5.2 Polynomial chaos expansion

PCE was originally used in the analysis of stochastic processes since it can converge to any process with finite second-order moments [Wiener, 1938; Xiu and Karniadakis, 2002]. A spectral method has been developed to use PCE in stochastic finite element analysis [Sudret and Der Kiureghian, 2002; Ghanem and Spanos, 2003], which represents the input random field with a finite expansion (such as a Karhunen-Loeve expansion) and represents the system response using a polynomial chaos expansion. This method can also be applied to construct a PCE surrogate model for any target computer model.

Consider again $y = G(\boldsymbol{x})$ as the target model. First, the input vector $\boldsymbol{x}$ needs to be transformed into standard normal random variables, which requires the probability distributions of $\boldsymbol{x}$ being available. Several methods can be used to transform non-normal random variables into normal [Rebba, 2005], including Rosenblatt transformation [Rosenblatt, 1952], Nataf transformation [Nataf, 1962], and Power and Modulus transformations [Box and Cox, 1964]. If some elements of the input vector $\boldsymbol{x}$ are random processes or random fields, Karhunen-Loeve expansion can be used to approximate these random processes/fields using a limited number of standard normal random variables [Ghanem and Spanos, 2003; Huang et al., 2001].

After the input $\boldsymbol{x}$ is converted into a set of standard normal random variables (denoted as $\boldsymbol{\zeta}$), the model output y can then be represented using a Hermite-chaos expansion, which is a summation of Hermite polynomial functions of $\boldsymbol{\zeta}$. A P^{th} order PCE surrogate model for y can be written as

$$y(\boldsymbol{x}) \approx \mathrm{PCE}(\boldsymbol{\zeta}) + \varepsilon_m = \sum_{i=1}^{M} a_i H_i(\boldsymbol{\zeta}) + \varepsilon_m \qquad M = \begin{pmatrix} n_{\boldsymbol{\zeta}} + P \\ n_{\boldsymbol{\zeta}} \end{pmatrix} \tag{2.38}$$

where H_i's are Hermite polynomials; $n_{\boldsymbol{\zeta}}$ is the dimension of the transformed input $\boldsymbol{\zeta}$ and P is the order of the polynomial; ε_m is the error of the PCE model. Because of the orthogonality of Hermite polynomials, the coefficients a_i's can be obtained as generalized Fourier coefficients [Huang et al., 2007]

$$a_i = \frac{\int \mathrm{PCE}(\boldsymbol{\zeta})\ H_i(\boldsymbol{\zeta})\ \mathrm{d}\boldsymbol{\zeta}}{\int H_i^2(\boldsymbol{\zeta})\ \mathrm{d}\boldsymbol{\zeta}} \tag{2.39}$$

Numerical integration methods, such as quadrature rule, smolyak grid [Gerstner and Griebel, 1998], and stochastic collocation [Huang et al., 2007], can be used to compute the integration in Eq. 2.39. Training data of the GPC surrogate model are thus needed on the grid points of these integration algorithms.

For any given input $\boldsymbol{x_k}$ and the corresponding transformed input $\boldsymbol{\zeta}_k$, the prediction of GPC surrogate model $\sum_{i=1}^{M} a_i H_i(\boldsymbol{\zeta_k})$ in Eq. 2.38 is deterministic, whereas the residual term ε_m is random. Under the Gauss-Markov assumption, ε_m asymptotically follows a normal distribution with zero mean, and the variance can be estimated as [Liang and Mahadevan, 2011; Seber and Wild, 2003]

$$\sigma_m^2 = \sigma^2[1 + \boldsymbol{H}^T(\boldsymbol{\zeta_k})(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}(\boldsymbol{\zeta_k})] \tag{2.40}$$

where σ_m^2 is a function of the transformed input $\boldsymbol{\zeta_k}$ (and thus a function of $\boldsymbol{x}_k$); the vector $\boldsymbol{H}(\boldsymbol{\zeta_k}) = [H_1(\boldsymbol{\zeta_k}),\ H_2(\boldsymbol{\zeta_k}),\ ...,\ H_M(\boldsymbol{\zeta_k})]^T$; the matrix $\boldsymbol{H} = [\boldsymbol{H}(\boldsymbol{\zeta_1}),\ \boldsymbol{H}(\boldsymbol{\zeta_2}),\ ...,\ \boldsymbol{H}(\boldsymbol{\zeta_N})]^T$; and $\sigma^2 = 1/(N-M)\sum_{j=1}^{N}[\mathrm{GPC}(\boldsymbol{\zeta_j}) - y(\boldsymbol{x_j})]^2$. $\{\boldsymbol{x}_j, \boldsymbol{y(x_j)}\}_{j=1}^{N}$ is the training data set, and $\{\boldsymbol{\zeta}_j\}_{j=1}^{N}$ is the corresponding transformed input set.

2.6 Summary

This chapter introduced the basic methods that will be used in uncertainty quantification and time-dependent reliability analysis. The Bayesian networks described in Section 2.1 are powerful tools to integrate models and experimental data at multiple levels of the system. The use of Bayesian networks to facilitate the integration of reliability prediction and UQ activities are demonstrated in subsequent chapters. Section 2.2.1 introduced Bayesian model calibration under the KOH framework, which explicitly accounts for model form uncertainty in the estimation of model parameters and can help avoid over-fitting. Section 2.3 introduced various quantitative model validation methods, which assess the predictive capability of model and can help quantify the uncertainty in model prediction. Several challenging issues in Bayesian calibration and in quantitative model validation are addressed respectively in Chapter III and Chapter IV. Section 2.5 illustrated two efficient surrogate modeling techniques, namely PCE and GP, that can facilitate the time dependent analysis described in Section 2.4. Inclusion of these UQ activities and surrogate models in time-dependent multi-physics system reliability analysis, and to prognosis with health monitoring data are investigated in Chapters V - VII.

CHAPTER III

CHALLENGING ISSUES IN BAYESIAN MODEL CALIBRATION

3.1 Introduction

As discussed in Section 2.2, Bayesian model calibration based on the KOH framework has been successfully applied in many problems. However, several challenging issues still remain in practical applications. Brynjarsdottir and O'Hagan [Brynjarsdottir and O'Hagan, 2013] showed that the calibration results may not be satisfying if the prior assumption of discrepancy does not capture the physics missing in the model. Moreover, the addition of a discrepancy term to the physics model may sometimes lead to parameter non-identifiability [Renard et al., 2010; Arendt et al., 2012; Ling and Mahadevan, 2013a]. Several different prior assumptions of model discrepancy have been used in previous studies, including constant [Arhonditsis et al., 2008], physics-based deterministic function [DeCarlo et al., 2013], Gaussian random variable [Renard et al., 2010; Sankararaman et al., 2011b; Sarkar et al., 2012; Koutsourelakis, 2009; Park et al., 2010; Riley and Grandhi, 2011], uncorrelated random vector [Bower et al., 2010], random walk [Arhonditsis et al., 2008], and Gaussian random process [McFarland and Mahadevan, 2008; Higdon et al., 2008; Arendt et al., 2012]. However, a rigorous comparison between these different prior formulations of model discrepancy has not been conducted, and a general guideline for choosing the formulations is not currently available.

In Section 3.2.1, we will investigate Bayesian calibration with five different prior formulations of model discrepancy function: (1) constant, (2) Gaussian random variable with fixed mean and variance, (3) Gaussian random variable with input-dependent mean and variance, (4) Gaussian random process with stationary covariance function, and (5) Gaussian random process with non-stationary covariance function.

A three-step method is proposed in Section 3.2.2 in order to assess these prior formulations of discrepancy, and also obtain a probability distribution of model parameters that accounts for the uncertainty due to the lack of knowledge of model inadequacy. First, Bayesian calibration is performed using each of the five options of model discrepancy function, and five posterior distributions of model parameters and model discrepancy are obtained. In the second step, we use the reliability-based metric introduced in Section 2.3.2.1 to assess the posterior model predictions corresponding to the distributions of model parameters and discrepancy obtained in the first step. In the third step, the five distributions of model parameters and discrepancy are combined into a single distribution based on the total probability theorem and the weights derived from the validation metric.

Section 3.3 gives a overview on the issue of model parameter identifiability, with focus on the non-identifiability caused by adding a model discrepancy function to the original physics model. A first-order Taylor series expansion-based method is developed in order to help choose a model discrepancy function without causing parameter non-identifiability.

The wide use of multi-physics simulations in engineering also poses another challenge to Bayesian calibration, since the dimension of parameter space is often large and the evaluation of likelihood function involves running multiple physics simulations which could become computationally expensive. Thus, simultaneous calibration of such multi-physics systems may not always be feasible. In Section 3.4, we will examine the calibration of coupled physics models using a Bayesian network-based approach. Strategies of sequential calibration are developed, which can provide accurate posterior distributions for unknown model parameters while keeping the dimension of the parameter space relatively low in each calibration.

3.2 Formulations and selection of model discrepancy in Bayesian calibration

3.2.1 Formulations of model discrepancy function

Section 2.2.1 reviewed the basic framework of Bayesian model calibration with a generalized discrepancy term. To establish the likelihood in Eq. 2.6, the model discrepancy function δ need to be formulated prior to calibration. This section explores several options to parameterize model discrepancy, and the corresponding likelihood functions are derived. These formulations can be used to represent the modeler's prior knowledge of model inadequacy, and the coefficients of the parameterized model discrepancy function can be included in Bayesian calibration in order to obtain *a posteriori* estimation of model discrepancy.

3.2.1.1 Model discrepancy as a constant bias

The simplest formulation is to treat model discrepancy as a constant, i.e., the bias between model prediction and the real process is assumed to be independent of the input $\boldsymbol{x}$. Using the assumption that measurement error $\varepsilon_{obs} \sim \mathcal{N}(0, \sigma_{obs}^2)$ and the linear relationship specified in Eqs. 2.3 and 2.4, experimental observation $(\boldsymbol{y}_D | \boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta)$ is normally distributed:

$$(\boldsymbol{y}_D | \boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \delta) \sim \mathcal{N}(G(\boldsymbol{X}_D; \boldsymbol{\theta}) + \delta, \sigma_{obs}^2 \boldsymbol{I}) \tag{3.1}$$

where $\boldsymbol{I}$ represents identity matrix, i.e., the elements of $\boldsymbol{y}_D$ are conditionally independent of each other. The likelihood function in Eq. 2.6 can then obtained by evaluating the PDF of the above Gaussian distribution with $\boldsymbol{y}_D = \boldsymbol{D}_i$.

3.2.1.2 Model discrepancy as i.i.d. Gaussian random variables with fixed mean and variance

Bias between model and experiment is rarely an input-independent constant. Instead, it may be different at different input settings. If the variation is purely random, model discrepancy can be treated as independent and identically distributed (i.i.d.) Gaussian random variables over the input domain, with fixed mean μ_δ and variance σ_δ. Therefore, for any input $\boldsymbol{x}$, we have

$$\delta(\boldsymbol{x}) \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2) \tag{3.2}$$

This type of model discrepancy not only adds a mean correction (μ_δ) to the model prediction but also increases the variance of $\boldsymbol{y}_D$ by σ_δ^2. The conditional probability distribution of $\boldsymbol{y}_D$ then becomes

$$(\boldsymbol{y}_D|\boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \mu_\delta, \sigma_\delta) \sim \mathcal{N}(G(\boldsymbol{X}_D, \boldsymbol{\theta}) + \mu_\delta, (\sigma_{obs}^2 + \sigma_\delta^2)\boldsymbol{I}) \tag{3.3}$$

3.2.1.3 Model discrepancy as Gaussian random variables with input-dependent mean and variance

If there is some evidence/knowledge suggesting the existence of input dependency in the model discrepancy function, the assumption of i.i.d. random variables may no longer be valid. The input dependency can be explicitly accounted for by assuming the mean and variance of δ to be functions of $\boldsymbol{x}$, i.e.,

$$\delta \sim \mathcal{N}(\mu_\delta(\boldsymbol{x}; \boldsymbol{\phi}), \sigma_\delta^2(\boldsymbol{x}; \boldsymbol{\varphi})) \tag{3.4}$$

where $\boldsymbol{\phi}$ is the set of coefficients in the mean function $\mu_\delta(*)$, and $\boldsymbol{\varphi}$ is the set of coefficients in the variance function $\sigma_\delta^2(*)$. The choice of the form of the mean and variance functions can be rather subjective and is typically problem-specific. One such choice will be demonstrated in the numerical example.

In this case, the conditional probability distribution of $\boldsymbol{y}_D$ becomes

$$(\boldsymbol{y}_D|\boldsymbol{X}_D,\boldsymbol{\theta},\sigma_{obs},\boldsymbol{\phi},\boldsymbol{\varphi}) \sim \mathcal{N}(G(\boldsymbol{X}_D,\boldsymbol{\theta}) + \mu_\delta(\boldsymbol{X}_D;\boldsymbol{\phi}), [\sigma_{obs}^2 + \sigma_\delta^2(\boldsymbol{X}_D;\boldsymbol{\varphi})]\boldsymbol{I}) \tag{3.5}$$

3.2.1.4 Model discrepancy as a Gaussian process with stationary covariance function

The aforementioned formulations of model discrepancy assume that model discrepancies at different input points are statistically independent of each other. However, if a model makes poor prediction at one input point, it is not uncommon to find that it also fails at a nearby input point, which suggests the existence of statistical correlation between model discrepancies at these two input points. Instead of a set of independent random variables, it may be desirable to treat δ as a Gaussian process

$$\delta \sim \mathcal{N}(m(\boldsymbol{x};\boldsymbol{\phi}), k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\varphi})) \tag{3.6}$$

where $m(*)$ is the mean function of this Gaussian process δ, and $\boldsymbol{\phi}$ is the set of coefficients of $m(*)$; $k(*)$ is the covariance function of δ, and $\boldsymbol{\varphi}$ is the set of coefficients of $k(*)$. Covariance functions with different properties (e.g., order of continuity, stationary/nonstationary, isotropic/anisotropic) have been developed. The most widely used one within the machine learning field is the squared exponential function, which is infinitely differentiable, stationary, and isotropic [Rasmussen and Williams, 2006]. An example squared exponential function can be written as

$$k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\varphi}) = \lambda \exp\Big(-\sum_{i=1}^{q}\frac{(x_i - x_i')^2}{2l_i^2}\Big) \tag{3.7}$$

where $\boldsymbol{\varphi} = [\lambda, l_1, l_2, ..., l_q]$, and q is the dimension of the inputs $\boldsymbol{x} = [x_1, x_2, ..., x_q]$; λ is the variance of this Gaussian process; l_i is the length-scale parameter corresponding to the input variable x_i. Higher values of l_i indicate higher statistical correlation along the input dimension x_i.

As the discrepancy function is assumed to be a Gaussian process, the elements of $\boldsymbol{y}_D$ become correlated, and they follow a multivariate Gaussian distribution as

$$(\boldsymbol{y}_D|\boldsymbol{X}_D, \boldsymbol{\theta}, \sigma_{obs}, \boldsymbol{\varphi}) \sim \mathcal{N}(G(\boldsymbol{X}_D; \boldsymbol{\theta}) + m(\boldsymbol{X}_D; \boldsymbol{\phi}), \Sigma + \sigma_{obs}^2 \boldsymbol{I}) \tag{3.8}$$

where

$$m(\boldsymbol{X}_D; \boldsymbol{\phi}) = \begin{bmatrix} m(\boldsymbol{x}_{D1}; \boldsymbol{\phi}) \\ \vdots \\ m(\boldsymbol{x}_{Dn}; \boldsymbol{\phi}) \end{bmatrix}, \quad \Sigma = \begin{bmatrix} k(\boldsymbol{x}_{D1}, \boldsymbol{x}_{D1}; \boldsymbol{\varphi}) & \dots & k(\boldsymbol{x}_{D1}, \boldsymbol{x}_{Dn}; \boldsymbol{\varphi}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_{Dn}, \boldsymbol{x}_{D1}; \boldsymbol{\varphi}) & \dots & k(\boldsymbol{x}_{Dn}, \boldsymbol{x}_{Dn}; \boldsymbol{\varphi}) \end{bmatrix} \tag{3.9}$$

In order to compute the likelihood function $\mathcal{L}(\boldsymbol{\theta}, \sigma_{obs}, \delta)$, the PDF of the multivariate Gaussian distribution in Eq. 3.8 needs to be evaluated, which requires computing the determinant and inverse of the covariance matrix $\Sigma + \sigma_{obs}^2 \boldsymbol{I}$. As the number of data points increases, this covariance matrix may become ill-conditioned and lead to significant numerical errors in the computation of the likelihood function. In this paper, we use a Cholesky decomposition-based strategy to compute the determinant and inverse of the covariance matrix, which helps mitigate the numerical difficulty in likelihood evaluation. [Haarhoff et al., 2013]

3.2.1.5 Model discrepancy as a Gaussian process with non-stationary covariance function

The Gaussian process in Section 3.2.1.4 is stationary since the squared exponential covariance function satisfies $k(\boldsymbol{x}, \boldsymbol{x'}; \varphi) = k(\boldsymbol{x} + \boldsymbol{\Delta x}, \boldsymbol{x'} + \Delta \boldsymbol{x}; \varphi)$, i.e., the marginal variances of model discrepancy at different input points are the same. If we consider variance as an indicator of the degree of uncertainty, it may be desirable to allow the variance of model discrepancy to vary with model input, since the uncertainty regarding model discrepancy at an input setting depends on the amount of data points available and the smoothness of the underlying discrepancy function. In order to account for the non-stationary trend of variance in the Gaussian process discrepancy formulation, we can add an input-dependent term to the squared exponential covariance function. For example, if time is one of the input variables and one has reasons to believe that the variance of discrepancy decreases with time, the following covariance function may be used

$$k(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\varphi}) = \begin{cases} k_1(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\varphi}_1), & t \neq t' \\ k_1(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\varphi}_1) + k_2(t; \boldsymbol{w}), & t = t' \end{cases} \tag{3.10}$$

where $k_1(*)$ is the squared exponential function in Eq. 3.7; the second term $k_2(t; \boldsymbol{w}) = w_1 \exp(-w_2 t)$ is an exponential decreasing function with respect to time when $w_1 > 0$ and $w_2 > 0$, which essentially reduces the values of the diagonal entries of the covariance matrix Σ in Eq. 3.9.

Note that other types of non-stationary covariance functions, such as linear/polynomial dot product kernels and neural network kernels [Rasmussen and Williams, 2006], can also be used depending on the features of a specific problem.

Substituting the new covariance function into Eq. 3.9, we can obtain the corresponding covariance matrix, which can be further used to compute the likelihood function as described in Section 3.2.1.4.

3.2.2 Assessment and combination of calibration results using a three-step approach

If the discrepancy term explains all the missing physics in a model, the posterior distribution of model parameters is expected to converge to the true distribution given sufficient data. However, precise knowledge of model discrepancy is rarely available in practice. In this section, we propose a three-step heuristic approach based on a quantitative model validation metric and the total probability theorem, as explained below.

(1) The available experimental data are partitioned into two sets, one of which will be used for calibration (denoted as $\{\boldsymbol{X}_D^C, \boldsymbol{y}_D^C\}$ whereas the other one will be used for validation (denoted as $\{\boldsymbol{X}_D^V, \boldsymbol{y}_D^V\}$).

(2) We perform Bayesian calibration on model parameters with the various formulations of model discrepancy function using data set D_C as discussed in Section 3.2.1. By imposing the i-th prior formulation of the discrepancy function (denoted as δ_i) and adding it to the model G, we obtain the corrected model $M_i = G + \delta_i$. The corresponding posterior probability distribution of model parameters and discrepancy is denoted as $\pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C, M_i)$. Model predictions based on $\pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C, M_i)$ are then validated using a reliability-based metric, which calculates the probability of model predictions being within a specified tolerance from the validation data.

(3) The probability of model prediction satisfying a specified tolerance can be used in the selection of the model discrepancy formulation, or can be further used to obtain an "average" posterior distribution of model parameters based on the total probability theorem.

3.2.2.1 Reliability-based model validation metric

The purpose of validation activity in this section is to assess the quality of predictions resulting from calibration with different prior formulations ofthe model discrepancy function. The reliability-based metric is illustrated in Section 2.3.2.1 which can be used as a measure of model predictive capability. It is defined as the probability of the absolute difference between model prediction and observed data being less than a specified tolerance ϵ, i.e.,

$$r(\boldsymbol{x}) = \Pr(|y_D - y_m| < \epsilon) \tag{3.11}$$

where $r(\boldsymbol{x})$ is the reliability metric for a given input point $\boldsymbol{x}$ within the validation domain, i.e., $\boldsymbol{x} \in \boldsymbol{X}_D^V$; y_D is the observation corresponding to $\boldsymbol{x}$ and thus $y_D \in \boldsymbol{y}_D^V$; y_m is the model prediction at $\boldsymbol{x}$. The computation of $r(\boldsymbol{x})$ requires the probability distributions of y_D and y_m, which are discussed below.

We first discuss how to obtain the prediction y_m and the corresponding probability distribution. As illustrated in Sections 2.2.1 and 3.2.1, the computer model $G(\boldsymbol{x};\boldsymbol{\theta})$ and the discrepancy function $\delta(\boldsymbol{x})$ can be calibrated using the observed values of $\boldsymbol{y}_D^C$ (denoted as $\boldsymbol{D}_C$), and then we obtain the posterior probability distribution $\pi(\boldsymbol{\theta}, \delta|\boldsymbol{D}_C)$ of model parameters and δ. The prediction y_m in Eq. 3.11 is based on the extrapolation of the calibrated computer model and discrepancy function into the validation domain, i.e.,

$$\begin{aligned} & y_m|\boldsymbol{x}, \boldsymbol{\theta}, \delta = G(\boldsymbol{x};\boldsymbol{\theta}) + \delta(\boldsymbol{x}), \quad \boldsymbol{x} \in \boldsymbol{X}_D^V \\ & \pi(y_m|\boldsymbol{x}, \boldsymbol{D}_C) = \int \pi(y_m|\boldsymbol{x}, \boldsymbol{\theta}, \delta)\pi(\boldsymbol{\theta}, \delta|\boldsymbol{D}_C)\ \mathrm{d}\boldsymbol{\theta}\ \mathrm{d}\delta \end{aligned} \tag{3.12}$$

Note that we cannot yet use the relationship specified in Eqs. 2.3 and 2.4 to obtain the probability distribution of y_D, since the model and discrepancy function have not been validated with $\{\boldsymbol{X}_D^V, \boldsymbol{y}_D^V\}$ and thus Eq. 2.4 may not be valid in the validation domain. However, Eq. 2.3 still applies since it is independent of model prediction. If the true value y is treated as a constant for the given input $\boldsymbol{x}$, and the measurement error is treated as a Gaussian random variable with zero mean and fixed variance, y_D will also be a Gaussian random variable. The mean and variance of y_D can be estimated using repetitive measurements at the same input $\boldsymbol{x}$. If only one measurement is taken at $\boldsymbol{x}$, we may assume that $y_D \sim \mathcal{N}(D, \sigma_{obs}^2)$, where D is the measured value of y_D, and σ_{obs}^2 is the variance of the measurement error obtained in the calibration step, i.e., the measurement error is assumed to be the same in the calibration and validation domain.

Eq. 3.11 defines the reliability metric as a function of the input $\boldsymbol{x}$, and we can compute the values of r at each point within the validation domain, e.g., $r(\boldsymbol{x}_{D1})$, $r(\boldsymbol{x}_{D2})$, ... These individual values of r can inform decisions on point-wise model selection [Hombal and Mahadevan, 2013a]. However, the focus of this section is to assess the formulations of model discrepancy in Bayesian calibration based on validation results. Thus, it is desirable to establish a single measure of model predictive capability over the entire validation domain, which can be achieved by treating the input $\boldsymbol{x}$ as a random variable. The probability distribution of $\boldsymbol{x}$ is determined by a certain design of experiment and the elements of $\boldsymbol{X}_D^V$ are random samples from this distribution. As a function of $\boldsymbol{x}$, the reliability metric r also becomes a random variable, and $\{r(\boldsymbol{x}_{Di})|\boldsymbol{x}_{Di} \in \boldsymbol{X}_D^V\}$ are samples of this random variable. These samples can be used to estimate the mean of r (denoted as μ_r), which represents the expected probability of prediction being tolerable over the validation domain, and thus can be considered as an overall validation metric.

By substituting the results of Bayesian calibration into Eq. 3.12, the validation metric μ_r can be computed with respect to each of the various model discrepancy formulations illustrated in Section 3.2.1. If the μ_r corresponding to a particular formulation δ_i is significantly higher than the others, it is suggested that the calibration using δ_i leads to better prediction in the validation domain, and thus δ_i may be the best approximation to the actual model discrepancy. However, different formulations can sometimes lead to similar values of μ_r, in which case one may use a total probability theorem-based approach as discussed below.

3.2.2.2 Combining the posterior probability distributions of model parameters and discrepancy

When predictions based on different formulations of model discrepancy function lead to similar validation results, it is not clear which formulation of δ and the corresponding calibration result one should select for further prediction. In such cases, combining the various posterior distributions of model parameters and discrepancy may be considered. This approach accounts for the uncertainty induced by the lack of knowledge regarding model discrepancy formulation, and can be viewed as an extension to the Bayesian framework proposed by Sankararaman and Mahadevan [Sankararaman and Mahadevan, 2012a], which demonstrated the use of model validation results to combine the prior and posterior probability distributions of model parameters in multi-level systems [Sankararaman, 2012; Mullins et al., 2013].

Since the overall reliability metric μ_r described in the previous section is a probabilistic metric which quantifies the predictive capability of M_i, we assume that the probability of the model M_i being correct is proportional to the corresponding reliability metric μ_r^i, i.e., $\Pr(M_i) \propto \mu_r^i$. By assuming further that $\{\delta_i\}_{i=1}^N$ is the set of all the possible model discrepancy formulations, we can obtain the corresponding set of possible corrected models $\{M_i\}_{i=1}^N$. Thus, normalization of μ_r^i leads to the probability of M_i as

$$\Pr(M_i) = \frac{\mu_r^i}{\sum_{j=1}^N \mu_r^j} \tag{3.13}$$

Each M_i corresponds a posterior probability distribution of model parameters and discrepancy $\pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C, M_i)$. Based on the total probability theorem, the density functions of these posterior distributions can be combined to obtain a single density function $\pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C)$ as

$$\pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C) = \sum_{i=1}^N \pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C, M_i) \Pr(M_i) \tag{3.14}$$

Note that the probability density function $\pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C)$ is not conditioned on any specific model discrepancy formulation, and it is expected to be wider than the conditional density functions $\{\pi(\boldsymbol{\theta}, \delta | \boldsymbol{D}_C, M_i)\}_{i=1}^N$ since the uncertainty due to model discrepancy formulation is included.

3.2.3 Numerical example

The various options of model discrepancy function are investigated by considering the calibration of Young's modulus using an Euler-Bernoulli beam model and synthetic data. A example with a more complicated physics model and real experimental data will be showed in Section 5.3.1.

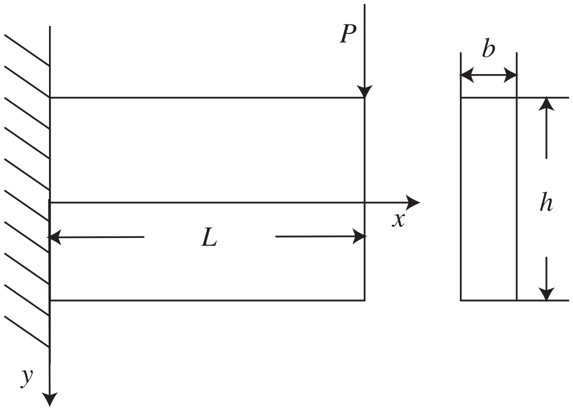


Figure 3.1: A thick cantilever subjected to point load at the free end

A thick microcantilever beam subjected to a point load P is considered (Fig. 3.1, $L = h = 400\ \mu m$, $b = 1\ \mu m$). By solving the Euler-Bernoulli equation, the static vertical deflection u_y along the x-axis ($y = 0$) can be obtained as [Gere and Goodno, 2009]

$$u_y = \frac{P(3L - x)x^2}{6EI} \tag{3.15}$$

The solution in Eq. 3.15 does not account for shear deformation. A more accurate solution of u_y can be obtained from the Timoshenko beam theory as [Timoshenko and Goodier, 1970; Augarde and Deeks, 2008]

$$u_y = \frac{P}{6EI}\left[(4 + 5\nu)\frac{h^2 x}{4} + (3L - x)x^2\right] \tag{3.16}$$

In this example, we assume for the sake of illustration that the solution in Eq. 3.16

represents the true value of u_y. 20 experimental data points are synthetically generated by adding a measurement noise term $\varepsilon_{obs} \sim \mathcal{N}(0, 0.1)$ to Eq. 3.16 with Young's modulus $E = 200$ GPa and the applied load $P = 2.5$ μN. Note that this value of Young's modulus and the true solution in Eq. 3.16 are assumed to be unknown in the calibration exercise. The Euler-Bernoulli beam solution in Eq. 3.15 is assumed to be the available physics model, and the synthesized data are used to calibrate E using Eq. 3.15. The discrepancy function between the physics model and the true solution is a linear function of x as

$$\delta_{true} = \frac{P}{6EI}(4 + 5\nu)\frac{h^2 x}{4} \tag{3.17}$$

Assuming δ_{true} in Eq. 3.17 is unknown, a common practice to acquire prior knowledge of model discrepancy is by examining the difference between calibration data and the least-squares model prediction (i.e., prediction using Eq. 3.15 with E estimated using the least-squares method). The upper subplot of Fig. 3.2 shows the calibration data and least-squares model prediction, and the lower subplot shows the actual model discrepancy δ_{true} and the least-squares model discrepancy δ_{ls}. We can observe that both δ_{true} and δ_{ls} have functional dependence on the input x. However, δ_{ls} appears to be a nonlinear function of x, whereas the dependence of δ_{true} on x is linear, which suggests that we may not be able to infer the actual form of model discrepancy function from δ_{ls}.

In order to study the effect of model discrepancy assumptions on the calibration results, we consider the various options discussed in Section 3.2.1: (1) an unknown constant δ_1; (2) i.i.d. Gaussian random variables with input-independent mean and variance, i.e., $\delta_2 \sim \mathcal{N}(\mu_{\delta_2}, \sigma^2_{\delta_2})$; (3) Gaussian random variables with input-dependent mean and variance, i.e., $\delta_3 \sim \mathcal{N}(\mu_{\delta_3}(x), \sigma_{\delta_3}(x)$; (4) a Gaussian process with squared exponential covariance function, i.e., $\delta_4 \sim \mathcal{N}(m(x), k(x, x'))$. The dependence of model discrepancy on the input x which we observed in the plot of δ_{ls} can be accounted for by using δ_3 or δ_4, since both options include input-dependent mean and variance function. In this example, $\mu_{\delta_3}(x)$, $\sigma_{\delta_3}(x)$, and $m(x)$ are

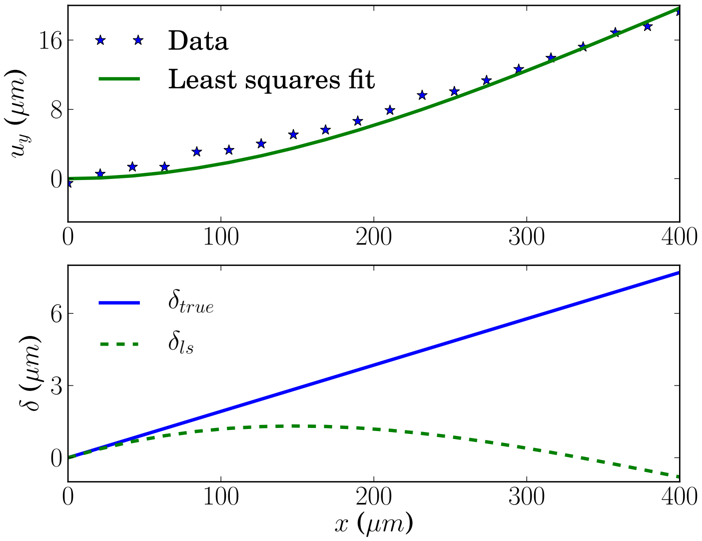


Figure 3.2: Actual model discrepancy versus least squares model discrepancy

assumed to be linear functions of x. The unknown coefficients in these model discrepancy formulations are calibrated along with Young's modulus E, and independent uniform prior distributions are used.

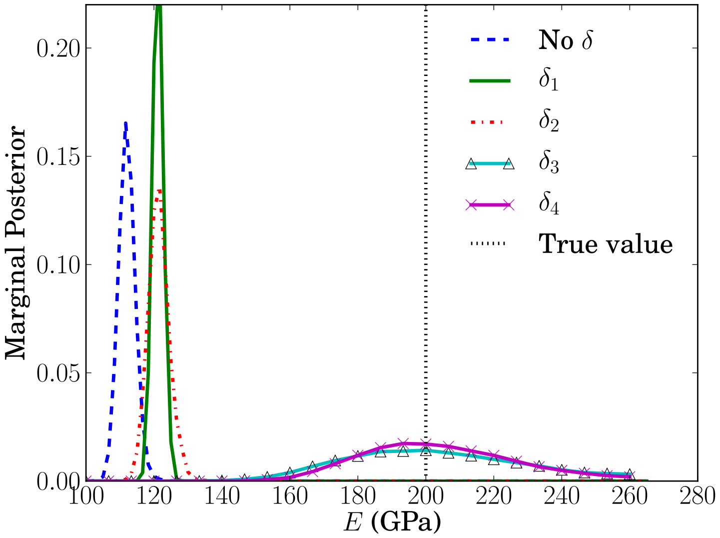


Figure 3.3: Marginal posterior PDFs of Young's modulus

The marginal posterior PDFs of Young's modulus corresponding to the various options of model discrepancy are shown in Fig. 3.3. It can be observed that calibration without model discrepancy results in the most biased estimation of E, which is expected since the physics model is imperfect. The posteriors of E obtained from calibration with δ_1 or δ_2 are also away from the true value, since these two options assume that model discrepancy is input-independent whereas the actual δ is input-dependent. Note that the use of δ_2 leads to a wider posterior of E compared to δ_1, since treating model discrepancy as i.i.d. random variables (δ_2) overestimates the uncertainty due to model error. Calibration with δ_3 or δ_4 gives identical posteriors of E, and both posterior PDFs cover the true value. We can also observe that these two posterior PDFs are significantly wider than the others, which is mainly due to the uncertainty on the coefficients of $\mu_{\sigma_3}(x)$ and $m(x)$. If stronger priors are enforced on these coefficients, i.e., the ranges of their uniform prior distributions are set to be smaller, the posterior variance of E will decrease correspondingly.

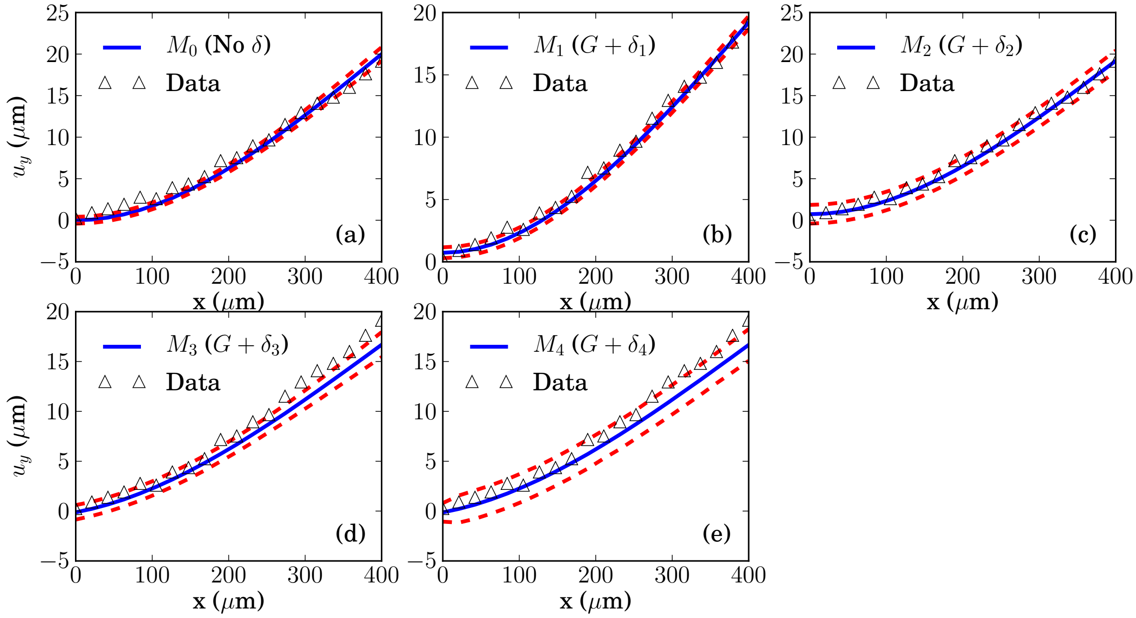


Figure 3.4: Comparison between the predictions of calibrated beam model and calibration data

We verify the posterior distributions of E and δ by comparing the calibrated model prediction against the calibration data, as shown in Fig. 3.4. The solid lines in Fig. 3.4 are the mean predictions of the model corrected by the various choices of δ, and the dashed lines are the corresponding 90% probability bounds. We can observe that all the mean predictions match the calibration data well, whereas only the 90% probability bounds resulting from δ_2, δ_3, and δ_4 are able to account for the variation in data.

Further, we assess the various options of model discrepancy formulations by examining the predictive capability of the calibrated model outside the domain of calibration ($P = 2.5$ μN). 20 validation data points are generated with the applied load $P = 3.5$ μN, and the three-step approach proposed in Section 3.2.2 is implemented. A graphical comparison between the validation data and the calibrated model predictions ($M_i = G + \delta_i, i = 0, 1, 2, 3, 4$) is shown in Fig. 3.5 (a)-(e), where the solid lines represent mean prediction and the dashed lines are 90% probability bounds.

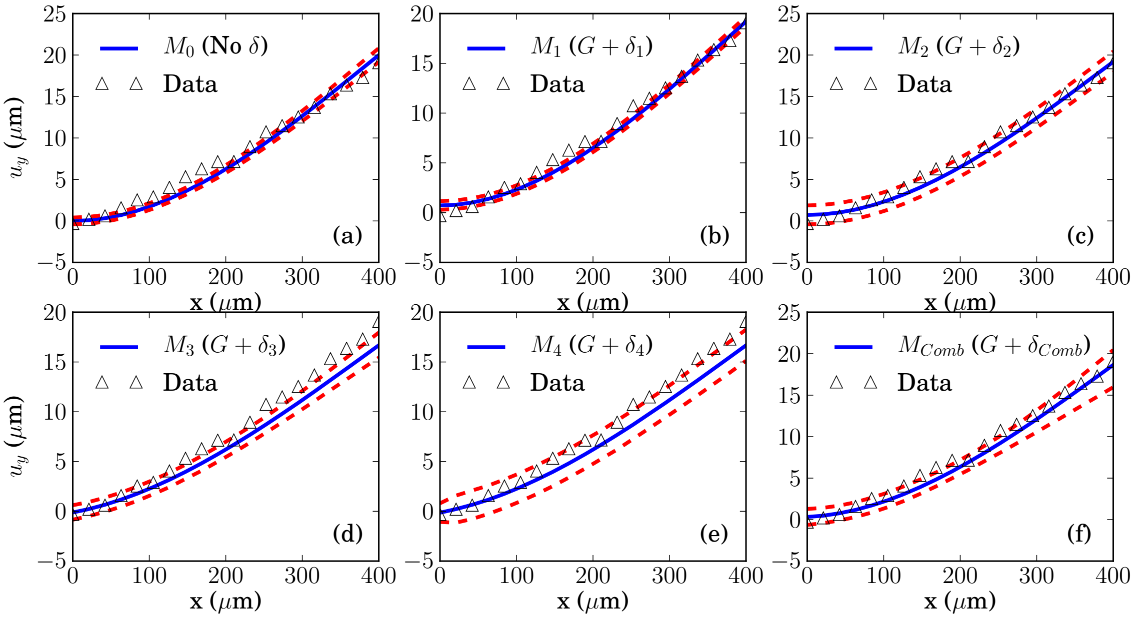


Figure 3.5: Comparison between the predictions of calibrated beam model and validation data

It is interesting to observe from Fig. 3.5 that the predictions M_3 and M_4 are outperformed by the other three in the validation domain, although M_3 and M_4 include better estimations of E and more accurate assumptions on model discrepancy. The reason why M_3 and M_4 fail to match the validation data is that δ_3 and δ_4 only depend on x, whereas the actual δ

is a function of both x and P. Thus, when we extrapolate the model from the calibration domain ($P = 2.5$ μN) to the validation domain ($P = 3.5$ μN), δ_3 and δ_4 are no longer good representations of model discrepancy. Since the model with the true value of E has significant errors as shown in Fig. 3.2, inaccurate estimation of the model error will not be able to correct the model prediction. In contrast, calibration without δ or with a simple formulation of δ may produce a model with relatively small errors (e.g., δ_{ls} in Fig. 3.2), and the prediction is more sensitive to the physics model than to the model discrepancy function δ. Consequently, we may see more consistent performance from the models calibrated with simple δ or without δ given that the behavior of the true physical quantity does not change dramatically outside the calibration domain (e.g., from linear to highly nonlinear).

Table 3.1 lists the values of the overall validation metric μ_r computed with the tolerance level $\epsilon = 0.75$ μm for the model predictions plotted in Fig. 3.5. The model calibrated with δ_1 (constant bias) is found to have the highest value of reliability. If the reliability of M_1 is not considered high enough to rule out other choices of model discrepancy, we can combine the various posterior distributions of E and δ via the method proposed in Section 3.2.2.2 and Eq. 3.14. The combined PDF of E is shown in Fig. 3.6, and the prediction (M_{Comb}) based on the combined distribution of E and δ is shown in Fig. 3.5 (f). The combined distribution of E becomes bimodal and covers the true value of E, and the prediction M_{comb} has a wider probability bound, which accounts for the uncertainty in model discrepancy formulation.

Table 3.1: Overall reliability of model predictions

	M_0 (No δ)	M_1	M_2	M_3	M_4
μ_r^i	0.50	0.66	0.55	0.36	0.34

Discussion In this numerical example, we demonstrated how the various choices of the model discrepancy function (δ) can affect the calibration result, and how the three-step approach proposed in Section 3.2.2 can help the selection of discrepancy formulation. We

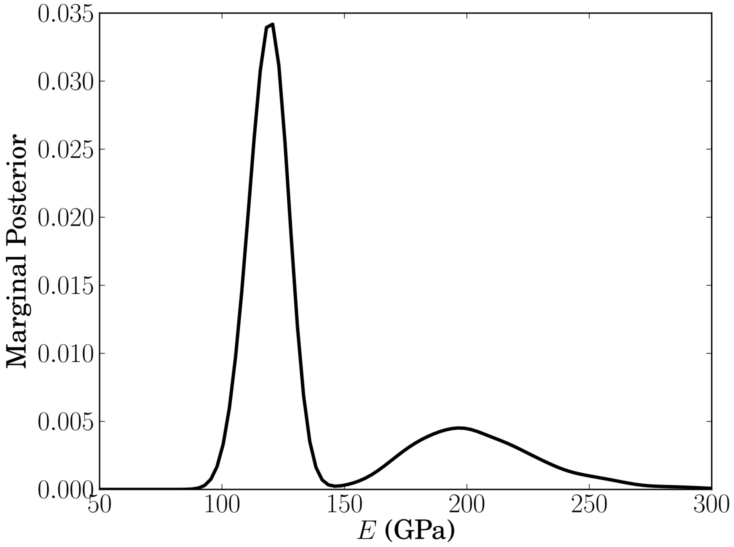


Figure 3.6: Combined marginal PDF of Young's modulus

showed that calibration with inaccurate assumption on δ can lead to biased estimation of physical parameters. However, posterior predictions based on simpler and less accurate assumptions of δ can still match data well outside the calibration domain. This is because the prediction is more sensitive to δ if complicated formulations are used, and thus the capability of δ to extrapolate from the calibration domain becomes crucial for achieving accurate predictions. In contrast, prediction based on a simpler formulation of δ tends to rely more on the physics model, which may have better extrapolation capability than a purely statistical term δ.

3.3 Identifiability of model parameters

Before implementing the calibration of a model, it is often of interest to determine whether we can extract useful information from the calibration data. A model is non-identifiable if there are infinite "best" (depending on the criterion chosen) estimates for the model parameters. In the Bayesian model calibration framework, a typical sign of non-identifiabilty is that the posterior PDFs of some of the model parameters are close to the prior PDFs, which indicates that the marginal likelihoods of these parameters are nearly flat leading to an infinite number of maximum likelihood estimates of the model parameters. In general, model non-identifiability can be classified into two types, namely structural non-identifiability and practical non-identifiability [Raue et al., 2009]. The first type of non-identifiability, structural non-identifiability is due to the redundant parameterization of the model structure. Even if the model is structurally identifiable, a second type of non-identifiability, practical non-identifiability, can still arise due to the insufficient amount and quality of observation data. The quality of data is related to the bias and noise in the data due to the imprecision of measurement techniques. Successful detection of structural non-identifiability may help reduce model redundancy. Also, by detecting the existence of practical non-identifiability, analysts may be able to overcome it by developing better design of experiments or improving data quality [Arendt et al., 2011].

It is usually straightforward to detect structural non-identifiability if the analytical expression of a model is available; however, in many problems, the analytical expression of the model is not readily available. One example is a dynamics model without an explicit steady state solution. Another possible case occurs when we add a discrepancy function to the numerical solutions of some governing equations, which in fact forms a new model without any analytical expression to consider [Arendt et al., 2010]. Various analytical and numerical methods have been developed to detect the structural non-identifiability of dynamics models [Grewal and Glover, 1976; Walter and Pronzato, 1996; Jia-fan et al., 2011], whereas the second possible case does not appear to have been studied in the literature. This section addresses this case.

Since the second level of non-identifiability, practical non-identifiability, is related to both model structure and observation data, it is necessary to inspect the likelihood function in order to determine whether some parameters of a model are practically non-identifiable.

In fact, rigorous definitions of model non-identifiability can be constructed based on the analytical properties of likelihood functions [Gu and Lu, 1994; Paulino and de Bragança Pereira, 1994; Little et al., 2010]. In addition to the theoretical analysis of likelihood functions, Raue et al. [Raue et al., 2009, 2011] developed a numerical approach based on the concept of ”profile likelihood” [Murphy and Van der Vaart, 2000], which has been shown to be effective in detecting practical non-identifiability. When the analytical expression of the likelihood function is available, or its numerical evaluation is trivial, it may be preferable to apply the profile likelihood-based method and determine the practical non-identifiability directly. But this method becomes cumbersome when the construction of the likelihood function is computationally expensive, since repetitive evaluations of the likelihood function are required to compute the profile likelihood.

Given the above observations, we propose a first-order Taylor series expansion-based method, which can detect structural non-identifiability for models without analytical expressions, and can detect practical non-identifiability due to insufficient amount of data. This method does not involve computation of the likelihood function, and thus is simpler to implement and less computationally demanding. The limitations of this method are: (1) it uses a linear approximation of the model, and hence may fail to detect non-identifiability if the model is highly nonlinear; (2) it can only detect local non-identifiability as the Taylor series expansion is constructed based on the derivatives at a single point; (3) it does not apply to statistical models; and (4) it does not cover practical non-identifiability due to the *quality* of data.

Suppose the physics model to be calibrated is $y_m = G(\boldsymbol{x};\boldsymbol{\theta})$, and a Gaussian process discrepancy function $\delta \sim \mathcal{N}\big(m(\boldsymbol{x};\boldsymbol{\phi}), k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\varphi})\big)$ is added to the model. Thus, a new model is formed as $G_{new}(\boldsymbol{x};\boldsymbol{\psi}) = G(\boldsymbol{x};\boldsymbol{\theta}) + m(\boldsymbol{x};\boldsymbol{\phi})$, where $\boldsymbol{\psi} = [\boldsymbol{\theta},\boldsymbol{\phi}]$ includes the parameters of the physics model ($\boldsymbol{\theta}$) and the parameters of the mean of the discrepancy function ($\boldsymbol{\phi}$). In the case that the measurement noise is a zero mean random variable, G_{new} is the expectation of observation y_D according to Eqs. 2.3 and 2.4. Further assume that the analytical form of G_{new} is not available. In such cases, the first-order Taylor series expansion of this model (as shown in Eq. 3.18) can be used as an efficient approximation when the model is believed to be not highly nonlinear:

$$E[y_D] = G_{new}(\boldsymbol{x}_D; \boldsymbol{\psi}) \approx G_{new}(\boldsymbol{x}_D; \hat{\boldsymbol{\psi}}) + \sum_{i=1}^{p} \left.\frac{\partial G_{new}}{\partial \psi_i}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} \psi_i \tag{3.18}$$

where $\hat{\boldsymbol{\psi}}$ is the mean value of the prior of $\boldsymbol{\psi}$, and p is the number of model parameters.

Suppose there are n data points available, i.e., experimentally observed values $\boldsymbol{D} = [D_1, D_2, ..., D_n]$ corresponding to different input settings $\boldsymbol{X}_D = [\boldsymbol{x}_{D1}, \boldsymbol{x}_{D2}, ..., \boldsymbol{x}_{Dn}]$. Without considering measurement uncertainty and the variance of the model discrepancy function, we can construct a linear system as

$$\boldsymbol{A}\boldsymbol{\psi}^T = \boldsymbol{D}, \quad \boldsymbol{A} = \begin{bmatrix} \left.\frac{\partial G_{new}}{\partial \psi_1}\right|_{\boldsymbol{x}_{D1}, \hat{\boldsymbol{\psi}}} & \cdots & \left.\frac{\partial G_{new}}{\partial \psi_p}\right|_{\boldsymbol{x}_{D1}, \hat{\boldsymbol{\psi}}} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial G_{new}}{\partial \psi_1}\right|_{\boldsymbol{x}_{Dn}, \hat{\boldsymbol{\psi}}} & \cdots & \left.\frac{\partial G_{new}}{\partial \psi_p}\right|_{\boldsymbol{x}_{Dn}, \hat{\boldsymbol{\psi}}} \end{bmatrix} \tag{3.19}$$

The linear system in Eq. 3.19 can be underdetermined or determined, depending on the rank of the matrix $\boldsymbol{A}$ (denoted as r_A). If $r_A < p$, the system is underdetermined and there will be an infinite number of $\boldsymbol{\psi}$ values satisfying Eq. 3.19; if $r_A = p$, the system is determined and there will be a unique vector $\boldsymbol{\psi}$ satisfying Eq. 3.19. The latter case suggests that the model is practically identifiable given the available data points (assuming the quality of the data does not cause non-identifiability). The former case suggests that the model is non-identifiable either due to the model structure or insufficient data. If the inequality $r_A < p$ continues to hold as we increase the number of observation data, then it can be inferred that the model is structurally non-identifiable.

In order to help the analyst reduce model redundancy once a model is detected as structurally non-identifiable, it may be of interest to know which set of parameters can/cannot be identified. Using the formulation of the linear system in Eq. 3.19, we can retrieve this information by checking the linear dependency between the column vectors of the matrix $\boldsymbol{A}$, since the i-th column of $\boldsymbol{A}$ corresponds to the parameter ψ_i. For example, if the i-th column vector $\boldsymbol{a}_i$ and the j-th column vector $\boldsymbol{a}_j$ are linearly dependent, it is apparent that the corresponding parameters ψ_i and ψ_j are non-identifiable using the linear model in Eq. 3.18. We can also find one set of identifiable parameters using the simple algorithm below (note that there may be multiple sets of identifiable parameters):

Algorithm 1 Find one set of identifiable parameters

```
Input: The first-order derivative matrix $\boldsymbol{A}$
Output: The index set of identifiable parameters $\boldsymbol{I}$
  $\boldsymbol{A}_{temp} = \boldsymbol{A}$
  $\boldsymbol{I}$ = empty set
  for $i = 1$ to $p$ do
    $r_1$ = the rank of $\boldsymbol{A}_{temp}$
    Remove the $i$-th column from $\boldsymbol{A}_{temp}$
    $r_2$ = the rank of $\boldsymbol{A}_{temp}$ (with the $i$-th column removed)
    if $r_1 > r_2$ then
      Add the value of $i$ to the set $\boldsymbol{I}$ as an element
    end if
  end for
  return $\boldsymbol{I}$
```

In order to illustrate the proposed method, consider a simple example:

$$\begin{aligned} y_m &= \theta_1 x + \theta_2 x^2 \\ \delta &= \phi_1 + \phi_2 x \end{aligned} \tag{3.20}$$

It is not difficult to see from the above expressions that the two parameters of model (θ_1 and θ_2) are identifiable given no less than two pairs of input-output data. However, if we add the model discrepancy δ to the model, θ_1 and ϕ_2 will become unidentifiable no matter how many data points are available.

Let $\boldsymbol{\psi} = [\theta_1, \theta_2, \phi_1, \phi_2]$. Suppose measurements are available at five input points: $x_D^1 = -1.0$, $x_D^2 = 2.0$, $x_D^3 = 3.5$, $x_D^4 = 4.0$, and $x_D^5 = 6.0$. We can calculate the derivatives $\partial(y_m + \delta)/\partial\psi_i$ numerically (e.g., forward difference or central difference) at these input points for given values of the parameters, and thus obtain the matrix $\boldsymbol{A}$. For example, for $[\theta_1, \theta_2, \phi_1, \phi_2] = [2.0, 0.5, 4.0, 1.5]$, the matrix $\boldsymbol{A}$ for the combination of the physics model and δ is

$$\boldsymbol{A} = \begin{bmatrix} -1.0 & 1.0 & 1.0 & -1.0 \\ 2.0 & 4.0 & 1.0 & 2.0 \\ 3.5 & 12.25 & 1.0 & 3.5 \\ 4.0 & 16.0 & 1.0 & 4.0 \\ 6.0 & 36.0 & 1.0 & 6.0 \end{bmatrix}$$

The rank of $\boldsymbol{A}$ is equal to 3, indicating that the parameters become unidentifiable after the combination of y_m and δ. We can also use the program in Algorithm 1 to infer that the parameters will become identifiable once the term $\phi_1 x$ is removed from the model discrepancy function.

3.4 Calibration of multi-physics computational models

Multi-physics modeling usually involves the combination of several models from different individual physics analyses. Ideally, these models would be calibrated separately with input-output experimental data corresponding to individual models. But in practice the experimental data may not be sufficient or available for some of the models. To calibrate all the models with limited information, a Bayesian network-based method is proposed below.

3.4.1 Integration of multi-physics models and experimental data via Bayesian network

Suppose we have two physics models $y_{m1} = G_1(\boldsymbol{x}_1; \boldsymbol{\theta}_1, \boldsymbol{\theta}_{12})$ and $y_{m2} = G_2(\boldsymbol{x}_2; \boldsymbol{\theta}_2, \boldsymbol{\theta}_{12})$. Note that these two models have different input variables ($\boldsymbol{x}_1$ versus $\boldsymbol{x}_2$) and parameters ($\boldsymbol{\theta}_1$ versus $\boldsymbol{\theta}_2$), but they also share some common parameters $\boldsymbol{\theta}_{12}$. Based on the discussion in Section 2.2, two Bayesian networks can be constructed for these two models individually. Further, due to the existence of the common parameters, these two networks can be connected to form a full network as shown in Fig. 3.7, which enables information flow from one network to the other.

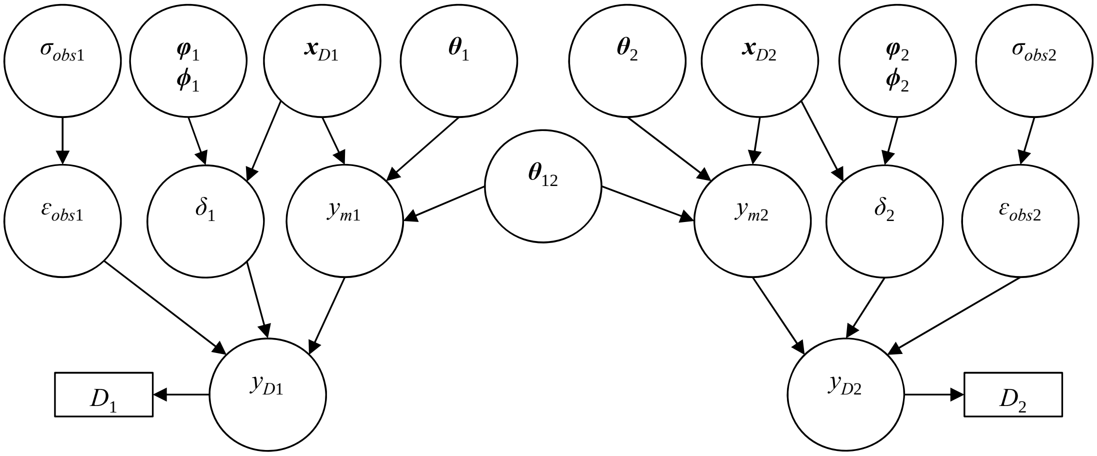


Figure 3.7: Bayesian network for two physics models

3.4.2 Strategy of Bayesian calibration for multi-physics models

If both the observation data D_1 and D_2 are available, we have three options for model calibration based on the Bayesian networks in Fig. 3.7, as presented below.

The first option is to calibrate the two models simultaneously. Let $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ represents the calibration parameters of the two networks respectively except for the common parameters

$\boldsymbol{\theta}_{12}$, i.e., $\boldsymbol{\Phi}_1 = [\boldsymbol{\theta}_1, \sigma_{obs1}, \boldsymbol{\phi}_1, \boldsymbol{\varphi}_1]$, $\boldsymbol{\Phi}_2 = [\boldsymbol{\theta}_2, \sigma_{obs2}, \boldsymbol{\phi}_2, \boldsymbol{\varphi}_2]$.

$$\begin{aligned} &\pi(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \boldsymbol{\theta}_{12} | y_{D1} = D_1, y_{D2} = D_2) \\ &= \frac{\pi(y_{D1} = D_1, y_{D2} = D_2 | \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \boldsymbol{\theta}_{12})\ \pi(\boldsymbol{\Phi}_1)\ \pi(\boldsymbol{\Phi}_2)\ \pi(\boldsymbol{\theta}_{12})}{\pi(y_{D1} = D_1, y_{D2} = D_2)} \end{aligned} \tag{3.21}$$

From the first option, we can obtain the posterior PDFs of all the parameters using both D_1 and D_2. However, this option can be computationally expensive because of the high dimensional parameter space.

The second option is to let information flow from left to right, conducting a two-step calibration. Following the procedure of Bayesian calibration for a single model presented in Section 2.2, $\boldsymbol{\Phi}_1$ and $\boldsymbol{\theta}_{12}$ are first calibrated using the observation data D_1. Then, the marginal posterior PDF of $\boldsymbol{\theta}_{12}$, $\pi(\boldsymbol{\theta}_{12} | y_{D1} = D_1)$, is used as the prior when we calibrate the parameters of the other network ($\boldsymbol{\Phi}_2$ and $\boldsymbol{\theta}_{12}$). Applying Bayes' theorem, we have

$$\begin{aligned} &\pi(\boldsymbol{\theta}_{12} | y_{D1} = D_1) = \frac{\int \pi(y_{D1} = D_1 | \boldsymbol{\Phi}_1, \boldsymbol{\theta}_{12})\ \pi(\boldsymbol{\Phi}_1)\ \pi(\boldsymbol{\theta}_{12})\ \mathrm{d}\boldsymbol{\Phi}_1}{\int \pi(y_{D1} = D_1)} \\ &\pi'(\boldsymbol{\Phi}_2, \boldsymbol{\theta}_{12} | y_{D1} = D_1, y_{D2} = D_2) \\ &= \frac{\pi(y_{D2} = D_2 | \boldsymbol{\Phi}_2, \boldsymbol{\theta}_{12})\ \pi(\boldsymbol{\Phi}_2)\ \pi(\boldsymbol{\theta}_{12} | y_{D1} = D_1)}{\int \pi(y_{D2} = D_2 | \boldsymbol{\Phi}_2, \boldsymbol{\theta}_{12})\ \pi(\boldsymbol{\Phi}_2)\ \pi(\boldsymbol{\theta}_{12} | y_{D1} = D_1)\ \mathrm{d}\boldsymbol{\Phi}_2\ \mathrm{d}\boldsymbol{\theta}_{12}} \\ &= \frac{\pi(y_{D2} = D_2 | \boldsymbol{\Phi}_2, \boldsymbol{\theta}_{12})\ \pi(\boldsymbol{\Phi}_2)\ \pi(\boldsymbol{\theta}_{12} | y_{D1} = D_1)}{\pi(y_{D2} = D_2 | y_{D1} = D_1)} \end{aligned} \tag{3.22}$$

We can prove that Eq. 3.22 gives the same joint posterior PDF of $\boldsymbol{\Phi}_2$ and $\boldsymbol{\theta}_{12}$ as from

Eq. 3.21. By combining the two expressions in Eq. 3.22, we have

$$
\begin{aligned}
&\pi'(\mathbf{\Phi}_2, \boldsymbol{\theta}_{12} | y_{D1} = D_1, y_{D2} = D_2) \\
&= \frac{\pi(y_{D2} = D_2 | \mathbf{\Phi}_2, \boldsymbol{\theta}_{12}) \pi(\mathbf{\Phi}_2) \int \pi(y_{D1} = D_1 | \mathbf{\Phi}_1, \boldsymbol{\theta}_{12}) \pi(\mathbf{\Phi}_1) \pi(\boldsymbol{\theta}_{12}) \mathrm{d}\mathbf{\Phi}_1}{\pi(y_{D2} = D_2 | y_{D1} = D_1)\ \pi(y_{D1} = D_1)} \\
&= \int \frac{\pi(y_{D1} = D_1 | \mathbf{\Phi}_1, \boldsymbol{\theta}_{12})\ \pi(y_{D2} = D_2 | \mathbf{\Phi}_2, \boldsymbol{\theta}_{12})\ \pi(\mathbf{\Phi}_1)\ \pi(\mathbf{\Phi}_2)\ \pi(\boldsymbol{\theta}_{12})}{\pi(y_{D2} = D_2 | y_{D1} = D_1)\ \pi(y_{D1} = D_1)}\ \mathrm{d}\mathbf{\Phi}_1 \\
&= \int \frac{\pi(y_{D1} = D_1, y_{D2} = D_2 | \mathbf{\Phi}_1, \mathbf{\Phi}_2, \boldsymbol{\theta}_{12})\ \pi(\mathbf{\Phi}_1)\ \pi(\mathbf{\Phi}_2)\ \pi(\boldsymbol{\theta}_{12})}{\pi(y_{D2} = D_2 | y_{D1} = D_1)\ \pi(y_{D1} = D_1)}\ \mathrm{d}\mathbf{\Phi}_1 \\
&= \int \pi(\mathbf{\Phi}_1, \mathbf{\Phi}_2, \boldsymbol{\theta}_{12} | y_{D1} = D_1, y_{D2} = D_2)\ \mathrm{d}\mathbf{\Phi}_1
\end{aligned}
\tag{3.23}
$$

Therefore, the second option provides us the posterior PDF of $\mathbf{\Phi}_1$ based on D_1, and the posterior PDFs of $\mathbf{\Phi}_2$ and $\boldsymbol{\theta}_{12}$ based on both D_1 and D_2. Note that the computational effort in the second option can be much smaller than in the first option, due to the reduced number of parameters in each step of the calibration.

The third option is similar to the second one, except that the information flows from right to left, i.e., $\mathbf{\Phi}_2$ and $\boldsymbol{\theta}_{12}$ are first calibrated using the observation data D_2, and then the marginal posterior PDF of $\boldsymbol{\theta}_{12}$, $\pi(\boldsymbol{\theta}_{12} | y_{D2} = D_2)$, is used as prior in the calibration of $\mathbf{\Phi}_2$ and $\boldsymbol{\theta}_{12}$ using the data D_1. Hence, the posterior PDFs of $\mathbf{\Phi}_1$ and $\boldsymbol{\theta}_{12}$ are obtained using both D_1 and D_2, whereas the posterior PDF of $\mathbf{\Phi}_2$ is only based on D_2.

Note that although the above Bayesian network-based method is presented using a two-model problem, it can be extended to the cases that $N (N \geq 2)$ physics models are to be calibrated with limited information, and there will be up to $N! + 1$ options of calibration available, depending on the existence of common parameters between different models.

3.5 Conclusion

This chapter first investigated various prior formulations of the model discrepancy function in Bayesian calibration. A first-order Taylor series expansion-based method was developed to determine the identifiability of model parameters, especially when one tries to add a model discrepancy function to a model. The numerical example showed that different choices of model discrepancy (δ) formulation can result in significantly different calibration results, and that the estimation of physical parameters may be biased if the formulation is over-simplified. It is also observed that the prediction based on complicated model discrepancy formulations may not match data well outside the calibration domain. In order to facilitate the selection of model discrepancy formulation and also accommodate the uncertainty due to this selection, we developed a three-step approach using a reliability-based model validation metric and the total probability theorem. This approach combines the posterior probability distributions of model parameters and δ resulting from the various options of δ into a single distribution, which is useful especially when the various options of δ have similar performances.

This chapter also proposed a Bayesian network-based approach in order to simplify the calibration of a multi-physics system. Multiple options of calibration exist when various physics models are connected via common parameters, and more efficient calibration may be achieved with sequential calibration strategies. This Bayesian network-based approach using a set of multi-physics models and data for two types of MEMS devices is illustrated in Chapter V.

CHAPTER IV

INTERPRETATIONS, RELATIONSHIPS, AND APPLICATION ISSUES IN MODEL VALIDATION

4.1 Introduction

This chapter is motivated by several issues that remain unclear in the practice of model validation: (1) validation with fully characterized, partially characterized, or uncharacterized experimental data; (2) validation of constant vs. stochastic model predictions; (3) accounting for the existence of directional bias; and (4) interpretation and selection of thresholds in different validation metrics.

The four issues cannot be discussed without an introduction to the terminology concerning model and validation experiments. When a model is developed, the physical quantity of interest Y is postulated to be a function of a set of variables $\{\boldsymbol{x}, \boldsymbol{\theta}\}$, i.e., $Y = f(\boldsymbol{x}, \boldsymbol{\theta})$. This function is not exactly known and hence is approximated using a model with output Y_m, i.e., $Y_m = g(\boldsymbol{x}, \boldsymbol{\theta})$. Y is observable through some experiments and $\boldsymbol{x}$ is the set of input variables to the experiments. The term ‘‘input’’ is referred to as the variables in a model that can be measured in experiments. We assume that the same set of variables goes into the model and validation experiments as inputs (i.e., the terms ‘‘model input’’ and ‘‘experimental input’’ represent the same set of variables), and we are comparing the outputs of the model and experiments during validation. Although there is an infinite set of variables in nature that can affect validation experiments, ‘‘model input’’ and ‘‘experimental input’’ only contain a subset of these variables due to the assumptions and approximations that are made in modeling. Model parameter set ($\boldsymbol{\theta}$) contains variables that may be measurable in concept but are usually difficult or impractical to directly measure in validation experiments. Therefore, the values/distributions of these variables are usually obtained from model calibration (performed prior to validation with a different set of input-output experimental data), and assumed invariant during multiple replicates of the validation experiment. For example, the maximum deflection (w) of a cantilever beam subjected to a point load at the free end is the quantity of interest, and it is known to be a function of the load value P, beam geometry, and Young's modulus (E). The solution $w = PL^3/(3EI)$ based on Euler-Bernoulli beam theory is a model that approximates this unknown function. The model input set $\boldsymbol{x}$ includes the load P, the moment of inertia I, and beam length L; the model parameter set $\boldsymbol{\theta}$ contains only one variable E.

The division of variables into two sets (model input $\boldsymbol{x}$ and model parameter $\boldsymbol{\theta}$), which is usually decided by modelers, helps classify the sources of uncertainty involved in validation. If a variable is measured and reported as point value, the uncertainty in this value is due to measurement error; if a variable is not measured but a range/distribution is assigned, the uncertainty is due to imprecise data; if a variable is inferred from model calibration, the sources of uncertainty include natural variability (the variability between different experiments), data uncertainty (uncertainty due to measurement error and insufficient data) and model uncertainty. It should be noted that the diagnostic quality and the bias in experiments are not considered as "input". Instead, they are classified as components of the measurement error, which is represented by ε_D in this thesis.

With the terminology introduced above, we can now continue with the discussion of the four issues. First, validation data can be collected from three possible types of experiments: (1) fully characterized (i.e., all the input variables of individual experiments are measured and reported as point values), (2) partially characterized (i.e., some inputs 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, $\boldsymbol{x}$ is treated as a random vector due to the lack of measurements on the input

Table 4.1: Three types of validation experiments and the corresponding input-output data

Fully characterized	Input	$\boldsymbol{x}_{D1}$	$\boldsymbol{x}_{D2}$	...	$\boldsymbol{x}_{Dn}$
	Output	$\boldsymbol{y}_{D1}$	$\boldsymbol{y}_{D2}$	...	$\boldsymbol{y}_{Dn}$
Partially characterized	Input	$\pi(\boldsymbol{x}_{D1})$	$\pi(\boldsymbol{x}_{D2})$	...	$\pi(\boldsymbol{x}_{Dn})$
	Output	$\boldsymbol{y}_{D1}$	$\boldsymbol{y}_{D2}$	...	$\boldsymbol{y}_{Dn}$
Uncharacterized	Input	$\pi(\boldsymbol{x}_{D})$			
	Output	$\boldsymbol{y}_{D1}$	$\boldsymbol{y}_{D2}$	...	$\boldsymbol{y}_{Dn}$

set $\boldsymbol{x}$. The reported intervals and expert opinions (if available) are needed to construct a probability distribution of $\boldsymbol{x}$, and the corresponding density function is denoted as $\pi(\boldsymbol{x})$. Note that in the Bayesian approach, the lack of knowledge (epistemic uncertainty) is represented through a probability distribution (subjective probability). Non-probabilistic approaches have also been proposed to handle the epistemic uncertainty; in this thesis, we only focus on probabilistic methods. For partially characterized validation data, input distributions are assigned to different experiments separately, and these distributions represent input data uncertainty in each individual experiment. For uncharacterized validation data, a single input distribution is assigned to multiple experiments, and this distribution represents the uncertainty due to both natural variability and input data uncertainty. Table 4.1 shows a typical format of input-output data collected from the three types of experiments. Although uncertainty exists in the inputs of partially characterized or uncharacterized experiments, the resulting data may still be considered for validation by practitioners if the input uncertainty is limited. While most of the previous studies only focus on validation with fully characterized experimental data, this chapter explores the use of all three types of data in various validation methods.

Second, due to the existence of aleatory and epistemic uncertainty, both the model prediction (denoted as Y_m) and the physical quantity to be predicted (denoted as Y) can be uncertain, and this has been the dominant case studied in the literature [Hills and Leslie, 2003; Zhang and Mahadevan, 2003; Rebba et al., 2006; Rebba and Mahadevan, 2006, 2008;

Ferson et al., 2008; Ferson and Oberkampf, 2009]. However, in practice it is possible that either Y_m or Y can be considered as deterministic. Deterministic Y_m implies that for given values of the model input variables, the output prediction of the model is deterministic. The application of various validation methods to these different cases is examined in this chapter.

Third, in this study, we define two terms to characterize the difference between model prediction and validation data - bias and directional bias. Bias is defined as simply the difference between the mean value of model prediction and the statistical mean value of experiment data, and the term ‘‘directional bias” implies the persistence of bias in one direction as one varies the inputs of model and experiment. This chapter explores whether various validation metrics are able to account for the existence of directional bias.

Fourth, although different validation metrics are developed to measure the agreement between model prediction and validation data from different perspectives, this chapter shows that under certain conditions some of the validation metrics can be mathematically related. These relationships may help decision makers to select appropriate validation metrics and the corresponding model acceptance/rejection thresholds.

Various quantitative validation metrics, including the p-value in classical hypothesis testing [Schervish, 1996], the Bayes factor in Bayesian hypothesis testing [O'Hagan, 1995], a reliability-based metric [Rebba and Mahadevan, 2008], and an area metric [Ferson et al., 2008; Ferson and Oberkampf, 2009], are investigated in this chapter. Based on the original definition of Bayes factor, we formulate two types of Bayesian hypothesis testing, one on the accuracy of the predicted mean and standard deviation of model prediction, and the other one on the entire predicted probability distribution of model prediction. These two formulations of Bayesian hypothesis testing can be applied to both fully characterized and partially characterized experiments. The use of these two types of experimental data in the other validation methods is also investigated. The first formulation of Bayesian hypothesis testing, along with the modified reliability-based method and the area metric-based method, takes into account the existence of directional bias. The mathematical relationships among the metrics used in classical hypothesis testing, Bayesian hypothesis testing, and the reliability-based method are investigated.

Section 4.2 discusses the possible scenarios and the general procedure of quantitative model validation in the presence of uncertainty. Sections 4.3 and 4.4 investigate the aforementioned model validation methods for fully characterized, partially characterized, and uncharacterized validation experiments. The mathematical relationships among some of these validation methods are presented in Section 4.5. Interpretations of the various validation methods are given in Section 4.6. Section 4.7 illustrates the detection of directional bias using Bayesian interval hypothesis testing, the reliability metric, and the area-based metric.

4.2 Scenarios and decision process of model validation under uncertainty

The development of quantitative validation metrics is usually based on assumptions regarding the quantity of interest Y, the corresponding model prediction Y_m, and the experimental observation Y_D. These assumptions relate to the various sources of uncertainty and the types of available validation data. In order to select appropriate validation methods, the first step is to identify the sources of uncertainty and the type of validation data.

As mentioned earlier, the available validation data can be from fully characterized, partially characterized, or uncharacterized experiments. In the case of fully characterized experiments, the model/experimental inputs $\boldsymbol{x}$ are measured and reported as point values. The value of Y corresponding to each of the experiments is a constant (although unknown). Y_m is also a constant if both the model parameters and model form error are treated as constants, otherwise Y_m is stochastic. If the experiment is partially characterized, the uncertainty in $\boldsymbol{x}$ is represented by treating $\boldsymbol{x}$ as a random vector. Since the model output Y_m is considered

Table 4.2: Scenarios of validation and the

Experimental data	Quantity of interest Y	Model prediction Y_m
Fully characterized	Constant Constant	Constant Stochastic
Partially characterized	Constant	Stochastic
Uncharacterized	Stochastic	Stochastic

Note: Y_D is always treated as a random variable due to measurement error

as a function of $\boldsymbol{x}$, Y_m becomes stochastic even with constant model parameters and model form error. Y is still a constant for individual experiments since the epistemic uncertainty in the input does not make the true value stochastic. If the experiments are uncharacterized, the sources of uncertainty in $\boldsymbol{x}$ include the variability across multiple experiments and the lack of precise measurement. Correspondingly, Y needs to be treated as a random variable to account for the variability of the quantity of interest across multiple experiments.

Note that Y_D results from the addition of measurement error to the true value of the physical quantity Y, i.e., $Y_D = Y + \varepsilon_D$, where ε_D represents measurement error. Hence, the uncertainty in the experimental observation (Y_D) can be split into two parts, the variability in the physical system response (Y) and the measurement error in experiments (ε_D). It should be noted that experimental data with poor quality can hardly provide any useful information on the validity of a model. The discussions in this chapter are restricted to the cases when uncertainty in data (due to the error in measuring experimental input and output variables) is limited.

Table 4.2 summarizes the stochasticity of Y, Y_m, and Y_D in different scenarios as discussed above, and the applicability of various validation methods to these scenarios will be presented in Section 4.4.

After selecting a validation method and computing the corresponding metric, another important aspect of model validation is to decide if one should accept or reject the model prediction based on the computed metric and the selected threshold. Section 4.5 and 4.6

will provide some discussions on the decision threshold based on the relationship between various validation metrics. The flowchart in Fig. 4.1 describes a systematic procedure for quantitative model validation.

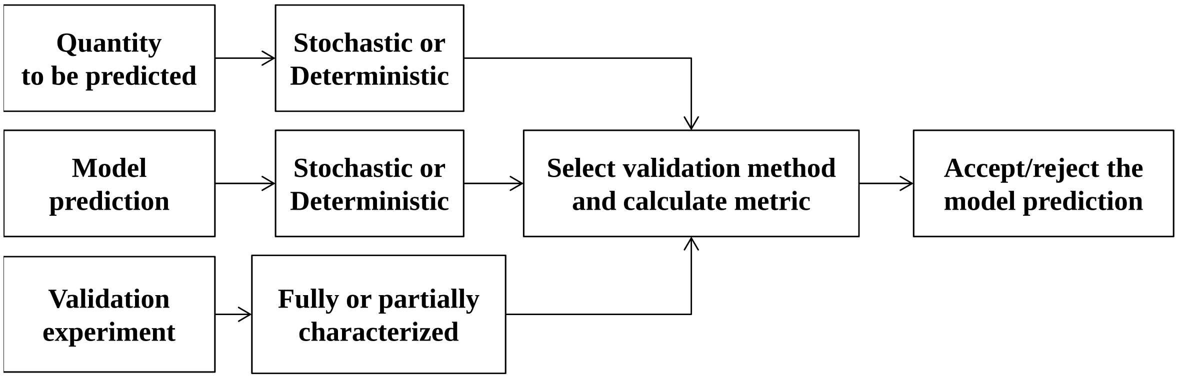


Figure 4.1: Decision process in quantitative model validation (Note: The last two steps involve decision making)

4.3 Development of Bayesian interval/equality hypothesis testing

When both the quantity of interest Y and the corresponding model prediction Y_m are considered to be random variables, the equality and interval hypotheses formulated in [Rebba and Mahadevan, 2008; Rebba et al., 2006] are no longer suitable. In this section, we develop extensions to the equality and interval hypotheses so that they can be applied to the above case: (1) the extended interval hypothesis states that the difference between the means of Y_m and Y, and the difference between the standard deviations of Y_m and Y, are within desired intervals respectively; (2) the extended equality hypothesis states that the PDF of Y_m is equal to the PDF of Y. With the first formulation, it is straightforward to derive the likelihood functions under the null and alternative hypothesis, and the existence of directional bias can be reflected in the test, as will be shown below. The advantages of the second formulation are that it avoids the setting of interval width in the first formulation, and leads to a direct test on probability distributions instead of distribution parameters. For the case that either Y or Y_m is deterministic, the first formulation can still be applicable by setting the standard deviation of the deterministic quantity to be zero; however, the second formulation only applies to the case when both Y and Y_m are stochastic. These two formulations are applicable to both fully characterized and partially characterized experiments. Note that in the case where the model output follows a tail-heavy distribution, formulating hypotheses on higher order moments (instead of the mean and standard deviation) may be necessary in order to assess the validity of the model.

4.3.1 Interval hypothesis on distribution parameters

The interval hypothesis can be formulated as $H_0 : \epsilon_{\mu 1} \leq \mu_m - \mu \leq \epsilon_{\mu 2}, \epsilon_{\sigma 1} \leq \sigma_m - \sigma \leq \epsilon_{\sigma 2}$, and $H_1 : \mu_m - \mu > \epsilon_{\mu 2}$ or $\mu_m - \mu < \epsilon_{\mu 1}, \sigma_m - \sigma > \epsilon_{\sigma 2}$ or $\sigma_m - \sigma < \epsilon_{\sigma 1}$. μ_m and μ are the means of Y_m and Y respectively, and σ_m and σ are the standard deviations of Y_m and Y respectively. $\epsilon_{\mu 1}$, $\epsilon_{\mu 2}$, $\epsilon_{\sigma 1}$ and $\epsilon_{\sigma 2}$ are constants which define the width of interval. Note that $\epsilon_{\mu 1} < \epsilon_{\mu 2}$, $\epsilon_{\sigma 1} < \epsilon_{\sigma 2}$.

Under the interval hypothesis H_0, μ can be any value between $[\mu_m - \epsilon_{\mu 2},\ \mu_m - \epsilon_{\mu 1}]$. So $\mu \sim \text{Unif}(\mu_m - \epsilon_{\mu 2}, \mu_m - \epsilon_{\mu 1})$, and the PDF $\pi_0(\mu|\mu_m) = 1/(\epsilon_{\mu 2} - \epsilon_{\mu 1})$. Similarly, $\sigma \sim \text{Unif}(\sigma_m - \epsilon_{\sigma 2}, \sigma_m - \epsilon_{\sigma 1})$, and the PDF $\pi_0(\sigma|\sigma_m) = 1/(\epsilon_{\sigma 2} - \epsilon_{\sigma 1})$. Thus

$$\begin{aligned} \pi_0(y|\mu_m, \sigma_m) &= \int\int \pi(y|\mu, \sigma)\pi_0(\mu|\mu_m)\pi_0(\sigma|\sigma_m) d\mu d\sigma \\ &= \frac{1}{(\epsilon_{\mu 2} - \epsilon_{\mu 1})(\epsilon_{\sigma 2} - \epsilon_{\sigma 1})} \int_{\sigma_m - \epsilon_{\sigma 2}}^{\sigma_m - \epsilon_{\sigma 1}} \Big\{ \int_{\mu_m - \epsilon_{\mu 2}}^{\mu_m - \epsilon_{\mu 1}} \pi(y|\mu, \sigma) d\mu \Big\} d\sigma \qquad (4.1) \end{aligned}$$

In the presence of measurement error, the experimental observation is a random variable with conditional probability $\Pr(y_D|y)$. Hence, the likelihood function under the null hypothesis

H_0 can be derived as

$$\Pr(y_D|H_0) \;=\; \int \Pr(y_D|y)\pi_0(y|\mu_m, \sigma_m)dy \tag{4.2}$$

Under the alternative hypothesis H_1, μ can be any value outside $[\mu_m - \epsilon_{\mu 2},\, \mu_m - \epsilon_{\mu 1}]$, but the uniform distribution is not applicable to infinite space in practical cases. To avoid this issue, we can assume that the possible values of μ are within a finite interval $[\mu_l,\, \mu_u]$ based on the underlying physics. Therefore $\mu \sim \mathrm{Unif}(\mu_l, \mu_m - \epsilon_{\mu 2}) \cup (\mu_m - \epsilon_{\mu 1}, \mu_u)$, and the PDF $\pi_1(\mu|\mu_m) = 1/(\mu_u - \mu_l + \epsilon_{\mu 1} - \epsilon_{\mu 2})$. Similarly, $\sigma \sim \mathrm{Unif}(\sigma_l, \sigma_m - \epsilon_{\sigma 2}) \cup (\sigma_m - \epsilon_{\sigma 1}, \sigma_u)$, and the PDF $\pi_1(\sigma|\sigma_m) = 1/(\sigma_u - \sigma_l + \epsilon_{\sigma 1} - \epsilon_{\sigma 2})$. thus

$$\begin{aligned}\pi_1(y|\mu_m, \sigma_m) &= \int\int \pi(y|\mu, \sigma)\pi_1(\mu|\mu_m)\pi_1(\sigma|\sigma_m)d\mu d\sigma \\ &= \frac{A}{(\mu_u - \mu_l + \epsilon_{\mu 1} - \epsilon_{\mu 2})(\sigma_u - \sigma_l + \epsilon_{\sigma 1} - \epsilon_{\sigma 2})}\end{aligned} \tag{4.3}$$

where A is calculated as

$$\begin{aligned}A = &\int_{\sigma_l}^{\sigma_m - \epsilon_{\sigma 2}} \Big\{ \int_{\mu_l}^{\mu_m - \epsilon_{\mu 2}} \pi(y|\mu, \sigma)d\mu + \int_{\mu_m - \epsilon_{\mu 1}}^{\mu_u} \pi(y|\mu, \sigma)d\mu \Big\} d\sigma + \\ &\int_{\sigma_m - \epsilon_{\sigma 1}}^{\sigma_u} \Big\{ \int_{\mu_l}^{\mu_m - \epsilon_{\mu 2}} \pi(y|\mu, \sigma)d\mu + \int_{\mu_m - \epsilon_{\mu 1}}^{\mu_u} \pi(y|\mu, \sigma)d\mu \Big\} d\sigma\end{aligned} \tag{4.4}$$

The likelihood function under H_1 can then be derived as

$$\Pr(y_D|H_1) \;=\; \int \Pr(y_D|y)\pi_1(y|\mu_m, \sigma_m)dy \tag{4.5}$$

The Bayes factor for the Bayesian interval hypothesis testing can be calculated by dividing $\Pr(y_D|H_0)$ in Eq. 4.2 by $\Pr(y_D|H_1)$ in Eq. 4.5.

4.3.2 Equality hypothesis on probability density functions

To further validate the entire distribution of Y_m predicted by a probabilistic model, H_0 or H_1 can be formulated correspondingly as the predicted distribution $\pi(y_m)$ being or not being the true distribution of the quantity to be predicted Y, i.e., $H_0 : \pi(y) = \pi(y_m)$, and $H_1 : \pi(y) \neq \pi(y_m)$. The Bayes factor in this case becomes

$$B = \frac{\Pr(y_D|H_0)}{\Pr(y_D|H_1)} = \frac{\int \Pr(y_D|y)\pi_0(y)dy}{\int \Pr(y_D|y)\pi_1(y)dy} \tag{4.6}$$

where $\Pr(y_D|y)$ is the conditional probability of observing noisy data y_D given that the actual value of Y is y; $\pi_0(y)$ is the PDF of Y under the null hypothesis H_0 and hence $\pi_0(y) = \pi(y_m)$; $\pi_1(y)$ is the PDF of Y under the alternative hypothesis H_1. If no extra information about $\pi_1(y)$ is available, it can be assumed as a non-informative uniform PDF. Note that the bounds of this uniform distribution will affect the value of the estimated Bayes factor, and thus it should be carefully selected based on available information.

Note that $\Pr(y_D|y)$ is proportional to the value of the PDF of Y_D conditioned on y which is evaluated at $Y_D = y$, i.e., $\Pr(y_D|y) \propto \pi(y_D|y)$. Therefore, Eq. 4.6 can be rewritten as

$$B = \frac{\int \pi(y_D|y)\pi_0(y)dy}{\int \pi(y_D|y)\pi_1(y)dy} \tag{4.7}$$

4.3.3 Bayesian hypothesis testing with multiple data points

If fully or partially characterized validation experiments are conducted for multiple test input combinations (including replicas), we can first compute individual Bayes factors B_i's ($i = 1, 2, ..., N$) corresponding to each data point based on the methods describe above, and then multiply these individual Bayes factors to obtain an overall Bayes factor $B = \prod_{i=1}^{N} B_i$, assuming model predictions (as well as experimental measurements) corresponding to different data points are independent. If there is a large difference between Bayes factors corresponding to different experiments, we can express these Bayes factors on a logarithmic scale in order to achieve better visualization (e.g., Figs. 5.13 and 5.14).

Another way to obtain an overall Bayes factor is by treating these fully/ partially characterized experiments as uncharacterized. That is to say, the individual measurements or probability distributions of the input $\boldsymbol{x}$ for different experiments are combined into a single probability distribution, and the output measurement data are treated as replicas. This approach is called "ensemble validation" in this dissertation, and the resulting overall Bayes factor in a Bayesian interval/equality hypothesis testing can be calculated as

$$B = \frac{\int \big[\prod_{i=1}^{N} \pi(y_D^i|y)\big]\pi_0(y)\ \mathrm{d}y}{\int \big[\prod_{i=1}^{N} \pi(y_D^i|y)\big]\pi_1(y)\ \mathrm{d}y} \tag{4.8}$$

If the assumption of independence between model predictions at different input points does not hold, a more rigorous approach would be to account for the statistical correlation between different data points in calculating the Bayes factor [Sankararaman, 2012], where the model prediction Y_m is treated as a random process over the input field $\boldsymbol{x}$. The equality null hypothesis H_0: $y|\boldsymbol{x}_i = y_m|\boldsymbol{x}_i, i = 1, 2, ..., N$ approximates the quantity of interest Y with a random process Y_m, and the alternative hypothesis becomes H_1: $y|\boldsymbol{x}_i \neq y_m|\boldsymbol{x}_i, i = 1, 2, ..., N$. The likelihoods of H_0 and H_1 aggregating multiple data points $\boldsymbol{y_D}$ can be obtained as [Sankararaman, 2012]

$$\begin{aligned}
\mathcal{L}(H_0) &\propto \Pr(\boldsymbol{y_D}|H_0)\\
&= \int \Big(\int \big[\prod_{i=1}^{N} \pi(y_D^i|\boldsymbol{x}_i, \boldsymbol{\theta})\pi(\boldsymbol{x}_i)\big]\mathrm{d}\boldsymbol{x}_i\Big)\pi(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}\\
\mathcal{L}(H_1) &\propto \Pr(\boldsymbol{y_D}|H_1)\\
&= \int \Big(\prod_{i=1}^{N} \pi(y_D^i|y_i)\Big)\pi_1(\boldsymbol{y})\ \mathrm{d}\boldsymbol{y} \qquad (4.9)
\end{aligned}$$

4.4 Validation with fully characterized, partially characterized, or uncharacterized experimental data

As discussed in Section 4.2, both the quantity of interest Y and the corresponding model prediction Y_m are stochastic when uncharacterized experimental data are considered. In such cases, all the five methods illustrated in Sections 2.3 and 4.3 can be used as long as the probability distribution of Y_m are available. In order to obtain the probability distribution of Y_m, we need to first construct the probability distributions of $\boldsymbol{x}$ based on the reported intervals or expert opinions [Sankararaman and Mahadevan, 2011a]. Assuming that the probability density function of $\boldsymbol{x}$ is $\pi(\boldsymbol{x})$, one can calculate the unconditional PDF of model prediction $\pi(y_m)$ via propagating uncertainty from $\boldsymbol{x}$ to model output Y_m as

$$\pi(y_m) = \int [(G(\boldsymbol{x}) = y_m)]\pi(\boldsymbol{x})d\boldsymbol{x} \tag{4.10}$$

where $G(\boldsymbol{x})$ denotes the model. The above distribution of Y_m can then be substituted into the various methods to compute the corresponding validation metrics.

The probability distribution of Y_m in the case of partially characterized experiments can also be obtained using the procedure described above. Below We will discuss the use of various validation methods with fully and partially characterized experimental data.

Classical hypothesis testing Classical hypothesis testing can be applied to fully or partially characterized experiments by comparing the data against the model predictions. Since Y is constant given input $\boldsymbol{x}$, the standard deviation σ in the equations shown in Section 2.3.1.1 becomes zero; if Y_m is constant, σ_m becomes zero. The computation procedure of p-value remains the same.

Bayesian interval hypothesis testing The interval assumption for the case when Y is constant will only be made on μ and μ_m, since we know σ is zero. For fully characterized experiments, if Y_m is constant, σ_m is set to be zero and the rest of the computation remains the same; if Y_m is stochastic, μ_m and σ_m are computed from probability distribution of Y_m given the measured $\boldsymbol{x}$. For partially characterized experiments, Y_m is always stochastic, and μ_m and σ_m are computed from probability distribution of Y_m obtained in Eq. 4.10.

Bayesian equality hypothesis testing As showed in Section 4.3.2, the extended equality hypothesis treats both Y and Y_m as random variables. However, as discussed in Section 4.2, Y is in fact a constant given a specific input $\boldsymbol{x}$ when the validation experiments are fully or partially characterized. The equality hypothesis that $\pi(y) = \pi(y_m)$ can be viewed as an approximation of the unknown constant Y using a random variable. In other words, the epistemic uncertainty about Y is represented by a subjective probability distribution following the Bayesian way of thinking.

However, for fully characterized experiments, if both Y and Y_m are constant (i.e., $Y = y$, $Y_m = y_m = G(\boldsymbol{x})$), we can modify the equality hypothesis to be $y = y_m = G(\boldsymbol{x})$ instead of $\pi_0(y) = \pi(y_m)$. Let the alternative hypothesis be $y = y_1$, the corresponding Bayes factor is

$$B = \frac{\pi\big(y_D|y = G(x)\big)}{\pi(y_D|y = y_1)} \tag{4.11}$$

Reliability metric The computation of the reliability metric only requires the probability distributions of Y_D and Y_m. Therefore, the reliability-based validation method can be applied to the cases when the validation experiments are fully/partially characterized. Since the quantity of interest Y is constant, σ is set be zero while computing $\sigma^2_{Y_D}$; when the model prediction Y_m is constant, σ_m is set to be zero.

Area-based metric Similar to the reliability metric, calculating the area-based metric only requires the probability distributions of Y_D and Y_m, and thus it can also be applied to the cases when the validation experiments are fully/partially characterized. When the model prediction Y_m is constant, the area metric-based method can be used by considering the model output to follow a degenerate distribution, i.e., $F_{Y_m}(y) = 0$ for $y < y_m$, $F_{Y_m}(y) = 1$ for $y \geq y_m$.

4.5 Relationship between p-value, Bayes factor, and the reliability metric

4.5.1 Relationship between p-value and Bayes factor

Although the p-value in classical hypothesis testing and the Bayes factor B are based on different philosophical assumptions and formulated differently, it has been shown that these two metrics can be mathematically related for some special cases [Rouder et al., 2009]. In the discussion below, the Bayes factor based on the hypothesis of probability density functions for a fully characterized experiment is found related to the p-value in t-test and z-test, if the model prediction Y_m is a normal random variable with mean μ_m and standard deviation σ_m.

Starting from the formula of Bayes factor in Eq. 4.7, since we assume that the PDF of the quantity to be predicted Y under the alternative hypothesis H_1 is uniform, the integration term in the denominator is not affected by the target model and thus can be treated as a constant $1/C$. Based on the null hypothesis H_0, the quantity of interest $Y \sim N(\mu_m, \sigma_m^2)$. Recall the relationship $Y_D = Y + \varepsilon_D$, and $\varepsilon_D \sim N(0, \sigma_D^2)$, we know that $Y_D \sim N(\mu_m, \sigma_m^2 + \sigma_D^2)$. Thus the numerator of Eq. 4.7 can be calculated as

$$\int \pi(y_D|y)\pi_0(y|\boldsymbol{x})dy = \frac{1}{\sqrt{\sigma_m^2 + \sigma_D^2}}\phi(\frac{y_D - \mu_m}{\sqrt{\sigma_m^2 + \sigma_D^2}}) \tag{4.12}$$

where $\phi(*)$ is the PDF of the standard norm random variable.

If the variance of measurement noise is negligible compared to the variance of Y_m, i.e., $\sigma_D^2 \ll \sigma_m^2$, we have $\sigma_m^2 + \sigma_D^2 \approx \sigma_m^2$. Also note that for a single data point $\bar{Y}_D = y_D$. Therefore Eq. 4.7 becomes

$$B = \frac{C}{\sigma_m} * \phi(\frac{\bar{Y}_D - \mu_m}{\sigma_m}) \tag{4.13}$$

Based on Eqs. 2.14 and 2.16, we have

$$\bar{Y}_D - \mu_m = \begin{cases} t * S_D/\sqrt{n} & \text{, for } t\text{-test} \\ z * \sigma_{Y_D}/\sqrt{n} & \text{, for } z\text{-test} \end{cases} \tag{4.14}$$

Substituting Eq. 4.14 into Eq. 4.13, we obtain

$$B = \begin{cases} C/\sigma_m * \phi[(t * S_D)/(\sigma_m\sqrt{n})] & \text{, for } t\text{-test} \\ C/\sigma_m * \phi[(z * \sigma_{Y_D})/(\sigma_m\sqrt{n})] & \text{, for } z\text{-test} \end{cases} \tag{4.15}$$

where ϕ is the probability density function of a standard normal variable.

From Eq. 4.15, we can see that the Bayes factor can be related to either the z statistic or the t statistic, and hence it can be related to the p-value in both z-test and t-test. Combining Eqs. 2.17 and 4.15, we obtain the relation between Bayes factor and the p-value in the z-test as

$$B = \frac{C}{\sigma_m} * \phi[\Phi^{-1}(\frac{p}{2})\frac{\sigma_{Y_D}}{\sigma_m\sqrt{n}}] \tag{4.16}$$

where Φ^{-1} is the inverse standard normal CDF. Similarly, the relation between Bayes factor and the p-value in the t-test can be obtained by combining Eqs. 2.15 and 4.15 as

$$B = \frac{C}{\sigma_m} * \phi\{[S_D * F_{T,n-1}^{-1}(\frac{p}{2})]/(\sigma_m\sqrt{n})\} \tag{4.17}$$

where $F_{T,n-1}^{-1}$ is the inverse CDF of a t-distribution with $(n-1)$ degrees of freedom.

If the chosen significance level in z-test or t-test is α, the corresponding threshold Bayes factor B_{th} can be calculated using Eq. 4.16 or 4.17 by letting $p = \alpha$. In that case, the z-test/t-test with significance level α and Bayesian hypothesis testing with the corresponding threshold value B_{th} will give the same model validation result.

4.5.2 Relationship between p-value and the reliability metric

Recall that experimental data are considered as samples of the random variable Y_D. If the size of experimental data set (n) is relatively large, e.g., $n > 30$, the sample variance S_D^2 can be assumed to be a good estimator of $\sigma_{Y_D}^2$ (the true variance of Y_D), which is needed to compute the reliability metric. If n is small and no prior information on σ is available, we can assume that $\sigma = \sigma_m$, which is the same assumption used in z-test. By assuming further that the mean of validation data $\bar{Y}_D$ is equal to μ, Eq. 2.24 can be re-written as

$$r \;=\; \Phi[\frac{\epsilon - (\bar{Y}_D - \mu_m)}{\sigma_d}] - \Phi[\frac{-\epsilon - (\bar{Y}_D - \mu_m)}{\sigma_d}] \tag{4.18}$$

By substituting Eq. 4.14 into Eq. 4.18, the relation between the reliability-based metric r and the test statistic in the t-test or z-test is obtained as

$$r \;=\; \begin{cases} \Phi[(\epsilon - t * S_D/\sqrt{n})/\sigma_d] + \Phi[(\epsilon + t * S_D/\sqrt{n})/\sigma_d] - 1, \text{ for } t\text{-test} \\ \Phi[(\epsilon - z * \sigma_{Y_D}/\sqrt{n})/\sigma_d] + \Phi[(\epsilon + z * \sigma_{Y_D}/\sqrt{n})/\sigma_d] - 1, \text{ for } z\text{-test} \end{cases} \tag{4.19}$$

By combining Eqs. 2.15, 2.17 and 4.19, the reliability-based metric can be further related to the p-value in the t-test or z-test as

$$r = \begin{cases} \Phi[(\epsilon - F_{T,n-1}^{-1}(p/2) * S_D/\sqrt{n})/\sigma_d]+ \\ \qquad \Phi[(\epsilon + F_{T,n-1}^{-1}(p/2) * S_D/\sqrt{n})/\sigma_d] - 1, , \text{ for } t\text{-test} \\ \Phi[(\epsilon - \Phi^{-1}(p/2) * \sigma_{Y_D}/\sqrt{n})/\sigma_d]+ \\ \qquad \Phi[(\epsilon + \Phi^{-1}(p/2) * \sigma_{Y_D}/\sqrt{n})/\sigma_d] - 1, , \text{ for } z\text{-test} \end{cases} \tag{4.20}$$

If one chooses to test models based on a threshold reliability value r_{th} calculated by letting $p = \alpha$ in Eq. 4.20 above, the result of model validation will be the same as that in the t-test or z-test with significance level α.

4.6 Interpretation of quantitative model validation results

Classical hypothesis testing Failing to reject H_0 indicates that the accuracy of the model is acceptable, but it does not prove that H_0 is true. The comparison between p-value and significance level α becomes meaningless when the sample size of experimental data is large. Since almost no null hypothesis H_0 is true, the p-value will decrease as the sample size increases, and thus H_0 will tend to be rejected at a given significance level α as the sample size grows large [Marden, 2000]. In addition, the over-interpretation of p-value and the corresponding significance testing result can be misleading and dangerous for model validation. Criticisms on over-stressing p-value and significance level can be found in [Ziliak and McCloskey, 2008; Ambaum, 2010].

Bayesian hypothesis testing If the Bayes factor calculated is greater than 1, it is indicated that the data favor the null hypothesis; if the Bayes factor is less than 1, it is indicated that the data favor the alternative hypothesis. In addition, Jeffreys [Jeffreys, 1983] gave a heuristic interpretation of Bayes factor in terms of the level of support that the hypotheses obtain from data. The threshold value of Bayes factor B_{th} can be related to the so-called Bayes risk in detection theory [Kay, 1998; Jiang and Mahadevan, 2007], which is the sum of costs due to different decision scenarios - failing to reject the true/wrong hypothesis and rejecting the true/wrong hypothesis. It has been shown that appropriate selection of B_{th} can help reduce the Bayes risk [Kay, 1998]. If one assumes that the cost of making correct decisions (failing to reject the true hypothesis or rejecting the wrong hypothesis) is zero, the costs of type I and type II error are the same, and the prior probabilities of the null and alternative hypothesis being true are equal, then the resulting $B_{th} = 1$ [Jiang and Mahadevan, 2007]. However, It should be noted that as a part of the decision making process, the choice of thresholds for Bayes factor inevitably contains subjective elements.

Before collecting validation data, there may be no evidence to support or reject the model. In such cases, it may be reasonable to assume that the prior probabilities of the null hypothesis and alternative hypothesis are equal (= 0.5), and thus a simple expression of the posterior probability of the null hypothesis can be derived in terms of the Bayes factor [Rebba et al., 2006], which is a convenient metric to assess the confidence in model prediction:

$$\begin{aligned}
\Pr(H_0|y_D) &= \frac{\Pr(y_D|H_0)\Pr(H_0)}{\Pr(y_D|H_0)\Pr(H_0) + \Pr(y_D|H_1)\Pr(H_1)} \\
&= \frac{\Pr(y_D|H_0)}{\Pr(y_D|H_0) + \Pr(y_D|H_1)} \\
&= \frac{B}{1+B}
\end{aligned} \tag{4.21}$$

An advantage of Bayesian hypothesis testing is that the posterior probabilities of H_0 and H_1 obtained from the validation exercise can both be used through a Bayesian model-averaging approach [Hoeting et al., 1999; Zhang and Mahadevan, 2000, 2003] to reflect the effect of the model validation result on the uncertainty in model output, as shown in Eq. 4.22

$$\bar{\pi}(y) = \pi_0(y)\mathrm{Pr}(H_0|y_D) + \pi_1(y)\mathrm{Pr}(H_1|y_D) \tag{4.22}$$

where $\bar{\pi}(y)$ is the predicted PDF of Y combining the PDFs of Y under the null and alternative hypothese. Therefore, instead of completely accepting a single model, one can include the risk of using this model in further calculations. This helps to avoid both Type I and Type II errors, i.e., accepting an incorrect model or rejecting a correct model.

Reliability metric The threshold r_{th} used in the reliability-based method represents the minimum probability of the difference d falling within an interval $[-\epsilon, \epsilon]$, and the decision of accepting/rejecting a model can be made based on the decision maker's acceptable level of model reliability.

Area-based metric Different from the validation metrics in hypothesis testing methods and the reliability-based method, the area metric has no probability interpretation; it is the difference between two CDFs; its physical unit is the same as for the quantity of interest (Y), and thus the area metric can be viewed as a direct measure of prediction error.

The area-based metric defined in u-space based on Eq. 2.25 can be transformed back to physical space to retrieve its physical interpretation. As suggested in [Ferson et al., 2008], one can use the CDF of model output (G_y) at a certain point to perform a back-transformation: $y_i = G_y^{-1}(u_i)$, and then compute the area metric in the physical space

$$d(F_y, G_y) = \int |F_y - G_y| \; \mathrm{d}y \tag{4.23}$$

where y_i is the transformed variable with the physical unit of the quantity of interest, and F_y is the empirical CDF of y_i.

Since the area metric has the physical unit of the quantity of interest and represents the prediction error of a model, the threshold of model rejection/acceptance can be set up based on the error tolerance limit in the prediction domain.

4.7 Detection of the directional bias

Bayesian interval hypothesis testing The directional bias defined in Section 4.1 can be captured by conducting two separate Bayesian interval hypothesis tests. In the first test, we set $\epsilon_{\mu 1} = -\epsilon_\mu$ and $\epsilon_{\mu 2} = 0$, and thus under the null hypothesis $-\epsilon_\mu \leq \mu_m - \mu \leq 0$. In the second test, we set $\epsilon_{\mu 1} = 0$ and $\epsilon_{\mu 2} = \epsilon_\mu$, and thus under the null hypothesis $0 \leq \mu_m - \mu \leq \epsilon_\mu$. The model will fail if any of these two null hypotheses fails the corresponding test. Therefore, the existence of directional bias will increase the chance of a model to fail the combined test. Fig. 4.2 illustrates this combined test using the concept of data space. Suppose Z is the overall validation data space, Z_1 is the set of data which does not support the model in the first Bayesian interval hypothesis test, and Z_2 is the set of data which does not support the model in the second test. Then, the union of Z_1 and Z_2 is the set of data that does not support the model combining these two tests.

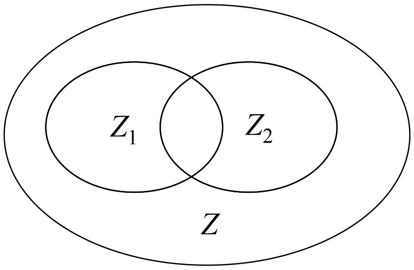


Figure 4.2: Graphical illustration of the combined test

Reliability-based metric Since the reliability-based metric is the probability of d being within a given interval, it can also reflect the existence of directional bias by modifying the intervals. Similar to the Bayesian interval hypothesis testing, we can take two different intervals $[0, \epsilon]$ and $[-\epsilon, 0]$, and calculate the corresponding values of metric r^1 and r^2 as:

$$\begin{aligned} r^1 &= \Phi[\frac{\epsilon - (\mu - \mu_m)}{\sigma_d}] - \Phi[\frac{-(\mu - \mu_m)}{\sigma_d}] \\ r^2 &= \Phi[\frac{-(\mu - \mu_m)}{\sigma_d}] - \Phi[\frac{-\epsilon - (\mu - \mu_m)}{\sigma_d}] \end{aligned} \tag{4.24}$$

By comparing the values of r^1 and r^2 against the threshold $r_{th}/2$ (half of the original threshold value because the width of intervals considered is half of the original one), the model may be judged to have failed the validation test if either r^1 or r^2 is less than $r_{th}/2$.

The area-based metric The area-based metric can reflect the existence of directional bias, i.e., when the experimental observations are consistently below or above the corresponding mean predictions of numerical model. For example, if the model outputs at different test combinations are normal random variables, the values of $F^m_{\boldsymbol{xi}}(y_{Di})$ will all be less than 0.5 if y_{Di}'s are smaller than the mean of the corresponding normal variables. Therefore, instead of being uniformly distributed between [0,1], u_i's are distributed between [0,0.5], causing a large area between the empirical CDF of u_i and the standard uniform CDF.

4.8 Conclusion

This chapter explored various quantitative validation methods, including classical hypothesis testing, Bayesian hypothesis testing, a reliability-based method, and an area metric-based method, in order to validate computational model prediction.

A Bayesian interval hypothesis testing-based method was formulated, which validates the accuracy of the predicted mean and standard deviation from a model, taking into account the existence of directional bias. Further, Bayesian hypothesis testing to validate the entire PDF of model prediction was formulated. These two formulations of Bayesian hypothesis testing can be used in the case when multiple validation points are available. We also discussed the applicability of the various validation methods to the cases when validation data are collected from fully characterized, partially characterized, or uncharacterized experiments.

It is shown that while the classical hypothesis testing is subject to type I and type II error, the Bayesian hypothesis testing can minimize such risk by (1) selecting a risk-based threshold, and (2) subsequent model averaging using posterior probabilities. It is observed that under some conditions, the p-value in the z-test or t-test can be mathematically related to the Bayes factor and the reliability-based metric.

The area metric in the transformed probability space (u-space) is shown to be sensitive to the direction of bias between model predictions and experimental data, and so are the Bayesian interval hypothsis testing-based method and the reliability-based method. The Bayesian model validation result and reliability-based metric can be directly incorporated in long-term failure and reliability analysis of the device, thus explicitly accounting for model uncertainty [Sankararaman, 2012]. In addition, due to the use of likelihood function in the Bayesian hypothesis testing, the Bayesian model validation method can be extended to the case that the validation data is in the form of interval, as shown in [Sankararaman and Mahadevan, 2011a,b].

CHAPTER V

TIME-DEPENDENT RELIABILITY ANALYSIS: APPLICATION TO MULTI-PHYSICS MEMS DEVICE

5.1 Introduction

As introduced in Chapter II, time-dependent reliability analysis for a complicated engineering system may involve the use of many physics-based models and surrogate models, which bring in various types of uncertainty. This chapter develops a Bayesian network-based probabilistic framework, which integrates model calibration, model validation, and surrogate modeling techniques, in order to address the issues listed below.

Quantification of uncertainty in model parameters and model form The estimation of model parameters is subject to various sources of uncertainty as outlined in [Kennedy and O'Hagan, 2001]. Quantifying the uncertainty due to the use of imprecise physics model is in particular challenging. Bayesian calibration under the KOH framework introduced in Section 2.2.1 has the capability to account for the various sources of uncertainty in the construction of the posterior probability distribution of model parameters, with the inclusion of a model discrepancy term to represent model form uncertainty.

Quantification of uncertainty in model predictive capability By applying quantitative model validation methods illustrated in Chapter IV, the predictive capability of physics-based or surrogate models are assessed. The results of model validation can be used to inform the decision of accepting or rejecting the models. Bayesian hypothesis testing 2.3.1.2 and the reliability-based method 2.3.2.1 can further provide probabilistic measure of confidence in model prediction.

Propagation of uncertainty from the component level to the system level By performing model calibration and model validation, uncertainty in the component (or model) level can be quantified. As illustrated in Sections 2.4 and 2.5, the cost of propagating uncertainty from the component level to the system level can be prohibitive in a time-dependent system. Therefore, efficient surrogate modeling techniques, such as Gaussian process (GP) interpolation and polynomial chaos expansion (PCE), will be used to facilitate the propagation of uncertainty and obtain the probability of failure.

Quantification of uncertainty in system level reliability prediction If validation experimental data are available at the system level, the prediction of failure probability can be assess again by using the quantitative model validation methods. As mentioned above, we can obtain a probabilistic measure of prediction confidence via Bayesian hypothesis testing or the reliability-based method.

Background of application Despite many desirable features such as the superior performance provided in terms of signal loss and isolation compared with silicon devices [Rebeiz, 2003], microelectromechanical system (MEMS) devices have been known for the existence of large variability in the performance and life time due to difficulty in controlling the manufacturing process [Guo et al., 2010]. In addition, multiple physical phenomena behind the behavior of MEMS devices have not been fully understood. Therefore, quantification of aleatory and epistemic uncertainty emerges as an essential part in the design and reliability analysis of MEMS devices.

In this chapter, we consider a contacting radio frequency (RF) capacitive MEMS switch as the target device. The overall prediction goal is the device life after a certain period of usage. A variety of degradation mechanisms have been identified for this type of device, including dielectric charging caused by trapped charges inside the thin dielectric layer, contact area damage and wear, mechanical creep development at very high or very low temperatures, etc.

Two performance metrics that relate to these failure mechanisms, pull-in and pull-out voltage, are used to predict the failure probability. The prediction of pull-in and pull-out voltage involves the simulation of the device displacement as a function of time, which is further based on multiphysics and multiscale modeling, such as microscale dynamic structural model, fluid damping model, creep model, dielectric charging model, and mesoscale contact model. Besides the length scale, the time scale used in different models can also be different. In addition to these physics models, experimental data with different forms (point/interval/time series), and corresponding to different physics and different generations of devices have also been collected.

In Section 5.2, we construct an overall Bayesian network, which includes the physics models involved in different levels of the system, and the corresponding experimental data. Using the overall BN as a platform, we apply the Bayesian calibration methods developed in Chpater III to the calibration of multiple physics models related to the failure mechanisms of RF MEMS switches, as will be showed in Section 5.3. In Section 5.4.1, the various quantitative model validation methods discussed in Chapter IV are used to validate a gas damping model. The Bayesian equality hypothesis testing illustrated in Section 4.3.2 is also applied to assess the predictability of several other physics models using fully and partially characterized validation data. Further, the failure probability of the target device is computed based on the the calibrated and validated models, as will be showed in Section 5.5.

5.2 Construction of a Bayesian network based on multi-scale and multi-physics models

The target MEMS device of this chapter is used as a capacitive switch. The moving top electrode (modeled as membrane/beam) deflects under some applied voltage, and will be pulled into contact with the dielectric pad when the applied voltage exceeds a certain threshold (pull-in voltage V_{pi}). The contact between top electrode and dielectric pad forms a large capacitance, and thus presents a small impedance in an AC circuit. The top electrode will return to the original position when the applied voltage is reduced to a certain threshold (pull-out voltage V_{po}). Both pull-in and pull-out voltages are important metrics in the reliability analysis of the device after a certain period of usage. Prediction of pull-in/pull-out voltage involves the use of multiple physics models, including electrostatic model, dielectric charging model, damping model, contact model, creep model, and structural model. The electrostatic model takes applied voltage as inputs, and calculates electrostatic loading as output. The dielectric model [Palit and Alam, 2012] predicts the number of charges trapped in the dielectric pad after contact, which can affect the electrostatic field. The gas squeeze-film damping model provides an expression for the damping coefficient, which can be further used to compute damping force [Guo and Alexeenko, 2009]. The contact model [Kim et al., 2012] models the dynamic interaction between the top electrode and the dielectric pad based on molecular dynamics (MD) simulation. The Coble creep model [Coble, 1963; Hsu et al., 2011] calculates the plastic deformation of the device under long-term loading. The electrostatic force, damping force, contact force, creep deformation are the inputs to the device level simulation. Two structural models with different fidelity can be considered to predict the dynamic behavior of the device: (1) a 1-D Euler-Bernoulli beam model [Ayyaswamy and Alexeenko, 2010], and (2) a 3-D membrane model [Das et al., 2012].

Combining all the aforementioned physics models, a schematic illustration of the overall BN is shown in Fig. 5.1. Because this BN is constructed based on physics models, the conditional probability associated with each directed edge can be obtained by considering the physical relationships between model input and output variables.

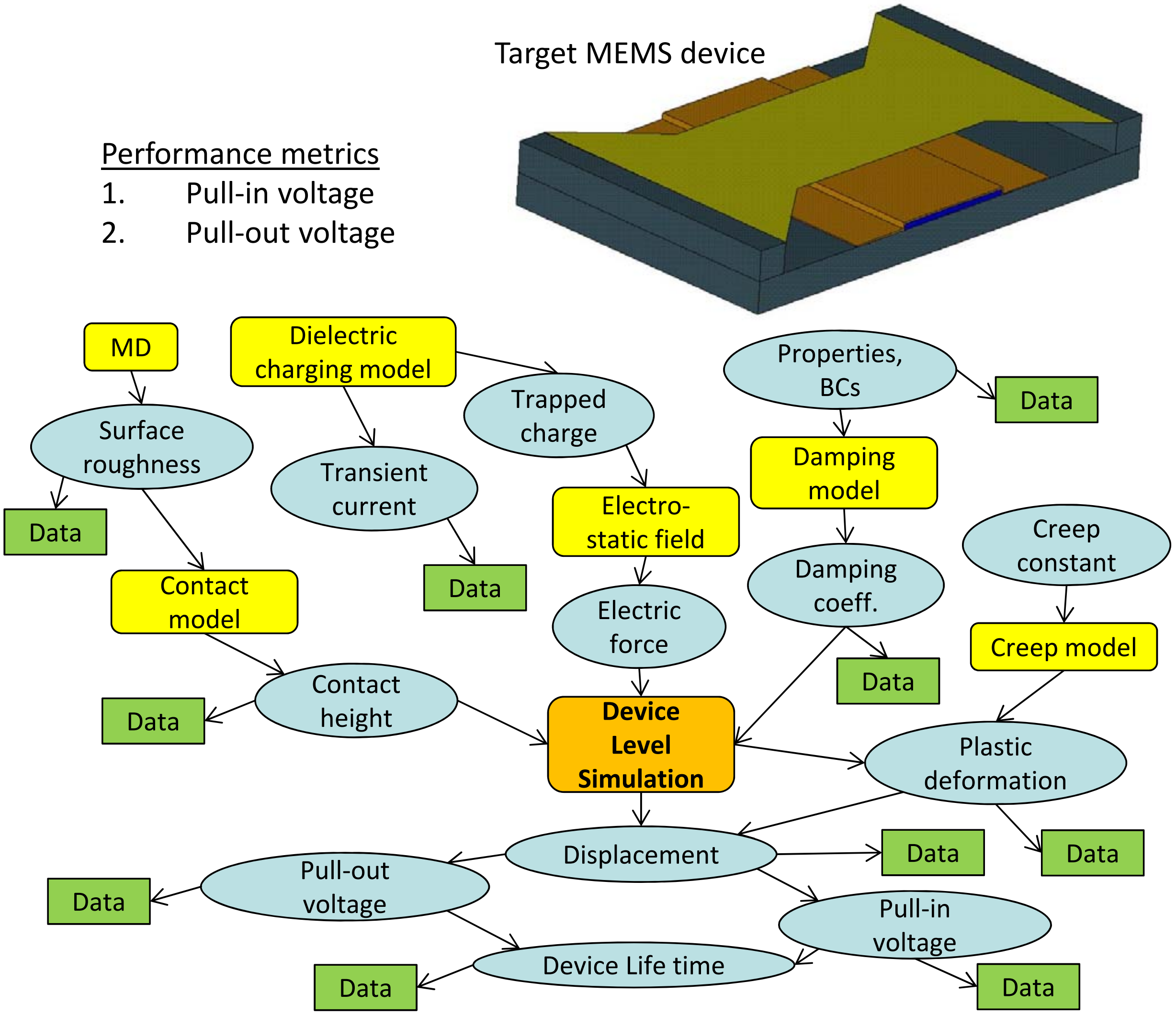


Figure 5.1: Bayes network integration of various models and data for device-level uncertainty quantification

5.3 Bayesian network-based model calibration

We illustrate the model calibration methods presented in the previous chapters using two numerical examples. In Section 5.3.1, we calibrate a dielectric charging compact model [Palit and Alam, 2012] with time series data to illustrate the Bayesian approach discussed in

Section 2.2.3 and the formulations of model discrepancy proposed in Section 3.2.1. In Section 5.3.2, we implement Bayesian calibration of multi-physics models, based on the methods presented in Section 2.2.2 and 3.4. Two types of radio-frequency (RF) microelectromechanical system (MEMS) devices are used in this example. Examination of model parameter identifiability is performed in both examples using the first-order Taylor series expansion-based method proposed in Section 3.3.

5.3.1 Calibration of dielectric charging parameters using a compact model

Dielectric charging has been identified as an important failure mechanism of RF MEMS switches, causing the switches to either remain stuck or fail to actuate [Jain et al., 2011]. In this section, we will focus on the calibration of a compact dielectric charging model developed in [Palit and Alam, 2012]. The model has three input variables (voltage V, temperature T, and time t), seven unknown parameters (trap density N_T, barrier height Φ_B, capture cross section σ, Frenkel-Poole (FP) attempt frequency γ, high frequency dielectric constant ε_{INF}, effective mass m^*, and trap activation energy E_A), and a single output variable (transient current density J). Experiments were conducted on a 200-nm silicon nitride (Si_3N_4) dielectric with 2 mm*2 mm area for 12 different combinations of V and T, and the transient current density was measured at about 190 discrete time points between 0 and 100 seconds. These experiments were repeated four times, and thus a data set with size $n = 12 * 4 * 190 = 9120$ is available. Assuming that the measurement error depends only on V and T, we can estimate the variance of measurement error for a particular combination of V and T using Eq. 2.12.

We again consider the various options of model discrepancy discussed in Section 3.2.1, and the same notations are used as in Section 3.2.3 (i.e., δ_1, δ_2, ...). Note that we also include δ_5 as a Gaussian process with non-stationary covariance function. The mean function and variance function of δ_3 are chosen to be linear functions of the model input combined with exponential functions of time as

$$\begin{aligned}\mu_{\delta_3} &= \phi_1 V + \phi_2 T + \phi_3 t + \phi_4 \exp(\phi_5 t)\\ \sigma^2_{\delta_3} &= \varphi_0 + \varphi_1 V + \varphi_2 T + \varphi_3 t + \varphi_4 \exp(\varphi_5 t)\end{aligned} \tag{5.1}$$

The same form of mean function is used in δ_4 and δ_5. We still use a squared exponential function as the covariance function of δ_4, and Eq. 3.10 is used as the covariance function of δ_5. Based on the first-order Taylor series expansion-based method developed in Section 3.3, a check of identifiability is performed while selecting the form of the mean function μ_{δ_3}, since the addition of the discrepancy function to the original model may cause non-identifiability. In this example, there are 12 unknown parameters in $G + \mu_{\delta_3}$, i.e., $p = 12$, and the corresponding matrix $\boldsymbol{A}$ is full rank, which suggests that the combination of the dielectric model and this mean function is identifiable. In fact, the reason that there is no constant term in the mean function is because the constant term is found to be unidentifiable.

Note that the model needs to be evaluated at all calibration experiment input points in order to compute the likelihood function for a given parameter set. Due to limited computational resources, we use only a subset of the experimental data (40 time points for each combination of V and T; total number of data points $= 12 * 40 = 480$). As discussed in Section 3.2.2, we further partition the selected data set into two sets (each with 240 data points), one for calibration and the other for validation.

Due to the high number of parameters (up to 19), we use the Metropolis-Hastings Markov chain Monte Carlo (MCMC) method [Hastings, 1970] to generate samples from the posterior distribution. 500,000 samples are collected and the fitted marginal posterior PDFs of the dielectric charging model parameters are shown in Fig. 5.2. Note that calibration using δ_1 cannot be implemented in this example because the likelihood function remains zero for any set of parameters, which suggests that the difference between the model and data is

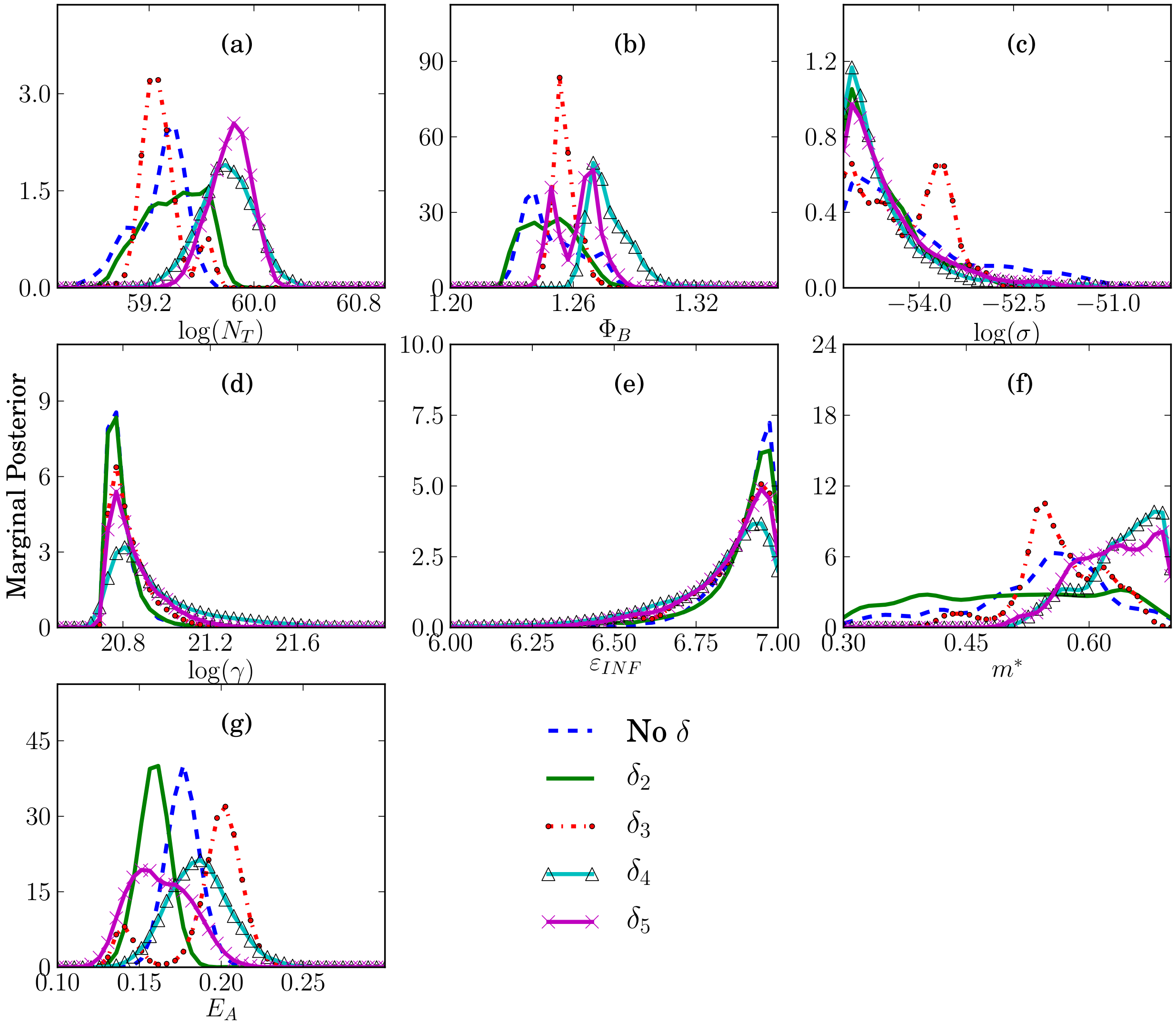


Figure 5.2: Marginal PDFs of dielectric charging model parameters

input-dependent and the assumption of constant model error is not valid. It can be observed that the various choices of model discrepancy give significantly different posteriors for some parameters. However, it is not clear which option is better until we perform the validation step.

The aforementioned validation data set (240 data points) is used to assess the predictive capability of the calibrated models. A graphical comparison between model predictions and data is shown in Fig. 5.3 (a)-(e), and the corresponding reliability metrics are shown in Table 5.1.

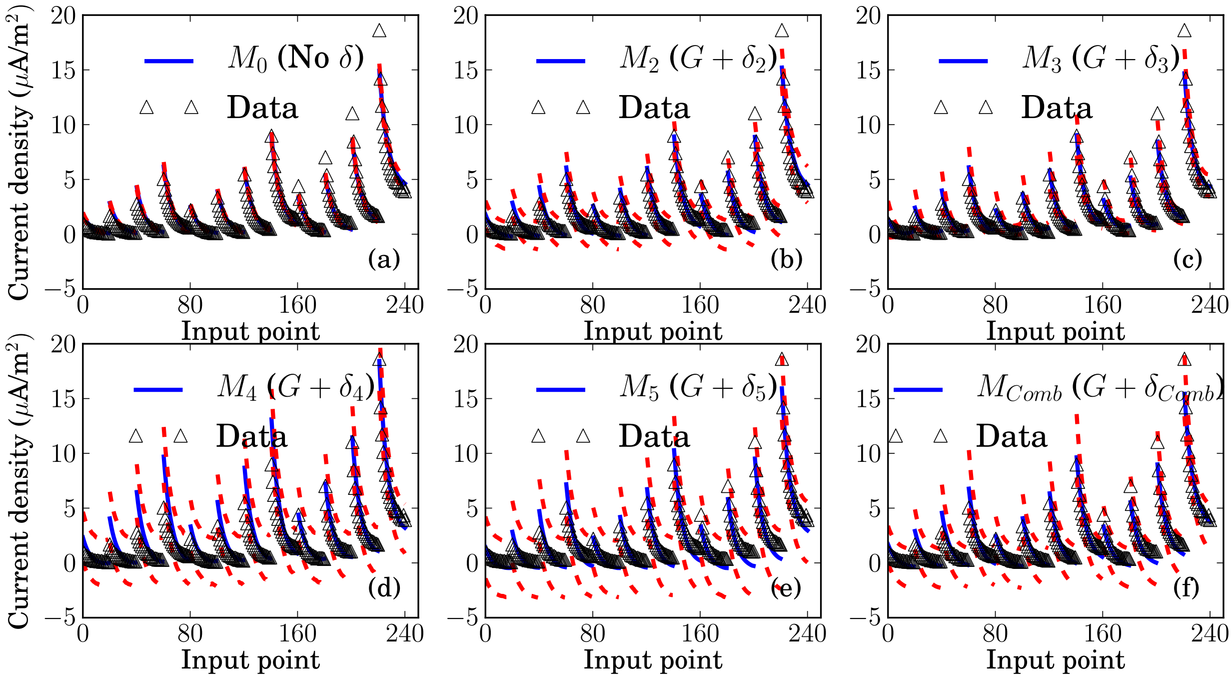


Figure 5.3: Comparison between the predictions of calibrated dielectric charging model and validation data

The graphical comparison and the validation metric suggest that calibration without *delta* or with δ_3 (input-dependent Gaussian random variables) gives the best posterior prediction of current density. However, it can be observed from Fig. 5.3 that the 95% probability bounds of M_0 and M_3 are both small, whereas the probability bounds of M_2, M_4, and M_5 give better coverage of the validation data. Thus, it is preferable to conduct the third step of the proposed method in Section 3.2.2.2, which combines the various posteriors of model parameters and model discrepancy based on the validation results. The model prediction based on the combined distribution of model parameters and δ is shown in Fig. 5.3 (f), of which the probability bound also gives a good coverage of the data.

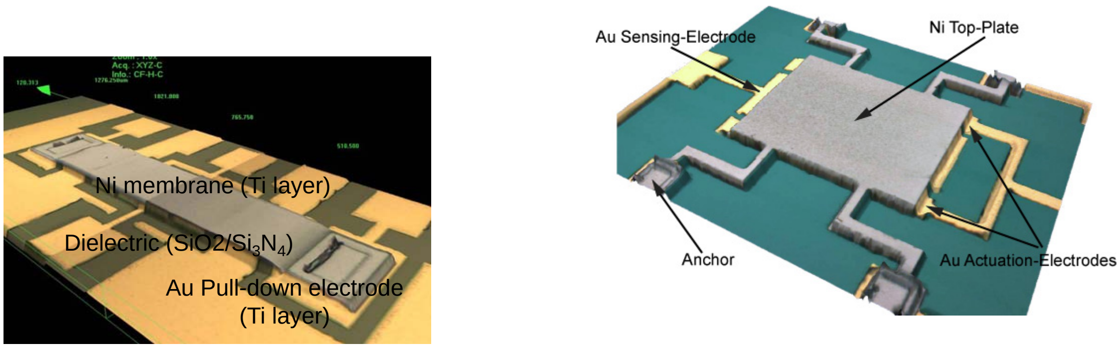
Table 5.1: Overall reliability of model predictions

	M_0 (No δ)	M_1	M_2	M_3	M_4	M_5
μ_r^i	0.86	-	0.58	0.82	0.37	0.36

Discussion In this numerical example, a complicated physics model with little information on the model error is calibrated with experimental data. We applied the first-order Taylor series expansion-based method in order to avoid parameter non-identifiability due to the addition of a model discrepancy function. It is observed that the various options of δ can lead to different calibration results, and the proposed three-step approach is implemented to combine the posterior distributions of model parameters and model discrepancy, which leads to accurate prediction while also accounting for the uncertainty in the form of model error.

5.3.2 Calibration of multi-physics models using interval and point data

The target MEMS device of this example (denoted as Dev-1) shown in Fig. 5.4(a) is used as a switch. The membrane deflects under some applied voltage, and will contact the dielectric pad when the applied voltage exceeds a certain threshold. This threshold voltage is called pull-in voltage (V_{pl}), and the device will be closed when the contact occurs. Pull-in voltage is an important metric in the reliability analysis of the device after a certain period of usage. Several models are needed to calculate the pull-in voltage, namely dynamic model, electrostatic model, damping model, and creep model. A 1-D Euler-Bernoulli beam model is used to simulate the dynamic behavior of the MEMS device [Ayyaswamy and Alexeenko, 2010]. The electrostatic model takes applied voltage and air gap (g) as inputs, and calculates electrostatic loading as output. The damping model considers the gas pressure and air gap, and the corresponding damping force is computed [Alexeenko et al., 2011]. The electrostatic loading, damping force, device geometry, material property, boundary condition, and time are the inputs of the dynamic model. The creep model calculates the plastic deformation of the device under long-term loading, and is coupled with the dynamic model. The unknown parameters include Young's modulus (E) and residual stress (σ_{rs}) in the dynamic model, and the creep coefficient A_c in the Coble creep model [Coble, 1963; Hsu et al., 2011]. To predict the pull-in voltage, an iterative method is used by varying the values of applied voltage, and calculating the resulting maximum deflection of the beam. The pull-in voltage is equal to the minimum value of applied voltage that causes the beam to be in contact with the dielectric pad.



(a) Dev-1: Contacting capacitive RF MEMS switch

(b) Dev-2: RF MEMS varactor

Figure 5.4: Example RF MEMS devices (Courtesy: Purdue PRISM center)

5.3.2.1 Different data on two devices

Due to the limitation of experimental resources, currently only the measurement data of pull-in voltage at an early time point is available, and the data are collected on 17 Dev-1 devices with different geometries and initial positions. Because the pull-in voltage data are obtained by keeping increasing the applied voltage by 5 volts until the switch becomes closed, the data are reported in the form of *intervals*.

Study of creep modeling has been separately performed for another type of device (denoted as Dev-2, which has different boundary conditions from Dev-1 as shown in Fig. 5.4(b)), and measurements of device deflection under constant voltage for a relative long time period (∼700 hours) are available. Since these two types of devices are made of the same material, the material-related parameters E and A_c can be considered as the same. A polynomial chaos expansion (PCE) surrogate model is constructed based on 3-D membrane simulation for Dev-2, with E and A_c as inputs and the deflection at three different time points $\boldsymbol{t} = [2lead00, 400, 600]$ hours as output, i.e., $g_{t2} = \mathrm{PCE}(A_c, E) + \delta_2$. δ_2 is the model discrepancy term.

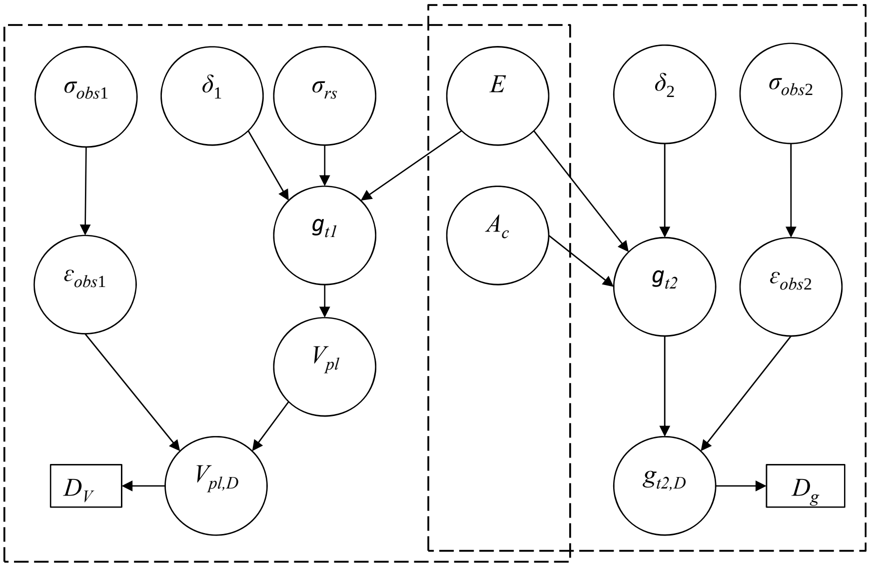


Figure 5.5: Bayesian network

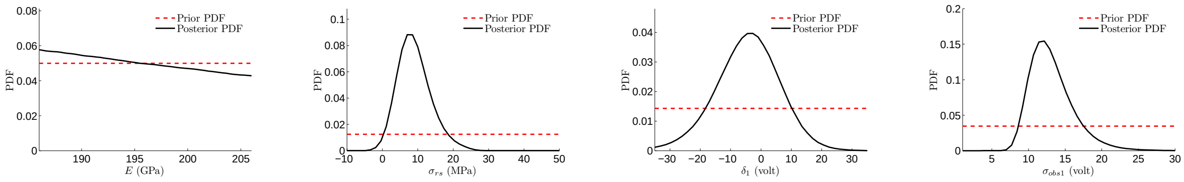
Based on the aforementioned models and data, we construct a Bayesian network as shown in Fig. 5.5. Note that A_c is not directly related to pull-in voltage, since the calculation of pull-in voltage at a given time point only requires dynamic simulation within microseconds, and creep is negligible in such a short time period. Therefore the only common parameter between the two physics models is E. The second and the third options presented in Section 3.4.2 are both implemented for the purpose of comparison in this example, although the first option is not considered due to its higher computation cost.

The identifiability of the calibration parameters in the Bayesian network given the available experimental data is checked using the first-order Taylor series expansion-based method presented in Section 3.3. Since the measurement data of pull-in voltage for 17 Dev-1 devices will be directly used to calibrate the parameters (E, σ_{rs}, δ_1) in the left half of the Bayesian network in Fig. 5.5, we obtain a $17 * 3$ first-order derivative matrix $\boldsymbol{A}$ with rank $r_A = 3$, i.e., E, σ_{rs}, and δ_1 are identifiable with these 17 data points of pull-in voltage. We also examine the identifiability of parameters E, A_c and δ_2 in the right half of the Bayesian network with the deflection data of Dev-2 at the three test time points (200, 400, and 600 hours). In this case, the size of the matrix $\boldsymbol{A}$ is $3 * 3$ and the rank of $\boldsymbol{A}$ is 3, which indicates that E, A_c and δ_2 are all identifiable with the deflection data. Note that this method is not applicable for σ_{obs1} and σ_{obs2}, since the standard deviations of measurement noise are the parameters of statistical models as stated in Section 3.3.

5.3.2.2 Calibration with information flowing from left to right in the Bayesian network

Following the second option presented in Section 3.4.2, the left half of the Bayesian network is considered first, i.e., the parameters E, σ_{rs}, δ_1, and σ_{obs1} are calibrated using the pull-in voltage data. The prior and marginal posterior PDFs of E, σ_{rs}, δ_1, and σ_{obs1} are plotted in Fig. 5.6. The prior PDFs are shown as red dashed lines, whereas the posterior PDFs are shown as black solid lines (the same format applies to Figs. 5.7, 5.8, and 5.9). The corresponding statistics are shown in Table 5.2. Note that all the prior PDFs used in this example are assumed to be uniform, except for the prior PDF of the common parameter E in the second step calibration, which is the posterior PDF obtained in the first step calibration.

Then, the parameters in the right half of the Bayesian network are calibrated using the deflection data of Dev-2, and the posterior PDF of E obtained in the first step is used as



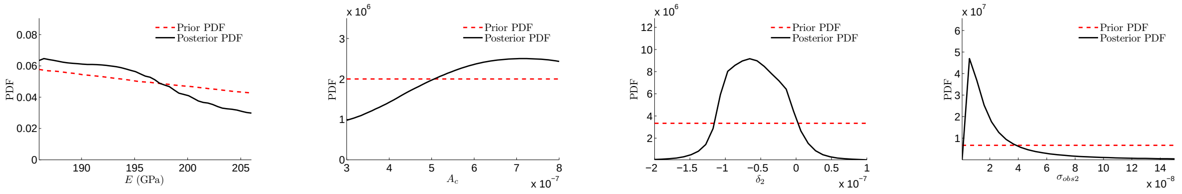
(a) E (b) σ_{rs} (c) δ_1 (d) σ_{obs1}

Figure 5.6: Calibration of parameters using pull-in voltage data

Table 5.2: Prior and posterior statistics of parameters (with data on Dev-1)

	Mean		Standard deviation	
	Prior	Posterior	Prior	Posterior
E (GPa)	196.0	195.5	5.78	5.76
σ_{rs} (MPa)	10.00	9.07	23.10	4.76
δ_1 (Volt)	0	-4.27	20.22	10.35
σ_{obs1} (Volt)	15.50	13.16	8.38	2.98

prior. Fig. 5.7 shows the prior and marginal posterior PDFs of E, A_c, δ_2, and σ_{obs2}, and Table 5.3 contains the corresponding statistics.



(a) E (b) A_c (c) δ_2 (d) σ_{obs2}

Figure 5.7: Calibration of parameters using deflection data

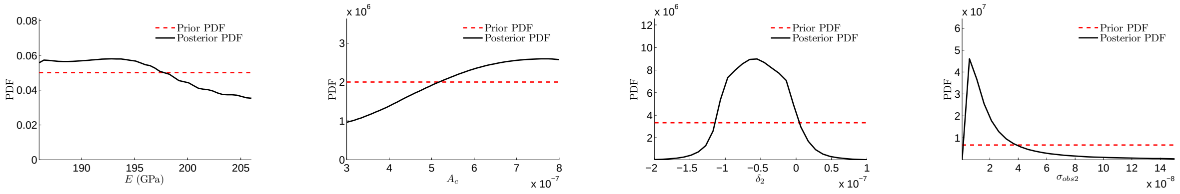
5.3.2.3 Calibration with information flowing from right to left in the Bayesian network

Following the third option presented in Section 3.4.2, the sequence of calibration in the previous section is now reversed. First, the calibration parameter (E, A_c, δ_2, and σ_{obs2}) in

Table 5.3: Prior and posterior statistics of parameters (with data on Dev-2)

	Mean		Standard deviation	
	Prior	Posterior	Prior	Posterior
E (GPa)	195.5	194.7	5.76	5.51
A_c	5.50e-7	5.85e-7	1.44e-7	1.35e-7
δ_2 (μm)	-0.050	-0.057	0.087	0.040
σ_{obs2} (μm)	0.075	0.026	0.043	0.027

the right half of the Bayesian network in Fig. 5.5 are calibrated with the deflection data of Dev-2. The prior and marginal posterior PDFs, and the corresponding statistics are shown in Fig. 5.8 and Table 5.4.



(a) E (b) A_c (c) δ_2 (d) σ_{obs2}

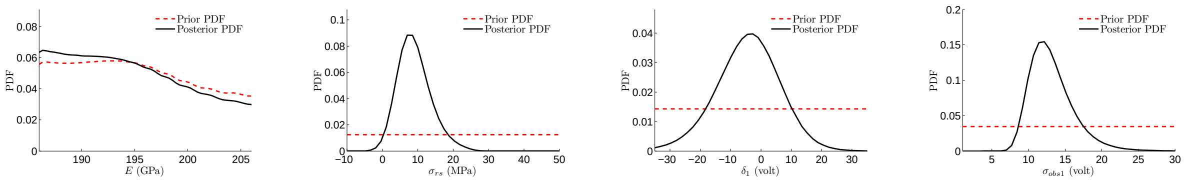
Figure 5.8: Calibration of parameters using deflection data

Table 5.4: Prior and posterior statistics of parameters (with data on Dev-2)

	Mean		Standard deviation	
	Prior	Posterior	Prior	Posterior
E (GPa)	196 .0	195.1	5.78	5.57
A_c	5.50e-7	5.88e-7	1.44e-7	1.35e-7
δ_2 (μm)	-0.050	-0.054	0.087	0.040
σ_{obs2} (μm)	0.075	0.025	0.043	0.027

Similarly to the previous section, the posterior PDF of the common parameter E obtained in the first step of calibration is used as prior, and the parameters in the left half of the

Bayesian network are calibrated using the pull-in voltage data of Dev-1. The calibration results can be found in Fig. 5.9 and Table 5.2.



(a) E (b) σ_{rs} (c) δ_1 (d) σ_{obs1}

Figure 5.9: Calibration of parameters using pull-in voltage data

Table 5.5: Prior and posterior statistics of parameters (with data on Dev-1)

	Mean		Standard deviation	
	Prior	Posterior	Prior	Posterior
E (GPa)	195.1	194.7	5.56	5.51
σ_{rs} (MPa)	10.00	9.08	23.10	4.76
δ_1 (Volt)	0	-4.22	20.22	10.35
σ_{obs1} (Volt)	15.50	13.16	8.38	2.98

.

5.3.2.4 Discussion

In this example, the posterior PDFs of the parameters are computed directly using trapezoidal integration rule as only 4 parameters need to be calibrated at one time. Uniform grids are used for the numerical integration over the parameters, and the number of grid points for each parameter is selected based on the convergence of the posterior density computation. By comparing Figs. 5.7 and 5.8, or Tables. 5.3 and 5.4, we observe that the second and the third options give similar posterior PDFs and statistics of the calibration parameters. The same observation can be drawn from the comparison between Figs. 5.6 and 5.9, or Tables. 5.2 and 5.5. This is due to the fact that the posterior PDFs of the common parameter E obtained in the first step of these two options are not significantly different from the uniform prior PDFs. Hence, the calibration in the second step, which uses the posterior PDF of E obtained from the first step as prior, will give similar results to the case that the uniform prior PDF is used. The relatively small difference between the prior and posterior PDFs of E indicates that the available experimental data are insufficient to reduce significantly the uncertainty about E. In addition, it can be observed from Tables. 5.3 and 5.5 that both the second and the third options give the same posterior statistics of E after the two-step calibration, which is expected since in theory both options should give $\pi(E|D_1, D_2)$ as the calibrated PDF of E (D_1 denotes the pull-in voltage data of Dev-1, and D_2 denotes the deflection data of Dev-2).

5.4 Bayesian network-based model validation

5.4.1 Validation of a damping model

In this section, the aforementioned model validation methods presented in Section 2.3 and Chapter IV are demonstrated via an application example on damping prediction for MEMS switches. The quantity of interest is the damping coefficient, and the corresponding computational model is a polynomial chaos expansion surrogate model. The validation data are obtained from fully characterized experiments, and it is found that the directional bias defined in Section 4.1 exists between model prediction and validation data.

5.4.1.1 Modeling of micro-scale squeeze-film damping

Within the framework of uncertainty quantification in the modeling of RF MEMS switches, the validation of squeeze-film damping model emerges as a crucial issue due to two factors: (1) damping strongly affects the dynamic behavior of the MEMS switch and therefore its lifetime [Snow and Bajaj, 2010]; (2) it is difficult to accurately model micro-scale fluid damping and available models are applicable to limited regimes [Bidkar et al., 2009].

Two major sources of uncertainty have been shown to affect the prediction of gas damping [Guo et al., 2010]. The first one is epistemic uncertainty related to the lack of understanding of fundamental failure modes and related physics models. The second one is aleatory uncertainty in model parameters and inputs due to variability in either the fabrication process or in the operating environment. Uncertainty quantification approaches usually require large numbers of deterministic numerical simulations. In order to reduce the computational cost, a polynomial chaos expansion (PCE) surrogate model is constructed and trained using solutions of the Navier-Stokes (N-S) equation for a few input combinations, thus avoiding repetitively solving the N-S equation.

Based on the calculated damping coefficient values $y(\boldsymbol{x_j})$ at the quadrature nodes $\boldsymbol{x_j}$ by solving the Navier-Stokes Slip Jump model [Gad-el Hak, 2005], the PCE model $y_m(\boldsymbol{x})$ can be constructed using Eqs. 2.38 and 2.39.

Therefore, for a given input combination $\boldsymbol{x_k}$, the prediction of damping coefficient based on the PCE model is a random variable with normal distribution $N(\mu_m(\boldsymbol{x_k}), \sigma_m(\boldsymbol{x_k}))$.

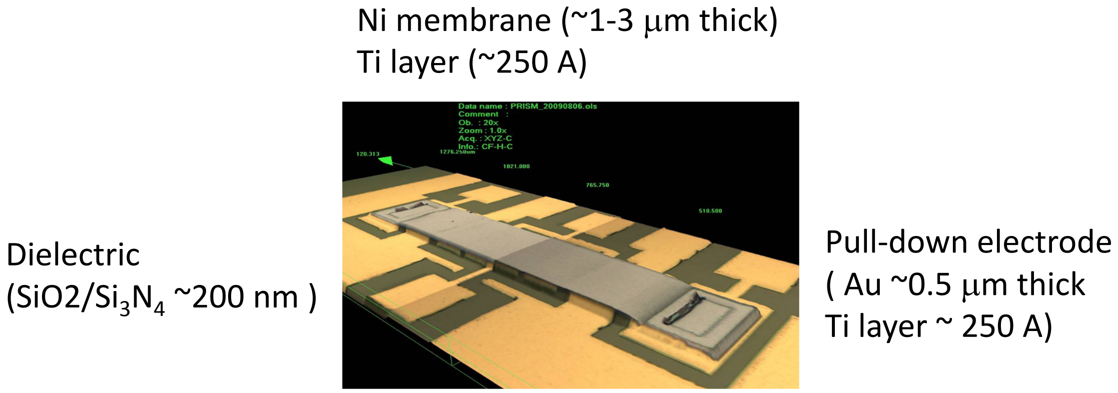


Figure 5.10: Example RF MEMS switch (Courtesy: Purdue PRISM center)

The example RF MEMS switch modeled as a membrane is shown in Fig. 5.10. To construct a PCE model for the damping coefficient, the variables $\boldsymbol{x}$ need to be specified first. A probabilistic sensitivity analysis shows that the membrane thickness t, the gap height g, and the frequency ω are the major sources of variability in the damping coefficient. Hence, these three variables are included in the PCE model and they are all measured in the validation experiment, i.e., $\boldsymbol{x} = [t, g, \omega]$. The coefficients a_i's in Eq. 2.38 are the parameters ($\boldsymbol{\theta}$) of the PCE model. Note that the parameters of this PCE surrogate model are estimated using the simulation data $\{\boldsymbol{x}_j, y(\boldsymbol{x}_j)\}$ from the Navier-Stokes Slip Jump model as shown in Eq. 2.39, instead of using experimental data. Four different gas pressures - 18798.45 Pa, 28664.31 Pa, 43596.41 Pa, and 66661.19 Pa - are considered and correspondingly four PCE models are constructed. This example uses a third order PCE model with Legendre polynomial bases [Guo et al., 2010]. The representation accuracy of the surrogate model can be quantified by the standard deviation (σ_m) of the surrogate model error term (ε_m) in Eq. 2.40. In this example, the magnitude of σ_m is limited to less than 5% of the model prediction over the training (sampling) domain, which we consider acceptable. However, it should be noted that the validity of the surrogate model does not guarantee the validity of the original model. We only have access to the surrogate model and validation experimental data; therefore in this example we are only assessing the validity of the surrogate model.

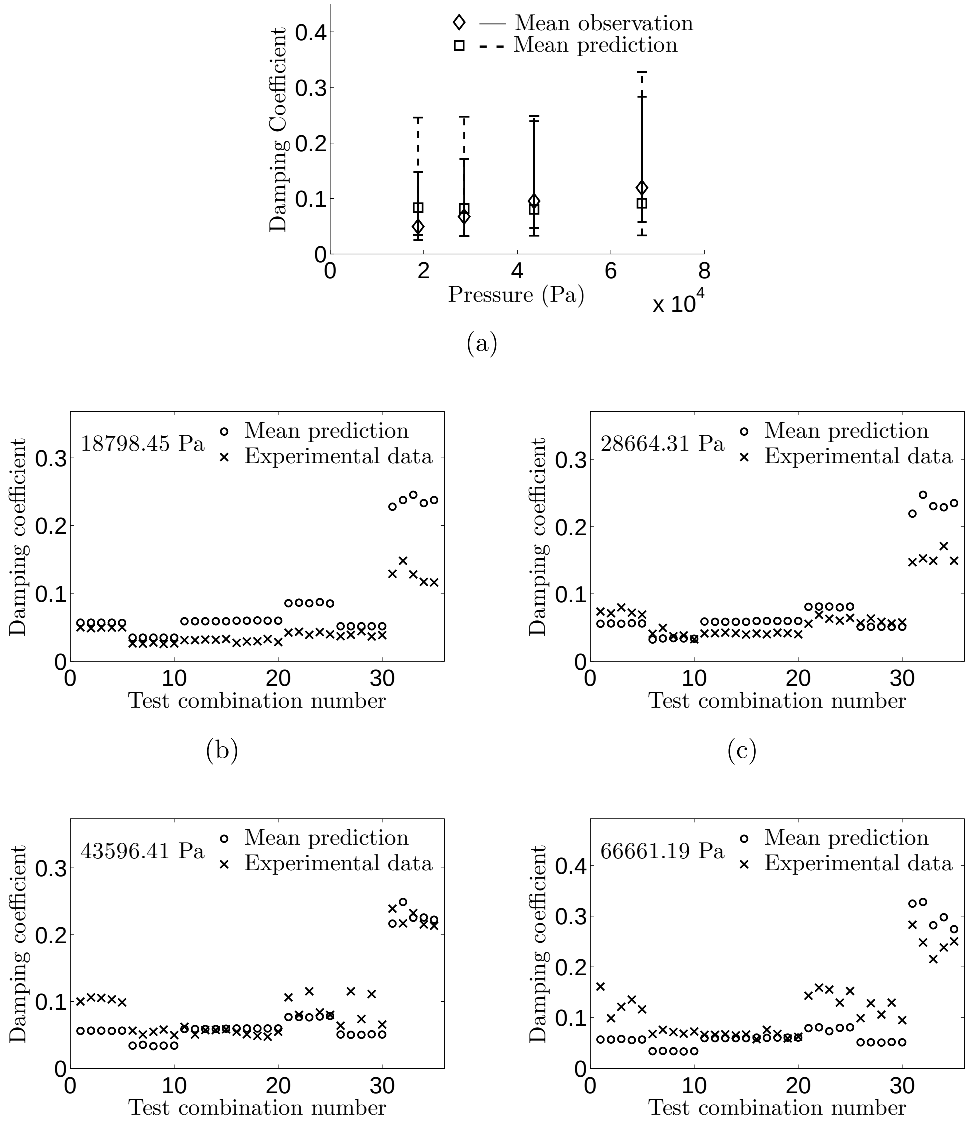
If the original model is to be validated, the number of model evaluations needed to compute a validation metric may be of interest in practice as the original model could be computationally demanding in some problems. In general, z-test, t-test, and Bayesian interval hypothesis testing require less number of model evaluations, since only the mean and variance of the model output are used to compute the validation metric. More model evaluations are needed in Bayesian equality hypothesis testing, the reliability-based method, and the area metric-based method, since the entire probability distribution of model output is needed. In this example, the output of each PCE model follows a normal distribution as shown in Eq. 2.38, which is fully described by the mean and variance. Therefore, the number of surrogate model evaluations needed in each validation approach is the same.

5.4.1.2 Experimental data for validation

In the experiment, seven devices with different geometric dimensions are considered. For each of the four pressures mentioned above, 5 tests are conducted on each of the seven devices with slightly different frequencies, and hence 140 data points are collected. Since the input set $[t, g, \omega]$ are recorded for each of the data points, these experiments are fully characterized and the 140 data points correspond to 140 different test input combinations. That is to say, there are 140 sample sets and each set contains only one sample. We assume that the variability of samples in each sample set is due to measurement error, and measurement errors for different test combinations are treated as statistically independent. Therefore, the sample sets are also statistically independent from one another.

Fig. 5.11(a) shows a graphical comparison between the mean PCE model prediction and experimental data under the four different pressures by aggregating predictions and data with respect to the 35 test combinations for each pressure value. The top/bottom points are correspondingly the maximum/minimum value of model mean predictions and experimental data, and the square/diamond markers are the average values of predictions/data on the 35 test combinations. Note that Fig. 5.11(a) ignores the difference between the seven devices, and thus should not be considered as a rigorous comparison. A more detailed graphical comparison showing mean prediction of the PCE model vs. experimental data on each of the individual test combinations is provided in Figs. 5.11(b)-(e).

From the graphical comparison, we can see that the PCE model performs better under the middle two values of pressure. Also note that there is a systematic bias between the PCE model and experimental observations at the low pressure value (18798.45 Pa), i.e., the mean predictions of the PCE model are always larger than the experimental data.



(a)

(b)

(c)

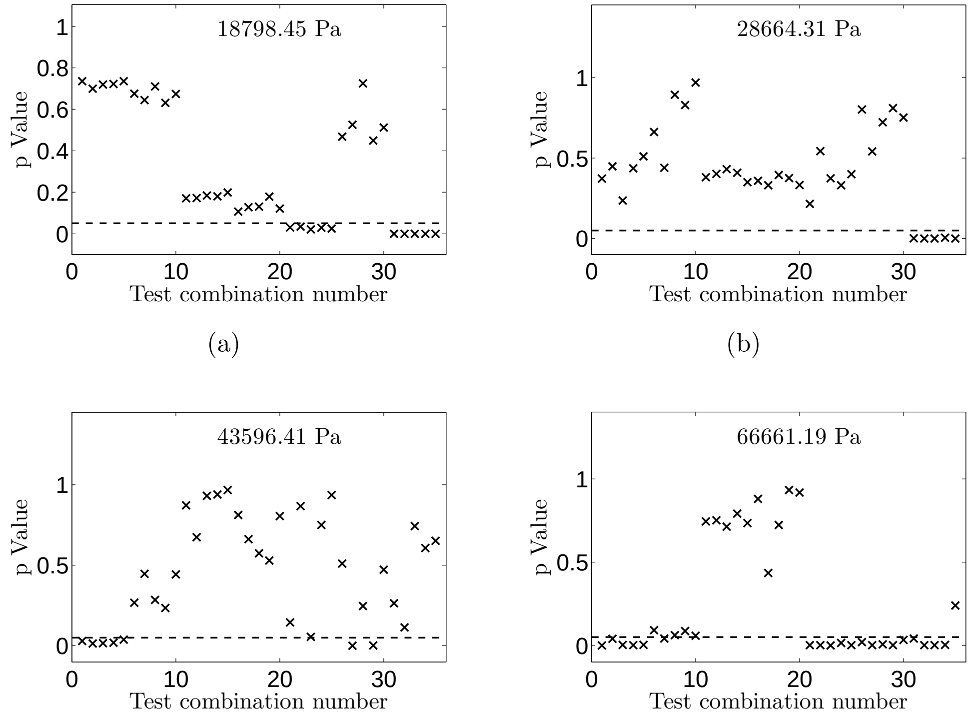
(d)

(e)

Figure 5.11: Graphical comparisons between PCE predictions and experimental data

5.4.1.3 Classical hypothesis testing

Because the sample size for each experimental combination is only 1, the t-test is not applicable and instead z-test is used. The p-values calculated using Eq. 2.17 are shown in Fig. 5.12. The dashed lines in Fig. 5.12 represent the significance level $\alpha = 0.05$. The model is considered to have failed at the experimental combinations where the corresponding p-values fall below the dashed line. Note that a more rigorous test will need to include the probability of making type II error (β). The individual numbers of failures of the four PCE models are shown in Table 5.6.



(a) (b)

(c) (d)

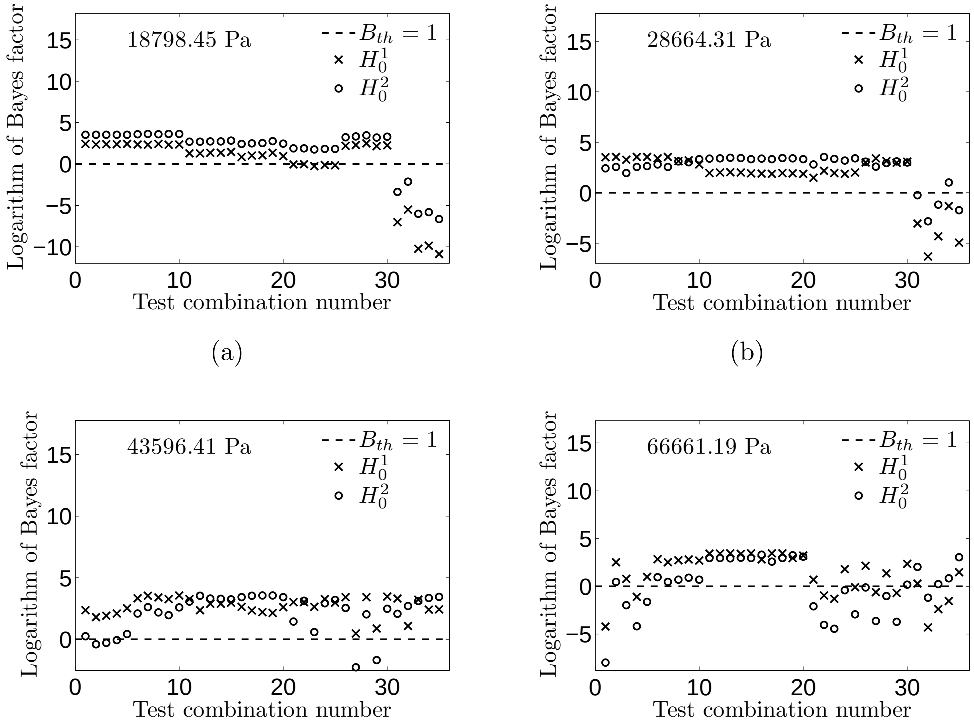
Figure 5.12: p-value of z-test

Table 5.6: Performance of PCE models in z-test with $\alpha = 0.05$

Pressure (Pa)	18798.45	28664.31	43596.41	66661.19
Number of failures	10	5	7	20
Failure percentage	28.6%	14.3%	20.0%	57.1%

5.4.1.4 Bayesian hypothesis testing

Interval hypothesis on distribution parameters As discussed in Section 4.7, combination of two Bayesian hypothesis tests based on the interval null hypotheses H_0^1 and H_0^2 respectively can reflect the existence of directional bias. In practice, the parameters ϵ_μ, $\epsilon_{\sigma 1}$, and $\epsilon_{\sigma 2}$ that define the intervals can be determined based on the strictness requirement of validation. For the purpose of illustration, we set $\epsilon_\mu = 0.025$, $\epsilon_{\sigma 1} = -0.005$, and $\epsilon_{\sigma 2} = 0.005$. μ_l and μ_u that define the possible range of μ are set as 0 and 1 respectively since the MEMS device considered is under-damped. σ_l and σ_u are set to be 0.001 and 0.05 respectively. The results of Bayesian interval hypothesis testings are calculated using Eq. 4.1 - 4.5, and are shown in Fig. 5.13 and Table 5.7.



(a)

(b)

(c)

(d)

Figure 5.13: Bayes factor in interval-based hypothesis testing (on logarithmic scale)

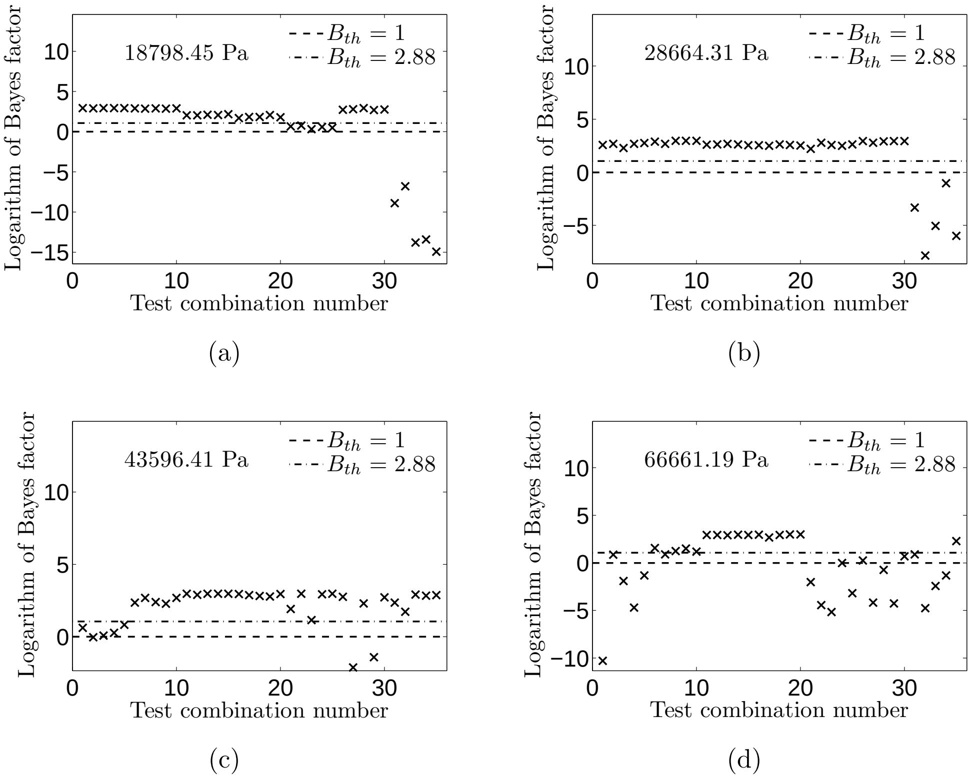
Table 5.7: Performance of PCE models in interval-based Bayesian hypothesis testing with $\log B_{th} = 0$

Pressure (Pa)			18798.45	28664.31	43596.41	66661.19
H_0^1 :	$-\epsilon_\mu \leq \mu_m - \mu \leq 0$	# of failures	10	5	0	10
	$\epsilon_{\sigma 1} \leq \lvert\sigma_m - \sigma\rvert \leq \epsilon_{\sigma 2}$	Overall B	3.1	58.3	92.9	44.1
H_0^2 :	$0 \leq \mu_m - \mu \leq \epsilon_\mu$	# of failures	5	4	5	14
	$\epsilon_{\sigma 1} \leq \lvert\sigma_m - \sigma\rvert \leq \epsilon_{\sigma 2}$	Overall B	63.9	87.1	74.1	1.4
Combined test		# of failure	10	5	5	16
		Failure %	28.6%	14.3%	14.3%	45.7%

Equality hypothesis on probability density functions The possible values of damping coefficient range from 0 to 1 since the system is under-damped. Hence the limit of integration in the denominator of Eq. 4.7 is [0, 1], while the limit of integration in the numerator is $[-\infty, +\infty]$.

The performance of the PCE models in Bayesian hypothesis testing are shown in Fig. 5.14 and Table 5.8. The values of Bayes factor are calculated using Eq. 4.7, and the threshold Bayes factor $B_{th} = 1$ (this threshold value is chosen based on the discussion in Section 4.6). Although the performance of the PCE model in terms of failure percentage is different for the two hypothesis tests as shown in Table 5.6 and Table 5.8, if one increases the threshold Bayes factor B_{th} to 2.88, which is calculated using Eq. 4.16 with $p = 0.05$ in Section 4.5.1, the result of Bayesian hypothesis testing in terms of the number of failures becomes the same as in the z-test in Section 5.4.1.3. The reason for this coincidence has been explained in Section 4.5.1. Note that the performance of the second PCE model (for pressure = 28664.31 Pa) remains the same when B_{th} is raised from 1 to 2.88, and this can be easily observed from Fig. 5.14(b).

By comparing the results based on interval hypothesis on distribution parameters and equality hypothesis on probability density functions (Tables 5.7 and 5.8), it can be observed that the performance of the PCE model for pressure 18798.45 Pa in the first test is significantly



(a) (b)

(c) (d)

Figure 5.14: Bayes factor in equality-based hypothesis testing (on logarithmic scale)

worse than in the second test, while the models for the other three pressures have similar failure percentages in these two tests. As shown in Fig. 5.11(b), the data are all located below the mean predictions of this PCE model, which indicates the existence of directional bias, and thus the PCE model for pressure 18798.45 Pa performs worse in the Bayesian interval hypothesis testing.

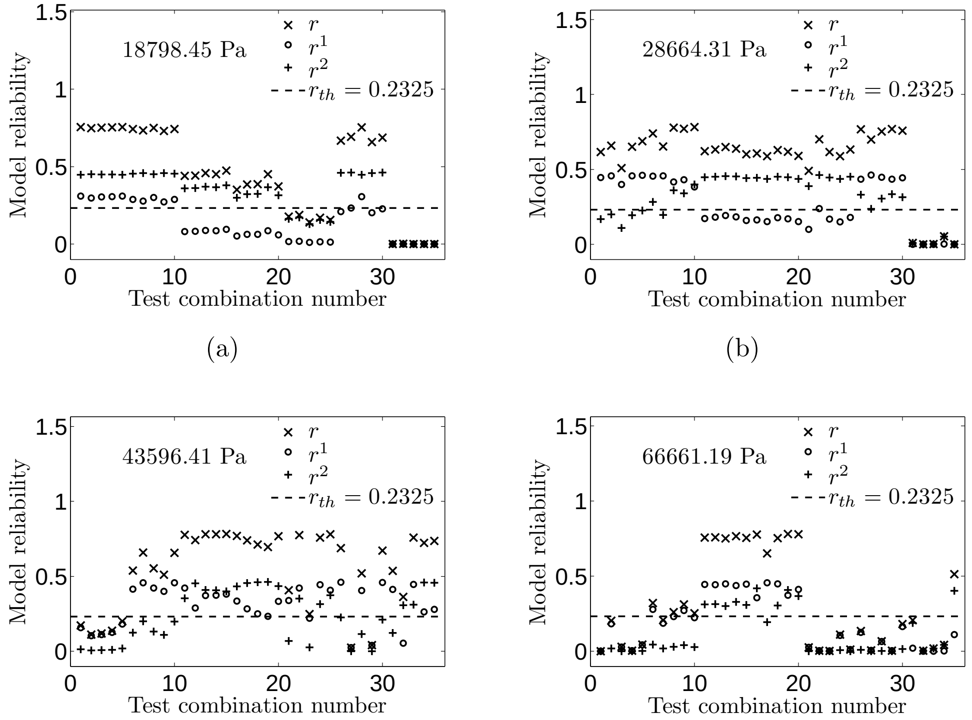
5.4.1.5 Reliability-based metric

Fig. 5.15 and Table 5.9 show the calculated values of the reliability-based metric r, r^1 and r^2 (Eq. 2.24 and 4.24), the failure percentage of models with $\epsilon = 0.025$ and the decision criterion $r_{th} = 0.2325$. This decision criterion is obtained using Eq. 4.20 with the significance level $\alpha = 0.05$, and thus the results of validation (comparing r with r_{th}) in terms of failure

Table 5.8: Performance of PCE models in equality-based hypothesis testing with $\log B_{th} = 0$

Pressure (Pa)	18798.45	28664.31	43596.41	66661.19
Number of failures	5	5	3	15
Failure percentage	14.3%	14.3%	8.6%	42.9%
Overall Bayes factor (log-scale)	7.4	57.2	72.3	-10.2

percentage are the same as in the z-test in Section 5.4.1.3. It can also observed that the failure percentage of the PCE model for pressure 18798.45 Pa increases significantly in the test that comparing r^1 and r^2 with $r_{th}/2$ due to the existence of directional bias.



(a) (b)

(c) (d)

Figure 5.15: Reliability-based metric

Table 5.9: Performance of PCE models in reliability-based method with $r_{th} = 0.69$

	Pressure (Pa)	18798.45	28664.31	43596.41	66661.19
r vs. r_{th}	Number of failures	10	5	7	20
	Failure percentage	28.6%	14.3%	20.0%	57.1%
r^1 and r^2 vs. $r_{th}/2$	Number of failures	20	7	12	25
	Failure percentage	57.1%	20.0%	34.3%	71.4%

5.4.1.6 Area metric-based method

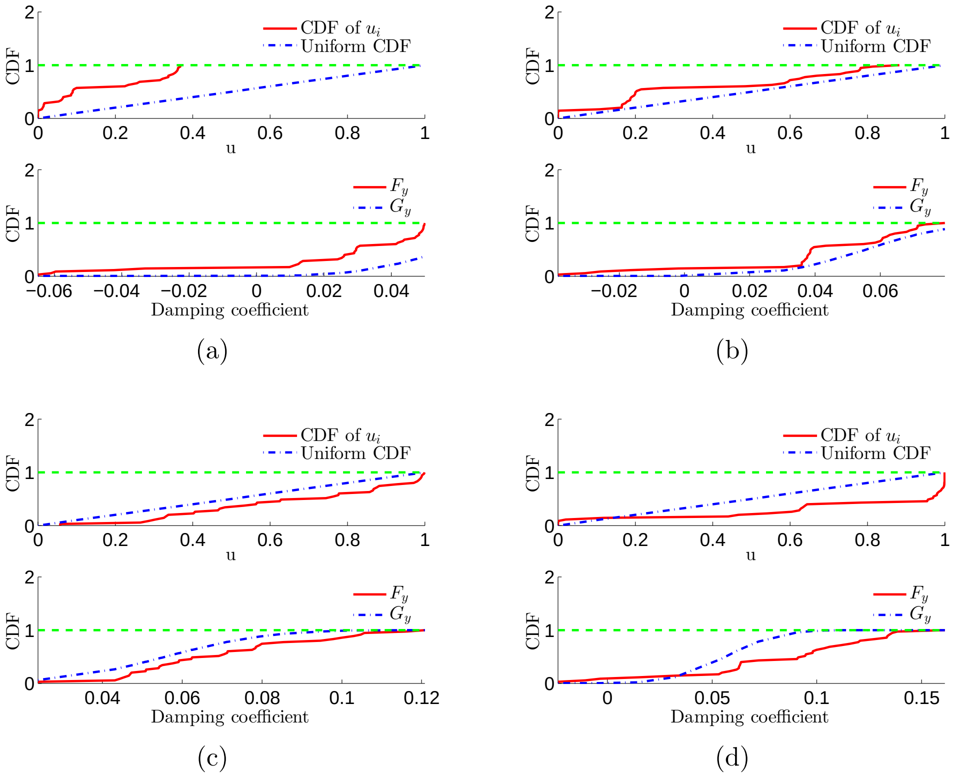
The area metrics for the four PCE models in both u-space and physical space are computed using Eqs. 2.25 and 4.23, and the results are shown in Fig. 5.16 and Table 5.10. Note that the PCE model for pressure 18798.45 Pa has the highest area value in u-space. This is due to the directional bias between mean predictions and experimental data, and it is reflected in the area metric as discussed in Section 4.7. Since the area metric in physical space ($d(F_y, G_y)$) can be interpreted as prediction error, the decision of rejecting/accepting the models can be made by comparing the values of $d(F_y, G_y)$ against a specified tolerance limit. If we use the same tolerance limit $\epsilon = 0.025$ as in the reliability-based method, the PCE model for pressure 66661.19 Pa will be rejected as the corresponding area metric ($= 0.033$) is larger than 0.025, whereas the other three PCE models will be accepted.

Table 5.10: Area metric for PCE models

Pressure (Pa)	18798.45	28664.31	43596.41	66661.19
$d(F_u, S_u)$	0.343	0.139	0.151	0.249
$d(F_y, G_y)$	0.024	0.014	0.013	0.033

5.4.1.7 Discussion

This section validated the PCE surrogate models for the MEMS switch damping coefficient using the validation methods presented in Section 2.3 and Chapter IV, and 140 fully



(a) (b)

(c) (d)

Figure 5.16: Comparison of CDFs in the u-space and the physical space

characterized experimental data points. Based on the performance of the PCE model in these validation tests, it can be concluded that the prediction of the PCE model has better agreement with observation under the middle two values of pressure (28664.31 Pa and 43596.41 Pa), whereas less agreement can be found under the lowest and highest pressure values (18798.45 Pa and 66661.19 Pa). The decision on model acceptance can be formed based on the failure percentages with the hypothesis testing methods and the reliability-based method, and the values of area metric, given the desired prediction error tolerance. It is shown that the z-test and the reliability-based metric give the same results in terms of failure percentage when r_{th} is selected corresponding to the significance level α used in z-test. Similarly, classical and Bayesian hypothesis testing give the same result by choosing a specific threshold Bayes factor as shown in Section 4.5.1. It is also shown that the existence of directional bias can be reflected in the Bayesian interval hypothesis testing, reliability-based method with modified intervals, and the area metric-based method. Models that have directional bias will perform worse in these three validation methods than in classical hypothesis testing and in Bayesian hypothesis testing with equality hypothesis on probability density functions.

5.4.2 Validation of calibrated dielectric charging model

The physical parameters calibrated using the compact model in Section 5.3.1 can be used as inputs to a full simulation model, which predicts the number of trapped charges in the dielectric pad after a period of sustained contact with the top electrode, and the trapped charges affect the pull-out voltage. PCE surrogate models are constructed with these parameters as inputs for different contact durations, and with change in pull-out voltage as output. In order to validate the calibrated parameter distributions, changes in the pull-out voltage of target devices due to 1, 10, 100, and 1000 millisecond of sustained contact with holddown voltage = 120V are measured in experiments. 13 repetitive tests are performed on 5 identical devices. The normalized histogram of experimental data and the predicted probability density function of model prediction are shown in Fig. 5.17.

By treating the experimental data as uncharacterized, we can obtain the overall Bayes factors for the four PCE surrogate models via the ensemble validation approach discussed in Section 4.3.3, as showed in the individual titles of Fig. 5.17. The Bayes factors suggest that the predictive capability of the model is acceptable (> 1) when the contact time is shorter than 10ms, but the quality of model prediction gradually decreases with time.

5.4.3 Validation of calibrated creep model

The target MEMS device has also tested in order to validate the creep model calibrated in Section 5.3.2. In the validation experiment, the creep deformations of four devices under different constant amplitude voltages are recored for ~ 50 hours. The collected deflection

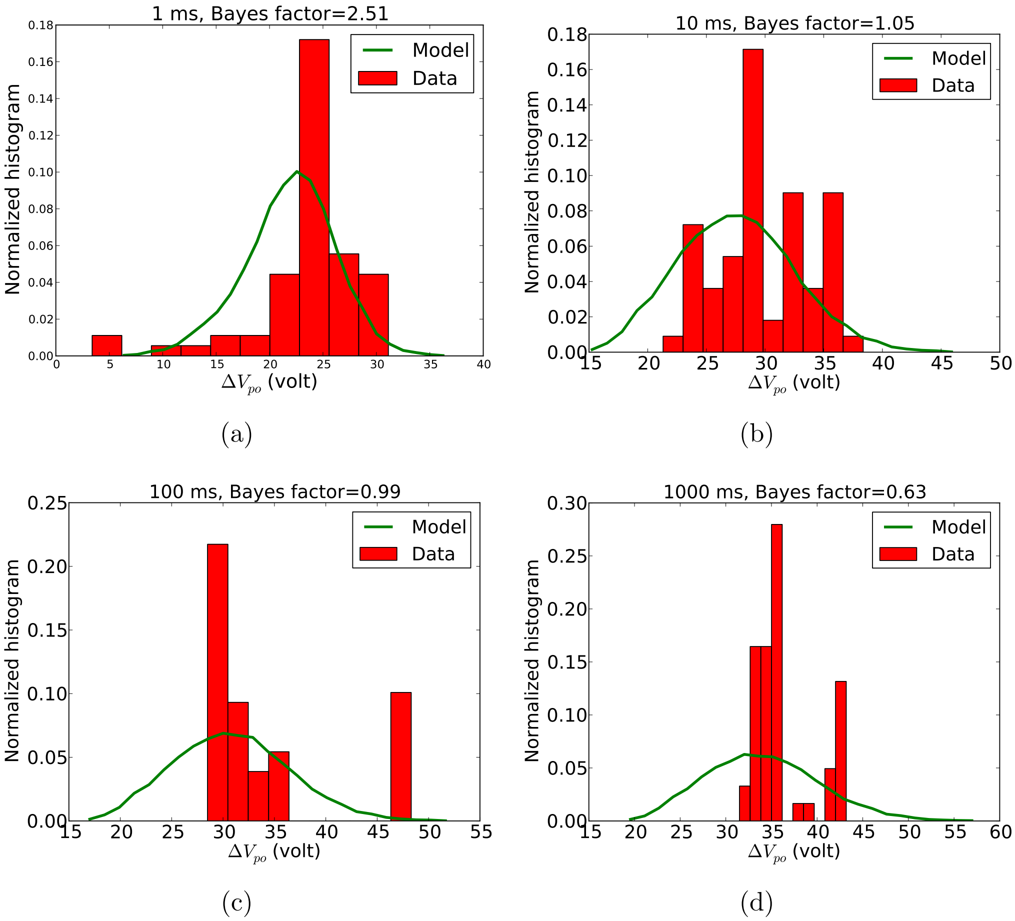
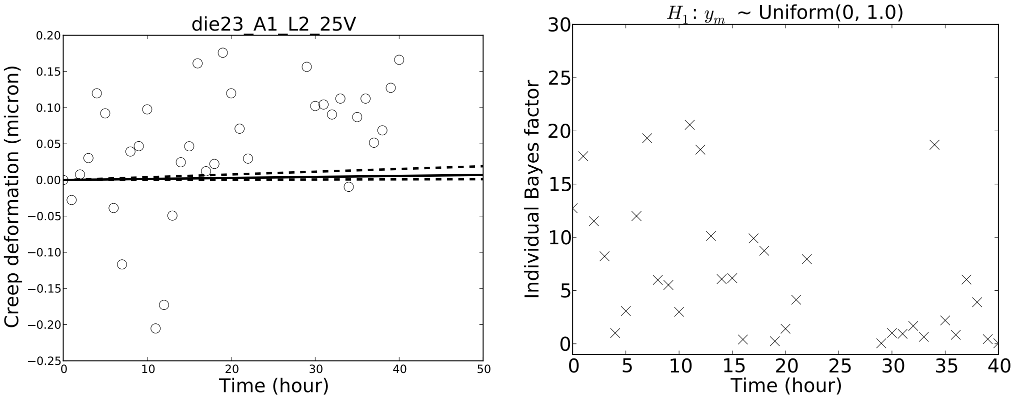


Figure 5.17: Validation of the calibrated dielectric charging parameters using pull-out voltage data

data and the corresponding model predictions are shown in Figs. 5.18(a) - 5.21(a). Bayesian equality hypothesis testing is applied and Bayes factors are computed for each time point, and the values are shown in Figs. 5.18(b) - 5.21(b).

It can be observed from the graphical comparisons in Figs. 5.18(a) - 5.21(a) that the model predictions do not match well with the validation data. This is in part due to the fact that the experiments only last for less than 50 hours and the resulting creep deformations are relatively small compared against measurement noise. The alternative hypothesis in the Bayesian equality hypothesis testing is chosen to be a uniform distribution ranging from [0,



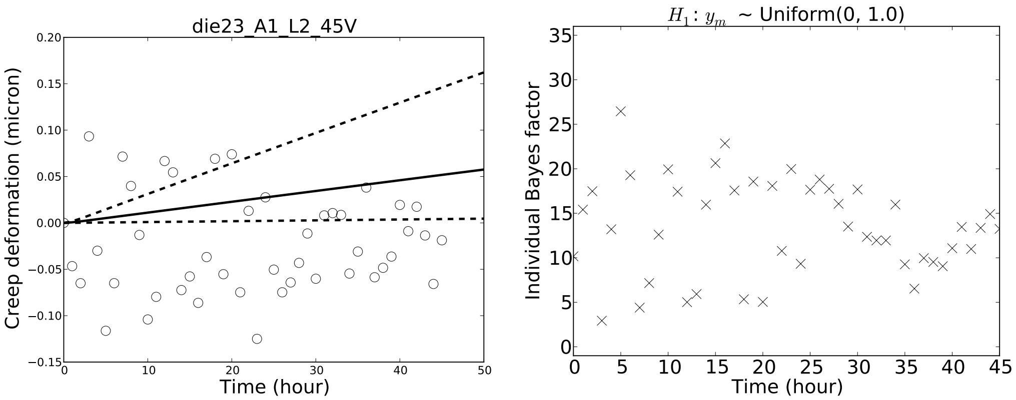
(a) (b)

Figure 5.18: Validation of the calibrated creep model using 25V data

1]. This is a weak alternative hypothesis since the range is relatively large compared against the range of experimental data, which is the cause of high Bayes factor values at some time points even when the graphical comparison is not satisfying. A more rigorous validation of the creep model will require data collected from a longer time scale and more careful selection of the alternative hypothesis.

5.4.4 Validation of device level simulation

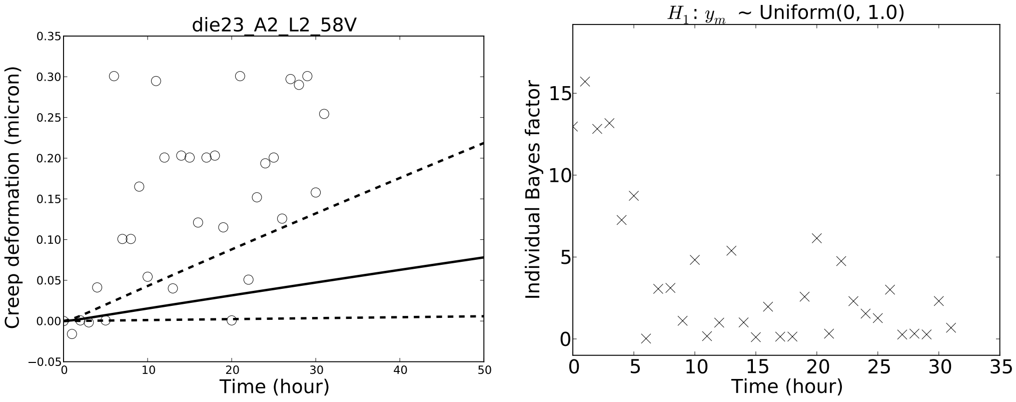
By combining multiple physics models mentioned in Section 5.2, device level simulation predicts pull-in and pull-out voltage. In the corresponding validation experiments, the pull-in and pull-out voltages of seven target device are measured, and the test is repeated for 20 times. The histograms of the data are shown in Fig. 5.22. Because the 3-D membrane model in the device level simulation is computationally demanding, two surrogate modeling techniques, polynomial chaos expansion (PCE) and Gaussian process regression (GPR) are used to approximate the original simulation with efficient surrogates. The 1-D Euler-Bernoulli beam-based device level simulation is also approximated using a GPR surrogate model for the purpose of comparison. Five variables, membrane thickness h, gap between one end of



(a) (b)

Figure 5.19: Validation of the calibrated creep model using 45V data

the membrane and the substrate g_1, gap between the other end of the membrane and the substrate g_2, Young's modulus E, contact height d_c, are identified as inputs to the model and experiments. Due to the precision of the measurement techniques, the geometry parameters h, g_1, and g_2 are reported as intervals for each of the five devices. Direct measurements of E and d_c are not available, but the ranges of these two variables can be obtained via multi-scale simulations [Koslowski and Strachan, 2011; Kim et al., 2012]. As described above, each device corresponds to a combination of the input set [h, g_1, g_2, E, d_c], which is in the form of intervals. Thus, the validation data are partially characterized with 20 replicas. Bayesian equality hypothesis testing is performed for each input combinations. Fig. 5.22(a) shows a graphical comparison between the PDF of pull-in voltages generated using the PCE surrogate model, and the PDF of pull-in voltages generated using the GPR surrogate model (GPR_MEMOSA denotes the surrogate for the 3-D membrane model-based simulation, and GPR_CG denotes the surrogate for the 1-D beam model-based simulation), and the histogram of experimental data. Similarly, Fig. 5.22(b) shows a graphical comparison between pull-out voltage predictions and experimental data. The corresponding Bayes factors are showed in the titles of the sub-figures of Fig. 5.22.

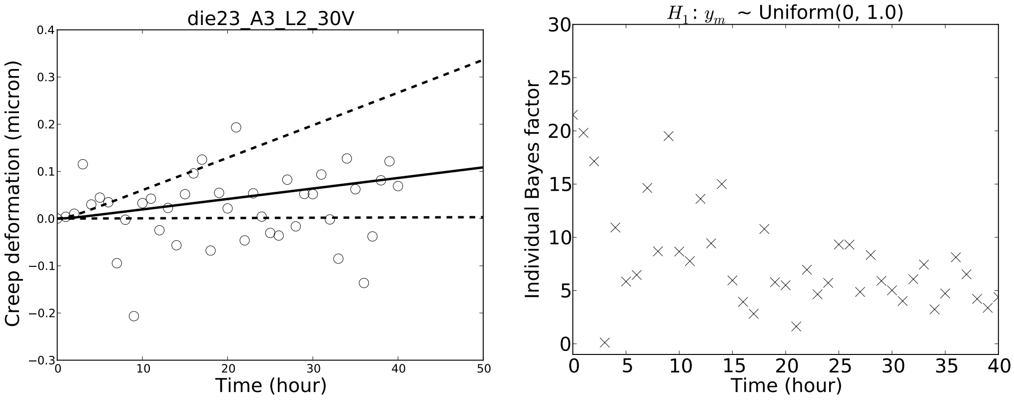


(a) (b)

Figure 5.20: Validation of the calibrated creep model using 58V data

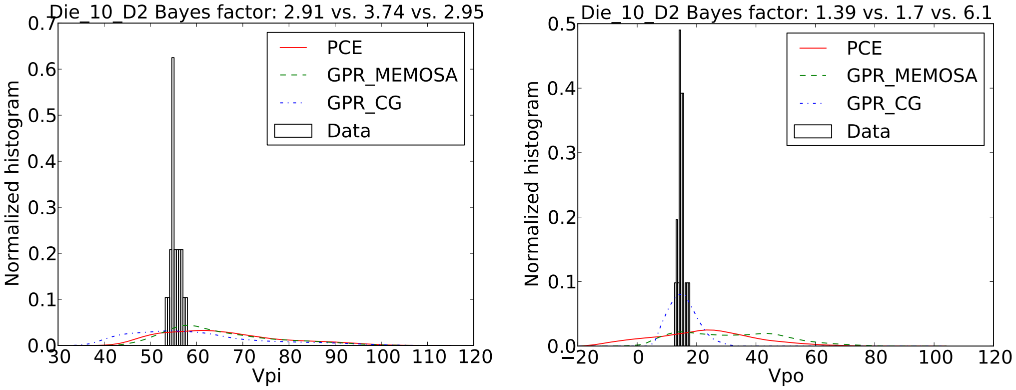
We also apply the ensemble validation approach to obtain the overall Bayes factor. Each of the five input variables has seven individual intervals corresponding to the seven devices, and these intervals are combined into a single one by simply taking the minimum value of the lower bounds and the maximum value of the upper bounds. Unconditional model predictions are obtained by treating the input variables as uniformly distributed with the combined ranges, and propagating the uncertainty from the input to the output as showed in Eq. 4.10. Fig. 5.23 shows the graphical comparisons between the unconditional model predictions and the histogram of all experimental data, and the corresponding Bayes factors are showed in the titles of the sub-figures.

The validation results showed in Fig. 5.22 and Fig. 5.23 suggest that the overall performance of the device level simulation is not satisfying although the prediction matches well with the data at some input combinations. These results could be partly due to surrogate model errors, since the current surrogate models are constructed based on a limited number of training points. Adaptive training algorithms [Hombal and Mahadevan, 2011] for Gaussian process surrogate model are currently under investigation in order to obtain a more accurate approximation of the original simulation.



(a) (b)

Figure 5.21: Validation of the calibrated creep model using 30V data

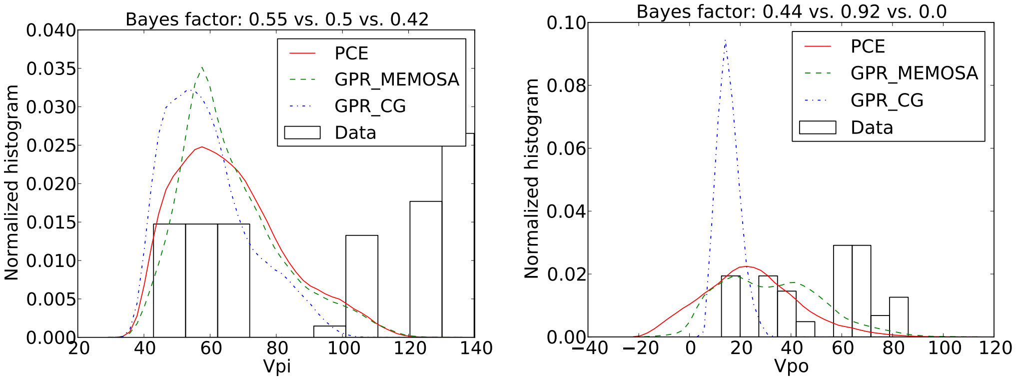


(a) (b)

Figure 5.22: (a) Validation of pull-in voltage prediction. (b) Validation of pull-out voltage prediction

5.5 Reliability of the target device

In this section, we study the reliability of the target device considering failure due to dielectric charging, which causes the performance variables, the pull-in and pull-out voltages, to decrease with time. For the purpose of illustration, the device "Die_10_D2" is assumed to be the target device, and the surrogate model "GPR_CG" of device level simulation validated in Section 5.4.4 is used to predict the initial pull-in and pull-out voltage for given values



(a) (b)

Figure 5.23: (a) Ensemble validation of pull-in voltage prediction. (b) Ensemble validation of pull-out voltage prediction

of h, g_1, g_2, E, and d_c. "GPR_CG" is selected because the corresponding Bayes factor for pull-out voltage prediction of the device "Die_10_D2" is significantly higher than the other two surrogate models. The Bayes factors obtained in the validation exercise can be used to derive the posterior probabilities of model being correct $\Pr(H_0|D)$ and model being incorrect $\Pr(H_1|D)$. In order to include the uncertainty in the model prediction, the Bayesian model averaging approach described in Section 4.6 can be used to obtained a combined probability distribution of V_{pi} or V_{po} as

$$\begin{aligned} \bar{\pi}(V_{pi}) &= \pi_0(V_{pi})\Pr(H_0|D) + \pi_1(V_{pi})\Pr(H_1|D) \\ \bar{\pi}(V_{po}) &= \pi_0(V_{po})\Pr(H_0|D) + \pi_1(V_{po})\Pr(H_1|D) \end{aligned} \tag{5.2}$$

where $\pi_0(V_{pi})$ and $\pi_0(V_{po})$ are the probability density functions of V_{pi} and V_{po} predicted using the model; $\pi_1(V_{pi})$ and $\pi_1(V_{po})$ are the probability density functions of V_{pi} and V_{po} considered in the alternative hypothesis H_1. The combined PDFs $\bar{\pi}(V_{pi})$ and $\bar{\pi}(V_{po})$ are showed in Fig. 5.24,

Further, surrogate models are constructed for changes in pull-in and pull-out voltages due

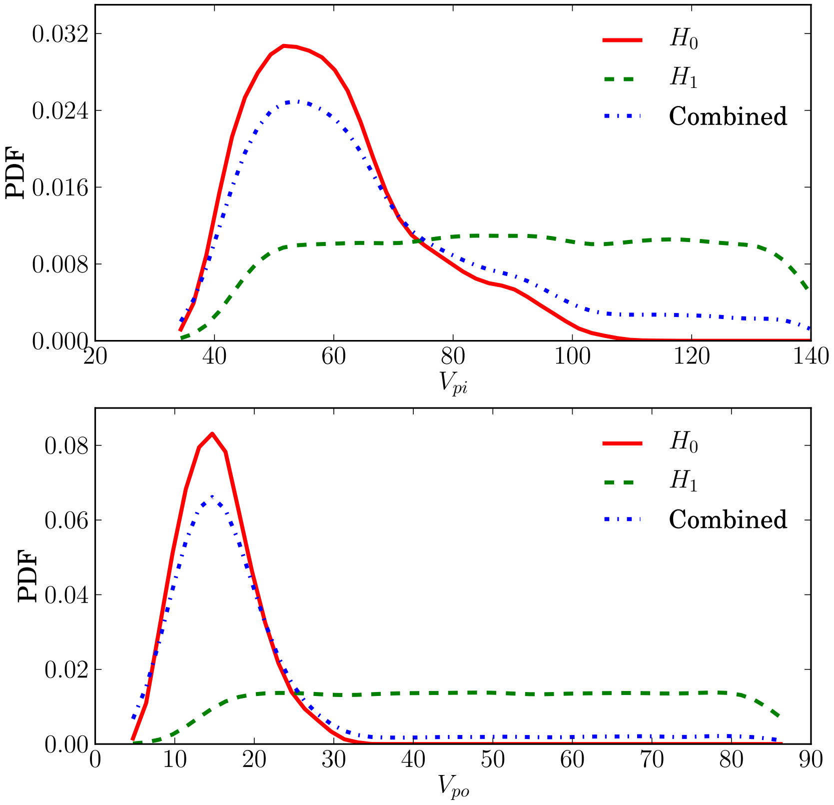


Figure 5.24: Inclusion of model uncertainty in pull-in and pull-out voltage prediction

to dielectric charging, with the dielectric charging model parameters and contact duration as input arguments. By incorporating the probability distribution of the device level simulation input $[h,\ g_1,\ g_2,\ E,\ d_c]$ and the calibrated dielectric charging model parameters, we can obtain the failure probability of the target device as a function of time. For the purpose of illustration, the hold-down voltage is set to be 120V, and the failure criteria are set to be $V_{pi} < 50.0V$ or $V_{po} < 0.1V$. The prediction of failure probability without the inclusion of model uncertainty in the prediction of initial V_{pi} or V_{po} is showed in Fig. 5.25, whereas the prediction of failure probability that includes model uncertainty in predicting initial V_{pi} or V_{po} is showed in Fig. 5.26.

It can be observed from Fig. 5.24 that the combined probability density functions (PDF)

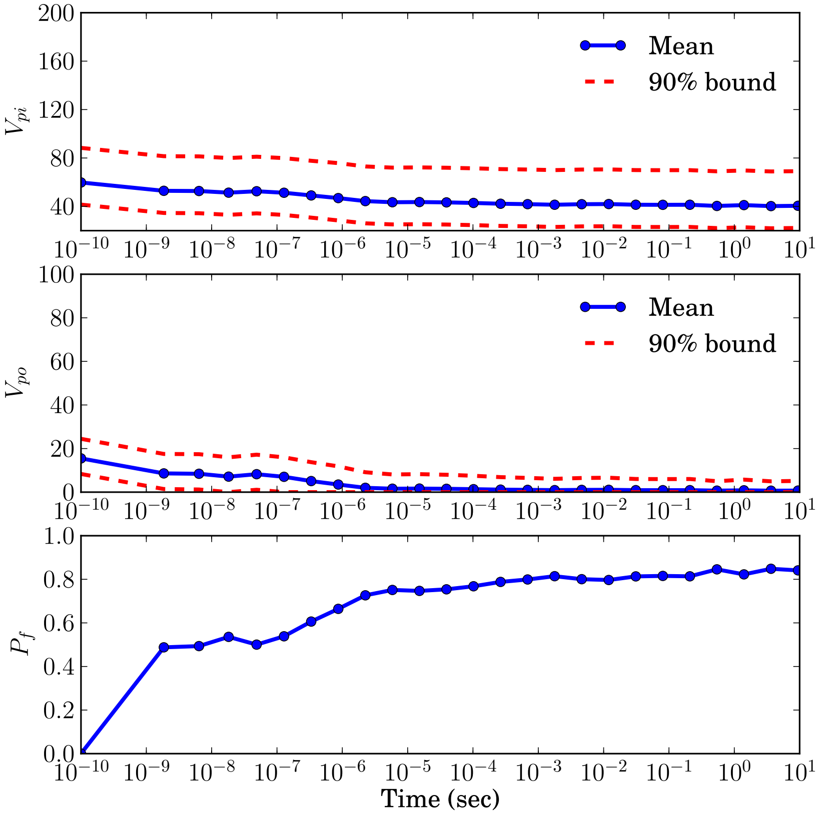


Figure 5.25: Probability of failure of the target device without the inclusion of model uncertainty

of the initial pull-in and pull-out voltage are significantly wider than the surrogate model prediction. Correspondingly, more uncertainty can be observed in the predictions of V_{pi} and V_{po} at future time instants, as showed in Fig. 5.26.

5.6 Conclusion

In this chapter, we constructed an overall Bayesian network approach to integrate various aspects of uncertainty quantification towards the reliability prediction of a MEMS device.

Bayesian model calibration was extended to multi-physics models to quantify the uncertainty in model parameters and model form. Various options of model discrepancy were

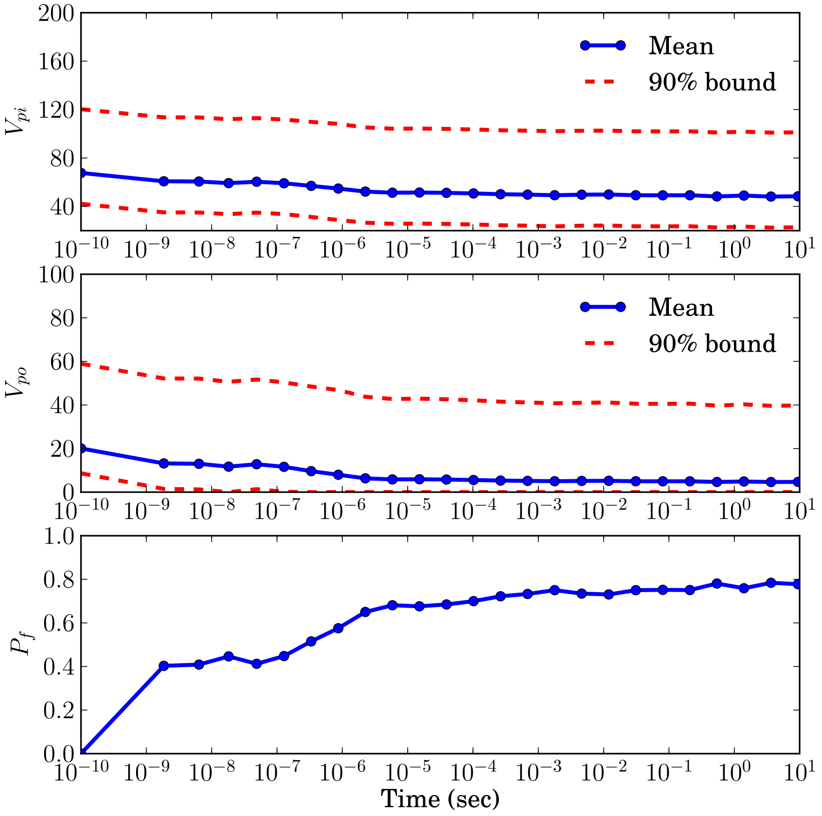


Figure 5.26: Probability of failure of the target device with the inclusion of model uncertainty

investigated in the calibration of the dielectric charging model, and different posterior probability distributions of model parameters and discrepancy were obtained. The three-step approach proposed in Section 3.2.2 was applied to assess the calibrated model in order to help the analyst select or improve the model discrepancy formulation. The sequential strategies outlined in Section 3.4 were implemented to calibrate residual stress and creep coefficient using two physics models and the corresponding experimental data.

Predictive capability of the simulations at the model level and the device level was assessed using quantitative model validation methods with fully and partially characterized experimental data. Ensemble validation was also performed to obtain an overall measure of model validity.

Based on the calibrated and validated physics-based models and surrogate models, prediction for the time-varying reliability of the target device was obtained, which incorporates the uncertainty from different levels of the system.

CHAPTER VI

INCLUSION OF TIME-DEPENDENT INPUT MONITORING DATA

6.1 Introduction

Quantifying the uncertainty due to the use of approximate models in time-dependent reliability analysis was the focus of Chapter V. In addition, time-varying load applied to engineering systems as an input may also contribute to the uncertainty in reliability prediction. Tremendous research efforts have been devoted to the modeling and simulation of random load using the theory of stochastic process, especially in the area of earthquake and foundation engineering [Shinozuka and Deodatis, 1991; Deodatis, 1996; Phoon et al., 2002]. Based on the established load models, this chapter will further focus on quantifying the uncertainty in time-dependent load via the inclusion of real-time monitoring data, with application developed for fatigue loading.

Background of application The fatigue process of mechanical components under service loading is stochastic in nature. Therefore, methodologies for uncertainty quantification and propagation have been developed [Doebling and Hemez, 2001; Farrar and Lieven, 2007; Gupta and Ray, 2007; Pierce et al., 2008; Sankararaman et al., 2009] for probabilistic crack growth and life prediction. Among the various uncertainty sources considered, including natural variability, data uncertainty and model uncertainty, past experimental studies have suggested that the variability and uncertainty in the load spectrum have considerable influence on crack growth behavior and fatigue life [Moreno et al., 2003; Zapatero et al., 2005; Wei et al., 2002]. In addition to the extensive efforts that have been devoted to generate deterministic load spectra [Heuler and Klatschke, 2005; Xiong and Shenoi, 2008], it is desirable to characterize the uncertainty in the load spectrum based on data from monitoring systems, and develop robust future load prediction capability for damage prognosis.

Two types of methods have been developed to model the fatigue load spectrum, namely cycle counting methods and random process methods. The cycle counting methods extract counting matrices from load amplitude data, containing the information on the number of cycles, the mean value and the range of each cycle [ASTM, 2005]. Among various cycle counting methods, has been considered as the most efficient and accurate [Dowling, 1972]. Further, following certain rules, load history can be regenerated from the counting matrices [Khosrovaneh and Dowling, 1990]. The cycle counting methods can be directly applied to the estimation of fatigue damage by using an S-N curve-based fatigue damage cumulative law, such as the Palmgren-Miner linear rule [Miner, 1945]. The random process methods characterize fatigue load spectrum as a stochastic process. The Markov chain method treats loading as a discrete time Markov chain with stationary transition probabilities [Krenk and Gluver, 1989; Rychlik, 1996], retaining the correlation between adjacent turning points (load extrema). Note that load amplitudes are discretized into different levels in both the Markov chain method and the cycle counting method, and hence a relatively large transition matrix is required if the variation of loading amplitude is high. Further, it is assumed that the next turning point depends only on the previous turning point. This assumption may not be valid if strong autocorrelation exists in the load spectrum.

Frequency domain-based methods and time domain-based methods have been investigated to model the load spectrum as a random process with continuous state space (i.e., load extrema are not discrete) and more flexible autocorrelation assumption. Frequency domain-based methods characterize loading with power spectral density, spectral moments and bandwidth parameters, and these characteristics are related directly to fatigue damage estimation [Tovo, 2002; Benasciutti and Tovo, 2005, 2007]. The application of frequency domain-based methods to fatigue damage prognosis is not straightforward since prognosis-related issues, such as risk assessment and management, inspection and maintenance scheduling, operational decision-making, etc., are mainly defined in the time domain. The autoregressive moving average (ARMA) method [Box et al., 1994] is based on time series analysis and characterizes the fatigue load spectrum in the time domain. Available studies identify the order of the ARMA model based on some criteria, and then estimate the value of the unknown parameters of the model using time series data. The parameters are assumed to be deterministic and the variability and the uncertainty of loading are represented by a random noise term [Leser et al., 1994, 1998]. Several important issues remain unresolved. (1) It is unclear what order of model should be selected when several model orders have similar performance under identification criteria, and sometimes it is desired to incorporate multiple competing models. (2) The variability in loading comes from various environmental factors and the underlying mechanisms of the load spectrum can be complicated, and hence it may not be appropriate to lump all the variability into one single noise term. (3) The data used to estimate the ARMA model coefficients may not be sufficient and cause additional uncertainty.

It should be noted that all of the three aforementioned methods are applicable for stationary load spectra, i.e., the statistics of loading are assumed to be constant with respect to time. In practice, the loading condition may be non-stationary. Leser et al. [Leser et al., 1994, 1998] used a truncated Fourier series to account for the non-stationary part of the loading history and model the stationary part with an ARMA model. The Fourier series fitted from the time series data is periodic with the length of the data as period, but the real load history may not be periodic. Therefore, a more general modeling framework that accounts for the non-stationary load spectrum is desired.

While all the three above methods can predict future loading, it is also beneficial to know which method provides more accurate prediction before a particular method is chosen to apply to prognosis. The assessment of the confidence in the predictions of these methods can be done using model validation [Sargent, 2005]. Several types of validation techniques, such as conceptual, graphical, and quantitative validation, have been developed based on comparisons between data and model predictions [Knepell and Arangno, 1993]. Conceptual and graphical validation methods are shown inadequate in some conditions [Defense Modeling and Simulation Office, 1996]. Rebba et al. [Rebba et al., 2006] proposed a quantitative metric based on Bayesian hypothesis testing to validate computational models with controlled input, considering the measurement uncertainty in experimental data. This metric needs to be extended to assess the confidence in the prediction of fatigue loading using time series data, and the uncertainty in the model needs to be included.

The first part of this chapter investigates stochastic characterization and reconstruction of fatigue load history using the aforementioned methods, i.e., rainflow counting, the Markov chain method, and the ARMA method. The ARMA approach is found to be most versatile and useful for prognosis, and is adapted to include all three types of uncertainty -- natural variability, data uncertainty, and model uncertainty. The parameters of the ARMA model are assumed to be random variables with unknown probability distributions that are constructed to represent both the natural variability of loading and the uncertainty from sparse data. A probabilistic weighting method is applied to the ARMA model to incorporate multiple competing models (model uncertainty).

In the second part of this chapter, a continuous model updating approach based on real-time monitoring data is proposed to account for the time-variant features of the load history. Direct updating of the characteristic matrices is applied to the rainflow counting method and the Markov chain method. A Bayesian approach is used to update the probabilistic weights and coefficients of ARMA model. The relation between updating interval and the accuracy of model predictions is evaluated quantitatively. Further, a quantitative model validation method based on Bayesian hypothesis testing is proposed to assess the confidence in the prediction from the three different loading models, which can help in the selection of the most suitable model for damage prognosis. The proposed techniques are illustrated with numerical examples.

6.2 Characterization of load history

Section 6.2 discusses three different methods to characterize applied load based on available history data. Based on the models constructed, random samples of the anticipated load spectrum can be generated and used in stochastic fatigue damage prognosis. Note that the focus of this section is to stochastically characterize a random loading history, i.e., the real and reconstructed load histories are expected to be statistically equivalent, which is in terms of the characteristic matrix (i.e., counting matrix in the rainflow method and transition probability matrix in the Markov Chain method). The uncertainty quantification in loads will help further quantify the uncertainty in the prognosis output, like the remaining useful life.

6.2.1 Rainflow counting method and stochastic reconstruction

Among the well-established counting methods, the two-parameter rainflow counting method has the greatest significance for fatigue life prediction as it fully captures the basic damaging elements (the number, amplitudes, and mean values of load cycles) of a load history [Heuler and Klatschke, 2005], and can be used for uncertainty quantification of the variable amplitude load spectrum.

Following a certain set of rules [ASTM, 2005], the rainflow counting method extracts and counts cycles of various amplitudes and mean values, leaving only a residue behind. These load cycles are considered to be the basic elements of a load sequence. The final counting result is contained in a matrix A of size $k*k$, in which the element a_{ij} gives the number of counted cycles from load level i to load level j, and k is the number of load discretization levels [Amzallag et al., 1994].

Rainflow reconstructions are based on the idea of extracting cycles from the rainflow matrix and placing them in valid locations [Dressler et al., 1997]. Several rules are defined to ensure that cycles are inserted in such a way as to yield a similar rainflow matrix as the original signal [Khosrovaneh and Dowling, 1990]. Instead of calculating the fatigue life of a component based on a single load sequence, stochastic reconstruction allows for statistical evaluation of the fatigue life based on numerous load sequences that have the same statistical properties as the original spectrum. It should be noted that this method assumes that the original spectrum is representative of the typical load spectrum experienced by the component since all reconstructions are based on the rainflow matrix calculated from the original signal. The regenerated loading histories are equivalent in terms of the damage calculated by some ‘‘macroscopic” methods like the Palmgren-Miner rule, but will yield different crack growth results if fracture mechanics-based models are adopted. A modified rainflow algorithm accounting for the load sequence effect has also been developed [Anthes, 1997], and the regenerated loading histories can yield similar crack growth results using fracture mechanics-based models.

6.2.2 Markov chain method and transition probability matrix

For a realistic loading history, not only is the load amplitude at a certain time instant random, but the load amplitudes at adjacent time instants may also be correlated, e.g., the amplitude at time instant T_k can affect the amplitude at time instant T_{k+1}. Given this assumption, fatigue loading history with m discrete load levels is modeled as a discrete time Markov chain $\{X_n\}$, which is a Markov stochastic process whose state space (the set of discretized load levels) is a finite set, and for which n is a discrete time instant ($n = T_0$, T_1, T_2, . . . ,) [Karlin, 1966]. Let event $E_{k,i}$ denote that the loading amplitude at time instant

T_k is equal to load level i, and let event $E_{k+1,j}$ denote that the loading amplitude at T_{k+1} is equal to load level j. A one-step transition probability $P_{i,j}^{k,k+1}$ between $E_{k,i}$ and $E_{k+1,j}$ is defined as the probability of $E_{k+1,j}$, given $E_{k,i}$, i.e.,

$$P_{i,j}^{k,k+1} = \Pr(E_{k+1,j}|E_{k,i}) \tag{6.1}$$

With a further assumption that the one-step transition probability is independent of the time instants, i.e., the transition probability between $E_{k,i}$ and $E_{k+1,j}$ depends on i and j only, $P_{i,j}^{k,k+1}$ becomes stationary and can be simplified as $P_{i,j}$. which forms a stationary Markov chain transition probability matrix **P**. Note that the fatigue load spectrum is a series of extreme points, i.e., it is formed by minimum -- maximum -- minimum - . . . , etc. Hence, the transition matrix **P** is split into two triangular matrices $\mathbf{P}^u$ and $\mathbf{P}^l$. The elements of the upper triangular matrix $\mathbf{P}^u$ are the transition probabilities from minima to maxima, whereas the elements of the lower triangular matrix $\mathbf{P}^l$ are the transition probabilities from maxima to minima.

Given a load spectrum with discrete load levels from time T_0 to T_e, $P_{i,j}$ can be estimated using the number of occurrences that the event $E_{k,i}$ is followed by the event $E_{k+1,i}$, i.e.,

$$C_{i,j} = \sum_{k=0}^{e-1} I_{i,j}(k) \tag{6.2}$$

$$P_{i,j} = \begin{array}{ll} C_{i,j}/\sum\limits_{j=i+1}^{m} C_{i,j} & i < j, P_{i,j} \in P^u \\ C_{i,j}/\sum\limits_{j=1}^{i-1} C_{i,j} & i > j, P_{i,j} \in P^l \end{array} \tag{6.3}$$

where $I_{i,j}(k)$ is an indicator function:

$$I_{i,j}(k) = \begin{array}{ll} 0 & \text{if } E_{\mathrm{k,i}} \text{ is followed by } E_{k+1,\mathrm{j}} \\ 1 & \text{if } E_{\mathrm{k,i}} \text{ is not followed by } E_{k+1,\mathrm{j}} \end{array} \tag{6.4}$$

For example, consider a load history data containing four load levels [2, 1, 3, 4, 1, 4, 1, 4, 1, 3, 1, 3, 4], the corresponding **C** matrix is calculated as: [0, 0, 3, 2; 1, 0, 0, 0; 1, 0, 0, 2; 3, 0, 0, 0], and the resulting **P** matrix is: [0,0, 0.6, 0.4; 1, 0, 0, 0; 1, 0, 0, 1; 1, 0, 0, 0].

Once the transition matrix is obtained, random samples of loading history can be conveniently generated from a given initial extreme point.

6.2.3 ARIMA process loading

6.2.3.1 Autoregressive integrated moving average (ARIMA) model

The autoregressive integrated moving average (ARIMA) model is widely used in time series analysis for its flexibility. The expression of ARIMA model can be better explained by first introducing the ARMA model, which is a mix of the autoregressive (AR) and moving average (MA) models. AR model represents the value at the current time instant in terms of the values at the previous time instants. Hence, it is capable of capturing the autocorrelation between time series. MA model represents the deviation of the series at the current time instant from its mean value as a linear combination of errors in the past time instants. Combining a p^{th} order AR model and a q^{th} order MA model, an ARMA(p, q) model is obtained as [Hanke and Wichern, 2005]

$$\begin{aligned} Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + ... + \varphi_p Y_{t-p} + \\ \varepsilon_t - \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} - ... - \omega_q \varepsilon_{t-q} \end{aligned} \tag{6.5}$$

where Y_t is the value at time instant t; Y_{t-i} is the value at time instant $t-i$ (there are i time lags before t); φ_i is the coefficient of the AR model; ε_t is the random noise term with respect to time instant t; ε_{t-i} is the random noise term at time instant $t-i$; ω_i is the coefficient of the MA model.

ARMA models are used to describe stationary time series, i.e., the mean of the underlying process does not change with time. For processes with time-varying mean, we can use an ARIMA model with order (p, d, q), which is defined by summing (or "integrating") an ARMA(p, q) process d times [Box et al., 1994]. For example

$$\begin{aligned} Y_{t+1} &= Y_t + \nabla Y_t && \text{if} \quad d=1 \\ Y_{t+1} &= 2Y_t - Y_t + \nabla^2 Y_t && \text{if} \quad d=2 \end{aligned}$$

where $\nabla^d Y_t$ is an ARMA process (the expression can be obtained by replacing Y_t in Eq. 6.5 with $\nabla^d Y_t$).

To account for the natural variability in loading and the uncertainty due to insufficient data, the coefficients of the ARIMA model are assumed as random variables. Therefore the output of an ARIMA(p, d, q) model (denoted as M_i) also becomes an random variable, of which the probability density function can be written as

$$\begin{aligned} \pi(Y_t|M_i) &= \int \Big(\pi(Y_t|M_i, \boldsymbol{\varphi}^i, \boldsymbol{\omega}^i, \boldsymbol{Y_{-t}}) \\ &\qquad \pi(\boldsymbol{\varphi}^i, \boldsymbol{\omega}^i)\pi(\boldsymbol{Y_{-t}}|M_i)\Big) d\boldsymbol{\varphi}^i d\boldsymbol{\omega}^i d\boldsymbol{Y_{-t}} \end{aligned} \tag{6.6}$$

where $\boldsymbol{Y_{-t}}$ is the vector containing the model outputs in the previous time steps Y_{t-1}, Y_{t-2}, ..., Y_{t-p}. $\pi(Y_t|M_i, \boldsymbol{\varphi}^i, \boldsymbol{\omega}^i, \boldsymbol{Y_{-t}})$ is the conditional probability density function of output Y_t for this ARIMA model M_i and its associated parameters $\boldsymbol{\varphi}^i$ and $\boldsymbol{\omega}^i$, which can be derived from Eq. 6.5. $\pi(\boldsymbol{Y_{-t}}|M_i)$ is the joint probability density distribution of model outputs in the previous time steps, which is obtained using Eq. 6.6 in the previous time steps. $\pi(\boldsymbol{\varphi}^i, \boldsymbol{\omega}^i)$ is the probability density function of $\boldsymbol{\varphi}^i$ and $\boldsymbol{\omega}^i$, which can be assumed as uniform at the beginning of prognosis if no prior information is available.

6.2.3.2 Identification of ARIMA model

The first step to build an ARIMA model is to identify its orders p, d, and q. The order of differentiation d can be determined by examining if the time series become stationary after the d-th differentiation. The sample autocorrelation function and the sample partial autocorrelation function of the stationary time series data obtained from differentiation can be used to help identify p and q.

The autocorrelation function (ACF) for a stationary time series Y with mean μ and standard deviation σ is defined as:

$$\rho_k = \frac{E[(Y_t - \mu)(Y_{t+k} - \mu)]}{\sigma^2} \tag{6.7}$$

where the operator E refers to the expected value; ρ_k is the autocorrelation function for time lag k, i.e., the correlation between Y_t and Y_{t+k}.

The partial autocorrelation function (PACF) at time lag k is defined as a measure of the correlation between Y_t and Y_{t+k} without accounting for the effects of the values at intermediate time instants $(Y_{t+1}, \ldots, Y_{t+k-1})$. Derivation and estimation of the partial autocorrelation function are given in [Box et al., 1994].

It is known that each ARMA model has a unique pattern for its ACF and PACF [Hanke and Wichern, 2005]. For an ARMA(p, 0) model, its ACF dies out gradually while the PACF shows a cutoff after time lag p. For an ARMA(0, q) model, its ACF cuts off after time lag q while the PACF dies out slowly. For an ARMA(p, q) model, both the ACF and the PACF exhibit a smoothly decaying pattern [Box et al., 1994]. Hence, if the sample ACF and the sample PACF of the time series data match the theoretical pattern of a certain ARMA(p, q) model, this ARMA model with orders p and q is identified as the model desired.

However, subjectivity is involved while visually comparing the sample ACF and PACF with the theoretical values. To address the identification problem, a tentative initial model is first selected, and then the parameters and residuals associated with the selected model are estimated using the least square method. Hereafter, some statistics or criteria are used to check for adequacy; if the tentative model is shown to be inadequate to represent the data, it may be replaced by another model. The Ljung-Box Q statistic [Ljung and Box, 1978], which is a function of the residual autocorrelations and is approximated as a chi-square random variable, is used here to check the adequacy of the tentative model. The formula for Q is:

$$Q_m = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{n-k} \tag{6.8}$$

where Q_m is a chi-square variable with m-r degrees of freedom, and r is the number of the estimated parameters in the ARMA model; r_k is the residual autocorrelation at time lag k; n is the number of residuals; and m is the number of time lags considered in this test. If the p-value, which is equal to (1 - the cumulative probability of Q evaluated at Q_m), is not large enough, this tentative model is rejected.

6.2.3.3 Uncertainty in the ARIMA model

Real-time monitoring of the loads which will be used to estimate the accumulated fatigue damage is commonly known as Operational Loads Monitoring (OLM), which is part of a Health and Usage Monitoring System (HUMS) [Staszewski et al., 2004]. Two techniques have been applied to implement OLM, namely flight parameters-based loads monitoring and strain gauge-based loads monitoring. The flight parameters-based loads monitoring method uses built-in sensors of the aircraft to measure the parameters related to acceleration and mass, and then this information is used to estimate the loading sequence experienced by the components. The strain gauge-based loads monitoring method requires additional strain gauges to measure strain sequences at some chosen discrete locations, and then the recorded data is input to a pre-built load model to obtain loading sequence at any location of interest on the components. The measurement technique and the projection from raw data to a well-defined loading history inevitably cause variations and errors. Hence, the use of the loading sequence data from OLM in fatigue damage prognosis becomes a source of uncertainty.

In previous studies [Leser et al., 1994, 1998], the coefficients of ARIMA model have been treated as deterministic and estimated through a moment estimator. The inherent variability of loading amplitude and uncertainty from data are incorporated into the single noise term ε_t, which is assumed as an independently and identically distributed random process with zero mean and constant variance. Mechanical components usually work under complicated operating environments and many factors contribute to the variability of loading amplitudes. The uncertainty due to limited amount of data can also be significant. Therefore, a single noise term is not always sufficient. To accurately capture the aforementioned variability and uncertainty, the parameters - φ and ω - of ARIMA model, along with the noise term ε_t, can be assumed to be random variables. At the beginning of prognosis, if no information about the probability distributions of φ and ω is available, a uniform prior distribution may be assumed and further calibrated by usage monitoring data based on Bayes theorem, which will be explained in detail in Section 6.3.2.

Besides inherent variability of loading amplitudes and uncertainty from data, additional uncertainty arises in the selection of the appropriate ARIMA model, which is referred to as model form uncertainty. When multiple competing models can be considered, the tentative model identification method with the Q statistics can help eliminate models that are insufficient to represent the data, and there may still be several competing models left. The risk of choosing a single incorrect model can be reduced by considering several possible models. A straightforward way to incorporate multiple models is to assign a probabilistic weight to each of the competing models [Zhang and Mahadevan, 2000]. The probabilistic weight P_{Mi} for the model M_i represents the probability of the model M_i being correct. Combining the uncertainty in the ARIMA model parameters and the probabilistic weights, the probability density function of the fatigue loading amplitude at time t is

$$f(Y_t) = \sum_{i=1}^{n} P_{\mathrm{Mi}} \int f(Y_t|M_i, \varphi^i, \omega^i, Y_{-t}) f(\varphi^i, \omega^i) f(Y_{-t}) d\varphi^i d\omega^i dY_{-t} \tag{6.9}$$

where $\mathbf{Y}_{\text{-}t}$ is the vector containing the model outputs in the previous time steps $Y_{t\text{-}1}$, $Y_{t\text{-}2}$, …, $Y_{t\text{-}p}$, and M_i's (i = 1, 2, …, n) are the competing models. $f(Y_t|M_i,\varphi^i,\omega^i,\mathbf{Y}_{\text{-}t})$ is the conditional probability density function of loading amplitude Y_t for a given ARIMA model M_i and its associated parameters φ^i and ω^i, which can be derived from Eq. 6.5. $f(\mathbf{Y}_{\text{-}t})$ is the joint probability density distribution of loading amplitudes in the previous time steps, which is obtained using Eq. 6.9 in the previous time steps. $f(\varphi^i,\omega^i)$ is the probability density function of φ^i and ω^i, which can be assumed as uniform at the beginning of prognosis if no prior information is available. Similarly, all the values of probabilistic weights P_{Mi}'s can be assumed as $1/n$ if no information is available to support any single model. These prior assumptions can be calibrated based on usage monitoring data, as will be discussed in Section 6.3.2.

Once the probability distribution of loading amplitude with respect to time instant t is obtained, samples of future anticipated loading can be generated and applied in probabilistic fatigue prognosis.

6.3 Statistical updating of load models based on real-time monitoring data

The samples of anticipated load history required in stochastic fatigue prognosis can be generated through any of the three methods presented in Section 6.2 based on the available data. An assumption underlying the application of these methods is that the available data fully represent the load history and provides sufficient information to predict future loading. This assumption is challenged when the available data is limited and significant uncertainty exists. Further, the characteristics of loading may vary gradually with time, due to the change in operating environments of mechanical components. A continuous updating approach for each of the three load modeling methods is therefore proposed in this section based on real-time monitoring data. Section 6.3.1 presents a direct updating scheme for both the rainflow counting method and the Markov chain method. Section 6.3.2 presents a Bayesian approach for updating the ARIMA model.

6.3.1 Direct updating of the characteristic matrix

Both the rainflow counting method and the Markov chain method characterize fatigue load spectrum with a single matrix. In the rainflow counting method, the counting matrix stores the number of cycles from one load level to another load level; in the Markov chain method, the transition probability matrix stores the transition probability from one load level to another load level. As mentioned in Section 6.2.1 and 6.2.2, the elements of these two characteristic matrices are obtained based on the available load amplitude data, and samples of load spectrum can be generated. Once a new set of data is collected, the rainflow counting method and the Markov chain method are applied to obtain updated characteristic matrices. If the pattern of new data is different from the previous data, it can then be incorporated into the updated characteristic matrices. For the rainflow counting method, a new counting matrix can be obtained from the new data set, and it can be added directly to the previous characteristic matrix to obtain an updated matrix. Similarly for the Markov chain method, a new $\mathbf{C}$ matrix can be derived from the new data set, and then the addition of the new $\mathbf{C}$ matrix and the previous $\mathbf{C}$ matrix forms an updated $\mathbf{C}$ matrix. The elements $C_{i,j}$ are calculated by summing the indicator functions as shown in Eqs. 6.2 and 6.4. An updated transition probability matrix can be obtained from this updated $\mathbf{C}$ matrix as shown in Eq. 6.3. The updated characteristic matrices of these two methods can then be used to generate samples of load spectrum for the next period of prognosis.

6.3.2 Bayesian updating of the ARIMA model

6.3.2.1 Model Calibration based on the Bayes Theorem

Considering one of the competing ARIMA models M_i, with the associated parameters φ^i and ω^i, the load amplitude Y_t is predicted as:

$$Y_t = M_i(\varphi^i, \omega^i, t, Y_{-t}) \tag{6.10}$$

Note that the model M_i contains uncertainty from the random noise terms ε_t, $\varepsilon_{t\text{-}1}$, ..., ε_0, in addition to the variability in its parameters. The noise terms are assumed to be normal random variables with zero mean and variance σ_ε^2. σ_ε^2 can be estimated by comparing the model predictions with available data set $\mathbf{D}$, as follows:

$$\sigma_\varepsilon^2 = \frac{1}{n-1}\sum_k^n [Y_{\mathrm{Dk}} - Y_{\mathrm{tk}})]^2 \tag{6.11}$$

where Y_{Dk} is the k^{th} element in the data set, and Y_{tk} is the corresponding prediction of the model M_i with given values of parameters φ^i and ω^i; n is the number of the data points considered.

The conditional probability density function $f(Y_t|M_i,\varphi^i,\omega^i,\mathbf{Y}_{\text{-}t})$ of the model output Y_t of M_i for given values of φ^i and ω^i can be constructed based on Eqs. 6.10 and 6.11 and the probability distributions of the noise terms. Monte Carlo simulation can be used to estimated $f(Y_t|M_i,\varphi^i,\omega^i,\mathbf{Y}_{\text{-}t})$.

Assuming a joint prior distribution $f(\varphi^i,\omega^i)$ for φ^i and ω^i, the calibrated distribution of

φ^i and ω^i given a collection of data **D**, $f(\varphi^i, \omega^i\ |\mathbf{D})$, is obtained as:

$$f(\varphi^i,\omega^i|D) = \frac{\mathcal{L}(\varphi^i,\omega^i)f(\varphi^i,\omega^i)}{\int \mathcal{L}(\varphi^i,\omega^i)f(\varphi^i,\omega^i)d\varphi^i d\omega^i} \tag{6.12}$$

where the likelihood function of φ^i and ω^i, $L(\varphi^i,\omega^i)$, is the probability of observing the collected data for given values of φ^i and ω^i, which is calculated as:

$$\mathcal{L}(\varphi^i,\omega^i) = \int f(Y_D|M_i,\varphi^i,\omega^i,Y_{-t})f(Y_{-t})dY_{-t} \tag{6.13}$$

By assuming the data points are independent of each other, Eq. 6.13 can be further written as:

$$\mathcal{L}(\varphi^i,\omega^i) = \prod_{k=1}^{n}\int f(Y_{\mathrm{Dk}}|M_i,\varphi^i,\omega^i,Y_{-t})f(Y_{-t})dY_{-t} \tag{6.14}$$

The probabilistic weight of M_i, i.e., the probability of M_i being the correct model, can be calibrated using Bayes' theorem as:

$$P(M_i|D) = \frac{\mathcal{L}(M_i)P(M_i)}{\sum_i \mathcal{L}(M_i)P(M_i)} \tag{6.15}$$

where $P(M_i)$ and $P(M_i|\mathbf{D})$ are the prior weight and updated weight, respectively; $\mathcal{L}(M_i)$ is the likelihood function of M_i, that is, the probability of observing the data with the assumption that M_i is the correct model. $\mathcal{L}(M_i)$ is calculated as:

$$\mathcal{L}(M_i) = \iint \mathcal{L}(\varphi^i,\omega^i)f(\varphi^i,\omega^i)d\varphi^i d\omega^i \tag{6.16}$$

6.3.2.2 Continuous Bayesian updating of the ARIMA model

Model calibration based on the Bayes theorem can be applied to the ARIMA model continuously with real-time monitoring data. The updated model can then represent the pattern of the latest data without losing information contained in the previous data sets.

In addition, Bayesian updating can reduce the uncertainty in the model coefficients and the selection of models as more data are used. The continuous Bayesian updating can be implemented following the five steps shown below:

(1) At the beginning of prognosis, by using the initial set of data, the ARIMA models satisfying the Q statistic based-criteria (Eq. 6.8) are identified as competing models. If no prior information about the probability distribution of the corresponding model parameters, uniform distributions are first assumed as the priors. Note that these uniform distributions are independent because the model parameters are random variables independent from each other. Similarly, if no model is preferable from the prior knowledge, they are assumed to be equally weighted.

(2) The probability distributions of model parameters and probabilistic model weights are calibrated using Eq. 6.12 and Eq. 6.15 as mentioned in Section 6.3.2.1.

(3) With the estimated distributions of model parameters and weights, the probability distribution of predicted loading amplitude with respect to time is obtained using Eq. 6.9. Samples of the load spectrum are then generated with sampling techniques and applied in fatigue prognosis.

(4) After a new set of monitoring data is collected, Step (2) is again conducted by assuming the previously estimated ARIMA model parameter distributions and model weights as priors.

(5) Repeat Steps (2) to (4), until the end of the prognosis.

In the above continuous updating procedure, the length of the updating interval remains unclear. A shorter period length means monitoring data is retrieved more frequently and so is the updating. The increased data transmission activities will lower the battery life of the monitoring device, and more updating will increase the computational effort. It is useful to find an optimum time interval that balances prediction accuracy and efficiency. Therefore, it is desired to investigate the effect of the updating interval length on the accuracy of the ARIMA model prediction.

Note that the output of the ARIMA model is a random process indexed by time, and hence two quantitative statistical metrics -- mean square error MSE_p of mean prediction with respect to load history data, and the width W_p of the 95% prediction bounds are used to evaluate the accuracy of model output for a selected updating interval:

$$\mathrm{MSE}_p = (E(Y_t) - Y_{\mathrm{Dt}})^2 \tag{6.17}$$

$$W_p = F_{\mathrm{Yt}}^{-1}(0.975) - F_{\mathrm{Yt}}^{-1}(0.025) \tag{6.18}$$

$E(Y_t)$ is the mean prediction of ARIMA model at time t, whereas Y_{Dt} is the load amplitude data at time t; F_{Yt}^{-1} is the inverse cumulative probability function of Y_t, e.g., $F_{Yt}(F_{Yt}^{-1}(0.975)) = 0.975$. If the value of MSE_p is small, the prediction of ARIMA model is close to the real data, i.e., the prediction is accurate with the corresponding updating interval. If the value of W_p is small, the uncertainty in the prediction of ARIMA model with the corresponding updating interval is also small.

6.3.3 Summary

A continuous model updating approach has been developed in this section, which is capable of including information from real-time load monitoring. Direct updating of characteristic matrices is applied to the rainflow counting method and the Markov chain method. A Bayesian updating approach is applied to the ARIMA model through calibrating the probability distributions of the model coefficients and the values of the probabilistic weights. The effect of the updating interval on the accuracy of model prediction is investigated quantitatively.

6.3.4 Confidence assessment in load history prediction

In fatigue damage prognosis, the aforementioned models are used for the prediction of future loading, and it is desired to validate the prediction. Here we use the Bayesian equality hypothesis testing-based method illustrated in Section 4.3.2.

As discussed in the previous sections, the loading amplitude predicted by each of the three methods at a given time instant is a random variable due to various sources of uncertainty. Let $f(Y_t)$ be the probability density function of the output of a model M at time t. For ARIMA models, $f(Y_t)$ is obtained as shown in Eq. 6.9; for rainflow counting and the Markov chain method, $f(Y_t)$ can be estimated by directly sampling load histories from the corresponding characteristic matrices. Hence, the Bayes factor for a model M at time t can be derived as:

$$B(t) = \frac{P(Y_{\mathrm{Dt}}|H_0)}{P(Y_{\mathrm{Dt}}|H_1)} = \frac{\int L(Y_t) f(Y_t) \mathrm{dY}_t}{\int L(Y_t) g(Y_t) \mathrm{dY}_t} \tag{6.19}$$

where Y_{Dt} is the data collected at time t, and y_t is a particular value of Y_t at time t; $L(y_t)$ is the likelihood function of Y_t, which is the conditional probability of observing the data Y_{Dt} with a given value of Y_t; $g(Y_t)$ is the prior probability density of Y_t under the alternative hypothesis H_1, and it can be assumed as a uniform density function if no information on $g(Y_t)$ is available.

Since the collection of data is usually accompanied with measurement noise, which is assumed to be a lognormal random variable with unit mean and a relatively small standard deviation σ_m (<0.30), the likelihood function can be calculated approximately as [Haldar and Mahadevan, 2000]:

$$L(y_t) = f(Y_{\mathrm{Dt}}|Y_t) \approx \frac{1}{\mathrm{Y}_t\sqrt{2\pi\sigma_m^2}} \exp\{-\frac{1}{2}\frac{[\ln(Y_{\mathrm{Dt}}/Y_t)]^2}{\sigma_m^2}\} \tag{6.20}$$

The confidence metric (posterior probability) of the model M at time t, i.e., $C(t)$, can be calculated by substituting $B(t)$ from Eq. 6.19 into Eq. 4.21. Further, if the data is a discrete time signal collected with a sampling interval Δt_s, the overall confidence metric for the model M can be derived by averaging $C(t)$ over time as:

$$C_M = \frac{\sum C(t)\Delta t_s}{T} \tag{6.21}$$

6.4 Numerical example

A scaled helicopter combat maneuver loading history data including 510 cycles (1020 extrema/turning points) [Khosrovaneh et al., 1989] as shown in Fig. 6.1 is used for investigating the rainflow counting and reconstruction method, the Markov chain method, and the ARIMA model method. It can be observed from the data plot that the load history shows a time-variant pattern, and hence the proposed continuous updating approach is applied to the aforementioned three methods.

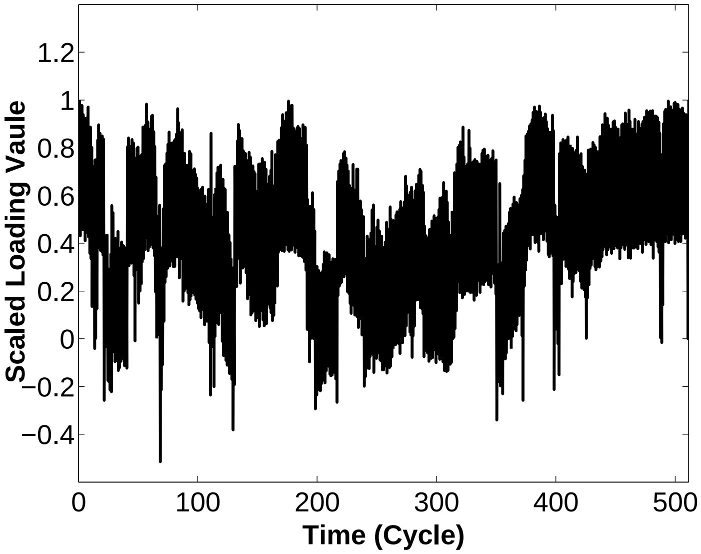
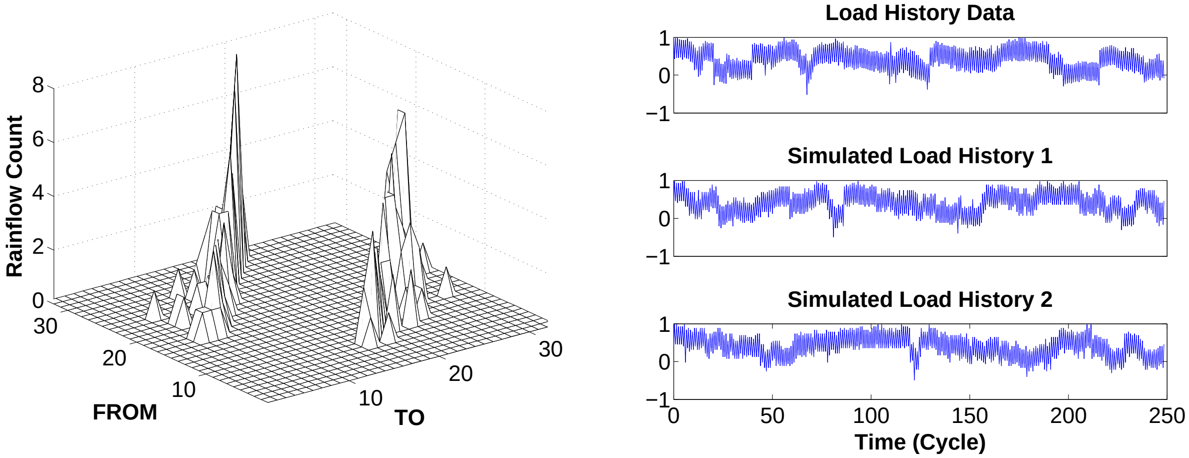


Figure 6.1: A Scaled Helicopter Combat Maneuver Load History Data

6.4.1 Rainflow counting, stochastic reconstruction and updating

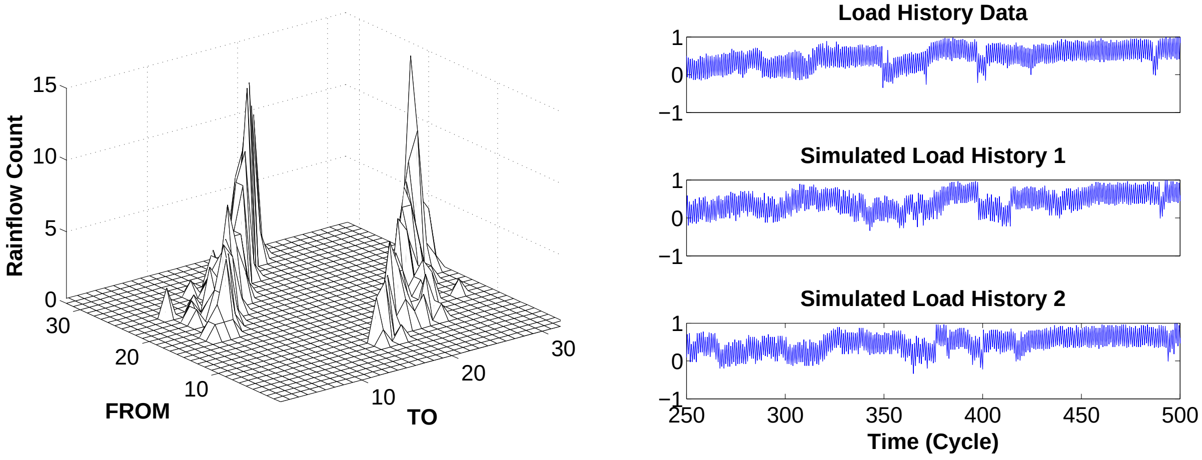
For the purpose of illustration, two subsets of the original data set are used, 1-250 cycles and 251-500 cycles. The first subset (1-250 cycles) is assumed to be the data currently available and is used to conduct the initial rainflow counting. A graphical representation of the counting matrix is shown in Fig. 6.2(a). Samples of simulated load history are generated from the counting matrix using the reconstruction technique introduced in Section 6.2.1, as shown in Fig. 6.2(b). The samples of load history show random rearrangements of the cycles extracted from the data. For the purpose of prognosis, the generated samples can be used as the prediction for future loading cycles, i.e., load amplitudes during 251-500 cycles, before the new real-time monitoring data is collected.



(a) (b)

Figure 6.2: (a) Graphical representation of rainflow counting matrix from the first subset of the load history data (1-250 cycles); (b) comparison of the load history data and two samples of simulated load history (1- 250 Cycles)

Consider the second subset of data (251-500 cycles) as the newly collected monitoring data, and then the direct updating method presented in Section 6.3.1 can be applied. First the rainflow counting technique is implemented on the new data and a new cycle counting matrix is obtained. This new matrix is added to the matrix counted from the previous data set (1-250 cycles) and then an updated counting matrix is obtained, as shown in Fig. 6.3(a). Further, samples of loading history are generated based on the updated counting matrix, and these samples can be again considered as predictions for future loading and used in probabilistic fatigue damage prognosis.

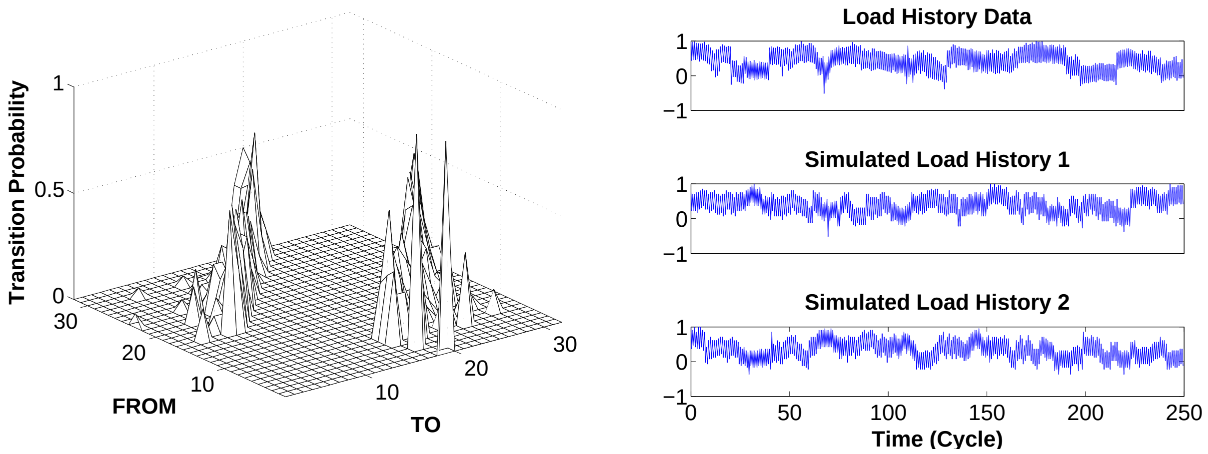


(a) (b)

Figure 6.3: (a) Graphical representation of the updated rainflow counting matrix using the second subset of the load history data; (b) comparison of the load history data and two samples of simulated load history (251-500 cycles)

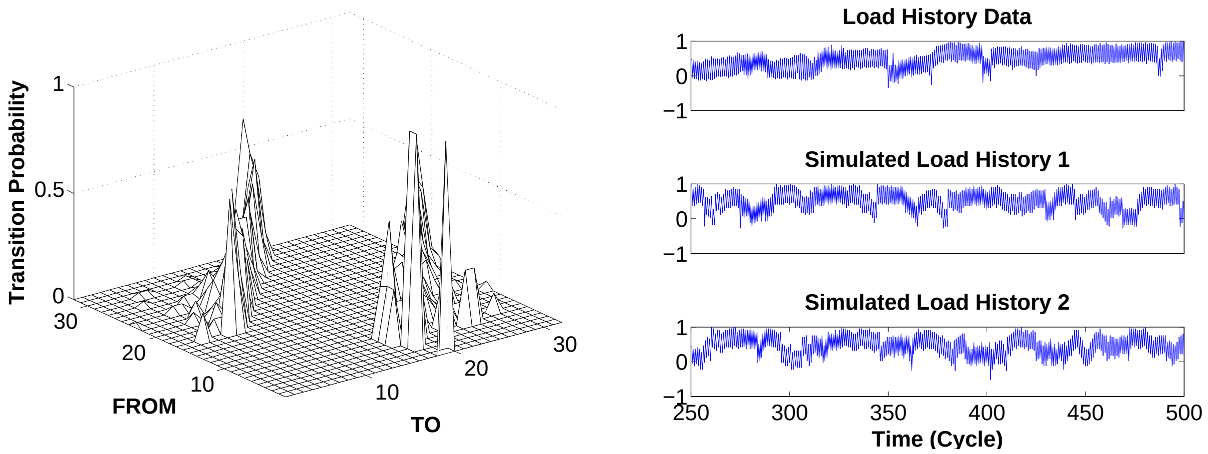
6.4.2 Markov chain method

The two subsets of data in Section 6.4.1 are also used to illustrate the Markov chain method and the updating of the transition probability matrix. The first data subset is assumed as the initially available data set, and the second data subset is the monitoring data set obtained later. Following the method presented in Section 6.2.2, the initial transition probability matrix is estimated using the first data subset and samples of the simulated load spectrum are generated as shown in Figs. 6.4(a)-(b). The generated samples are considered as the prediction of future loading and used in prognosis for the next time interval (251-500 cycles). After a new set of monitoring data is obtained (the second data subset), the initial transition probability matrix is updated and then predictions for future loading can be again generated for prognosis, as shown in Figs. 6.5(a)-(b).



(a) (b)

Figure 6.4: (a) Graphical representation of Markov chain transition probability matrix using the first subset of the load history data (1-250 cycles); (b) comparison of the load history data and two samples of simulated load history (1-250 cycles)



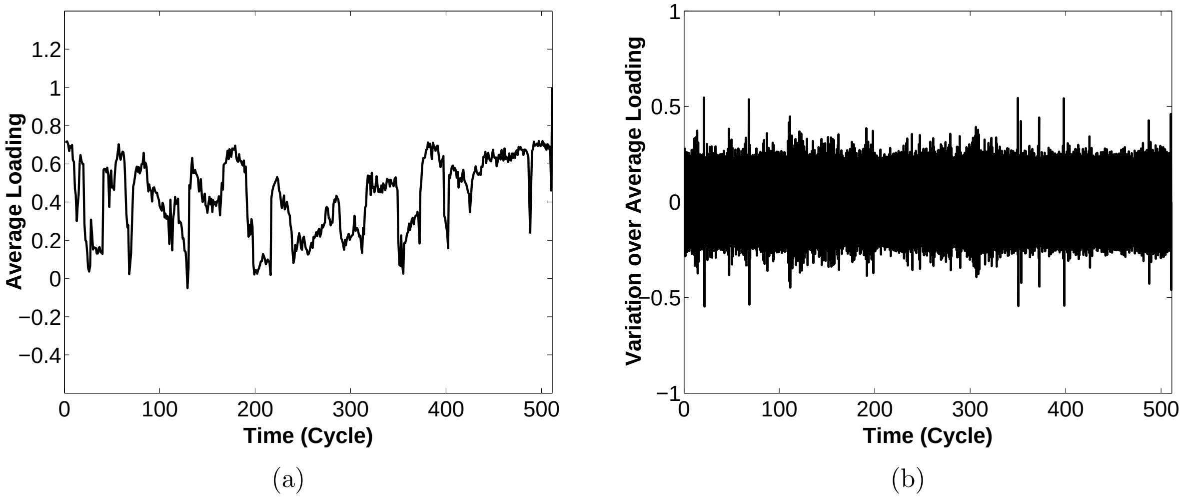
(a) (b)

Figure 6.5: (a) Graphical representation of updated Markov chain transition probability matrix using the second subset of the load history data; (b) comparison of the load history data and two samples of simulated load history (251-500 cycles)

6.4.3 ARIMA model method

6.4.3.1 Partition of data set and initial model identification

The entire data set (510 cycles) is used to demonstrate the extended ARIMA model method, the Bayesian approach, and the model verification and validation methodology. Due to the cyclic nature of fatigue loading, the load spectrum is split into two parts, the mean amplitude and the cycle variation, as shown in Fig. 6.6.

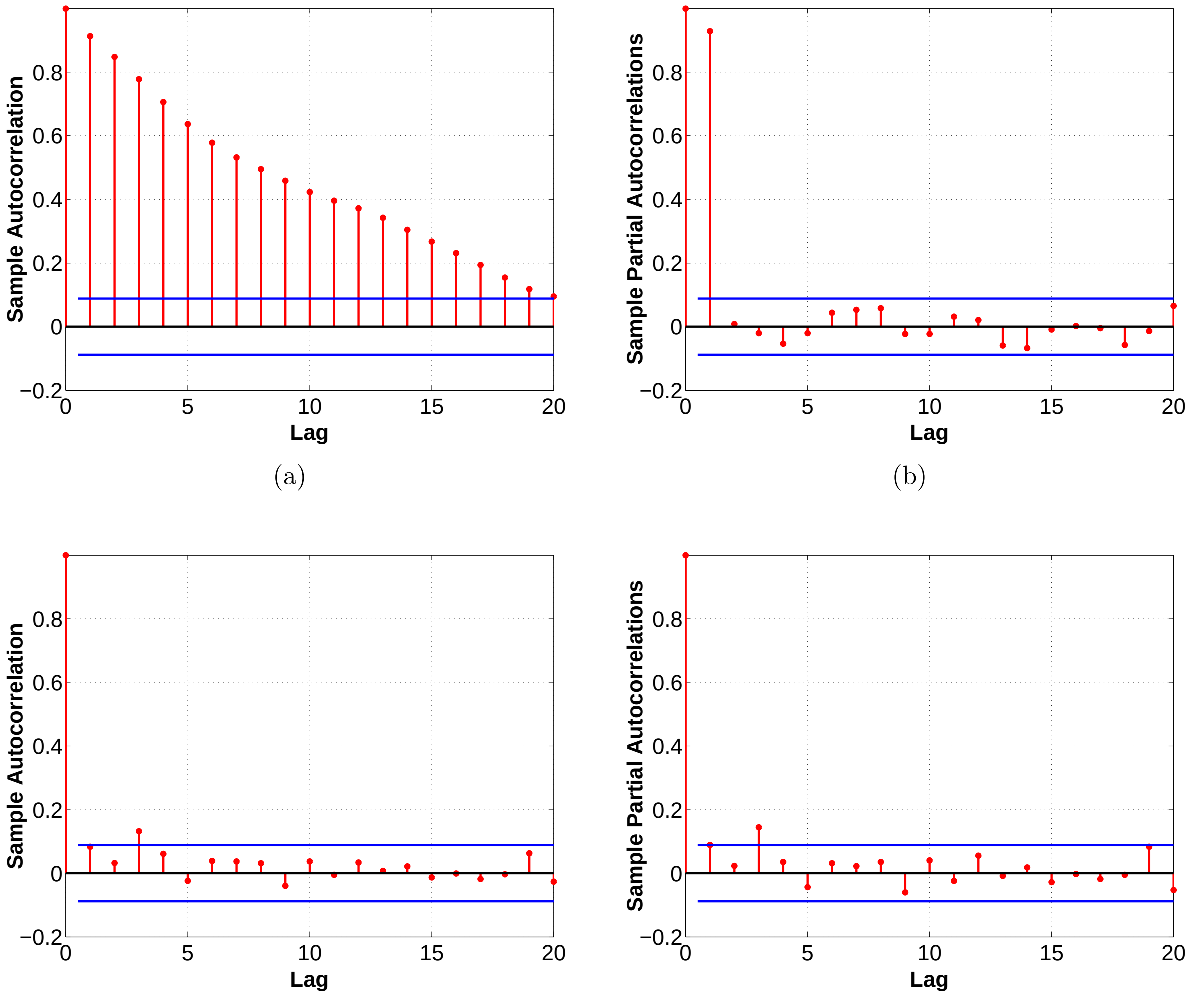


(a) (b)

Figure 6.6: (a) The mean amplitude of the load spectrum; (b) the cycle variation of the load spectrum

The sample ACF and PACF of the mean amplitude data, as shown in Figs. 6.7(a)-(d), suggest that the mean amplitude of the load spectrum can be modeled as an ARIMA process, whereas the lack of ACF and PACF in the cycle variation suggest that it can be modeled as a white noise.

To illustrate the continuous updating approach, the original data set is devided into several subsets. The first data set (1-250 cycles) is considered as the initially available data, and the following data sets (data set 2, 3, . . . , N) are assumed as real-time monitoring data retrieved subsequently.



(a) (b) (c) (d)

Figure 6.7: (a) Sample autocorrelation function (ACF) of the mean amplitude of the load history data; (b) sample partial autocorrelation function (PACF) of the mean amplitude of the load history data; (c) sample ACF of the cycle variation of the load history data; (d) sample PACF of the cycle variation of the load history data

The initial data set is used to identify possible ARMA models based on the Q statistics and the associated p-values presented in Section 6.2.3.2. As shown in Table. 6.1, both ARIMA(1,0,0) and ARIMA(2,0,0) pass the chi-square test since the corresponding p-values are significant. Therefore ARIMA(1,0,0) and ARIMA(2,0,0) are considered as candidate models for the load spectrum.

Table 6.1: Calculated Q statistics and the associated p-values

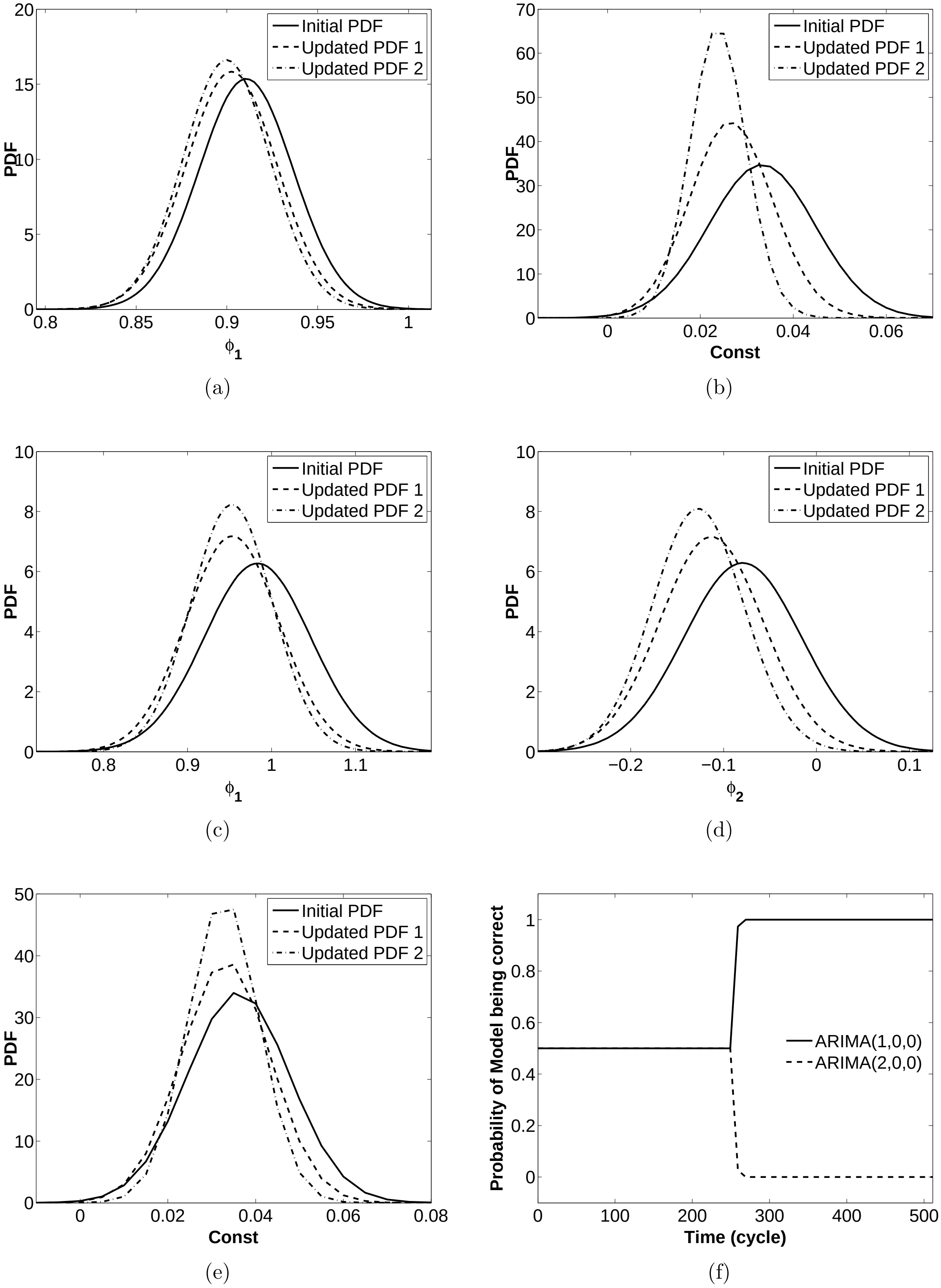
	Time Lag	12	24	36	48
ARIMA(1,0,0)	Ljung-Box Q	10.84	22.78	31.11	36.64
	p-value	0.37	0.41	0.61	0.84
ARIMA(2,0,0)	Ljung-Box Q	8.45	20.19	28.87	34.57
	p-value	0.58	0.57	0.71	0.89

6.4.3.2 Continuous updating of model parameters and probabilistic weights

The parameters of ARIMA(1,0,0), i.e., φ_0 and φ_1, and the parameters of ARMA(2,0,0), i.e., φ_0, φ_1, and φ_2, are treated as random variables. Following the Bayesian approach in Section 6.3.2.1, the initial probability distributions of these variables can be estimated by combining the likelihood functions from the initial data set and non-informative priors. Probabilistic weights are assigned to ARIMA(1,0,0) and ARIMA(2,0,0) respectively. The initial values of the weights are assumed equal to each other, i.e., the two candidate models are initially assumed to have equal probability of being the correct model for the loading history.

With the monitoring data set retrieved subsequently, the probability distributions of ARIMA model parameters and the probabilistic weights are continuously updated, as presented in Section 6.3.2.1 and 6.3.2.2. For the purpose of illustration, the plot of initial probability distributions, updated distributions using the second monitoring data set, and the updated distributions using the third monitoring data set are shown in Figs. 6.8(a)-(e). The plot of the updated values of the probabilistic weights is also shown in Fig. 6.8(f).

As shown in Figs. 6.8(a)-(e), the widths of the probability distribution functions of the ARIMA model parameters shrink gradually, i.e., the uncertainty due to sparse data decreases as more data is retrieved. The increasing values of the probabilistic weight for ARIMA(1,0,0) as shown in Fig. 6.8(f) suggest that ARIMA(1,0,0) obtained increasing support from monitoring data.

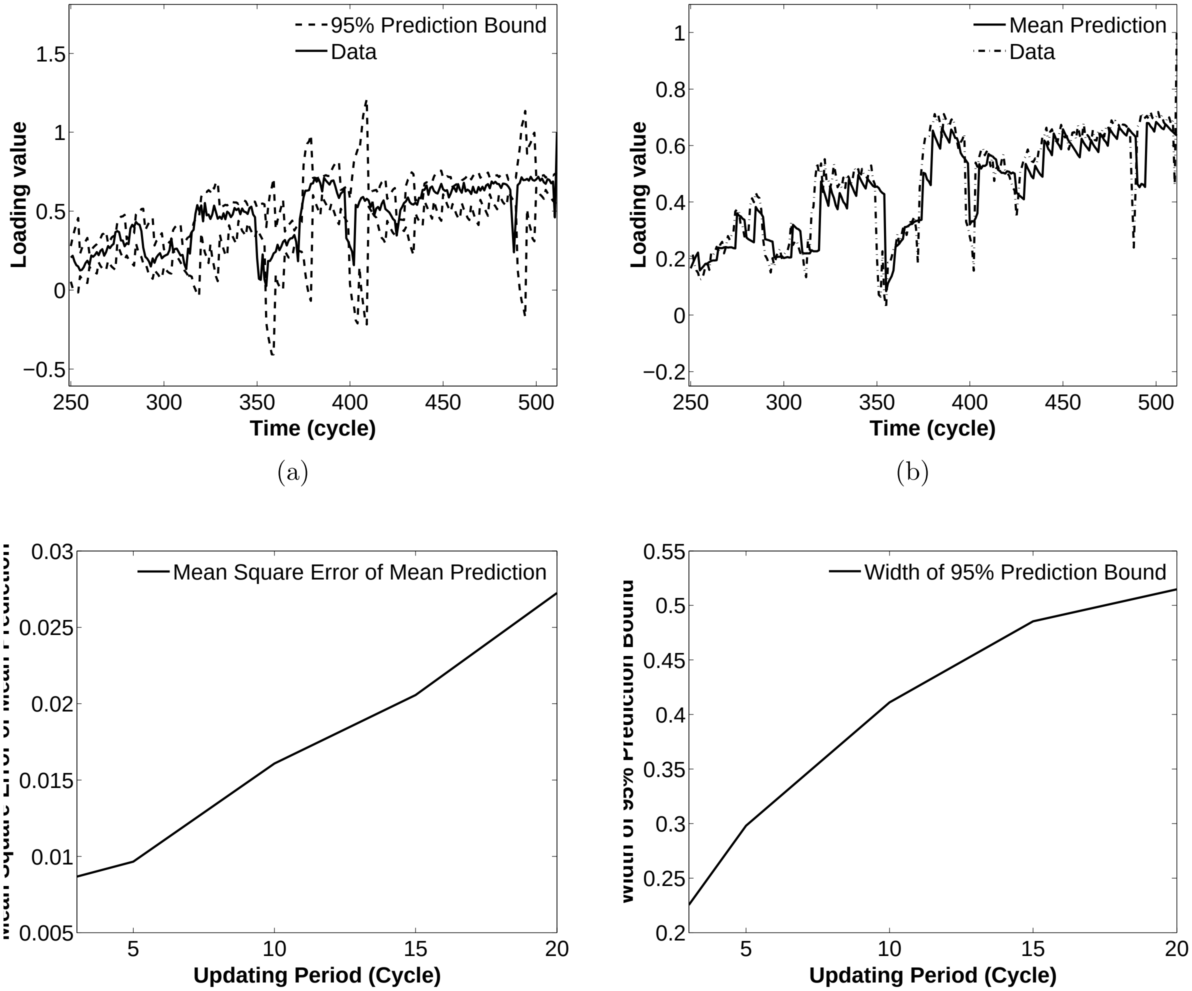


(a) (b)

(c) (d)

(e) (f)

Figure 6.8: (a) - (e) Initial probability distribution functions of the ARIMA model parameters -- φ_0, φ_1 of ARIMA(1,0,0), φ_0, φ_1, φ_2 of ARIMA(2,0,0) - and the updated distributions with newly collected data sets; (f) the updated values of the probabilistic weights versus time



(a) (b)

(c) (d)

Figure 6.9: (a)-(b) Prediction bound and mean prediction versus data when the updating interval is 5 cycles; (c) $\mathrm{MSE_p}$ of mean prediction versus the updating interval; (d) width of 95% prediction bound $\mathrm{W_p}$ versus the updating interval

The two metrics presented in Section 6.3.2.2, MSE_p of mean prediction with respect to load history data, and the width of 95% prediction bound W_p, are calculated to investigate quantitatively the relationship between the model prediction accuracy and the model updating interval. Figs. 6.9(a)-(b) give a graphical illustration of the two metrics when the updating interval is five cycles. Figs. 6.9(c)-(d) plot the relations between the two metrics and the

Table 6.2: Overall predictive confidence for the three methods

	Rainflow	Markov chain	ARIMA
Overall predictive confidence	0.68	0.61	0.71

updating interval. It is observed that W_p decreases with smaller updating interval, which suggests that the uncertainty in model prediction can be reduced by more frequent updating; similarly, MSE_p also decreases with smaller updating interval, which suggests that the accuracy of model prediction can be improved by more frequent updating. It is also seen that the model prediction can capture the time-variant feature of data by continuous updating.

6.4.4 Confidence assessment of model prediction

The predictions from the three methods above are validated using the confidence assessment method presented in Section 6.3.4. The data is partitioned in the same way as Section 6.4.3.1. The first data set (1-250 cycles) is considered as the initially available data, and the following data sets (data set 2, 3, . . . , N) are assumed as the monitoring data retrieved subsequently. To make a fair comparison between these three methods, each of them generates predictions for loading with a 5-cycle updating interval, i.e., each model is updated every 5 cycles and the models make predictions for loading in the next 5 cycles. By assuming that the measurement noise follows lognormal distribution with unit mean and standard deviation equal to 0.1, the Bayes factor value and the degree of confidence in predictions for 251-500 cycles are calculated using Eqs. 6.19-6.21.

As shown in Fig. 6.10, the Bayes factor values are higher than unity at most of the time points, which indicates that the predictions from all the three methods have good support from the time series data.

As shown in Table. 6.2, the overall predictive confidence for the ARIMA method is the

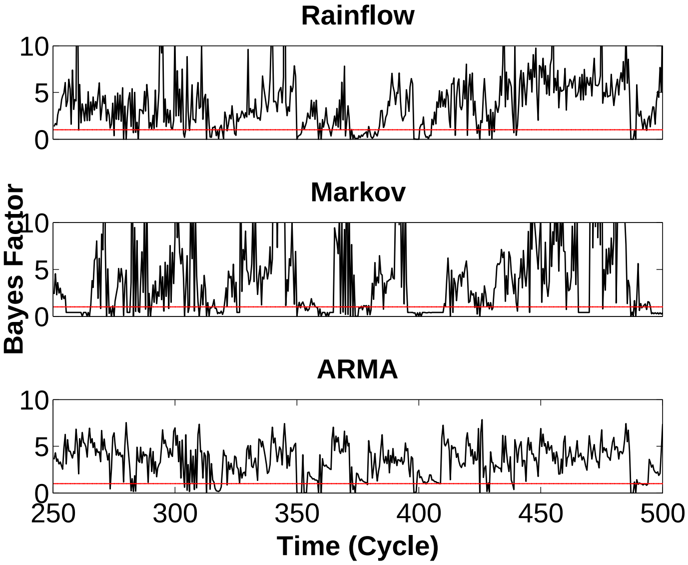


Figure 6.10: Bayes factor versus time

highest, which suggests that the ARIMA method has the best performance in the prediction of load amplitudes in this example.

6.5 Conclusion

Three different methods, namely rainflow counting method, Markov chain method and ARIMA model method, to characterize and reconstruct fatigue load spectra for use in prognosis were investigated. The ARIMA method was extended through random parameters and probabilistic weights to accommodate the inherent variability in loading, the uncertainty due to sparse data, and the uncertainty in model selection. A continuous model updating approach with real-time monitoring data was developed, including direct updating of the characteristic matrices for the rainflow counting method and the Markov chain method, and a Bayesian updating approach for the ARIMA model method. The relation between prediction accuracy and updating interval was investigated quantitatively. It is shown in Section 6.4.3.2

that the continuous updating approach can help the ARIMA model method to capture the time-variant feature of fatigue loading and also reduce the uncertainty in prediction due to limited data. A quantitative validation method based on Bayesian hypothesis testing was proposed to assess the predictive confidence of the three methods discussed. It is shown in Section 6.4.4 that all the three methods with the continuous updating approach perform comparably well for the load history example, as the Bayes factor values are larger than unity at most of the time points. The overall confidence metric suggests that the prediction of the ARIMA method has the best support from load amplitude data in the numerical example presented.

CHAPTER VII

INCLUSION OF TIME-DEPENDENT SYSTEM HEALTH MONITORING DATA

7.1 Introduction

Time-dependent reliability analysis is useful for risk management of a system in several ways: (1) optimum design of the system with reliability as a constraint [Kuschel and Rackwitz, 2000; Singh et al., 2010; Wang and Wang, 2012], (2) prognosis of the system, e.g., remaining useful life [Kulkarni and Achenbach, 2008; Zio, 2009], and (3) scheduling of inspection, maintenance, and repair based on the results of prognosis [Kong and Frangopol, 2003; Zhou et al., 2007; Niu et al., 2010]. This chapter will focus on uncertainty quantification in prognosis with application to aging aerospace mechanical components under fatigue loading.

The objective of fatigue prognosis is to obtain accurate estimation of remaining useful life (RUL), which has been a challenge due to the complexity and uncertainty in service environments and multidisciplinary damage mechanisms. The emerging techniques in both the areas of structural health monitoring (SHM) and fatigue damage prognosis (FDP) provide a promising future for tackling this challenge. Note that SHM and FDP are connected in nature, and a robust FDP relies on knowledge of the current status of components and service environment monitored by SHM system. Hence, in addition to the extensive research efforts conducted separately in SHM and FDP, integration of these two technologies is desired [Farrar and Lieven, 2007].

Based on the sensor-monitored data of external loading applied on mechanical components, different methods have been used to characterize and predict loading for fatigue damage prognosis, including rainflow counting [Amzallag et al., 1994], the Markov chain method [Rychlik, 1996], and ARMA (autoregressive moving average) modeling [Leser et al., 1998], etc. In Section VI, we have extended the ARMA modeling method to account for various uncertainty sources in service loading and developed a Bayesian approach to update the load model with real-time monitoring data. On-ground damage inspection for aerospace mechanical components using non-destructive inspection (NDI) techniques includes crack detection and size measurement. Different techniques using crack size measurement (CSM) data to infer the probability distribution of an equivalent initial flaw size (EIFS), which is the starting point of fatigue crack growth analysis, have been developed [Makeev et al., 2007; Cross et al., 2007; Sankararaman et al., 2010], including Bayesian approaches to account for multiple sources of uncertainty [Sankararaman et al., 2010, 2011b].

In the area of FDP, numerous fracture mechanics-based crack propagation models have been proposed to analyze the behavior of metal fatigue, and a summary of these models can be found in [Schijve, 2003]. Due to the stochastic nature of fatigue crack growth, a probabilistic prognosis method is desired. Studies have been conducted on probabilistic damage prognosis accounting for physical variability [Farrar and Lieven, 2007; Pierce et al., 2008; Gupta and Ray, 2007]. In a recent paper, we have developed a detailed uncertainty quantification approach for fatigue crack growth modeling that includes physical variability, data uncertainty and model uncertainty [Sankararaman et al., 2011a].

The purpose of this chapter is to develop a probabilistic methodology to integrate SHM results into a fracture mechanics-based FDP for aerospace mechanical components in a fleet, accounting for various sources of uncertainty and errors.

First, uncertainty quantification approaches for SHM and FDP are investigated. Two types of SHM data - real-time load monitoring data and on-ground crack inspection data - are considered, and the uncertainty due to the monitoring technique is quantified in Section 7.2. In Section 7.3, crack growth prognosis for mechanical components with realistic geometry and subjected to multi-axial variable amplitude loading is presented, with a focus on uncertainty quantification. Various sources of uncertainty and errors in prognosis are quantified, including physical variability in loading and material properties, data uncertainty due to the use of the structural health monitoring data and insufficient data, and model uncertainty and errors due to the use of various models in prognosis (crack growth model, loading model, finite element discretization error, etc.). A Bayesian network is constructed to systematically integrate the various uncertainties and errors, and a global sensitivity analysis is performed to identify the contributions of these sources to the uncertainty in the prognosis results (the predicted crack size after a number of loading cycles).

Section 7.4 proposes a framework to integrate SHM data with FDP. The fatigue loading sequence is characterized and predicted using an ARIMA (autoregressive integrated moving average) modeling method based on real-time load monitoring data. A Bayesian updating approach is used to estimate the coefficients of the ARIMA model and a probabilistic model averaging method is used to account for load model uncertainty. The probability distributions of EIFS and current crack sizes are inferred from the on-ground crack inspection data via a Bayesian method. The application of this integrated framework is shown for both individual components and a fleet of components. Sometimes it may be expensive to implement load monitoring and comprehensive inspection for the entire fleet. Only some of components may be selected for load monitoring and detailed examination, including crack detection and measurement, and the health status of the other components may have to be inferred combining existing data and model based-prognosis. Strategies of FDP for components in the fleet with different monitoring status are proposed. The prognosis results are validated using a Bayesian hypothesis testing method when new crack inspection data become available. A numerical example is presented in Section 7.5 to illustrate the overall framework of integrating prognosis with structural health monitoring under uncertainty.

7.2 Crack inspection data

Both real-time load monitoring and on-ground crack inspection are considered in this application since they are directly relevant to fatigue damage prognosis (FDP) for aerospace mechanical components. However, load monitoring was discussed in detail in Chapter VI. Therefore, this section will focus on crack inspection data.

On-ground damage inspection is a critical aspect of mechanical components maintenance. Various non-destructive inspection (NDI) techniques have been developed and applied to mechanical systems, including visual inspection, ultrasonics, eddy current, acoustic emission, X-ray, thermography and shearography [Staszewski et al., 2004]. However, the performance of NDI techniques is affected by many uncertain factors, such as the geometry of inspected components, sensitivity of inspection equipment, location of damage, operator skills, etc. Several probabilistic metrics have been developed to evaluate the performance of NDI, including probability of detection (POD), flaw size measurement accuracy, and false call probability (FCP) [Zhang and Mahadevan, 2001]. These criteria are developed from different methods, and they are used to evaluate different aspects of NDI performance. In the context of fatigue crack growth prognosis, FCP is not considered in this application since it is assumed that cracks exist in all the components inspected.

POD is usually a monotonic function of crack size and is used to represent the uncertainty in the crack detection. The result of crack detection (denoted as I_d) can be considered as a binary variable

$$I_d = \begin{cases} = 1 & \text{Crack detected} \\ = 0 & \text{Crack not detected} \end{cases} \tag{7.1}$$

The probability of the two possible values of I_d can be expressed in terms of POD

$$\begin{aligned} \Pr(I_d = 1|a_N) &= \mathrm{POD}(a_N) = f(a_N) \\ \Pr(I_d = 0|a_N) &= 1 - \mathrm{POD}(a_N) = 1 - f(a_N) \end{aligned} \tag{7.2}$$

where ”Pr” stands for ”Probability”. The value of POD function evaluated at crack size a_N, denoted as $f(a_N)$, can be obtained either by pure empirical methods [Spencer and Schurman, 1995] or model-assisted methods [Smith et al., 2007]. In this application, a statistical representation is adopted by treating $f(a_N)$ as a standard normal cumulative distribution function (CDF) [Berens and Hovey, 1983]

$$f(a_N) = \Phi(\alpha + \beta a_N) = \frac{1}{2}[1 + \mathrm{erf}(\frac{\alpha + \beta a_N}{\sqrt{2}})] \tag{7.3}$$

where Φ is the standard normal CDF, erf is the Gauss error function, and α and β are the parameters of this POD model.

The size measurement accuracy is used to quantify the uncertainty in experimental crack growth data, with the following expression determined by regression analysis [Zhang and Mahadevan, 2001]

$$a_m = \beta_0 + \beta_1 a + \varepsilon_m \tag{7.4}$$

where a is the actual crack size, a_m is the measured crack size, ε_m is the measurement noise term, and β_0 and β_1 are regression parameters. Note that this linear regression model is used only for the purpose of illustration. In practice, the size measurement accuracy can also be represented by other probabilistic models depending on the actual experimental data of a specific NDI technique [Heasler et al., 1993].

7.3 Fatigue damage prognosis under uncertainty

To quantify the uncertainty in fatigue damage prognosis, a significant amount of crack growth simulation analyses are required. However, for mechanical components with complicated geometry under service loading conditions (multi-axial, variable amplitude loading), cycle-by-cycle crack growth analyses and finite element analyses are usually needed to obtain accurate prognostic results, which render the uncertainty quantification extremely expensive. To address this issue, an efficient deterministic fatigue crack growth simulation method is presented in Section 7.3.1. Following it, detailed uncertainty quantification in fatigue crack growth prediction is presented in Section 7.3.2. Various sources of uncertainty and error are connected through a Bayesian network, and then a global sensitivity analysis is performed to identify the most important sources of uncertainty in fatigue crack growth prognosis.

7.3.1 Fatigue crack growth simulation under multi-axial variable amplitude loading

More than 20 different fracture mechanics-based fatigue crack growth models have been developed to simulate propagation of long cracks, including Paris' law [Paris and Erdogan, 1963], a modified Paris' law [Donahue et al., 1972], Forman's equation [Forman et al., 1967], Weertman's equation [Weertman, 1966], NASGRO equation [NASA and Southwest Research Institute, 2010], etc. Note that for a specific application, it may be unclear which model is correct, and the selection of model brings in uncertainty. This issue will be discussed in Section 7.3.2.1. For the sake of illustration, a modified Paris' law [Donahue et al., 1972] is used in this example, and a Wheeler's retardation model [Yuen and Taheri, 2006] is incorporated to account for the crack growth retardation effect due to the existence of overloads in a variable amplitude loading history

$$\frac{da}{dN} = \varphi^r C(\Delta K)^n (1 - \frac{\Delta K_{th}}{\Delta K})^m \tag{7.5}$$

where da/dN is the crack growth rate per loading cycle; φ^r is the retardation parameter and its calculation can be found in [Yuen and Taheri, 2006]; C, n, and m are parameters of this crack growth model, which can be estimated from existing crack growth data; ΔK is the stress intensity factor (SIF), and ΔK_{th} is the threshold SIF.

Note that Eq. 7.5 is not applicable to the short crack growth regime, i.e., the initial value of this differential equation - initial crack size a_0 - has to exceed the characteristic length of material microstructure. The concept of an equivalent initial flaw size (EIFS) was proposed to bypass short crack growth analysis and make direct use of a long crack growth law for total fatigue life prediction (including both the short crack growth and the long crack growth regime) [Johnson, 2010]. Note that this example adopts the damage-tolerance approach to fatigue life, and thus the total fatigue life does not include crack initiation period. Liu and Mahadevan [Liu and Mahadevan, 2009] related EIFS a_0 to material properties and component geometry based on the Kitagawa-Takahashi diagram [Kitagawa and Takahashi, 1976] and the El Haddad equation [El Haddad et al., 1979], and derived the relationship

$$a_0 = \frac{1}{\pi}(\frac{\Delta K_{th}}{Y \Delta \sigma_f})^2 \tag{7.6}$$

where ΔK_{th} is the threshold stress intensity factor and $\Delta \sigma_f$ is the fatigue limit (both are considered as material properties); Y is the geometry correction factor.

Further, in the case of multi-axial loading, Liu and Mahadevan [Liu and Mahadevan, 2007] adopted a characteristic plane method [Liu and Mahadevan, 2007] to combine the mode-I mode-II and mode-III SIFs (K_I, K_{II}, K_{III}) into a single equivalent SIF based on a fatigue limit criterion under multi-axial loading for 3D stress condition [Liu and Mahadevan, 2007], as

$$K_{mixed,eq} = \frac{1}{B}\sqrt{(k_1)^2 + (\frac{k_2}{s})^2 + (\frac{k_3}{s})^2 + A(k^H)^2} \tag{7.7}$$

where k_1, k_2, k_3 are the parameters associated with modes I, II, and III loading, respectively. k^H is related to hydrostatic stress. s is the ratio of Modes II and I threshold SIF. A and B are material parameters. Note that the formulation of this equivalent SIF is based on the principle of dimension reduction, i.e., finding a scalar SIF that can represent the multi-axial loading condition. Liu and Mahadevan [Liu and Mahadevan, 2007] verified the accuracy of this approach for several materials commonly used in the aerospace industry, including aluminum alloys and steel. Further verification is needed for other types of materials.

Analytical solutions of ΔK are usually unavailable for mechanical components with complicated geometry and under multi-axial loading. Therefore numerical methods like finite element analysis (FEA) are needed. However, as used in Eq. 7.5, the SIF ΔK needs to be evaluated for each loading cycle, which means thousands of FEA runs have to be performed for a high-cycle fatigue problem. To reduce the computation cost and render the uncertainty quantification analysis in the following section possible, the Gaussian process surrogate modeling technique illustrated in Section 2.5.1 is used to replace FEA in the cyclic crack growth simulation. In order to construct a GP model for the calculation of SIF, a few FEA runs are performed with different crack configurations and load combinations (including current crack size, aspect ratio, amplitudes of the applied bending and torsion moments) and the corresponding SIFs are calculated. Eq. 7.7 is used to calculate the equivalent SIF $\boldsymbol{K}$ values. Based on these training points with input combinations $\boldsymbol{x}$ and the corresponding SIF solutions $\boldsymbol{K}$, the parameters of the GP surrogate model are estimated, including the coefficients of its trend function $\boldsymbol{\beta}$, the process variance λ and the parameters of correlation function $\boldsymbol{\xi}$. Then, the mean prediction and variance of the SIF $K(x^*)$ for a particular input combination x^* (i.e., crack configuration and loading) are calculated as

$$\begin{aligned} \mathrm{E}[K(x^*)|\boldsymbol{K}] &= \boldsymbol{f}^T(x^*)\boldsymbol{\beta} + \boldsymbol{r}^T\boldsymbol{R}^{-1}(\boldsymbol{K} - \boldsymbol{F}\boldsymbol{\beta}) \\ \mathrm{Var}[K(x^*)|\boldsymbol{K}] &= \lambda(1 - \boldsymbol{r}^T\boldsymbol{R}^{-1}\boldsymbol{r}) \end{aligned} \tag{7.8}$$

where $\boldsymbol{f}(x)$ is the trend function; $\boldsymbol{r}$ is the vector of correlations between x^* and each of the training points, and $\boldsymbol{R}$ is the matrix of correlations among the training points. Both $\boldsymbol{r}$ and $\boldsymbol{R}$ are obtained by evaluating the correlation function $\boldsymbol{\xi}$. $\boldsymbol{F}$ is the matrix of the trend function $\boldsymbol{f}(x)$ at each of the training points.

Note that for deterministic crack growth simulation, only the mean prediction $\mathrm{E}[K(x^*)|\boldsymbol{K})$ is needed. The calculation of variance is needed in the following section to account for surrogate model uncertainty.

7.3.2 Uncertainty quantification in fatigue crack growth simulation

The fatigue crack growth analysis presented in Section 7.3.1 is deterministic and does not account for errors and uncertainty. An investigation on various sources of uncertainty in fatigue crack growth simulation is presented in this section.

7.3.2.1 Classification of uncertainty sources

The sources of uncertainty can be classified into three different types - physical variability, data uncertainty and model uncertainty - as discussed below. A graphical summary of the uncertainty sources classification is shown in Fig. 7.1.

(1) Physical or natural variability

The external loading applied on aircraft mechanical components in service condition is a result of multiple factors, such as air pressure, air flow, friction, and weights of other components, etc., and therefore is inherently stochastic. To represent this randomness, ARIMA models with a Gaussian random noise term is used as illustrated in Section 6.2.3.

Similarly, material properties could also be affected by many factors, such as manufacturing process, sample composition, micro-structure, etc., and variability in measurements. Fatigue limit and threshold stress intensity factor, two material parameters that are used to derive the prior estimation of EIFS, are treated as random variables to account for the natural variability in material properties. (Note: the geometry correction factor Y, and other material properties such as Young's modulus and Poisson's ratio can also be considered as random variables if the variations are not negligible.)

(2) Data uncertainty

The uncertainty associated with loading data includes the measurement noise and the error due to projecting raw data to a full loading sequence in OLM. To account for this uncertainty, the coefficients of ARIMA models estimated from this data are assumed as random variables. The uncertainty in crack inspection data includes crack detection and measurement uncertainty due to the use of NDI techniques, as discussed in Section 7.2. Also, the probability distributions of some material properties are inferred from laboratory experiments. This data may be sparse and cause uncertainty regarding the probability distribution type and parameters. Two probabilistic methods have been developed to account for sparse data, namely a flexible probability distributions-based method [Zaman et al., 2010] and a likelihood-based method [Sankararaman and Mahadevan, 2011a]. The uncertainty due to data size is taken into account while estimating the probability distribution of the threshold stress intensity factor and the fatigue limit, and this uncertainty is propagated to EIFS through the relation given in Eq. 7.6.

(3) Model uncertainty and errors

This study uses several models (crack growth model, ARIMA model, finite element model, surrogate model, etc.), and each of these models has its own error/uncertainty. Some of these errors are deterministic while others are stochastic; these need to be treated in different ways. The errors and uncertainties occur at different stages of the analysis, and may be combined in nonlinear, iterative or nested manner.

Model uncertainty comes from two sources: model form, and solution approximations. Model form uncertainty can be subdivided into two types: uncertainty in model coefficients and model fitting residual. The uncertainty in the coefficients of the modified Paris law - C, n and m - can be represented through their probability distributions. To represent the residual due to model fitting, a multiplicative lognormal random error term ε_{cg} with unit mean is introduced

$$\frac{da}{dN} = \varphi^r C(\Delta K)^n (1 - \frac{\Delta K_{th}}{\Delta K})^m \varepsilon_{cg} \tag{7.9}$$

The issue of model selection also exists while using ARIMA models to characterize and predict future loading. As shown in Section 6.2.3.3, a probabilistic model averaging method is used to address this uncertainty.

Finite element discretization error can be quantified using several methods [Ainsworth and Oden, 1997]. This chapter uses the Richardson extrapolation (RE) method, because it directly estimates the actual discretization error [Richardson, 1911]. Note that the use of RE requires the model solution to be monotonically convergent and the domain to be discretized uniformly (uniform meshing) [Rebba et al., 2006]. With the RE method, the discretization error for a coarse mesh is calculated as

$$\varepsilon_h = \frac{f_1 - f_2}{r^p - 1} \tag{7.10}$$

where f_1 and f_2 are solutions for a coarse mesh and a fine mesh respectively. If the corresponding mesh sizes are denoted by h_1 and h_2, then the mesh refinement ratio r is calculated as h_2/h_1. The order of convergence of p is calculated as

$$p = \frac{log(\frac{f_3 - f_2}{f_2 - f_1})}{log(r)} \tag{7.11}$$

where f_3 represents the solution for a finer mesh of size h_3, with the same mesh refinement ratio, i.e., r=h_2/h_1=h_3/h_2.

In this chapter, the role of FEA is to provide training data for the Gaussian process (GP) surrogate model as mentioned in Section 7.3.1. To incorporate the discretization errors, three FEA runs are perform for each of the training points to obtain f_1, f_2 and f_3. Then, the RE discretization errors are calculated and added to the coarse mesh solutions f_1 to approximate the true solutions, which become the training data for GP surrogate model.

The uncertainty due to the use of a GP model is reflected in the variance term in Eq. 7.8. Hence, by considering the SIF predicted from the GP model as a Gaussian random variable for a given crack configuration with the mean and variance values from Eq. 7.8, the uncertainty contribution of the surrogate model is also included.

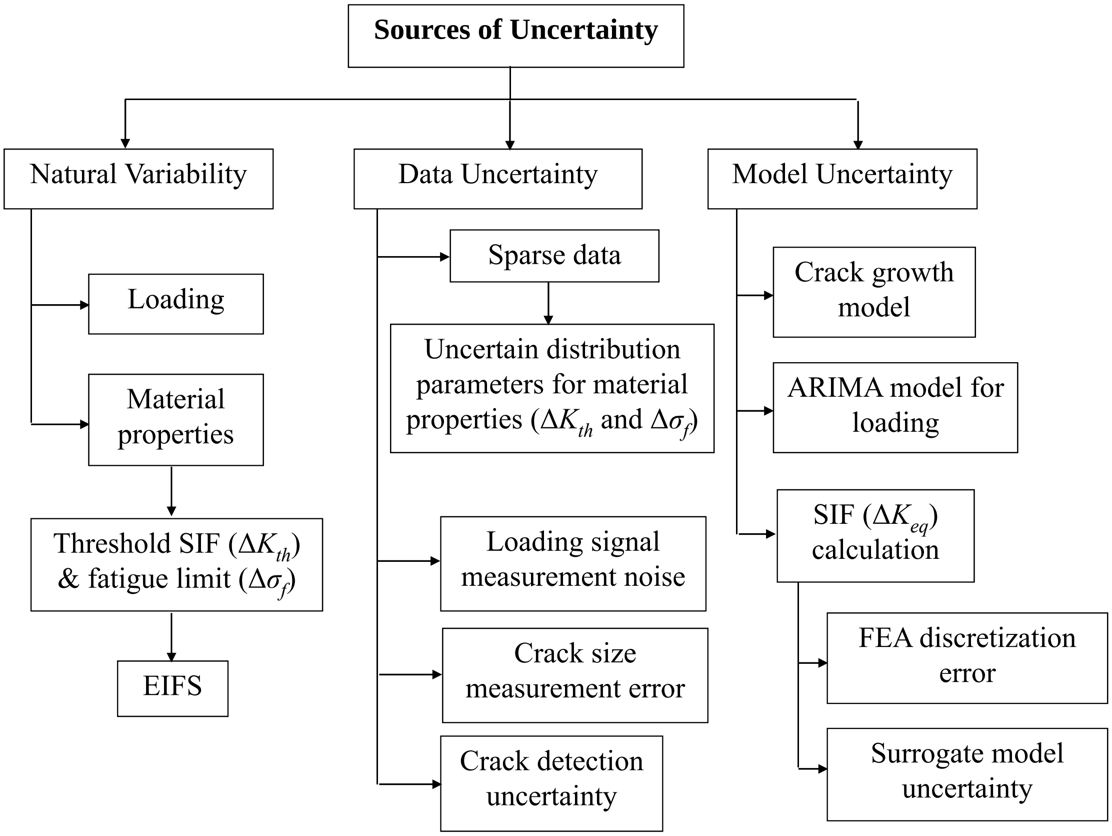


Figure 7.1: Classification of uncertainty sources in fatigue crack growth analysis

7.3.2.2 Connection of uncertainty sources using a Bayesian network

As introduced in Section 2.1, a Bayesian network is a graphical representation of uncertain quantities that explicitly incorporates the probabilistic causal dependence between the variables as well as the flow of information in the model. In this section, we construct a

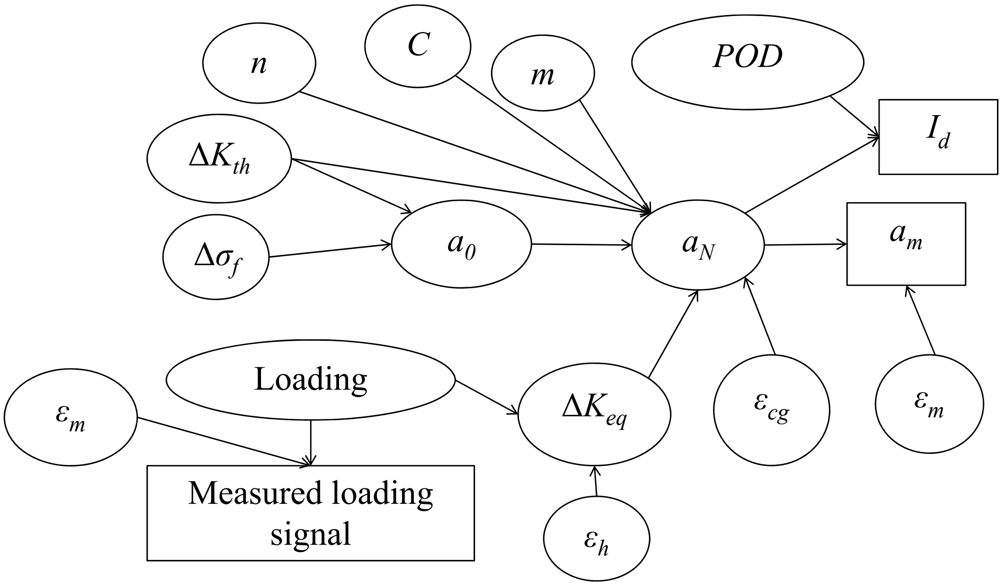


Figure 7.2: Bayesian network to connect sources of uncertainty and errors

Bayesian network for fatigue crack growth analysis under various sources of uncertainty and errors as shown in Fig. 7.2. The conditional probability between each node is given by the equations and models in the aforementioned crack growth analysis.

7.3.2.3 Probabilistic sensitivity analysis

A probabilistic sensitivity analysis (PSA) is required to study the contributions from each of the uncertainty sources to the overall uncertainty in the prognostic output - the crack size after a particular number of cycles of operation. Identifying the uncertainty sources that have significant contributions can help in allocating resources to monitor the important uncertainty sources and hence improve the efficiency and accuracy of the overall SHM-FDP system. Various methods have been developed to perform PSA, including variance-based global sensitivity analysis [Saltelli et al., 2008] and the Kullback-Leibler (K-L) entropy-based method [Liu et al., 2006]. The variance-based global sensitivity analysis provides a global measure of the contribution from a particular source and is adopted in this chapter.

Consider a model with n random input variables X_1, X_2, ..., X_n

$$Y = G(X_1, X_2, X_3, ..., X_n) \tag{7.12}$$

The first-order effect index quantifies the individual contribution from an input variable X_i without interaction with other variables, and is given by

$$S_i^1 = \frac{\mathrm{Var}_{X_i}[\mathrm{E}_{X_{\sim i}}[Y|X_i]]}{\mathrm{Var}[Y]} \tag{7.13}$$

where E is the expectation operator, Var is the variance operator, and $X_{\sim i}$ stands for all the input variables except for X_i. $\mathrm{E}_{X_{\sim i}}$ calculates the expected value of the model output Y when X_i is fixed at a specific value; Var_{X_i} computes the variance of this expected value when the randomness of X_i is included.

The contribution from the variable X_i including its interaction with all other variables is known as the total effects index and can be calculated as

$$S_i^T = 1 - \frac{\mathrm{E}_{X_{\sim i}}[\mathrm{Var}_{X_i}[Y|X_{\sim i}]]}{\mathrm{Var}[Y]} \tag{7.14}$$

where Var_{X_i} computes the variance of Y when all the input variables except for X_i are fixed at specific values; $\mathrm{E}_{X_{\sim i}}$ calculates the expected value of this variance considering the randomness of $X_{\sim i}$.

In this chapter, all X_i's denote the various sources of uncertainty in fatigue crack growth prognosis as classified in Section 7.3.2.1. The calculation of sensitivity indices in Eqs. 7.13 and 7.14 by brute force nested Monte Carlo simulation requires a large number of model evaluations, and hence an efficient method reported in [Saltelli, 2002] is used, which designs an approximation-based strategy to reduce the number of model evaluations required by an order of magnitude.

7.4 Use of crack inspection data within prognosis

System monitoring provides data to quantify the uncertainty in two of the most important inputs in fracture mechanics-based fatigue crack growth prognosis - external loading and the equivalent initial flaw size (EIFS). In this application, we adopt the ARIMA model-based approach as illustrated in Chapter VI, which provides a rigorous framework for uncertainty quantification of loading using real-time monitoring data. The estimation of EIFS using crack inspection data is illustrated below.

7.4.1 Inference of EIFS using crack inspection data

As mentioned in Section 7.3.1, the concept of the equivalent initial flaw size (EIFS) is used to bypass the complexity due to small crack growth and a long crack growth model can be used regardless of the scale of the actual initial crack size. Since the EIFS is not a physical quantity and hence cannot be measured directly. Eq. 7.6 gives an estimation of EIFS considering two material properties (the threshold SIF and the fatigue limit) and the geometry of component. Fatigue analysis based on this EIFS estimation has been validated using experimental data under constant amplitude, uni-axial and multi-axial loading [Liu and Mahadevan, 2009; Xiang et al., 2010; Lu et al., 2010]. However, its accuracy for variable amplitude loading is not clear so far. Meanwhile, approximation is needed while calculating the geometry factor for components with realistic geometry. Given these considerations, it is useful to consider the probability distribution of EIFS estimated from Eq. 7.6 as a prior and then update this distribution using Bayes' theorem

$$\pi(a_0|D) = \frac{\prod_i^M Pr(D_i|a_0)\pi(a_0)}{\int \prod_i^M Pr(D_i|a_0)\pi(a_0)da_0} \tag{7.15}$$

Consider three possible outcomes of crack inspection: (1) no crack is detected; (2) crack is detected but the size is not measured; (3) crack is detected and the size is measured. The likelihood function $\prod_i^M Pr(D_i|a_0)$ is the conditional probability of M crack inspection results given a particular EIFS value a_0, which can be written for each of the three possible outcomes as

$$\begin{aligned} \text{Case 1:} \quad & Pr(D_i|a_0) &&= \int f(a_N)\pi(a_N|a_0)da_N && (7.16) \\ \text{Case 2:} \quad & Pr(D_i|a_0) &&= \int [1 - f(a_N)]\pi(a_N|a_0)da_N && (7.17) \\ \text{Case 3:} \quad & Pr(D_i|a_0) &&\propto \pi(a_m|a_0) = \int \pi(a_m|a_N)\pi(a_N|a_0)da_N && (7.18) \end{aligned}$$

where $f(a_N)$ is the probability of detection (POD) and can be obtained from Eq. 7.3; $\pi(a_m|a_N)$ is the conditional probability density function of measured crack size for a given actual crack size, which can be derived from Eq. 7.4. By assuming the regression coefficients β_0 and β_1 as constants, and assuming the residual term ε_m as a zero-mean normal random variable, this conditional probability can be derived as

$$\pi(a_m|a_N) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{a_m - \beta_0 - \beta_1 a_N)^2}{2\sigma_m^2}] \qquad (7.19)$$

where σ_m is the standard deviation of ε_m.

7.4.2 Strategy of prognosis for components in a fleet

For a fleet of aerospace mechanical components in service, an ideal case will be to conduct structural health monitoring for each component, followed by individual prognosis. However, due to budget or technical limitations, load monitoring data and crack inspection data may or may not be available for every component in the fleet. Four scenarios of prognosis can be classified based on the availability of monitoring data as shown below.

Scenario 1: Prognosis for components with OLM and CSM

This is the best situation where prognosis is tied to the individual component. The future loading can be predicted using the Bayesian ARIMA model presented in Sections 6.2.3 and 6.3.2, and the initial values of the ARIMA model are the latest recorded loading amplitudes.

Before crack inspection, the prognosis starts from the EIFS with the prior probability distribution derived from Eq. 7.6. After the crack inspection, the prognosis of future crack growth starts from the current crack size, which can be obtained using the measured crack size by rewriting Eq. 7.4 as

$$a = \frac{1}{\beta_1}(a_m - \beta_0 - \varepsilon_m) \tag{7.20}$$

Scenario 2: Prognosis for components without OLM but with CSM

For components without loading monitored, future loading predictions are generated with random initial values using the ARIMA model estimated in Scenario 1. The extra uncertainty due to the random initial value will require more Monte Carlo simulations for prognosis than in Scenario 1.

The use of the prior EIFS distribution before inspection and the estimation of the current crack size based on measured crack size is the same as in Scenario 1.

Scenario 3: Prognosis for components with OLM but without CSM

In this case, the prediction for future loading is the same as in Scenario 1. The use of the prior EIFS distribution before inspections is the same as in Scenario 1. However, after inspections, since the crack size is not measured, crack growth simulation based on updated EIFS is used to infer the probability distribution of current crack size. Note that this procedure also applies to (i) components without any crack inspection and (ii) components with crack detection but without size measurement. The updated probability distribution of EIFS is obtained using crack inspection data via the Bayesian approach as presented in Section 7.4.1.

Scenario 4: Prognosis for components with no OLM or CSM

This is the worst case, where no load monitoring data or inspection data are available for an individual component. In such a situation, the prognosis for such a component has to be based on data obtained on other components. In this case, the prediction for future loading is the same as in Scenario 2. The use of the prior EIFS distribution before inspections is the same as in Scenario 1, and the estimation of the current crack sizes is the same as in Scenario 3.

7.4.3 Validation of prognosis with new crack inspection data

It is important to assess the validity and performance of prognostic algorithms based on the comparison between predictions and observed data, and various graphical and quantitative methods have been developed for this purpose [Oberkampf and Barone, 2006; Hills and Leslie, 2003; Saxena et al., 2010; Rebba et al., 2006]. A detailed illustration and discussion of the existing quantitative model validation methods has been given in Section 2.3 and Chapter IV. The Bayesian equality hypothesis testing-based approach illustrated in Section 4.3 is used here since it takes into account the entire probability distribution of model output, instead of only the distribution parameters, and a confidence metric for model prediction can be easily derived based on the calculated validation metric. Two hypotheses are compared, namely H_0 - the null hypothesis that the proposed method gives correct predictions, and H_1 - the alternative hypothesis that the proposed method gives incorrect predictions. The validation metric - Bayes factor - is equal to the ratio of the likelihood functions of these two hypotheses [O'Hagan, 1995]

$$B = \frac{Pr(D|H_0)}{Pr(D|H_1)} = \frac{\prod_i \int Pr(D_i|a_N)\pi_0(a_N)da_N}{\prod_i \int Pr(D_i|a_N)\pi_1(a_N)da_N} \tag{7.21}$$

where $Pr(D_i|a_N)$ is the conditional probability of obtaining the inspection data D_i for a given actual crack size. Eqs. 7.2 and 7.19 are used to calculate this conditional probability for three different types of inspection results. In Eq. 7.21, $\pi_0(a_N)$ is the PDF of the actual crack size under the null hypothesis, and $\pi_1(a_N)$ is the PDF of the actual crack size under the alternative hypothesis. $\pi_0(a_N)$ is the same as the PDF of crack size predicted by the prognosis. In order to calculate $\pi_1(a_N)$, we assume that under the alternative hypothesis the crack size follows a uniform distribution, i.e., the crack sizes within a certain interval are equally probable. Note that the boundaries of this interval will affect values of the estimated Bayes factor.

In addition to crack inspection data, load monitoring data can also be used for validation. An example of using loading data to validate the prediction of the ARMA model can be found in Section 6.3.4. If new SHM data on system health status or damage response are available, further validation can be performed by comparing the system response estimated based on crack size predictions against the SHM data.

7.5 Numerical example

This section provides a numerical example of the proposed methodology for model-based fatigue damage prognosis, which incorporates information from real-time loading monitoring and on-ground crack inspection.

7.5.1 Problem description

A two-radius hollow cylindrical component subjected to bending and torsion is considered, similar to a rotorcraft mast. (The mast also experiences an axial thrust load; the effect of this load is much smaller than that of bending and torsion in this problem and therefore the axial load is ignored in this example). Assume that the crack inspection is for a possible elliptical surface crack in the fillet radius region. Nominally identical specimens of the cylindrical component may be used in a fleet of rotorcraft. Our interest is in crack growth prognosis for individual components under the four scenarios discussed in section 7.4.2, by making use of the load monitoring and crack inspection data.

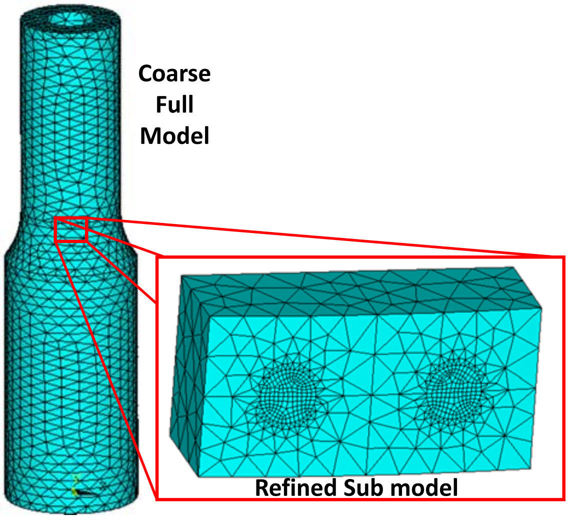


Figure 7.3: Surface crack model

Crack growth prognosis in this example uses linear elastic fracture mechanics (LEFM), and assumes a planar crack. The commercial finite element software ANSYS (version 11.0) is used to calculate the mode I, mode II, and mode III stress intensity factors. A sub-modeling technique is used to facilitate computational efficiency in finite element analysis, as shown in Fig. 7.3. First the entire structure is modeled with a coarse mesh and solved; this is referred to as the full model. Then the region surrounding the crack is modeled using a refined mesh, which is called the sub-model. The boundary conditions of the sub-model are obtained from the solution of the full model.

Table 7.1 list the material and geometrical properties of the components under study, and the values are assumed as constant in this example. Note that in reality, the parameters in Tables 7.1 may be random and require probabilistic treatment.

A modified Paris law is used for crack growth analysis, combined with a Wheeler's retardation model, as described in section 7.3.1. Cracks are assumed as elliptical (characterized by the crack sizes along x-axis and y-axis). Five hundred and two runs of FEA are conducted, accounting for different combinations of input variables - crack sizes along x-axis, aspect

Table 7.1: Material and geometrical properties

Material properties of Al 7075-T6		Geometrical properties of cylinder	
Modulus of Elasticity	71.7 GPa	Length	0.152 m
Poisson Ratio	0.33	Inside radius	7.62 mm
Yield Stress	691 MPa	Outside radius (narrow sect)	15.24 mm
Ultimate Stress	764 MPa	Outside radius (wide sect)	20.3 mm

ratios, bending moments (torsion is assumed to be proportional to bending), and the SIFs at crack tip are calculated.

7.5.2 Uncertainty and error quantification

The various sources of uncertainty and errors discussed in Section 7.3.2.1 are quantified for this application example as discussed below. Details are provided in Table 7.2. Note that some of the statistics in Table 7.2 can be found in the literature, such as the statistics of ΔK_{th} and $\Delta\sigma_f$, which are material properties. Some of the statistics, e.g., the statistics of model uncertainty ε_{cg}, are assumed for the sake of illustration, and they can be conveniently replaced in practice if more information is available.

Physical variability The probability distributions of the fatigue limit and threshold stress intensity factor are found in [Liu and Mahadevan, 2009]. Further, the prior distribution of EIFS is derived from Eq. 7.6, which is a lognormal distribution with parameters λ and ζ in this example. The loading is characterized using the ARIMA modeling method presented in Section 6.2.3, based on load monitoring data. The Gaussian noise term ε_t in ARIMA models represents the physical variability in loading. In this example, the data are simulated using an ARIMA(2,1,0) process. Note that these synthetic data are only used for the purpose of illustration, and could be replaced by real world load monitoring data if available.

Data uncertainty The coefficients in the expression of POD and the CSM uncertainty are assumed as shown in Table 7.2. To account for uncertainty due to sparse data, the two distribution parameters of EIFS - λ and ζ - are considered as random variables. Their prior distributions can be estimated using a re-sampling technique. For the purpose of illustration, it is assumed that λ and ζ are both normal random variables.

Model uncertainty and errors The parameters of the crack growth law are estimated by fitting the model using the experimental data of aluminum 7075-T6. Because it is difficult to determine simultaneously the statistics of all the parameters, n $(= 3.9)$ and p $(= 0.75)$ are treated as deterministic quantities [Liu and Mahadevan, 2009]. The parameter C is assumed to have a lognormal distribution (since $C > 0$) and the corresponding statistics are estimated. The uncertainty in the crack growth model (ε_{cg}) is treated as a lognormal random variable. The uncertainty due to the surrogate model is quantified using the expected output and the corresponding variance. For each FEA solution, three different meshes were considered and the discretization error was quantified as explained in Section 7.3.2.1. The discretization errors are added to the finite element solutions and the Gaussian process model is trained to predict the stress intensity factor. The uncertainty due to the selection of ARIMA loading model will be considered in the next subsection.

Probabilistic sensitivity analysis (global) is performed to evaluate the significance of the uncertainty sources quantified in this example. It is found that the variations in the crack growth simulation are mainly contributed from three sources - the parameter of the modified Paris law C, EIFS, and loading. The corresponding sensitivity indices are shown in Figs. 7.4(a)-(c). The sensitivity indices of C and loading increase over time, while the sensitivity indices of EIFS decrease over time. This is because C and loading are input of the crack growth model in every cycle while EIFS is only used at the initial time point.

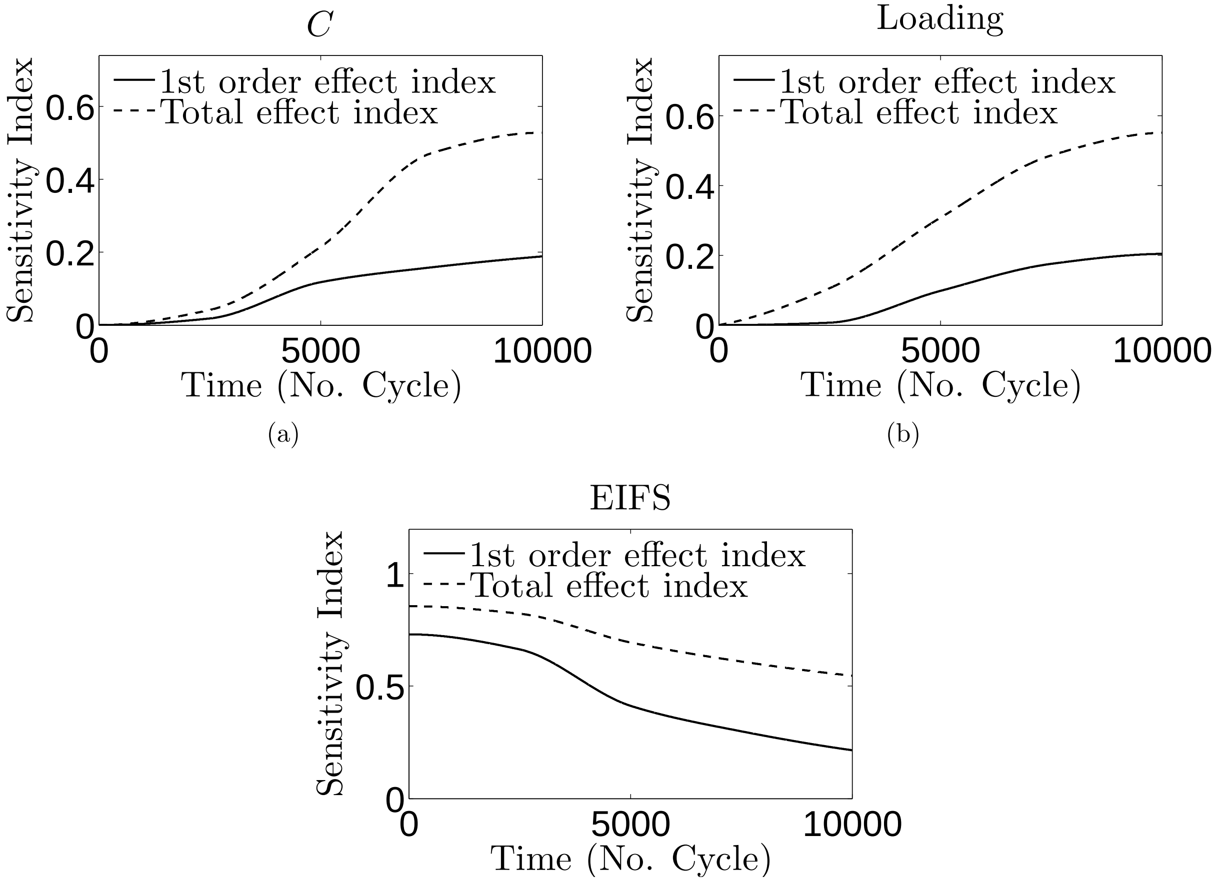
Table 7.2: Uncertainty quantification and associated statistics

Uncertainty classification	Unceratinty source	Random variable	Distribution type	Mean	Standard deviation
Physical variability	Material properties	ΔK_{th} (MPam$^{0.5}$)	Lognormal	5.66	0.268
		$\Delta \sigma_f$ (MPa)	Lognormal	201	20.1
	Loading ARIMA(2,1,0)	ε_t	Normal	0	0.1
Data uncertainty	POD	α	Constant	-2.1	0
		β	Constant	8.3E3	0
	Crack size measurement	β_0	Constant	0	0
		β_1	Constant	0	0
		ε_m (m)	Normal	0	0.04E-3
	Prior EIFS	λ	Normal	-7.60	0.1
		ζ	Normal	0.22	0.05
	Loading ARIMA(2,1,0)	φ_0	Uniform	0	0.05
		φ_1	Uniform	0.67	0.1
		φ_2	Uniform	0.3	0.05
Model uncertainty	Crack growth model	C (m/cycle)	Lognormal	6.54E-13	4.01E-13
		ε_{cg}	Lognormal	1	0.1

7.5.3 Usage of structural health monitoring data

It is assumed that 20 components in the fleet are monitored. Real-time load monitoring is applied to components No. 1, 2, 3, 4, 10, 11, 12, 13, 15, and 16. Crack detections are performed for all components after 10,000 cycles of operation, and no crack is detected on components No. 2 and 6 (crack is detected on all the other components). Among the components with crack detected, crack size measurements are reported on components No. 1, 3, 4, 5, 7, 8, 9, 10, and 14.

The load monitoring data (maximum bending moment in loading cycles, load ratio R=0) are generated from an ARIMA(2,1,0) model shown in Table 7.2. The crack inspection data are generated by crack propagation from an EIFS distribution (this distribution is assumed as the ”actual” distribution of EIFS for this fleet of components). The various scenarios of crack detection are sampled based on POD. The measured crack sizes are generated by



(a)

(b)

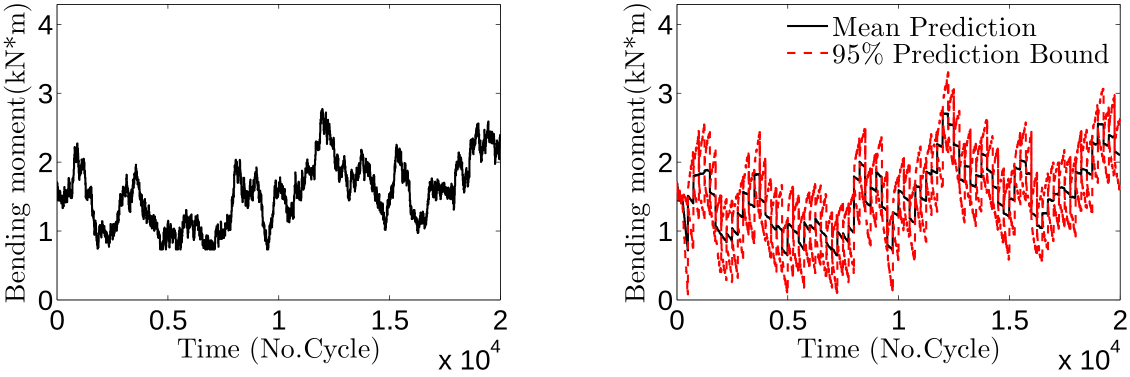
(c)

Figure 7.4: Sensitivity indices

adding sampled measurement noises to the actual crack sizes after 10,000 load cycles as shown in Eq. 7.4.

7.5.3.1 Characterization and prediction of loading sequence

For a component with real-time load monitoring, the ARIMA models used to characterize and predict the loading sequence are updated using the load monitoring data of that component, as discussed in Section 6.2.3. For example, the load monitoring data of Component No. 10 are shown in Fig. 7.5(a), and it is assumed that these data are collected every 250 cycles. By comparing the patterns of ACF and PACF of the data and ARIMA models and calculating the Q statistics, two possible ARIMA models are chosen as ARIMA(1,1,0) and ARIMA(2,1,0). The coefficients and weights of these two models are updated continuously using the Bayesian approach presented in Section 6.3.2. The updating of the probabilistic model weights are shown in Fig. 7.6(a). It can be observed that the weight of the ARIMA(2,1,0) model increases to nearly one in a short time, implying that the ARIMA(2,1,0) model fits the data better. This implication is valid since the data was generated from an ARIMA(2,1,0) model. The prior and updated PDFs of the coefficient φ_1 of the ARIMA(2,1,0) model are shown in Fig. 7.6(b) as an example, and it is shown that the uncertainty in the coefficient shrinks after updating. Based on the estimated coefficients and model weights, the probability distribution of the loading amplitude at a future time point can be obtained using Eqs. 6.6 and 6.9, and the mean value and 95% probability bound of the predicted loading amplitude are shown in Fig. 7.5(b).

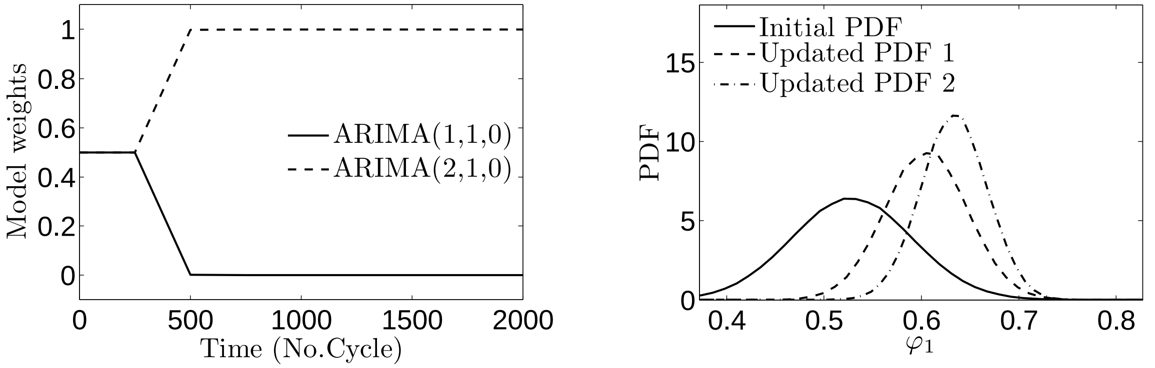


(a) (b)

Figure 7.5: (a) Real-time load monitoring data of component No. 10; (b) Prediction of future loading based on available data

7.5.3.2 Inference of EIFS

Following the Bayesian method presented in Section 7.4.1, the likelihood function of EIFS is constructed incorporating the three types of inspection results (crack is not detected,



(a) (b)

Figure 7.6: Continuous updating of ARIMA model weights and ARIMA model coefficient φ_1

crack is detected but not measured, and crack is detected and measured). The prior and updated distributions of EIFS are shown in Fig. 7.7. Recall that the crack inspection data are generated from an assumed EIFS distribution (Gaussian, with mean $\mu = 0.38$ and standard deviation $\sigma = 0.03$). It is observed that the updated distribution of EIFS using inspection data is close to the generating distribution.

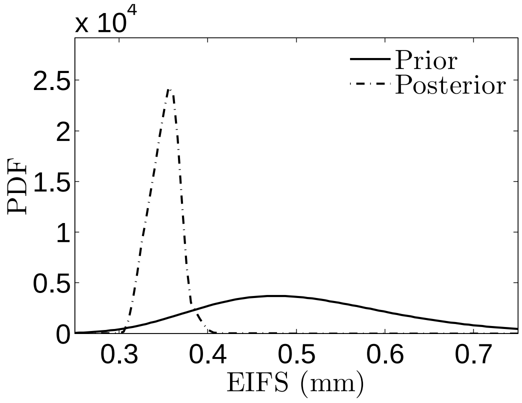


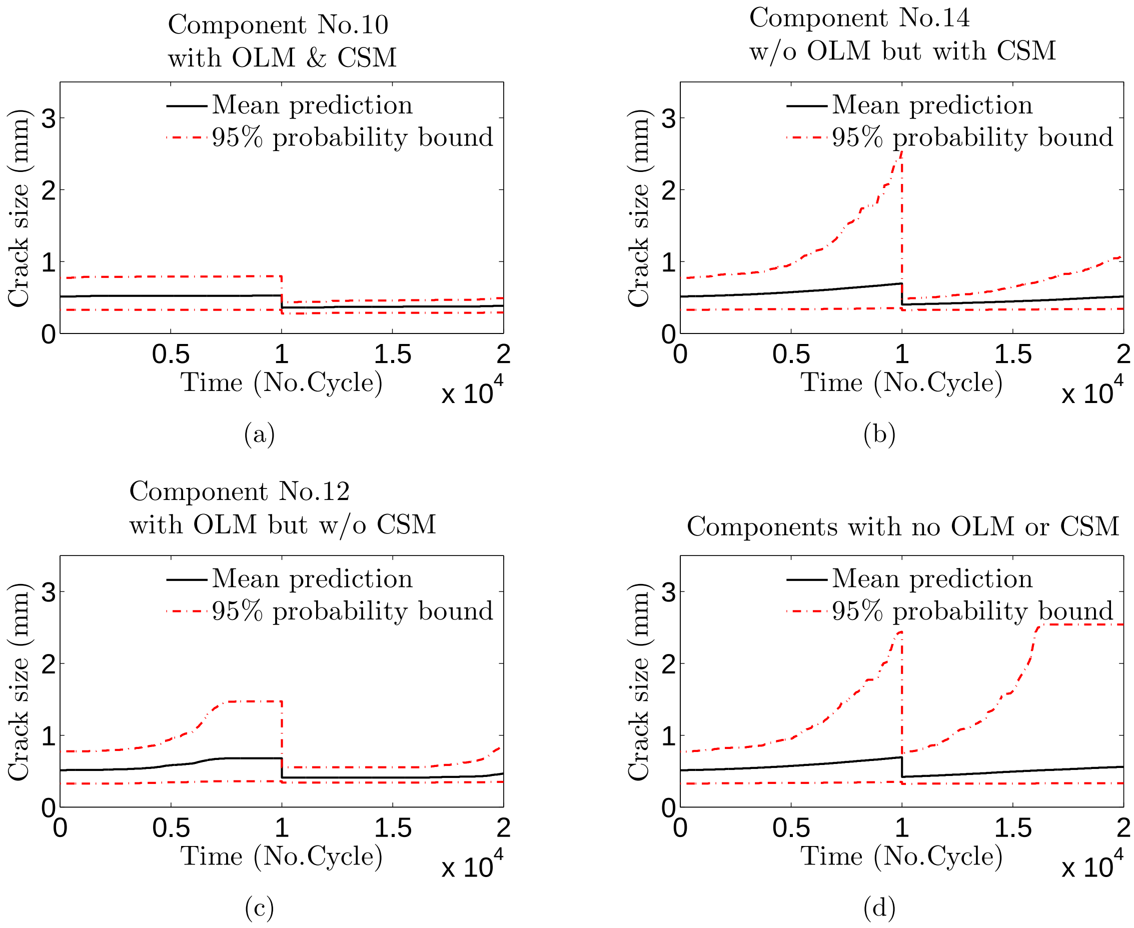
Figure 7.7: Prior and posterior PDF of EIFS

7.5.4 Prognosis for components in a fleet

Four different scenarios of prognosis are possible for the 20 inspected components described in Section 7.5.3, and other components in the fleet without any inspection, as discussed in Section 7.4.2. Component numbers 10, 14 and 12 are selected as examples of Scenarios 1, 2 and 3, respectively. The prognosis in Scenario 4 is for components without any crack inspection or load monitoring, and hence is not associated with any of the 20 inspected components. Prognostic results for all four scenarios are shown in Figs. 7.8(a)-(d). It can be observed from the results that the uncertainty in the prognosis of the four scenarios decreases significantly after inspection. Among the four scenarios, uncertainty in the prognosis of Scenario 1 is the smallest while the prognosis of Scenario 4 has the largest scatter, as expected. This indicates that the load monitoring and crack inspection for an individual component significantly reduce the uncertainty in the prognosis for this component.

7.5.5 Validation of prognosis results

Once new inspection data are available, the prognosis can be validated using Bayesian hypothesis testing as discussed in Section 7.4.3. In this example, it is assumed that 5 of the 20 components considered in Section 7.5.3 (component No. 10, 11, 12, 13, and 14) are detected again for crack after 5,000 cycles of operation, and the sizes of cracks on component No. 10 and 14 are measured. These new inspection data are used for validation. Since the crack sizes are measured for components 10 and 14, the prognoses in the Scenario 1 (for component No. 10) and the Scenario 2 (for component No. 14) are validated using the measured crack size of the corresponding components, while the prognoses in the Scenario 3 and 4 are validated using the inspection data of all the five components. The calculated Bayes factors and the confidence metrics are shown in Table 7.3. The prognoses are well supported by the data as the Bayes factors are all larger than 1, and the high values of confidence metric ($B/B + 1$) indicates that we have high confidence in the prognosis.



(a) (b)

(c) (d)

Figure 7.8: Results for the four scenarios of prognosis

Table 7.3: Bayes factor and confidence assessment

Scenario	Component #	Bayes factor	Confidence metric
1	No. 10	15.1	0.94
2	No. 14	7.30	0.88
3	No. 12	45.7	0.98
4	-	41.3	0.98

7.6 Conclusion

This chapter developed a probabilistic framework to integrate fatigue damage prognosis (FDP) of mechanical components with structural health monitoring (SHM) data, accounting for various sources of uncertainty and error. Efficient crack growth analysis for components with realistic geometry and multi-axial variable amplitude loading was presented using a GP surrogate model. The concept of EIFS was used to bypass micro-crack growth analysis. The uncertainty and errors related to the prognosis were classified into three categories: physical variability, data uncertainty and model uncertainty/errors, and these uncertainty sources are connected through a Bayesian network. A global sensitivity analysis was performed to evaluate the significance of these uncertainty sources.

The real-time load monitoring data were used to characterize and predict loading sequence using the ARIMA model-based approach. The crack inspection data were used to update the distribution of EIFS used in crack growth calculation. According to the availability of the real-time load monitoring data and on-ground crack inspection data, four different scenarios of prognosis were considered for a fleet of components, and appropriate strategies were developed for each scenario. The prognosis results were validated using a Bayesian hypothesis testing method.

CHAPTER VIII

SUMMARY AND FUTURE WORK

The overall goal of this research is to develop a rigorous probabilistic framework to quantify the uncertainty in reliability prediction for practical engineering applications. This target is approached by carrying out studies regarding the state-of-the-art uncertainty quantification and propagation techniques, with Bayesian network as the platform that integrates various sources of information. The accomplishment of this dissertation and some potential future work are outlined below.

Bayesian model calibration In Chapter III, methods were developed to address three challenging issues regarding Bayesian model calibration: (1) formulation and selection of model discrepancy to account for model form uncertainty, (2) detection of model parameter identifiability, and (3) development of calibration strategy for multi-physics system.

First, Bayesian calibration with various prior formulations of the model discrepancy function (δ) was studied, and numerical examples suggest that different choices of formulation can result in significantly different calibration results. If the formulation of δ is over-simplified, the estimation of physical parameters may be biased. However, if the formulation of δ is complicated but fails to cover all the missing physics, the resulting calibrated model may be outperformed by models calibrated using simpler formulation of δ outside the calibration domain. In order to help the analyst formulate and select the model discrepancy function, a three-step approach was developed using a reliability-based model validation metric and the total probability theorem. This approach provides an assessment of calibrated models using different options of δ, and it is capable of combining the posterior probability distributions of model parameters and δ resulting from the various options of δ into a single distribution, which is useful especially when the various options of δ have similar performances.

Second, a first-order Taylor series expansion-based method was developed to detect structural non-identifiability for models without analytical expressions, and detect practical non-identifiability due to insufficient amount of data. This method does not require the computation of likelihood function, and thus is simpler to implement and less computationally demanding. As pointed out in Section 3.3, the limitations of this method include: (1) it uses a linear approximation of the model, and hence may fail to detect non-identifiability if the model is highly nonlinear; (2) it can only detect local non-identifiability as the Taylor series expansion is constructed based on the derivatives at a single point; (3) it does not apply to statistical models; and (4) it does not cover practical non-identifiability due to the *quality* of data.

Third, a Bayesian network-based approach was proposed in order to develop insights into the calibration of a multi-physics system. We showed by using a Bayesian network that multiple options of calibration exist when various physics models are connected via common parameters, and more efficient calibration may be achieved with sequential calibration strategies. Preliminary study has been perform to the calibration of two physics-based models using data collected from two types of MEMS devices, as showed in Section 5.3.

Future research efforts may include: (1) examining more options of model discrepancy, and investigating the extrapolation capability of model discrepancy to higher level of system hierarchy, (2) extending the first-order Taylor series expansion-based method to models containing hyperparameters (e.g., when the model discrepancy is approximated using a Gaussian process, the coefficients of the covariance function need to be taken into account), and (3) searching for an optimal calibration strategy for systems with more complicated structures (e.g., multi-scale, multi-level).

Quantitative model validation Chapter IV was devoted to four issues existing in quantitative model validation methods: (1) validation with fully characterized, partially characterized, or uncharacterized experimental data; (2) validation of constant vs. stochastic model predictions; (3) accounting for the existence of directional bias; and (4) interpretation and selection of thresholds in different validation metrics.

The above issues were examined with various quantitative validation methods, including classical hypothesis testing, Bayesian hypothesis testing, a reliability-based method, and an area metric-based method.

The applicability of the various validation methods to the cases when validation data are collected from fully characterized, partially characterized, or uncharacterized experiments were discussed. Two hypotheses were developed for Bayesian hypothesis testing for validation with uncharacterized experimental data: (1) interval hypothesis on the distribution parameters of model prediction, which accesses the accuracy of the predicted mean and standard deviation from a model, taking into account the existence of directional bias, and (2) equality hypothesis on the entire PDF of model prediction. We showed that these two formulations of Bayesian hypothesis testing can be used in the case when multiple validation points are available. We also showed that under some conditions, the p-value in the z-test or t-test can be mathematically related to the Bayes factor and the reliability-based metric.

It was observed that the area metric in the transformed probability space (u-space) is sensitive to the direction of bias between model predictions and experimental data, and so are the Bayesian interval hypothsis testing-based method and the reliability-based method. Thus, these three methods can be used to detect the existence of the directional bias.

Future research efforts may include (1) investigating the case when the available validation data is a mixture of fully characterized, partially characterized, and uncharacterized, and (2) incorporation of the Bayesian model validation result and reliability-based metric in long-term failure and reliability analysis of engineering systems, which explicitly accounts for model uncertainty [Sankararaman, 2012].

Multi-physics system reliability prediction In Chapter V, the procedure of integrating various sources of uncertainty and information across multiple levels and physics based on

Bayesian networks was systematically demonstrated. Model calibration and model validation techniques developed in previous chapters were applied to quantify the uncertainty at different levels of the system, including uncertainty in model parameters and model form, and uncertainty in the predictive capability of the calibrated models. Efficient surrogate models were constructed to replace expensive physics-based dynamic simulations. Considering dielectric charging as the failure mechanism, the reliability of the target MEMS device was computed with the pull-in and pull-out voltage as the performance metrics.

Future work may include: (1) considering system reliability applications with multiple failure mechanisms, (2) exploring the inverse reliability problems for the multi-physics system, including resource allocation and optimal design, (3) performing rigorous surrogate model verification in order to justify the use of the surrogate models in reliability prediction, and (4) quantifying the uncertainty due to extrapolation from the calibration and validation domain to prediction domain.

Quantifying uncertainty in time-varying input using monitoring data Chapter VI investigated three different methods, namely rainflow counting, Markov chain method, and ARIMA model method, to characterize and reconstruct the fatigue load history using real-time monitoring data. The ARIMA method was extended through random parameters and probabilistic weights to accommodate the inherent variability in loading, the uncertainty due to sparse data, and the uncertainty in model selection. A continuous model updating approach with real-time monitoring data was developed, including direct updating of the characteristic matrices for the rainflow counting method and the Markov chain method, and a Bayesian updating approach for the ARIMA model method. The relation between prediction accuracy and updating interval was investigated quantitatively. We showed that the continuous updating approach can help the ARIMA model method to capture the time-variant feature of fatigue loading and also reduce the uncertainty in prediction due to limited data. A Bayesian hypothesis testing-based method was used to assess the predictive confidence of the three methods discussed. It is shown in the numerical example that all the three methods with the continuous updating approach perform comparably well, as the Bayes factor values are larger than unity at most of the time points. The overall confidence metric suggests that the prediction of the ARIMA method has the best support from load amplitude data in the numerical example presented.

The methods developed for fatigue loading prediction in this dissertation are particularly suitable for a single component under continuous monitoring, i.e., the load model is updated based on observed load history data. The future work of interest is to generate fleet-level loading predictions based on monitoring of several components under service, incorporating information in multiple load histories from these 1components.

Inclusion of health monitoring data in reliability analysis Chapter VII developed a probabilistic framework to integrate failure probability prediction of mechanical components under fatigue load with structural health monitoring (SHM) data. The framework facilitates a more informative and robust estimation of remaining useful life for mechanical components. Two features - the use of a surrogate model to significantly reduce the computational effort, and the quantification of uncertainty and confidence in prognosis - make it possible to explore online health assessment and decision making under uncertainty during a mission, regarding flight maneuvers based on the current state of health of the aircraft.

The proposed framework was illustrated with linear elastic fracture mechanics-based crack growth prognosis, assuming planar crack growth. Various sources of uncertainty and error, including natural variability, data uncertainty and model uncertainty/errors, were considered in the fatigue prognosis application. These uncertainty sources were connected through a Bayesian network, and the Sobol index-based global sensitivity analysis was performed to evaluate the significance of these uncertainty sources. A GP surrogate model was developed to expedite crack growth analysis for components with realistic geometry and multi-axial variable amplitude loading.

Two types of SHM data were considered: (1) real-time load monitoring data, which were used to characterize and predict loading sequence using the ARIMA model-based approach, and (2) on-ground crack inspection data, which were used to update the distribution of equivalent initial crack size. According to the availability of the real-time load monitoring data and on-ground crack inspection data, four different scenarios of prognosis were identified for a fleet of mechanical components, and appropriate strategies were developed for each scenario. The Bayesian hypothesis testing-based method was used to validate the prognosis results.

Future efforts may incorporate advanced material models and crack growth models that consider plastic deformation, and more realistic 3-D non-planar crack growth based on advanced surrogate modeling techniques developed in [Hombal et al., 2012; Hombal and Mahadevan, 2013b]. Future efforts may also investigate risk management decision-making based on the integration of diagnosis and prognosis.

Concluding remarks This dissertation has focused on the quantification and inclusion of model uncertainty in time-dependent reliability analysis, while also accounting for natural variability and data uncertainty. Although the state-of-the-art UQ methods often work well with simple mathematical problems, many interesting yet challenging issues have been identified in more realistic and complicated applications, as summarized in this chapter. The biggest challenge in a practical application may be the lack of experimental data at the device/system level, which are needed in the UQ framework proposed in this dissertation to assess the quality of the reliability prediction. Development of UQ methods to quantify the confidence in system-level prediction via information fusion from multiple sources in heterogeneous formats and multiple levels of fidelity will be of high value to the decision makers.

BIBLIOGRAPHY

AIAA (1998). *AIAA Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*. American Institute of Aeronautics and Astronautics, AIAA-G-077-1998, Reston, VA.

Ainsworth, M. and Oden, J. T. (1997). A posteriori error estimation in finite element analysis. *Computer Methods in Applied Mechanics and Engineering*, 142(1-2):1--88.

Alexeenko, A., Chigullapalli, S., Zeng, J., Guo, X., Kovacs, A., and Peroulis, D. (2011). Uncertainty in microscale gas damping: Implications on dynamics of capacitive MEMS switches. *Reliability Engineering and System Safety*, 96(9):1171--1183.

Ambaum, M. (2010). Significance tests in climate science. *Journal of Climate*, 23(22):5927--5932.

Amzallag, C., Gerey, J., Robert, J., and Bahuaud, J. (1994). Standardization of the rainflow counting method for fatigue analysis. *International Journal of Fatigue*, 16(4):287--293.

Andrieu-Renaud, C., Sudret, B., and Lemaire, M. (2004). The PHI2 method: a way to compute time-variant reliability. *Reliability Engineering & System Safety*, 84(1):75--86.

Angus, J. (1994). The probability integral transform and related results. *SIAM review*, 36(4):652--654.

Anthes, R. (1997). Modified rainflow counting keeping the load sequence. *International Journal of Fatigue*, 19(7):529--535.

Arendt, P. D., Apley, D. W., and Chen, W. (2012). Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability. *Journal of Mechanical Design*, 134(10):100908.

Arendt, P. D., Chen, W., and Apley, D. W. (2010). Updating Predictive Models: Calibration, Bias Correction and Identifiability. In *Proceedings of the ASME 2010 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2010)*, pages 1089--1098, Montreal, Quebec, Canada. ASME.

Arendt, P. D., Chen, W., Apley, D. W., and Zhu, P. (2011). Multiple Responses and Design of Experiments for Improving Identifiability in Model Calibration. In *9th World Congress on Structural and Multidisciplinary Optimization*, Shizuoka, Japan.

Arhonditsis, G., Papantou, D., Zhang, W., Perhar, G., Massos, E., and Shi, M. (2008). Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. *Journal of Marine Systems*, 73(1-2):8--30.

ASME (2006). *Guide for Verification and Validation in Computational Solid Mechanics*. American Society of Mechanical Engineers, ASME Standard V&V 10-2006, New York, NY.

ASTM (2005). E1049 - 85, Standard Practices for Cycle Counting in Fatigue Analysis.

Augarde, C. E. and Deeks, A. J. (2008). The use of Timoshenko's exact solution for a cantilever beam in adaptive analysis. *Finite Elements in Analysis and Design*, 44(9-10):595--601.

Ayyaswamy, V. and Alexeenko, A. (2010). Coarse-grained Model for RF MEMS Device.

Babuska, I. and Oden, J. T. (2004). Verification and validation in computational engineering and science: basic concepts. *Computer Methods in Applied Mechanics and Engineering*, 193(36-38):4057--4066.

Benasciutti, D. and Tovo, R. (2005). Cycle distribution and fatigue damage assessment in broad-band non-Gaussian random processes. *Probabilistic Engineering Mechanics*, 20(2):115--127.

Benasciutti, D. and Tovo, R. (2007). On fatigue damage assessment in bimodal random processes. *International Journal of Fatigue*, 29(2):232--244.

Bensi, M. T. (2010). *A Bayesian Network Methodology for Infrastructure Seismic Risk Assessment and Decision Support*. PhD thesis, University of California, Berkeley,.

Beran, P. S., Pettit, C. L., and Millman, D. R. (2006). Uncertainty quantification of limit-cycle oscillations. *Journal of Computational Physics*, 217(1):217--247.

Berens, A. P. and Hovey, P. (1983). Statistical Methods for Estimating Crack Detection Probabilities. In *Probabilistic Fracture Mechanics and Fatigue Methods: Applications for Structural Design and Maintenance*, pages 79--94.

Berry, R. D., Najm, H. N., Debusschere, B. J., Marzouk, Y. M., and Adalsteinsson, H. (2012). Data-free inference of the joint distribution of uncertain model parameters. *Journal of Computational Physics*, 231(5):2180--2198.

Bichon, B., McFarland, J., and Mahadevan, S. (2011). Efficient surrogate models for reliability analysis of systems with multiple failure modes. *Reliability engineering & systems safety*, 96(10):1386--1395.

Bidkar, R., Tung, R., Alexeenko, A., Sumali, H., and Raman, A. (2009). Unified theory of gas damping of flexible microcantilevers at low ambient pressures. *Applied Physics Letters*, 94(16):163117.

Blatman, G. (2009). *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. PhD thesis.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., and Mugunthan, P. (2008). Bayesian Calibration and Uncertainty Analysis for Computationally Expensive Models Using Optimization and Radial Basis Function Approximation. *Journal of Computational and Graphical Statistics*, 17(2):270--294.

Bower, R. G., Vernon, I., Goldstein, M., Benson, a. J., Lacey, C. G., Baugh, C. M., Cole, S., and Frenk, C. S. (2010). The parameter space of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, 407(4):2017--2045.

Box, G., Jenkins, G., and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 3 edition.

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211--252.

Brynjarsdottir, J. and O'Hagan, A. (2013). Learning about physical parameters : The importance of model discrepancy. *Submitted to SIAM/ASA Journal of Uncertainty Quantification.*

Campbell, K. (2006). Statistical calibration of computer simulations. *Reliability Engineering & System Safety*, 91(10):1358--1363.

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167--174.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *American Statistician*, 49(4):327--335.

Coble, R. L. (1963). A Model for Boundary Diffusion Controlled Creep in Polycrystalline Materials. *Journal of Applied Physics*, 34(6):1679--1682.

Cross, R., Makeev, A., and Armanios, E. (2007). Simultaneous uncertainty quantification of fracture mechanics based life prediction model parameters. *International Journal of Fatigue*, 29(8):1510--1515.

Das, S., Mathur, S. R., and Murthy, J. Y. (2012). Finite-Volume Method for Structural Analysis of RF MEMS Devices Using the Theory of Plates. *Numerical Heat Transfer, Part B: Fundamentals*, 61(1):1--21.

DeCarlo, E. C., Mahadevan, S., and Smarslok, B. P. (2013). Bayesian Calibration of Aerothermal Models for Hypersonic Air Vehicles. In *54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, Boston, MA.

Defense Modeling and Simulation Office (1996). *Verification, validation, and accreditation (VV&A) recommeded practices guide.* Office of the Director of Defense Research and Engineering, Alexandria, VA.

Deodatis, G. (1996). Non-stationary stochastic vector processes: seismic ground motion applications. *Probabilistic Engineering Mechanics*, 11(3):149--167.

Der Kiureghian, A. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105--112.

Dey, A. and Mahadevan, S. (2000). Reliability estimation with time-variant loads and resistances. *Journal of Structural Engineering*, 126(5):612--620.

Ditlevsen, O. and Madsen, H. O. (1996). Introduction to process descriptions. In *Structural Reliability Methods*, chapter 15. John Wiley & Sons, 1 edition.

Doebling, S. and Hemez, F. (2001). Overview of Uncertainty Assessment for Structural Health Monitoring. In *Proceedings of the 3rd Intl Workshop on Structural Health Monitoring*, volume 6950, Stanford University, Stanford, California.

Donahue, R., Clark, H., Atanmo, P., Kumble, R., and McEvily, A. (1972). Crack opening displacement and the rate of fatigue crack growth. *International Journal of Fracture Mechanics*, 8(2):209--219.

Dowling, N. (1972). Fatigue failure predictions for complicated stress-strain histories. *J Mater JMLSA*, 7(1):71--87.

Dressler, K., Hack, M., and Krüger, W. (1997). Stochastic Reconstruction of Loading Histories from a Rainflow Matrix. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 77(3):217--226.

El Haddad, M., Topper, T., and Smith, K. (1979). Prediction of non propagating cracks. *Engineering Fracture Mechanics*, 11(3):573--584.

Faravelli, L. (1989). Response-surface approach for reliability analysis. *Journal of Engineering Mechanics*, 115(12):2763--2781.

Farrar, C. and Lieven, N. (2007). Damage prognosis: the future of structural health monitoring. *Philosophical Transactions of the Royal Society A*, 365(1851):623--32.

Ferson, S. and Oberkampf, W. (2009). Validation of imprecise probability models. *International Journal of Reliability and Safety*, 3(1):3--22.

Ferson, S., Oberkampf, W., and Ginzburg, L. (2008). Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29-32):2408--2430.

Finkel, D. and Kelley, C. (2004). Convergence analysis of the DIRECT algorithm. *Optimization Online*, pages 1--10.

Forman, R., Kearney, V., and Engle, R. (1967). Numerical analysis of crack propagation in cyclic-loaded structures. *Journal of Basic Engineering*, 89:459--464.

Gad-el Hak, M., editor (2005). *MEMS: Introduction and Fundamentals*. CRC Press, 2 edition.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient metropolis jumping rules. *Bayesian statistics*, 5:599--607.

Gere, J. M. and Goodno, B. J. (2009). *Mechanics of Materials*. Cengage Learning, 7 edition.

Gerstner, T. and Griebel, M. (1998). Numerical integration using sparse grids. *Numerical algorithms*, 18(3-4):209--232.

Ghanem, R. and Spanos, P. (2003). *Stochastic Finite Elements: A Spectral Approach*. Dover Pubns.

Gomes, H. M. and Awruch, A. M. (2004). Comparison of response surface and neural network with other methods for structural reliability analysis. *Structural Safety*, 26(1):49--67.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711--732.

Grewal, M. and Glover, K. (1976). Identifiability of linear and nonlinear dynamical systems. *Automatic Control, IEEE Transactions on*, 21(6):833--837.

Gu, M. and Lu, J. (1994). A note on identifiability of the regression parameter and validity of the partial likehood approach in general relativerisk regression. *Biometrika*, 81(4):802--806.

Guan, X., Jha, R., and Liu, Y. (2009). Probabilistic fatigue damage prognosis using maximum entropy approach. *Journal of Intelligent Manufacturing*, 23(2):163--171.

Guo, X. and Alexeenko, A. (2009). Compact model of squeeze-film damping based on rarefied flow simulations. *Journal of Micromechanics and Microengineering*, 19(4):045026.

Guo, X., Li, J., Xiu, D., and Alexeenko, A. (2010). Uncertainty quantification models for micro-scale squeeze-film damping. *International Journal for Numerical Methods in Engineering*, 84(10):1257--1272.

Gupta, S. and Ray, A. (2007). Real-time fatigue life estimation in mechanical structures. *Measurement Science and Technology*, 18(7):1947--1957.

Haarhoff, L. J., Kok, S., and Wilke, D. N. (2013). Numerical Strategies to Reduce the Effect of Ill-Conditioned Correlation Matrices and Underflow Errors in Kriging. *Journal of Mechanical Design*, 135(4):044502.

Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339--354.

Haldar, A. and Mahadevan, S. (2000). *Probability, reliability, and statistical methods in engineering design*. Wiley, New York.

Hanke, J. and Wichern, D. (2005). *Business Forecasting*. Pearson/Prentice Hall, Upper Saddle River, NJ, 8 edition.

Hartmann, C., Smeyers-Verbeke, J., Penninckx, W., Vander Heyden, Y., Vankeerberghen, P., and Massart, D. (1995). Reappraisal of hypothesis testing for method validation: detection of systematic error by comparing the means of two methods or of two laboratories. *Analytical Chemistry*, 67(24):4491--4499.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97--109.

Heasler, P., Taylor, T., and Doctor, S. (1993). Statistically based reevaluation of PISC-II round robin test data. Technical report, Nuclear Regulatory Commission, Richland, WA.

Heuler, P. and Klatschke, H. (2005). Generation and use of standardised load spectra and loadtime histories. *International Journal of Fatigue*, 27(8):974--990.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer Model Calibration Using High-Dimensional Output. *Journal of the American Statistical Association*, 103(482):570--583.

Hills, R. G. and Leslie, I. H. (2003). Statistical validation of engineering and scientific models: validation experiments to application. *Sandia technical report*, (SAND2003-0706).

Hills, R. G. and Trucano, T. G. (1999). Statistical Validation of Engineering and Scientific Models : Background. *Sandia technical report*, (SAND99-1256).

Hills, R. G. and Trucano, T. G. (2002). Statistical Validation of Engineering and Scientific Models : A Maximum Likelihood Based Metric. *Sandia technical report*, (SAND2001-1783).

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical science*, 14(4):382--401.

Hombal, V. K., Ling, Y., Wolfe, K., and Mahadevan, S. (2012). Two-stage planar approximation of non-planar crack growth. *Engineering Fracture Mechanics*, 96:147--164.

Hombal, V. K. and Mahadevan, S. (2011). Bias Minimization in Gaussian Process Surrogate Modeling for Uncertainty Quantification. *International Journal for Uncertainty Quantification*, 1(4):321--349.

Hombal, V. K. and Mahadevan, S. (2013a). Model Selection Among Physics-Based Models. *Journal of Mechanical Design*, 135(2):021003.

Hombal, V. K. and Mahadevan, S. (2013b). Surrogate modeling of 3D crack growth. *International Journal of Fatigue*, 47:90--99.

Hsu, H., Koslowski, M., and Peroulis, D. (2011). An Experimental and Theoretical Investigation of Creep in Ultrafine Crystalline Nickel RF-MEMS Devices. *Microwave Theory and Techniques, IEEE Transactions on*, 59(10):2655--2664.

Huang, S. P., Mahadevan, S., and Rebba, R. (2007). Collocation-based stochastic finite element analysis for random field problems. *Probabilistic Engineering Mechanics*, 22(2):194--205.

Huang, S. P., Quek, S. T., and Phoon, K. K. (2001). Convergence study of the truncated KarhunenLoeve expansion for simulation of stochastic processes. *International Journal for Numerical Methods in Engineering*, 52(9):1029--1043.

Hurtado, J. E. (2004). An examination of methods for approximating implicit limit state functions from the viewpoint of statistical learning theory. *Structural Safety*, 26(3):271--293.

Hurtado, J. E. and Alvarez, D. a. (2001). Neural-network-based reliability analysis: a comparative study. *Computer Methods in Applied Mechanics and Engineering*, 191(1-2):113--132.

Jain, A., Palit, S., and Alam, M. A. (2011). A Physics-Based Predictive Modeling Framework for Dielectric Charging and Creep in RF MEMS Capacitive Switches and Varactors. *Microelectromechanical Systems, Journal of*, (99):1--11.

Jeffreys, H. (1983). *Theory of probability.* Oxford University Press, USA, London, 3 edition.

Jensen, F. and Nielsen, T. (2007). *Bayesian networks and decision graphs (Information Science and Statistics).* Springer, 2 edition.

Jia-fan, Z., Qing-hua, Y., and Tong, Z. (2011). Numerical Approach to Identifiability Test of Parametric Models in Nonlinear Mechanical Systems. *Journal of Dynamic Systems, Measurement, and Control*, 133(5):051002.

Jiang, X. and Mahadevan, S. (2006). Bayesian cross-entropy methodology for optimal design of validation experiments. *Measurement Science and Technology*, 17(7):1895.

Jiang, X. and Mahadevan, S. (2007). Bayesian risk-based decision method for model validation under uncertainty. *Reliability Engineering & System Safety*, 92(6):707--718.

Johnson, W. S. (2010). The history, logic and uses of the Equivalent Initial Flaw Size approach to total fatigue life prediction. *Procedia Engineering*, 2(1):47--58.

Karlin, S. (1966). *A First Course in Stochastic Processes.* Academic Press, New York, NY, 1 edition.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773--795.

Kay, S. M. (1998). *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory.* Prentice Hall, 1 edition.

Kaymaz, I. (2005). Application of kriging method to structural reliability problems. *Structural Safety*, 27(2):133--151.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 63(3):425--464.

Khosrovaneh, A. and Dowling, N. (1990). Fatigue loading history reconstruction based on the rainflow technique. *International Journal of Fatigue*, 12(2):99--106.

Khosrovaneh, A. K., Dowling, N. E., Berens, A. P., and Gallagher, J. P. (1989). *Fatigue life estimates for helicopter loading spectra*. NASA contractor report 181941.

Kim, H., Venturini, G., and Strachan, A. (2012). Molecular dynamics study of dynamical contact between a nanoscale tip and substrate for atomic force microscopy experiments. *Journal of Applied Physics*, 112(9):094325.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671--680.

Kitagawa, H. and Takahashi, S. (1976). Applicability of fracture mechanics to very small cracks or the cracks in the early stage. In *Second International Conference on Mechanical Behavior of Materials. ASM, Metals Park, Ohio*, pages 627--631.

Knepell, P. and Arangno, D. (1993). *Simulation validation: a confidence assessment methodology*. Wiley-IEEE Computer Society Press, 1 edition.

Kong, J. S. and Frangopol, D. M. (2003). Life-Cycle Reliability-Based Maintenance Cost Optimization of Deteriorating Structures with Emphasis on Bridges. *Journal of Structural Engineering*, 129(6):818--828.

Koslowski, M. and Strachan, A. (2011). Uncertainty propagation in a multiscale model of nanocrystalline plasticity. *Reliability Engineering & System Safety*, 96(9):1161--1170.

Koutsourelakis, P. (2009). A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters. *Journal of Computational Physics*, 228(17):6184--6211.

Krenk, S. and Gluver, H. (1989). A Markov matrix for fatigue load simulation and rainflow range evaluation. *Structural Safety*, 6(2-4):247--258.

Kulkarni, S. and Achenbach, J. (2008). Structural Health Monitoring and Damage Prognosis in Fatigue. *Structural Health Monitoring*, 7(1):37--49.

Kuschel, N. and Rackwitz, R. (2000). Optimal design under time-variant reliability constraints. *Structural Safety*, 22(2):113--127.

Langseth, H. and Portinale, L. (2007). Bayesian networks in reliability. *Reliability Engineering & System Safety*, 92(1):92--108.

Lehmann, E. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, 3 edition.

Leser, C., Juneja, L., Thangjitham, S., and Dowling, N. (1998). On Multi-axial Random Fatigue Load Modeling. *SAE transactions*, 107:481--494.

Leser, C., Thangjitham, S., and Dowling, N. (1994). Modeling of random vehicle loading histories for fatigue analysis. *International Journal of Vehicle design*, 15:467--483.

Li, H.-s., Lü, Z.-z., and Yue, Z.-f. (2006). Support vector machine for structural reliability analysis. *Applied Mathematics and Mechanics*, 27(10):1295--1303.

Liang, B. and Mahadevan, S. (2011). Error and uncertainty quantification and sensitivity analysis in mechanics computational models. *International Journal for Uncertainty Quantification*, 1(2):147--161.

Ling, Y. and Mahadevan, S. (2013a). Challenging issues in Bayesian calibration of multiphysics models. In *54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, Boston, MA.

Ling, Y. and Mahadevan, S. (2013b). Quantitative model validation techniques: New insights. *Reliability Engineering & System Safety*, 111:217--231.

Little, M., Heidenreich, W., and Li, G. (2010). Parameter identifiability and redundancy: theoretical considerations. *PloS one*, 5(1):e8915.

Liu, H., Chen, W., and Sudjianto, A. (2006). Relative Entropy Based Method for Probabilistic Sensitivity Analysis in Engineering Design. *Journal of Mechanical Design*, 128(2):326.

Liu, Y., Chen, W., and Arendt, P. (2011). Toward a Better Understanding of Model Validation Metrics. *Journal of Mechanical Design*, 133(7):071005.

Liu, Y. and Mahadevan, S. (2007). Threshold stress intensity factor and crack growth rate prediction under mixed-mode loading. *Engineering fracture mechanics*, 74(3):332--345.

Liu, Y. and Mahadevan, S. (2009). Probabilistic fatigue life prediction using an equivalent initial flaw size distribution. *International Journal of Fatigue*, 31(3):476--487.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297--303.

Lu, Z., Xiang, Y., and Liu, Y. (2010). Crack growth-based fatigue-life prediction using an equivalent initial flaw model. Part II: Multiaxial loading. *International Journal of Fatigue*, 32(2):376--381.

Lucas, L., Owhadi, H., and Ortiz, M. (2008). Rigorous verification, validation, uncertainty quantification and certification through concentration-of-measure inequalities. *Computer Methods in Applied Mechanics and Engineering*, 197(51-52):4591--4609.

Ma, X. and Zabaras, N. (2009). An efficient Bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method. *Inverse Problems*, 25(3):035013 (27pp).

Mahadevan, S. and Dey, A. (1997). Adaptive Monte Carlo simulation for time-variant reliability analysis of brittle structures. *AIAA journal*, 35(2):321--326.

Mahadevan, S., Zhang, R., and Smith, N. (2001). Bayesian networks for system reliability reassessment. *Structural Safety*, 23(3):231--251.

Maître, O. L., Najm, H., Pébay, P., Ghanem, R., and Knio, O. (2007). Multi-resolution-analysis scheme for uncertainty quantification in chemical systems. *SIAM Journal on Scientific Computing*, 29(2):864--889.

Makeev, A., Nikishkov, Y., and Armanios, E. (2007). A concept for quantifying equivalent initial flaw size distribution in fracture mechanics based life prediction models. *International Journal of Fatigue*, 29(1):141--145.

Marden, J. (2000). Hypothesis testing: from p values to Bayes factors. *Journal of the American Statistical Association*, 95(452):1316--1320.

Marzouk, Y. M. and Najm, H. N. (2009). Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862--1902.

McFarland, J. and Mahadevan, S. (2008). Multivariate significance testing and model calibration under uncertainty. *Computer Methods in Applied Mechanics and Engineering*, 197(29-32):2467--2479.

McFarland, J. M. (2008). *Uncertainty analysis for computer simulations through validation and calibration*. PhD thesis, Vanderbilt University.

Melchers, R. (1999). *Structural Reliability Analysis and Prediction (Civil Engineering)*. John Wiley, 2 edition.

Miner, M. A. (1945). Cumulative damage in fatigue. *Journal of Applied Mechanics*, 12:A159--A164.

Moreno, B., Zapatero, J., and Dominguez, J. (2003). An experimental analysis of fatigue crack growth under random loading. *International Journal of Fatigue*, 25(7):597--608.

Mullins, J., Li, C., Sankararaman, S., Mahadevan, S., and Urbina, A. (2013). Probabilistic integration of validation and calibration results for prediction level uncertainty quantification: Application to structural dynamics. In *54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Boston, MA.

Murphy, S. and Van der Vaart, A. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449--465.

Najm, H. N. (2009). Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics. *Annual Review of Fluid Mechanics*, 41(1):35--52.

NASA and Southwest Research Institute (2010). NASGRO Crack Growth Equation.

Nataf, A. (1962). Détermination des distributions de probabilité dont les marges sont données. *Comptes Rendus de lAcadémie des Sciences*, 225:42--43.

Neal, R. (2003). Slice sampling. *The annals of statistics*, 31(3):705--741.

Niu, G., Yang, B.-S., and Pecht, M. (2010). Development of an optimized condition-based maintenance system by data fusion and reliability-centered maintenance. *Reliability Engineering & System Safety*, 95(7):786--796.

Oberkampf, W. and Barone, M. (2006). Measures of agreement between computation and experiment: Validation metrics. *Journal of Computational Physics*, 217(1):5--36.

Oberkampf, W. and Trucano, T. (2002). Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences*, 38(3):209--272.

Oberkampf, W. and Trucano, T. (2004). Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, 57(5):345--384.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):99--138.

O'Hagan, A. and Oakley, J. E. (2004). Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety*, 85(1-3):239--248.

Palit, S. and Alam, M. (2012). Theory of charging and charge transport in intermediate thickness dielectrics and its implications for characterization and reliability. *Journal of Applied Physics*, 111(5):054112.

Paris, P. and Erdogan, F. (1963). A critical analysis of crack propagation laws. *Journal Of Basic Engineering*, 85(4):528--534.

Park, I., Amarchinta, H. K., and Grandhi, R. V. (2010). A Bayesian approach for quantification of model uncertainty. *Reliability Engineering & System Safety*, 95(7):777--785.

Paulino, C. and de Bragança Pereira, C. (1994). On identifiability of parametric statistical models. *Statistical Methods and Applications*, 3(1):125--151.

Pericchi, L. R. (2005). Handbook of Statistics, Volume 25: Bayesian Thinking, Modeling and Computation. chapter 6, pages 115--149. North Holland, 1 edition.

Phoon, K., Huang, S., and Quek, S. (2002). Simulation of second-order processes using KarhunenLoeve expansion. *Computers & Structures*, 80(12):1049--1060.

Pierce, S., Worden, K., and Bezazi, A. (2008). Uncertainty analysis of a neural network used for fatigue lifetime prediction. *Mechanical Systems and Signal Processing*, 22(6):1395--1411.

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

Rackwitz, R. (2001). Reliability analysisa review and some perspectives. *Structural Safety*, 23(4):365--395.

Rangavajhala, S., Sura, V. S., Hombal, V. K., and Mahadevan, S. (2011). Discretization Error Estimation in Multidisciplinary Simulations. *AIAA Journal*, 49(12):2673--2683.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923--1929.

Raue, A., Kreutz, C., Maiwald, T., Klingmuller, U., and Timmer, J. (2011). Addressing parameter identifiability by model-based experimentation. *Systems Biology, IET*, 5(2):120--130.

Reagan, M. T., Najm, H. N., Ghanem, R. G., and Knio, O. M. (2003). Uncertainty quantification in reacting-flow simulations through non-intrusive spectral projection. *Combustion and Flame*, 132(3):545--555.

Rebba, R. (2005). *Model validation and design under uncertainty*. PhD thesis, Vanderbilt University.

Rebba, R. and Mahadevan, S. (2006). Validation of models with multivariate output. *Reliability Engineering & System Safety*, 91(8):861--871.

Rebba, R. and Mahadevan, S. (2008). Computational methods for model reliability assessment. *Reliability Engineering & System Safety*, 93(8):1197--1207.

Rebba, R., Mahadevan, S., and Huang, S. (2006). Validation and error estimation of computational models. *Reliability Engineering & System Safety*, 91(10-11):1390--1397.

Rebeiz, G. M. (2003). *RF MEMS: Theory, Design, and Technology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1 edition.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5):W05521.

Richardson, L. F. (1911). The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations, with an Application to the Stresses in a Masonry Dam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 210(459-470):307--357.

Riley, M. and Grandhi, R. (2011). Quantification of model-form and predictive uncertainty for multi-physics simulation. *Computers & Structures*, 89(11-12):1206--1213.

Rocco, C. and Moreno, J. (2002). Fast Monte Carlo reliability evaluation using support vector machine. *Reliability Engineering & System Safety*, 76(3):237--243.

Romero, V., Swiler, L., and Giunta, A. (2004). Construction of response surfaces based on progressive-lattice-sampling experimental designs with application to uncertainty propagation. *Structural Safety*, 26(2):201--219.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470--472.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225--237.

Roussouly, N., Petitjean, F., and Salaun, M. (2013). A new adaptive response surface method for reliability analysis. *Probabilistic Engineering Mechanics*, 32:103--115.

Roy, C. and Oberkampf, W. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25-28):2131--2144.

Rychlik, I. (1996). Simulation of load sequences from rainflow matrices: Markov method. *International Journal of Fatigue*, 18(7):429--438.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145:280--297.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*, volume 76. John Wiley & Sons, Inc.

Sankararaman, S. (2012). *Uncertainty Quantification and Integration in Engineering Systems*. PhD thesis, Vanderbilt University.

Sankararaman, S., Ling, Y., and Mahadevan, S. (2010). Statistical inference of equivalent initial flaw size with complicated structural geometry and multi-axial variable amplitude loading. *International Journal of Fatigue*, 32(10):1689--1700.

Sankararaman, S., Ling, Y., and Mahadevan, S. (2011a). Uncertainty quantification and model validation of fatigue crack growth prediction. *Engineering Fracture Mechanics*, 78(7):1487--1504.

Sankararaman, S., Ling, Y., Shantz, C., and Mahadevan, S. (2009). Uncertainty quantification in fatigue damage prognosis. In *Annual Conference of the Prognostics and Health Management Society*, San Diego, CA.

Sankararaman, S., Ling, Y., Shantz, C., and Mahadevan, S. (2011b). Inference of equivalent initial flaw size under multiple sources of uncertainty. *International Journal of Fatigue*, 33(2):75--89.

Sankararaman, S. and Mahadevan, S. (2011a). Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data. *Reliability Engineering & System Safety*, 96(7):814--824.

Sankararaman, S. and Mahadevan, S. (2011b). Model validation under epistemic uncertainty. *Reliability Engineering and System Safety*, 96(9):1232--1241.

Sankararaman, S. and Mahadevan, S. (2012a). Comprehensive framework for integration of calibration, verification and validation. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Honolulu, HI.

Sankararaman, S. and Mahadevan, S. (2012b). Distribution type uncertainty due to sparse and imprecise data. *Mechanical Systems and Signal Processing*, 37(1-2):182--198.

Sargent, R. (2005). Verification and validation of simulation models. *of the 37th conference on Winter simulation*, pages 130--143.

Sarkar, S., Kosson, D., Mahadevan, S., Meeussen, J., der Sloot, H. V., Arnold, J., and Brown, K. (2012). Bayesian calibration of thermodynamic parameters for geochemical speciation modeling of cementitious materials. *Cement and Concrete Research*, 42(7):889--902.

Saxena, A., Celaya, J., Saha, B., Saha, S., and Goebel, K. (2010). Metrics for Offline Evaluation of Prognostic Performance. *International journal of prognostics and health management*, (1):1--20.

Schervish, M. J. (1996). P Values: What They Are and What They Are Not. *The American Statistician*, 50(3):203--206.

Schijve, J. (2003). Fatigue of structures and materials in the 20th century and the state of the art. *International Journal of Fatigue*, 25(8):679--702.

Seber, G. and Wild, C. (2003). *Nonlinear regression (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Seghouane, A., Bekara, M., and Fleury, G. (2005). A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence. *Signal processing*, 85(7):1405--1417.

Shinozuka, M. and Deodatis, G. (1991). Simulation of stochastic processes by spectral representation. *Applied Mechanics Reviews*, 44(4):191--204.

Singh, A., Mourelatos, Z. P., and Li, J. (2010). Design for Lifecycle Cost Using Time-Dependent Reliability. *Journal of Mechanical Design*, 132(9):091008.

Smith, K., Thompson, B., Meeker, B., Gray, T., and Brasche, L. (2007). Model-Assisted Probability of Detection Validation for Immersion Ultrasonic Application. In *AIP Conference Proceedings*, pages 1816--1822.

Snow, M. and Bajaj, A. (2010). Uncertainty Quantification Study For A Comprehensive Electrostatic MEMS Switch Model. In *Third International Conference on Uncertainty in Structural Dynamics - USD2010*, Department of Mechanical Engineering, Katholieke Universiteit, Leuven, Belgium.

Spencer, F. and Schurman, D. (1995). Reliability Assessment at Airline Inspection Facilities, Volume III: Results of an Eddy Current Inspection Reliability Experiment. *DOT/FAA/CT-92/12*, III(May).

Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. Wiley-Interscience, 1 edition.

Staszewski, W., Boller, C., and Tomlinson, G., editors (2004). *Health Monitoring of Aerospace Structures: Smart Sensor Technologies and Signal Processing*. John Wiley & Sons, Inc., 1 edition.

Straub, D. and Der Kiureghian, A. (2010). Bayesian network enhanced with structural reliability methods: methodology. *Journal of Engineering Mechanics*, 136(10):1248--1258.

Sudret, B. and Der Kiureghian, A. (2002). Comparison of finite element reliability methods. *Probabilistic Engineering Mechanics*, 17(4):337--348.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents series)*. The MIT Press.

Timoshenko, S. and Goodier, J. (1970). *Theory of elasticity*. McGraw-Hill, New York, 3 edition.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211--244.

Torres-Toledano, J. and Sucar, L. (1998). Bayesian networks for reliability analysis of complex systems. *Progress in Artificial Intelligence - IBERAMIA 98*, pages 465--465.

Tovo, R. (2002). Cycle distribution and fatigue damage under broad-band random loading. *International Journal of Fatigue*, 24(11):1137--1147.

Urbina, A., Paez, T. L., Hasselman, T., Wathugala, W., and Yap, K. (2003). Assessment of model accuracy relative to stochastic system behavior. In *44 th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 10(5):988--99.

Walter, E. and Pronzato, L. (1996). On the identifiability and distinguishability of nonlinear parametric models. *Mathematics and Computers in Simulation*, 42(2):125--134.

Wang, Z. and Wang, P. (2012). A Nested Extreme Response Surface Approach for Time-Dependent Reliability-Based Design Optimization. *Journal of Mechanical Design*, 134(12):121007.

Weber, P. and Jouffe, L. (2006). Complex system reliability modelling with Dynamic Object Oriented Bayesian Networks (DOOBN). *Reliability Engineering & System Safety*, 91(2):149--162.

Weertman, J. (1966). Rate of growth of fatigue cracks calculated from the theory of infinitesimal dislocations distributed on a plane. *International Journal of Fracture*, 2(2):460--467.

Wei, L., Delosrios, E., and James, M. (2002). Experimental study and modelling of short fatigue crack growth in aluminium alloy Al7010-T7451 under random loading. *International Journal of Fatigue*, 24(9):963--975.

Wiener, N. (1938). The homogeneous chaos. *American Journal of Mathematics*, 60(4):897--936.

Wojtkiewicz, S., Eldred, M., Field, R., Urbina, A., and Red-Horse, J. (2001). Uncertainty quantification in large computational engineering models. In *42rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*.

Xiang, Y., Lu, Z., and Liu, Y. (2010). Crack growth-based fatigue life prediction using an equivalent initial flaw model. Part I: Uniaxial loading. *International Journal of Fatigue*, 32(2):341--349.

Xiong, J. and Shenoi, R. (2008). A load history generation approach for full-scale accelerated fatigue tests. *Engineering Fracture Mechanics*, 75(10):3226--3243.

Xiu, D. and Karniadakis, G. (2002). The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619--644.

Yuen, B. and Taheri, F. (2006). Proposed modifications to the Wheeler retardation model for multiple overloading fatigue life prediction. *International Journal of Fatigue*, 28(12):1803--1819.

Zabaras, N. and Ganapathysubramanian, B. (2008). A scalable framework for the solution of stochastic inverse problems using a sparse grid collocation approach. *Journal of Computational Physics*, 227(9):4697--4735.

Zaman, K., Rangavajhala, S., McDonald, M. P., and Mahadevan, S. (2010). A probabilistic approach for representation of interval uncertainty. *Reliability Engineering & System Safety*, 96(1):117--130.

Zapatero, J., Moreno, B., Gonzalezherrera, A., and Dominguez, J. (2005). Numerical and experimental analysis of fatigue crack growth under random loading. *International Journal of Fatigue*, 27(8):878--890.

Zhang, R. and Mahadevan, S. (2000). Model uncertainty and Bayesian updating in reliability-based inspection. *Structural Safety*, 22(2):145--160.

Zhang, R. and Mahadevan, S. (2001). Fatigue reliability analysis using nondestructive inspection. *Journal of Structural Engineering*, 127(8):957.

Zhang, R. and Mahadevan, S. (2003). Bayesian methodology for reliability model acceptance. *Reliability Engineering & System Safety*, 80(1):95--103.

Zhou, X., Xi, L., and Lee, J. (2007). Reliability-centered predictive maintenance scheduling for a continuously monitored system subject to degradation. *Reliability Engineering & System Safety*, 92(4):530--534.

Ziliak, S. and McCloskey, D. (2008). *The cult of statistical significance.* University of Michigan Press.

Zio, E. (2009). Reliability engineering: Old problems and new challenges. *Reliability Engineering & System Safety*, 94(2):125--141.

Zuev, K. M., Beck, J. L., Au, S.-K., and Katafygiotis, L. S. (2012). Bayesian post-processor and other enhancements of Subset Simulation for estimating failure probabilities in high dimensions. *Computers & Structures*, 92-93:283--296.

Zwillinger, D. (1995). *Handbook of Integration.* Jones and Bartlett Publishers, Boston, MA, 1 edition.