DEVELOPMENTS IN COMPUTER MODEL CALIBRATION

by

Matthew T. Pratola
B.Sc., Brock University, 2003
M.Sc., Simon Fraser University, 2006

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in the Department
of
Statistics and Actuarial Science

© Matthew T. Pratola 2010
SIMON FRASER UNIVERSITY
Summer 2010

All rights reserved. This work may not be
reproduced in whole or in part, by photocopy
or other means, without the permission of the author.

APPROVAL

Name: Matthew T. Pratola

Degree: Doctor of Philosophy

Title of Thesis: Developments in Computer Model Calibration

Examining Committee: Dr. Gary Parker
Chair

Dr. Derek Bingham, Senior Supervisor

Dr. Tim Swartz, Supervisor

Dr. Jiguo Cao, Supervisor

Dr. David Campbell, Supervisor

Dr. Brian Williams, External Examiner
Los Alamos National Laboratory, Los Alamos, NM

Date Approved: _____

Abstract

Computer models enable scientists to investigate real-world phenomena virtually using computer experiments. Recently, statistical calibration enabled scientists to incorporate field data and estimate unknown physical constants. In this thesis, we outline three new developments for the statistical calibration of computer models.

The first development is a practical approach for calibrating large and non-stationary computer model output. We present a new computationally efficient approach using a criterion that measures the discrepancy between the computer model and field data. One can then construct empirical distributions for the parameters and sequentially add design points to improve these estimates. The strength of this approach is its simple computation using existing algorithms. Our method also provides good parameter estimates for large and non-stationary data.

The second development deals with incorporating derivative information from the computer model into a calibration experiment. Many computer models are governed by differential equations, and including this derivative information can be helpful. Although incorporating such information has garnered a lot of attention in some areas of statistics, such as penalized regression, it has been largely ignored in computer experiments. We develop a new statistical methodology for the calibration of a computer model when derivative information is additionally available.

The final development deals with extending the methodology incorporating derivatives to

allow for the inclusion of possible bias in the computer model. A statistical model accounting for such bias was previously proposed, but heavily criticised as not being identifiable. We develop a model that accounts for this possible bias while simultaneously including the derivative information from the computer model in the hopes that such identifiability issues can be reduced or eliminated. Our results indicate some modest improvements over the previous approach in some experimental conditions. Proving exact conditions where such models can be identified remains interesting and challenging research to explore in future work.

Keywords: Calibration Parameter Estimation; Gaussian Process; Sequential Design; Derivative Information; Bayesian Modelling; Kriging

Dedication

To Jenn

Acknowledgments

There are more people to thank than I could possibly remember who have made this journey possible, but I will mention a few. First, I would like to thank my parents for always giving both myself and my sister the opportunity to follow our interests. I am grateful for the support and patience of my supervisor, Derek Bingham, who encouraged me to explore my research interests. I also wish to thank the members of my examining committee for their thoughtful feedback on my thesis, it was greatly appreciated. Thanks also go to Stephen Sain for my visit to NCAR, and many other faculty members and administrators in our department who have helped me on my way. Finally, I am indebted to Richard Lockhart for the many insightful discussions.

Last but certainly not least are my lab-mates: Joslin Goh, Ryan Lekivetz, Jervyn Ang, Shirin Golchi and Flavia Moser. I wish to thank each of you for such a fun work environment these last few years. It would not have been the same without you!

Contents

Approval ii

Abstract iii

Dedication v

Acknowledgments vi

Contents vii

List of Tables x

List of Figures xiii

1 Introduction 1

2 Background 5

2.1 Gaussian Process Regression in Computer Experiments 6

2.2 Calibration Experiments 9

2.3 Problems Addressed 14

3 Calibration for Large and Non-Stationary Fields 16

3.1 Motivating Example 17

3.2 Fast Model Calibration 21

3.2.1 Criterion 21

3.2.2 Modelling the Δ Criterion . . . 24
3.2.3 Multivariate Responses . . . 25
3.2.4 Discussion . . . 26
3.3 Sequential Design . . . 27
3.4 Simulation Study . . . 28
3.5 Modelling the Earth's Magnetosphere . . . 30
3.6 Discussion . . . 34

4 Theoretical Support for Fast Calibration 36
4.1 A Linear Covariance Model . . . 36
4.2 Gaussian Process Field Assumption and Theorem 1 . . . 40
4.3 Discussion . . . 42

5 Calibration with Derivative Information 44
5.1 Connection to Penalized Regression . . . 45
5.2 Modelling Derivatives with Gaussian Processes . . . 49
5.3 Bayesian Calibration with Derivative Information . . . 51
5.4 Example . . . 54
5.5 Calibration with Incomplete Derivatives . . . 59

6 Simulation Study: Calibration with Derivatives 65
6.1 Results: Estimating $\boldsymbol{\theta}$ . . . 69
6.2 Results: Mean Squared Prediction Error and Relative Efficiency . . . 76
6.3 Results: Prediction Interval Width and Coverage Probability . . . 85
6.4 Discussion . . . 91

7 Calibration with Derivatives: Computer Model Bias 96
7.1 Example . . . 100
7.2 Simulation Study . . . 103
7.2.1 Results: Estimating $\boldsymbol{\theta}$ . . . 105

7.2.2 Results: Mean Squared Prediction Error and Relative Efficiency for the Field . . . 107
7.2.3 Results: Mean Squared Prediction Error and Relative Efficiency for the Calibrated Computer Model . . . 112
7.3 Discussion . . . 116

8 Conclusions **117**

A Simulation-Based Inference **120**

B Properties of Multivariate Normals **122**

Bibliography **123**

List of Tables

5.1 Geweke diagnostics for the standard calibration model and the calibration model with derivatives 56

5.2 Raftery-Lewis diagnostics for the standard calibration model and the calibration model with derivatives 57

6.1 Setting levels for the 2-dimensional calibration simulation study. 68

6.2 Setting levels for the 4-dimensional calibration simulation study. 68

6.3 Setting levels for the 6-dimensional calibration simulation study. 68

6.4 Reported $MedSE(\theta)$ for $\lambda_f = 1000$ from 2-d calibration simulation study. . . 71

6.5 Reported $MedSE(\theta)$ for $\lambda_f = 10$ from 2-d calibration simulation study. . . . 71

6.6 Reported $MedSE(\theta)$ for (θ_1, θ_2) and $\lambda_f = 1000$ from 4-d calibration simulation study. 74

6.7 Reported $MedSE(\theta)$ for (θ_1, θ_2) and $\lambda_f = 10$ from 4-d calibration simulation study. 74

6.8 Reported $MedSE(\theta)$ for $(\theta_1, \theta_2, \theta_3)$ and $\lambda_f = 10$ from 6-d calibration simulation study. 76

6.9 Reported $MedMSPE$ and eff for $\lambda_f = 1000$ from 2-d calibration simulation study 78

6.10 Reported $MedMSPE$ and eff for $\lambda_f = 10$ from 2-d calibration simulation study 80

6.11 Reported $MedMSPE$ and eff for $\lambda_f = 1000$ from 4-d calibration simulation study 82

6.12 Reported $MedMSPE$ and eff for $\lambda_f = 10$ from 4-d calibration simulation study . . . 83
6.13 Reported $MedMSPE$ and eff for $\lambda_f = 10$ from 6-d calibration simulation study . . . 84
6.14 Reported $MedPIW$ and $MedCP$ for $\lambda_f = 1000$ from 2-d calibration simulation study . . . 86
6.15 Reported $MedPIW$ and $MedCP$ for $\lambda_f = 10$ from 2-d calibration simulation study . . . 88
6.16 Reported $MedPIW$ and $MedCP$ for $\lambda_f = 1000$ from 4-d calibration simulation study . . . 89
6.17 Reported $MedPIW$ and $MedCP$ for $\lambda_f = 10$ from 4-d calibration simulation study . . . 90
6.18 Reported $MedPIW$ and $MedCP$ for $\lambda_f = 10$ from 6-d calibration simulation study . . . 91

7.1 Geweke diagnostics for the standard calibration with bias model and the calibration with bias and derivatives model . . . 101
7.2 Raftery-Lewis diagnostics for the standard calibration with bias model and the calibration with bias and derivatives model . . . 102
7.3 Setting levels for the 2-dimensional calibration with bias simulation study. . . 105
7.4 Setting levels for the 4-dimensional calibration with bias simulation study. . . 105
7.5 Setting levels for the 6-dimensional calibration with bias simulation study. . . 106
7.6 Reported $MedSE(\theta)$ from 2-d calibration simulation study. . . . 106
7.7 Reported $MedSE(\theta)$ from 4-d calibration simulation study. . . . 111
7.8 Reported $MedSE(\theta)$ from 6-d calibration simulation study. . . . 111
7.9 Reported $MedMSPE$ and eff from the 2-d calibration with bias simulation study . . . 112
7.10 Reported $MedMSPE$ and eff from the 4-d calibration with bias simulation study . . . 113

7.11 Reported $MedMSPE$ and eff from the 6-d calibration with bias simulation study . . . 113
7.12 Reported $MedMSPE\eta$ and $eff\eta$ from the 2-d calibration with bias simulation study . . . 114
7.13 Reported $MedMSPE\eta$ and $eff\eta$ from the 4-d calibration with bias simulation study . . . 115
7.14 Reported $MedMSPE\eta$ and $eff\eta$ from the 6-d calibration with bias simulation study . . . 115

List of Figures

2.1 Decay of the squared-exponential covariance function 9
2.2 Fitted GP model using squared exponential covariance 10

3.1 Observed flux and energy space-time fields of a geomagnetic storm 20
3.2 Response surface for velocity of a simple harmonic oscillator 31
3.3 Raw deviations of calibration parameter estimate 32
3.4 Univariate and bivariate distributions of α, β, r 33

5.1 Traceplots of fitted standard calibration model and calibration model with derivatives 61
5.2 Autocorrelation plots for standard calibration model and calibration model with derivatives 62
5.3 Calibration of a 1-dimensional surface with a 1-dimensional calibration parameter 63
5.4 Calibrated 1-dimensional surface with a 1-dimensional calibration parameter 64

6.1 Boxplots of deviations of calibration parameter estimates and empirical mean squared prediction errors with correlation setting of 0.05 94
6.2 Boxplots of deviations of calibration parameter estimates and empirical mean squared prediction errors with correlation setting of 0.6 95

7.1 Traceplots for the standard calibration with bias model and the calibration with bias and derivatives model 108

7.2 Autocorrelation plots for the standard calibration with bias model and the calibration with bias and derivatives model 109
7.3 Calibration of a 1-dimensional surface with a 1-dimensional calibration parameter when the computer model is biased 110

Chapter 1

Introduction

It is typical in modern multidisciplinary scientific investigations to have a hypothesized or even accurate mathematical model of a physical process which one would like to fit to observational data. For instance, a system of equations may be known or proposed to describe the physics of a physical process. Usually, such a mathematical model is parametrized by physical constants whose values are unknown, but which scientists would like to learn. The purpose of such experiments are many, for instance, estimating the physical constants, verifying the mathematical model, predicting the physical process based on a fitted model, understanding the effects of certain inputs on the variability of the physical process, and so on.

In statistics, such mathematical models of a physical process are usually referred to as computer models (Sacks et al., 1989b). When the computer model is too costly to run at all settings desired, a statistical surrogate model is used in its place. The design and analysis of such a statistical model is known as a computer experiment (Santner et al., 2003). When, in addition to estimating the computer model with a statistical surrogate, real world data is additionally combined for the purposes of prediction and parameter estimation, it is said that one performs a computer model calibration experiment (Kennedy and O'Hagan, 2001). In this type of experiment, one hopes to simultaneously leverage the computer model to aid in predicting the physical process, and use the observations of the physical process to

estimate the unknown parameters governing the computer model.

The focus of this thesis is on developments related to computer model calibration experiments. The relevance of such statistical methodology has grown in importance in recent years as the ability to model natural processes using complex mathematical models has started to seem achievable. However, from a statistical perspective the challenges are significant. For instance, while the field or computer model observations analysed in such experiments are typically modelled as realizations of random fields with complex structure, we are usually dealing with a small sample size (often only n=1) of these realizations. Even so, one quickly runs into computational limitations with the statistical approaches typically used. Understanding the asymptotic behaviour of such models is also hard both because the statistical models can be difficult to manipulate, and the usual asymptotic framework is typically not applicable. The efficient design of such experiments also suffers from the so-called curse of dimensionality.

While we certainly cannot hope to resolve all these issues here, we aim to offer new insight by investigating some well-defined problems. In the first part of this thesis, we investigate the calibration of a computer model whose outputs are too large to handle with the usual approach to calibration. This leads to the development of a primary-stage approach to computer model calibration of large, non-stationary fields with a simple sequential design strategy. As a by-product, we are able to develop some understanding of the statistical properties of the calibration parameter estimate in this case. To our knowledge, no results on the statistical properties of the calibration parameter estimate have been made in the literature.

The second part of the thesis focuses on the development of new statistical models for the calibration of computer experiments that incorporate knowledge of the derivatives of the computer model into the calibration. We develop Bayesian models to incorporate the observed derivative fields from the computer model into a calibration experiment. The inclusion of this higher-order information provides refined parameter estimates of the fitted

model and a potentially large reduction in prediction error while also improving the second order properties of the predictor - i.e. the pointwise prediction intervals. On the other hand, the degree of measurement error has a diminishing effect on these benefits.

In the final part of this thesis, we extend the Bayesian models of the second part to include the possibility of systemic bias between the computer model and the real-world process. Modelling this bias has previously been done by using an additive model who's identifiability was questionable. It is hoped that by introducing derivative information into the model, the identifiability issue could be improved or resolved. To this end, we compare these models which include bias with and without derivative information via simulation and arrive at some important conclusions in the application of these models.

Before delving into the results of this thesis, we first provide a background summary of the required methodologies in computer experiments on which our methods are based. In Chapter 2 we outline the usual approach to Gaussian Process regression, and extend that to regression with derivatives. We then introduce the calibration model.

In Chapter 3, we develop a criterion-based approach to model calibration that is applicable with large and non-stationary datasets. In Chapter 4, we provide some theory to support that the criterion-based approach of Chapter 3 is effective. In particular, we are able to prove the asymptotic consistency of the calibration parameter estimate, in the infill sense (Stein, 1999). We have not previously seen any results on the consistency of this parameter in the literature.

In Chapter 5, we develop a Bayesian model for calibration that allows for the inclusion of derivative observations from the computer model. The novelty of our approach as opposed to, for instance, approaches from the penalized regression literature, is that in the latter case, the presence of derivatives of the process only serves to provide better estimates of various model parameters, while in our approach the derivative process additionally lends

predictive information for the process. This difference stems, in part, from the scenarios where the respective modelling strategies apply; in penalized regression (Ramsay and Silverman, 2005), one is usually data rich, while in our applications, we are relatively data poor. We investigate this relationship to try to understand what benefits the inclusion of derivatives brings. This is followed up in Chapter 6 by an extensive simulation study where we investigate 2-dimensional, 4-dimensional and 6-dimensional calibration problems, and compare the usual model to a model including full derivative information as well as the sometimes more practical case of partial derivative information. In Chapter 7, we extend the models of Chapter 5 to account for the possibility of computer model bias, and compare these models via simulation study.

Finally, in Chapter 8, we offer conclusions and directions for future research in computer model calibration experiments.

Chapter 2

Background

The statistical analysis of computer experiments has as its foundation problems arising from the industrial engineering sciences, and as a solution, methodology based on statistical approaches first made popular in spatial statistics. The motivation from industrial engineering arises from problems where the physical behaviour of some process has been implemented in a computer model, η. Such computer models are often used to study real-world processes that are too difficult to investigate directly using traditional experimental methods. The models attempt to incorporate all known relevant behaviour of the physical system and allow the investigator to explore the impact on this process of changing the levels of the inputs without directly obtaining physical measurements (i.e., perform a *computer experiment*).

The statistical problem arises due to the cost of obtaining data from computer models. There are many reasons for this cost to negate the possibility of sampling from these models whenever one would like. For instance, the computing requirements of the model may be so high as to only allow limited sampling, even on the most recent supercomputers; running the model may require specialized personnel; the number of runs available are insufficient to answer the question of interest (maximization, sensitivity analysis, etc.); and so on. Computer models also have the interesting property that they are typically deterministic. That is, at a given input setting, they produce the same output no matter how often you run the model at that setting.

As a result, the purpose of a statistical model for such computer models is to capture and describe the "functional" or "process" variability of the physical process described by the computer model. The methodology that has become popular is known as Gaussian Process (GP) regression, sometimes called Kriging. This model was originally proposed by the engineer Joseph Krige (Krige, 1966) and applied to a mining problem. Its application to spatial statistics is summarized in Cressie (1993) and Ripley (1981). In this context, the data $y(\boldsymbol{x})$ observed at spatial location $\boldsymbol{x} \in \mathcal{D}$ is modelled as $y(\boldsymbol{x}) = z(\boldsymbol{x}) + \epsilon(\boldsymbol{x})$ so that $y(\boldsymbol{x})$ is viewed as being a realization of an unknown process $z(\boldsymbol{x})$ in the presence of some error process $\epsilon(\boldsymbol{x})$, where $\mathcal{D}$ is the domain where the process is observed.

Usually, the unknown process $z(\boldsymbol{x})$ is modelled as a continuous GP (Stein, 1999) over $\mathcal{D}$ with mean function $\mu(\boldsymbol{x})$ and covariance $C(z(\boldsymbol{x}_i), z(\boldsymbol{x}_j)) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where the degree of smoothness of the process can be controlled by choosing K accordingly. The error process is usually modelled as being mean-zero independent and identically distributed (i.i.d.) Gaussian error independent of z with covariance $C(\epsilon(\boldsymbol{x}_i), \epsilon(\boldsymbol{x}_j)) = \sigma^2 \mathcal{I}_{\boldsymbol{x}_i, \boldsymbol{x}_j}$, where $\mathcal{I}_{\cdot,\cdot}$ is the Kronecker delta function. The statistical model for observed data $\mathbf{Y} = [y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_N)]^T$ is then

$$\mathbf{Y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}^{\epsilon}\right),$$

where $\boldsymbol{\mu} = [\mu(\boldsymbol{x}_1), \ldots, \mu(\boldsymbol{x}_N)]^T$, $\boldsymbol{\Sigma}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and $\boldsymbol{\Sigma}^{\epsilon}_{i,j} = \sigma^2 \mathcal{I}_{xi,xj}$.

2.1 Gaussian Process Regression in Computer Experiments

There are many reasons why GP's are a reasonable statistical model for describing data collected in a computer experiment. For example, the deterministic nature of computer models can be accommodated by simply setting $\epsilon(\boldsymbol{x}) = 0$. In this case, the GP has the nice behaviour that it will interpolate observations collected in a computer experiment (Jones et al., 1998). Another benefit is the small number of parameters as compared to spline-based approaches. This occurs as most of the modelling is done through specification of the

covariance function, which incorporates assumptions on smoothness and correlation without requiring one to estimate individual weights in a high-order expansion of basis functions. In fact, in computer experiments it is usually the case that the mean term is also modelled as $\mu(\boldsymbol{x}) = 0$, unless the particular problem at hand warrants otherwise.

Additional assumptions usually made in applying GP's to computer experiments are stationarity and anisotropy. Stationarity is satisfied by specifying $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K(|\boldsymbol{x}_i - \boldsymbol{x}_j|)$. That is, the covariance structure only depends on the distance between the observation sites but not on the particular location of a site itself. Anisotropy allows for different covariation along different dimensions, e.g. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \prod_{d=1}^{D} K_d(x_{id}, x_{jd})$, where $K_p \neq K_q$ for $p \neq q$. This allows for a response that has a different behaviour for different input factors.

Finally, there are numerous covariance models parametrized to allow different degrees of smoothness, or differentiability, and correlation: linear, cubic, exponential, Matern, and squared exponential, amongst others (Rasmussen and Williams, 2006). The most popular of these is the anisotropic squared exponential model,

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_z^2 e^{\sum_{d=1}^{D} \phi_d (x_{id} - x_{jd})^2}, \tag{2.1}$$

which defines a process that is infinitely differentiable in the mean-squared sense as well as in the path-wise sense, has variance $\sigma_z^2 > 0$, and correlation scale parameter $\phi_d > 0$ for each of the $d = 1, \ldots, D$ dimensions. In the computer experiments literature, this model has been widely used (Sacks et al., 1989a,b). Since computer models are usually systems of differential equations, the notion of the process being at least a few times differentiable is almost always reasonable. However, the stronger assumption of infinite differentiability is not without its critics (Stein, 1999).

Let us now introduce some further notation and standard results for the statistical model of a computer experiment using the squared exponential covariance function. Refer to the computer model as $\eta(\boldsymbol{x})$, which at this point depends on a D-dimensional input setting $\boldsymbol{x}$, $D \geq 1$. The computer model can be run at input settings $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, where $\boldsymbol{x}_i \in \mathcal{D}$, for all i

and $\mathcal{D} \subset \mathcal{R}^D$ is the domain of possible settings of these continuous inputs. An observation, $y(\boldsymbol{x})$ is modelled as

$$y(\boldsymbol{x}) = \quad z(\boldsymbol{x}) \tag{2.2}$$

where $z(\boldsymbol{x})$ is modelled as an unknown mean-zero GP with squared exponential covariance (2.1), and $z(\boldsymbol{x}_i) = \eta(\boldsymbol{x}_i)$, that is the unknown process interpolates observations collected from the computer model. Since the process is observed error-free, there is no difference in the specification of the distribution of $\mathbf{Z} = [z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_N)]^T$ or of $\mathbf{Y} = [y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_N)]^T$, so we will proceed using notation based on the latter. Then,

$$\mathbf{Y} \sim \mathcal{N}(0, \sigma_z^2 \prod_{d=1}^{D} \rho_d^{(x_{id}-x_{jd})^2}), \tag{2.3}$$

where we will often use the re-parametrization $\rho_d = e^{-\phi_d}, 0 < \rho_d < 1$, of the squared exponential covariance model for convenient interpretation. For instance, $\rho_d \approx 0$ implies low correlation between observations while $\rho_d \approx 1$ implies high correlation between observations. The behaviour of this decay is shown in Figure 2.1.

The model can be fit using either maximum likelihood or in a Bayesian framework. For now, we will limit ourselves to stating the standard MLE results for the models in this section. For model (2.3), it has been shown (Jones et al., 1998) that

$$\widehat{\sigma_z^2} = \mathbf{Y}^T \mathbf{R}^{-1} \mathbf{Y},$$

where $\mathbf{R}_{i,j} = \prod_{d=1}^{D} \rho_d^{(x_{id}-x_{jd})^2}$, while the MLE of the correlation parameters ρ_d have no known closed-form solution, so must be estimated numerically. The best linear unbiased predictor (BLUP) of this model at a new location $\boldsymbol{x}^*$ is

$$\widehat{y}(\boldsymbol{x}^*) = \quad \boldsymbol{r}^T \mathbf{R}^{-1} \mathbf{Y} \tag{2.4}$$

with the variance of the predictor given as

$$s^2(\boldsymbol{x}^*) = \quad \sigma_z^2 \left(1 - (\boldsymbol{r}^T \mathbf{R}^{-1} \boldsymbol{r})^{-1} \boldsymbol{r}^T \mathbf{R}^{-1} \mathbf{Y}\right) \tag{2.5}$$

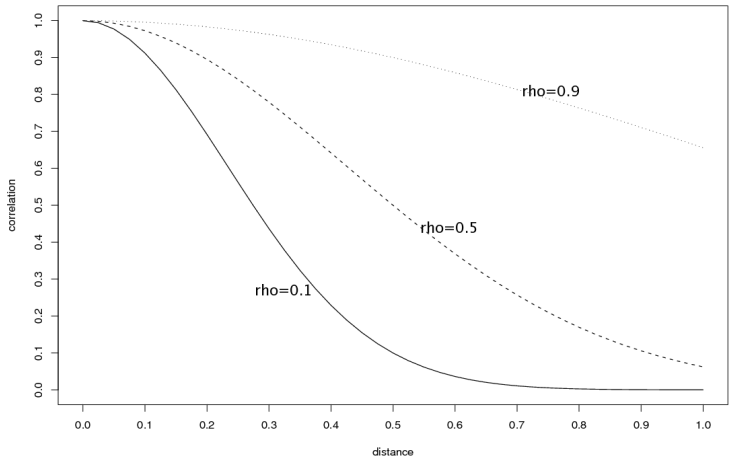


Figure 2.1: Decay of the squared-exponential covariance function with a variance of 1 and ρ =0.1, 0.5 and 0.9 on the unit interval [0,1].

where $\boldsymbol{r}^T = [corr(y(\boldsymbol{x}^*), y(\boldsymbol{x}_1)), \ldots, corr(y(\boldsymbol{x}^*), y(\boldsymbol{x}_N))]$ is the vector of correlations between the process at the prediction location and the observed data (Sacks et al., 1989b). A 1-dimensional example of this model fit to $N = 5$ data points is shown in Figure 2.2. The data points are the red dots while the prediction is the solid red line. The blue lines denote the 95% pointwise prediction interval. Note how the prediction interpolates the data points, and that the variance of the predictor is zero at these points. One can also see that where there are many data points (left), the variance of the predictor is small as there are many observations lending information in that region, while the variance is large where there is a lack of observations (right).

2.2 Calibration Experiments

Now we build on (2.1)-(2.5) to allow for a model that considers computer model observations, field data and unknown computer model parameters that govern the field. Such models are the main focus of this thesis. In our setting, the inputs to computer models can be broadly classified into two groups. The first group consists of the design variables, $\boldsymbol{x}$, which have

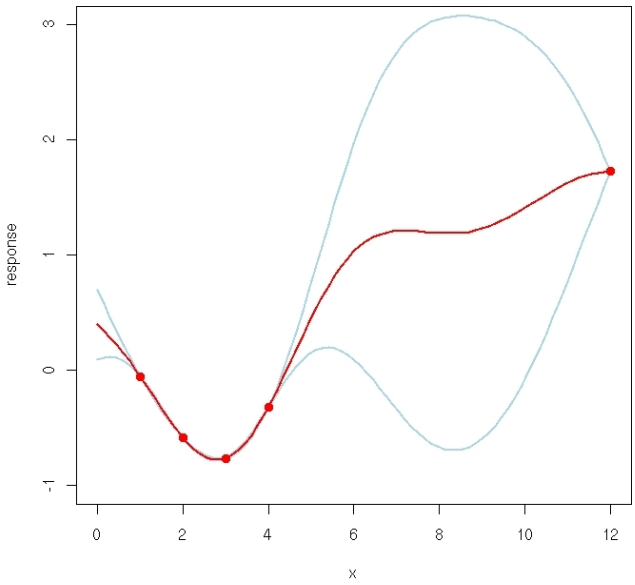


Figure 2.2: GP model with squared exponential covariance fit to $N = 5$ data points (red). The predictor (red) is shown along with the 95% pointwise prediction interval (blue).

already been introduced. These are inputs that can be measured and/or adjusted in the physical system, such as the spatial locations that have motivated our development so far. The second set of variables, called *calibration parameters* (Kennedy and O'Hagan, 2001), are unknown physical constants that impact the system.

So far, we have only considered modelling scenarios where the first type of input is present. Yet, when running the computer model, some value will have to be inserted for the calibration parameters to run the code. An important goal for experimenters is estimating the calibration parameters using computer model outputs and field observations. In the case of an easy to sample computer model, estimating the calibration parameter $\boldsymbol{\theta}$ is conceptually simple. For instance, Higdon et al. (2004) suggest modelling the observed field data as centered on the computer model evaluated at $\boldsymbol{\theta}$, i.e. $y^f(\boldsymbol{x}) = \eta(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon$. Then,

placing a prior on $\boldsymbol{\theta}$, one can arrive at the posterior distribution of the calibration parameter. This inverse problem can be viewed as a Bayesian approach to non-linear regression. While this provides a clear conceptual understanding of a calibration experiment, the approach is rarely practical due to the cost of evaluating η.

The essential framework to computer model calibration was proposed by Kennedy and O'Hagan (2001). Their method was adapted by Higdon et al. (2004) where a single Bayesian hierarchical model is used for model calibration. The latter model (denoted hereafter as the "standard model") is specified as:

$$\begin{aligned} y^f(\boldsymbol{x}) = &\quad z(\boldsymbol{x}, \boldsymbol{\theta}) + b(\boldsymbol{x}) + \epsilon, \\ y^c(\boldsymbol{x}, \boldsymbol{t}) = &\quad z(\boldsymbol{x}, \boldsymbol{t}), \end{aligned} \tag{2.6}$$

where the $\boldsymbol{x}$'s denote design locations for the field observation and design inputs for the computer trial, $\boldsymbol{t}$ denotes the calibration inputs for the computer trial, $z(\boldsymbol{x}_i, \boldsymbol{t}_i) = \eta(\boldsymbol{x}_i, \boldsymbol{t}_i)$ where η denotes the output from the computer simulator, $b(\boldsymbol{x})$ represents a systemic bias between the computer model output and field observation, and $\epsilon \sim N\left(0, \sigma^2\right)$ represents i.i.d. measurement error. The partially observed process z is modelled as a GP with mean zero and covariance $\Sigma = \begin{bmatrix} \Sigma^f & \Sigma^{fc} \\ \Sigma^{cf} & \Sigma^c \end{bmatrix}$, where the covariance matrix has been partitioned into covariances for the observed computer model runs, Σ^c, the unobserved computer model runs at the unknown true calibration parameter $\boldsymbol{\theta}$ denoted by Σ^f, and the cross-covariance between the observed and unobserved runs, Σ^{cf}. The model for the bias is also taken to be a GP with mean zero and covariance Σ^b. The computer model observations, bias and measurement error are assumed to be mutually independent, giving the model for the observations $\mathbf{Y}^f = [y^f(\boldsymbol{x}_1), \ldots, y^f(\boldsymbol{x}_N)]^T$ and $\mathbf{Y}^c = [y^c(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, y^c(\boldsymbol{x}_M, \boldsymbol{t}_M)]^T$ as

$$\begin{pmatrix} \mathbf{Y}^f \\ \mathbf{Y}^c \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma^f & \Sigma^{fc} \\ \Sigma^{cf} & \Sigma^c \end{bmatrix} + \begin{bmatrix} \Sigma^b & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \sigma^2 \mathbf{I}_N & 0 \\ 0 & 0 \end{bmatrix}\right), \tag{2.7}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix, and the covariances are again usually modelled using the squared exponential:

$$\Sigma^b_{i,j} = \quad \sigma_b^2 \prod_{d=1}^{D} \psi_d{}^{(x_{id} - x_{jd})^2},$$

with correlation parameters $\psi_d, d = 1, \ldots, D$ for the D space-time inputs, and,

$$\Sigma_{i,j} = \sigma_z^2 \prod_{d=1}^{D} (\rho_d)^{(x_{id}-x_{jd})^2} \prod_{l=1}^{L} (\phi_l)^{(t_{il}-t_{jl})^2} \tag{2.8}$$

where L are the number of calibration inputs, with corresponding correlation scale parameters $\rho_d, d = 1, \ldots, D$ and $\phi_l, l = 1, \ldots, L$ respectively.

When model (2.6) was first introduced, it was not without some criticism (Kennedy and O'Hagan, 2001). In particular, the identifiability of the bias term, $b(\boldsymbol{x})$, was questioned, and it is commonly believed that this term is not identifiable. A simple analogy is the model $y = a + b + \epsilon$. In this case, one can never identify the true value of a or b no matter what is observed, but of course identifiability of the sum $a + b$ poses no problem. On the other hand, for the purposes of prediction, identifiability may not be an issue.

In practice, a simpler version of model (2.6) is often used, and it is this model that we define now, and which will be our starting point for the models developed in this thesis. Consider the case where it is reasonable to assume $b(\boldsymbol{x}) = 0$. In other words, it is believed that the computer model is able to reproduce all necessary aspects of the physical process, and hence no bias in its simulation of reality is expected. Or, the magnitude of the bias is in the same order as the magnitude of the measurement error. In this case,

$$\begin{aligned} y^f(\boldsymbol{x}) &= z(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon, \\ y^c(\boldsymbol{x}, \boldsymbol{t}) &= z(\boldsymbol{x}, \boldsymbol{t}), \end{aligned} \tag{2.9}$$

where $z(\boldsymbol{x}_i, \boldsymbol{t}_i) = \eta(\boldsymbol{x}_i, \boldsymbol{t}_i)$, and $\mathbf{Y}^f = [y^f(\boldsymbol{x}_1), \ldots, y^f(\boldsymbol{x}_N)]^T$, $\mathbf{Y}^c = [y^c(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, y^c(\boldsymbol{x}_M, \boldsymbol{t}_M)]^T$, leading to

$$\begin{pmatrix} \mathbf{Y}^f \\ \mathbf{Y}^c \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma^f & \Sigma^{fc} \\ \Sigma^{cf} & \Sigma^c \end{bmatrix} + \begin{bmatrix} \sigma^2 \mathbf{I}_N & 0 \\ 0 & 0 \end{bmatrix}\right), \tag{2.10}$$

where the covariances are modelled using the squared exponential as above (2.8).

The idea of this approach is to jointly model the field data and computer model outputs using stationary Gaussian process models. Markov Chain Monte Carlo (MCMC) is used to

arrive at the posterior distribution of the calibration parameters and a predictive distribution for the field observations. These two goals are, of course, tightly intertwined. When the number of field observations is relatively small, as is expected in a calibration experiment, then prediction of the field process is improved by borrowing information from the computer model. On the other hand, without enough field observations, a good estimate of the calibration parameter will not be possible, thereby hindering the ability of the computer model to predict the field process.

We will now try to develop this intuition more formally. In trying to understand the properties of correlated models such as the above, a simplification that is frequently made (at least initially) is to consider the behaviour when the error term is absent. Suppose we want to predict the field at a new location $\boldsymbol{x}^*$, and denote its unknown value as $y^f(\boldsymbol{x}^*)$. Writing the vector of covariances between $y^f(\boldsymbol{x}^*)$ and the observed field and computer model data as

$$\begin{aligned} c^T = \; & [cov(y^f(\boldsymbol{x}^*), y^f(\boldsymbol{x}_1)), \ldots, cov(y^f(\boldsymbol{x}^*), y^f(\boldsymbol{x}_N)), \\ & \quad cov(y^f(\boldsymbol{x}^*), y^c(\boldsymbol{x}_1, \boldsymbol{t}_1)), \ldots, cov(y^f(\boldsymbol{x}^*), y^c(\boldsymbol{x}_M, \boldsymbol{t}_M))] \\ = \; & [c^{fT}, c^{cT}], \end{aligned}$$

then after some algebra it can be shown that

$$\begin{aligned} \widehat{y^f}(\boldsymbol{x}^*) = \; & E[y^f(\boldsymbol{x}^*)|\mathbf{Y}^f, \mathbf{Y}^c, \cdot] \\ = \; & \left(c^{fT} - c^{cT}(\Sigma^c)^{-1}\Sigma^{cf}\right)\left(\Sigma^f - \Sigma^{fc}(\Sigma^c)^{-1}\Sigma^{cf}\right)^{-1}\mathbf{Y}^f \\ & + \left(c^{cT} - c^{fT}(\Sigma^f)^{-1}\Sigma^{fc}\right)\left(\Sigma^c - \Sigma^{cf}(\Sigma^f)^{-1}\Sigma^{fc}\right)^{-1}\mathbf{Y}^c \\ = \; & \textstyle\sum_{i=1}^N w_i^f(\boldsymbol{\theta}) y^f(\boldsymbol{x}_i) + \sum_{j=1}^M w_j^c(\boldsymbol{\theta}) y^c(\boldsymbol{x}_j, \boldsymbol{t}_j) \end{aligned}$$

This result shows that the field process is predicted as a weighted combination of the field and computer model observations. The role of the calibration parameter comes through two components, namely the vector c^c and the matrix Σ^{cf}, which both depend on the unknown $\boldsymbol{\theta}$. For instance, if the estimated calibration parameter indicates that the field observation is "far" from the calibration settings that the computer model was run at (i.e. $|\hat{\boldsymbol{\theta}} - \boldsymbol{t}_i|$ is

large for all i), then these components will be relatively small (i.e. note that the covariance decreases with the squared distance in equation (2.8)). In other words, in this situation the predicted value would be based mainly on the field observations. In the extreme case of $c^c = \mathbf{0}$ and $\Sigma^{cf} = \mathbf{0}$, the formula above would degenerate to just $\widehat{y^f}(\boldsymbol{x}^*) = {c^f}^T(\Sigma^f)^{-1}\mathbf{Y}^f$, the usual predictor for a GP model of the field. Of course, in practice one would never arrive at this case. On the other hand, if the calibration parameter was poorly estimated, then the prediction might be unduly influenced by the computer model observations if they receive too much weight (i.e. if the poorly estimated calibration parameter suggested that the field data was "closer" to some of the computer model outputs than it actually is).

2.3 Problems Addressed

The previous section introduced computer model calibration and motivated the idea of a calibration experiment: to borrow information from the computer model to improve predictions of the field process while also using the field observations to obtain estimates of the unknown calibration parameter. In practice, there are two main areas where the calibration models (2.6) and (2.9) can be improved which we address in this thesis.

One problem that can occur is if the ability to sample both the field process and computer model are highly limited. In such scenarios, the computer model runs that are obtained may have settings of the calibration parameter inputs too far from the true calibration parameter that governs the field. In other words, the amount of useful information that can be borrowed from the computer model runs for predicting the field process may be very small. At the same time, a good estimate of the calibration parameter may not be possible. In Chapters 5-7, we consider including derivative information in computer model calibration as one way to improve these situations.

Another problem occurs when the data structure being calibrated is very large, leading to covariance matrices in (2.7) and (2.10) that are simply too big to manipulate, store and

invert on computer. The covariance of such data structures may also change as the calibration parameter is varied, which violates the stationarity assumption of the usual approach to calibration. In the next chapter, we propose a method for estimating the calibration parameter in such a scenario.

Chapter 3

Calibration for Large and Non-Stationary Fields

The goal of the current chapter is to develop a design and analysis strategy for estimating the calibration parameters in a computer model calibration experiment. However, the case we consider in this chapter is unique for a few reasons. First, we are only interested in estimating the calibration parameters because they are needed for future runs of the computer model rather than building a predictive model of the physical system. Second, it is felt that the computer model is essentially unbiased so that running the simulator at the true value of the calibration parameters will reproduce the behaviour of the physical system, up to observational error. Furthermore, the inputs to the computer model are only calibration parameters. This represents the scenario where there are no design parameters or they are held constant across experiments.

It also turns out that the outputs from the computer model and the field observations from the physical system are both large space-time fields. Modelling of such data structures typically requires the specification of a GP model as introduced in Chapter 2. The use of large covariance matrices can be impractical to store and invert when evaluating the likelihood for the GP. Moreover, the variability of the computer model output, which is the solution to a system of partial differential equations, can change drastically for even small

changes in the calibration parameters. This presents another challenge, since the usual approach to calibration assumes that the model outputs can be represented by a stationary GP.

The final feature in our setting is the high computational demands of the computer models being investigated. These models can only be run at a limited number of calibration settings on large supercomputers. So, efficient design of the computer experiment is necessary.

We propose new methodology for the design and analysis of computer model calibration experiments that is effective for computer models which generate large outputs and, possibly, exhibit non-stationarity as the calibration parameters are varied. Our approach models the difference between the computer model output and observed field data at each setting of the calibration parameters. The idea underlying the proposed approach is that at the true value of the calibration parameters, the difference between the computer model output and the field observations should follow independent white noise. A new criterion, that can be viewed as a form of the Kullback-Liebler (KL) divergence (see Kullback (1997)) between two densities, is proposed. The criterion is then modelled as a function of the computer model inputs, and we show how minimizing this function can give estimates of the calibration parameters. Furthermore, this approach has the added benefit that one can easily use existing sequential design methodology (see Jones et al. (1998); Williams et al. (2000)) to propose new computer model runs. The approach greatly reduces the computational burden present when modelling very large response fields, and handles the presence of non-stationarity of the response in a simple manner.

3.1 Motivating Example

The motivation for the work in this chapter comes from a computer simulator of the effect of electron storms on the upper atmosphere. In extreme cases, these storms are observed as the well known aurora borealis. The Lyon-Fedder-Mobary (LFM) model (Fedder et al.

(1995); Lyon et al. (2004)) is a model of the magnetosphere - the region near the Earth where the magnetic field forms a protective bubble which impedes the transfer of energy and momentum from solar wind plasma. Understanding how the electric currents and fields behave in the magnetosphere has practical consequences with satellite orbits, radio communication and GPS systems, especially during solar disturbances.

The LFM model is a global magnetohydrodynamic (MHD) simulation which depends on three unknown parameters, (α, β, r). These are what we refer to as calibration parameters. The parameter, α, is related to the effects determining electron energy from the fluid MHDs, β helps govern the effects of plasma anisotropy, and r is related to the effects of current direction and strength. The model has two outputs: (i) average energy and (ii) flux of precipitating electrons. Both are space-time fields, which constitute the design variables. This particular example is somewhat atypical in the sense that a single high-density field observation is sampled by satellite, and the resulting large data structure presents challenges in estimating the usual calibration model (2.9). In addition, presence of non-stationarity in the simulators outputs is not accounted for in the standard approach.

For notational purposes, the computer code takes in an $L \times 1$ input vector $\boldsymbol{t}$, and outputs an $N \times 1$ vector $\mathbf{Y}^c$ representing a space-time field. Denote the true (but unknown) values of the calibration parameters by the $L \times 1$ vector $\boldsymbol{\theta}$. The inputs, $\boldsymbol{t}$, are used in place of $\boldsymbol{\theta}$ when running the code. The outputs, $\mathbf{Y}^c$, can be thought of as a stacked vector representation of a D-dimensional space-time field where the stacking has been done by varying the spatial variables first and the time variable last.

The LFM model is computationally intensive, so efficient use of resources is essential. To gain some understanding of how the response varies as a function of the inputs, a limited number of initial runs, n, was made with inputs chosen via a space filling design (e.g. Johnson et al. (1990)), giving computer model runs $\mathbf{Y}_1^c, \ldots, \mathbf{Y}_n^c$ at locations $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$.

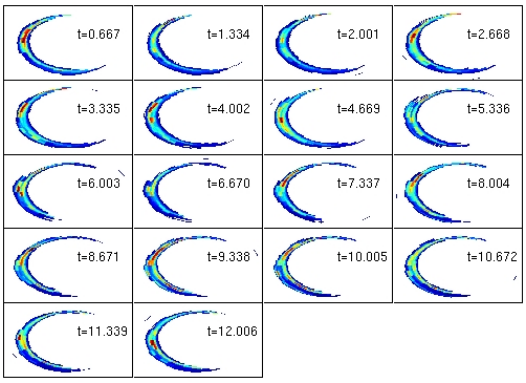
In order to learn information about the parameters $\boldsymbol{\theta} = (\alpha, \beta, R)$ for the LFM computer model, measurements were made by the Polar Ultraviolet Imager (UVI) during a geomagnetic storm on January 10, 1997. The UVI recorded flux and energy fields of the geomagnetic storm over the Earth's northern polar region for a period of about 2 hours. These observations are shown in Figure 3.1(a) and Figure 3.1(b). Denote a single space-time field observation recorded by the UVI as the $N \times 1$ vector $\mathbf{Y}^f$, which has the same stacked construction as the computer model output. The space-time field is observed over an equidistant grid constructed over 9 and 12 locations of the two space coordinates respectively and 18 equally spaced time points. The computer model outputs are constructed over this same grid with the size of these vector outputs for our problem given as $N = 9 \times 12 \times 18 = 1944$. Each of these 1944×1 vectors represent either computer model outputs at different settings of the calibration parameters, or the actual observed flux or energy fields.

The aim at this point is to use the field and computer model data to help estimate the unknown parameters, $\boldsymbol{\theta}$, and also to propose new computer trials to improve the estimate. The LFM model is believed to exhibit little or no bias, and thus we operate under the assumption that $b(\boldsymbol{x}) = 0$. The calibration model can then be written in vector notation as,

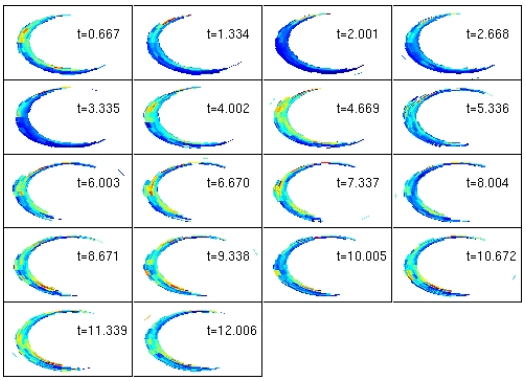
$$\begin{aligned} \mathbf{Y}^f &= \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}, \\ \mathbf{Y}_i^c &= \boldsymbol{\eta}(\boldsymbol{t}_i), \end{aligned} \tag{3.1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$. The assumption that there is no systemic bias has significant implications. Most importantly, when the computer model is run at the true value of the calibration parameter, the computer model output will exactly reproduce the observed field data, up to the measurement error, $\boldsymbol{\epsilon}$.

The standard model proceeds, in this case, by jointly modelling the field observations $\mathbf{Y}^f$ and computer model outputs $\mathbf{Y}^c = \left(\mathbf{Y}_1^{cT}, \ldots, \mathbf{Y}_n^{cT}\right)^T$ with a stationary Gaussian Process model as a function of the calibration parameters (2.10). The correlation model typically employed is the powered-exponential form (2.1). Here, the practical limitation is computing the inverse and determinant of the $N(n+1) \times N(n+1)$ covariance matrix when evaluating



(a)



(b)

Figure 3.1: Observed flux (a) and energy (b) space-time fields of a geomagnetic storm over the north pole. Each box plots the observed flux or energy in radial coordinates, and the change in these responses over time is seen by scanning from the top left corner to the bottom right corner of each plot. The scaled time units of when observations were taken are printed in each box.

the likelihood. For the LFM model, this amounts to manipulating a $40,824 \times 40,824$ covariance matrix at each step of the MCMC, which is not computationally feasible. Although it is sometimes possible to work with large datasets by using a Kronecker product form for the covariance, the lack of stationarity of our data made this approach problematic. And, as higher resolution models become available, even larger datasets can be expected.

These limitations result in the standard model being difficult to implement, or often impossible, in many practical situations. In the next section, a new approach is proposed that works well in the situation described.

3.2 Fast Model Calibration

In this section, a new approach to computer model calibration that is effective for computer simulators with large and/or non-stationary outputs is proposed. Rather than modelling the joint distribution of the field observations and computer model output directly, a univariate representation for the discrepancy between the computer model output and the field observations is proposed. By modelling this discrepancy surface as a function of the calibration parameters, the calibration problem is simplified to a straightforward minimization problem.

3.2.1 Criterion

Consider model (3.1). We begin by computing the difference between each field observation and each computer model output,

$$\boldsymbol{\delta}(\boldsymbol{t}_i) = \mathbf{Y}^f - \mathbf{Y}_i^c = \boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{t}_i) + \boldsymbol{\epsilon}.$$

If we are lucky enough to run the computer model at the correct setting of the calibration parameter on the ith trial (i.e. $\boldsymbol{t}_i = \boldsymbol{\theta}$), the difference between $\mathbf{Y}^f$ and $\mathbf{Y}_i^c$ should follow the measurement error process $\boldsymbol{\epsilon}$. Thus, the difference can be written as $\boldsymbol{\delta}(\boldsymbol{t}_i) \sim N(0, \sigma^2 \mathbf{I}_N)$, where $\sigma^2 > 0$ is the variance of the measurement error. We will call this model the restricted

model.

Now suppose $\boldsymbol{t}_i \neq \boldsymbol{\theta}$. In this case, the difference $\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{t}_i)$ will be non-zero and can be assumed to exhibit some space-time dependence. In this case, we model the difference field, $\boldsymbol{\delta}(\boldsymbol{t}_i)$, as a GP. Specifically, let the model be $\boldsymbol{\delta}(\boldsymbol{t}_i) \sim N(\boldsymbol{\mu}^i, \sigma^2(\mathbf{I}_N + \gamma^i \mathbf{R}(\boldsymbol{\phi}^i)))$ where $\boldsymbol{\mu}^i$, σ^2, γ^i represent the mean (usually taken to be constant), measurement error variance and the process to error variance ratio. The correlation between the pth and q^{th} space-time locations of the difference field are modelled using the squared exponential correlation, $\left[\mathbf{R}(\boldsymbol{\phi}^i)\right]_{p,q} = exp\left(-\sum_{d=1}^{D} \phi_d^i \left(x_{pd} - x_{qd}\right)^2\right)$ where $\boldsymbol{x}_p = (x_{p1}, \ldots, x_{pD})$ and $\boldsymbol{x}_q = (x_{q1}, \ldots, x_{qD})$ are D-dimensional space-time coordinates, and $0 < \phi_d^i < \infty$ represents the correlation scale parameter for the dth space-time variable. Note that we have indexed the parameters here by superscript i only to emphasize the dependency on each run of the computer model. We refer to the model in this case as the unrestricted model.

The basic idea behind our approach is that at a "good" estimate of the calibration parameter, $\boldsymbol{\theta}$, the two models will be indistinguishable. We propose to measure the level of (dis)agreement between the two models using the discrepancy criterion:

$$\Delta(\boldsymbol{t}_i) = -2\ log\left(\frac{L_r^*(\boldsymbol{\delta}(\boldsymbol{t}_i))}{L_u^*(\boldsymbol{\delta}(\boldsymbol{t}_i))}\right), \tag{3.2}$$

where $L_r^*(\boldsymbol{\delta}(\boldsymbol{t}_i))$ is the maximized likelihood when the difference is modelled as a measurement-error process (i.e. the restricted model), and $L_u^*(\boldsymbol{\delta}(\boldsymbol{t}_i))$ is the maximized likelihood when the difference is modelled as having space-time dependence (i.e. the unrestricted model).

The criterion can be recognized as the log of the likelihood ratio test statistic, well known for its application in deciding between two models when one model (the restricted model) is nested under a more general model (the unrestricted model). This corresponds to the present case, as taking $\gamma^i = 0$ in the unrestricted model reduces to the restricted model.

So, when $\boldsymbol{t} = \boldsymbol{\theta}$, the ratio in (3.2) is expected to be relatively small. Similarly, when $\boldsymbol{t} \neq \boldsymbol{\theta}$, then the ratio in (3.2) should be relatively large. Thus, it is tempting to estimate $\boldsymbol{\theta}$

with the value, $\widehat{\boldsymbol{\theta}}$, that minimizes (3.2), and indeed we do just that. Specifically, we claim that

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{t}} \quad \Delta(\boldsymbol{t}) \tag{3.3}$$

is a consistent estimator of $\boldsymbol{\theta}$ (see Theorem 1 below). To prove that $\widehat{\boldsymbol{\theta}}$ in (3.3) is consistent for $\boldsymbol{\theta}$, we rely on the consistency of maximum likelihood estimators (MLE's) for the unrestricted model. For instance, consistency of the parameter estimates for the anisotropic exponential covariance model in 1-dimension and 2-dimensions was investigated by Chen et al. (2000), while the consistency of parameter estimates for the Matern covariance model, of which the exponential and squared exponential models are special cases, was investigated in 1-dimension by Zhang (2004). The results, obtained for the case of sampling a single realization of a stochastic process on a fixed domain (so-called infill asymptotics), usually rely on the maximization of the likelihood over a compact region of the parameter space assumed to contain the true parameter value, and taking observations on a regular grid. In any case, when the MLEs for the parameters are consistent, we arrive at the following theorem, which we will show in 1-dimension.

Theorem 1:
Suppose that the MLE's under the unrestricted model are consistent. Furthermore, assume that the unrestricted model for the discrepancy is correct. Under these conditions, $\hat{\theta} \longrightarrow \theta$ *in probability as* $N \to \infty$.

We will defer the proof to Chapter 4.

The theorem supports the general strategy where $\boldsymbol{\theta}$ is estimated by the minimum of the proposed discrepancy. That is, we choose the estimate of $\boldsymbol{\theta}$ where the white noise model for the difference is best supported over one displaying space-time dependence in the difference between the observed data and computer model output.

If we could repeatedly sample the simulator, an algorithm that minimized $\Delta(\boldsymbol{t})$ as a

function of $\boldsymbol{t}$ could be devised. However, the computer simulator will typically be too slow to consider such a strategy. Instead, we propose to model $\Delta(\boldsymbol{t})$ and choose $\widehat{\boldsymbol{\theta}}$ as the minimum of the estimated response surface.

3.2.2 Modelling the Δ Criterion

The field observations and computer model runs can be used to compute the discrepancy criterion values $\boldsymbol{\Delta} = (\Delta(\boldsymbol{t}_1), \ldots, \Delta(\boldsymbol{t}_n))^T$, however, the smallest possible discrepancy is unlikely to be included in this set. Since the number of computer trials is limited, we also model $\Delta(\boldsymbol{t})$ as a GP at unsampled inputs. The GP model treats the criterion as a realization of a random function by modelling the joint distribution of the observed criteria values as $\boldsymbol{\Delta} \sim N\left(\mu_\Delta, \sigma^2_\Delta \mathbf{R}_\Delta(\boldsymbol{\gamma}_\Delta)\right)$, where μ_Δ is the mean of the discrepancy criterion, σ^2_Δ is the variance and the correlation matrix is usually taken to have anisotropic squared-exponential form parametrized by $\boldsymbol{\gamma}_\Delta$. Essentially, we are treating $\Delta(\boldsymbol{t})$ as a computer model in its own right, and using the usual approach (Sacks et al., 1989a,b) for interpolation.

At some new value of the calibration parameter, $\boldsymbol{t}^*$, the criterion can be predicted. Denote this prediction as $\widehat{\Delta}(\boldsymbol{t}^*)$ (2.4) with mean squared prediction error $s^2_\Delta(\boldsymbol{t}^*)$ (2.5) as derived earlier in Chapter 2. Our point estimate of the calibration parameter is then found as

$$\widehat{\boldsymbol{\theta}}_{(n)} := \arg\min_{\boldsymbol{t}} \ \widehat{\Delta}(\boldsymbol{t}).$$

By minimizing $\widehat{\Delta}(\boldsymbol{t})$ over $\boldsymbol{t}$, we obtain an estimate of $\boldsymbol{\theta}$, but how do we assess the uncertainty in $\widehat{\boldsymbol{\theta}}$? To represent the uncertainty, we construct an empirical distribution for $\widehat{\boldsymbol{\theta}}$ by drawing J realizations (see Appendix A) of $\widehat{\Delta}(\boldsymbol{t})$, conditional on the data and estimated GP parameters, and estimating the calibration parameters for each of these realizations. This approach is demonstrated using a simulated example in Section 3.4.

To summarize, our approach consists of calculating the discrepancy criterion for each run of the computer model, constructing a predictive model for the criterion, and estimating

$\widehat{\boldsymbol{\theta}}$ as the minimum of $\widehat{\Delta}(\boldsymbol{t})$. An empirical distribution for the estimate of $\boldsymbol{\theta}$ is determined by repeatedly simulating the discrepancy criterion from $\widehat{\Delta}(\boldsymbol{t})$ and re-estimating $\boldsymbol{\theta}$ for each simulation.

3.2.3 Multivariate Responses

Up to this point, our approach has been formulated considering only a single space-time field response variable. Calibrating with a single response variable is common in many practical settings. The reason for this is the high difficulty and/or cost associated with observing the phenomena of interest. In some instances, several response variables are available which depend on the same calibration parameters. For instance, in the LFM model, both responses (energy and flux) are available to help estimate the calibration parameters.

Augmenting earlier notation, let $\mathbf{Y}_v^f$ denote the field observation of the vth response variable where $v = 1, \ldots, n_v$. Let $\mathbf{Y}_{v,i}^c$ denote the vth computer model output collected at input setting $\boldsymbol{t}_i$, where $i = 1, \ldots, n$. The discrepancy of the field observation for the vth response at the ith computer model trial is defined as $\boldsymbol{\delta}_v(\boldsymbol{t}_i) = \mathbf{Y}_v^f - \mathbf{Y}_{v,i}^c$. With a model for the cross-covariance between the discrepancies for the n_v responses, our approach would proceed by replacing the unrestricted model maximized likelihood in (3.2) with the appropriate unrestricted model maximum likelihood for the joint distribution of the n_v discrepancies. However, specifying a cross-covariance model between responses might not be straightforward. For simplicity, we treat the different responses as independent. The joint likelihood will then be computationally amenable, only depending on manipulating matrices of size $N \times N$. Accordingly, the final form of our discrepancy criterion has a simple form:

$$\Delta(\boldsymbol{t}_i) = -2 \sum_{v=1}^{n_v} log \left(\frac{L_v^*(\boldsymbol{\delta}_v(\boldsymbol{t}_i)}{L_u^*(\boldsymbol{\delta}_v(\boldsymbol{t}_i))} \right),$$

where $L_v^*(\boldsymbol{\delta}_v(\boldsymbol{t}_i))$ and $L_u^*(\boldsymbol{\delta}_v(\boldsymbol{t}_i))$ are the maximized likelihoods under the restricted measurement error model and the unrestricted GP model respectively, as described earlier.

So, extending the approach to multiple response variables is straightforward. Once the discrepancy criterion is computed as above, one can proceed by minimizing this criterion over the calibration parameters to obtain an estimate of the true parameter value.

3.2.4 Discussion

The proposed approach allows us to avoid fitting a hierarchical model and thus the related numerical issues. Instead, calibration is accomplished indirectly by transforming the problem into a optimization problem. This approach has some advantages. First, the proposed method avoids some of the computational challenges faced when fitting the standard calibration model. This advantage stems from fitting n simpler models involving $D+3$ parameters and $N \times N$ covariance matrices. Incidentally, fitting these n models can easily be done in parallel. Solving the L-dimensional minimization problem (3.3) leads to the calibration parameter estimates. The minimization occurs over a much smaller parameter space, and can be solved by existing methods. For the LFM example introduced earlier where outputs at $n = 20$ settings of the calibration inputs were available, our approach changes a problem involving a $40,824 \times 40,824$ correlation matrix and 12 parameters into 20 smaller problems, each of which only involves a 1944×1944 correlation matrix and 6 parameters.

The second main advantage of the proposed approach is the ability to implicitly handle space-time fields that exhibit non-stationarity as the calibration parameters change. This is accomplished by allowing each of the n simpler models to have unique mean and covariance parameters. This avoids a complicated model structure that explicitly accounts for non-stationarity. Instead, the proposed approach is still able to recover good estimates of the true calibration parameter setting in the presence of non-stationarity, while retaining it's computationally efficient form. The suggestion is that our method can provide improved calibration parameter estimates as compared to the standard approach which assumes stationarity. This is observed in the simulation study performed in Section 3.4.

Another aspect of the approach is the ability to simultaneously calibrate multiple responses. If a plausible model for the cross-covariances can be proposed, the approach outlined can incorporate this covariance structure. However in our experience, reasonable cross-covariance models may not be obvious. In practice, we have proceeded under the simplifying assumption that the responses are independent. In either case, because we model the difference, the discrepancies for the multiple responses will all follow the measurement error process and be independent of one another at the true calibration parameter setting.

When computing the criterion (3.2), the maximization of the likelihoods usually need be done over an unbounded domain for each parameter (i.e. the entire real line), while in practice only a subset of this parameter space is searched. Our practical experience has not indicated any problems with this approach, and indeed the consistency results for parameters in the unrestricted model, on which we rely, are often proved by assuming the true parameter value lies in a bounded subset of the overall parameter space.

Lastly, it is worth noting that there is a straightforward connection between the proposed criterion and the Kullback-Leibler (KL) Divergence (Kullback, 1997), namely $E(\Delta(\boldsymbol{t})) = \frac{1}{2}KL(\boldsymbol{\delta}(\boldsymbol{t}) \parallel \boldsymbol{\delta}(\boldsymbol{\theta}))$. Finding the value of $\boldsymbol{t}$ at which the KL is zero is equivalent to minimizing the divergence since it is a known property that $KL \geq 0$. It follows that our choice of estimate for $\boldsymbol{\theta}$ given by $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{t}} \ \widehat{\Delta}(\boldsymbol{t})$ is equivalent to minimizing the observed divergence at the observed data locations $\boldsymbol{t}_i$, and minimizing the estimated divergence otherwise.

3.3 Sequential Design

An advantage of the proposed approach is that choosing new computer trials to further refine the estimate of the calibrations parameters is fairly straightforward. Indeed, we have found it useful to run a subset of the available computer trials and then sequentially add new trials to improve the estimate of $\boldsymbol{\theta}$ using the approach proposed by Jones et al. (1998). That approach is now outlined.

For any potential new design point, $\boldsymbol{t}^*$, define the improvement function as $I = max\left(min(\boldsymbol{\Delta}) - \widehat{\Delta}(\boldsymbol{t}^*), 0\right)$, where $min(\boldsymbol{\Delta})$ is the smallest value of $\boldsymbol{\Delta}(\boldsymbol{t})$ at the observed computer trials. Design points that have a predicted discrepancy that is less than the observed minimum will have a positive improvement. The improvement is then averaged, giving the expected improvement,

$$E[I(\boldsymbol{t}^*)] = (min(\boldsymbol{\Delta}) - \widehat{\Delta}(\boldsymbol{t}^*))F\left(\frac{min(\boldsymbol{\Delta}) - \widehat{\Delta}(\boldsymbol{t}^*)}{s_\Delta(\boldsymbol{t}^*)}\right) + s_\Delta(\boldsymbol{t}^*)f\left(\frac{min(\boldsymbol{\Delta}) - \widehat{\Delta}(\boldsymbol{t}^*)}{s_\Delta(\boldsymbol{t}^*)}\right),$$

whose closed-form solution depends on $\widehat{\Delta}(\boldsymbol{t}^*)$ and $s^2_\Delta(\boldsymbol{t}^*)$ (see Jones et al. (1998)) and where F and f are the distribution and density functions of the standard normal distribution respectively. The optimal choice of a new computer trial will take place at the point, $\boldsymbol{t}_{(n+1)}$, that maximizes the expected improvement.

The maximum expected improvement criterion has a simple interpretation in this setting. Notice that $E[I(\boldsymbol{t}^*)]$ will tend to be large when a new design point gives a substantial gain on the observed minimum (i.e., $min(\boldsymbol{\Delta}) - \widehat{\Delta}(\boldsymbol{t}^*)$ is large) or when the uncertainty at the design point is large (i.e., $s^2_\Delta(\boldsymbol{t}^*)$ is large). This has the effect of favouring design points close to the current optimum (or optima) or in sparsely sampled regions of the input space. We use this approach to propose new computer trials in the upcoming simulation study and also for the application that motivated the current work.

3.4 Simulation Study

The new methodology is now demonstrated with a simple example motivated by the classical harmonic oscillator. To make this example as straightforward as possible, it is constructed so that there is one calibration parameter input and a single response variable (i.e. a curve in this case). The computer code, $\eta(t)$, models the velocity over time, x, for some particle of interest as a function of the calibration parameter t. The task is to recover a good estimate $\hat{\theta}$ of the true calibration parameter using the field observations, which are generated by

adding white noise to the computer model output.

Suppose the computer model is $\eta(t) = t\ sin(tx)$ where $x = (0, 0.5, \ldots, 10.0)$ and it is known that the calibration parameter is in the interval $[0.5, 10.5]$. The computer model response is shown in Figure 1. Suppose $n = 11$ computer trials are performed at design points $(t_1, \ldots, t_{11}) = (0.5, 1.5, \ldots, 10.5)$, respectively, giving outputs $\mathbf{Y}_1^c, \ldots, \mathbf{Y}_{11}^c$. For each simulation, a uniformly distributed calibration parameter, $\theta \in [0.5, 10.5]$, is selected. The field observations are specified as $\mathbf{Y}^f = \boldsymbol{\eta}(\theta) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N\left(0, \rho * var\left(\boldsymbol{\eta}(\theta)\right)\mathbf{I}_{11}\right)$ is the observational error having magnitude $\rho \times 100\%$ of the variance of the true response. We will vary ρ in the simulation to investigate the impact of observational error on the procedure. Thus, the computer model is simply the specified sine function, and the field observations will be simulated with randomly varying θ and adding measurement error.

The methodology is applied by constructing the discrepancy criterion based on the available data, $\mathbf{Y}^f$ and $\mathbf{Y}^c$. For each iteration of the simulation study, the current best estimate $\hat{\theta}_{(11)}$ is found as described earlier. To improve the initial estimate of the calibration parameter, $m = 2$ new runs of the computer code are obtained using the using the maximum expected improvement criterion. The new computer trials are then performed and the improved estimate, $\hat{\theta}_{(13)}$, is computed.

In contrast, the standard statistical approach (2.9) jointly models the field observations and computer outputs, $\left(\mathbf{Y}^{fT}, \mathbf{Y}^{cT}\right)^T$, and estimates the calibration parameter, $\hat{\theta}_{std}$. The point estimate for comparison is taken to be the median of the posterior distribution of the calibration parameter from a Bayesian implementation of the standard statistical model.

We performed 1000 repetitions of this procedure for each of $\rho = 0.10$ and $\rho = 0.25$. Boxplots of the difference between the calibration parameter estimate and the truth for the simulations are shown in Figure 3.3(a) and Figure 3.3(b). A quick glance at the plots reveals that the standard model appears to have some difficulty in obtaining a good estimate of the

calibration parameter. The deviations from the truth are often fairly large, and the mean squared error of the estimate was found to be 7.77 with $\rho = 0.10$ and 8.92 with $\rho = 0.25$. By comparison, we observe that the proposed approach had smaller deviations from the true calibration parameter value, and the mean squared error was calculated to be 0.0240 and 0.0228 for $\rho = 0.10$ and $\rho = 0.25$ respectively. Applying our sequential sampling strategy to obtain 2 new runs of the computer model resulted in a noticeably tighter estimate of the calibration parameter, with a resulting mean squared error of 0.0087 and 0.0082 for $\rho = 0.10$ and $\rho = 0.25$ respectively.

This simple example demonstrates the behaviour of both methods on a response surface that cannot be represented by a stationary covariance, leading to the poor performance of the standard calibration model. Nonetheless, our method performs well, providing good estimates of the calibration parameter without requiring a complex model that directly incorporates this non-stationarity into its structure.

3.5 Modelling the Earth's Magnetosphere

We now return to the LFM model that motivated the current chapter. The field observations are single space-time fields representing a geomagnetic storm over the Earth's northern polar region for a period of about 2 hours, as shown in Figure 3.1(a) and Figure 3.1(b) respectively. In addition, an $n = 20$ run space filling design was constructed over a plausible range of the calibration parameters, giving computer model runs $\mathbf{Y}_1^c, \ldots, \mathbf{Y}_{20}^c$ at locations $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{20}$. There is additional interest in seeing if the calibration parameter estimates agree for the differing response fields, flux and energy. Therefore, we use the proposed approach to estimate the calibration parameter in three settings: (i) calibrating on flux only; (ii) calibrating on energy only; and (iii) calibrating on both flux and energy. Thus, for each discrepancy field, the Δ criterion was computed for each computer model run, giving $\boldsymbol{\Delta}^{flux} = \big(\Delta^{flux}(\boldsymbol{t}_1), \ldots, \Delta^{flux}(\boldsymbol{t}_{20})\big)$, and similarly for $\boldsymbol{\Delta}^{energy}$. Note that the criterion in Section 3.2.3 simplifies to $\boldsymbol{\Delta}^{flux,energy} = \boldsymbol{\Delta}^{flux} + \boldsymbol{\Delta}^{energy}$, and is used for calibrating on flux

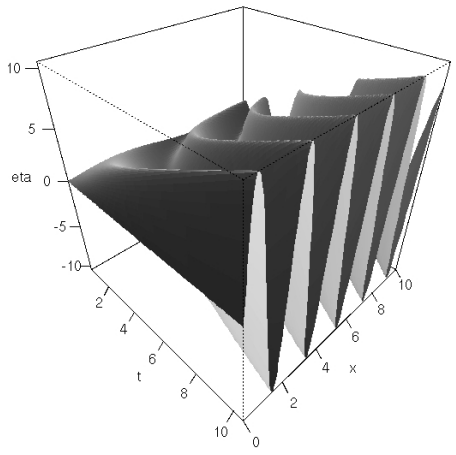
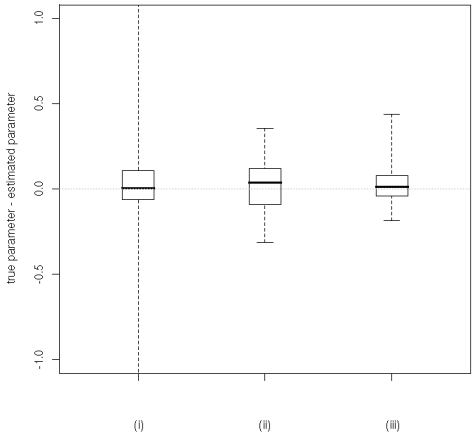


Figure 3.2: Response surface for velocity (η) as a function of time (x) and calibration parameter (t) for the simple harmonic oscillator.

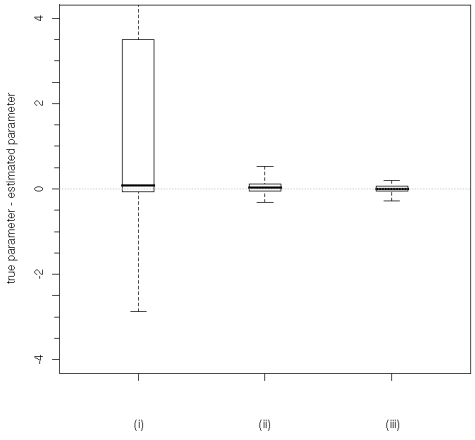
and energy simultaneously. In each case, the estimate, $\widehat{\boldsymbol{\theta}}$, is the value of $\boldsymbol{t}$ that minimizes the Δ criterion.

For each of the three cases, we also found empirical distributions for the parameter estimates using the approach described in Section 3.2.2 and also Appendix A. Specifically, 1000 realizations of the each discrepancy surface are drawn and estimates of the calibration parameter are found. The empirical univariate and bivariate distributions of the calibration parameters for each of these three settings are shown in Figures 3.4(a) and 3.4(b). The first row of each figure describes the estimates obtained using the flux response, the second row corresponds to the energy response while the third row corresponds to calibrating on both flux and energy. The red dots correspond to the locations of the $n = 20$ space-filling design for the calibration parameter settings used in the LFM computer model runs.

From these results, information about the calibration parameters can be inferred. For



(a) Raw deviations when $\rho = 0.10$. The whiskers in (i) extend to -4.44 and 8.62.

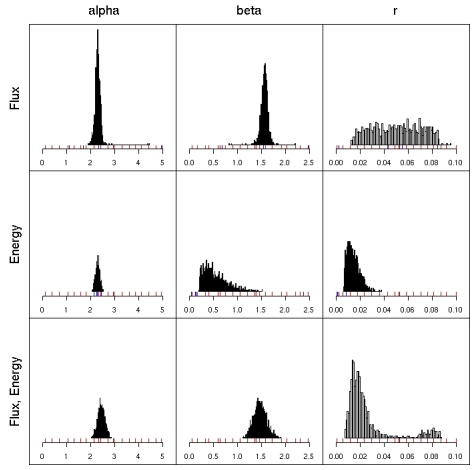


(b) Raw deviations when $\rho = 0.25$. The upper whisker in (i) extends to 8.38.

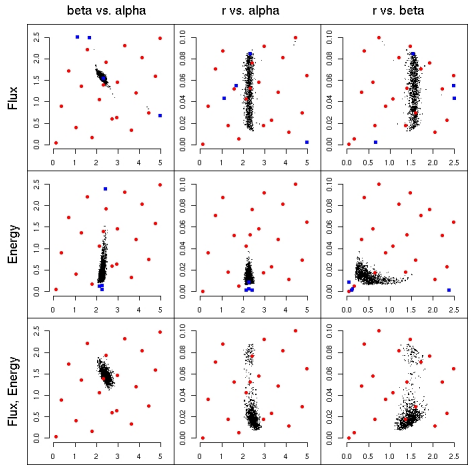
Figure 3.3: Resulting raw deviations of the calibration parameter estimate from the truth for (i) the standard statistical model, (ii) the proposed discrepancy model and (iii) the discrepancy model after $m = 2$ additional computer code runs obtained using the proposed sequential sampling strategy.

instance, the α parameter seems to be well estimated at a similar level for each of the models considered. On the other hand, the mode of the β parameter is noticeably lower when using the energy response as compared to the other responses. However, the β distributions of all three approaches do overlap. For the flux response, it appears that the r parameter has little effect as its distribution is fairly uniform over the range investigated. Instead, the r parameter seems to be much more constrained for the energy response, and again the univariate distributions of r overlap for all three models.

See that jointly calibrating appears to provide more constrained estimates of β and r than simply considering either flux alone or energy alone. In addition, the estimates seem reasonable when compared to the estimates obtained using flux or energy alone. For instance, the β distribution has been shifted left when compared to the distribution using flux, while its spread has increased, and the r distribution is very close to that arrived at using energy alone, although there is a very slight presence of bimodality in the right tail.



(a) Red ticks on the rug correspond to values of the calibration parameters where computer model output was obtained, while the blue ticks represent the 4 new locations suggested by our sequential sampling approach.



(b) Red dots represent the values of the calibration parameters where computer model output was obtained, while the blue squares represent the 4 new locations suggested by our sequential sampling approach.

Figure 3.4: (a) Univariate and (b) bivariate distributions of α, β, r constructed from 1000 conditional simulations of $\widehat{\Delta}^{flux}$, $\widehat{\Delta}^{energy}$ and $\widehat{\Delta}^{flux,energy}$.

The behaviour of the bivariate distributions follow similar trends.

To improve the estimates of the calibration parameters, our sequential sampling strategy was applied to select new settings of the calibration parameters at which to perform new runs of the computer model. For this example, $m = 4$ additional runs were desired using the energy and flux outputs, respectively. The calibration parameter settings that maximize the expected improvement are shown as blue ticks on the rug of the univariate distributions of Figure 3.4(a), and as blue squares in the bivariate distributions of Figure 3.4(b). The locations of these new points demonstrate that the sampling methodology balances between selecting new locations which may yield a new maxima of the criterion predictor $\hat{\Delta}(\boldsymbol{t})$, and locations where the uncertainty in this predictor is large.

3.6 Discussion

The computer model calibration method outlined is more flexible than the standard calibration model for many real world problems when the data to be modelled is a large space-time field or exhibits non-stationarity in the field as the calibration parameters are varied. By modelling the distribution of the difference between field observations and computer model outputs rather than the joint distribution of the field observations and computer model outputs, the methodology becomes computationally feasible for the larger space-time geophysical fields of interest. In addition, our criterion approach has a simple sequential design strategy for refining our estimate of the calibration parameter. This is perhaps more desirable than the standard calibration model, where a closed form for the estimate of the calibration parameter cannot be expressed, making a good approach for follow-up designs to likely be very computationally challenging.

The proposed criterion has both a simple interpretation and connections to many other common criteria used in comparing distributions. It can be interpreted as the log odds in favour of the assumed residual noise model over the unrestricted model. The Δ criterion is the well known log-likelihood ratio statistic, a common measure of goodness of fit. If the maximized likelihoods are replaced with marginal likelihoods from a Bayesian analysis of the differences, then the criterion can be viewed as a Bayes Factor (Schabenberger and Gotway, 2005). Averaging the criterion over the observed data leads to the Kullback-Liebler divergence (Kullback, 1997), another common measure of distributional dissimilarity. Here, our use of the Δ criterion as a means for sequential design is in the same spirit as the Kullback-Liebler based criterion used in deciding the equivalence and orthogonality of Gaussian measures in the infill asymptotics framework of Stein (1999).

The motivation for using the Δ criterion as described is its easy implementation for the practitioner, as well as the relative ease with which it can be computed. At the same time, in the problems investigated it appears to behave very reasonably. The advantage in computational efficiency comes from only having to invert $N \times N$ correlation matrices in

the calculation of the likelihood rather than the $N(n+1) \times N(n+1)$ correlation matrices in the standard calibration model.

In addition to the computational advantages, our method is also more flexible in accommodating the presence of non-stationarity. This flexibility results from the modelling approach taken, which allows the model parameters to vary as the calibration parameter is varied.

In terms of predicting the physical process, our view is that typically a practitioner is only interested in making predictions at or nearby the true value of the calibration parameter. One could then view our methodology as a means of finding a good region in the calibration parameter space where it will be more reasonable to model the full joint distribution using the standard calibration model in a follow-up experiment.

In the next chapter, we will develop more technical results, including Theorem 1, which support the usage of the proposed criterion in the calibration scenario we have outlined in this chapter.

Chapter 4

Theoretical Support for Fast Calibration

4.1 A Linear Covariance Model

Proving that the method described in Chapter 2 results in a consistent estimator of the calibration parameter is quite difficult in general due to the need to manipulate the dense covariance matrix that is typical for spatial models. For instance, the squared exponential covariance model used so far does not have a known eigendecomposition to our knowledge, making it difficult to show the consistency of our method through direct manipulation of the proposed discrepancy criterion. So, first we consider the much simplified case of a linear covariance model, and for such a model we are able to completely derive the asymptotic consistency of our calibration parameter estimator.

Suppose the true difference process at a particular setting, $\boldsymbol{t}$, of the calibration parameter setting is $\mu = \eta(\boldsymbol{\theta}) - \eta(\boldsymbol{t})$. We will assume that μ is continuous and integrable. The difference process is observed at N space-time locations with normally distributed measurement error, so that the observed difference, $\boldsymbol{\delta}$, can be represented in vector notation as $\boldsymbol{\delta} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_N)$. Think of this as the mechanism that generates the observed difference process. We can then investigate the behaviour of the proposed calibration approach when a

simple covariance model for the unrestricted model is used. Also, the N space-time locations are sampled on a regular grid over the fixed domain of interest, so that our investigation corresponds to the case of dense infill sampling.

Consider a linear covariance for the unrestricted model of the observed difference with zero mean, $\boldsymbol{\delta} \sim \mathcal{N}\left(0, \sigma^2(\mathbf{I}_N + \gamma\mathbf{R}(\alpha,\beta))\right)$, where $\mathbf{R}(\alpha,\beta) = \alpha\mathbf{1}\mathbf{1}^T + \beta\mathbf{V}\mathbf{V}^T$, for some $\mathbf{V} \perp \mathbf{1}$, where $\perp$ means orthogonal. It turns out that we can write the eigenstructure of this matrix directly, giving

$$\mathbf{I}_N + \tilde{\alpha}\mathbf{1}\mathbf{1}^T + \tilde{\beta}\mathbf{V}\mathbf{V}^T = \quad \mathbf{U}\boldsymbol{\Omega}\mathbf{U}^T,$$

where $\tilde{\alpha} = \gamma\alpha, \tilde{\beta} = \gamma\beta, \mathbf{U} = \left[\frac{\mathbf{1}}{\sqrt{N}}, \frac{\mathbf{V}}{||\mathbf{V}||}, \mathbf{e}_{N-2}\right]$, where $\mathbf{e}_{N-2}$ are $N-2$ arbitrary eigenvectors orthogonal to the first two (and eachother), and $\boldsymbol{\Omega} = diag(\tilde{\alpha}+1, \tilde{\beta}+1, \mathbf{1}_{N-2})$. From this decomposition, it follows that

$$\begin{aligned} log|\mathbf{I}_N + \tilde{\alpha}\mathbf{1}\mathbf{1}^T + \tilde{\beta}\mathbf{V}\mathbf{V}^T| = &\quad log(\tilde{\alpha}+1) + log(\tilde{\beta}+1), \\ (\mathbf{I}_N + \tilde{\alpha}\mathbf{1}\mathbf{1}^T + \tilde{\beta}\mathbf{V}\mathbf{V}^T)^{-1} = &\quad \mathbf{U}\boldsymbol{\Omega}^{-1}\mathbf{U}^T \end{aligned}$$

and

$$\boldsymbol{\delta}^T(\mathbf{I}_N + \tilde{\alpha}\mathbf{1}\mathbf{1}^T + \tilde{\beta}\mathbf{V}\mathbf{V}^T)^{-1}\boldsymbol{\delta} = \quad \tfrac{d_1^2}{\tilde{\alpha}+1} + \tfrac{d_2^2}{\tilde{\beta}+1} + \textstyle\sum_{i=3}^N d_i^2,$$

where $\mathbf{d} = \boldsymbol{\delta}^T\mathbf{U}$. Using these equations, it is possible to analytically solve for the MLE's of σ^2, $\tilde{\alpha}$ and $\tilde{\beta}$. Combining those results with the MLE of σ^2 under the restricted model, we arrive at the following expression for the criterion (3.2):

$$\begin{aligned} \Delta = \quad & Nlog\left(1 - \tfrac{2}{N} + \frac{\frac{S_1}{N\sigma^2}}{\frac{T}{(N-2)\sigma^2} - \frac{S_1}{(N-2)\sigma^2} - \frac{S_2}{(N-2)\sigma^2}} + \frac{\frac{S_2}{N\sigma^2}}{\frac{T}{(N-2)\sigma^2} - \frac{S_1}{(N-2)\sigma^2} - \frac{S_2}{(N-2)\sigma^2}}\right) \\ & -log\left(\frac{\frac{S_1}{\sigma^2}}{\frac{T}{(N-2)\sigma^2} - \frac{S_1}{(N-2)\sigma^2} - \frac{S_2}{(N-2)\sigma^2}}\right) - log\left(\frac{\frac{S_2}{\sigma^2}}{\frac{T}{(N-2)\sigma^2} - \frac{S_1}{(N-2)\sigma^2} - \frac{S_2}{(N-2)\sigma^2}}\right), \end{aligned} \tag{4.1}$$

where $T = \sum_{i=1}^N d_i^2, S_1 = d_1^2$ and $S_2 = d_2^2$. Note that under the data-generating process, it is easy to see that these statistics, normalized by the variance, all have non-central χ^2

distributions. Namely,

$$\begin{aligned}
\frac{S_1}{\sigma^2} &\sim \chi^2_{nc}\left(df = 1, \lambda_1 = \frac{\left(\sum_{i=1}^N \mu_i\right)^2}{N\sigma^2}\right), \\
\frac{S_2}{\sigma^2} &\sim \chi^2_{nc}\left(df = 1, \lambda_2 = \frac{\left(\sum_{i=1}^N \mu_i V_i\right)^2}{\sigma^2 \sum_{i=1}^N V_i^2}\right), \\
\frac{T}{\sigma^2} &\sim \chi^2_{nc}\left(df = N, \lambda_T = \frac{\sum_{j=1}^N \left(\sum_{i=1}^N \mu_i U_{ji}\right)^2}{\sigma^2}\right).
\end{aligned}$$

Using these facts, we arrive at the following two lemmas.

Lemma 1: Under the data-generating mechanism, we have that

$$\begin{aligned}
\frac{S_1}{N\sigma^2} &\longrightarrow \Lambda_1 = \frac{\left(\int \mu\right)^2}{\sigma^2}, \\
\frac{S_1}{(N-2)\sigma^2} &\longrightarrow \Lambda_1 = \frac{\left(\int \mu\right)^2}{\sigma^2}, \\
\frac{S_2}{N\sigma^2} &\longrightarrow \Lambda_2 = \frac{\left(\int \mu V\right)^2}{\sigma^2 ||V||^2}, \\
\frac{S_2}{(N-2)\sigma^2} &\longrightarrow \Lambda_2 = \frac{\left(\int \mu V\right)^2}{\sigma^2 ||V||^2},
\end{aligned}$$

and

$$\frac{T}{(N-2)\sigma^2} \longrightarrow \Lambda_T = 1 + \frac{\int \mu^2}{\sigma^2}$$

in probability.

Proof: We will prove the case for $\frac{S_1}{N\sigma^2}$, the rest follow similarly. Note that $lim_{N\to\infty} E\left[\frac{S_1}{N\sigma^2}\right] = lim_{N\to\infty} \frac{1}{N} + \frac{\left(\sum_{i=1}^N \mu_i\right)^2}{N^2\sigma^2} \to \frac{\left(\int \mu\right)^2}{\sigma^2}$ and $lim_{N\to\infty} Var\left[\frac{S_1}{N\sigma^2}\right] = lim_{N\to\infty} \frac{4\left(\int \mu\right)^2}{N\sigma^2} \to 0$. The

lemma then follows by the usual application of the Cauchy-Swartz inequality:

$$\begin{aligned} lim_{N\to\infty} P\left(\left(\frac{S_1}{N\sigma^2} - \frac{\left(\int \mu\right)^2}{\sigma^2}\right)^2 \geq c^2\right) &\leq lim_{N\to\infty} \frac{E\left[\left(\frac{S_1}{N\sigma^2} - \frac{(\int \mu)^2}{\sigma^2}\right)^2\right]}{c^2} \\ &= \lim_{N\to\infty} \frac{V\left[\frac{S_1}{N\sigma^2}\right]}{c^2} \\ &\to 0 \end{aligned}$$

Lemma 2: Under the data-generating mechanism, we have that $\frac{S_1}{\sigma^2} = \mathcal{O}_p\left(1 + \lambda_1\right)$ and $\frac{S_2}{\sigma^2} = \mathcal{O}_p\left(1 + \lambda_2\right)$.

Proof: Recall that $\frac{S_1}{\sigma^2} = \mathcal{O}_p\left(1 + \lambda_1\right)$ iff for all $\epsilon > 0$ there exists $C < \infty$ such that for all N, $P\left[\frac{|\frac{S_1}{\sigma^2}|}{|1+\lambda_1|} \geq C\right] \leq \epsilon$. But by Markov's inequality, take $\epsilon = \frac{E\left[\frac{\frac{S_1}{\sigma^2}}{1+\lambda_1}\right]}{C} = \frac{1}{C}$, so taking $C = \frac{1}{\epsilon}$ completes the proof. The proof for $\frac{S_2}{\sigma^2}$ similarly follows.

Based on Lemma 1 and Lemma 2, we then have that

$$\Delta \longrightarrow Nlog\left(1 - \tfrac{2}{N} + \tfrac{\lambda_1}{\Lambda_T - \Lambda_1 - \Lambda_2}\right) - log\left(\tfrac{1+\lambda_1}{\Lambda_T - \Lambda_1 - \Lambda_2}\right) - log\left(\tfrac{1+\lambda_2}{\Lambda_T - \Lambda_1 - \Lambda_2}\right). \tag{4.2}$$

First, we note that the first term in (4.2) dominates when N is large. Next, we can see that if $\mu = 0$, as would occur under our assumptions if we are at the true calibration parameter setting, then $\Delta \to -2$ in probability. On the other hand, when $\mu \neq 0$, then generally $\Delta \to \infty$. However, there is another possibility that indicates the limitation of using such a simple linear covariance model. Namely, if $\mu \perp \mathbf{1}$ and $\mu \perp \mathbf{V}$, then we will have $\Delta \to 2log(\Lambda_T)$. Therefore, the implicit assumption is either that the eigenfunctions defined by the covariance matrix are enough to capture most of the variability of μ, or better, that the covariance function is the true covariance of the observed process μ. Another way of stating the latter is making the assumption that the unrestricted model for the difference process is correct.

4.2 Gaussian Process Field Assumption and Theorem 1

As mentioned earlier, in general it is not possible, or at least not very easy, to analyse the behaviour of the criterion (3.2) as we did in the previous section. This is because the covariance functions used in practice, such as the squared exponential, usually don't have an explicit eigendecomposition. However, there have been some results showing the asymptotic consistency of parameters for these models in at least some cases. Motivated by the previous section, let us assume that the unrestricted model for the difference process is correct, and consider what happens if we do have consistent estimators for the unrestricted model's parameters.

For simplicity, we will develop the proof assuming a zero-mean term in the unrestricted model, $\boldsymbol{\mu}(t_i) = \mathbf{0}$. In this case, our criterion becomes

$$\Delta = N log \left(\frac{\frac{\boldsymbol{\delta}^T \boldsymbol{\delta}}{N}}{\frac{\boldsymbol{\delta}^T (\mathbf{I}_N + \hat{\gamma}\mathbf{R}(\hat{\phi}))^{-1} \boldsymbol{\delta}}{N}} \right) - log|\mathbf{I}_N + \hat{\gamma}\mathbf{R}(\hat{\phi})|,$$

where $\hat{\gamma}$ and $\hat{\phi}$ are the MLE's from the unrestricted model. In what follows, suppose the difference process $\boldsymbol{\delta}$ is observed over a fixed 1-dimensional domain $\mathcal{D}$ at N observation locations $x_1, \ldots, x_N$ forming an equidistant grid. Write $\boldsymbol{\delta} = \boldsymbol{f} + \boldsymbol{\epsilon}$ where $\boldsymbol{f} = \boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(t)$ and assume $\int_{x \in \mathcal{D}} f^2(x) dx < \infty$. Furthermore, assume that $\int_{x \in \mathcal{D}} f^2(x) dx = 0$ if and only if $t = \theta$. We will derive a proof demonstrating that as the number of sampling locations $N \to \infty$, $\Delta(t) \longrightarrow \infty$ when $t \neq \theta$ and $\Delta(t) \longrightarrow 0$ when $t = \theta$, assuming that conditional on t the unrestricted model for the observations is correct. Then, by the continuous mapping theorem, it follows that $\hat{\theta} \longrightarrow \theta$ in probability.

First let us state the following lemmas:

Lemma 3: Note that $\frac{1}{N}\boldsymbol{\delta}^T\boldsymbol{\delta} = \frac{1}{N}\sum_{i=1}^N f(x_i)^2 + \frac{1}{N}\sum_{i=1}^N \epsilon_i^2 + \frac{2}{N}\sum_{i=1}^N f(x_i)\epsilon_i \longrightarrow \int f(x)^2 dx + \sigma^2$ in probability.

Proof:

First, it is straightforward that $\frac{1}{N}\sum_{i=1}^{N} f(x_i)^2 \longrightarrow \int_{x\in\mathcal{D}} f(x)^2 dx$ and $\frac{1}{N}\sum_{i=1}^{N}\epsilon_i^2 \longrightarrow \sigma^2$ in probability. For the third term, note that $E\left[\frac{2}{N}\sum_{i=1}^{N} f(x_i)\epsilon_i\right] = 0$ and $Var\left[\frac{2}{N}\sum_{i=1}^{N} f(x_i)\epsilon_i\right] = \frac{4\sigma^2}{N}\frac{\sum_{i=1}^{N} f(x_i)^2}{N} \longrightarrow 0$ as $N \longrightarrow \infty$ by the finiteness of $\int_{x\in\mathcal{D}} f(x)^2 dx$. It follows by the Cauchy-Schwartz inequality that $\frac{2}{N}\sum_{i=1}^{N} f(x_i)\epsilon_i \longrightarrow 0$ in probability, thus proving the lemma.

Lemma 4: Let λ_i be the ith eigenvalue of the matrix $\mathbf{R}(\hat{\phi})$. Then, $-log|\mathbf{I}_N + \hat{\gamma}\mathbf{R}(\hat{\phi})| \geq -\hat{\gamma}\sum_{i=1}^{N}\lambda_i$.

Proof:

Note that $|\mathbf{I}_N + \hat{\gamma}\mathbf{R}(\hat{\phi})| = \prod_{i=1}^{N}(1+\hat{\gamma}\lambda_i)$. Recall the common logarithm inequality $log(a) \leq a-1$ for some $a > 0$. Then,

$$\begin{aligned} -log|\mathbf{I}_N + \hat{\gamma}\mathbf{R}(\hat{\phi})| = & -\sum_{i=1}^{N} log(1+\hat{\gamma}\lambda_i) \\ \geq & -\hat{\gamma}\sum_{i=1}^{N}\lambda_i. \end{aligned}$$

We now prove the main result.

Proof of Theorem 1:

First we consider the case when $t \neq \theta$. By Lemma 4, we have

$$\Delta \geq Nlog\left(\frac{\frac{\boldsymbol{\delta}^T\boldsymbol{\delta}}{N}}{\frac{\boldsymbol{\delta}^T(\mathbf{I}_N+\hat{\gamma}\mathbf{R}(\hat{\phi}))^{-1}\boldsymbol{\delta}}{N}}\right) - \hat{\gamma}\sum_{i=1}^{N}\lambda_i.$$

Under our assumptions, we know that the denominator in the first term, $\frac{\boldsymbol{\delta}^T(\mathbf{I}_N+\hat{\gamma}\mathbf{R}(\hat{\phi}))^{-1}\boldsymbol{\delta}}{N}$, is the MLE of the variance under the unrestricted model and converges in probability to σ^2. Also, by Lemma 3, we know that $\frac{1}{N}\boldsymbol{\delta}^T\boldsymbol{\delta} \longrightarrow \int_{x\in\mathcal{D}} f(x)^2 dx + \sigma^2$ in probability. So the *log* term converges in probability to $log\left(\frac{\int_{x\in\mathcal{D}} f(x)^2 dx}{\sigma^2} + 1\right)$. Now consider the second term. Let $\Lambda = \sum_{i=1}^{\infty}\lambda_i < \infty$. Here, the finiteness of $\sum_{i=1}^{\infty}\lambda_i$ is a requirement of the infinite-dimensional correlation operator, see for instance Ibragimov and Rozanov (1978). Therefore, the second term converges in probability to $-\gamma\Lambda$. Hence, by the assumption

$\int_{x\in\mathcal{D}} f^2(x)dx > 0$ when $t \neq \theta$, the quantity in the log is greater than 1 and so $\Delta \to \infty$ as $N \to \infty$.

On the other hand, when $t = \theta$ we have $\hat{\gamma} \longrightarrow 0$ by the consistency of the unrestricted model parameters. So, from Lemma 3 and Lemma 4, $\Delta \geq Nlog\left(\frac{\frac{\boldsymbol{\delta}^T\boldsymbol{\delta}}{N}}{\frac{\boldsymbol{\delta}^T(\mathbf{I}_N+\hat{\gamma}\mathbf{R}(\hat{\phi}))^{-1}\boldsymbol{\delta}}{N}}\right) - \hat{\gamma}\sum_{i=1}^{N}\lambda_i \longrightarrow 0$ in probability. Also, note that $log|I + \hat{\gamma}\mathbf{R}(\hat{\phi})| = \sum_{i=1}^{N} log(1 + \hat{\gamma}\lambda_i) \geq 0$. Then $\Delta \leq Nlog\left(\frac{\frac{\boldsymbol{\delta}^T\boldsymbol{\delta}}{N}}{\frac{\boldsymbol{\delta}^T(\mathbf{I}_N+\hat{\gamma}\mathbf{R}(\hat{\phi}))^{-1}\boldsymbol{\delta}}{N}}\right) \longrightarrow 0$ in probability by Lemma 3 and the consistency of the unrestricted model parameters. Therefore, $\Delta \longrightarrow 0$ in probability when $t = \theta$, proving the theorem.

4.3 Discussion

The result from Theorem 1 indicates that if the parameters in the unrestricted model can be consistently estimated, then our methodology will provide a consistent estimate of the unknown calibration parameter θ. For the consistency requirements on $\hat{\gamma}$ and $\hat{\phi}$ in the proof of Theorem 1, it has been shown that they can be satisfied for the popular covariance models used in practice in greater than one dimension in some situations. For instance, Ying (1991) shows the consistency of the product $\gamma\phi$ in 1-D for the exponential covariance model. The consistency of the parameters for a 2-D exponential covariance model were later shown in Ying (1993). In the presence of measurement error, in 1-D the product $\widehat{\gamma\phi}$ and the measurement error $\hat{\sigma}^2$ is consistent (Chen et al., 2000; Zhang, 2004). Similar results are shown for the error-free Matern covariance model by Zhang (2004) in the 1-D case. Finally, Stein (1999) suggests that the measurement error σ^2 is always consistently estimable, and the parameters that were consistent in the error-free case can still be estimated consistently in the presence of measurement error.

The more general results of Stein (1999) suggest that the parameters for these covariance models should be consistently estimable in arbitrary dimensions as long as the parameters satisfy certain microergodicity conditions. In 1-D cases where only the product $\gamma\phi$ is consistent, we believe our method remains valid since the product can only converge to zero

when $\hat{\gamma}$ goes to zero. This issue might also be improved when there are multiple fields being calibrated, or if there are replicate observations.

The result found using a particular linear covariance model is more intuitive to interpret. It suggests that if the eigenbasis resulting from the covariance model is able to capture the variability of the process, then our method should be able to estimate the calibration parameter. Our practical experiences described in the previous chapter suggest that the method works quite well in general.

Chapter 5

Calibration with Derivative Information

Our focus now changes to developing a Bayesian statistical model that augments the usual calibration framework (2.9) to allow for the possible observation of some or all of the partial derivatives of the computer model. Initially, our approach directly extends the standard Bayesian calibration model of Kennedy and O'Hagan (2001) in the setting of no bias, i.e. $b(\boldsymbol{x}) = 0$. Incorporating derivatives into statistical modelling has increasingly become popular in various areas of the statistical literature. The motivation can sometimes be ad-hoc, for instance penalizing the second derivative in a cubic spline regression to arrive at a smooth predictor that does not "over-fit" the data (Ramsay and Silverman, 2005). Another situation where derivatives play a role is when knowledge of the derivative process is known a priori. The penalized regression methods of Ramsay and Silverman (2005) allow one to form a predictor that accounts for this knowledge. Finally, the process being observed may be hypothesized to be a noisy observation of the solution trajectory of a system of ordinary differential equations or partial differential equations. These differential equations are governed by some unknown parameters, $\boldsymbol{\theta}$, which are to be estimated (Ramsay et al., 2007). This is the setting that is most synonymous with the computer model calibration problem discussed in this thesis.

In the approach we present, again assume that at some level the computer model is being treated as a black-box. That is, we don't explicitly rely on knowledge of the derivative equations, but we want to be able to account for derivative information in the model to improve the ability to estimate the calibration parameters $\boldsymbol{\theta}$ and make more accurate predictions of the process. This approach has a different scope of application than the explicit methods such as Ramsay et al. (2007) which make direct use of the differential equations. For instance, derivative information can be extracted from the computer model by using automatic differentiation (Griewank and Walther, 2008). In other scenarios, the computer model already outputs derivatives, and one would like to make use of that information in a calibration experiment rather than throw it away. Sometimes the computer model and its derivatives may be high-dimensional fields, which is also generally easier to accommodate with our GP-based approach. But perhaps the most relevant situation stems from the reason calibration models are primarily used in the first place: to borrow information from the computer model for estimating $\boldsymbol{\theta}$ and predicting the field process because the field process can only be sparsely sampled. Hence our methodology applies to a very real scenario where other methods may have difficulties. Nonetheless, let us briefly outline some connections between the GP models used here and penalized regression.

5.1 Connection to Penalized Regression

As alluded to earlier, there is an interesting known connection between the GP model used in the proposed calibration methodology and penalized spline smoothing approaches. The latter attempt to smooth the field data while taking into account a known differential operator $\mathcal{L}$ of which the computer model response η is a solution. Think of $\mathcal{L}$ as a residual making operator. As a simple example, for the linear differential equation $\frac{d\eta(s)}{ds} = g(\theta)h(s)$, then $\mathcal{L}\eta(s) := \frac{d\eta(s)}{ds} - g(\theta)h(s)$ which will equal zero if $\eta(s)$ solves the differential equation.

In the penalized regression approach, the data $y(s)$ observed at location s is modelled as $y(s) = \boldsymbol{c}^T\boldsymbol{\psi}(s) + \epsilon(s)$ where $\boldsymbol{c}$ is a $m \times 1$ coefficient vector and $\boldsymbol{\psi}$ is a $m \times 1$ vector of

the basis functions evaluated at the location s, while $\epsilon(s)$ is the measurement error term, which is usually i.i.d. $\mathcal{N}(0, \frac{1}{\lambda_f})$ for all s (Ramsay and Silverman, 2005). Here, we use a general location index s to avoid confusion with the models developed in previous chapters. The basis functions $\boldsymbol{\psi}$ are chosen a priori, and an appropriate penalty is applied so that the regression coefficients, $\boldsymbol{c}$, make an appropriate trade-off between fitting the observed data and satisfying the penalty associated with the differential equation of interest. The fitting criterion used is

$$J = \lambda_f \left(\mathbf{Y} - \boldsymbol{\Psi}\boldsymbol{c}\right)^T \left(\mathbf{Y} - \boldsymbol{\Psi}\boldsymbol{c}\right) + \lambda_z \boldsymbol{c}^T \mathbf{K} \boldsymbol{c},$$

where $\boldsymbol{\Psi}$ is the $N \times m$ matrix of the basis functions evaluated at observation locations $s_1, \ldots, s_N$ (Ramsay and Silverman, 2005). The first term is proportional to the usual log-likelihood for the observed data in a normal basis regression model. It has been shown (Ramsay and Silverman, 2005) that the second term, called the penalty term, arises from the differential operator $\mathcal{L}$ associated with the differential equation, namely

$$\begin{aligned} \int (\mathcal{L}y(s))^2 ds &= \boldsymbol{c}^T \left[\int \mathcal{L}\boldsymbol{\psi}(s) \mathcal{L}\boldsymbol{\psi}(s)^T ds\right] \boldsymbol{c} \\ &= \boldsymbol{c}^T \mathbf{K} \boldsymbol{c}. \end{aligned}$$

The parameter λ_z serves as a trade-off between fitting the data and fitting the model defined by the differential operator: as $\lambda_z \to 0$, we fit the data at the expense of model bias, while as $\lambda_z \to \infty$, we fit the model, at the expense of increased error variance. In this framework, the unknown parameter of interest (i.e. the calibration parameter) is contained in the equation defining the differential operator, so we can re-write $\mathcal{L} = \mathcal{L}_\theta$ for clarity. The generalized profiling methodology of Ramsay et al. (2007) enables one to fit such a model and arrive at estimates of the parameters $\boldsymbol{c}$, $\boldsymbol{\theta}$ and λ_f, while λ_z is typically chosen via cross-validation.

One can write a Bayesian model similar to the above. Consider the model $y(s) = z(s) + \epsilon(s)$ where $z(s) = \boldsymbol{c}^T \boldsymbol{\psi}(s) = \sum_{j=1}^N c_j k(s, s_j)$ (i.e. we take $\psi_j(s) = k(s, s_j)$ and we have the same number of basis functions as observations, so $m = N$ in this setup). Next, we place a prior on $\boldsymbol{c}$, namely $\boldsymbol{c} \sim \mathcal{N}(0, \frac{1}{\lambda_z}\mathbf{K}^{-1})$, where $\mathbf{K}_{ij} = k(s_i, s_j) = \psi_j(s_i)$. Ignoring

constant terms, we have

$$-l_{\boldsymbol{c}|\mathbf{Y}} \propto -l_{\mathbf{Y},\boldsymbol{c}} = -l_{\mathbf{Y}|\boldsymbol{c}} - l_{\boldsymbol{c}} \quad = \lambda_f \left(\mathbf{Y}-\mathbf{Z}\right)^T \left(\mathbf{Y}-\mathbf{Z}\right) + \lambda_z \boldsymbol{c}^T \mathbf{K} \boldsymbol{c}$$
$$= \lambda_f \left(\mathbf{Y}-\boldsymbol{\Psi}\boldsymbol{c}\right)^T \left(\mathbf{Y}-\boldsymbol{\Psi}\boldsymbol{c}\right) + \lambda_z \boldsymbol{c}^T \mathbf{K} \boldsymbol{c},$$

which is the same as the penalized regression criterion J from above (Nychka, 2000). Note that in this setup, the covariance kernel $k(\cdot,\cdot)$ needs to be chosen to satisfy the differential operator $\mathcal{L}_{\boldsymbol{\theta}}$, so we can re-write it as $k = k_{\boldsymbol{\theta}}$ for clarity.

In the case of GP regression models, this is abstracted further, so that we consider only a prior on $\mathbf{Z}$ rather than expressing it as a combination of particular basis functions. If the prior is taken to be $\mathbf{Z} \sim \mathcal{N}(0, \frac{1}{\lambda_z}\mathbf{K})$, then again ignoring constants,

$$-l_{\mathbf{Z}|\mathbf{Y}} \propto -l_{\mathbf{Y},\mathbf{Z}} = -l_{\mathbf{Y}|\mathbf{Z}} - l_{\mathbf{Z}} = \lambda_f \left(\mathbf{Y}-\mathbf{Z}\right)^T \left(\mathbf{Y}-\mathbf{Z}\right) + \lambda_z \mathbf{Z}^T \mathbf{K}^{-1} \mathbf{Z}.$$

Ideally the covariance $\mathbf{K}$ is again chosen according to the differential operator $\mathcal{L}_{\boldsymbol{\theta}}$. In the practical application of GP regression, the model is chosen to give predictions which behave reasonably 'nice'. For instance, by choosing the exponential covariance model, one gets predictions which correspond to a continuous but nowhere differentiable stochastic process. Choosing the cubic covariance model leads to predictions which correspond to a process satisfying an operator that penalizes the second derivative, i.e. $\mathcal{L} = \int (\frac{\partial^2}{ds^2} z(s))^2 ds$. The very commonly used squared exponential covariance model leads to predictions corresponding to a process that is infinitely differentiable.

Of course, in the practical application of GP regression, the covariance model is never chosen based on a particular differential operator $\mathcal{L}_\theta$. For instance, most operators will not yield a closed-form solution for the covariance kernel, rendering their usage impractical. Rather, the popular models such as exponential, squared exponential or Matern are chosen on a more abstract belief of the smoothness of the process because of the convenient implementation of these covariance models. This convenience is of little help if $\boldsymbol{\theta}$ is unknown and required for predicting the observed physical process. It is in such situations that the calibration model (2.9) outlined earlier is applicable. In this model setup, it is assumed that

although the computer model may be expensive to run, it is still much cheaper to sample the computer model than to sample observations of the physical process. The idea is, instead of modelling how the *covariance* of the process varies with $\boldsymbol{\theta}$ as above, we model how the *process* varies with $\boldsymbol{\theta}$ as in (2.10) and arrive at an estimate of the calibration parameter and predictions of the observed field process in this way.

Nonetheless, in the practical case of finite data, one may expect that using the wrong covariance in the calibration model (i.e. unrelated to $\mathcal{L}_\theta$) will affect estimates of the calibration parameter and predictions of the physical process. However, by simultaneously modelling the derivative process while still using these convenient covariance kernels, we can at least correct their behaviour to some degree. For instance, by including observations of the first derivative process, we can improve the predictions and parameter estimates arrived at by using the "wrong" covariance model. Suppose we observe these derivative observations $y'(s)$ independently of $y(s)$ and modelled as $y'(s) = z'(s) + \epsilon'(s)$ where $\epsilon'(s) \sim \mathcal{N}(0, \frac{1}{\lambda_d})$, with the prior of $\mathbf{Z'} = [z'(s_1), \ldots, z'(s_N)]$ being $\mathbf{Z'} \sim \mathcal{N}(0, \frac{1}{\lambda_z}\mathbf{K'})$, independent of $\mathbf{Z}$. In this simple independent setup, we have (ignoring constants) $-l_{\mathbf{Y},\mathbf{Y'},\mathbf{Z},\mathbf{Z'}} = -l_{\mathbf{Y}|\mathbf{Z}} - l_{\mathbf{Y'}|\mathbf{Z'}} - l_{\mathbf{Z}} - l_{\mathbf{Z'}}$.

Now for a model that is governed by a differential operator, consider the distinction between the operator induced by the covariance function chosen, $\mathcal{L}_K$, and the operator that is actually of interest, $\mathcal{L}_\theta$. Then the term $-l_{\mathbf{Y'}|\mathbf{Z'}} = \lambda_d(\mathbf{Y'} - \mathbf{Z'})^T(\mathbf{Y'} - \mathbf{Z'})$ can be viewed as a finite approximation to $\int(\mathcal{L}_\theta y(s))^2 ds$ when $\mathcal{L}_\theta$ is a first order differential operator. We still have penalties on the smoothness of $\mathbf{Z}$ and $\mathbf{Z'}$ as defined by the priors (which is a penalty based on the incorrect operator $\mathcal{L}_K$), but in this model we can correct this smoothness assumption to some degree by having also gathered some observations of the derivative process. The degree to which this is corrected is controlled by λ_d. For instance, as $\lambda_d \to \infty$ the derivative process will be interpolated.

In reality, the process and it's derivatives are not independent, and we take this into account in our calibration model. This allows us to borrow information from the derivatives

for both parameter estimation and prediction. We will see in the subsequent simulation studies of Chapter 6 reductions in prediction and parameter estimate errors, in some cases by many orders of magnitude. Next, let us introduce the proposed calibration model which incorporates derivative information.

5.2 Modelling Derivatives with Gaussian Processes

The notion of incorporating derivative information into statistical modelling techniques has received growing interest in the statistics community in recent years. This has been driven by the type of scientific problems introduced earlier in this chapter. In the framework of GP models, derivative information has received relatively less attention than in other areas of statistical modelling. Yet, we will see that the benefits of including this information when it is available can be very noticeable.

The foremost paper extending the GP model used in computer experiments to the case of derivative information is by Morris et al. (1993). In it, they adapt the results in Parzen (1962) to computer experiments that incorporate derivatives in the error-free observations setting (2.3). They specifically consider a model where the first derivatives with respect to the computer model inputs are available. While it is feasible to consider higher-order derivatives in the same way, our work in this thesis is also restricted to first derivatives.

Consider again, the unknown process $z(\boldsymbol{s})$. It has been shown (Morris et al., 1993) that

$$E[\frac{d}{ds_l}z(\boldsymbol{s})] = \frac{d}{ds_l}E[z(\boldsymbol{s})],$$

$$Cov(\frac{d}{ds_l}z(\boldsymbol{s}), z(\boldsymbol{s}')) = \frac{d}{ds_l}Cov(z(\boldsymbol{s}), z(\boldsymbol{s}')),$$

and

$$Cov(\frac{d}{ds_l}z(\boldsymbol{s}), \frac{d}{ds'_m}z(\boldsymbol{s}')) = \frac{d^2}{ds_l ds'_m}Cov(z(\boldsymbol{s}), z(\boldsymbol{s}')).$$

This allows us to easily extend the usual GP approach of computer experiments to the case where derivative information is included in the analysis. In what follows, we will use subscript notation to denote differentiation with respect to a given input. For instance, $y_0(\boldsymbol{s}) =$

$y(\boldsymbol{s})$, $y_l(\boldsymbol{s}) = \frac{d}{ds_l} y(\boldsymbol{s})$, and $K_{lm}(\boldsymbol{s}, \boldsymbol{s}') = \frac{d^2}{ds_l ds'_m} K_{00}(\boldsymbol{s}, \boldsymbol{s}')$, while $[\Sigma_{lm}]_{i,j} = K_{lm}(\boldsymbol{s}_i, \boldsymbol{s}_j)$.

Recall the squared exponential covariance (2.1) which we can re-express as

$$K_{00}(\boldsymbol{s}, \boldsymbol{s}') = \frac{1}{\lambda_z} \prod_{k=1}^{D} \phi_k^{(s'_k - s_k)^2},$$

where $\lambda_z > 0$ is the process precision and $0 < \phi_k < 1$ for all $k = 1, \ldots, D$ is the correlation parameter. With this as the covariance model for our process, one can show the partial derivative processes have covariances given by:

$$K_{l0}(\boldsymbol{s}, \boldsymbol{s}') = -2\frac{1}{\lambda_z} log(\phi_l)(s'_l - s_l) \prod_{k=1}^{D} \phi_k^{(s'_k - s_k)^2} \tag{5.1}$$

$$K_{0m}(\boldsymbol{s}, \boldsymbol{s}') = -2\frac{1}{\lambda_z} log(\phi_m)(s'_m - s_m) \prod_{k=1}^{D} \phi_k^{(s'_k - s_k)^2} \tag{5.2}$$

and,

$$K_{lm}(\boldsymbol{s}, \boldsymbol{s}') = 4\tfrac{1}{\lambda_z} log(\phi_l) log(\phi_m)(s'_l - s_l)(s'_m - s_m) \textstyle\prod_{k=1}^{D} \phi_k^{(s'_k - s_k)^2} \tag{5.3}$$

In the context of GP regression, Morris et al. (1993) report a marked improvement in prediction performance when using the model with derivative information in the sense of mean-squared prediction error (MSPE). Mitchell et al. (1994) also show that their results on space-filling designs in the usual case extend to designing a computer experiment with derivative outputs. Yet besides these early results, there does not seem to have been much further development of models which include derivative information in computer experiments. This would seem to be in contrast with other areas of statistical literature, such as the penalized likelihood methods discussed earlier, where models incorporating derivative information have recently been an active area of research. Next, we will develop a calibration model which incorporates derivatives by using the above results.

5.3 Bayesian Calibration with Derivative Information

As described in Chapter 2, our scenario consists of field observations $y^f(\boldsymbol{x})$ observed at locations $\boldsymbol{x}_i \in \mathcal{D}$ for $i = 1, \ldots, N$ and some fixed domain $\mathcal{D}$, and also computer model observations $y^c(\boldsymbol{x}, \boldsymbol{t})$ observed at locations $\boldsymbol{x}_j \in \mathcal{D}$ and at calibration settings $\boldsymbol{t}_j \in \mathcal{T}$ for some fixed domain $\mathcal{T}$ where $j = 1, \ldots, M$. Suppose there are D space-time dimensions, so that each location $\boldsymbol{x}$ is a $D \times 1$ vector, and suppose there are L calibration parameter dimensions, so that each calibration setting $\boldsymbol{t}$ is an $L \times 1$ vector. Then, when all the partial first derivative processes are available, our model is:

$$\begin{aligned} y^f(\boldsymbol{x}) = &\quad \eta(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon, \\ y^c(\boldsymbol{x}, \boldsymbol{t}) = &\quad \eta(\boldsymbol{x}, \boldsymbol{t}) \\ y^c_1(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial x_1} \\ &\vdots \\ y^c_D(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial x_D} \\ y^c_{D+1}(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial t_1} \\ &\vdots \\ y^c_{D+L}(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial t_L}, \end{aligned} \tag{5.4}$$

where there are D partial derivative processes with respect to the space-time dimensions and L partial derivative processes with respect to the calibration parameter dimensions. Note that we have overloaded the subscripts to make the notation compact and readable. That is, if the subscript is given by u, then $1 \leq u \leq D$ represents differentiation with respect to the uth dimension of $\boldsymbol{x}$, while $(D+1) \leq u \leq (D+L)$ represents differentiation with respect to the $(u-D)^{th}$ dimension of $\boldsymbol{t}$. For example, if $D = L = 2$ and $u = 1$, then $y^c_u(\boldsymbol{x}, \boldsymbol{t}) = \frac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial x_1}$, while if $u = 3$ then $y^c_u(\boldsymbol{x}, \boldsymbol{t}) = \frac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial t_1}$.

Let $\mathbf{Y}^f = [y^f(\boldsymbol{x}_1), \ldots, y^f(\boldsymbol{x}_N)]$ be the $N \times 1$ vector of field observations, $\mathbf{Y}^c = [y^c(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, y^c(\boldsymbol{x}_M, \boldsymbol{t}_M)]$ be the $M \times 1$ vector of computer model observations and $\mathbf{Y}^c_u = [y^c_u(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, y^c_u(\boldsymbol{x}_M, \boldsymbol{t}_M)]$ be the $M \times 1$ vector of computer model derivative observations where the

derivative is with respect to the uth input. Then, letting $\mathbf{Y} = [\mathbf{Y}^f, \mathbf{Y}^c, \mathbf{Y}^c_1, \ldots, \mathbf{Y}^c_D, \mathbf{Y}^c_{D+1}, \ldots, \mathbf{Y}^c_{D+L}]$, the joint distribution of the data conditional on the parameters is

$$\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \lambda_f, \lambda_z \sim \mathcal{N}\left(0, \tfrac{1}{\lambda_z}\mathbf{R}^Y + \tfrac{1}{\lambda_f}\begin{bmatrix} \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)$$

where

$$\mathbf{R}^Y = \begin{bmatrix} \mathbf{R}^f_{00} & \mathbf{R}^{fc}_{00} & \mathbf{R}^{fc}_{01} & \ldots & \mathbf{R}^{fc}_{0(D+L)} \\ \mathbf{R}^{cf}_{00} & \mathbf{R}^{c}_{00} & \mathbf{R}^{c}_{01} & \ldots & \mathbf{R}^{c}_{0(D+L)} \\ \mathbf{R}^{cf}_{10} & \mathbf{R}^{c}_{10} & \mathbf{R}^{c}_{11} & \ldots & \mathbf{R}^{c}_{1(D+L)} \\ \vdots & \ldots & \vdots & \ddots & \vdots \\ \mathbf{R}^{cf}_{(D+L)0} & \mathbf{R}^{c}_{(D+L)0} & \ldots & & \mathbf{R}^{c}_{(D+L)(D+L)} \end{bmatrix}, \tag{5.5}$$

and $\boldsymbol{\rho}$ is an $D \times 1$ vector of correlation parameters for the space-time dimensions, $\boldsymbol{\phi}$ is an $L \times 1$ vector of correlation parameters for the calibration parameter dimensions, λ_f is the precision of the measurement error process, λ_z is the precision of the stochastic process model for η, and $\boldsymbol{\theta}$ is an $L \times 1$ vector of the unknown calibration parameter settings.

In (5.5), superscripts denote correlation amongst field observations, computer model observations, or the cross-correlation between field and computer model observations, while the subscripts identify which derivative processes a particular sub-matrix of the full correlation matrix is calculated for. Again note that here we have overloaded the subscripts to make the notation more compact and readable. That is, if the subscripts are represented by u and v, then $1 \leq u \leq D$ represents differentiation with respect to the uth dimension of $\boldsymbol{x}$, while $(D+1) \leq u \leq (D+L)$ represents differentiation with respect to the $(u-D)^{th}$ dimension of $\boldsymbol{t}$. Similarly for v. For instance, if $D = L = 2$ and $u = 1, v = 3$, then $\mathbf{R}^c_{uv} = Corr\left(\mathbf{Y}^c_1, \mathbf{Y}^c_3\right)$, where $\mathbf{Y}^c_1 = \left[\frac{\partial\eta(\boldsymbol{x}_1,\boldsymbol{t}_1)}{dx_1}, \ldots, \frac{\partial\eta(\boldsymbol{x}_M,\boldsymbol{t}_M)}{dx_1}\right]$ and $\mathbf{Y}^c_3 = \left[\frac{\partial\eta(\boldsymbol{x}_1,\boldsymbol{t}_1)}{dt_1}, \ldots, \frac{\partial\eta(\boldsymbol{x}_M,\boldsymbol{t}_M)}{dt_1}\right]$. All these correlation matrix entries can be computed following the formulas (5.1), (5.2) and (5.3) above.

In specifying the priors, first note that $0 < \rho_d < 1$ for all $d = 1, \ldots, D$ and similarly $0 < \phi_l < 1$ for all $l = 1, \ldots, L$. Priors for these correlation parameters often put more weight towards 1 to emphasize a smoother and less variable process unless the data indicates

otherwise. For instance, a Beta($\alpha = 5, \beta = 1$) prior is a reasonable choice that places most of the distributions weight on correlation parameter values greater than 0.5. This is due to a preference for parsimony and also because for small sample sizes, there is usually not enough information in the data to suggest more complex responses. Of course, the precision parameters are required to be positive, $\lambda_f > 0, \lambda_z > 0$. We usually use Gamma priors for these parameters, and assume the process is scaled to have variance 1 for simplicity. For the calibration parameter $\boldsymbol{\theta}$, we will typically assume that these settings have been scaled to the unit hypercube $[0,1]^L$ which covers the part of this parameter space where the experimenter knows the truth lies. Reasonable priors for θ_l might be a Uniform(0,1), or a truncated Normal distribution. This corresponds to a belief that the true calibration parameter lies within the range specified by the experimenter, and in the latter case that the truth is additionally believed to be centered in the plausible range. Our particular choices of prior distributions will be specified in more detail depending on the problem at hand later. We treat all these parameters as independent of one another, so that the prior distribution for the model parameters is a product of 1-dimensional priors.

The posterior distribution of the parameters,

$$\begin{aligned} \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \lambda_f, \lambda_z | \mathbf{Y} \propto \quad & L(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \lambda_f, \lambda_z) \times \textstyle\prod_{l=1}^{L} \pi(\theta_l) \times \prod_{k=1}^{K} \pi(\rho_k) \\ & \times \textstyle\prod_{l=1}^{L} \pi(\phi_l) \times \pi(\lambda_f) \times \pi(\lambda_z) \end{aligned} \tag{5.6}$$

is sampled using Markov Chain Monte Carlo (MCMC). We use an adaptive Metropolis-Hastings algorithm with uniform proposal distributions and let the step size adapt for the first 50% of the iterations before allowing the chain to burn-in. We take a sample of size S from the end of the chain as draws from the stationary distribution. The specific details are specified in the examples and simulations.

Predictions of the field, $\mathbf{Y}^* = \left[\eta(\boldsymbol{x}_1^*, \boldsymbol{\theta}), \ldots, \eta(\boldsymbol{x}_{N_p}^*, \boldsymbol{\theta})\right]$, are constructed at the N_p prediction locations $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{N_p}^*)$ similar to the formulas from Chapter 2. First, note that the joint distribution of the process at the prediction locations and the observed data

is specified as

$$(\mathbf{Y}^*, \mathbf{Y})|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \lambda_f, \lambda_z \sim \mathcal{N}\left(0, \frac{1}{\lambda_z}\begin{bmatrix} \mathbf{R}^* & \mathbf{R}^{*Y} \\ \mathbf{R}^{Y*} & \mathbf{R}^{Y} \end{bmatrix} + \frac{1}{\lambda_f}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix}\right),$$

where $\mathbf{W} = diag(\mathbf{I}_N, \mathbf{0}_{M\times(D+L+1)})$, so that the predictive distribution for the process at the unobserved locations, $\mathbf{Y}^*|\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \lambda_f, \lambda_z$, can be expressed using the usual properties of multivariate normal distributions (see Appendix B). The density of the posterior predictive distribution is then

$$f(\mathbf{Y}^*|\mathbf{Y}) = \int\int\int\int\int f(\mathbf{Y}^*|\mathbf{Y}, \cdot)\pi(\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \lambda_f, \lambda_z|\mathbf{Y})d\boldsymbol{\theta}d\boldsymbol{\rho}d\boldsymbol{\phi}d\lambda_f d\lambda_z$$

which can be used to construct the mean estimate of the predictor and pointwise prediction intervals using quantiles of the distribution. In practice, these estimates are approximated using samples from the posterior predictive distribution constructed from the samples of the posterior distribution for the parameters (5.6).

5.4 Example

Let us now demonstrate the application of the model through a 2-dimensional example, where the first dimension represents spatial location and the second dimension is the univariate calibration parameter. In other words, the physical process is a curve parametrized by the calibration parameter, and the collection of possible curves are represented by a surface. For this example, we need to calculate the computer model $\eta(\boldsymbol{x}, \boldsymbol{t})$ at $M = 5$ settings of (x, t) and we simulate the field process $\eta(\boldsymbol{x}, \boldsymbol{\theta})$ at $N = 3$ locations. This was done by drawing a realization of the calibration model with derivatives (7.1) with $\lambda_z = 1, \rho = 0.05$ and $\phi = 0.05$, where the M computer trials and N field observation locations were selected by a space-filling design (Johnson et al., 1990) and choosing $\theta = 0.5$ as the true calibration parameter setting. We also simultaneously draw observations of the field over an equi-distant prediction grid of $N_p = 30$ locations, which we will use later for comparing the predictive performance of our calibrated estimates.

The field observations were constructed by adding i.i.d. normally distributed mean-zero error with precision $\lambda_f = 10/v_f$ to the drawn field process, where v_f is the sample variance of the true field process calculated using the $N_p = 30$ prediction grid locations. In other words, we add observation error with a variance that is approximately 10% of the variance of the underlying true field process. So, our data consists of the computer model observations $\mathbf{Y}^c = [\eta(x_1, t_1), \ldots, \eta(x_5, t_5)]$ and field observations $\mathbf{Y}^f = [\eta(x_6, \theta) + \epsilon_1, \ldots, \eta(x_8, \theta) + \epsilon_3]$ where the ϵ_i's are i.i.d. draws from $\mathcal{N}(0, \frac{1}{\lambda_f})$. In addition, we have the derivative observations $\mathbf{Y}_1^c = \left[\frac{\partial\eta(x_1,t_1)}{\partial x}, \ldots, \frac{\partial\eta(x_5,t_5)}{\partial x}\right]$ and $\mathbf{Y}_2^c = \left[\frac{\partial\eta(x_1,t_1)}{\partial t}, \ldots, \frac{\partial\eta(x_5,t_5)}{\partial t}\right]$.

The prior distribution on λ_z was taken as Gamma($\alpha_z = 5, \beta_z = 5$) and for λ_f was taken as Gamma($\alpha_f = 25, \beta_f = 2.5$), so that the means of both priors are located at the true values, and are conservative enough so that the model will move from the prior if the data suggests. The priors for both ρ and ϕ are Beta($\alpha = 5, \beta = 1$), which places more weight on correlation parameters closer to 1, which indicates a prior belief in a smoother process. Finally, the prior for θ was taken to be Normal($\mu_\theta = 0.5, \sigma_\theta^2 = 0.0625$) which is centered at 0.5 and has 95% of the weight of the distribution in the plausible interval $[0, 1]$ and was truncated to this interval.

The MCMC algorithm is implemented as follows. The chain was initialized with the values $\lambda_z = 1, \lambda_f = 10, \rho = 0.5, \phi = 0.5$, and $\theta = 0.5$. The algorithm performs Metropolis-Hastings steps for each parameter using uniform proposal distributions. The stepwidths for the uniform proposals were initialized as $w_{\lambda_z} = 0.5, w_{\lambda_f} = 50, w_\rho = 0.03, w_\phi = 0.1$, and $w_\theta = 0.15$. For each proposal of each parameter, the ratio

$$r = min\left(1, \frac{G^*}{G}\right)$$

is calculated, where G^* is proportional to the posterior distribution (5.6) calculated with the new proposed parameter, and G is proportional to the posterior distribution (5.6) calculated with the existing parameters. We then accept the proposal with probability r and reject it with probability $1 - r$, and continue to the next parameter.

Model	λ_z	λ_f	ρ	ϕ	θ
Standard	0.07	0.06	0.37	0.59	0.23
Derivatives	0.32	0.62	0.91	0.88	0.84

Table 5.1: Geweke diagnostics (p-values) for the standard calibration model and the calibration model with derivatives comparing the first 20% of the MCMC chain to the last 50% of the MCMC chain.

We ran the algorithm for $S_0 = 5000$ iterations, and allowed the stepwidths to be adapted every $S_0/20$ iterations for the first $S_0/2$ iterations. The stepwidths were adapted to target the 0.44 accept/reject ratio as discussed in Gelman et al. (2004). After these $S_0/2$ iterations, the stepwidths are held fixed and the chain is allowed to burn-in.

There are a few popular approaches to deciding if the MCMC chain has reached the stationary distribution. We will consider the Geweke diagnostic (Geweke, 1992), the Raftery-Lewis diagnostic (Raftery and Lewis, 1992, 1995) as well as the visual inspection of traceplots and autocorrelation function (ACF) plots.

The Geweke diagnostic aims to compare two estimates of the mean of the MCMC chain taken from the first and last part of the chain. If the mean has converged then we don't expect there to be a difference between these estimates. The approach estimates the variance using spectral approaches so that possible correlation between the sampled values of the chain is accounted for. The test statistic is then asymptotically normal and a p-value for the difference of means test is reported. We applied the Geweke diagnostic as implemented in the R (Team, 2007) package CODA (Plummer et al., 2009), comparing the first 20% of each MCMC chain to the last 50%. The p-values are reported in Table 5.1, from which we can see the statistics are not significantly different from zero at the 0.05 level. This suggests that the MCMC chains are converging quickly.

The Raftery-Lewis diagnostic estimates the number of burn-in iterations and total iterations required to estimate a quantile based on a sample from the chain. The diagnostic

Model	Quantile	L/Burn	λ_z	λ_f	ρ	ϕ	θ
Standard	0.025	L	937	937	937	937	937
	0.025	Burn	15	10	14	19	13
	0.975	L	937	937	937	937	937
	0.975	Burn	16	14	13	12	13
Derivatives	0.025	L	937	937	937	937	937
	0.025	Burn	29	10	14	13	11
	0.975	L	937	937	937	937	937
	0.975	Burn	24	10	20	22	13

Table 5.2: Raftery-Lewis diagnostics for minimum chain length (L) and burn-in (Burn) for the standard calibration model and the calibration model with derivatives with $r = 0.01, s = 0.95$ for both the 0.025 and 0.975 quantiles.

accepts a quantile q, the desired precision r with which we want to estimate the quantile, and at what probability s this level of precision should be attained. In other words, the diagnostic suggests the burn-in and total iterations required so the estimate of q is within $(q-r, q+r)$ with probability s. Typically, one will run this test on the full chain, then drop the suggested burn-in and re-run the diagnostic to see if the suggested burn-in is closer to zero. Then the chain can be run to the suggested total iterations to get a highly accurate estimate of the quantile if desired.

We applied the Raftery-Lewis diagnostic as implemented in the R (Team, 2007) package CODA (Plummer et al., 2009) to our fit of the standard model and the model with derivatives from the described example, specifying $q = (0.025, 0.975), r = 0.01$ and $s = 0.95$. The results shown in Table 5.2 also indicate that the chain has burned in quite quickly as few iterations are suggested to be dropped as burn-in. Also, the calculation suggests that a sample of size $S = 1000$ from the chain sufficiently describes the true posterior distribution at the prescribed accuracy. We re-calculated both diagnostics using the last $S = 1000$ iterations as our sample from the posterior. Both diagnostics showed little change, suggesting the MCMC chains have reached their stationary distributions.

Finally, we can also visually inspect the traceplots shown in Figure 5.1, where the traces for the model with derivatives are in the left column, and for the standard model are in the right column. The traceplots appear reasonable, suggesting that the MCMC has reached the stationary distribution. The true parameter values are plotted as solid red lines. From these lines, one can immediately notice that the standard model does not appear to have converged to the true values of the correlation parameters ρ and ϕ. In contrast, the model that includes derivatives is able to estimate the correlation parameters with a much better accuracy.

The autocorrelation (ACF) plots are also shown in Figure 5.2, where the left column shows the ACF's for the model with derivative information and the right column corresponds to the standard model. Note the presence of autocorrelation in the chosen posterior sample. Using a thinning interval of 10 was found to remove this autocorrelation, and in practice one would run the chain much longer to get an improved sample of the posterior after thinning. However, for our purposes, this is simply not computationally possible if one wants to perform a simulation study in a reasonable amount of time, so we proceed using the posterior sample selected.

The calibrated responses are shown in Figure 5.3 and Figure 5.4. In Figure 5.3, the black surface represents the true computer model η, and the yellow dots represent the $M = 5$ observations of the computer model and also the locations where the two partial derivatives of the computer model were observed. The red dots with stems are the $N = 3$ field observations that were simulated from the computer model by setting the true calibration parameter to be 0.5 and adding normally distributed error. Realizations of the predictive process for both the standard model and the model with derivatives were constructed by taking samples from the posterior distribution that fall within the empirical 95% interval for θ and plotting the prediction of the process for each such sample. These realizations are shown in blue for the standard model and in green for the model with derivatives.

One can see that the standard model is less accurate in estimating θ as the blue realizations in Figure 5.3 are spread over a wider range of values of the calibration parameter (t), while the green curves from the model with derivatives are more closely centered around the true value of 0.5. The second thing that can be noticed is how the realizations from the model with derivatives are more curved, seemingly following the true surface with greater accuracy than the standard model. This behaviour is more clearly seen in Figure 5.4. Here, the curve representing the true field process is plotted in black while the field observations are represented by red dots. The blue and green curves represent the average predictions of the standard model and model with derivatives where the average was computed over the posterior samples. The dotted lines represent 95% prediction intervals computed using the pointwise quantiles of the posterior samples of the predictor. Here we see more clearly that the standard model completely misses the path of the true field process over most of the range of x, and the intervals do not include the true process for much of this range as well. In contrast, the model with derivatives provides a very accurate prediction of the true process and the prediction intervals are very reasonable. As an indication of how poor the standard model behaves in this example, the empirical mean squared prediction error of fit for the standard model was 2.26×10^{-1} while for the model with derivatives it was 5.57×10^{-3}, indicating that the model with derivatives was around 40 times as efficient in this case. Note that the true field process has a variance of approximately 1 in this example.

We will investigate the performance of our model in greater detail through an extensive simulation study in the next chapter, but first we mention a final technical point in regards to calibrating with derivative information.

5.5 Calibration with Incomplete Derivatives

A final practical consideration in applying our calibration model that incorporates derivative information is the possibility that not all of the partial first derivatives of the computer model with respect to the inputs may be available. In our experience, this situation can

sometimes occur in calibration experiments depending on how the computer model is implemented. It is quite typical for the implementation to compute η and it's derivatives over a grid of space-time locations given a single setting of the calibration parameter. For instance, the LFM example from Chapter 3 computed outputs over a space-time grid for a single setting of the calibration inputs. In this case, subsequent runs of the computer model can be run at various settings of the calibration parameter, but only the derivatives with respect to the space-time inputs are available for the calibration experiment. Nonetheless, we can still make use of the incomplete derivative information available to us by simply taking the appropriate sub-matrix of (5.5).

Now we continue in the next chapter to investigate and compare the standard calibration model and the proposed calibration model with derivative information via simulation.

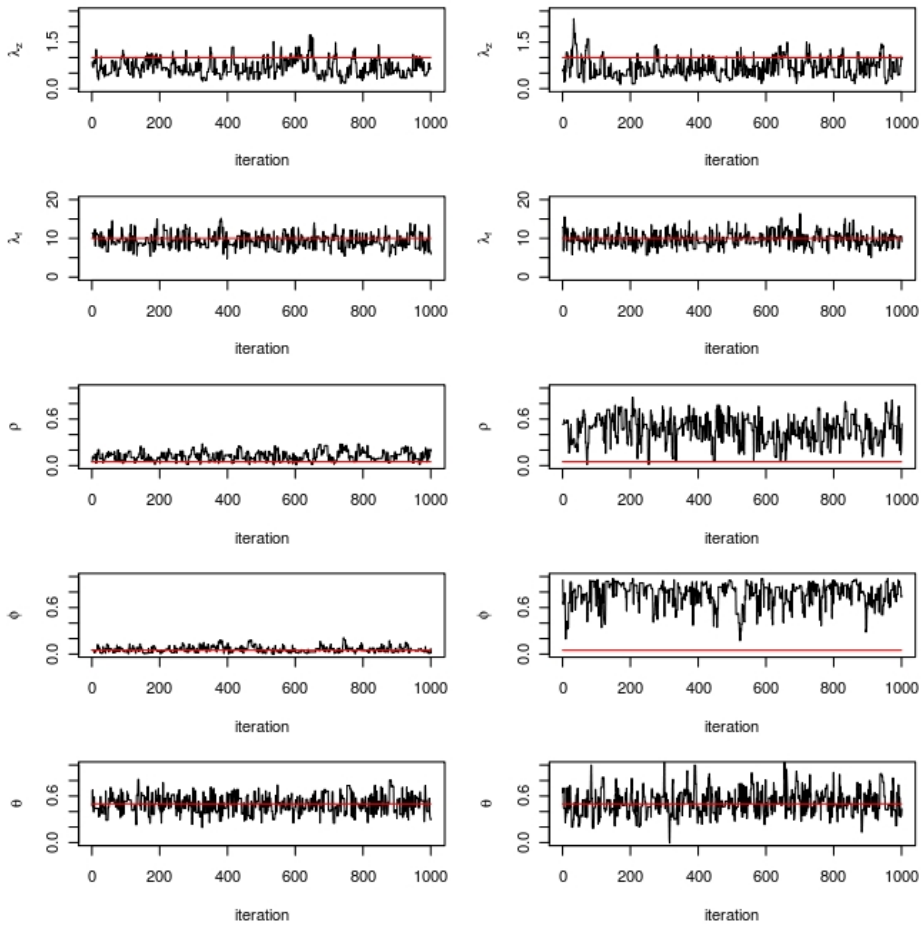


Figure 5.1: Traceplots for parameters of fitted calibration model with derivatives (left column) and standard calibration model (right column). True parameter values are shown as red lines.

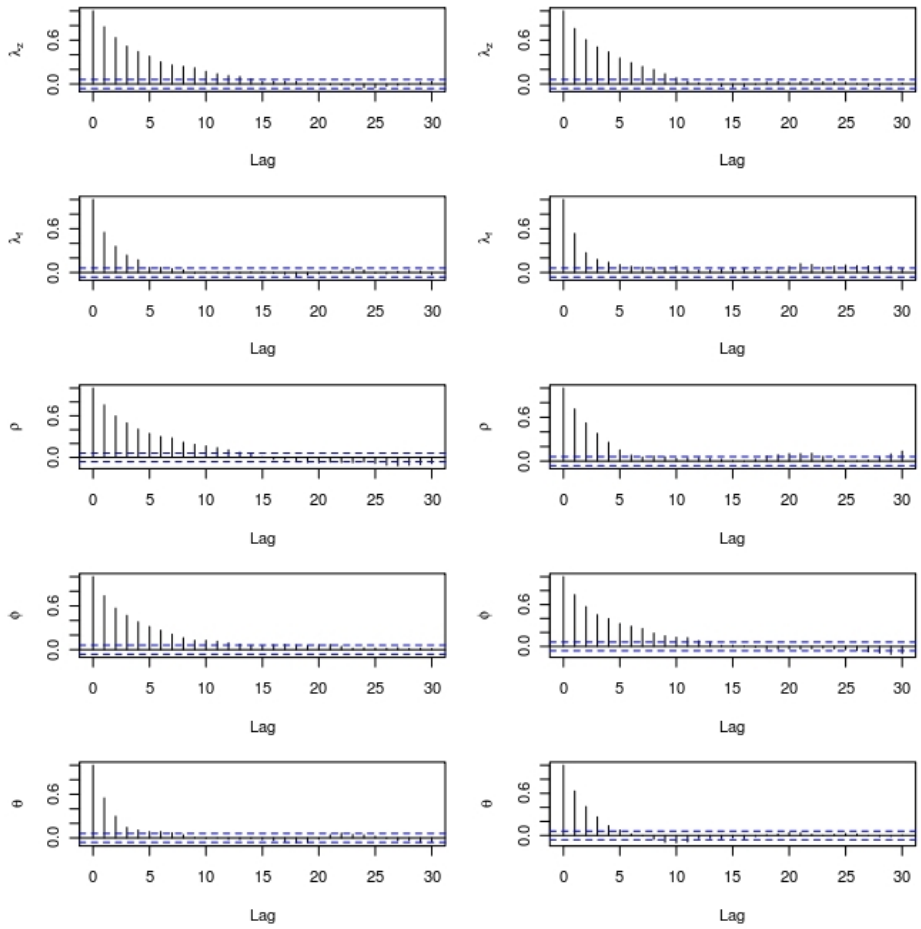


Figure 5.2: Autocorrelation plots for parameters of fitted calibration model with derivatives (left column) and standard calibration model (right column).

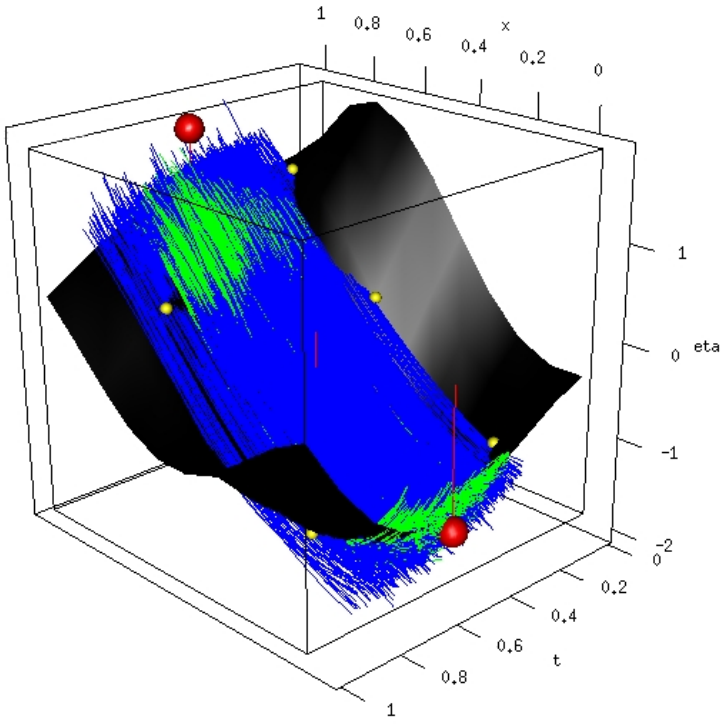


Figure 5.3: Calibration of a 1-dimensional surface with a 1-dimensional calibration parameter. Here, $M = 5$ computer model trials were run (yellow dots) and $N = 3$ field observations were observed (red dots with stems). The true calibration parameter was set as $\theta = 0.5$, and draws from the calibrated field predictions for the standard model are shown in blue while draws for the model with derivatives are shown in green.

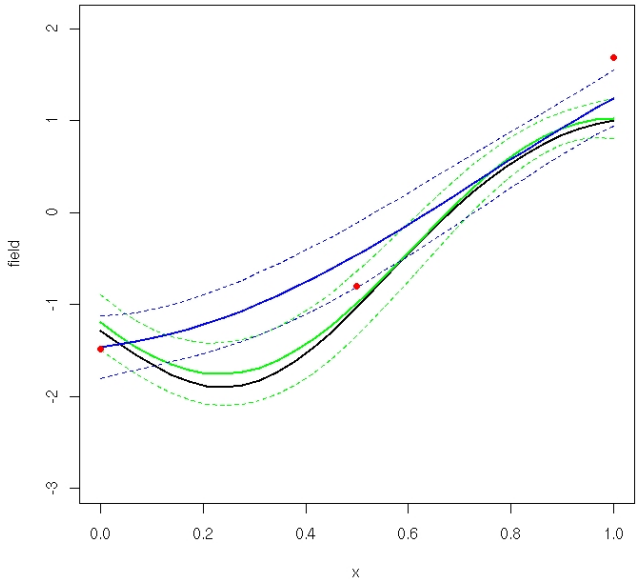


Figure 5.4: Calibration of a 1-dimensional surface with a 1-dimensional calibration parameter. Here, the true underlying process is shown as the black curve, and the $N = 3$ field observations from this curve are shown as the red dots. The mean of the calibrated field predictions for the standard model is shown as the blue line with 95% prediction intervals shown as the dashed blue lines. Similarly, the mean of the prediction from the model with derivatives is shown as the green line, with 95% prediction intervals shown as the dashed green lines.

Chapter 6

Simulation Study: Calibration with Derivatives

In this chapter we will investigate the behaviour of the proposed calibration model that includes derivative information by means of a simulation study. With this study we will also compare our model to the same model with incomplete derivative information as well as the standard calibration model. There are many aspects one could investigate when making a comparative investigation of models. In particular, the current simulation will look at five measures of model performance that are of primary interest in a computer model calibration experiment. We report the median of these measures over the simulation replicates as it is robust to outliers.

The first measure is the median of the squared errors of the calibration parameter estimate. For the ith simulation run, the squared error is calculated as

$$SE(\theta_i) = (\bar{\theta^*}_i - \theta_i)^2$$

where θ_i is the true value of the calibration parameter for the ith simulation, and $\bar{\theta^*}_i$ is the posterior mean for θ from the ith simulation. Then

$$MedSE(\theta) = \quad median(SE(\theta_1), \ldots, SE(\theta_{N_r})) \tag{6.1}$$

is the median of the $SE(\theta_i)$'s over N_r simulation runs.

The next two measures deal with evaluating the predictive performance of the models by considering the mean squared error of the calibrated field prediction as well as it's relative efficiency to the standard calibration model. For the ith simulation, we calculate the empirical mean squared prediction error as

$$MSPE_i = \frac{1}{N_p}(\widehat{\mathbf{Y}^*} - \mathbf{Y}^*)^T(\widehat{\mathbf{Y}^*} - \mathbf{Y}^*)$$

where $\widehat{\mathbf{Y}^*} = \widehat{E}[\mathbf{Y}^*|\mathbf{Y}]$ is the $N_p \times 1$ vector of calibrated field predictions calculated at the N_p prediction locations and $\mathbf{Y}^*$ is the $N_p \times 1$ vector of the true values of the field at these same locations. The median over all the simulations of these empirical mean squared prediction errors is reported as

$$MedMSPE = \quad median(MSPE_1, \ldots, MSPE_{N_r}). \tag{6.2}$$

The relative efficiency of the calibration model that includes derivative information or partial derivative information versus the standard calibration model is calculated as

$$eff = \quad \frac{MedMSPE_{std}}{MedMSPE}, \tag{6.3}$$

where $MedMSPE_{std}$ is the median of the mean squared prediction errors when using the standard calibration model, while the denominator is the median of the mean squared prediction errors when including (partial) derivative information in the calibration model. If $eff > 1$, then this means that the model with derivatives had a lower median of the average squared prediction errors as compared to the standard model, and vice-versa. For instance, if $eff = 2$, then the median of the average squared prediction errors for the model with derivatives was half that of the standard model.

The final two measures considered are coverage probability and prediction interval width. We estimate the 95% coverage probability for the ith simulation, CP_i, as the proportion of the N_p prediction points where the true process at these points lies outsize the 95% pointwise

prediction interval of the posterior predictor. Then, the median coverage probability is reported as

$$MedCP = \quad median(CP_1, \ldots, CP_{N_r}). \tag{6.4}$$

The average 95% prediction interval width for the ith simulation, PIW_i is calculated as the average of the 95% pointwise prediction intervals constructed at the N_p prediction points. At each prediction point, the 95% prediction interval was taken using the appropriate quantiles from the posterior samples of the predictor at that point. The median of these average prediction interval widths over the N_r simulations is then reported as

$$MedPIW = \quad median(PIW_1, \ldots, PIW_{N_r}). \tag{6.5}$$

In addition to the above numerical measures, boxplots of the estimated calibration parameters and empirical mean squared prediction errors are used for visualization. We remove outliers from such plots for clarity.

The simulations investigated 2-dimensional, 4-dimensional and 6-dimensional calibration problems using realizations randomly drawn from the calibration model with full derivative information. In the 2-dimensional case, there is 1 space-time dimension and 1 calibration parameter dimension. In the 4-dimensional case, there are 2 space-time dimensions and 2 calibration parameter dimensions, while in the 6-dimensional case, there are 3 space-time dimensions and 3 calibration parameter dimensions. For each of these basic calibration problem types, $N_r = 100$ replicate simulations were performed at a fixed level of parameter settings. The settings investigated in the simulation included: the number of computer model runs, M; the number of field observations, N; the correlation parameter for the space-time dimensions, $\boldsymbol{\rho}$; the correlation parameter for the calibration parameter dimensions, $\boldsymbol{\phi}$; the precision of the measurement error, λ_f; and the precision of the process, λ_z. All inputs were scaled to $[0, 1]$, and the true calibration parameters $\boldsymbol{\theta}$ for a given replicate of a simulation run was drawn from the uniform(0,1) distribution. The setting levels investigated for each problem type are summarized in Table 6.1, Table 6.2 and Table 6.3 below. The simulation experiment consisted of running the full factorial of these setting levels with

Setting	Level 1	Level 2	Level 3
M	5	10	15
N	3	7	15
ρ,ϕ	0.05	0.2	0.6
λ_f	1000	10	3
λ_z	1	n/a	n/a

Table 6.1: Setting levels for the 2-dimensional calibration simulation study.

Setting	Level 1	Level 2	Level 3
M	20	30	n/a
N	7	15	n/a
ρ,ϕ	0.05	0.2	0.6
λ_f	1000	10	3
λ_z	1	n/a	n/a

Table 6.2: Setting levels for the 4-dimensional calibration simulation study.

$N_r = 100$ replicates each.

For these simulations, we used the adaptive MCMC algorithm as described in the example of Section 5.4. In particular, we use $5,000$ MCMC steps where the first $2,500$ allow the proposal distributions to be adapted, while we report the last $1,000$ as draws from the posterior. The prior distributions were also the same as described in our earlier example.

Setting	Level 1	Level 2
M	30	45
N	11	23
ρ,ϕ	0.2	n/a
λ_f	10	n/a
λ_z	1	n/a

Table 6.3: Setting levels for the 6-dimensional calibration simulation study.

Three calibration models were compared in these simulations: (i) the calibration model with full derivative information; (ii) the calibration model with partial derivative information; and (iii) the standard calibration model. For the model with partial derivative information, we chose to include the derivatives with respect to the space-time dimensions while excluding those for the calibration parameter dimensions. This is based on our experience with common computer model implementations, although studying a few examples where the calibration parameter derivatives were included instead of the space-time derivatives suggested that the effect of having incomplete derivative information was similar in either case.

For such an extensive simulation study, we highlight important findings rather than exhaustively listing results in the upcoming sections.

6.1 Results: Estimating θ

A good estimate of the unknown true calibration parameter is one of the most important goals in a calibration experiment. This is because the physical interpretation of the estimated calibration parameter can be important for subject-area scientists. An accurate estimate is also important for future runs of the computer model, and even can be used in subsequent studies of other physical processes where these parameters may be required, but it is not possible or efficient to estimate them.

The value of $MedSE(\theta)$ was calculated according to (6.1) for the various settings explored in Tables 6.1-6.3. The results for the 2-dimensional simulations are shown below in Table 6.4 and Table 6.5 for small ($\lambda_f = 1000$) and large ($\lambda_f = 10$) levels of measurement error. The column header of Table 6.4 shows the settings of the correlation parameters in the first row, and the second row indicates the number of computer model runs. The trials were selected using a space-filling design as described in our example. The row header indicates the type of model summarized, where "i" indicates the calibration model with full

derivative information, "ii" indicates the calibration model with space-time derivative information, and "iii" indicates the standard calibration model with no derivative information. Finally, the second column indicates the number of field observations, N, which were also selected using a space-filling design for each simulation replicate. The table entries list the value of $MedSE(\theta)$ calculated over $N_r = 100$ simulation replicates.

To compare the three different models, one can read a given column down the rows. So, for instance, the first column represents a correlation parameter setting of 0.6 and $M = 5$ computer model observations and $N = 3, 7, 15$ field observations. With $N = 3$, the $MedSE(\theta)$ for the full derivatives model of $1.36e-05$ is about 33 times smaller than the $MedSE(\theta)$ for the standard model, which is $4.54e-04$ in this case. By comparison, the model with partial derivative information reduces the $MedSE(\theta)$ by around 2.5 times as compared to the standard model.

Moving one column right along the same rows, one can see the effect of increasing the number of computer model observations. For instance, for $M = 10$, the full derivatives model has about a 5.5 times smaller error when compared to the standard model. In contrast, the model with partial derivative information demonstrates similar performance as the standard model. Moving one more column to the right, we see that all three models offer similar performance with $M = 15$ computer model observations. This decrease in the improvement realized with derivatives is also seen as we hold the number of computer model observations to a small level and increase the number of field observation (eg - looking down the first column). This effect is not surprising, since as we increase the number of observations to a fairly high level ($N = 15, M = 15$), the uncertainty in the response is likely so low that the inclusion of derivatives does not lend much additional information.

	ρ, ϕ :	0.6			0.2			0.05		
Model	M :	5	10	15	5	10	15	5	10	15
i.		1.36e-05	8.81e-06	1.29e-05	5.84e-05	6.42e-06	7.62e-06	8.51e-05	1.61e-05	1.41e-05
ii.	$N = 3$	1.80e-04	3.85e-05	1.38e-05	7.82e-04	2.11e-05	1.99e-05	1.09e-03	6.57e-05	1.47e-05
iii.		4.54e-04	4.91e-05	1.26e-05	1.96e-03	1.44e-04	3.17e-05	3.24e-03	1.54e-04	9.75e-05
i.		4.82e-06	6.27e-06	5.12e-06	1.07e-05	3.93e-06	3.74e-06	2.31e-05	8.28e-06	2.83e-06
ii.	$N = 7$	7.15e-05	5.30e-06	5.82e-06	3.23e-04	1.09e-05	3.31e-06	7.88e-04	2.57e-05	6.31e-06
iii.		2.60e-04	2.12e-05	9.15e-06	1.43e-03	2.14e-05	6.80e-06	2.34e-03	1.24e-04	1.03e-05
i.		5.91e-06	2.46e-06	2.39e-06	1.06e-05	1.65e-06	2.80e-06	1.34e-05	2.54e-06	1.32e-06
ii.	$N = 15$	1.33e-04	1.93e-06	1.68e-06	4.27e-04	5.55e-06	2.94e-06	1.30e-03	1.95e-05	5.34e-06
iii.		2.16e-04	1.46e-05	2.59e-06	9.39e-04	3.42e-05	5.06e-06	5.33e-03	9.97e-05	1.44e-05

Table 6.4: Reported $MedSE(\theta)$ for $\lambda_f = 1000$ from 2-d calibration simulation study.

	ρ, ϕ :	0.6			0.2			0.05		
Model	M :	5	10	15	5	10	15	5	10	15
i.		1.68e-03	1.25e-03	9.82e-04	1.07e-03	1.01e-03	1.11e-03	9.43e-04	8.53e-04	7.79e-04
ii.	$N = 3$	2.38e-03	1.26e-03	1.17e-03	2.52e-03	9.51e-04	1.04e-03	3.87e-03	9.39e-04	7.63e-04
iii.		3.63e-03	1.51e-03	9.74e-04	3.55e-03	9.84e-04	1.48e-03	6.44e-03	1.75e-03	8.11e-04
i.		8.31e-04	4.83e-04	2.93e-04	4.74e-04	9.63e-04	9.92e-04	4.05e-04	2.21e-04	5.34e-04
ii.	$N = 7$	1.02e-03	4.51e-04	2.86e-04	1.61e-03	1.12e-03	1.04e-03	2.11e-03	4.20e-04	4.49e-04
iii.		1.35e-03	4.21e-04	2.46e-04	2.21e-03	9.25e-04	8.38e-04	6.13e-03	1.19e-03	6.34e-04
i.		7.39e-05	3.11e-04	2.13e-04	4.01e-04	1.28e-04	2.07e-04	3.98e-04	1.71e-04	2.64e-04
ii.	$N = 15$	4.08e-04	2.23e-04	1.86e-04	1.64e-03	1.26e-04	1.57e-04	2.21e-03	4.07e-04	3.55e-04
iii.		5.08e-04	4.46e-04	2.08e-04	2.14e-03	2.33e-04	2.73e-04	3.46e-03	5.90e-04	3.99e-04

Table 6.5: Reported $MedSE(\theta)$ for $\lambda_f = 10$ from 2-d calibration simulation study.

As the correlation parameter is decreased, the variability of the simulated computer model responses and simulated field observations increases, representing more complex functions. In such cases, the gains that can be realized by including derivative information are larger. For instance, in the small sample case where $N = 3$, $M = 5$ and $\rho = \phi = 0.05$, we see the model with derivative information has a 38 times smaller error than the standard model. Only including partial derivatives in this case reduces the error by about 3 times.

The largest improvement for these more complex responses was seen with $N = 7$ field observations and $M = 5$ computer model runs, where the $MedSE(\theta)$ of the model with derivative information was about 100 times smaller than the standard calibration model. For these complex responses, we also see that as the number of observations increases, there is still a noticeable improvement with the derivative information. For instance, with $N = 15, M = 15$, the model with derivatives had a $MedSE(\theta)$ approximately 11 times smaller as compared to the standard model. This is likely due to more complex responses requiring more observations to be sufficiently described, whether derivatives are included or not. Nonetheless, the benefits of including derivative information are quite stark. As an example, with $M = 15$ computer model runs, the derivative model has similar performance with $N = 3$ field observations as the standard model with $N = 15$ field observations. Similarly, for $M = 10$ computer model runs, the derivative model's error with $N = 3$ observations was 6 times smaller than the standard models error using $N = 15$ observations. Finally, with only $M = 5$ computer model runs, the derivative model's $MedSE(\theta)$ with $N = 3$ observations was about 62 times smaller than that achieved using the standard model with $N = 15$ field observations.

As the measurement error is increased, the benefits seen with derivative information is diminished due to the uncertainty in the field observations. This is particularly true for the very simple responses generated with a correlation parameter of 0.6, as summarized in the first three columns of Table 6.5. Here, the setting of $\lambda_f = 10$ represents adding a measurement error with variance about 10% of the variance of the true but unobserved field

response. Note that all three models exhibit similar performance for these simple responses, except for the case of $M = 5$ and $N = 7$ or $N = 15$, where the model with derivative information starts to provide some benefit in the estimation of the calibration parameter.

As the correlation parameter decreases, leading to more complex responses, the benefits of including the derivative information again become clear. With a correlation parameter of 0.05, the derivative information provides notable improvement except at the largest sample sizes. For instance, when $M = 15$ and $N = 15$ the model with derivatives delivers similar performance as the standard model. In contrast, with $M = 5$ and $N = 3$, the model with derivatives reduces the error in the calibration parameter estimate by a factor of 7 when compared to the standard model. Here, the model with partial derivative information saw a small but noticeable 1.7 times reduction as compared to the standard model.

When comparing the model with derivatives using $M = 5$ computer model runs and $N = 3$ field observations to the standard model with $N = 15$ field observations, the model with derivatives has a 3.7 times smaller $MedSE(\theta)$. This suggests that the model with derivatives is able to borrow information from the computer model runs much more efficiently than the standard model as it still has a much lower error in the calibration parameter estimate even though the standard model has 5 times the number of field observations in that particular comparison. Overall, the strongest reduction in $MedSE(\theta)$ for the case of a correlation parameter of 0.05 occurred for $M = 10$ and $N = 7$, in which case the model with derivatives had achieved a 15 times reduction compared to the standard model. We also saw similar behaviour for the case of higher measurement error ($\lambda_f = 3$, results not shown) in that an increase in observations reduces the difference between models while an increase in the complexity of the responses leads to more accurate estimation of the calibration parameter when using the model with derivative information.

	ρ, ϕ :	0.6				0.2				0.05			
	M :	20		30		20		30		20		30	
Model		θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
i.		3.32e-04	4.89e-04	4.43e-04	6.47e-04	2.71e-04	5.91e-04	2.23e-04	1.25e-04	7.87e-04	5.44e-04	5.03e-04	4.27e-04
ii.	$N = 7$	4.78e-04	7.55e-04	4.44e-04	4.98e-04	1.84e-03	1.56e-03	5.76e-04	6.69e-04	3.20e-03	2.77e-03	1.17e-03	2.36e-03
iii.		1.42e-03	1.91e-03	1.28e-03	1.99e-03	4.12e-03	9.92e-03	2.11e-03	2.10e-03	9.57e-03	1.18e-02	4.20e-03	7.91e-03
i.		2.20e-04	1.20e-04	2.08e-04	2.77e-04	1.90e-04	1.44e-04	8.80e-05	8.68e-05	5.67e-04	5.36e-04	1.85e-04	2.31e-04
ii.	$N = 15$	5.96e-04	5.58e-04	3.50e-04	4.94e-04	8.85e-04	1.12e-03	3.04e-04	3.00e-04	2.03e-03	1.74e-03	8.32e-04	9.78e-04
iii.		2.26e-03	2.99e-03	1.01e-03	1.08e-03	3.64e-03	5.76e-03	1.52e-03	2.05e-03	5.86e-03	7.90e-03	4.40e-03	4.52e-03

Table 6.6: Reported $MedSE(\theta)$ for (θ_1, θ_2) and $\lambda_f = 1000$ from 4-d calibration simulation study.

	ρ, ϕ :	0.6				0.2				0.05			
	M :	20		30		20		30		20		30	
Model		θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
i.		2.77e-03	1.92e-03	1.81e-03	2.25e-03	1.76e-03	1.75e-03	9.48e-04	1.51e-03	1.68e-03	1.97e-03	2.03e-03	8.64e-04
ii.	$N = 7$	3.01e-03	3.22e-03	1.83e-03	1.89e-03	1.93e-03,	3.86e-03	1.82e-03	2.45e-03	4.86e-03	6.31e-03	3.03e-03	2.39e-03
iii.		9.42e-03	6.67e-03	2.91e-03	4.24e-03	7.13e-03	1.06e-02	5.69e-03	5.56e-03	1.03e-02	1.33e-02	8.71e-03	7.67e-03
i.		1.81e-03	2.08e-03	9.32e-04	6.07e-04	5.14e-04	6.47e-04	6.96e-04	7.67e-04	9.13e-04	5.57e-04	1.27e-03	9.66e-04
ii.	$N = 15$	2.23e-03	1.82e-03	7.32e-04	5.45e-04	2.16e-03	1.54e-03	9.84e-04	1.15e-03	2.31e-03	3.28e-03	2.01e-03	2.18e-03
iii.		5.40e-03	3.57e-03	1.09e-03	9.50e-04	8.82e-03	6.23e-03	3.87e-03	4.06e-03	6.38e-03	1.07e-02	9.02e-03	7.95e-03

Table 6.7: Reported $MedSE(\theta)$ for (θ_1, θ_2) and $\lambda_f = 10$ from 4-d calibration simulation study.

Results for the 4-dimensional calibration study with small measurement error are shown in Table 6.6, and exhibit similar behaviour as the 2-dimensional results already described. For instance, with $M = 20$ computer model runs and $N = 15$ field observations, if the response is fairly simple as occurs with a correlation parameter of 0.6, the calibration model with derivatives has a (10,25) times smaller $MedSE(\theta)$ than the standard model for θ_1 and θ_2 respectively. And, with a more complex response, as is the case when the correlation parameter is 0.05, we see a greater improvement. As an example, when $M = 20$ and $N = 15$, the model with derivatives reduces the error by around (24,20) times when compared to the standard model. Note that even with only $N = 7$ field observations, the $MedSE(\theta)$ is (9,11) times smaller than what the standard model achieves with more than twice the number of field observations.

As the measurement error is increased, the same change in behaviour is seen as in the 2-dimensional study. The results for the 4-dimensional study with $\lambda_f = 10$ are summarized in Table 6.7. Note that the improvement for the simple responses constructed using a correlation of 0.6 is reduced. For the more complex responses simulated using a correlation parameter of 0.05, the advantage of including derivatives increases though it is still less than the simulations using a smaller measurement error. For example, when $M = 20$ and $N = 15$, the model with derivatives has a (7,19) times smaller error than the standard model. With only $N = 7$ field observations, the derivative model has about (4,5) times smaller error than the standard model with more than twice the data. With only partial derivative information, we see only some gains in performance. In the case of $N = 15$ field observations, this model has a $MedSE(\theta)$ which is only about (3,3) times smaller than the standard model.

Finally, we also consider results from the 6-dimensional simulation with $\lambda_f = 10$ in Table 6.8. We only consider responses of moderate complexity with a correlation setting of 0.2 for computational reasons. Again, one can see the uniform improvement of using the calibration model with derivatives over the standard model, and smaller but consistent improvements

	ρ, ϕ :	0.2					
	M :	30			45		
Model		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
i.		3.75e-03	2.86e-03	5.87e-03	2.80e-03	2.71e-03	2.62e-03
ii.	$N = 11$	8.75e-03	8.03e-03	1.23e-02	9.04e-03	1.02e-02	5.13e-03
iii.		1.53e-02	2.69e-02	1.67e-02	2.43e-02	1.89e-02	2.30e-02
i.		1.76e-03	2.92e-03	2.29e-03	1.36e-03	1.93e-03	1.49e-03
ii.	$N = 23$	4.73e-03	7.40e-03	6.64e-03	3.78e-03	4.67e-03	4.08e-03
iii.		1.46e-02	3.06e-02	2.38e-02	5.98e-03	1.61e-02	1.36e-02

Table 6.8: Reported $MedSE(\theta)$ for $(\theta_1, \theta_2, \theta_3)$ and $\lambda_f = 10$ from 6-d calibration simulation study.

when only partial derivative information is available. As an example, in the case where there are $M = 30$ computer model runs and $N = 23$ field observations, the model with derivatives has errors about (8,10,10) times smaller than the standard calibration model. For the same situation, the model with partial derivative information saw a still notable reduction in error of approximately (3,4,4) times when compared to the standard model. One can also notice that the model with derivatives achieves better estimates of the calibration parameter with $N = 11$ field observations as compared to the standard model with $N = 23$ observations.

6.2 Results: Mean Squared Prediction Error and Relative Efficiency

In this section we consider the predictive performance of our calibration models. A good predictive model for the observed physical process is clearly one of the main goals of a calibration experiment, and depends on a good estimate of the calibration parameter. The results of the previous section suggest that the predictive performance of the proposed models that include derivative information should be markedly improved over the standard calibration model. We evaluate this predictive performance by calculating the median of the empirical mean squared prediction errors, $MedMSPE$ (6.2), and also the efficiencies of

these predictions, eff (6.3) relative to the standard calibration model.

The results for the 2-dimensional simulation study with small measurement error is shown in Table 6.9. Inspecting this table shows similar trends as our investigation of the calibration parameter estimate in the previous section. For instance, at small samples sizes of field observations and/or computer model runs, we see the largest improvements from including derivative information as compared to the standard model. This can be quickly discerned by scanning the values of the relative efficiencies denoted in brackets. For simple responses simulated using a correlation of 0.6 with $M = 5$ computer model runs and $N = 3$ field observations, the model with derivatives is 6.78 times more efficient as compared to the standard model. Here, the model with partial derivative information also provides good improvement with a 4.26 times reduction in prediction error.

As we increase either M or N or both, we see these improvements decrease to around a 1.5-2 times reduction in prediction error. One can note that with $N = 3$ field observations, the standard model would need to use about 3 times as many computer model runs to achieve similar predictive performance as the model with derivatives achieves with $M = 5$ computer model runs. Or, with $M = 5$ computer model runs, the standard model would need nearly 5 times the number of field observations to achieve similar predictive performance as achieved with the model using derivatives with $N = 3$ field observations.

	ρ, ϕ :	0.6			0.2			0.05		
Model	M :	5	10	15	5	10	15	5	10	15
i.		3.56e-04	1.83e-04	1.24e-04	1.39e-03	1.53e-04	2.19e-04	4.59e-03	4.18e-04	2.47e-04
		(6.78)	(6.45)	(2.26)	(4.98)	(48.12)	(4.97)	(6.38)	(35.05)	(14.82)
ii.	$N = 3$	5.67e-04	2.10e-04	1.16e-04	2.02e-03	5.42e-04	3.10e-04	4.91e-03	1.22e-03	4.30e-04
		(4.26)	(5.63)	(2.41)	(3.43)	(13.63)	(3.51)	(5.96)	(12.02)	(8.52)
iii.		2.42e-03	1.18e-03	2.80e-04	6.94e-03	7.38e-03	1.09e-03	2.93e-02	1.47e-02	3.66e-03
i.		1.50e-04	6.00e-05	1.07e-04	6.55e-04	1.10e-04	6.24e-05	8.95e-04	2.45e-04	8.72e-05
		(3.97)	(5.58)	(1.48)	(1.39)	(7.01)	(6.95)	(1.19)	(3.67)	(9.90)
ii.	$N = 7$	3.11e-04	7.04e-05	1.05e-04	7.49e-04	3.12e-04	1.08e-04	8.60e-04	4.90e-04	1.98e-04
		(1.92)	(4.76)	(1.51)	(1.21)	(2.47)	(4.02)	(1.23)	(1.83)	(4.36)
iii.		5.96e-04	3.35e-04	1.58e-04	9.10e-04	7.70e-04	4.34e-04	1.06e-03	8.97e-04	8.64e-04
i.		9.37e-05	4.86e-05	5.73e-05	2.36e-04	5.05e-05	3.67e-05	4.09e-04	1.21e-04	5.02e-05
		(2.63)	(5.35)	(1.66)	(1.63)	(5.84)	(6.78)	(1.22)	(3.94)	(7.52)
ii.	$N = 15$	1.51e-04	6.61e-05	3.90e-05	3.35e-04	1.33e-04	4.49e-05	4.22e-04	2.80e-04	1.52e-04
		(1.63)	(3.93)	(2.44)	(1.15)	(2.21)	(5.53)	(1.18)	(1.71)	(2.49)
iii.		2.46e-04	2.60e-04	9.50e-05	3.85e-04	2.95e-04	2.49e-04	5.00e-04	4.79e-04	3.77e-04

Table 6.9: Reported $MedMSPE$ for $\lambda_f = 1000$ from 2-d calibration simulation study. The efficiency relative to the standard model, eff, is shown in brackets.

For more complex responses simulated with a correlation parameter of 0.05, even higher levels of improvement are possible. As an example, with $N = 3$ field observations the model with derivatives is 35 times more efficient than the standard model when $M = 10$, and about 15 times more efficient when $M = 15$. The model with partial derivative information also realizes good gains in this case. Overall, the largest improvement in prediction error was seen with responses simulated with a correlation parameter of 0.2, $N = 3$ field observations and $M = 10$ computer model runs. In this case, the model with derivative information was 48 times more efficient than the standard calibration model. The model with derivatives is still more efficient than the standard calibration model when the standard calibration model is run with 5 times the number of field observations ($N = 15$) and 3 times the number of computer model runs ($M = 15$).

Of course, there are situations where the derivative model only offers performance that is comparable to the standard model. For instance, with a correlation parameter of 0.05, $N = 7$ field observations and $M = 5$ computer model runs, the model with derivatives has similar performance as the standard model. Such situations suggest that there are cases where the amount of data available is so small that the additional derivative information adds little value. However, note that for the same situation, if we increase the number of computer model runs to $M = 15$, the standard model does not improve greatly. Increasing the number of computer model runs for the model with derivatives by the same amount results in a 10 times greater efficiency than the standard model.

	ρ, ϕ :	0.6			0.2			0.05		
Model	M :	5	10	15	5	10	15	5	10	15
i.		1.78e-02	1.32e-02	1.52e-02	2.28e-02	1.33e-02	2.09e-02	4.11e-02	2.18e-02	1.56e-02
		(3.12)	(1.75)	(0.91)	(3.67)	(2.31)	(1.32)	(2.98)	(3.23)	(1.91)
ii.	$N = 3$	2.51e-02	1.36e-02	1.48e-02	3.49e-02	1.43e-02	2.10e-02	7.63e-02	2.86e-02	1.65e-02
		(2.22)	(1.71)	(0.94)	(2.40)	(2.14)	(1.31)	(1.61)	(2.45)	(1.81)
iii.		5.56e-02	2.32e-02	1.38e-02	8.37e-02	3.06e-02	2.75e-02	1.22e-01	7.03e-02	2.99e-02
i.		8.01e-03	6.26e-03	6.02e-03	1.03e-02	8.21e-03	9.30e-03	2.40e-02	4.66e-03	8.05e-03
		(3.95)	(1.66)	(1.12)	(3.66)	(2.86)	(1.21)	(2.76)	(7.58)	(2.00)
ii.	$N = 7$	1.38e-02	7.06e-03	7.32e-03	1.59e-02	8.51e-03	1.11e-02	3.65e-02	7.49e-03	8.99e-03
		(2.30)	(1.47)	(0.92)	(2.36)	(2.76)	(1.02)	(1.82)	(4.71)	(1.79)
iii.		3.17e-02	1.04e-02	6.72e-03	3.77e-02	2.35e-02	1.13e-02	6.63e-02	3.53e-02	1.61e-02
i.		3.79e-03	3.06e-03	3.81e-03	6.52e-03	2.75e-03	3.67e-03	1.48e-02	4.51e-03	3.14e-03
		(3.44)	(1.74)	(1.29)	(3.44)	(3.49)	(2.00)	(1.85)	(4.46)	(2.79)
ii.	$N = 15$	5.37e-03	2.87e-03	3.63e-03	1.22e-02	3.72e-03	3.17e-03	1.98e-02	9.02e-03	3.53e-03
		(2.43)	(1.86)	(1.36)	(1.83)	(2.58)	(2.32)	(1.38)	(2.23)	(2.48)
iii.		1.31e-02	5.33e-03	4.93e-03	2.24e-02	9.59e-03	7.36e-03	2.74e-02	2.01e-02	8.76e-03

Table 6.10: Reported $MedMSPE$ for $\lambda_f = 10$ from 2-d calibration simulation study. The efficiency relative to the standard model, eff, is shown in brackets.

The effect of increasing the measurement error can be seen by inspecting Table 6.10. The absence of large improvements in relative efficiency are immediately apparent, but there are nonetheless consistently good improvements over the standard calibration model. The pattern of improvements also mimic results from the previous table. In the case of a simple response simulated with a correlation setting of 0.6, using the model that includes derivative information with $N = 3$ field observations and $M = 5$ computer model runs provides similar performance as using the standard model with either $N = 15$ field observations or $M = 15$ computer model runs.

Overall, the largest improvement in efficiency and corresponding reduction in prediction error was realized for more complex responses simulated with a correlation parameter of 0.05. Here, with $N = 7$ field observations and $M = 10$ computer model runs, the model with derivative information is about 7.5 times more efficient than the standard model, while the model with partial derivative information is about 4.7 times more efficient. Note that the model with derivative information is still almost 2 times more efficient than when the standard calibration model is used with $N = 15$ field observations and $M = 15$ computer model runs.

The results from the 4-dimensional simulation studies are shown in Table 6.11 for $\lambda_f = 1000$. Here again we see very large improvements in efficiency and corresponding reductions in prediction errors when derivative information is included in calibration experiments. For simple responses generated with a correlation parameter of 0.6, we see that the model with derivatives is about 9 times more efficient than the standard model when $N = 7$ and $M = 20$. With only partial derivative information, we still obtain a 4 times improvement in efficiency. Note here that the model with derivatives is about 4 times more efficient than the standard model is with $N = 15$ field observations and $M = 30$ computer model runs.

Another interesting observation is to notice the behaviour as both the complexity of the

	ρ, ϕ :	0.6		0.2		0.05	
Model	M :	20	30	20	30	20	30
i.		2.53e-03	2.81e-03	1.39e-02	5.95e-03	4.88e-02	2.59e-02
		(9.23)	(5.54)	(6.17)	(13.12)	(3.75)	(6.25)
ii.	$N = 7$	5.85e-03	3.57e-03	2.49e-02	1.56e-02	7.42e-02	4.92e-02
		(3.99)	(4.36)	(3.44)	(5.01)	(2.47)	(3.29)
iii.		2.33e-02	1.56e-02	8.55e-02	7.81e-02	1.83e-01	1.62e-01
i.		1.41e-03	1.11e-03	7.88e-03	3.71e-03	2.90e-02	1.99e-02
		(11.72)	(8.74)	(5.62)	(9.84)	(2.86)	(3.79)
ii.	$N = 15$	2.67e-03	1.59e-03	1.44e-02	8.66e-03	3.87e-02	3.00e-02
		(6.21)	(6.13)	(3.09)	(4.21)	(2.14)	(2.50)
iii.		1.66e-02	9.71e-03	4.43e-02	3.65e-02	8.31e-02	7.52e-02

Table 6.11: Reported $MedMSPE$ for $\lambda_f = 1000$ from 4-d calibration simulation study. The efficiency relative to the standard model, eff, is shown in brackets.

response increases and the number of computer model runs increases. For instance, with $N = 7$ and simple responses using a correlation parameter of 0.6, there is not much difference between the performance of the model with derivatives with $M = 20$ or $M = 30$ computer model runs. At the intermediate level of complexity investigated with a correlation of 0.2, using $M = 30$ computer model runs is about 2.3 times as efficient as using $M = 20$ runs. Yet at the highest level of complexity with a correlation of 0.05, using $M = 30$ computer model runs is now only about 1.9 times as efficient as using $M = 20$ runs. Also, the magnitude of the prediction error at this higher level of response complexity is much larger than the intermediate and simple responses. These results suggest that for the simple response, there is not much benefit seen from increasing the number of computer model runs above $M = 20$. This is likely because for these simple responses, the additional information provided by the collected computer model runs and derivative observations is relatively small. However, at the intermediate complexity, a good improvement is seen indicating that the additional computer model runs are providing useful information. Yet, at the highest level of complexity, the improvement from additional computer model runs is not as strong as seen for the intermediate complexity responses. This might indicate that at this higher level of complexity, more than $M = 30$ computer model runs may provide useful additional

	ρ, ϕ :	0.6		0.2		0.05	
Model	M :	20	30	20	30	20	30
i.		1.45e-02	1.97e-02	3.51e-02	2.31e-02	7.28e-02	5.22e-02
		(3.54)	(2.13)	(2.91)	(3.96)	(2.92)	(3.16)
ii.	$N = 7$	2.00e-02	2.19e-02	5.65e-02	3.24e-02	1.05e-01	7.73e-02
		(2.57)	(1.92)	(1.81)	(2.83)	(2.02)	(2.13)
iii.		5.15e-02	4.21e-02	1.02e-01	9.17e-02	2.12e-01	1.65e-01
i.		8.99e-03	1.12e-02	2.27e-02	1.42e-02	4.81e-02	3.37e-02
		(3.58)	(2.81)	(3.02)	(3.95)	(1.93)	(2.64)
ii.	$N = 15$	1.04e-02	1.17e-02	3.65e-02	2.12e-02	5.97e-02	4.99e-02
		(3.10)	(2.69)	(1.88)	(2.64)	(1.56)	(1.79)
iii.		3.22e-02	3.16e-02	6.85e-02	5.60e-02	9.30e-02	8.91e-02

Table 6.12: Reported $MedMSPE$ for $\lambda_f = 10$ from 4-d calibration simulation study. The efficiency relative to the standard model, eff, is shown in brackets.

information in the model with derivatives to improve the calibrated predictions.

The standard model behaves differently. Here, with $N = 7$ field observations and simple responses simulated with a correlation parameter of 0.6, there is about a 1.5 times improvement when increasing then number of computer model runs from $M = 20$ to $M = 30$. However, for the intermediate and higher complexity responses simulated with correlations of 0.2 and 0.05 respectively, there is hardly any difference between using $M = 20$ and $M = 30$ computer model runs. Also, the magnitude of the prediction errors at these higher levels of complexity are much larger than seen for the simple responses, suggesting that either much more than $M = 30$ computer model runs may be needed for the standard model to improve its calibrated predictions in these cases, or an increase in the number of field observations or both.

The simulation study results for 4-dimensional calibration problems with larger measurement error $\lambda_f = 10$ are shown in Table 6.12, and demonstrate consistently good improvements. Here, the pattern is again much like has been observed thus far. For instance, for simple responses simulated with a correlation of 0.6, the model with derivatives is about

	ρ, ϕ :	0.2	
Model	M :	30	45
i.		1.14e-01	8.81e-02
		(2.08)	(2.52)
ii.	$N = 11$	1.32e-01	1.20e-01
		(1.80)	(1.84)
iii.		2.37e-01	2.22e-01
i.		6.73e-02	5.74e-02
		(1.88)	(2.02)
ii.	$N = 23$	8.29e-02	6.68e-02
		(1.53)	(1.74)
iii.		1.27e-01	1.16e-01

Table 6.13: Reported $MedMSPE$ for $\lambda_f = 10$ from 6-d calibration simulation study. The efficiency relative to the standard model, eff, is shown in brackets.

3.5 times more efficient than the standard model when $N = 7$ and $M = 20$. In addition, the model with derivatives is still about 2 times more efficient than when the standard model is used with $N = 15$ and $M = 30$. Similar improvements are seen with more complex responses. With responses generated using a correlation of 0.05, the model with derivatives is about 2.9 times as efficient as the standard model when $N = 7$ and $M = 20$, while the standard model with $N = 15$ and $M = 30$ provides comparable performance.

Finally, the results of our 6-dimensional simulation study are shown in Table 6.13 when $\lambda_f = 10$ and the correlation parameter is 0.2. Here we generally see a 2-2.5 times improvement in efficiency using the model with derivative information as compared to the standard model, while the model with partial derivative information provides about a 1.5-1.8 times improvement in efficiency. Note that with $N = 11$ and $M = 30$, the model with derivative information has similar performance as the standard model with $N = 23$ and $M = 45$.

Generally, we see that the model with derivative information provides consistent benefits; however, in some cases it provides exceptional improvements. These results imply that achieving exceptional improvements may require some thought placed into the design

of such calibration experiments and the trade-off between the number of field observations collected and the number of computer model runs obtained. Although not our focus in this thesis, this is an interesting observation to investigate in the future.

6.3 Results: Prediction Interval Width and Coverage Probability

In this section, we investigate second order properties of our calibration models, namely the width of the 95% point-wise prediction intervals for the calibrated response estimate, and the corresponding coverage probability of these prediction intervals. These second-order properties are important as the benefit of performing statistical calibration of computer models is to provide for a reasonable and believable measure of uncertainty in locations where data has not been observed so that appropriate inferences and interpretations can be made. We evaluate these second-order properties by calculating the median of the average 95% pointwise prediction interval widths ($MedPIW$) over $N_r = 100$ simulation replicates (6.5) and the median of the empirical 95% coverage probabilities ($MedCP$) taken over the simulation replicates (6.4).

Table 6.14 shows the resulting calculations for the 2-dimensional simulation study when $\lambda_f = 1000$. The first thing one can notice is that the empirical coverage probabilities, $MedCP$, are almost all 100%. In actuality, the coverage probabilities are likely high but not perfect. Unfortunately, we were not able to get an accurate enough estimate with this computationally-limited simulation study.

	ρ, ϕ :	0.6			0.2			0.05		
Model	M :	5	10	15	5	10	15	5	10	15
i.		7.87e-02	7.00e-02	2.84e-01	1.34e-01	6.84e-02	6.57e-02	2.37e-01	8.77e-02	6.92e-02
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
ii.	$N = 3$	1.14e-01	7.02e-02	9.16e-02	1.72e-01	1.00e-01	7.16e-02	2.84e-01	1.41e-01	8.39e-02
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
iii.		2.02e-01	1.31e-01	7.70e-02	3.33e-01	2.96e-01	1.40e-01	4.54e-01	4.03e-01	2.24e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(0.950)	(1.000)	(1.000)
i.		5.27e-02	4.64e-02	2.69e-01	8.98e-02	4.64e-02	4.25e-02	1.08e-01	6.22e-02	4.36e-02
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
ii.	$N = 7$	7.58e-02	4.62e-02	6.54e-02	1.01e-01	6.98e-02	4.80e-02	1.14e-01	9.01e-02	5.85e-02
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
iii.		9.99e-02	7.99e-02	5.21e-02	1.13e-01	1.07e-01	8.86e-02	1.21e-01	1.19e-01	1.08e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(0.967)	(1.000)	(1.000)
i.		4.36e-02	3.23e-02	2.95e-01	6.38e-02	3.19e-02	3.02e-02	7.44e-02	4.32e-02	3.10e-02
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
ii.	$N = 15$	5.77e-02	3.40e-02	6.33e-02	7.21e-02	5.02e-02	3.30e-02	7.85e-02	6.48e-02	4.80e-02
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
iii.		6.88e-02	5.99e-02	3.82e-02	7.72e-02	7.31e-02	6.23e-02	8.14e-02	7.94e-02	7.29e-02
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)

Table 6.14: Reported $MedPIW$ for $\lambda_f = 1000$ from 2-d calibration simulation study. The median coverage probability, $MedCP$, is shown in brackets.

The main difference between the model with derivatives and the standard model comes through the widths of the 95% pointwise prediction intervals. For instance, with $N = 3$ field observations and $M = 5$ computer model runs, the model with derivatives has a $MedPIW$ of $7.87e-02$, which is about 2.6 times smaller than the standard models $MedPIW$ of $2.02e-01$. Here, the model with partial derivative information has an approximately 1.8 times smaller interval width than the standard model. If we consider more complex responses, then with a correlation parameter of 0.05, $N = 3$ field observations and $M = 5$ computer model runs, the model with derivatives has a 1.9 times smaller prediction interval than the standard model.

As we increase the number of computer model runs to $M = 15$, the model with derivatives now has an interval width which is about 3.2 times smaller than the standard model. If instead we increase the number of field observations to $N = 15$ then the prediction interval width for the model with derivatives is similar to the standard model. If we consider both $N = 15$ and $M = 15$, then we see that the model with derivatives has a 2.3 times smaller interval than the standard model. Note here that increasing the number of computer model runs leads to only a small reduction in the prediction interval widths when using the standard model, while the model with derivatives realizes noticeable reductions in the prediction interval widths as the number of computer model runs increases. This corresponds with what we have seen earlier, namely that the model with derivative information is able to more efficiently borrow information from the computer model runs in a calibration experiment.

	ρ, ϕ :	0.6			0.2			0.05		
Model	M :	5	10	15	5	10	15	5	10	15
i.		6.51e-01	6.58e-01	8.26e-01	7.11e-01	6.21e-01	6.31e-01	7.74e-01	6.59e-01	6.38e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
ii.	$N = 3$	7.14e-01	6.33e-01	6.70e-01	8.26e-01	6.46e-01	6.49e-01	9.26e-01	7.53e-01	6.63e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
iii.		8.85e-01	7.13e-01	6.68e-01	1.04e+00	7.80e-01	7.11e-01	1.10e+00	9.18e-01	7.81e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(0.933)	(1.000)	(1.000)
i.		4.46e-01	4.31e-01	5.70e-01	4.75e-01	4.16e-01	4.21e-01	5.83e-01	4.27e-01	4.20e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
ii.	$N = 7$	5.01e-01	4.43e-01	4.53e-01	6.10e-01	4.65e-01	4.23e-01	7.37e-01	5.13e-01	4.35e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
iii.		6.70e-01	4.74e-01	4.52e-01	7.78e-01	5.87e-01	4.67e-01	8.41e-01	7.07e-01	5.54e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(0.900)	(1.000)	(1.000)
i.		3.12e-01	2.90e-01	5.56e-01	3.54e-01	3.05e-01	2.92e-01	4.49e-01	3.04e-01	2.85e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
ii.	$N = 15$	3.72e-01	2.91e-01	3.20e-01	4.53e-01	3.09e-01	2.97e-01	5.22e-01	3.90e-01	2.92e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)
iii.		4.97e-01	3.57e-01	3.11e-01	5.48e-01	4.40e-01	3.46e-01	6.07e-01	5.50e-01	4.05e-01
		(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(0.967)	(1.000)	(1.000)

Table 6.15: Reported $MedPIW$ for $\lambda_f = 10$ from 2-d calibration simulation study. The median coverage probability, $MedCP$, is shown in brackets.

	ρ, ϕ :	0.6		0.2		0.05	
Model	M :	20	30	20	30	20	30
i.		5.67e-01	5.77e-01	7.08e-01	6.46e-01	9.47e-01	8.18e-01
		(0.998)	(0.999)	(0.993)	(0.989)	(0.965)	(0.974)
ii.	$N = 7$	6.01e-01	5.86e-01	8.10e-01	7.55e-01	1.11e+00	9.57e-01
		(0.998)	(0.996)	(0.982)	(0.977)	(0.951)	(0.959)
iii.		7.90e-01	7.09e-01	1.08e+00	1.02e+00	1.26e+00	1.21e+00
		(0.980)	(0.982)	(0.936)	(0.932)	(0.859)	(0.850)
i.		3.91e-01	3.81e-01	5.02e-01	4.32e-01	7.15e-01	6.38e-01
		(0.999)	(0.999)	(0.976)	(0.986)	(0.960)	(0.966)
ii.	$N = 15$	4.12e-01	3.98e-01	5.86e-01	5.14e-01	8.12e-01	7.65e-01
		(0.996)	(0.993)	(0.986)	(0.986)	(0.957)	(0.972)
iii.		6.00e-01	5.07e-01	7.92e-01	7.40e-01	9.07e-01	9.10e-01
		(0.984)	(0.987)	(0.945)	(0.950)	(0.886)	(0.899)

Table 6.16: Reported $MedPIW$ for $\lambda_f = 1000$ from 4-d calibration simulation study. The median coverage probability, $MedCP$, is shown in brackets.

The results with an increased level of measurement error ($\lambda_f = 10$) are shown in Table 6.15. Here we again see that the coverage probabilities are too high to detect in all but a few cases. At this level of measurement error, there is not a consistent difference in prediction interval widths for the simple responses simulated using a correlation parameter of 0.6. However, as the responses become more complex, the model with derivative information and the model with partial derivative information can provide tighter prediction intervals. For instance, with responses generated using a correlation parameter of 0.05, $N = 3$ field observations and $M = 5$ computer model runs, the model with derivative information has a 1.4 times smaller prediction interval width than the standard model. The model with partial derivative information is not quite as good in this case. Increasing the number of field observations to $M = 15$ and computer model runs to $N = 15$, we see that the model with derivatives is again about 1.42 times smaller than the standard model, while the model with partial derivative information is about 1.38 times smaller.

Similar results are seen in the 4-dimensional simulation study where the results with $\lambda_f = 1000$ are shown in Table 6.16 and with $\lambda_f = 10$ in Table 6.17. In both cases the

	ρ, ϕ :	0.6		0.2		0.05	
Model	M :	20	30	20	30	20	30
i.		6.13e-01	6.21e-01	7.40e-01	6.57e-01	1.05e+00	9.10e-01
		(0.966)	(0.943)	(0.929)	(0.940)	(0.938)	(0.950)
ii.	$N = 7$	6.40e-01	6.33e-01	8.58e-01	7.55e-01	1.19e+00	1.09e+00
		(0.965)	(0.943)	(0.921)	(0.940)	(0.941)	(0.934)
iii.		9.01e-01	8.18e-01	1.20e+00	1.09e+00	1.41e+00	1.39e+00
		(0.956)	(0.948)	(0.914)	(0.912)	(0.878)	(0.893)
i.		4.23e-01	4.33e-01	5.80e-01	4.97e-01	8.13e-01	7.17e-01
		(0.956)	(0.942)	(0.952)	(0.922)	(0.938)	(0.923)
ii.	$N = 15$	4.45e-01	4.61e-01	7.14e-01	5.94e-01	9.17e-01	8.72e-01
		(0.955)	(0.935)	(0.949)	(0.942)	(0.930)	(0.946)
iii.		6.56e-01	7.33e-01	9.21e-01	8.58e-01	1.04e+00	1.06e+00
		(0.948)	(0.949)	(0.925)	(0.932)	(0.907)	(0.922)

Table 6.17: Reported $MedPIW$ for $\lambda_f = 10$ from 4-d calibration simulation study. The median coverage probability, $MedCP$, is shown in brackets.

coverage probabilities are mostly at or above the nominal 95% level. For more complex responses, we see that the coverage provided by the standard model can be poor with a smaller sample of field observations and computer model runs as compared to the models that include derivative information. When the measurement error is small ($\lambda_f = 1000$), the model with derivatives generally has prediction intervals that are around 1.2-1.7 times smaller than the standard model, while using only partial derivative information results in prediction intervals that are about 1.1-1.5 times smaller than the standard model. At a higher level of measurement error ($\lambda_f = 10$), the model with derivatives has prediction intervals approximately 1.3-1.75 times smaller than the standard model, while the model with partial derivative information has intervals about 1.1-1.6 times smaller than the standard model.

Finally, the results from the 6-dimensional simulation study with $\lambda_f = 10$ are shown in Table 6.18. Here again the coverage probabilities exhibit good behaviour for the model with derivative information and the model with partial derivative information, although the standard model is poor in all the cases investigated. The model with derivative information has

	ρ, ϕ :	0.2	
Model	M :	30	45
i.		1.18e+00	1.11e+00
		(0.926)	(0.937)
ii.	$N = 11$	1.28e+00	1.27e+00
		(0.918)	(0.927)
iii.		1.40e+00	1.42e+00
		(0.844)	(0.854)
i.		9.38e-01	8.66e-01
		(0.935)	(0.937)
ii.	$N = 23$	9.83e-01	9.73e-01
		(0.925)	(0.931)
iii.		1.07e+00	1.11e+00
		(0.866)	(0.892)

Table 6.18: Reported $MedPIW$ for $\lambda_f = 10$ from 6-d calibration simulation study. The median coverage probability, $MedCP$, is shown in brackets.

prediction intervals about 1.1-1.3 times smaller than the standard model, while the model with partial derivative information has prediction intervals that are about 1.1 times smaller.

6.4 Discussion

The results presented indicate that the proposed calibration model with full derivative information is uniformly better than the standard calibration model across a wide range of experimental conditions. Incorporating derivative information in the calibration enabled more accurate estimates of the unknown true calibration parameter, a reduction in mean squared prediction error, and tighter 95% pointwise prediction intervals. The largest improvements were realized with either a small sample of field observations (N), a small number of computer model runs (M), or both. In many cases, the standard model could only match the performance of the proposed model by collecting many times the amount of field observations, computer model runs, or both.

The improvements achieved using the model with full derivative information were also much more noticeable with the more complex simulated responses using a correlation of 0.2 or 0.05 than the simple responses constructed with a correlation setting of 0.6. An intuitive understanding of these levels of complexity can be discerned by looking at the example shown in Figure 5.3, which was constructed with a correlation of 0.05. In practice, one can easily imagine real-world computer models with similar or higher response complexity than shown in this figure, in which case one expects the model with full derivative information to enable consistently good levels of improvement, and in some cases exceptional levels of improvement, in a calibration experiment.

With only partial derivative information available, our results indicate that incorporating this information will still provide consistent benefits, but the level of improvement is much reduced when compared to the model that includes full derivative information. Whether the level of improvement is large enough to warrant our proposed approach is a consideration for the experimenter. Exploring this aspect for different experimental costs of using a regular computer model and a computer model that provides partial (or more) derivative information would be an interesting topic for future research.

There are some aspects of the relative performance of the proposed models with full derivative information or partial derivative information and the standard model that we have not shown simply because the number of tables and plots would be overwhelming. We summarize these now. First, fitting a given model to a particular dataset typically results in more accurate posterior distributions with reduced variability for all model parameters when using the models with full or partial derivative information as opposed to the standard model. Second, for the simulation studies summarized in the current chapter, in situations where the proposed models provided benefits, the variability of the point estimates were much lower than the standard calibration model. If instead the proposed models did not provide benefits, the variability of the point estimates were similar to the standard model.

For instance, for the 4-dimensional simulation study, we noted a large improvement in $MedSE(\theta)$ and $MedMSPE$ for the model with derivatives for a response simulated with a correlation setting of 0.05 with $N = 15$ field observations and $M = 20$ computer model runs. Figure 6.1 shows the boxplots of the difference between the estimated calibration parameters, and the empirical mean squared prediction error of the resulting calibrated prediction of the true field process over the 100 replicates from the simulation. These plots confirm that the variability of the point estimates from our simulations are much smaller for the models using derivative information in this case. On the other hand, with a correlation setting of 0.6, we had noted a minor improvement over the standard model when $N = 15$ and $M = 20$. The corresponding plots in Figure 6.2 now indicate that the variability of the calibration parameter estimates over the simulation study replicates is similar for all three models, although there is still a notable reduction for prediction error.

As a final point, one should note that in practice, the computer model and field process being investigated is unlikely to be a draw from the stationary GP model we use in this thesis, even though it is commonly believed to generally be a useful model. In these cases, one might expect that the addition of derivative information in a calibration experiment might provide an even greater benefit to practical calibration. Investigating such robustness of the three calibration models investigated to departures from the stationary GP assumption would be an interesting aspect to explore in the future.

In the next chapter, we will finally return to the original full calibration model (2.7) which attempts to account for possible bias in the computer model, and extend this model to incorporate derivative information.

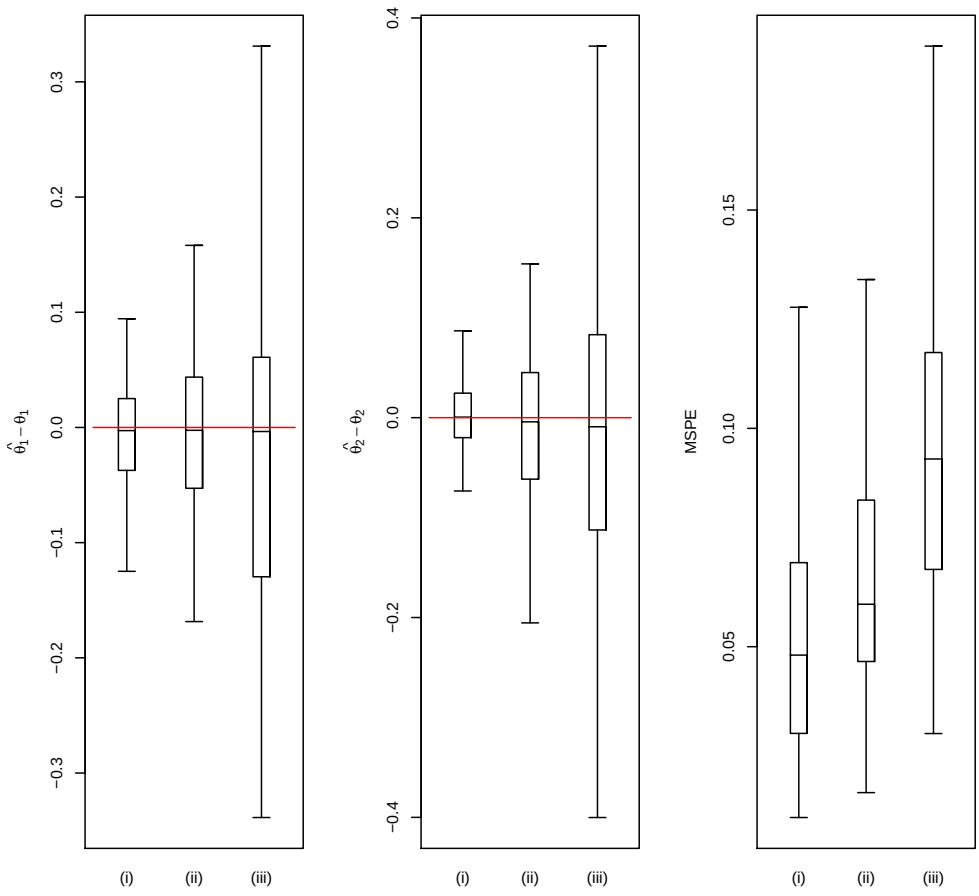


Figure 6.1: Boxplots of deviation of both calibration parameter estimates from true calibration parameter (left, middle panels) over $N_r = 100$ replicates for 4-dimensional simulation study with $\lambda_f = 10$ and a correlation setting of 0.05. Corresponding boxplot of empirical mean squared prediction errors for each simulation replicate shown in right panel.

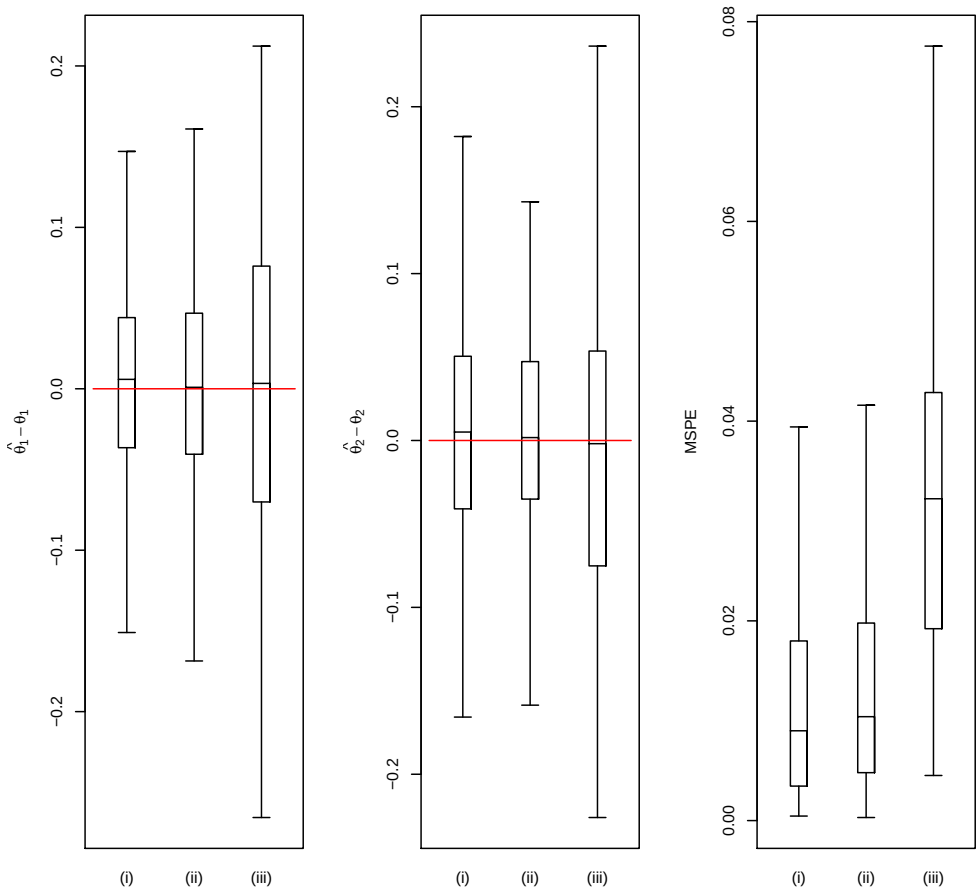


Figure 6.2: Boxplots of deviation of both calibration parameter estimates from true calibration parameter (left, middle panels) over $N_r = 100$ replicates for 4-dimensional simulation study with $\lambda_f = 10$ and a correlation setting of 0.6. Corresponding boxplot of empirical mean squared prediction errors for each simulation replicate shown in right panel.

Chapter 7

Calibration with Derivatives: Computer Model Bias

The final methodological development in this thesis aims to combine the calibration model that incorporates derivative information developed in section 5.3 with the most general version of the standard calibration model which allows for possible bias in the computer simulator. The idea underlying this model is a recognition that the computer model may not actually be able to represent reality exactly. A pragmatic view is to hope that the computer model gets close enough that the calibration parameter can still be reasonably estimated, and to add a bias term to account for the behaviour that is not well represented in the computer model so that the models' predictive performance is good. On the other hand, a pessimistic view is to expect that the computer model is such a poor description of reality that in actuality the bias term is largely responsible for achieving a good fit to the field observations, leading to poor estimation of the calibration parameter.

This issue is not new - as mentioned earlier in this thesis, it was noted in the discussion of Kennedy and O'Hagan (2001). And indeed, we do not resolve this issue in the current work. However, given the large improvements in parameter estimates realized for the calibration model with no bias in the previous two chapters, our aim is to investigate whether incorporating such derivative information into the full calibration model with bias can bring

practical improvements as evaluated via simulation.

The proposed model follows the development of section 5.3, where we again have field observations $y^f(\boldsymbol{x})$ observed at locations $\boldsymbol{x} \in \mathcal{D}$ for some fixed domain $\mathcal{D}$, and also computer model observations $y^c(\boldsymbol{x}, \boldsymbol{t})$ observed over the same domain and at calibration settings $\boldsymbol{t} \in \mathcal{T}$ for some fixed domain $\mathcal{T}$. Suppose there are D space-time dimensions, so that each location $\boldsymbol{x}$ is a $D \times 1$ vector, and suppose there are L calibration parameter dimensions, so that each calibration setting $\boldsymbol{t}$ is an $L \times 1$ vector. Then, when all the partial first derivative processes are available, our model is:

$$\begin{aligned}
y^f(\boldsymbol{x}) = &\quad \eta(\boldsymbol{x}, \boldsymbol{\theta}) + b(\boldsymbol{x}) + \epsilon, \\
y^c(\boldsymbol{x}, \boldsymbol{t}) = &\quad \eta(\boldsymbol{x}, \boldsymbol{t}) \\
y^c_1(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial x_1} \\
&\vdots \\
y^c_D(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial x_D} \\
y^c_{D+1}(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial t_1} \\
&\vdots \\
y^c_{D+L}(\boldsymbol{x}, \boldsymbol{t}) = &\quad \tfrac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial t_L},
\end{aligned} \tag{7.1}$$

where there are D partial derivative processes with respect to the space-time dimensions and L partial derivative processes with respect to the calibration parameter dimensions. Note that we have overloaded the subscripts to make the notation compact and readable. That is, if the subscript is given by u, then $1 \leq u \leq D$ represents differentiation with respect to the uth dimension of $\boldsymbol{x}$, while $(D+1) \leq u \leq (D+L)$ represents differentiation with respect to the $(u-D)^{th}$ dimension of $\boldsymbol{t}$. For example, if $D = L = 2$ and $u = 1$, then $y^c_u(\boldsymbol{x}, \boldsymbol{t}) = \frac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial x_1}$, while if $u = 3$ then $y^c_u(\boldsymbol{x}, \boldsymbol{t}) = \frac{\partial \eta(\boldsymbol{x}, \boldsymbol{t})}{\partial t_1}$.

Let $\mathbf{Y}^f = [y^f(\boldsymbol{x}_1), \ldots, y^f(\boldsymbol{x}_N)]$ be the $N \times 1$ vector of field observations, $\mathbf{Y}^c = [y^c(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, y^c(\boldsymbol{x}_M, \boldsymbol{t}_M)]$ be the $M \times 1$ vector of computer model observations and $\mathbf{Y}^c_u = [y^c_u(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, y^c_u(\boldsymbol{x}_M, \boldsymbol{t}_M)]$ be the $M \times 1$ vector of computer model derivative observations where the

derivative is with respect to the uth input. Then, letting $\mathbf{Y} = [\mathbf{Y}^f, \mathbf{Y}^c, \mathbf{Y}^c_1, \ldots, \mathbf{Y}^c_D, \mathbf{Y}^c_{D+1}, \ldots, \mathbf{Y}^c_{D+L}]$, the joint distribution of the data conditional on the parameters is

$$\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\psi}, \lambda_f, \lambda_z, \lambda_b \sim \mathcal{N}\left(0, \tfrac{1}{\lambda_z}\mathbf{R}^Y + \tfrac{1}{\lambda_b}\begin{bmatrix} \mathbf{Q}^f & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tfrac{1}{\lambda_f}\begin{bmatrix} \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)$$

where $\mathbf{R}^Y$ is as defined in (5.5), $\mathbf{Q}^f$ is the $N \times N$ correlation matrix for the bias process using the squared-exponential model (2.1) and parametrized by the $D \times 1$ vector of correlation parameters $\boldsymbol{\psi}$, λ_b is the precision of the bias process model, and all other parameters are as before (see section 5.3).

The posterior distribution of the parameters,

$$\begin{aligned} \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\psi}, \lambda_f, \lambda_z, \lambda_b|\mathbf{Y} \propto \; & L(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\psi}, \lambda_f, \lambda_z, \lambda_b) \times \textstyle\prod_{l=1}^{L} \pi(\theta_l) \times \prod_{k=1}^{K} \pi(\rho_k) \\ & \times \textstyle\prod_{l=1}^{L} \pi(\phi_l) \prod_{k=1}^{K} \pi(\psi_k) \times \pi(\lambda_f) \times \pi(\lambda_z) \times \pi(\lambda_b) \end{aligned} \tag{7.2}$$

is sampled using Markov Chain Monte Carlo (MCMC), where we again use an adaptive Metropolis-Hastings algorithm with uniform proposal distributions and allowing the step size to adapt for the first 50% of the iterations before allowing the chain to burn-in. We take a sample of size S from the end of the chain as draws from the stationary distribution.

Predictions of the field, $\mathbf{Y}^* = \left[\eta(\boldsymbol{x}^*_1, \boldsymbol{\theta}) + b(\boldsymbol{x}^*_1), \ldots, \eta(\boldsymbol{x}^*_{N_p}, \boldsymbol{\theta}) + b(\boldsymbol{x}^*_{N_p})\right]$, are constructed at the N_p prediction locations $\boldsymbol{x}^* = (\boldsymbol{x}^*_1, \ldots, \boldsymbol{x}^*_{N_p})$ similar to the formulas with no bias from Chapter 5. First, note that the joint distribution of the process at the prediction locations and the observed data is specified as

$$\begin{aligned} & (\mathbf{Y}^*, \mathbf{Y})|\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\psi}, \lambda_f, \lambda_z, \lambda_b \sim \\ & \mathcal{N}\left(0, \frac{1}{\lambda_z}\begin{bmatrix} \mathbf{R}^* & \mathbf{R}^{*Y} \\ \mathbf{R}^{Y*} & \mathbf{R}^Y \end{bmatrix} + \frac{1}{\lambda_b}\begin{bmatrix} \mathbf{Q}^* & \mathbf{Q}^{*f} & 0 \\ \mathbf{Q}^{f*} & \mathbf{Q}^f & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{\lambda_f}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix}\right), \end{aligned}$$

where $\mathbf{W} = diag(\mathbf{I}_N, \mathbf{0}_{M \times (D+L+1)})$, so that the predictive distribution for the process at the unobserved locations, $\mathbf{Y}^*|\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \boldsymbol{\psi}, \lambda_f, \lambda_z, \lambda_b$, can be expressed using the usual properties of multivariate normal distributions (see Appendix B). The density of the posterior

predictive distribution is then

$$f(\mathbf{Y}^*|\mathbf{Y}) = \int\int\int\int\int\int\int f(\mathbf{Y}^*|\mathbf{Y},\cdot)\pi(\boldsymbol{\theta},\boldsymbol{\rho},\boldsymbol{\phi},\boldsymbol{\psi},\lambda_f,\lambda_z,\lambda_b|\mathbf{Y})d\boldsymbol{\theta}d\boldsymbol{\rho}d\boldsymbol{\phi}d\boldsymbol{\psi}d\lambda_f d\lambda_z d\lambda_b$$

which can be used to construct the mean estimate of the predictor and pointwise prediction intervals using quantiles of the distribution. In practice, these estimates are approximated using samples from the posterior predictive distribution constructed from the samples of the posterior distribution for the parameters (7.2).

We may also construct predictions of the computer model, $\mathbf{Z}^* = \left[\eta(\boldsymbol{x}_1^*,\boldsymbol{\theta}),\ldots,\eta(\boldsymbol{x}_{N_p}^*,\boldsymbol{\theta})\right]$ at the N_p prediction locations. In this case, the joint distribution of the process at the prediction locations and the observed data is

$$(\mathbf{Z}^*,\mathbf{Y})|\boldsymbol{\theta},\boldsymbol{\rho},\boldsymbol{\phi},\boldsymbol{\psi},\lambda_f,\lambda_z,\lambda_b \sim \mathcal{N}\left(0, \frac{1}{\lambda_z}\begin{bmatrix}\mathbf{R}^* & \mathbf{R}^{*Y}\\ \mathbf{R}^{Y*} & \mathbf{R}^{Y}\end{bmatrix} + \frac{1}{\lambda_b}\begin{bmatrix}0 & 0 & 0\\ 0 & \mathbf{Q}^f & 0\\ 0 & 0 & 0\end{bmatrix} + \frac{1}{\lambda_f}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}\end{bmatrix}\right),$$

where $\mathbf{W} = diag(\mathbf{I}_N, \mathbf{0}_{M\times(D+L+1)})$, so that the predictive distribution, $\mathbf{Z}^*|\mathbf{Y},\boldsymbol{\theta},\boldsymbol{\rho},\boldsymbol{\phi},\boldsymbol{\psi},\lambda_f,\lambda_z,\lambda_b$, can be expressed using the usual properties of multivariate normal distributions (see Appendix B). The density of the posterior predictive distribution is then

$$f(\mathbf{Z}^*|\mathbf{Y}) = \int\int\int\int\int\int\int f(\mathbf{Z}^*|\mathbf{Y},\cdot)\pi(\boldsymbol{\theta},\boldsymbol{\rho},\boldsymbol{\phi},\boldsymbol{\psi},\lambda_f,\lambda_z,\lambda_b|\mathbf{Y})d\boldsymbol{\theta}d\boldsymbol{\rho}d\boldsymbol{\phi}d\boldsymbol{\psi}d\lambda_f d\lambda_z d\lambda_b$$

which can be used to construct the mean estimate of the predictor and pointwise prediction intervals using quantiles of the distribution. In practice, these estimates are also approximated using samples from the posterior predictive distribution $\mathbf{Z}^*|\mathbf{Y}$ constructed from the samples of the posterior distribution for the parameters (7.2).

We will compare the proposed calibration model incorporating derivative information with the presence of bias to the standard calibration model with bias (2.7). For both models, we use independent uniform(0,1) priors on the correlation parameters $\boldsymbol{\rho}$, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, as well as uniform(0,1) priors on the unknown calibration parameter $\boldsymbol{\theta}$. We also place uninformative

priors on the precision parameters, $\lambda_z \propto I(\lambda_z > 0)$, $\lambda_f \propto I(\lambda_f > 0)$ and $\lambda_b \propto I(\lambda_b \geq 0)$, where $I(\cdot)$ is the indicator function. These priors were selected for two related reasons. The first is that when making initial investigations using informative priors, the MCMC would often have difficulty converging and/or would not converge to the correct parameter values for both models. The second is that, given the perceived unidentifiability of the standard model, it would be useful to understand what differences (if any) are seen in posterior estimates of parameters between the standard model and the proposed model with derivatives when the prior lends no additional information on the true values of the parameters.

7.1 Example

Our example closely follows the bias-free example of section 5.4. Here again, we consider the 2-dimensional calibration problem where the first dimension represents spatial location and the second dimension is the univariate calibration parameter. The difference in this situation is that the computer model is a biased representation of reality. For this example, we calculated the computer model, $\eta(x,t)$, at $M = 15$ settings of (x,t) and $\eta(x,\theta)$ at $N = 7$ locations. This was done by drawing a realization of the calibration model with derivatives (7.1) with $\lambda_z = 1, \rho = 0.05$ and $\phi = 0.05$, where the M computer trials and N field locations were selected by a space-filling design (Johnson et al., 1990) and choosing $\theta = 0.5$ as the true calibration parameter setting. We also simultaneously draw observations of $\eta(x^*,\theta)$ over an equi-distant prediction grid of $N_p = 30$ locations, which we will use later for comparing the predictive performance of our calibrated estimates of the computer model and the field. Finally, we also draw a realization of the bias $b(x)$ from a GP with $\lambda_b = 1$ and correlation $\psi = 0.6$ at the N field locations and N_p prediction locations, and take the true field process to be $\eta(x,\theta) + b(x)$ calculated at the $N = 7$ observation locations and $N_p = 30$ prediction locations. Let $\mathbf{Z} = [\eta(x_1,\theta),\ldots,\eta(x_7,\theta)]$ be the vector of computer model runs generated at the true calibration setting θ observed at the field locations, while $\mathbf{B} = [b(x_1),\ldots,b(x_7)]$ is the vector of the true bias process also generated at the field locations.. Then $\mathbf{Z} + \mathbf{B}$ is the 7×1 vector of the true field process calculated at the field locations. Similarly, we have

Model	λ_z	λ_f	λ_b	ρ	ϕ	ψ	θ
Standard	0.03	0.03	0.03	<0.01	<0.01	0.52	0.02
Derivatives	0.06	0.11	<0.01	0.09	0.04	0.29	0.35

Table 7.1: Geweke diagnostics (p-values) for the standard calibration with bias model and the calibration with bias model with derivatives comparing the first 20% of the MCMC chain to the last 50% of the MCMC chain.

$\mathbf{Y}^* = \mathbf{Z}^* + \mathbf{B}^*$ as the 30×1 vector of the true field process calculated at the prediction locations.

The field observations, $\mathbf{Y}^f = \mathbf{Z} + \mathbf{B} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_7]^T$, were constructed by adding i.i.d. normally distributed mean-zero error with $\lambda_f = 10/v_f$ to the drawn field process $\mathbf{Z} + \mathbf{B}$, where v_f is approximately the variance of the true field process estimated using the sample variance of $\mathbf{Y}^*$ from the $N_p = 30$ prediction grid locations. In other words, we add observation error with a variance that is about 10% of the variance of the underlying true physical process. So, our data consists of field observations $\mathbf{Y}^f = [\eta(x_1, \theta) + b(x_1) + \epsilon_1, \ldots, \eta(x_7, \theta) + b(x_7) + \epsilon_7]$ where the ϵ_i's are drawn from $\mathcal{N}(0, \frac{1}{\lambda_f})$, and computer model observations $\mathbf{Y}^c = [\eta(x_8, t_8), \ldots, \eta(x_{22}, t_{22})]$. In addition, we have the derivative observations $\mathbf{Y}_1^c = \left[\frac{\partial \eta(x_8,t_8)}{\partial x}, \ldots, \frac{\partial \eta(x_{22},t_{22})}{\partial x}\right]$ and $\mathbf{Y}_2^c = \left[\frac{\partial \eta(x_8,t_8)}{\partial t}, \ldots, \frac{\partial \eta(x_{22},t_{22})}{\partial t}\right]$.

The calibration model with derivative information and bias as well as the standard model with bias was fit to this simulated data using the priors outlined above and the adaptive MCMC algorithm explained earlier (see section 5.4). Since the model with bias is generally more challenging to fit to data, our MCMC was run for $S_0 = 10,000$ iterations. The Geweke diagnostic is summarized in Table 7.1, from which we can see the statistics are often significantly different at the 0.05 level. This suggests that there is a difference between the start of the chain and the end, implying that the chain is not reaching the stationary distribution quickly.

We also applied the Raftery-Lewis diagnostic, specifying $q = (0.025, 0.975), r = 0.01$ and

Model	Quantile	L/Burn	λ_z	λ_f	λ_b	ρ	ϕ	ψ	θ
Standard	0.025	L	937	937	937	937	937	937	937
	0.025	Burn	846	20	33	22	15	14	16
	0.975	L	937	937	937	937	937	937	937
	0.975	Burn	24	12	85	265	187	18	33
Derivatives	0.025	L	937	937	937	937	937	937	937
	0.025	Burn	58	14	30	14	15	14	14
	0.975	L	937	937	937	937	937	937	937
	0.975	Burn	18	15	27	120	33	19	32

Table 7.2: Raftery-Lewis diagnostics for minimum chain length (M) and burn-in (Burn) for the standard calibration model and the calibration model with derivatives with $r = 0.01, s = 0.95$ for both the 0.025 and 0.975 quantiles.

$s = 0.95$. The results shown in Table 7.2 also indicate that the chain has not burned in. Although the calculation suggests that a sample of size $S = 1000$ from the chain would be sufficient, we take $S = 2000$ to be conservative since we will use the same settings in the upcoming simulation study. So, the first 8000 iterations were dropped as burn-in, and the last $S = 2000$ iterations were taken as our sample from the posterior. Re-calculating both diagnostics after dropping this burn-in now suggested that the chains had reached their stationary distributions (not shown).

The traceplots for the $S = 2000$ samples of the posteriors for both models are shown in Figure 7.1, where the traces for the model with derivatives are in the left column, and the standard model in the right column. The traceplots appear reasonable, although there is a departure around the 1500th iteration in the plot for λ_b in the standard model. The true parameter values are plotted as solid red lines, which suggest both models are able to estimate the parameters fairly well. The model with derivatives again appears to better estimate the correlation parameters for the computer model observations (i.e. ρ, ϕ). However, both models appear to have great difficulty estimating the correlation parameter of the bias, ψ. The autocorrelation plots for the parameters are shown in Figure 7.2. Note

that there still exists strong autocorrelation in the chosen posterior sample. Using a thinning interval of 10 was found to remove this autocorrelation, and in practice one would run the chain much longer to get a good sample of the posterior after thinning. However, for our purposes, this is simply not computationally possible if one wants to perform a simulation study in a reasonable amount of time, so we proceed using the posterior sample selected.

The calibrated field responses are shown in the left panel of Figure 7.3, while the calibrated computer model response is shown in the right panel. One can immediately note that the prediction of the field process was quite similar for both models, suggesting that there is little to be gained by including the derivative information. Similarly, the prediction of the computer model response at the unknown true value of the calibration parameter θ plotted in the right panel shows minor differences between the two models. This simple example foreshadows that the benefits of including derivative information when calibrating under the assumption of computer model bias may be subtle.

7.2 Simulation Study

Following the example just outlined, we proceed to investigate the behaviour of these two models via simulation study. For this study, our interest will be in five measures of the performance of these models. The first measure, (6.1), evaluates how well the calibration parameter is estimated. The next two are given in (6.2) and (6.3), where (6.2) calculates the median of the mean squared prediction errors of the field over the simulation study replicates while (6.3) reports the relative efficiency with respect to the standard model. Both of these measures were previously explained in Chapter 6.

The last two measures are simply the median of the mean squared prediction error for the calibrated prediction of the computer model at the unknown true calibration parameter, and the corresponding relative efficiency. For the ith simulation, we calculate the empirical

mean squared prediction error of the calibrated computer model as

$$MSPE\eta_i = \frac{1}{N_p}(\widehat{\mathbf{Z}^*} - \mathbf{Z}^*)^T(\widehat{\mathbf{Z}^*} - \mathbf{Z}^*)$$

where $\widehat{\mathbf{Z}^*} = \widehat{E}[\mathbf{Z}^*|\mathbf{Y}]$ is the $N_p \times 1$ vector of calibrated computer model predictions calculated at the N_p prediction locations and $\mathbf{Z}^*$ is the $N_p \times 1$ vector of the true values of the computer model evaluated at the true calibration parameter at these same locations. The median over all the simulations of these empirical squared prediction errors is reported as

$$MedMSPE\eta = \quad median(MSPE\eta). \tag{7.3}$$

The relative efficiency of the calibrated computer model that includes derivative information versus the standard calibration model is calculated as

$$eff\eta = \quad \frac{MedMSPE\eta_{std}}{MedMSPE\eta}, \tag{7.4}$$

where $MedMSPE\eta_{std}$ is the median of the mean squared prediction errors when using the standard calibration model, while the denominator is the mean squared prediction error when including derivative information in the calibration model.

The simulations investigated 2-dimensional, 4-dimensional and 6-dimensional calibration problems using realizations randomly drawn from the calibration model with bias and full derivative information as outlined in the above example. In the 2-dimensional case, there is 1 space-time dimension and 1 calibration parameter dimension. In the 4-dimensional case, there are 2 space-time dimensions and 2 calibration parameter dimensions, while in the 6-dimensional case, there are 3 space-time dimensions and 3 calibration parameter dimensions. For each of these basic calibration problem types, $N_r = 100$ replicate simulations were performed at a fixed level of parameter settings. The settings investigated in the simulation included: the correlation parameter for the space-time dimensions, $\boldsymbol{\rho}$ and for the calibration parameter dimensions, $\boldsymbol{\phi}$; and the correlation parameter for the bias, $\boldsymbol{\psi}$. All input parameters were scaled to $[0, 1]$, and the true calibration parameters $\boldsymbol{\theta}$ for a given replicate of a simulation run was drawn from the uniform(0,1) distribution. The setting levels investigated for each problem type are summarized in Table 7.3, Table 7.4 and Table

Setting	Level 1	Level 2
M	15	n/a
N	7	n/a
ρ,ϕ	0.05	0.6
ψ	0.05	0.6
λ_f	10	n/a
λ_z	1	n/a
λ_b	1	n/a

Table 7.3: Setting levels for the 2-dimensional calibration with bias simulation study.

Setting	Level 1	Level 2
M	30	n/a
N	15	n/a
ρ,ϕ	0.05	0.6
ψ	0.05	0.6
λ_f	10	n/a
λ_z	1	n/a
λ_b	1	n/a

Table 7.4: Setting levels for the 4-dimensional calibration with bias simulation study.

7.5 below. The simulation experiment consisted of running the full factorial of these setting levels with 100 replicates each.

7.2.1 Results: Estimating $\boldsymbol{\theta}$

First, we evaluate how well the calibration parameter is estimated when the computer model is biased. The results of fitting the model with derivative information (i.) and the standard calibration model (ii.) for the 2-dimensional simulation study are shown in Table 7.6. The first column corresponds to the case when both the computer model and the bias are simple responses ($\rho = \phi = \psi = 0.6$), while the last column considers the case where both are complex responses ($\rho = \phi = \psi = 0.05$). The centre two columns correspond to the cases where either the response is simple and the bias is complex (column 3), or vice-versa (column 4).

Setting	Level 1	Level 2
M	45	n/a
N	23	n/a
ρ,ϕ	0.05	0.6
ψ	0.05	0.6
λ_f	1000	10
λ_z	1	n/a
λ_b	1	n/a

Table 7.5: Setting levels for the 6-dimensional calibration with bias simulation study.

	ρ, ϕ :	0.6		0.05	
Model	ψ :	0.6	0.05	0.6	0.05
i.	λ_f=1000	2.88e-02	1.50e-03	4.03e-02	1.84e-02
ii.	λ_f=1000	3.25e-02	2.66e-03	3.89e-02	1.58e-02
i.	λ_f=10	6.52e-02	9.59e-03	3.19e-02	2.02e-02
ii.	λ_f=10	5.93e-02	8.53e-03	3.78e-02	2.38e-02

Table 7.6: Reported $MedSE(\theta)$ from 2-d calibration simulation study.

From this table, the two methods appear to provide similar accuracy in the estimation of the calibration parameter for all level combinations of the correlation parameters and at both levels of measurement error. The only situation where there is some difference is the case when the computer model has a simple response, the bias is complex and the measurement error is small ($\lambda_f = 1000$), in which case the model with derivatives saw the $MedSE(\theta)$ reduced by a factor of about 1.8.

The results for the 4-dimensional study are shown in Table 7.7. In this case, when $\rho = \phi = \psi = 0.6$, we again see little difference between the two calibration models. However, when the computer model was a simple response and the bias was complex there was a notable difference. In particular, with small measurement error ($\lambda_f = 1000$), the model with derivatives reduced the error by about (25,10) times for (θ_1, θ_2) respectively as compared to the standard model. This decreased to a reduction of (4.5,3) times when the measurement

error was increased ($\lambda_f = 10$). There was also some improvement when both the computer model and the bias were complex, but not as pronounced.

In the case of the 6-dimensional simulation results shown in Table 7.8, there is again a noticeable improvement when the computer model is a simple response and the bias is complex. In this case, the median error was reduced by factors of approximately (4,5,9) for $(\theta_1, \theta_2, \theta_3)$ respectively when the measurement error was small, and by factors of about (3,4,3) when the measurement error was large. There is also some improvement seen when both the computer model and bias are complex, but the results are more mixed: when the measurement error was small, the model with derivatives had $MedSE(\theta)$'s that were about (4,3,1) times smaller, and with large measurement error the $MedSE(\theta)$'s were around (1,4,3) times smaller. In other words, in this case, two of the calibration parameters saw notable reductions in their error of estimation, but one of them did not.

7.2.2 Results: Mean Squared Prediction Error and Relative Efficiency for the Field

Our next measure of comparative performance is to evaluate how well the field process is predicted. For the 2-dimensional study, the calculated medians of the mean square prediction errors and relative efficiencies are shown in Table 7.9. Here again the two methods appear to provide similar performance for all level combinations of the correlation parameters and at both levels of measurement error. The only noticeable difference is when the computer model has a simple response, the bias is complex and the measurement error is small, in which case the model with derivatives was twice as efficient as compared to the standard model.

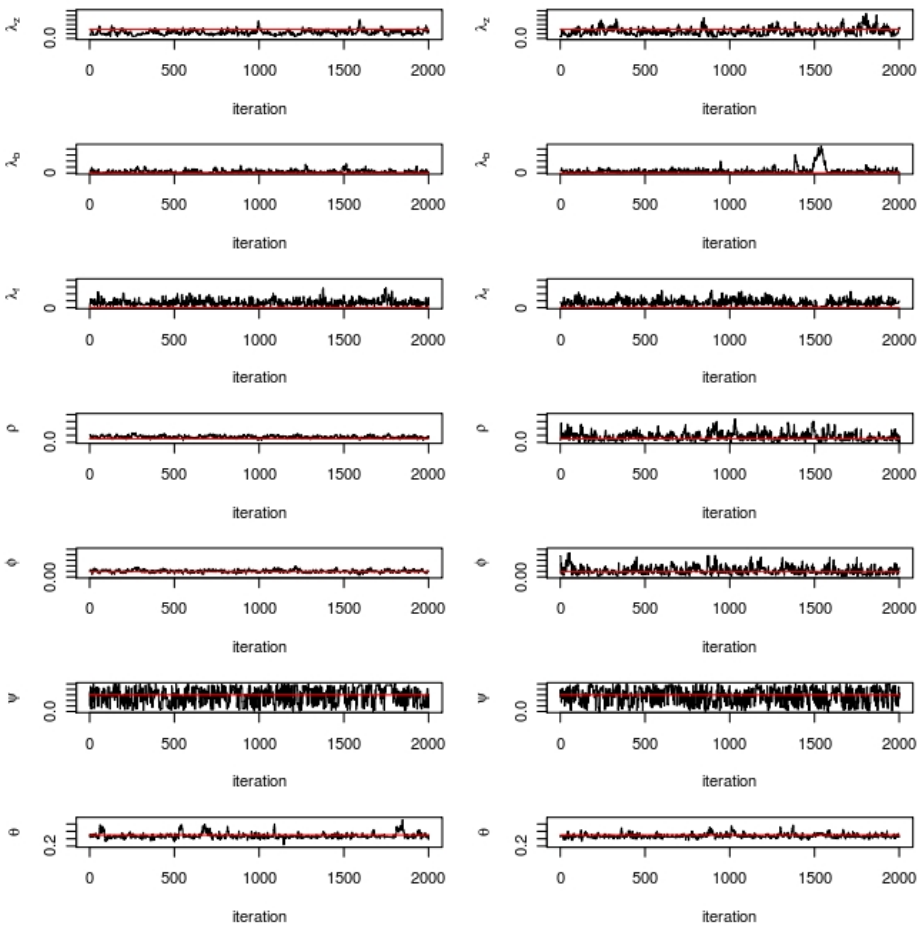


Figure 7.1: Traceplots for parameters of fitted calibration model with bias and derivatives (left column) and standard calibration with bias model (right column). True parameter values are shown as red lines.

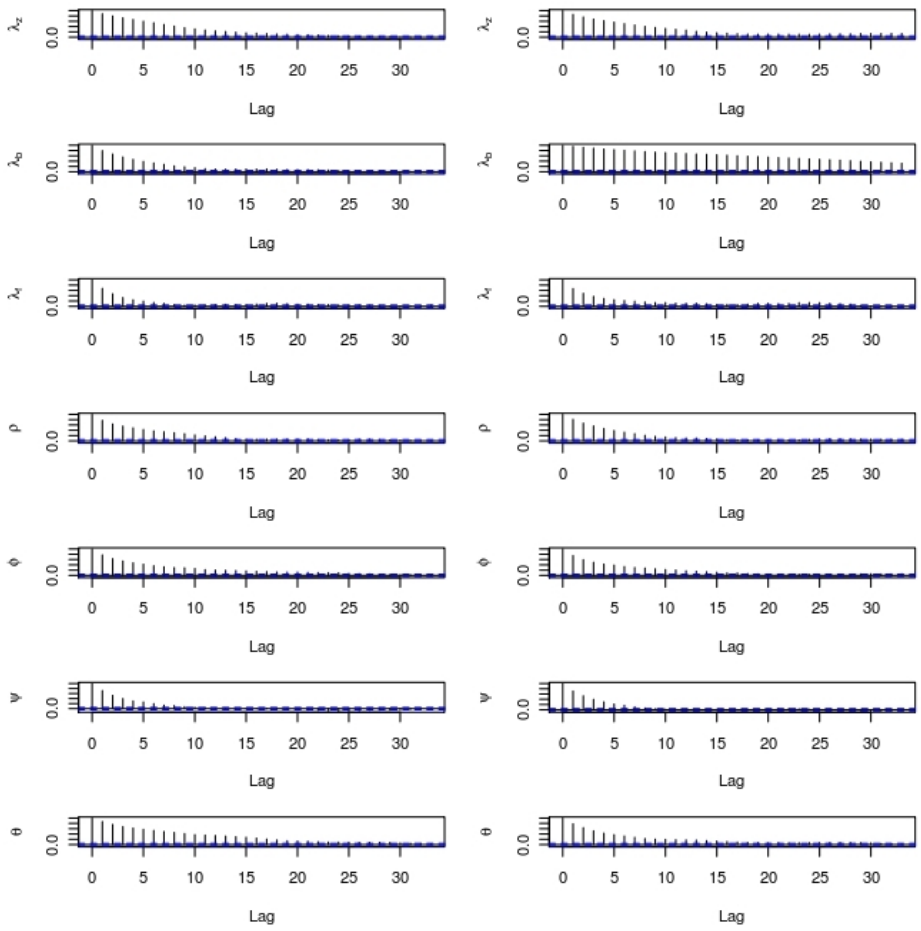


Figure 7.2: Autocorrelation plots for parameters of fitted calibration model with bias and derivatives (left column) and standard calibration with bias model (right column).

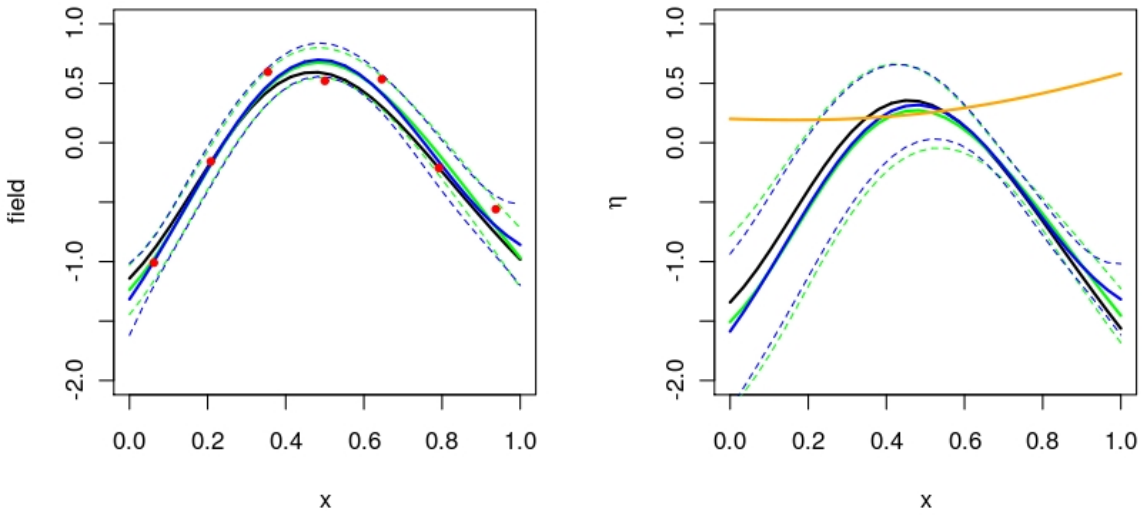


Figure 7.3: Calibration of a 1-dimensional surface with a 1-dimensional calibration parameter. In the left panel, the true underlying process is shown as the black curve, and the $N = 7$ field observations from this curve are shown as the red dots. The mean of the calibrated field predictions for the standard model is shown as the blue line with 95% prediction intervals shown as the dashed blue lines. Similarly, the mean of the prediction from the model with derivatives is shown as the green line, with 95% prediction intervals shown as the dashed green lines. In the right panel, the true underlying computer model is shown as the black curve. The mean of the calibrated computer model predictions for the standard model is shown as the blue line with 95% prediction intervals shown as the dashed blue lines. Similarly, the mean of the prediction from the model with derivatives is shown as the green line, with 95% prediction intervals shown as the dashed green lines. The bias is shown as the solid orange line.

	ρ, ϕ :	0.6				0.05			
	ψ :	0.6		0.05		0.6		0.05	
Model		θ_1	θ_2	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
i.	λ_f=1000	4.03e-02	2.46e-02	8.67e-04	1.61e-03	3.39e-02	5.38e-02	9.77e-03	5.86e-03
ii.	λ_f=1000	4.82e-02	3.36e-02	2.19e-02	1.56e-02	3.82e-02	4.92e-02	1.56e-02	2.60e-02
i.	λ_f=10	3.53e-02	4.74e-02	6.28e-03	8.77e-03	4.07e-02	4.90e-02	1.26e-02	1.52e-02
ii.	λ_f=10	3.27e-02	3.89e-02	2.87e-02	2.72e-02	3.64e-02	4.89e-02	2.80e-02	2.68e-02

Table 7.7: Reported $MedSE(\theta)$ from 4-d calibration simulation study.

	ρ, ϕ :	0.6						0.05					
	ψ :	0.6			0.05			0.6			0.05		
Model		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
i.	λ_f=1000	1.77e-02	1.65e-02	1.61e-02	4.10e-03	6.81e-03	4.82e-03	5.37e-02	3.74e-02	3.79e-02	9.54e-03	1.19e-02	1.46e-02
ii.	λ_f=1000	4.04e-02	2.84e-02	2.35e-02	1.84e-02	3.28e-02	4.56e-02	5.84e-02	4.01e-02	4.09e-02	3.85e-02	3.54e-02	1.58e-02
i.	λ_f=10	2.34e-02	2.90e-02	3.46e-02	1.01e-02	7.02e-03	9.31e-03	4.34e-02	5.20e-02	4.03e-02	1.79e-02	1.29e-02	1.26e-02
ii.	λ_f=10	3.33e-02	3.93e-02	3.18e-02	3.25e-02	3.12e-02	3.10e-02	4.08e-02	5.70e-02	3.68e-02	2.30e-02	5.77e-02	3.80e-02

Table 7.8: Reported $MedSE(\theta)$ from 6-d calibration simulation study.

	ρ, ϕ :	0.6		0.05	
Model	ψ :	0.6	0.05	0.6	0.05
i.	λ_f=1000	5.79e-05	1.51e-04	3.45e-04	4.32e-04
		(0.88)	(2.00)	(0.94)	(1.04)
ii.	λ_f=1000	5.12e-05	3.02e-04	3.25e-04	4.52e-04
i.	λ_f=10	4.77e-03	1.14e-02	2.51e-02	4.41e-02
		(0.96)	(1.12)	(0.96)	(1.06)
ii.	λ_f=10	4.56e-03	1.28e-02	2.41e-02	4.68e-02

Table 7.9: Reported $MedMSPE$ from the 2-d calibration simulation study. The efficiency for predicting the field relative to the standard model, eff, is shown in brackets.

The results for the 4-dimensional study are shown in Table 7.10. Again, with a simple computer model response and complex bias in the presence of small measurement error the model with derivatives is nearly twice as efficient as the standard model. However, this efficiency is lost when the measurement error is increased. In all other cases, the two models seem to provide similar performance in predicting the field process.

Finally, the results for the 6-dimensional study are shown in Table 7.11. Little improvement is seen using the model with derivatives at all settings investigated. The largest improvement again occurs when the computer model is a simple response, the bias is complex and the measurement error is small, but here the model with derivatives only achieves an eff of 1.42 as compared to the larger improvements seen in the 2-dimensional and 4-dimensional studies.

7.2.3 Results: Mean Squared Prediction Error and Relative Efficiency for the Calibrated Computer Model

The final measure of performance we are interested in is evaluating the prediction of the calibrated computer model. The calculated values of $MedMSPE\eta$ are shown in Table 7.12 for the 2-dimensional simulation study. Yet again, the only improvement seen occurred for the case of a simple computer model response, complex bias and small measurement error.

	ρ, ϕ :	0.6		0.05	
Model	ψ :	0.6	0.05	0.6	0.05
i.	λ_f=1000	4.99e-04	3.30e-03	6.07e-03	9.52e-03
		(1.07)	(1.96)	(1.03)	(1.27)
ii.	λ_f=1000	5.35e-04	6.47e-03	6.24e-03	1.21e-02
i.	λ_f=10	2.44e-02	5.32e-02	8.69e-02	1.05e-01
		(0.97)	(1.29)	(1.01)	(0.97)
ii.	λ_f=10	2.36e-02	6.86e-02	8.74e-02	1.02e-01

Table 7.10: Reported $MedMSPE$ from 4-d calibration simulation study. The efficiency for predicting the field relative to the standard model, eff, is shown in brackets.

	ρ, ϕ :	0.6		0.05	
Model	ψ :	0.6	0.05	0.6	0.05
i.	λ_f=1000	1.41e-03	5.62e-02	7.02e-02	1.13e-01
		(1.39)	(1.42)	(1.03)	(1.20)
ii.	λ_f=1000	1.96e-03	7.98e-02	7.22e-02	1.35e-01
i.	λ_f=10	3.86e-02	1.29e-01	1.55e-01	2.35e-01
		(1.15)	(1.29)	(1.00)	(1.11)
ii.	λ_f=10	4.45e-02	1.66e-01	1.55e-01	2.61e-01

Table 7.11: Reported $MedMSPE$ from 6-d calibration simulation study. The efficiency for predicting the field relative to the standard model, eff, is shown in brackets.

	ρ, ϕ :	0.6		0.05	
Model	ψ :	0.6	0.05	0.6	0.05
i.	λ_f=1000	1.64e-02	6.53e-03	1.54e-02	4.84e-02
		(0.93)	(1.84)	(0.94)	(0.82)
ii.	λ_f=1000	1.52e-02	1.20e-02	1.45e-02	3.95e-02
i.	λ_f=10	3.13e-02	4.86e-02	1.52e-02	8.42e-02
		(0.99)	(1.17)	(1.09)	(1.06)
ii.	λ_f=10	3.10e-02	5.70e-02	1.65e-02	8.93e-02

Table 7.12: Reported $MedMSPE\eta$ from 2-d calibration simulation study. The efficiency for predicting the calibrated computer model relative to the standard model, $eff\eta$, is shown in brackets.

All other settings show the model with derivative information provided performance similar to the standard model.

The results for the 4-dimensional study are shown in Table 7.13. Note the noticeable improvement when the computer model is simple and the bias is complex, or when both are complex. Here, the relative efficiency of the model with derivatives over the standard model is particularly large with small measurement error. In contrast, when the computer model is simple or complex and the bias is complex, we do not see any notable improvement when using the model with derivative information.

The 6-dimensional simulation study results are shown in Table 7.14. In this case the model with derivative information provides a relative efficiency of about 2 over the standard model in all cases except when the computer model is complex and the bias is simple. Here, the case of a simple computer model response and a simple bias shows improvements in the 6-dimensional case whereas little to no improvement was seen in the 2-dimensional and 4-dimensional studies.

	ρ, ϕ :	0.6		0.05	
Model	ψ :	0.6	0.05	0.6	0.05
i.	λ_f=1000	3.80e-02	2.02e-02	3.36e-02	9.54e-02
		(1.16)	(14.60)	(1.10)	(3.78)
ii.	λ_f=1000	4.42e-02	2.95e-01	3.71e-02	3.61e-01
i.	λ_f=10	6.03e-02	9.50e-02	6.53e-02	1.43e-01
		(1.02)	(3.71)	(1.01)	(3.23)
ii.	λ_f=10	6.15e-02	3.53e-01	6.61e-02	4.61e-01

Table 7.13: Reported $MedMSPE\eta$ from 4-d calibration simulation study. The efficiency for predicting the calibrated computer model relative to the standard model, $eff\eta$, is shown in brackets.

	ρ, ϕ :	0.6		0.05	
Model	ψ :	0.6	0.05	0.6	0.05
i.	λ_f=1000	4.38e-02	2.14e-01	7.59e-02	3.49e-01
		(2.84)	(1.99)	(1.12)	(1.95)
ii.	λ_f=1000	1.24e-01	4.27e-01	8.51e-02	6.79e-01
i.	λ_f=10	9.83e-02	3.20e-01	9.06e-02	3.87e-01
		(1.58)	(1.75)	(1.19)	(1.89)
ii.	λ_f=10	1.55e-01	5.59e-01	1.07e-01	7.32e-01

Table 7.14: Reported $MedMSPE\eta$ from 6-d calibration simulation study. The efficiency for predicting the calibrated computer model relative to the standard model, $eff\eta$, is shown in brackets.

7.3 Discussion

The results of our investigation comparing the calibration model with derivative information to the standard model when the computer model is biased is not as easy to interpret as when the computer model was not biased. For instance, the model with derivative information is not uniformly better than the standard model, although it is never worse.

In general, the results suggest that the benefits possible when including derivative information are largely dependent on the complexity of the computer model and the complexity of the bias. We consistently saw improvements in the estimation of the calibration parameters and in predicting the calibrated computer model when the computer model was a simple response and the bias was complex, and to a lesser extent when both were complex responses. However, the improvements that can be realized with small measurement error quickly decrease as the measurement error is increased.

Prediction of the calibrated computer model showed some improvements, particularly in the 4-dimensional simulations when the bias was complex. This in turn allows for a better understanding of how the computer model is biased, which is useful information for the experimenter.

In contrast, prediction of the field process was typically not improved, or only slightly improved, when using the model with derivative information. This is not surprising, since if the calibration parameter and resulting calibrated computer model are poorly estimated in the standard model, then the bias will adjust accordingly so the model fits the observed field reasonably well. This represents the known identifiability issue with the standard model. The results we have presented provide empirical evidence that incorporating derivative information may help reduce this issue, but only in certain cases. A more theoretical investigation into both the (un)identifiability of the standard model and the possible identifiability of the model with derivative information in some cases would be an interesting topic for future research.

Chapter 8

Conclusions

In this thesis, three new statistical methods for computer model calibration were developed. These methods were developed with the aim of addressing two issues that commonly arise in computer model calibration. The first problem is that the data structure being calibrated can be very large, leading to covariance matrices that are simply too big to manipulate, store and invert on computer. The covariance of such data structures may also exhibit non-stationarity as the calibration parameter is varied. This violates the stationarity assumption of the usual approach to calibration.

We investigated this problem in Chapter 3, which led to the development of a criterion-based calibration method suitable for large data structures that exhibit the described non-stationarity. In addition, our criterion approach enables a simple sequential design strategy for refining our estimate of the calibration parameter. Furthermore, theoretical evidence for the consistency of the calibration parameter estimate using our method was developed in Chapter 4. The proposed method was found to work well in all the examples investigated, and was computationally amenable.

The second problem that can occur is if the ability to sample both the field process and computer model are highly limited. In such scenarios, the amount of useful information that can be borrowed from the computer model runs for predicting the field process may

be very small. This can result in poor prediction of the field process and poor estimates of the calibration parameter. In Chapters 5-7, we consider including derivative information in computer model calibration as one way to improve the predictive and parameter estimation capabilities when the sample size is small. Since computer models often solve systems of differential equations to model the physical process of interest, such derivative information may be available.

Chapter 5 outlines a Bayesian model that allows for the inclusion of some or all of the first partial derivatives of the computer model into the calibration experiment under the assumption that the computer model is unbiased. We explored this model via simulation study in Chapter 6. We considered the predictive performance, estimation of the calibration parameter, prediction interval widths and coverage probabilities of three calibration models: (i) calibration with full derivative information; (ii) calibration with partial derivative information; and (iii) calibration with no derivative information. Our findings indicate that calibrating with full or partial derivative information was uniformly better than calibrating without this information. In particular, the improvements when using full derivative information were often quite large. The model with partial derivative information realized less striking results, offering comparable or modest improvement in performance relative to the standard model with no derivative information. In general, the results indicated that for more complex responses, including derivative information would yield greater benefits, while at higher levels of measurement error, the benefits were diminished.

In Chapter 7 we extended our model of Chapter 5 to the case of a biased computer model. We compared the behaviour of: (i) our proposed model including full derivative information and accounting for computer model bias; and (ii), the standard model allowing bias but no derivative information. The performance of calibration parameter estimates, predictive performance for the field, and predictive performance for the calibrated computer model were investigated via simulation study. In this case, the results were more mixed. In some situations, such as when the computer model was a simple response and the bias

was a complex response, modest improvements could be realized by including full derivative information. These improvements became less pronounced as the measurement error was increased. However, for other experimental conditions, the model with derivative information did not provide substantial gains over the usual approach.

The developments in this thesis suggest many avenues to explore in future research. For the fast calibration method outlined in Chapters 2 and 3, a natural extension is to consider a similar approach in the case of a biased computer model. Considering design in the method of Chapter 5 is also interesting. For instance, the cost of obtaining derivatives might have a noticeable affect on optimal sampling. Another consideration is allowing for the sampling of the derivative process at different locations than the process itself. Finally, resolving the challenge of calibrating a biased computer model remains an interesting problem. Investigating possible ways to improve on the results in Chapter 7 is a topic worth pursuing in future research.

Appendix A

Simulation-Based Inference

Suppose we have collected data $\mathbf{\Delta} = (\Delta(\mathbf{t}_1), \ldots, \Delta(\mathbf{t}_n))^T$ at locations $\mathbf{t}_i, i = 1, \ldots, n$, where the data is considered to have been observed from a realization of a random function $\Delta(\mathbf{t})$. Denote the covariance of the data by $\sigma^2_\Delta \mathbf{R}_{\mathbf{\Delta}} = Cov(\mathbf{\Delta}, \mathbf{\Delta}^T)$ and the mean by $\mu_\Delta = E(\mathbf{\Delta})$. Following Schabenberger and Gotway (2005), one can simulate a realization of the process $\Delta(\mathbf{t})$ at a grid of m locations using these steps:

1. Select a grid of m locations at which to observe realizations of the random function. Call these locations $\mathbf{t}_{n+1}, \ldots, \mathbf{t}_{n+m}$.

2. Generate a realization, $\mathbf{U}$, of a process having mean μ_Δ and covariance $\sigma^2_\Delta \mathbf{R}_\Delta$ at all $n+m$ sites. Let $\mathbf{U}_0 = [U(\mathbf{t}_1), \ldots, U(\mathbf{t}_n)]$ be the values of this realization at the original n data locations.

3. Calculate $\widehat{\Delta}(\mathbf{t}_i) = U(\mathbf{t}_i) + \mathbf{r}_i \mathbf{R}_\Delta^{-1}(\mathbf{\Delta} - \mathbf{U}_0)$, where $\mathbf{r}_i = [corr(U(\mathbf{t}_i), U(\mathbf{t}_1)), \ldots, corr(U(\mathbf{t}_i), U(\mathbf{t}_n))]$ is the vector of correlations between data observed at the new location $\mathbf{t}_i$ and the n original locations.

4. Finally, set $\widehat{\mathbf{\Delta}} = [\widehat{\Delta}(\mathbf{t}_1), \ldots, \widehat{\Delta}(\mathbf{t}_{n+m})]$.

Then $\widehat{\mathbf{\Delta}}$ are observations from a realization of the unknown random function constructed at the n original observation locations and the m additional grid locations. Schabenberger and Gotway (2005) show that this realization will interpolate the observed data $\mathbf{\Delta}$. The

realization will also better reflect the variability of the unknown random function at the m grid locations than simply drawing from the conditional distribution of the random function given the n observations.

Using the above algorithm, an empirical distribution for θ is constructed as follows:

1. Construct J realizations of the unknown random function at the $n + m$ locations by repeatedly applying the steps above.

2. For each realization, calculate $\hat{\theta}_{(n)} = \arg\min_{\mathbf{t}} \widehat{\Delta}(\mathbf{t})$ for $\mathbf{t} \in \{\mathbf{t}_1, \ldots, \mathbf{t}_{n+m}\}$.

This generates J estimates of the calibration parameter that incorporates uncertainty about $\Delta(\mathbf{t})$ from only observing the criterion at n locations. For instance, from these J estimates, one can construct the histograms shown in Figure 3.4.

Appendix B

Properties of Multivariate Normals

Suppose $\mathbf{Y}|\cdot \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{\Sigma}^{11} & \mathbf{\Sigma}^{12} \\ \mathbf{\Sigma}^{21} & \mathbf{\Sigma}^{22} \end{bmatrix}\right)$ where $\mathbf{Y} = \left[\mathbf{Y}^{1T}, \mathbf{Y}^{2T}\right]^T$. Then, $\mathbf{Y}^1|\mathbf{Y}^2, \cdot$ is multivariate normal with

$$E\left[\mathbf{Y}^1|\mathbf{Y}^2, \cdot\right] = \mathbf{\Sigma}^{12}\mathbf{\Sigma}^{22^{-1}}\mathbf{Y}^2,$$

and

$$V\left[\mathbf{Y}^1|\mathbf{Y}^2, \cdot\right] = \mathbf{\Sigma}^{11} - \mathbf{\Sigma}^{12}\mathbf{\Sigma}^{22^{-1}}\mathbf{\Sigma}^{21}.$$

Bibliography

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall / CRC Press, Bacon Raton, FL.

Chen, H., Simpson, D. G., and Ying, Z. (2000). Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica*, 10:141–156.

Cressie, N. A. C. (1993). *Statistics for Spatial Data, rev. edn.* John Wiley & Sons Inc., New York.

Fedder, J. A., Slinker, S. P., Lyon, J. G., and Elphinstone, R. D. (1995). Global numerical simulation of the growth phase and the expansion onset for substorm observed by Viking. *Journal of Geophysical Research*, 100:19083–19094.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis, 2nd edition*. Chapman & Hall, Boca Raton.

Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In bayesian statistics 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.)*. Clarendon Press, Oxford.

Griewank, A. and Walther, A. (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition*. SIAM, Philadelphia.

Higdon, D., Kennedy, M. C., Cavendish, J., Cafeo, J., and Ryne, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *Siam Journal on Scientific Computing*, 26:448–466.

Ibragimov, I. A. and Rozanov, Y. A. (1978). *Gaussian Random Processes*. Springer-Verlag, New York.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, 63:425–464.

Krige, D. G. (1966). Two-dimensional weighted moving average trend surfaces for ore evaluation. In *Proceedings of the symposium on mathematical statistics and computer applications in ore valuation*, pages 13–38, Johannesburg.

Kullback, S. (1997). *Information Theory and Statistics*. Courier Dover Publications.

Lyon, J. G., Fedder, J. A., and Mobarry, C. M. (2004). The Lyon-Fedder-Mobarry (LFM) global MHD magnetospheric simulation code. *Journal of Atmospheric and Solar-Terrestrial Physics*, 66:1333–1350.

Mitchell, T., Morris, M., and Ylvisaker, D. (1994). Asymptotically optimum experimental designs for prediction of deterministic functions given derivative information. *Journal of Statistical Planning and Inference*, 41:377–389.

Morris, M. D., Mitchell, T., and Ylvisaker, D. (1993). Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, 35:243–255.

Nychka, D. (2000). Spatial process estimates as smoothers. In *Smoothing and Regression: approaches, computation and application*, pages 393–424. Wiley, New York.

Parzen, E. (1962). *Stochastic Processes*. Holden-Day, San Francisco.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2009). *coda: Output analysis and diagnostics for MCMC*. R package version 0.13-4.

Raftery, A. E. and Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statistical Science*, 7:493–497.

Raftery, A. E. and Lewis, S. M. (1995). *The number of iterations, convergence diagnostics and generic metropolis algorithms. In practical markov chain monte carlo (W. R. Gilks, D. J. Spiegelhalter and S. Richardson, eds.)*. Chapman and Hall, London.

Rall, L. B. (1981). *Automatic Differentiation: Techniques and Applications*, volume 120 of *Lecture Notes in Computer Science*. Springer, Berlin.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society, Series B*, 69:741–796.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, 2nd. Edition*. Springer, New York.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

Ripley, B. D. (1981). *Spatial Statistics*. John Wiley & Sons, New York.

Sacks, J., Schiller, S. B., and Welch, W. J. (1989a). Designs for computer experiments. *Technometrics*, 31:41–47.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989b). Design and analysis of computer experiments. *Statistical Science*, 4:409–423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The design and analysis of computer experiments*. Springer-Verlag, New York.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, FL.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

Team, R. D. C. (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Williams, B. J., Santner, T. J., and Notz, W. I. (2000). Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, 10:1133–1152.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a gaussian process. *Journal of Multivariate Analysis*, 36:280–296.

Ying, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Annals of Statistics*, 21:1567–1590.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99:250–261.