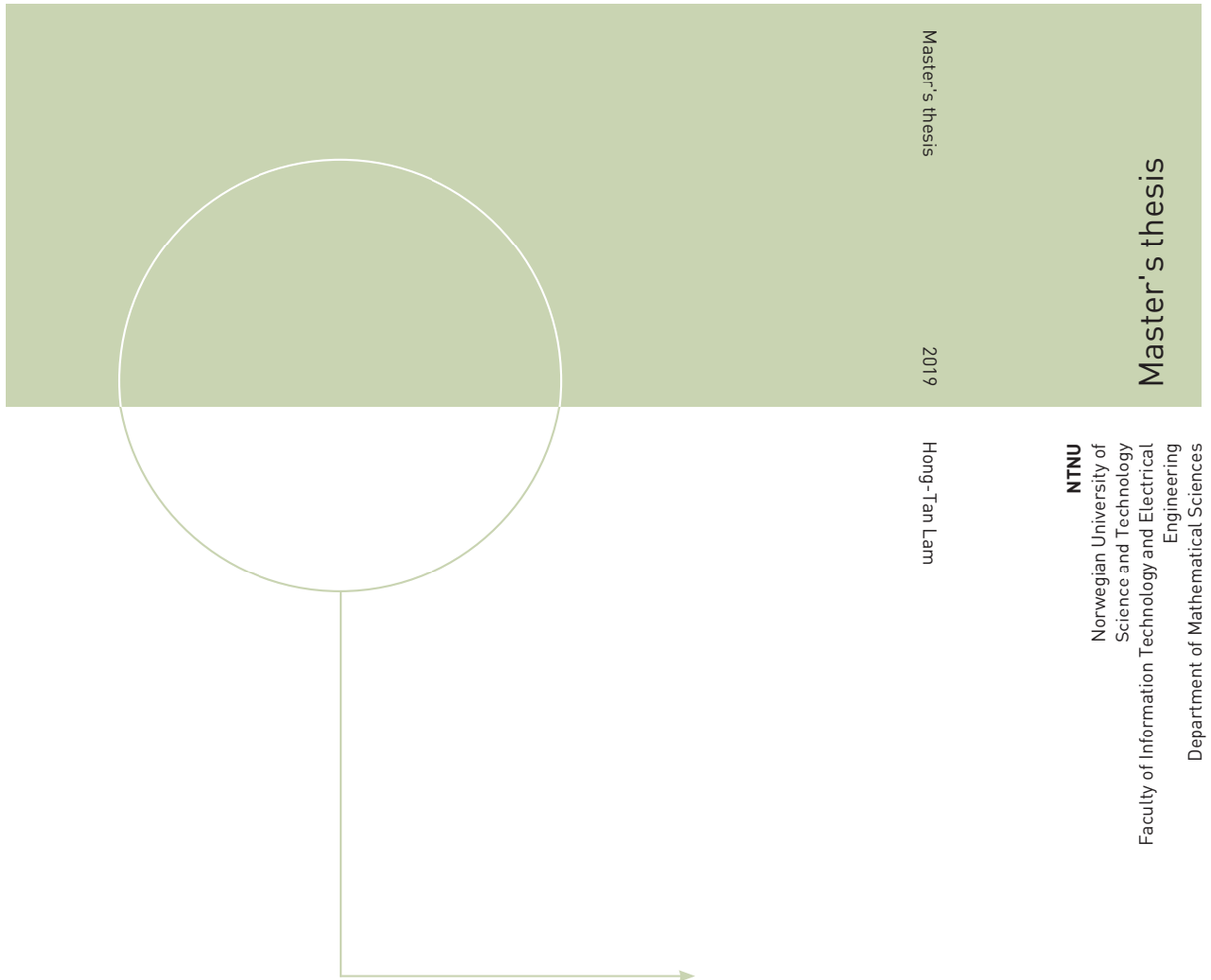


Master's thesis

2019

Hong-Tan Lam

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

Hong-Tan Lam

Bayesian Calibration and Inference for Multiple Machines

July 2019

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Bayesian Calibration and Inference for Multiple Machines

Hong-Tan Lam

Applied Physics and Mathematics
Submission date: July 2019
Supervisor: Ingelin Steinsland

Norwegian University of Science and Technology
Department of Mathematical Sciences

”If you ever think that you can only be one thing in life, think again! A prediction point can be both an interpolation point and an extrapolation point at the same time, depending on the machine you consider. Perspective is truly everything.”

Abstract

In this thesis, simulation models, also called simulators, are used to perform predictions on machines. One of the challenges with performing predictions using simulation models, is that they do not fully describe the true process of the machine. Often, there exists a model discrepancy: the difference between the simulator output and the true process. The working hypothesis is that, by using multiple machines with some parameters and hyperparameters in common, it is possible that the model learns from all the machines, and perform better predictions.

There are three different models used to perform inference. The first model, is a model from Brynjarsdóttir and O'Hagan (2014), slightly modified, with individual parameters and an individual discrepancy term. It is referred to as the individual model. Two new model frameworks are introduced in this thesis, referred to as Model 1 and Model 2. Model 1 assumes common parameters for the machines and has an individual discrepancy term. Model 2 assumes common parameters for the machines, and has a common discrepancy term and an individual discrepancy term. All three models are Latent Gaussian models, with their discrepancy terms following Gaussian fields with Matérn covariance functions. The last two models evaluates the multiple machines simultaneously, and are used to create observations, resulting in two new types of machines, referred to as the Ideal machines 1 and the Ideal machines 2 respectively. These ideal machines are created to learn more about the predictive perfomances of Model 1 and Model 2. Additionally, the three models are applied on multiple simple machines, the same machine described by Brynjarsdóttir and O'Hagan (2014).

There are four types of prediction points that are tested: interpolation points, extrapolation points, pseudo extrapolation points (prediction points that are extrapolation points for some machines and interpolation points for others) and prediction points that are located far from the observations. Different designs of where the observations created are located, are used to test the models predictive performance on the four types of prediction points. The prediction is performed using the R-package INLA, which is less computationally expensive than using Markov chain Monte Carlo algorithms.

For all four types of prediction points, evaluating the machines simultaneously gives better results than evaluating the machines individually. Model 2 has the best predictive performance for all three machines and all four types of prediction points. Model 1 has equally good predictive performance for the Ideal machines 1 as Model 2, and in general better predictive performance than the individual model. There is not one type of prediction point for any of the multiple machines, where the individual model has the best performance. The results are found using only synthetical machines and only when the calibration parameter has the same value for all machines. Suggestions for further research, are to use real machines or to vary the value of the calibration parameter for all machines.

Sammendrag

I denne oppgaven, blir simuleringsmodeller, også kalt simulatorer, brukt til utføre prediksjon på maskiner. En av utfordringene med å gjøre prediksjon med simuleringsmodeller, er at de ikke fullt beskriver den sanne prosessen til maskinen. Ofte eksisterer det et modellavvik: forskjellen mellom simulatoren og den sanne prosessen. Hypotesen, er at ved å bruke flere maskiner med noen parametere og hyperparametere til felles, er det mulig at modellen lærer fra alle maskinene, og utfører bedre prediksjon.

Det er tre forskjellige modeller brukt til inferens. Den første modellen, er en modell fra Brynjarsdóttir and O'Hagan (2014), litt modifisert, med individuelle parametere og et individuelt avviksledd. Den er referert til som den individuelle modellen. To nye modellrammeverk er introdusert i denne oppgaven, referert til som Modell 1 og Modell 2. Modell 1 antar felles parametere for maskinene, og har et individuelt avviksledd. Modell 2 antar felles parametere for maskinene, og har et felles avviksledd og et individuelt avviksledd. Alle tre modellene er latente Gaussiske modeller, med avviksledd som følger Gaussiske felt med Matérn kovariansfunksjoner. De siste to modellene evaluerer maskinene samtidig, og er brukt til å lage observasjoner, noe som resulterer i to nye typer maskiner, som henholdsvis er referert til som de Idelle maskiner 1 og de Idelle maskiner 2. Disse idelle maskinene er lagd for å lære mer om prestasjonene på prediksjonene til Modell 1 og Modell 2. I tillegg, er de tre modellene brukt på de multiple enkle maskinene, den samme maskinen beskrevet av Brynjarsdóttir and O'Hagan (2014).

Det er fire typer prediskjonspunkter som er testet: interpolasjonspunkter, ekstrapolasjonspunkter, pseudo-ekstrapolasjonspunkter (prediksjonspunkter som er ekstrapolasjonspunkter for noen maskiner og interpolasjonspunkter for andre) og prediksjonspunkter som ligger langt fra observasjonene. Forskjellige designer for hvor observasjonspunktene skal ligge, er brukt til å teste modellenes prestasjoner på prediksjonene på alle fire typer prediksjonspunkter. Prediksjonene er utført ved å bruke R-pakken INLA, som er mindre beregningskrevende enn å bruke Markov-kjede-Monte-Carlo-algoritmer.

For alle fire typer prediksjonspunkter, gir evalueringene av maskinene samtidig bedre resultater enn å evaluere maskinene individuelt. Modell 2 har best prestasjon på prediksjonene for alle tre maskiner og alle fire typer prediksjonspunkter. Modell 1 har like god prestasjon på prediksjonene for de Idelle maskiner 1 og de Idelle maskiner 2, og generelt bedre prestasjon på prediksjonene enn den individuelle modellen. Det finnes ikke en eneste type prediksjonspunkt for noen av de multiple maskinene hvor den individuelle modellen har best prestasjon. Resultatene er funnet bare ved å bruke syntetiske maskiner og bare når kalibreringsparameteren har den samme verdien for alle maskinene. Forslag til videre forskning, er å bruke ekte maskiner eller å variere verdien til kalibreringsparameteren for alle maskinene.

Preface

I remember the first three months working on my Master's thesis. It was mostly getting to know how INLA works, which was difficult, as it is a package with little documentation, and getting errors is very easy. I am not going to lie and say that it was all a walk in the park. At one point, I was stuck on the same bug for a month. In those couple of weeks, I was just forcing myself to sit at campus, watching the time pass by, as there was no progression in my work. I was really hating my thesis at this point, and my goal was just to deliver something acceptable, as I had already managed to get a job, and I did not bother too much with the grade.

However, there was this one day when everything changed. It was at the end of April, two months before the deadline. I decided to take a closer look at my code, trying to read the documentation for everything I had written, hoping that the documentation existed. And there, after a month, I finally found my bug. From there on, it kind of was like a walk in the park. That is just how good I am. No, but seriously, that is the first big lesson I have learned from my Master's thesis: that patience matters a lot. To stop and take a closer look, whether it is in a piece of code or if it is in life in general, is one of the things that is easy to neglect in our daily lives. And sometimes, a closer look, is all you need to resolve some of the most difficult problems you may face.

The second thing I have learned from my Master's thesis, is more on how research is done. For the last two years I have been preparing myself for a data analytics job, but I have been very curious about how it is to do research. That my Master's thesis gave me insight on how research is done, is something I have been very grateful for. It was this appreciation combined with the patience I gained, that finally made me enjoy working on my thesis.

There are several people I want to thank. Thank you to Ingelin Steinsland, who have been my supervisor. She has told me which experiments are interesting to perform, a little about how different designs can help produce informative results, as well as helping me improve my writing and the structure of this thesis. I also want to thank my friends and family. They are truly everything to me.

Hong-Tan Lam,
Trondheim, July 2019.

Table of Contents

Chapter 1

Introduction

Many physical sciences use simulation models to describe physical processes. According to Winsberg (2003), a simulation model is a model of the underlying physics, developed on a computer, and by using computationally intensive methods, it is possible to learn about the physical process. The applications of simulation models keeps growing. One of the reason for this, is that a simulation model can be a virtual model of what an Internet of Things (IoT) device is tracking, and the number of IoT devices installed worldwide is increasing (Statista, 2016). Such devices, can for example, track the performance of certain processes on a boat, the lifetime of a bridge or the blood pressure in a human. However, although simulation models are widely used, there can occur many uncertainties with simulation models. To illustrate what uncertainties might occur, the example of a simple machine is used.

The simple machine is a theoretical machine created by Brynjarsdóttir and O'Hagan (2014). It is a machine producing work as a function of effort and efficiency. Although the simple machine is a theoretical machine, and thus, the relation between work, effort and efficiency is known, the machine is pretended to be a real machine, where this relation is unknown. This relation is only used to create the observations, and is referred to as the true process of the machine. After the observations are created using the true process, parts of the true process is pretended to be unknown. Work is considered as the output, the quantity to be predicted, effort is considered as the control variable, a quantity controlling the output, and efficiency is considered as the calibration parameter, a parameter with unknown value. In this case, the simulation model is a computer model with incomplete information about the true process, and is referred to as simulator for the rest of the thesis. The simulator tries to learn about the true process by using (synthetical) observations, a process called calibration.

According to Kennedy and O'Hagan (2001), there are six types of uncertainties that can occur. For this thesis, only the three most relevant uncertainties are described:

- Model discrepancy: difference between the true process' mean value of output, and the simulator output with true input values. In the case of the simple machine, using the true value for efficiency, this is the difference between the prediction of work from the simulator and what the work really is (true process).
- Observation error: difference between the observations measured and the true output value. In the case of the simple machine, this is the difference between the work observed and the true work (true process). Since the true process is used to generate the observations for the simple machine, normally distributed errors are added to create synthetical observation errors.
- Parametric variability: uncertainty in the simulator output due to lack of information about some parameters. For the simple machine, this uncertainty occurs due to letting the efficiency follow a distribution with little to no information (e.g. by letting it follow a non-informative prior).

The main goal of the simulator, is to predict work, but this is difficult if the simulator cannot account for the abovementioned uncertainties. Kennedy and O'Hagan (2001) suggests a model they claim takes account of all the abovementioned uncertainties, as well as the three types of uncertainties not described in this thesis. The model suggests that work is the sum of three terms: the work predicted by the simulator (also called simulator output), the model discrepancy and the observation error. It uses a Bayesian framework, by setting priors for the parameteres and hyperparameters. Then, observations are used for calibration, i.e. to learn more about the unknown parameters and hyperparameters.

Brynjarsdóttir and O'Hagan (2014) uses the model of Kennedy and O'Hagan (2001) to perform prediction for the simple machine. The model they use that is of interest in this thesis, succeeds at interpolation. For extrapolation however, the posterior distribution does not center around the true value.

In this thesis, the main focus is to introduce Bayesian calibration model frameworks that considers multiple machines simultaneously, apply them to study predictions of multiple machines, and to study the differences between the models' predictive performances. The multiple machines studied in this thesis, all have the same value for the calibration parameter, while the true processes are different. Although the true processes for the machines are different, the multiple machines' parameters are still the same. If a parameter is different for the multiple machines, then the distribution generating this parameter, have the same hyperparameters.

Three models are used in this thesis. The first model is similar to the model Brynjarsdóttir and O'Hagan (2014) use, referred to as the individual model. The second model is a similar model, as it assumes that the machines have individual discrepancy. However, it also assumes that the machines have the same values for the same parameters and hyperparameters. It is referred to as Model 1. The third model, is similar to Model 1, but in addition to assuming that the machines have the same values for the parameters and hyperparameters and have an individual discrepancy term, it assumes the machines have a common discrepancy term that is identical for all the machines. It is referred to as Model 2. All of these models are latent Gaussian models, with the model discrepancy term(s) following

zero-mean Gaussian fields with Matérn covariance functions. For such models, it is possible to use the R-package INLA by Rue et al. (2009b). This package is used for this thesis, as it has a shorter computation time for these models, compared to Markov chain Monte Carlo algorithms.

Model 1 and Model 2 are two new model frameworks introduced in this thesis. They evaluate the machines simultaneously, unlike the individual model, which is fitted to the observations for each machine separately. To learn more about Model 1 and Model 2 and how well they perform predictions, the models are tested on observations generated by the models themselves. The theoretical machines generating these observations, are referred to as the Ideal machine 1 and the Ideal machine 2. Thus, there are three models applied on three types of machines in this thesis: the Ideal machines 1, the Ideal machines 2 and the multiple simple machines. The prediction points are divided into four different categories: interpolation points, extrapolation points, pseudo extrapolation points and points that are located far from the observations (at least far from the observations on one side). Pseudo extrapolation points is a term used by the author as prediction points that are extrapolation points for some machines, and interpolation points for other machines, all of which are evaluated simultaneously by Model 1 and Model 2 for prediction.

To evaluate the predictive performance of the models on these four types of prediction points, and in which case predictive performance is better for a specific type of prediction point, different designs deciding where the observations are located are needed. Hence, three designs are created. They are referred to as Design 1, Design 2 and Mixed design.

The aim of this thesis, is to introduce Bayesian calibration model frameworks that consider multiple machines simultaneously to perform predictions. These models can be seen as an extension of the framework introduced by Kennedy and O'Hagan (2001), and it is studied how and when it is beneficial to consider the multiple machines simultaneously with regards to predictive performance. For which of the four types of prediction points, and for which of the three machines are predictive performance improved?

Chapter 2 explains the simple machine in depth. In the end of the chapter, a description of multiple simple machines is given. Chapter 3 gives background information about Bayesian statistics, latent Gaussian models, Gaussian fields and how the predictive quality can be measured. It also includes a summary of the relevant results obtained by Brynjarsdóttir and O'Hagan (2014) for the simple machine. Chapter 4 describes the different models, the different machines, the different designs and an overview of how many times a model is applied to which machine for which design. Chapter 5 describes the results and finally, chapter 6 gives a discussion and conclusion of the results.

Chapter 2

Case: Multiple simple machines

2.1 The simple machine

One of the physical systems that is studied in this thesis, is the simple machine. It is an example found in Brynjarsdóttir and O'Hagan (2014), so the model, the supplied terminology, the experimental setup and the results presented in this section, are based on the paper of Brynjarsdóttir and O'Hagan (2014).
The simple machine is a physical system, where the work is modeled, and depends on the amount of effort put into the machine and the efficiency of the machine. Both the efficiency and effort put into the machine are inputs. The efficiency θ is a calibration parameter, a parameter with an unknown value, while the effort put into the machine is a control variable x, a variable controlling the output.

True process and data generation:
The work $\zeta(x)$ done by the simple machine is given by

$$\zeta(x) = \frac{\theta x}{1 + x/a}, \tag{2.1}$$

where

- $\zeta(x)$ represents the work done by the machine,
- x represents the effort,
- $\theta = 0.65$ represents the efficiency,
- $a = 20$ is a constant.

The observations are synthetic and are generated from

$$\begin{aligned} z_i &= \zeta(x_i) + \epsilon_i, \quad i = 1, ..., n \\ \epsilon_i &\sim \mathcal{N}(0, 0.01^2) \end{aligned} \tag{2.2}$$

where

- z_i represents the work of the i'th observation,
- $x_i \in [0.2, 4]$ represent the effort of the i'th observation,
- $\zeta(x_i)$ represents true work of the simple machine (equation (2.1)),
- ϵ_i's represents the observation errors and are independent and identically distributed (i.i.d.) normal random variables,
- $n = 11, 31$ and 61 represents the number of observations.

The experiment is performed three times. One time with 11 observations, one time with 31 observations and one time with 61 observations. For each experiment, the effort x_i's of the observations are evenly spaced over the interval $[0.2, 4]$.
After the observations are created, θ is treated as unknown, and is estimated later on. Notice how the mean of (2.1) is the true process of the simple machine, as the observation errors ϵ_i's have a mean of zero. The goal of predicting the work with a simulator (the simulation model used to learn about the physical process), is to predict the work from the true process.

Simulator:
The simulator for the simple machine, is set up so that effort x is proportional to the work, with the efficiency θ as the proportionality constant. With the simulator output $\eta(x, \theta)$ representing work, the equation for the simulator can be written as

$$\eta(x, \theta) = \theta x. \tag{2.3}$$

An important detail to notice, is that the efficiency θ in the simulator (2.3) represents the same θ as in the true process (2.1). From this fact, it is very obvious that there is a difference between the work predicted by the simulator and the work performed by the machine in the true process. Note that $\zeta(x)$ always has a smaller value than the linear equation (2.3) for the same value of θ. This is due to the fact that the denominator of (2.1) is greater than or equal to 1 for $x \geq 0$. As the denominator becomes larger with increasing x, the difference between the simulator output $\eta(x, \theta)$ and the work $\zeta(x)$ done by the true process, becomes larger. This difference is called the model discrepancy, and can be seen as loss of energy (e.g. due to friction). In **Fig. 2.1**, one can see the model discrepancy as the difference between the blue and red curve. The figure shows the work versus effort for the true process using equation (2.1) (blue curve) and the simulator using equation (2.3) (red curve). (2.2) is used to plot the observations ($n = 11$, black points) and the two red points at $x = 1.5$ and $x = 6$ are the prediction points. The curve of the true process, the simulator and the black points of the observations in the figure, all used the true value $\theta = 0.65$.

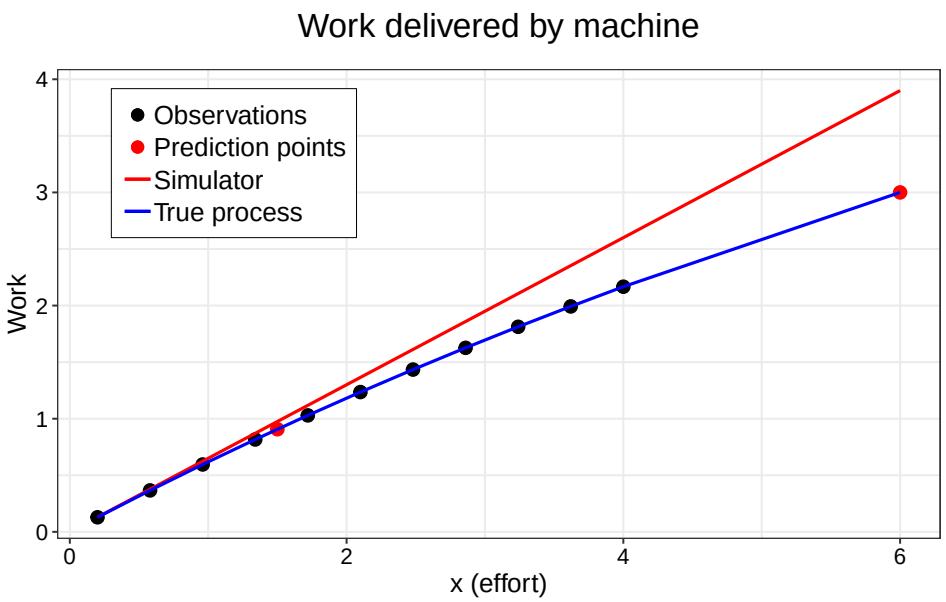


Figure 2.1: Work vs. effort for the simulator (red curve), the true process (blue curve) and $n = 11$ observations evenly spaced over the interval $[0.2, 4]$ (black points). The red points on the curve of the true process represent the prediction points. The calibration parameter is set to $\theta = 0.65$ for both the simulator and the true process, and $a = 20$ for the true process.

Prediction:
Brynjarsdóttir and O'Hagan (2014) perform interpolation at point $x = 1.5$ and extrapolation at point $x = 6$.

2.2 Multiple simple machines

One of the goals of this thesis, is to perform inference for multiple simple machines simultaneously, and see if the predictions of work becomes better than if the inference is performed on the simple machines individually. The simple machines have the same fixed true value for the efficiency θ, but different values for a. I.e. the machines are still simple machines, but with individual model discrepancies. The work $\zeta_j(x)$ done by machine number j, is given by

$$\begin{aligned} \zeta_j(x) &= \frac{\theta x}{1 + x/a_j}, \\ a_j &\sim \mathcal{N}(0, \sigma_a^2), \quad j = 1, 2, ..., N \end{aligned} \tag{2.4}$$

where θ and σ_a^2 have the same values for all the N machines, and the a_j's are generated from a normal distribution. For the multiple simple machines, the same simulator used for

the simple machine (equation (2.3)) is used.

Fig. 2.2, shows work vs. effort for 100 different simple machines. The machines have the same value for the efficiency $\theta = 0.65$, but the a_j's have been generated from a normal distribution $\mathcal{N}(0, 2^2)$, and differ for all the machines. The red line represents the simulator θx, with the true value of $\theta = 0.65$. Notice that, not only do the machines differ more from the simulator with larger values of x, but the simple machines differ more from each other with larger values of x as well. This can also be seen from the prediction points, which are the black points in the figure. The prediction points get more spread with larger values of x. In this thesis, observations are generally generated in the interval $x \in [0.2, 4]$, so the prediction points at $x = 1, 2$ and 3 are interpolation points, and the points at $x = 6$ and 8 are extrapolation points.

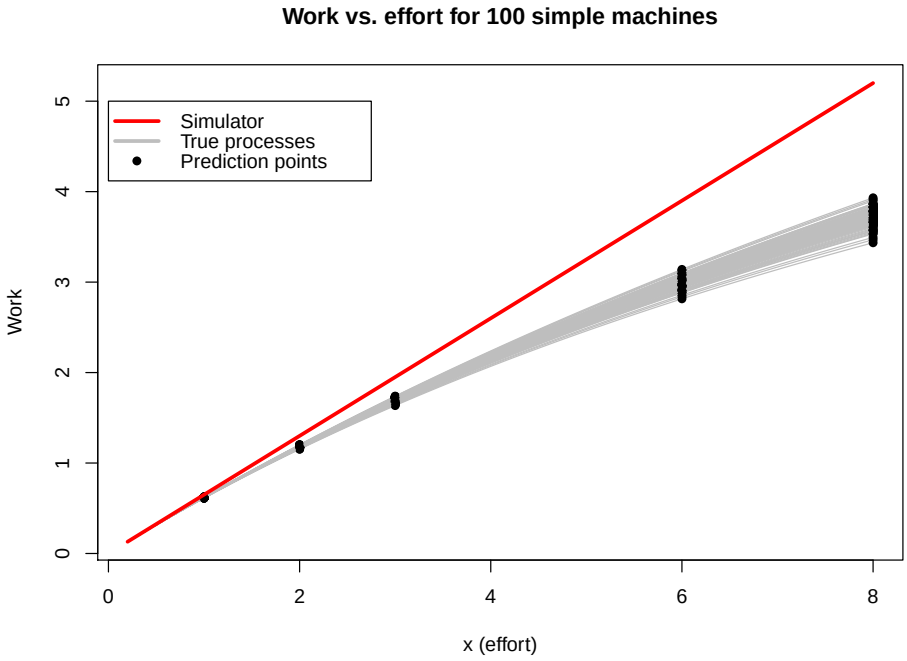


Figure 2.2: Work versus effort for 100 simple machines. The grey curves represent the true processes for the different simple machines, the red curve represent the simulator θx and the black points represent the prediction points. The efficiency has the same value for all the machines $\theta = 0.65$ and is also the same value that is used for the simulator in this plot. a varies for all the true processes of the different machines, and is generated from $\mathcal{N}(0, 2^2)$.

Chapter 3

Background

The models used in this thesis are called Latent Gaussian models (LGM), where one or two of the models' terms follow a Gaussian field (GF). To perform inference, Bayesian statistics is needed coupled with Markov chain Monte Carlo (MCMC) algorithms. However, there is a challenge with using MCMC algorithms, as they are computationally expensive, especially for a large system with many machines. To get around this, a package called INLA (stands for Integrated nested Laplace approximations) by Rue et al. (2009b) is used. INLA offers an alternative to MCMC algorithms and for the models used in this thesis, is much less computationally expensive. In the following sections, the concepts of Bayesian statistics, LGM and GF are explained, as well as a brief introduction to INLA.

3.1 Bayesian statistics

The brief overview in this section is based on Robert (2007).
The main idea of Bayesian statistics, is to merge prior knowledge about an event, and the data obtained, to make inferences. This is done to predict new data points, or to estimate parameters.

Parameter estimation:
Consider data points z (this can be a vector of values), with the pdf $f(z|\theta)$ where θ is a vector of parameters. Note that θ and z are general vectors of parameters and data points respectively, not necessarily the θ and z used in the case of the simple machine. Bayesian statistics is concerned with estimation of θ given the data z. The value of θ can be estimated with Bayes' theorem:

$$f(\theta|z) = \frac{f(z|\theta)f(\theta)}{\int_\Theta f(z|\theta)f(\theta)d\theta} \propto f(z|\theta)f(\theta), \tag{3.1}$$

where

- z represents the data
- θ represents the parameter(s)
- $f(\theta|z)$ is the posterior distribution (also called posterior) for θ, which is the distribution used to estimate θ.
- $f(\theta)$ is the prior distribution (also called prior) of θ. This represents the knowledge about θ before observing the data z.
- $f(z|\theta)$ is the distribution for the data given the model. It is called the likelihood function, and is the same likelihood function used in frequentist statistics.

The integral in the denominator integrates over the domain of θ, and because the integral becomes a constant, the posterior becomes proportional to the prior distribution and likelihood function. Notice that, as the posterior is proportional to a product of the prior and the likelihood function, it contains both information of what was thought to be the distribution of θ before obtaining the data, and information from the data itself. This is how the idea of merging prior knowledge and the information from the data to make inference (the main idea of Bayesian statistics) shows up in Bayes' theorem.

Prediction:
Bayesian statistics uses the posterior predictive distribution, to predict a new data point $\tilde{z}$. The posterior predictive distribution has the following form

$$f(\tilde{z}|x) = \int_{\Theta} f(\tilde{z}|\theta) f(\theta|z) d\theta,$$

where

- $\tilde{z}$ is the data point to be predicted.
- $f(\tilde{z}|z)$ is the posterior predictive distribution, the distribution used to predict the point $\tilde{z}$.
- $f(\theta|z)$ is the posterior distribution found in equation (3.1).
- $f(\tilde{z}|\theta)$ is the likelihood function.

Notice that the integral integrates over the domain of θ, accounting for the uncertainty in θ. Both the prior knowledge of θ and the information from the data can be found in the posterior distribution $f(\tilde{z}|x)$.

Credible interval (CI):
Credible intervals are the Bayesian statistics' analogue to frequentist statistics' confidence intervals. A $100(1-\alpha)\%$ credible interval is an interval covering $100(1-\alpha)\%$ of the posterior distribution. For prediction, it can be expressed as $(\tilde{z}_L, \tilde{z}_U)$, where

$$P(\tilde{z}_L < \theta < \tilde{z}_U|z) = \int_{\tilde{z}_L}^{\tilde{z}_U} f(\tilde{z}|z) d\theta = 1 - \alpha.$$

In this thesis, the mean and the 95% credible intervals using the 2.5th percentile to the 97.5th percentile of the posterior distributions, are used for predictions.

3.2 Latent Gaussian Models (LGM)

This section is based on Rue et al. (2009a) and (Blangiardo and Cameletti, 2015).
Let $\boldsymbol{z} = (z_1, ..., z_n)$ be the observed data, where n is the number of data points. A distribution for the data points $\boldsymbol{z}$ is specified, and in this thesis, the specified distribution is a Gaussian distribution. When the specified distribution is a Gaussian distribution, the mean $E[z_i]$ is equal to the additive predictor η_i. It can be written as

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(y_{li}),$$

where

- i is the index for the i'th data point
- β_0 represents the intercept,
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)$ represents the linear effects of explanatory variables,
- $x_{1i}, ..., x_{Mi}$ are covariates for data point i,
- $\boldsymbol{f} = (f_1(\cdot), ..., f_L(\cdot))$ are functions on covariates $y_{1i}, ..., y_{Li}$, that can take various forms (including non-linear effects),
- M and L represent the number of linear effects and number of functions respectively.

The model is a latent Gaussian model iff a Gaussian prior is assigned to β_0, the β_k's and the $f_l(\cdot)$'s. In this thesis $\boldsymbol{\tau} = (\beta_0, \boldsymbol{\beta}, \boldsymbol{f})$ is referred to as the latent field, and the other parameters $\boldsymbol{\psi} = (\psi_1, ..., \psi_K)$ are referred to as hyperparameters, where K is the number of hyperparameters. It is assumed that the observations are conditionally independent, and thus, the n observations have the following likelihood

$$p(\boldsymbol{z}|\boldsymbol{\tau}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p(z_i|\tau_i, \boldsymbol{\psi}).$$

Here, each z_i is connected to a τ_i, only one element in $\boldsymbol{\tau}$.
$\boldsymbol{\tau}$ follows a multivariate Normal prior $\boldsymbol{\tau} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\psi})^{-1})$, where the density function is given by

$$p(\boldsymbol{\tau}|\boldsymbol{\psi}) = (2\pi)^{-n/2} |\boldsymbol{Q}(\boldsymbol{\psi})|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{\tau}'\boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\tau}\right),$$

where $'$ denotes the transpose and $|\cdot|$ the discriminant. Supposedly, the components of $\boldsymbol{\tau}$ are conditionally independent, and this implies that $\boldsymbol{Q}(\boldsymbol{\psi})$ is a sparse matrix. With this specification, the latent Gaussian field becomes a Gaussian Markov random field (GMRF). Due to the precision matrix $\boldsymbol{Q}(\boldsymbol{\psi})$ being sparse, INLA saves considerable computation time as numerical linear algebra for sparse matrices can be used.

The objective is to find the marginal posterior distributions for each element in $\boldsymbol{\tau}$ and $\boldsymbol{\psi}$, which can be calculated as

$$p(\tau_i|\boldsymbol{z}) = \int p(\tau_i, \boldsymbol{\psi}|\boldsymbol{z}) d\boldsymbol{\psi} = \int p(\tau_i|\boldsymbol{\psi}, \boldsymbol{z}) p(\boldsymbol{\psi}|\boldsymbol{z}) d\boldsymbol{\psi}, \tag{3.2}$$

and

$$p(\psi_k|\boldsymbol{z}) = \int p(\boldsymbol{\psi}|\boldsymbol{y})d\boldsymbol{\psi}_{-k} \tag{3.3}$$

respectively. Here, τ_i is one element in the latent field $\boldsymbol{\tau}$ and the index $d\boldsymbol{\psi}_{-k}$ means that the integrand is integrated over all of $\boldsymbol{\psi}$ except ψ_k. Thus, to find the marginals $p(\tau_i|\boldsymbol{z})$ and $p(\psi_k|\boldsymbol{z})$, $p(\boldsymbol{\psi}|\boldsymbol{z})$ and $p(\tau_i|\boldsymbol{\psi}, \boldsymbol{z})$ need to be computed. To do this, the method of Laplace approximation is used.

Laplace approximation
The objective of Laplace approximation is to approximate the following integral

$$\int f(x)dx = \int \exp(\log(f(x))dx. \tag{3.4}$$

Here, $f(x)$ is the density function of a random variable X. $\log(f(x))$ can be represented as a Taylor series expansion

$$\log(f(x)) = \log(f(x^*)) + \frac{(x-x^*)^2}{2}\frac{\partial^2 \log(f(x))}{\partial x^2}\bigg|_{x=x^*} \tag{3.5}$$

evaluated at $x = x^*$, which is the mode of $f(x)$ ($x^* = \text{argmax}_x \log(f(x))$. Thus, the integral can be written as

$$\int f(x)dx = \exp(\log(f(x^*)) \int \exp\left(\frac{(x-x^*)^2}{2}\frac{\partial^2 \log(f(x))}{\partial x^2}\bigg|_{x=x^*}\right) dx, \tag{3.6}$$

which is found by substituting $f(x)$ by the exponential of the right side of (3.5). By letting $\sigma^{2^*} = -1/\frac{\partial^2 \log f(x)}{\partial x^2}\Big|_{x=x^*}$, the integrand becomes similar to the density of a normal distribution, and (3.6) can be written as

$$\int f(x)dx = \exp(\log(f(x^*)) \int \exp\left(-\frac{(x-x^*)^2}{2\sigma^{2^*}}\right) dx.$$

Note that the integrand can be seen as the kernel of a normal distribution, where x^* is the mean and σ^{2^*} is the variance. Hence, the integral evaluated in the interval (α, β) becomes

$$\int_\alpha^\beta f(x)dx = f(x^*)\sqrt{2\pi\sigma^{2^*}}(\Phi(\beta) - \Phi(\alpha)),$$

where $\Phi(\cdot)$ represents the cumulative density function of a normal distribution with mean x^* and σ^{2^*}. Note that, although Laplace approximation is an approximation method, only equality signs are used. This is because for Gaussian likelihoods, the expressions become equal.

Inference on $\boldsymbol{\tau}$ and $\boldsymbol{\psi}$:
To find the marginals $p(\tau_i|\boldsymbol{z})$ and $p(\psi_k|\boldsymbol{z})$, $p(\boldsymbol{\psi}|\boldsymbol{z})$ and $p(\tau_i|\boldsymbol{\psi}, \boldsymbol{z})$ need to be computed

(this can be seen in equation (3.2) and (3.3)). The joint posterior of the hyperparameters can be calculated as follows

$$\begin{aligned} p(\boldsymbol{\psi}|\boldsymbol{z}) &= \frac{p(\boldsymbol{\tau},\boldsymbol{\psi}|\boldsymbol{z})}{p(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{z})} \\ &= \frac{p(\boldsymbol{z}|\boldsymbol{\tau},\boldsymbol{\psi})p(\boldsymbol{\tau},\boldsymbol{\psi})}{p(\boldsymbol{z})}\frac{1}{p(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{z})} \\ &= \frac{p(\boldsymbol{z}|\boldsymbol{\tau},\boldsymbol{\psi})p(\boldsymbol{\tau}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{z})}\frac{1}{p(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{z})} \\ &\propto \frac{p(\boldsymbol{z}|\boldsymbol{\tau},\boldsymbol{\psi})p(\boldsymbol{\tau}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{z})} \\ &= \left.\frac{p(\boldsymbol{z}|\boldsymbol{\tau},\boldsymbol{\psi})p(\boldsymbol{\tau}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{\tilde{p}(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{z})}\right|_{\boldsymbol{\tau}=\boldsymbol{\tau}^*(\boldsymbol{\psi})} =: \tilde{p}(\boldsymbol{\psi}|\boldsymbol{z}), \end{aligned}$$

where $\tilde{p}(\boldsymbol{\tau}|\boldsymbol{\psi},\boldsymbol{z})$ is the Laplace approximation described in the last paragraph, and $\tau^*(\boldsymbol{\psi})$ is the mode given $\boldsymbol{\psi}$.

To compute $p(\tau_i|\boldsymbol{\psi},\boldsymbol{z})$ is more complicated, as there are generally more elements in $\boldsymbol{\tau}$ than in $\boldsymbol{\psi}$. One option is to rewrite $\boldsymbol{\tau} = (\tau_i, \boldsymbol{\tau}_{-i})$, where $\boldsymbol{\tau}_{-i}$ are all the components of $\boldsymbol{\tau}$ except τ_i, and then calculate the posterior conditional distributions $p(\tau_i|\boldsymbol{\psi},\boldsymbol{z})$ as follows

$$\begin{aligned} p(\tau_i|\boldsymbol{\psi},\boldsymbol{z}) &= \frac{p((\tau_i,\boldsymbol{\tau}_{-i})|\boldsymbol{\psi},\boldsymbol{z})}{p(\boldsymbol{\tau}_{-i},\tau_i,\boldsymbol{\psi},\boldsymbol{z})} \\ &= \frac{p(\boldsymbol{\tau},\boldsymbol{\psi}|\boldsymbol{z})}{p(\boldsymbol{\psi}|\boldsymbol{z})}\frac{1}{p(\boldsymbol{\tau}_{-i}|\tau_i,\boldsymbol{\psi},\boldsymbol{z})} \\ &\propto \frac{p(\boldsymbol{\tau},\boldsymbol{\psi}|\boldsymbol{z})}{p(\boldsymbol{\tau}_{-i}|\tau_i,\boldsymbol{\psi},\boldsymbol{z})} \\ &= \left.\frac{p(\boldsymbol{\tau},\boldsymbol{\phi}|\boldsymbol{z})}{\tilde{p}(\boldsymbol{\tau}_{-i}|\tau_i,\boldsymbol{\psi},\boldsymbol{z})}\right|_{\boldsymbol{\tau}_{-i}=\boldsymbol{\tau}^*_{-i}(\tau_i,\boldsymbol{\psi})} := \tilde{p}(\tau_i|\boldsymbol{\psi},\boldsymbol{z}). \end{aligned} \tag{3.7}$$

Here, $\tilde{p}(\boldsymbol{\tau}_{-i}|\tau_i,\boldsymbol{\psi},\boldsymbol{z})$ is the Laplace approximation described in the last paragraph, and $\boldsymbol{\tau}^*_{-i}(\tau_i,\boldsymbol{\psi})$ is its mode. However, the standard option to calculate $p(\tau_i|\boldsymbol{\psi},\boldsymbol{z})$ is to perform the simplified Laplace approximation, which is related to a Taylor's series expansion of $\tilde{p}(\tau_i|\boldsymbol{\psi},\boldsymbol{z})$ from (3.7). A more detailed explanation of this, can be found in Blangiardo and Cameletti (2015).

Once $\tilde{p}(\tau_i|\boldsymbol{\psi},\boldsymbol{z})$ and $p(\boldsymbol{\psi}|\boldsymbol{z})$ are obtained, $p(\tau_i|\boldsymbol{z})$ can be approximated by

$$\tilde{p}(\tau_i|\boldsymbol{z}) \approx \int \tilde{p}(\tau_i|\boldsymbol{\psi},\boldsymbol{z})\tilde{p}(\boldsymbol{\psi}|\boldsymbol{z})d\boldsymbol{\psi},$$

which are solved numerically through

$$\tilde{p}(\tau_i|\boldsymbol{z}) \approx \sum_j \tilde{p}(\tau_i|\boldsymbol{\psi}^{(j)},\boldsymbol{z})\tilde{p}(\boldsymbol{\psi}^{(j)}|\boldsymbol{z})\Delta_j,$$

for some integration points $\{\boldsymbol{\psi}^{(j)}\}$ with the related set of weights $\{\Delta_j\}$. To find each marginal posterior $\tilde{p}(\psi_k|\boldsymbol{z})$, an interpolation algorithm based on the density of $\tilde{p}(\boldsymbol{\psi},\boldsymbol{z})$

evaluated at the integration points $\{\boldsymbol{\psi}^{(j)}\}$ can be used. More on how INLA finds the integration points $\{\boldsymbol{\psi}^{(j)}\}$, can be found in Blangiardo and Cameletti (2015).

3.3 Gaussian fields (GF) and Stochastic partial differential equation (SPDE)

This section is based on Lindgren et al. (2015) and (Blangiardo and Cameletti, 2015).
A spatial process can be denoted as $\{\xi(x), x \in D \in \mathbb{R}^1\}$, where x is a spatial index and D is the fixed domain which x varies continuously on. If for any $n \in \mathbb{N}$ and any set of locations $(x_1, ..., x_n)$ the following vector $(\xi(x_1), ..., \xi(x_n))$ follows the multivariate normal distribution $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is a spatially structured covariance matrix, the spatial process is called a Gaussian field (GF). The elements of the covariance matrix $\boldsymbol{\Sigma}$ are defined by the covariance function $C(\cdot, \cdot)$ where $\Sigma_{ij} = \text{Cov}(\xi(x_i), \xi(x_j)) = C(\xi(x_i), \xi(x_j))$.

A spatial process is called second-order stationary if $\boldsymbol{\mu}(\boldsymbol{x}_i) = \mu$ for all i's, and $\boldsymbol{\Sigma}$ is only dependent on the distance vector $(x_i - x_j)$, such that $\text{Cov}(\xi(x_i), \xi(x_j)) = c(x_i - x_j)$. If the covariance function only depends on the euclidean distance $\|x_i - x_j\|$ (which is the absolute value as only one dimension is considered in this thesis), then the spatial process is called isotropic.

Gaussian fields can be seen as solutions to the linear fractional stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau\xi(x)) = \mathcal{W}(x), \tag{3.8}$$

where $\kappa > 0$ is the scale parameter, Δ is the Laplacian operator, α and τ control the smoothness and the variance respectively and $\mathcal{W}(x)$ is a Gaussian white noise process. The solution to this SPDE is a GF $\xi(x)$ with the matérn covariance matrix

$$\text{Cov}(\xi(x_i), \xi(x_j)) = \text{Cov}(\xi_i, \xi_j) = \frac{\sigma_0^2}{\Gamma(\lambda)2^{(\lambda-1)}}(\kappa|x_i - x_j|)^{\lambda}K_{\lambda}(\kappa|x_i - x_j|), \tag{3.9}$$

where $|\cdot|$ is the absolute value (euclidean distance in one dimension), σ_0^2 is the marginal variance and K_{λ} is the modified Bessel function of order $\lambda > 0$ and second kind. λ measures the degree of smoothness of the process and $\kappa > 0$ is a scaling parameter. It is related to the range r in the following way: $r = \frac{\sqrt{8\lambda}}{\kappa}$.

To link the SPDE from (3.8) with the parameters in the matérn covariance matrix (3.9), the following equations are used

$$\begin{aligned} \lambda &= \alpha - 1/2 \\ r &= \frac{\sqrt{8\lambda}}{\kappa} \\ \sigma_0^2 &= \frac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)^{1/2}\kappa^{2\lambda}\tau^2}. \end{aligned} \tag{3.10}$$

The default choice for α is 2 in R-INLA, and this is the value that is used in this thesis. According to Rasmussen (2004), When $\alpha \to \infty$, the matérn covariance matrix converges to the squared exponential covariance matrix with the form

$$\text{Cov}(\xi(x_i), \xi(x_j)) = \sigma_0^2 \exp\left(-2\frac{d^2}{r^2}\right), \tag{3.11}$$

where $d = |x_i - x_j|$. In R-INLA, $0 \leq \alpha \leq 2$. As 2 is the maximum value for α in R-INLA, it is set to $\alpha = 2$, so that the covariance matrix gets as similar to a squared exponential covariance matrix as possible. Hence, $\lambda = 3/2$ (as seen from (3.10)), and according to Rasmussen (2004), the covariance matrix obtains the following form

$$\text{Cov}(\xi(x_i), \xi(x_j)) = \sigma_0^2 \left(1 + \frac{2\sqrt{3}d}{r}\right) \exp\left(-\frac{2\sqrt{3}d}{r}\right). \tag{3.12}$$

The solution to the SPDE (the GF $\xi(x)$) can be approximated in the following way

$$\xi(x) = \sum_{g=1}^{G} \varphi_g(x)\tilde{\xi}_g.$$

Here, G represents the number of B-spline knot locations, the set $\{\tilde{\xi}_g\}$ are Gaussian distributed weights with mean zero and $\{\varphi_g\}$ is the set of deterministic basis functions. The precision matrix $\boldsymbol{Q}$ for $\tilde{\boldsymbol{\xi}} = \{\tilde{\xi}_1, ..., \tilde{\xi}_G\}$ is given by

$$\boldsymbol{Q} = \tau^2(\kappa^4\boldsymbol{C} + 2\kappa^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}),$$

where the elements of the diagonal matrix $\boldsymbol{C}$ is given by $C_{ii} = \int \varphi(x)ds$ and the elements of the sparse matrix $\boldsymbol{G}$ is given by $G_{ij} = \int \nabla\varphi_i(x)\nabla\varphi_j(x)ds$, where ∇ is the gradient operator. As the precision matrix $\boldsymbol{Q}$ is sparse, $\boldsymbol{\xi}$ is a GMRF and has the distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$. It is an approximation to the solution of the SPDE.

3.4 Bayesian calibration and inference for a single simple machine

Model

One of the models that Brynjarsdóttir and O'Hagan (2014) use to perform inference on the simple machine, the model that is of most interest in this thesis, is a model from Kennedy and O'Hagan (2001). After adjusting the model to fit the framework of the simple machine, the following model is obtained

$$z_i = \theta x_i + \delta(x_i) + \epsilon_i, \tag{3.13}$$

where z_i is the work for the i'th observation, θ is the efficiency and ϵ_i represents the observation error. $\delta(x_i)$ represents the model discrepancy and follows a zero-mean Gaussian

process with a squared exponential covariance function. Mathematically, it can be written as

$$\delta(\cdot) \sim \mathcal{GP}\left(0, \sigma_0^2 c(\cdot, \cdot|\psi)\right),$$

where

$$c(x_1, x_2|\psi) = \exp\left(-\left(\frac{x_1 - x_2}{\psi}\right)^2\right). \tag{3.14}$$

This is equivalent to a GF with a matérn covariance function, where $\alpha \to \infty$ (equation (3.11), where $2/r^2 = 1/\psi^2$). They let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, σ^2 and σ_0^2 follow a inverse gamma distribution, ψ follow a truncated gamma distribution and θ have a non-informative prior. By using Gibbs sampling and the Metropolis-Hastings algorithm, they perform inference to perform the prediction at $x = 1.5$ and $x = 6$, and the estimation of the calibration parameter θ.

Results:
The results of Brynjarsdóttir and O'Hagan (2014) that is of interest for this thesis, can be summarized as follows:

- The model succeeded at interpolating $x = 1.5$, as the posterior distribution centered around the true value, and the 90% credible interval became smaller with more observations.
- For extrapolation at $x = 6$, the posterior distribution did not center around the true value. However, the 90% credible interval covered the true value.

3.5 Evaluation of predictions

To measure the predictive performance, mean squared error (MSE), continuous rank probability score (CRPS), coverage probability and length of credible intervals are used.

Mean squared error (MSE):
According to Saigal and Mehrotra (2012), the MSE measures how accurate the predictions are and has the numerical value

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}(\hat{z}_j - z_j)^2.$$

Here, $\hat{z}_j$ and z_j represent the prediction (the mean of the predictive distribution) and the true value respectively. In this thesis, N is the number of machines and the index j represents the machine number. The closer the predictions are to the true values, the smaller MSE becomes. Thus, a smaller MSE indicates better predictions.

Continuous rank probability score (CRPS):
According to Gneiting and Raftery (2007), the CRPS can be defined as follows

$$\text{CRPS}(F, z) = \int_{-\infty}^{\infty}(F(y) - \mathbf{1}\{y \geq z\})^2 dy, \tag{3.15}$$

where $F(\cdot)$ is the cumulative density function of the predictive distribution, z is the true value of what is being predicted and $\mathbf{1}\{y \geq z\}$ represents the Heaviside function: the term takes value 1 if $y \geq z$ and 0 otherwise. According to Gneiting and Raftery (2007), the CRPS of a normal predictive distribution can be calculated as follows

$$\text{CRPS}(\mathcal{N}(\mu, \sigma^2), z) = \sigma \left[\frac{z-\mu}{\sigma} \left(2\Phi \left(\frac{z-\mu}{\sigma} \right) - 1 \right) + 2\phi \left(\frac{z-\mu}{\sigma} \right) - \frac{1}{\sqrt{\pi}} \right],$$

where μ and σ are the mean and standard deviation of the predictive distribution respectively, and ϕ and Φ are the probability density function and cumulative density function of a standard normal distribution respectively. A small CRPS indicates better predictions, as the term $(F(y) - \mathbf{1}\{y \geq z\})^2$ is smaller, which can intuitively be seen as the difference between the cumulative distribution function of the prediction and the cumulative distribution of the true value squared. CRPS can also be viewed as the sum of two terms, one representing sharpness and one representing calibration. According to Gneiting and Raftery (2007), sharpness is the concentration of the distribution, and calibration is the consistency between the predictive distribution and the true value.

Coverage probability:
According to Casella and Berger (2002), coverage probability is the probability that the credible interval contains the true value. Although the credible intervals of interest in this thesis are set to 95%, coverage probability is the true probability that the credible interval covers the true value, which might differ from 95%. To calculate this percentage, predictions are performed N times (one time for each machine) per prediction point. Then, by dividing the number of times the credible intervals cover the true value by N, an approximation of the coverage probability is found for the specific prediction point.

To evaluate the coverage probability found, the corresponding p-value is used. A p-value is the probability of obtaining the coverage probility found or a more extreme coverage probability, given that the true coverage probability is 95%. The p-value is calculated using only one tail. That is, if a coverage probability is smaller than 95% is found, the p-value is the probability of obtaining this specific coverage probability or a smaller coverage probability. If the coverage probability found is larger than 95%, then the p-value is the probability of obtaining this specific coverage probability or a larger coverage probability. If the p-value is smaller than 5%, it is concluded that the credible interval is not a 95%. Otherwice, it is concluded that the credible interval is 95% credible interval.

Length of CI:
The length of the credible intervals are also considered in this thesis. A smaller credible interval is considered as better than a larger credible interval, as the prediction is more certain of the true value. However, it is important to check if the credible intervals cover the true values for a method. Smaller credible intervals are not necessarily better if the intervals always miss the true values. Thus, it is important to also evaluate coverage probability, to check if the method producing shorter credible intervals does cover the true values the amount of times it is supposed to. To do this, the p-values of the coverage probabilities are evaluated.

Chapter 4

Bayesian calibration and inference for multiple machines

4.1 Multiple ideal machine models

There are three models that are used to perform inference on the multiple simple machines. The first model, referred to as the individual model, is the model from Brynjarsdóttir and O'Hagan (2014) with some modification to make it possible to run in INLA. This model assumes the machines have individual model discrepancy and individual parameters. The second model, referred to as Model 1, assumes the machines have individual discrepancy, while the parameters are the same. It is an extension of the individual model, as it is the same model, with the assumption that the parameters are identical included. The third model, referred to as Model 2, assumes the machines have an individual discrepancy term, common parameters and a common discrepancy term that is identical for all the machines. It is an extension of Model 1, as it is the same model with a common discrepancy term added. The individual model evaluates the machines individually, as it does not assume that the machines have anything in common. Model 1 and Model 2 evaluates the machines simultaneously, as they assume that the machines have common parameters. Additionally, Model 2 assumes that the machines have a common discrepancy term.

Individual model: Individual discrepancy, Individual parameters.
The individual model is the model from Brynjarsdóttir and O'Hagan (2014), and is the model from equation (3.13). However, it is modified to fit the framework of INLA: the Gaussian process is modeled with a Gaussian field (GF) with Matérn covariance function.

Mathematically, the individual model is written as

$$\begin{aligned} z_{ij} &= \theta_j x_{ij} + \xi_j(x_{ij}) + \epsilon_{ij}, \\ \xi_j(\cdot) &\sim \text{GF}(\tau_{\xi j}, \kappa_{\xi j}), \qquad\qquad i = 1, ...n_j, \quad j = 1, ..., N, \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma_j^2), \end{aligned}$$

where

- z represents work,
- x represents effort,
- θ_j represents efficiency (calibration parameter),
- $\xi_j(\cdot)$ represents the model discrepancy, and follows a Gaussian field with mean 0 and an exponential covariance function (equation (3.12)),
- ϵ_{ij} represents the observation error, and follows a normal distribution,
- $\tau_{\xi j}$ and $\kappa_{\xi j}$ represent the parameters of the Gaussian field,
- σ_j^2 represents the variance of the observation errors,
- i and j represent observation number and machine number respectively,
- n_j and N represent the number of observations for machine j and the number of machines respectively.

Note that all the parameters θ_j, $\tau_{\xi j}$, $\kappa_{\xi j}$ and σ_j^2 have different values for each machine. The discrepancy terms $\xi_j(\cdot)$'s are also different for each machine, and as it has an exponential covariance function, the model is different than the model of Brynjarsdóttir and O'Hagan (2014) (which uses a squared exponential covariance function). For the observation errors ϵ_{ij}, the values vary for each machine and for each observation, but follow the same distribution for the same machine (as σ_j^2 have the same value for the same machine).

Model 1: Individual discrepancy, Common parameters.
Model 1 is constructed by extending the individual model to consider all the machines simultaneously. Thus, it assumes that all the parameters have the same values accross all the machines. Mathematically, Model 1 is written as

$$\begin{aligned} z_{ij} &= \theta x_{ij} + \xi_j(x_{ij}) + \epsilon_{ij}, \\ \xi_j(\cdot) &\sim \text{GF}(\tau_{\xi}, \kappa_{\xi}), \qquad\qquad i = 1, ...n_j, \quad j = 1, ..., N, \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{4.1}$$

which is the exact same model as the individual model, except that the parameters θ, τ_{ξ}, κ_{ξ} and σ^2 have the same value for all the machines, and thus, do not have the subscript j. The model discrepancy term ξ_j are different for all the machines, following a GF with mean of 0, and an exponential covariance function (equation (3.12)). However, unlike the individual model, the model assumes the same parameter values τ_{ξ} and κ_{ξ} for all the machines. The observation errors ϵ_{ij} are i.i.d. for all observations for all machines.

Model 2: Common discrepancy, Individual discrepancy, Common parameters.
Model 2 is constructed by adding a common model discrepancy term to Model 1. Hence, it has two model discrepancy terms: one common and one individual. Mathematically, Model 2 is written as

$$\begin{aligned} z_{ij} &= \theta x_{ij} + \delta(x_{ij}) + \xi_j(x_{ij}) + \epsilon_{ij}, & \\ \delta(\cdot) &\sim \text{GF}(\tau_\delta, \kappa_\delta), & i = 1, ...n_j, \\ \xi_j(\cdot) &\sim \text{GF}(\tau_\xi, \kappa_\xi), & j = 1, ..., N, \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2), & \end{aligned}$$

where

- $\delta(\cdot)$ represents the model discrepancy term that is identical for all the machines, and follows a Gaussian field with mean 0 and an exponential covariance function (equation (3.12))
- τ_δ and κ_δ represent the parameters of the Gaussian field that $\delta(\cdot)$ follows,
- The other terms are exactly the same as the terms in Model 1 (equation (4.1)).

That $\delta(\cdot)$ is a common discrepancy term, means that not only are the values of τ_δ and κ_δ the same for all the machines, but $\delta(x)$ has the same value for all the machines for the same value of x. Other than the added common discrepancy term, Model 2 is identical with Model 1.

The motive behind constructing the model with two model discrepancy terms, is that $\delta(\cdot)$ might represent the common discrepancy that all the simple machines have (the difference between a simple machine and the simulator θx), while the $\xi_j(\cdot)$'s might represent the difference between each simple machine.

Multiple Ideal machines:
Before inference of the simple machines is performed using the individual model, Model 1 and Model 2, the models are tested on datasets generated from Model 1 and Model 2 themselves. The datasets generated from Model 1 and Model 2, represent work and effort for a machine referred to as the Ideal machine 1 and the Ideal machine 2 respectively. For the Ideal machines 1, N different GFs ($\xi_j(\cdot)$) are generated, one for each of the N machines. For the Ideal machines 2, $N + 1$ different GFs are generated, one for each of the N machines ($\xi_j(\cdot)$), and one common for all the machines ($\delta(\cdot)$). More on how the GFs are generated, can be found in section A.1.

One thing to note, is that these machines are conceptional. The machines might produce negative work for some values of effort, which is not realistic in the real world. The motive behind generating observations from Model 1 and Model 2, and performing inference on them, is to see if the predictions on these machines are successful, and if it is possible to learn something about the inference methods applied to those machines. Then, the three models are used to perform inference on the multiple simple machines.

4.2 Case studies:

There are three machines that are of interest in this thesis: the Ideal machine 1 and the Ideal machine 2 (both created using the framework of Model 1 and Model 2 from section 4.1 respectively), and the simple machine from Brynjarsdóttir and O'Hagan (2014).

4.2.1 True process, observations and simulator

The Ideal machine 1: Individual discrepancy, Common parameters
The Ideal machines 1 are created using the framework of Model 1. Their work are given by

$$\begin{aligned} \zeta_j(x) &= \theta x + \xi_j(x), \\ \xi_j(\cdot) &\sim \text{GF}(\tau, \kappa), \quad j = 1, ..., N, \end{aligned} \tag{4.2}$$

where

- $\zeta_j(x)$ represents work from the true process of the Ideal machines 1,
- x represents effort,
- θ represents the efficiency,
- $\xi_j(\cdot)$ represents the model discrepancy. It follows a GF and have a mean of 0 and an exponential covariance function (equation (3.12)),
- τ and κ represent the parameters of the Gaussian field,
- j and N represents the machine number and the number of machines respectively.

Note that, as with Model 1, the parameters θ, τ and κ have the same values for all the machines, and the GFs $\xi_j(\cdot)$ are different for each machine.

Fig. 4.1 shows work vs. effort for 100 different Ideal machines 1 represented by the grey lines, the simulator θx represented by the red line and the prediction points represented by the black dots. All the machines have the same value for the efficiency $\theta = 0.65$ and the same value is used for the simulator as well. As the $\xi_j(\cdot)$'s (model discrepancies) are different for all the 100 machines, the grey lines do not overlap. $\tau \approx 0.46$ and $\kappa \approx 2.45$ for all the Gaussian fields used to generate the model discrepancies.

The Ideal machine 2: Common discrepancy, Individual discrepancy, Common parameters.
The Ideal machines 2 are created using the framework of Model 2. Their true processes depend on a common model discrepancy term that is identical for all the N machines as well as a model discrepancy term that varies for each machine. The work from the true processes of the Ideal machines 2, are given by

$$\begin{aligned} \zeta_j(x) &= \theta x + c_1\delta(x) + c_2\xi_j(x), \\ \delta(\cdot) &\sim \text{GF}(\tau, \kappa), \\ \xi_j(\cdot) &\sim \text{GF}(\tau, \kappa), \quad j = 1, ..., N, \end{aligned} \tag{4.3}$$

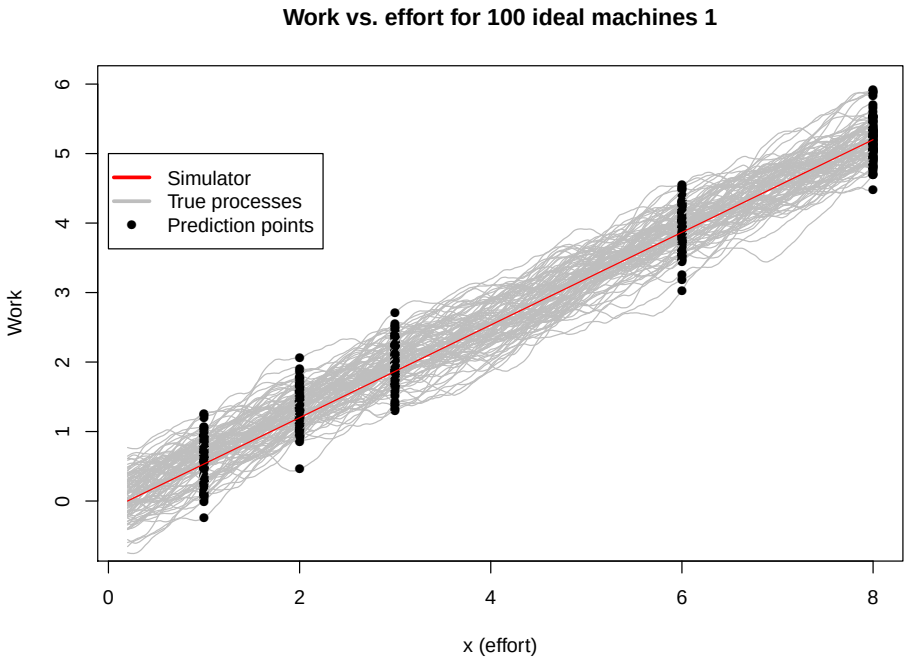


Figure 4.1: Work versus effort for 100 Ideal machines 1. The grey curves represent the true processes for the different simple machines, the red curve represent the simulator θx and the black dots represent the prediction points. The efficiency has the same value for all the machines $\theta = 0.65$ and is also the same value used for the simulator in this plot. The true processes have an individual discrepancy term, following a GF with mean 0 and an exponential covariance function. Its parameters have the values $\tau \approx 0.46$ and $\kappa \approx 2.45$.

where

- $\delta(\cdot)$ represents the model discrepancy that is identical for all the machines. It follows a GF and have a mean of 0 and an exponential covariance function (equation (3.12)).
- c_1 and c_2 are two constants to scale the variance of $\delta(x)$ and $\xi_j(x)$,
- The other terms represent the same as the terms of the Ideal machines 1 (equation (4.2)).

As with Model 2, the discrepancy term $\delta(x)$ not only have the same parameters τ and κ for all the machines, but the same value for a given x as well.

Note that τ and κ of the model discrepancy term $\delta(\cdot)$ have the same values as τ and κ for the model discrepancy term $\xi_j(\cdot)$. To make the two model discrepancies have different variances, different constants c_1 and c_2 are multiplied with the model discrepancy terms. The total variance of the model discrepancy terms var$(c_1\delta(x) + c_2\xi(x))$ are kept constant and equal to the variance of the model discrepancy term of the Ideal machines 1. This is

done by letting the constants satisfy $c_1^2 + c_2^2 = 1$, and the SPDE-parameters of the Ideal machines 1, have the same value as the SPDE-parameters of $\delta(\cdot)$ and $\xi_j(\cdot)$ of the Ideal machines 2 (an explanation of why this is the case, can be found in section A.2). The reason the total variance of the two model discrepancy terms of the Ideal machines 2, are set equal to the variance of the model discrepancy term of the Ideal machines 1, is because Model 1 is also applied to perform inference on the Ideal machines 2.

Fig. 4.2 shows work. vs effort for 100 different Ideal machines 2 represented by the grey lines, the simulator θx represented by the red line, the simulator and the common model discrepancy added ($\theta x + \delta(x)$) represented by the blue line and the prediction points represented by the black points. All the machines, the simulator and the simulator and model discrepancy added use the same value for the efficiency $\theta = 0.65$. The parameters τ and κ are set to 0.46 and 2.45 respectively for both the common model discrepancy term $\delta(\cdot)$ and the different model discrepancy terms $\xi_j(\cdot)$. As the $\xi_j(\cdot)$'s are different for all the N machines, the grey lines do not overlap. c_1 and c_2 are both set to $\sqrt{0.5}$, and thus, the variances of the two discrepancy terms are equal (var($c_1\delta(x)$) = var($c_2\xi(x)$)), as they are both generated from GFs with the same values for τ and κ.

Multiple simple machines:
Unlike the Ideal machines 1 and the Ideal machines 2, the multiple simple machines are not created using the framework of any of the models in this report. The true processes for multiple simple machines can be written as

$$\begin{aligned} \zeta_j(x) &= \frac{\theta x}{1 + x/a_j}, \\ a_j &\sim \mathcal{N}(0, \sigma_a^2), \quad j = 1, 2, ..., N \end{aligned} \tag{2.4}$$

for machine j, where θ and σ_a^2 have the same values for all the N machines, and the a_j's are generated from a normal distribution. More details about the true processes of the multiple simple machines and a work-versus-effort-plot can be found in section 2.2.

Observations
The observations for the three machines, are generated from

$$\begin{aligned} z_{ij} &= \zeta_j(x_{ij}) + \epsilon_{ij}, \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \qquad i = 1, ...n_j, \quad j = 1, ..., N, \end{aligned} \tag{4.4}$$

where

- z_{ij} represents the work for observation i and machine j,
- x_{ij} represents the effort for observation i and machine j,
- $\zeta_j(\cdot)$ represent the true process for machine j
- ϵ_{ij} represents the observation error for observation i and machine j. These differ for each observation and each machine, but have the same variance σ^2,
- n_j and N represent the number of observation for machine j and the number of machines respectively.

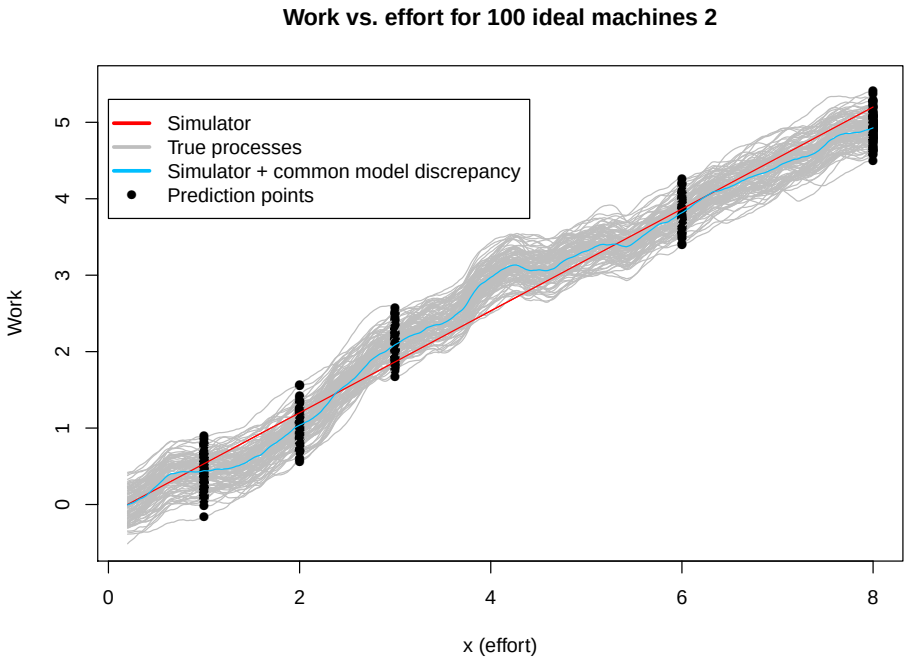


Figure 4.2: Work versus effort for 100 Ideal machines 2. The grey curves represent the true processes for the different simple machines, the red curve represents the simulator θx, the blue curve represents the simulator and the common model discrepancy term added $(\theta x + \delta(x))$ and the black dots represent the prediction points. The efficiency has the same value $\theta = 0.65$ for all the machines, the simulator and the simulator and common model discrepancy term added. The true processes have a common discrepancy term and an individual discrepancy term, both following a GF with mean 0 and an exponential covariance function. Their parameters have the same values $\tau \approx 0.46$ and $\kappa \approx 2.45$. $c_1 = \sqrt{0.5}$ and $c_2 = \sqrt{0.5}$, so the variances of the two discrepancy terms are equal (var$(c_1\delta(x))$ = var$(c_2\xi(x))$).

The true process $\zeta_j(\cdot)$ is substituted with (4.2), (4.3) or (2.4) when generating observations for the Ideal machines 1, the Ideal machines 2 or the multiple simple machines respectively. For the Ideal machines 1 and the Ideal machines 2, Gaussian fields are used to generate the model discrepancy terms (more on how GFs are generated, can be found in section A.1).

The true values for the parameters used to generate the observations, can be found in **Table. 4.1** for the Ideal machines 1, the Ideal machines 2 and the multiple simple machines. For the Ideal machines 2, three subcases are used with a unique pair of values for c_1 and c_2: $(c_1 = \sqrt{0.1}, c_2 = \sqrt{0.9})$, $(c_1 = \sqrt{0.5}, c_2 = \sqrt{0.5})$ and $(c_1 = \sqrt{0.9}, c_2 = \sqrt{0.1})$.

For both the Ideal machines 1 and the Ideal machines 2, the values for the parameters τ and κ are found by setting $\sigma_0^2 = 2 \cdot 0.2^2$ and $r = \sqrt{2}$ and solving the set of equations (3.10).

σ_0^2 is set to $2 \cdot 0.2^2$ because the prior Brynjarsdóttir and O'Hagan (2014) use for σ_0^2 for the simple machine, is an inverse gamma distribution with mode 0.2^2. The value is multiplied by 2, which makes the variance of each discrepancy term of the Ideal machines 2 subcase 2 equal to 0.2^2 (because var($c_1\delta(x)$) = c_1^2var($\delta(x)$) = $0.5 \cdot (2 \cdot 0.2^2) = 0.2^2$). For the spatial range, the value is found by setting the squared exponential covariance functions (3.11) and (3.14) equal, and $2/r^2 = 1/\psi^2$ is obtained. By setting $\psi = 1$ (Brynjarsdóttir and O'Hagan (2014) set the prior of ψ to a gamma distribution with mean 1) and solving the equation, $r = \sqrt{2}$ is obtained.

The variance of the observation error is set to $\sigma^2 = 0.01^2$ for all the machines, as this is what Brynjarsdóttir and O'Hagan (2014) use for the simple machine. If an entry in **Table. 4.1** is empty, then the parameter does not exist for this particular machine. All the values in the table are known when producing the observations, but are pretended to be unknown when inference is performed.

Simulator
The simulator $\eta(x, \theta)$ used for all three machines (the Ideal machines 1, the Ideal machines 2 and the multiple simple machines) is the same simulator that is used for the individual simple machine from section 2.1. It is given by

$$\eta(x, \theta) = \theta x,$$

where θ represents the efficiency (calibration parameter) for all the machines, and are set to the same value $\theta = 0.65$ (**Table. 4.1**).

Parameters	**Ideal machines 1**	**Ideal machines 2**	**Multiple simple machines**
θ	0.65	0.65	0.65
τ	0.46	0.46	
κ	2.45	2.45	
σ^2	0.01^2	0.01^2	0.01^2
σ_a^2			2^2
c_1		Subcase 1: $\sqrt{0.1}$ Subcase 2: $\sqrt{0.5}$ Subcase 3: $\sqrt{0.9}$	
c_2		Subcase 1: $\sqrt{0.9}$ Subcase 2: $\sqrt{0.5}$ Subcase 3: $\sqrt{0.1}$	

Table 4.1: The values set for the parameters when generating observations for the Ideal machines 1, the Ideal machines 2 and the multiple simple machines. If an entry is empty, then the parameter does not exist for the machine. For the Ideal machines 2, there are 3 subcases where constants c_1 and c_2 get different values. For example, for subcase 1, $c_1 = \sqrt{0.1}$ and $c_2 = \sqrt{0.9}$. All of the values in the tables are known when generating the observations, but are pretended to be unknown when inference is performed.

4.3 Model priors:

There are three models introduced in section 4.1: the individual model, Model 1 and Model 2. Each of these models, have a set of unknown parameters that are estimated during inference. For the individual model, there are $(4 \cdot N)$ unknown parameters: θ_j, σ_j^2, $\tau_{\xi j}$ and $\kappa_{\xi j}$ where $j = 1, ..., N$. For Model 1, there are 4 unknown parameters θ, σ^2, τ_ξ and κ_ξ, and for Model 2 there are 6 unknown parameters: θ, σ^2, τ_δ, κ_δ, τ_ξ and κ_ξ.

Table. 4.2 shows the priors of all the parameters for the three models. Although the individual model assumes different parameter values accross the N machines, the priors are set to the same distribution. Thus, the subscript j is dropped from the table, and the model's priors is set to the same priors as Model 1's priors. Model 2 have the same priors for θ and σ^2 as the individual model and Model 1, but different priors for the Gaussian field parameters.

For all three models, the prior of θ is set to $\mathcal{N}(0, 1000)$. Brynjarsdóttir and O'Hagan (2014) use a non-informative prior for θ, but in this thesis, a weakly informative prior is used, as it is not possible to use non-informative priors in INLA. The prior used is also the default prior in INLA for fixed effects, and thus, no specification in the code is needed for the prior.

The prior for the variance of the observations errors are set to
$\sigma^2 \sim$ Inv-Gamma$(181/19, 81/95000)$ for all three models. This is also the same prior that Brynjarsdóttir and O'Hagan (2014) use. The first parameter is the shape-parameter and the second parameter is the scale-parameter, and the values are chosen such that the distribution has a mode of 0.009^2 and a mean of 0.01^2. Note that the mean has the same value as the true value of the σ^2 used to generate the observations. The variance of the prior distribution is approximately $1.3 \cdot 10^{-9}$. Thus, it is an informative prior. Although the true value is pretended to be unknown during inference, it is an underlying assumption when performing inference that the variance of the observations errors has a value close to 0.01^2.

The parameters of the Gaussian fields for the individual model and Model 1 follow the same prior: a lognormal distribution. For all pairs of τ and κ, whether it is pair for a specific machine in the individual model or it is the pair for all machines in Model 1, the prior is $(\log(\tau), \log(\kappa)) \sim (\mathcal{N}(\log(0.46), 10), \mathcal{N}(\log(2.45), 10))$. Note that the log of the means have the same value as the values used to generate the observations for the Ideal machines 1. This is done to test if the models perform well on the Ideal machines 1 when the prior have the correct mean. The prior is a weakly informative prior, as it has the correct means, but large variances.

For Model 2, the log of the pair of parameters (τ_ξ, κ_ξ) and $(\tau_\delta, \kappa_\delta)$ have the same prior $(\mathcal{N}(\log(0.65), 10), \mathcal{N}(2.45, 10))$. Here, the log of mean of τ_ξ and τ_δ is not the same value as the value used to generate the observations for the Ideal machines 2. This is because when $\sigma_0^2 = 0.2^2$ and $r = \sqrt{2}$ (the values Brynjarsdóttir and O'Hagan (2014) use), $\tau \approx 0.65$ after solving (3.10). The reason why these values are not used for the individual model and Model 1, is that it is desired that the total variance of the model

discrepancies var($\xi(x) + \delta(x)$) of Model 2 is equal to the variance of the discrepancy term of the individiual model and Model 1. Thus, the marginal variance of the individual model and Model 1 is twice as large as each of the two marginal variances of Model 2: $\sigma_0^2 = 2 \cdot 0.2^2$. The prior is a weakly informative prior, as it has the correct means, but large variances.

Parameter(s)	**Individual model and Model 1**
θ	$\mathcal{N}(0, 1000)$
σ^2	IG$(181/19, 81/95000)$
$(\log(\tau_\xi), \log(\kappa_\xi))$	$(\mathcal{N}(\log(0.46), 10), \mathcal{N}(\log(2.45), 10))$
Parameter(s)	**Model 2**
θ	$\mathcal{N}(0, 1000)$
σ^2	IG$(181/19, 81/95000)$
$(\log(\tau_\xi), \log(\kappa_\xi))$	$(\mathcal{N}(\log(0.65), 10), \mathcal{N}(\log(2.45), 10))$
$(\log(\tau_\delta), \log(\kappa_\delta))$	$(\mathcal{N}(\log(0.65), 10), \mathcal{N}(\log(2.45), 10))$

Table 4.2: Priors of the parameters of the individual model, Model 1 and Model 2. For the individual model, the same priors are used accross all the N machines, and thus, the subscript j is neglected. For the three models, the prior of θ is a normal distribution, and the prior of σ^2 is an inverse gamma distribution, where the first parameter is the shape-parameter and the second parameter is the scale-parameter. The prior of the parameters of the GFs are lognormally distributed, and these are the only priors that is not the same for all models, as Model 2 have different parameter values.

4.4 Experimental design:

Data generation and design:
For all the experiments performed in this thesis, $N = 100$ machines are used. There are three different designs used to define the number of observations per machine and at which locations the observations are generated. The three designs are

- Design 1: 5 observations per machine. The 5 observations are generated from a uniform distribution $x_{ij} \sim \mathcal{U}(0.2, 4)$, and hence, the 100 machines have observations at different locations on this interval. The motivation behind using Design 1, is to evaluate if the coverage probabilities are large enough for the three models when there are few observations per machine.
- Design 2: 60 observations per machine. The 60 observations are generated from a uniform distribution $x_{ij} \sim \mathcal{U}(0.2, 4)$. This is the same design as Design 1, but with 60 observations instead of 5 observations for each of the 100 machines. The motivation behind using Design 2, is to compare the predictive performances of the three models for interpolation points and extrapolation points.
- Mixed design. The observations are generated from uniform distributions, but the intervals of the uniform distributions are not the same for each machine. In this design, the 100 machines are split into six cases, where the number of observations and/or the interval of the uniform distribution the observations are generated from, are different from case to case. The different cases can be described as follows:

  Case 1: The first 50 machines. 60 observations are generated from a uniform distribution in the interval $[0.2, 4]$ for each machine.

  Case 2: Machine number 51 to 60. 5 observations are generated from a uniform distribution in the interval $[0.2, 1.5]$ for each machine.

  Case 3: Machine number 61 to 70. 5 observations are generated from a uniform distribution in the interval $[1.5, 2.5]$ for each machine.

  Case 4: Machine number 71 to 80. 5 observations are generated from a uniform distribution in the interval $[2.5, 4]$ for each machine.

  Case 5: Machine number 81 to 90. 3 observations are generated from a uniform distribution in the interval $[0.2, 1]$ and 3 observations are generated from a uniform distribution in the interval $[3, 4]$ for each machine.

  Case 6: Machine number 91 to 100. 5 observations are generated from a uniform distribution in the interval $[0.2, 4]$ for each machine.

  Fig. 4.3 shows the interval of the uniform distributions the observations are generated from (the y-values of the grey bars) and the number of observations generated in each interval (the number inside the bars) for each case (name of x-values) for the Mixed design. The red horizontal lines represent the prediction points of interest. Note that the intervals of case 2, 3 and 4 cover one interpolation point each, and that the interpolation points that the intervals do not cover, are extrapolation points

for these particular machines. Hence, when evaluating the machines individually, these points are extrapolation points, but when the 100 machines are evaluated simultaneously, the observations around these prediction points are known for some of the machines. These prediction points, that is, those that are interpolation points for some machines and extrapolation points for other machines, are referred to as pseudo extrapolation points. To find out whether there is a large difference between evaluating pseudo extrapolation points individually and simultaneously, is the main reason why case 2, 3 and 4 are designed in this manner. For case 5, the prediction points $x = 1$, $x = 2$ and $x = 3$ are all interpolation points. However, point $x = 1$ and $x = 3$ are interpolation points that are are far from the observations at one side, while point $x = 2$ is far from the observations on both sides if the machines are evaluated individually. Case 1 is equivalent to Design 2, but with 50 machines instead of 100 machines, and is designed in this manner so that INLA have many data points in the whole range when evaluating the machines simultaneously. Case 6 is equivalent to Design 1, but with 10 machines instead of 100 machines. The motivation behind using Mixed design, is to compare the predictive performances for the three models at pseudo extrapolation points and interpolation points that are located far from the observations at at least one side.

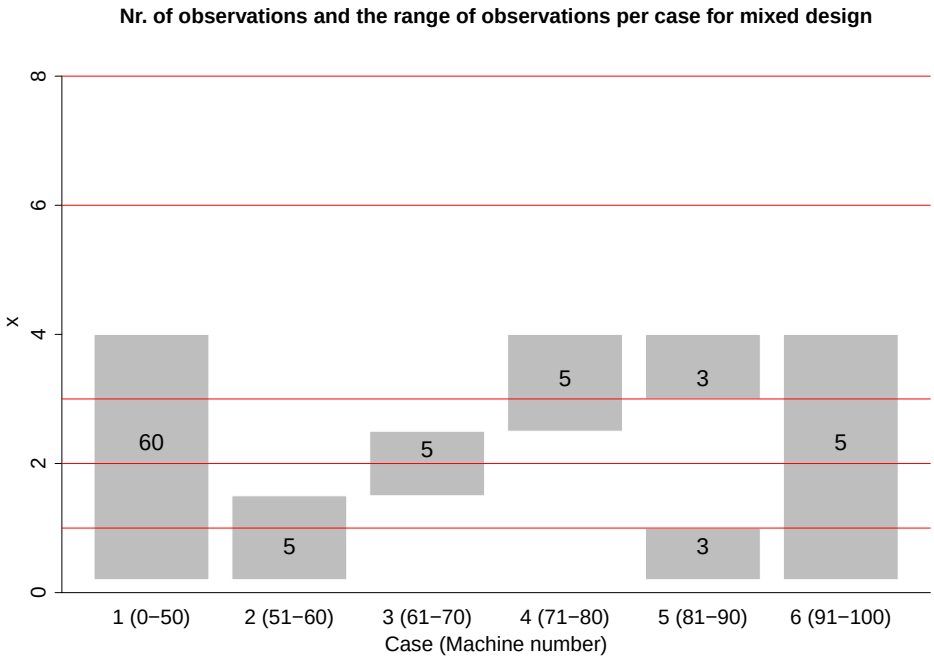


Figure 4.3: The number of observations and the range of observations per case for the Mixed design. For the two ends of each bar, the y-values show the minimum and the maximum value of the uniform distribution the observations are generated from, while the number inside each bar, is the number of observations generated. Cases are shown in the x-axis, where the machine numbers are shown in parenthesis. The red lines represent the prediction points of interest.

How each design is used to generate the locations (the effort) of the observations for each machine, can be summarized as follows:

- Ideal machine 1: All three designs are used to generate the locations of the observations for the Ideal machines 1.
- Ideal machine 2: All three designs are used to generate the locations of the observations for the Ideal machines 2 subcase 2 (when ($c_1 = \sqrt{0.5}, c_2 = \sqrt{0.5}$)). For subcase 1 and subcase 3 (when ($c_1 = \sqrt{0.1}, c_2 = \sqrt{0.9}$) and ($c_1 = \sqrt{0.9}, c_2 = \sqrt{0.1}$) repectively), only Design 2 and Mixed design is used (see **Table. 4.1** for an overview of the subcases).
- Simple machine: All three designs are used to generate the locations of the observations for the simple machines.

A summary of which of the designs are used to generate the observations for which combinations of machines and subcases, are shown in **Table. 4.3**.
After the locations (the effort) of the observations are generated, equation (4.4) is used to find the locations' corresponding values of work. The true processes $\zeta_j(x)$'s can either be replaced by (4.2), (4.3) or (2.4), depending on which machine the observations are generated for.

Prediction points
It is desired to perform predictions on four types of prediction points: interpolation points, extrapolation points, pseudo extrapolation points and points that are located far from the observations at at least one side. To evaluate the predictive performance on all four types of prediction points, five prediction points are considered: $x = 1, 2, 3, 6$ and 8. The first three points are interpolation points, while the last two points are extrapolation points. $x = 1$ and $x = 2$ are also used as pseudo extrapolation points. $x = 1$ is a pseudo extrapolation point when considering case 3 and case 4 for the Mixed design, while $x = 2$ is a pseudo extrapolation point when considering case 2 and case 4 for the Mixed design. When considering case 5, $x = 1$ is a point that is far from the observations at one side, while $x = 2$ is a point that is far from the observations at both sides. Thus, case 5 is also considered for $x = 1$ and $x = 2$ when Mixed design is applied.

4.5 Evaluation, model fitting

Experiments:
Table. 4.3 shows an overview of how many times a model is applied to perform inference on which combinations of machines, designs and subcases. For example, for the Ideal machines 2, subcase 2 with Mixed design, 100 experiments are performed with Model 2, while only 1 experiment is performed with the individual model and Model 1 each. The locations of the observations are set to be the same for all the 100 experiments. That is, for machine number j, the locations are the same for all the 100 experiments, but machine number j and machine number k have observations at different locations iff $j \neq k$. What is different for each experiment, is the model discrepancy term. For every experiment, new

Gaussian fields are generated: 100 Gaussian fields are generated for $\delta(\cdot)$ (one for each inference) and 10000 Gaussian fields are generated for $\xi(\cdot)$ (100 different GFs for each experiment, where each GF belongs to one machine). That the observations have the same locations for all experiments, but with different GFs for each experiment, is true whenever multiple experiments are performed. The motive behind performing multiple experiments with the same model, is to confirm the results of interest for the specific combination of model, design, subcase and machine.

Machine	**Design**	**Subcase**	**Individual model**	**Model 1**	**Model 2**
Ideal machines 1	1		10	1	1
Ideal machines 1	2		1	1	1
Ideal machines 1	Mixed		1	1	1
Ideal machines 2	1	2	10	1	1
Ideal machines 2	2	2	1	1	1
Ideal machines 2	Mixed	2	1	1	100
Ideal machines 2	2	1	1	1	1
Ideal machines 2	Mixed	1	1	1	1
Ideal machines 2	2	3	1	1	1
Ideal machines 2	Mixed	3	1	1	1
Simple machines	1		1	1	1
Simple machines	2		1	1	1
Simple machines	Mixed		1	1	1

Table 4.3: Overview of how many experiments are performed for a model on a specific combination of machine, design and subcase. If a model is applied to perform experiments multiple times (like Model 2 is applied to perform 100 experiments on the Ideal machines 2, Mixed design, subcase 2), the locations of the observations are the same, but the discrepancies vary for each experiment.

Evaluation of predictive performance

To evaluate the performance of the predictions, mean squared error (MSE), continuous rank probability score (CRPS), lengths of credible interval (CI) and coverage probability are used. Boxplots are used to study the distributions of the values of MSE, CRPS and lengths of CI. The smaller the spread of the distributions and the smaller the values of the distributions, the better the predictive performance. For coverage probabilities, p-values are used to decide if the CIs are 95% CIs. It is concluded that the CIs are not 95% CIs if the corresponding p-value is significant. If a model has the best CRPS, MSE and lengths of CI, but a coverage probability that is smaller than 95% and a significant p-value, then a model with worse CRPS, MSE and lengths of CI is considered a better model if its coverage probability is large enough. If a model has the best CRPS, MSE and lengths of CI, but a coverage probability that is larger than 95% and a significant p-value, then it is still considered the best model.

Chapter 5

Results

The individual model, Model 1 and Model 2 described in section 4.1, with the prior distributions shown in **Table. 4.2**, are used to perform predictions on the Ideal machines 1, the Ideal machines 2 and multiple simple machines described in section 4.2.1. The observations are generated from equation (4.4), using the true processes for the Ideal machines 1 (equation (4.2)), the Ideal machines 2 (equation (4.3)) and the multiple simple machines (equation (2.4)), the parameter values from table **Table. 4.1** and the designs explained in section 4.4. An overview can be found in **Table. 4.3**, where the number of times a model is applied to a specific machine with a specific design is shown. All the results found in this thesis, are found using code available at Lam (2019), written by the author. It is written in R, by R Core Team (2018). For the inference, the package INLA (by Rue et al. (2009b)) is used.

There are two classes of plots in this chapter, plots showing predictions and boxplots showing predictive performance. The first class of plots are referred to as prediction plots and the second class of plots are referred to as CRPS/MSE/CI boxplots for the rest of the thesis. For both classes of plots, blue represents the indvidual model, green represents Model 1 and black represents Model 2. If a plot contains true values of the machines, then these are represented by red points (red "x").

For the prediction plots, the x-axis represents the machine number (implying that it runs from 1 to 100) and the y-axis represents the work of the machine minus the prediction mean of Model 2. Hence, the prediction mean of Model 2 is always $x = 0$ for these plots, even though the prediction in general varies from machine to machine. The solid lines represent the prediction means for the three models, and the dashed lines represent the 95% CIs for the three models. For some plots, only solid lines are shown. These are referred to as coverage plots, and the solid lines represent the 95% CIs for the three models.

For the boxplots showing the predictive performance, the x-axis represents the prediction

points for Design 1 and Design 2, and case number for Mixed design. The y-axis represents the predictive measure (CRPS, MSE or length of CI), and the boxes show the spread of the predictive measures for all the 100 machines (Design 1 and Design 2) or the machines in the particular case (Mixed design).

There are two types of tables in this chapter, referred to as coverage tables: one type for Design 1 and Design 2, and one type for Mixed design. For Design 1 and Design 2, these tables show the coverage probabilities at each prediction point for each of the three models. Additionally, they show the total coverage probabilities using the coverage probabilities at all five prediction points. If multiple experiments are performed for a particular model, data from all of the experiments are used to calculate the coverage probabilities. The last column of these tables, shows the p-values of the total coverage probabilities.
For Mixed design, these tables show the coverage probability for the prediction points studied, and their respective p-values in parenthesis for each of the three models. If multiple experiments are performed for a particular model, then results from all of the experiments are used to calculate the coverage probabilites and p-values for the model.

5.1 Ideal Machines 1

5.1.1 Design 1:

The motivation behind using Design 1, is to check the coverage probability for the three models. With as few as 5 observations per machine, the coverage probability for the individual model is worse than the coverage probability for Model 1 and Model 2.

Fig. 5.1 shows the coverage plot for extrapolation at $x = 6$. In this plot, the credible intervals of Model 1 and Model 2 seem to overlap. They both cover 98% of the points, unlike the individual model, which only covers 88% of the points.

Table. 5.1 shows the coverage table. As can be seen from this table, the coverage probability is worse for the individual model, and the p-value is less than 5%. Hence, it is significant, and its credible intervals are conluded to be smaller than 95%. Note that, there are 10 experiments performed for the individual model, which affects its values in the table. Especially its p-value gets affected by this, as it is much smaller than if only 1 experiment is performed.

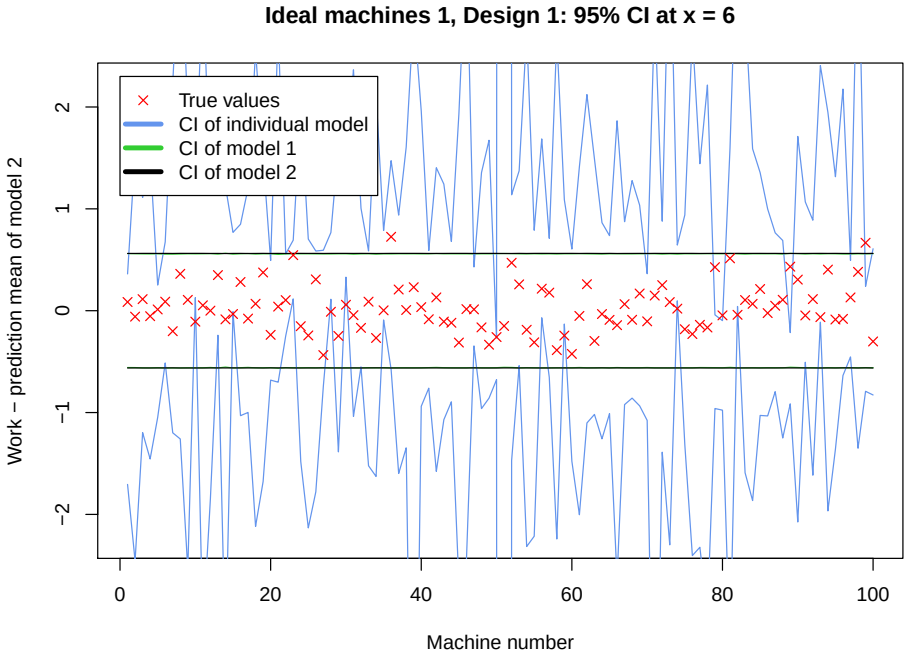


Figure 5.1: Coverage plot for the Ideal machines 1, Design 1 at $x = 6$. The blue, green and black curves represent the 95% credible intervals for the individual model, Model 1 and Model 2 respectively, and the red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	89.8	88.8	90.2	90.1	90.7	89.9	$2.5 \cdot 10^{-46}$
Model 1	94.0	92.0	97.0	98.0	96.0	95.4	39.0
Model 2	95.0	91.0	98.0	98.0	96.0	95.6	31.2

Table 5.1: Coverage table for the Ideal machines 1, Design 1. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. For the individual model, 10 experiments are used to calculate the values, while for Model 1 and Model 2, 1 experiment is used.

5.1.2 Design 2:

The motivation behind using Design 2, is to compare the predictive performance of the three models for interpolation and extrapolation. With 60 observation points per machine, there is no considerable difference between the individual model, Model 1 and Model 2 when performing interpolation. For extrapolation however, Model 1 and Model 2 have better predictive performances than the individual model.

Fig. 5.2 shows the prediction plots at $x = 2$ (interpolation) and $x = 8$ (extrapolation). The prediction means and CIs of the three models seem to overlap when interpolating at $x = 2$, and it is difficult to tell if any of the three models are better than the other. When extrapolating at $x = 8$, there is a clear difference between the individual model compared to Model 1 and Model 2. Model 1 and Model 2 seem to overlap, while the individual model seems to have larger lengths of CIs.

Fig. 5.3 shows the CRPS boxplots for the interpolation points and extrapolation points. The CRPS values for the extrapolation points are less spread and in general smaller for Model 1 and Model 2 compared to the individual model. For the interpolation points, there is no considerable difference between the distribution of the CRPS values for the three models.

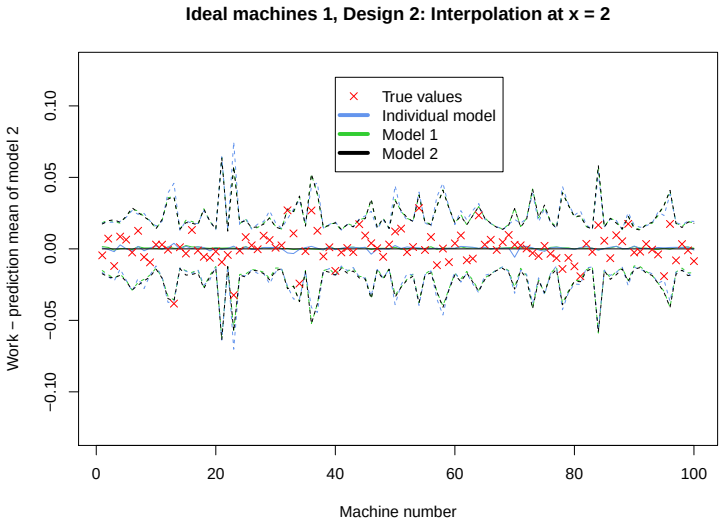
Fig. 5.4 shows the MSE boxplots for the interpolation and extrapolation points. Although it is not obvious from the prediction plots, the MSE for the extrapolation points are in general smaller and less spread for Model 1 and Model 2 compared to the individual model. For the interpolation points, there are no considerable difference between the MSE of the three models. These general conclusions, are similar to what is found for the CRPS.

Fig. 5.5 shows the CI boxplots for the interpolation and extrapolation points. As with the CRPS and MSE boxplots, there are no considerable difference between the three models when considering the interpolation points. However, when considering the extrapolation points, the lengths of CIs have a very small spread for Model 1 and Model 2 compared to the individual model. Not only is the spread very small, but the lengths of CIs values are also smaller than the medians of the individual model.

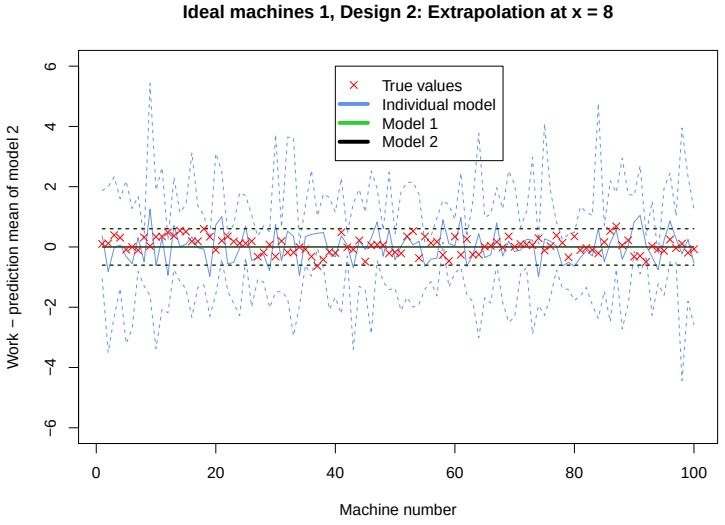
Table. 5.2 shows the coverage table. There are no significant p-values, and it is concluded that the CIs for the three models are 95% CIs. Hence, although the length of the CIs for the predictions of Model 1 and Model 2 are smaller than for the individual model's predictions, the coverage probabilites are still as good as the individual model's.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	94	98	95	96	90	94.6	36.9
Model 1	94	99	94	96	98	96.2	12.7
Model 2	94	99	95	96	98	96.4	8.6

Table 5.2: Coverage table for the Ideal machines 1, Design 2. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. The values are calculated using 1 experiment for each model.

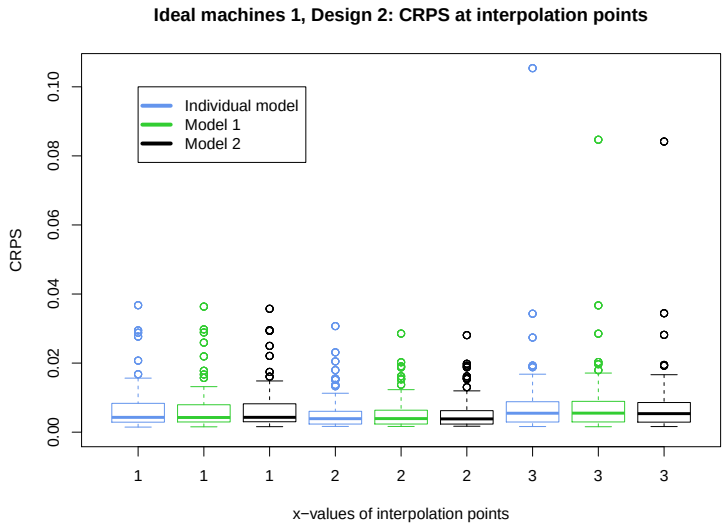


(a)

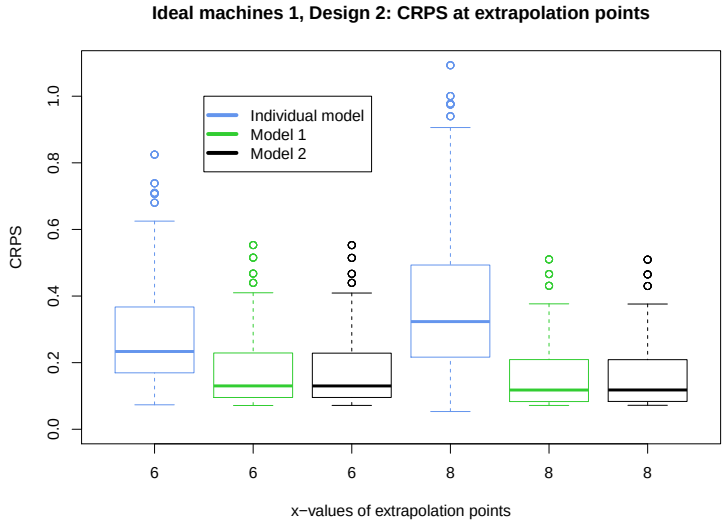


(b)

Figure 5.2: Prediction plots for the Ideal machines 1, Design 2 at $x = 2$ (a) and $x = 8$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

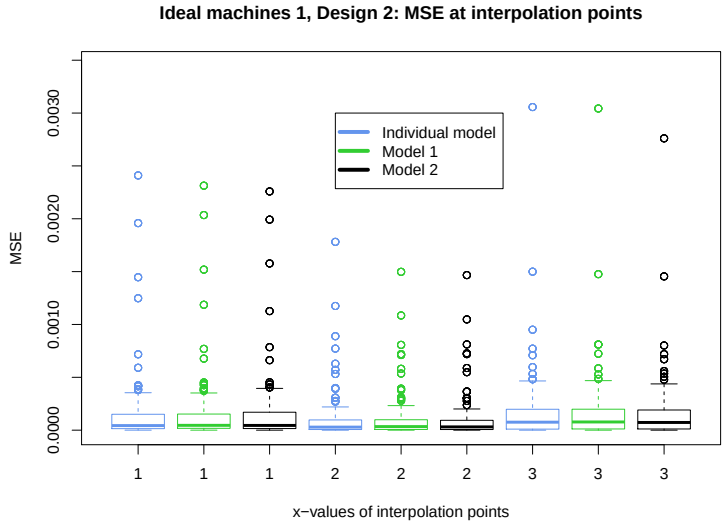


(a)

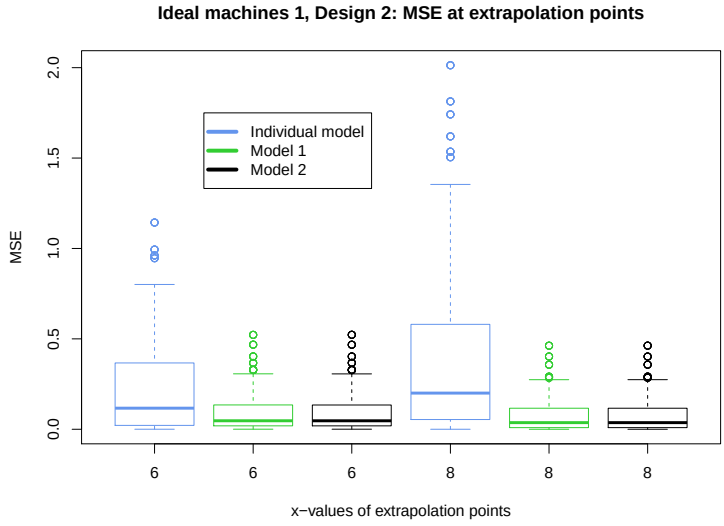


(b)

Figure 5.3: CRPS boxplots for the Ideal machines 1, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

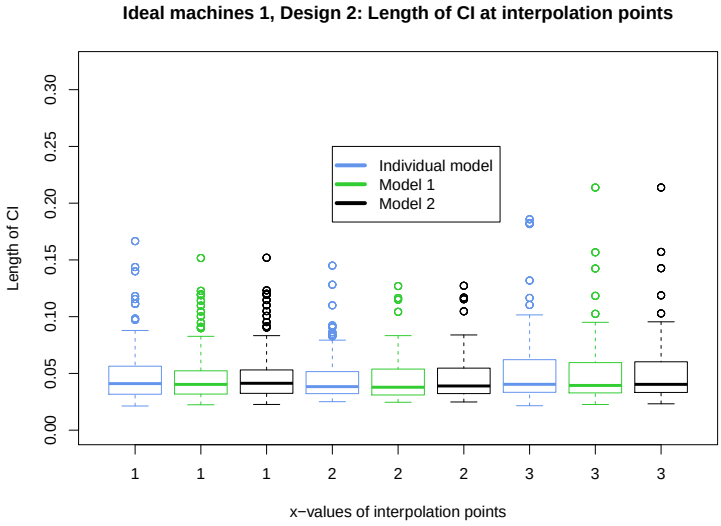


(a)

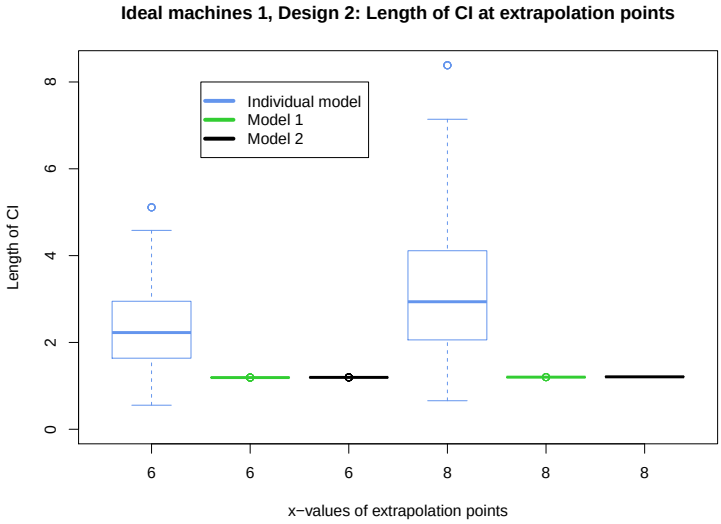


(b)

Figure 5.4: MSE boxplots for the Ideal machines 1, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.5: CI boxplots for the Ideal machines 1, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

5.1.3 Mixed design:

The motivation behind using Mixed design, is to compare the predictive performance of the three models at pseudo extrapolation points (prediction points that are extrapolation points for some machines and interpolation points for others), and at interpolation points that are far from the observations at at least one side. For both types of prediction points, Model 1 and Model 2 performs better than the individual model, and there are no considerable difference between the predictive performance of Model 1 and Model 2. The interpolation points considered in this subsection, are $x = 1$ and $x = 2$. $x = 1$ is a pseudo extrapolation point for case 3 and 4, while for case 5, the observations are close on one side and distant on the other side. $x = 2$ is a pseudo extrapolation point for case 2 and 4, while for case 5, the observations are distant on both sides. These are the only cases considered in this subsection.

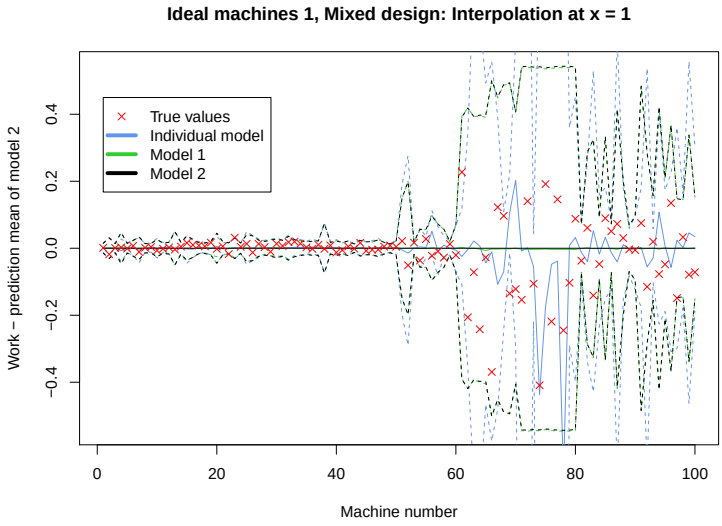
Fig. 5.6 shows the prediction plots at the interpolation points $x = 1$ and $x = 2$. There are considerable differences between the cases of the Mixed design for both the interpolation points. To look further into the details of each case, a closer look at the CRPS, MSE, CI boxplots and the coverage table is needed.

Fig. 5.7 shows the CRPS boxplots for the interpolation points $x = 1$ and $x = 2$. In general, the spread of CRPS and the CRPS values are smaller for Model 1 and Model 2 when the prediction point is a pseudo extrapolation point and for case 5. This can be seen for $x = 1$ case 3 and $x = 2$ case 5.

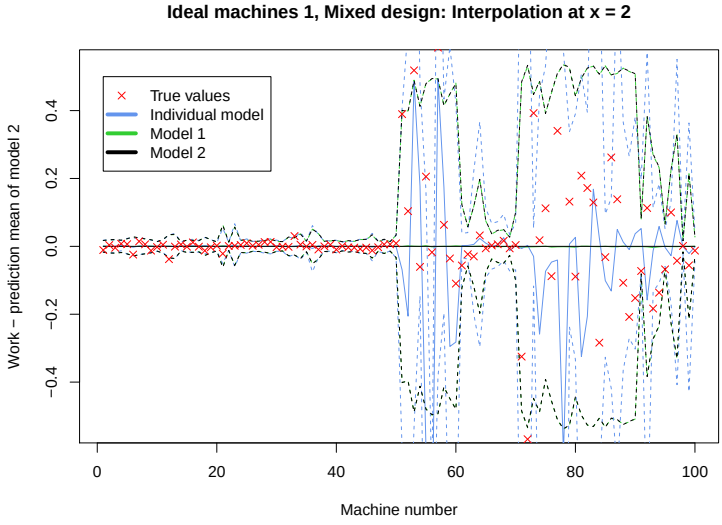
Fig. 5.8 shows the MSE boxplots for the interpolation points $x = 1$ and $x = 2$. It is not evident if Model 1 and Model 2 have better MSE values than the individual model. For example, for $x = 2$ case 4, the individual model seems to have better MSE, but for $x = 1$ case 3, Model 1 and Model 2 have better MSE. What seems to be consistent for the MSE boxplots, is that if the prediction point is a pseudo extrapolation point, then the MSE values are more spread for all models.

Fig. 5.9 shows the CI boxplots for the interpolation points $x = 1$ and $x = 2$. In general, the spreads and values of Model 1 and Model 2 are smaller than the spreads and values of the individual model the further away the observations are located from the prediction point. This is seen for all the cases considered: $x = 1$ case 3, 4 and 5, and $x = 2$ case 2, 4 and 5.

Table. 5.3 shows the coverage table. Note that, for all the relevant cases discussed in this subsection, the p-values are not significant. Thus, all the relevant CIs are concluded to be 95% CIs. Although Model 1 and Model 2 have smaller lengths of CI than the individual model has, the coverage probabilites of Model 1 and Model 2 are still large enough.

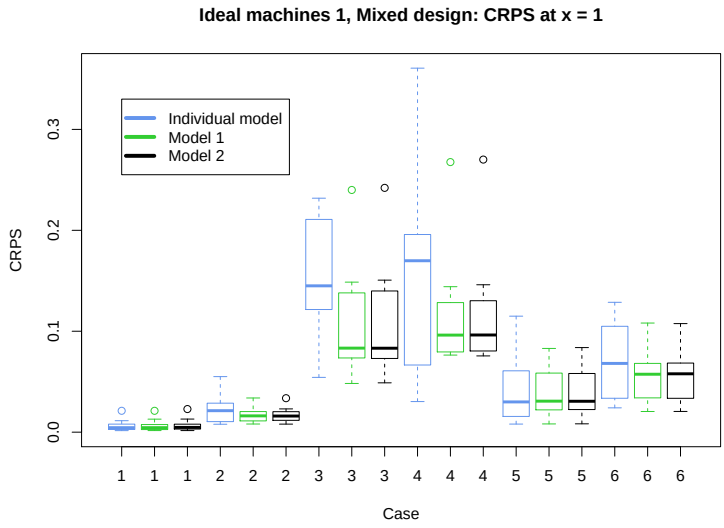


(a)

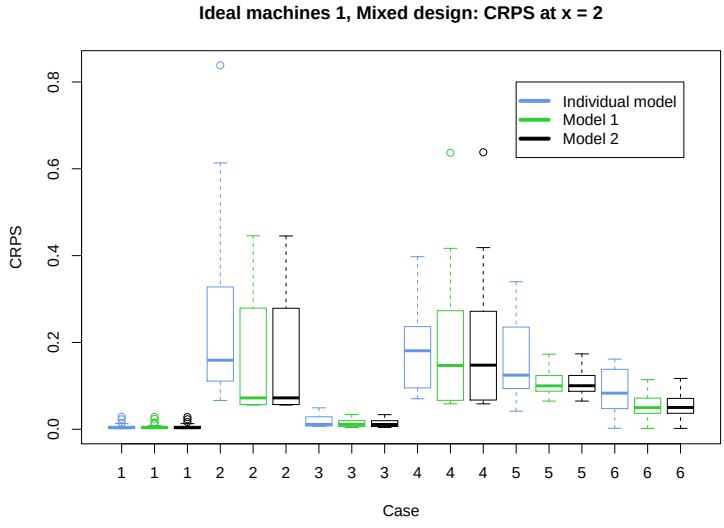


(b)

Figure 5.6: Prediction plots for the Ideal machines 1, Mixed design at $x = 1$ (a) and $x = 2$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

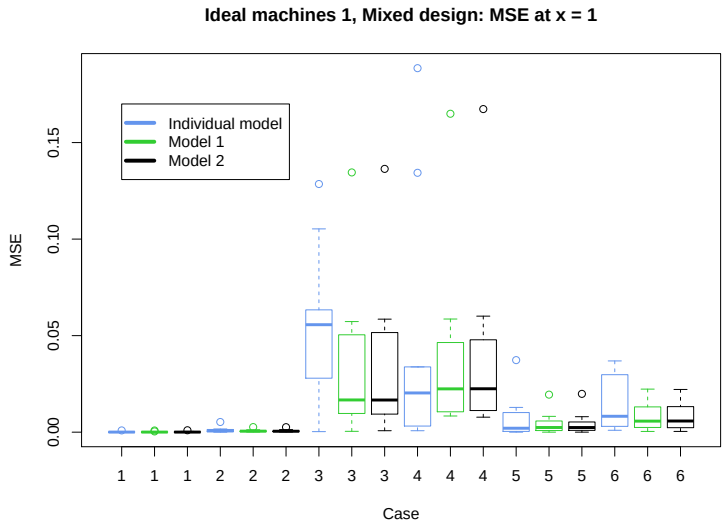


(a)

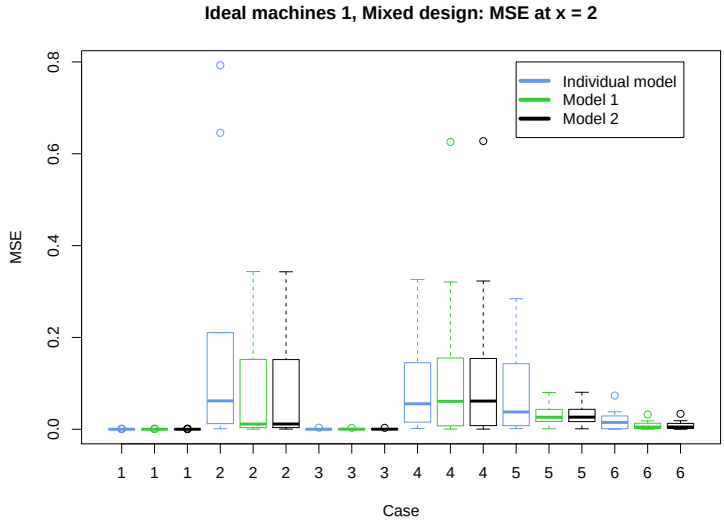


(b)

Figure 5.7: CRPS boxplots for the Ideal machines 1, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

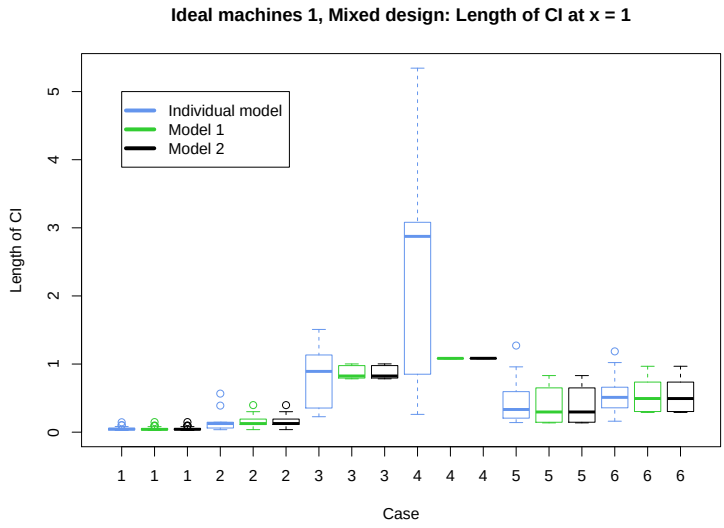


(a)

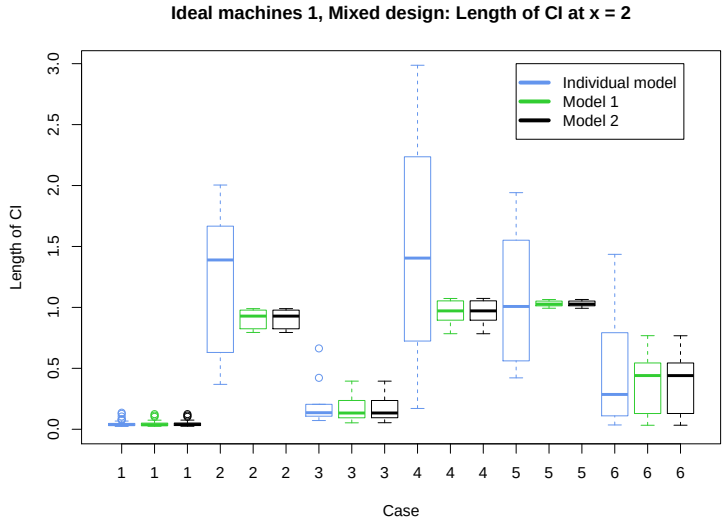


(b)

Figure 5.8: MSE boxplots for the Ideal machines 1, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.9: CI boxplots for the Ideal machines 1, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

3

Case	**Individual model**				**Model 1**				**Model 2**			
	$x = 1$		$x = 2$		$x = 1$		$x = 2$		$x = 1$		$x = 2$	
1	100	(8)	96	(54)	98	(28)	96	(54)	96	(54)	96	(54)
2	90	(40)	**80**	**(9)**	100	(60)	**80**	**(9)**	100	(60)	**80**	**(9)**
3	**80**	**(9)**	90	(40)	**100**	**(60)**	100	(60)	**100**	**(60)**	100	(60)
4	**100**	**(60)**	**90**	**(40)**	**100**	**(60)**	**80**	**(9)**	**100**	**(60)**	**80**	**(9)**
5	**100**	**(60)**	**90**	**(40)**	**100**	**(60)**	**100**	**(60)**	**100**	**(60)**	**100**	**(60)**
6	100	(60)	70	(1)	90	(40)	90	(40)	90	(40)	90	(40)

Table 5.3: Coverage table for the Ideal machines 1, Mixed design. The values show the coverage probabilities at the interpolation points $x = 1$ and $x = 2$, and the corresponding p-values in parenthesis, all given in percentages. Percentages in bold, represent the relevant cases discussed in this subsection. The values are calculated using 1 experiment for each model.

5.1.4 Summary, Ideal machines 1:

The motivation behind using Design 1, is to check the coverage probability for the three models. Their true processes is created using the framework of Model 1. Below follows a summary of the results from the Ideal machines 1.

- The CI of the predictions for the individual model do not cover 95% of the true values when there are few observations per machine.
- Interpolation is very similar for all three models when there are many observations per machine.
- Model 1 and Model 2 perform better extrapolation compared to the individual model when there are many observations per machine. They have better CRPS, MSE and length of CIs.
- Model 1 and Model 2 perform better pseudo extrapolation (prediction of points that are extrapolation points for some machines and interpolation points for others) and interpolation with points that are distant from the observations at at least one side, compared to the individual model. They have better CRPS and length of CIs, but it is not evident that they have better MSE.
- Model 1 and Model 2 have very similar predictive distributions for all experiments.

For the models considered to have better predictive performance based on the CRPS, MSE or lengths of CI, it is the spread and the values of the distributions of CRPS, MSE or lengths of CI that is considered. For example, if two models have the same median for the distributions of CRPS values, the one with the smaller spread is considered the best. This is because it has more stable results, and thus, it is considered better than a model that has really small CRPS values for some machines and really large CRPS values for other machines. If two models have approximately the same spreads on the distributions of the CRPS, then the one with the smaller values is considered the best. If a model has a larger spread, but smaller median than another model, then none of the two models are considered better than the other.

5.2 Ideal machines 2:

5.2.1 Subcase 2 ($c_1 = \sqrt{0.5}, c_2 = \sqrt{0.5}$), Design 1:

The motivation behind using Design 1, is to check the coverage probability of the individual model. With as few as 5 observations per machine, the coverage probability for the individual model is worse than the coverage probability for Model 1 and Model 2.

Fig. 5.10 shows the coverage plot for interpolation at $x = 2$. In this plot, the credible intervals of the individual model only covers 72% of the points, while Model 1 and Model 2 cover 90% and 94% of the points respectively.

Table. 5.4 shows the coverage table. As can be seen from this table, the coverage probability is worse for the individual model, and the p-value is less than 5%. Hence, it is significant, and its credible intervals are conluded to be smaller than 95%. Note that, there are 10 experiments performed for the individual model, which affects its values in the table. Especially its p-value gets affected by this, as it is much smaller than if only 1 experiment is performed. The p-values for Model 1 and Model 2 are also significant. Thus, it is concluded that the CIs of Model 1 and Model 2 are larger than 95% CIs.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	93.5	84.5	90.1	89.9	93.5	90.3	$3.0 \cdot 10^{-40}$
Model 1	100	90	99	100	100	97.8	0.1
Model 2	94	94	97	100	100	97	2

Table 5.4: Coverage table for the Ideal machines 2, subcase 2, Design 1. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. For the individual model, 10 experiments are used to calculate the values, while for Model 1 and Model 2, 1 experiment is used.

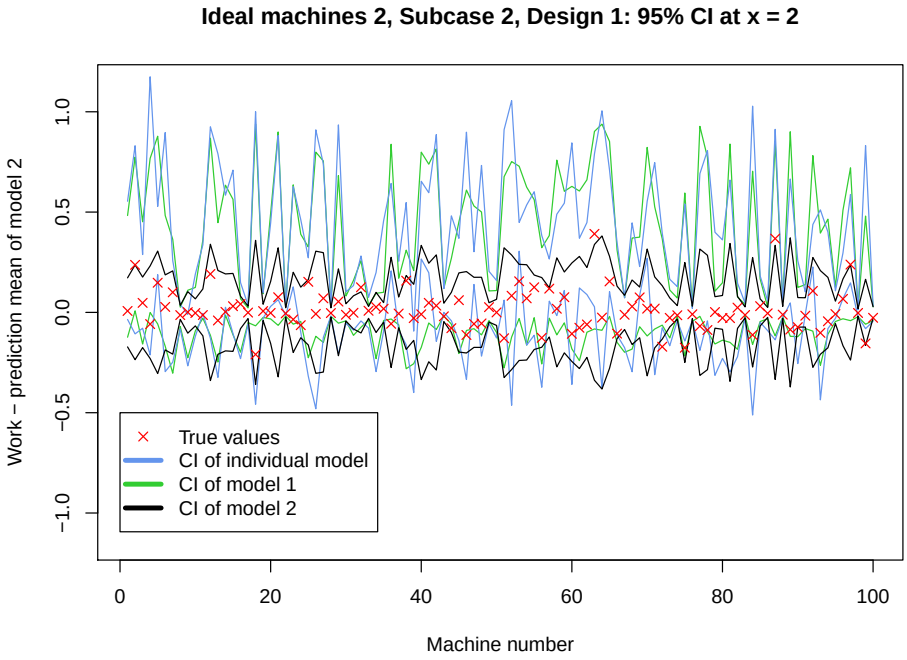


Figure 5.10: Coverage plot for the Ideal machines 2, subcase 2, Design 1 at $x = 2$. The blue, green and black curves represent the 95% credible intervals for the individual model, Model 1 and Model 2 respectively, and the red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

5.2.2 Subcase 2 ($c_1 = \sqrt{0.5}, c_2 = \sqrt{0.5}$), Design 2:

The motivation behind using Design 2, is to compare the predictive performance of the three models for interpolation and extrapolation. With 60 observation points per machine, the performance of the three models are very similar when performing interpolation, but Model 2 performs slightly better. For extrapolation however, Model 1 and Model 2 performs better than the individual model, with Model 1 being slightly better than Model 2.

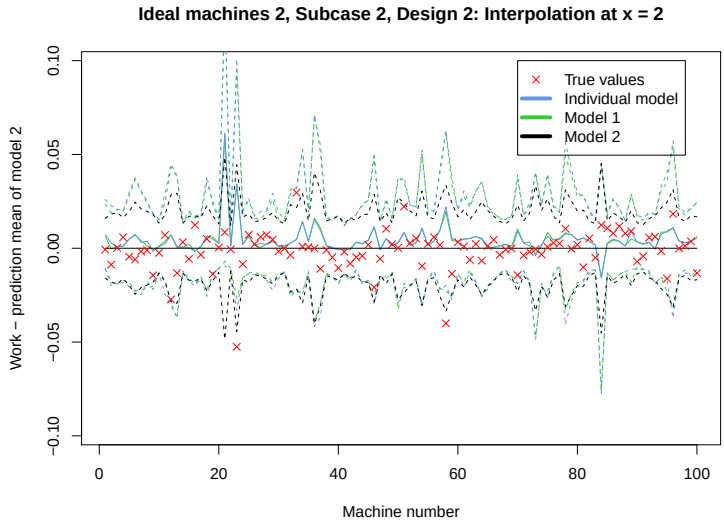
Fig. 5.11 shows the prediction plots at $x = 2$ (interpolation) and $x = 6$ (extrapolation). When interpolating at $x = 2$, there is a difference between the predictions of Model 2 compared to the individual model and Model 1. The individual model and Model 1 seem to have predictions that mostly overlap, while Model 2 have smaller length of CIs. There is a considerable difference between the three models when extrapolating at $x = 8$. The prediction means of Model 1 and Model 2 seem to overlap, but Model 1 has smaller length of CIs compared to Model 2. For the indvidual model, the length of CIs seems to be larger than both Model 1 and Model 2.

Fig. 5.12 shows the CRPS boxplots for the interpolation points and extrapolation points. The CRPS distributions for the interpolation points seem very similar for all three models, with Model 2 being slightly better (medians and largest CRPS values are smaller for Model 2). For the extrapolation points, Model 1 and Model 2 have smaller spreads and median, with Model 2 having the smallest spread. However, Model 1's medians are smaller than Model 2's medians, and it is difficult to decide which of the 2 models performs better.

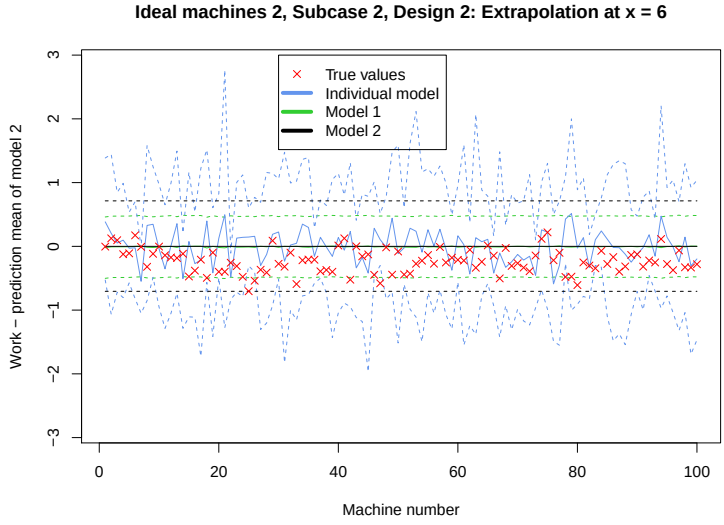
Fig. 5.13 shows the MSE boxplots for the interpolation and extrapolation points. Although it is not obvious from the prediction plots, the MSE for the extrapolation points are in general smaller and less spread for Model 1 and Model 2 compared to the individual model. For the interpolation points, the three models seem very similar, with Model 2 being slightly better (medians and largest CRPS values are in general smaller).

Fig. 5.14 shows the CI boxplots for the interpolation and extrapolation points. There is no considerable difference between the length of CIs for the individual model and Model 1 at the interpolation points. However, Model 2 has smaller values and a smaller spread than the individual model and Model 1 have. When considering the extrapolation points, the length of CI distributions have a very small spread for Model 1 and Model 2 compared to the individual model, with Model 1 performing better than Model 2, as its length of CIs are smaller than Model 2's.

Table. 5.5 shows the coverage table. The p-value for the individual model and Model 2 are significant (3.4% and 1.1% respectively), and it is concluded that their CIs are larger than 95% CIs. For Model 1, the p-value is not significant, and it is concluded that the CIs of Model 1 are 95% CIs. Thus, although Model 1 and Model 2 have smaller lengths of CI for extrpolation points and interpolation points respectively, the coverage probability is still large enough (the CIs still cover enough true values).

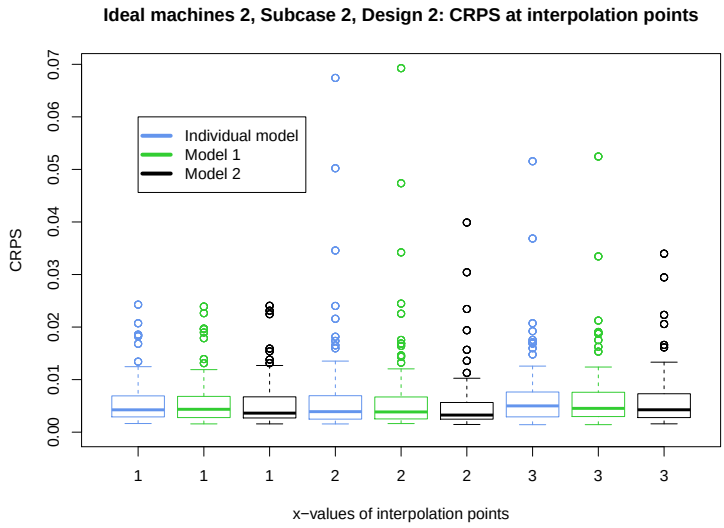


(a)

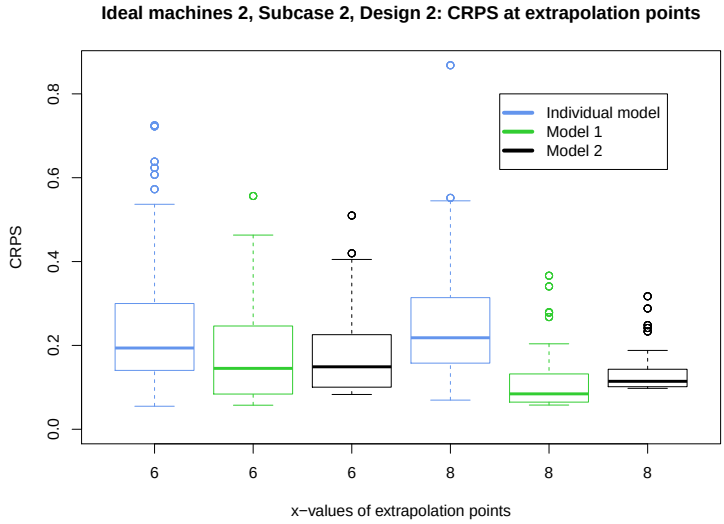


(b)

Figure 5.11: Prediction plots for the Ideal machines 2, subcase 2, Design 2 at $x = 2$ (a) and $x = 6$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

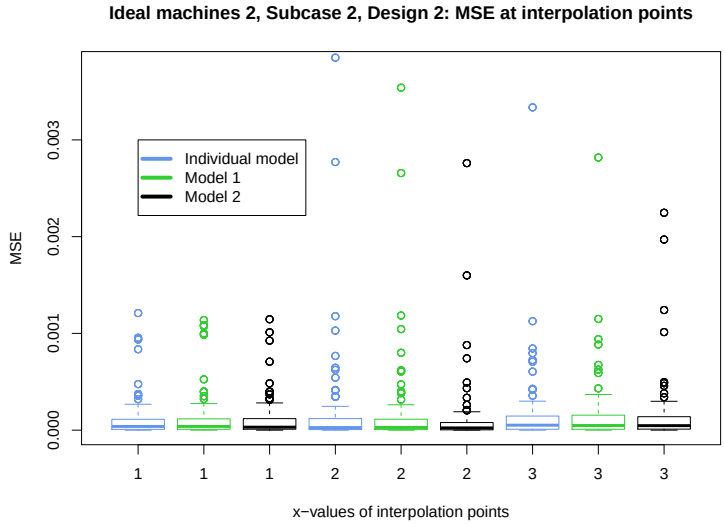


(a)

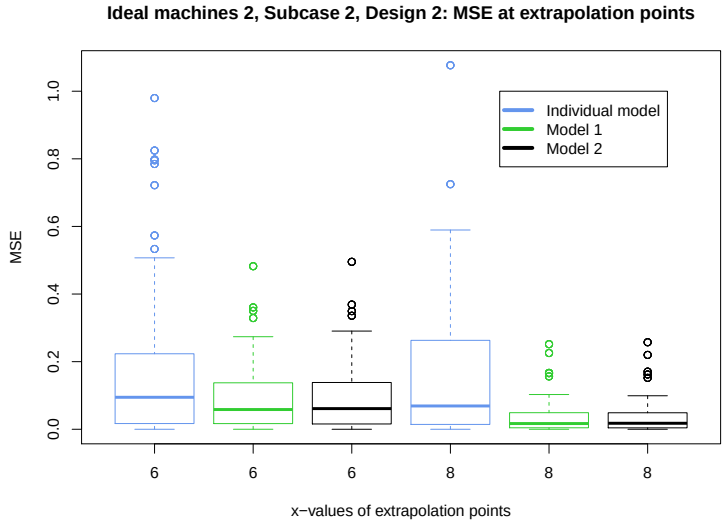


(b)

Figure 5.12: CRPS boxplots for the Ideal machines 2, subcase 2, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

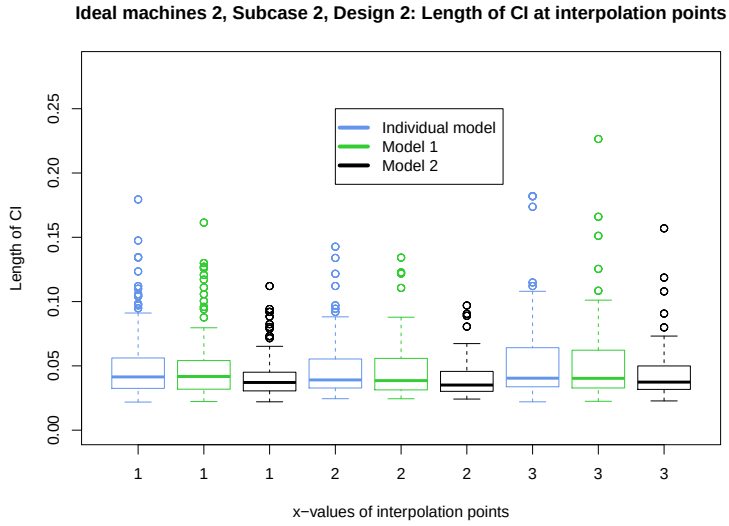


(a)

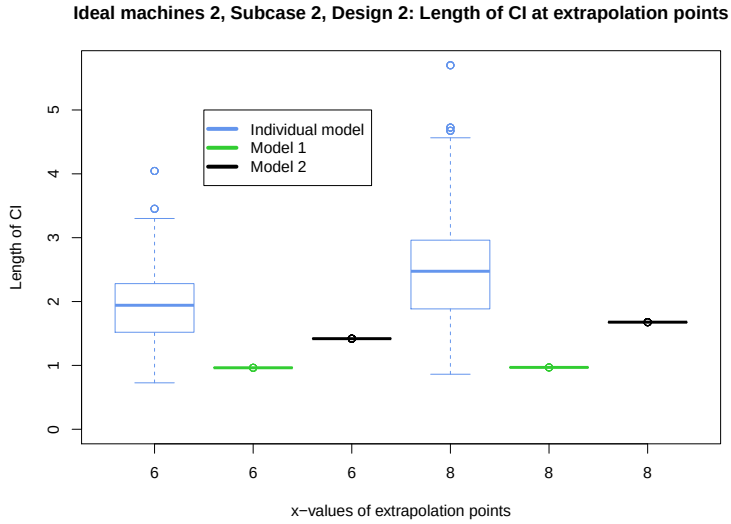


(b)

Figure 5.13: MSE boxplots for the Ideal machines 2, subcase 2, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.14: CI boxplots for the Ideal machines 2, subcase 2, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	98	97	98	94	97	96.8	3.4
Model 1	98	96	97	93	99	96.6	5.6
Model 2	94	97	95	100	100	97.2	1.1

Table 5.5: Coverage table for the Ideal machines 2, subcase 2, Design 2. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. The values are calculated using 1 experiment for each model.

5.2.3 Subcase 2 ($c_1 = \sqrt{0.5}, c_2 = \sqrt{0.5}$), Mixed design:

The motivation behind using Mixed design, is to compare the predictive performance of the three models at pseudo extrapolation points, and at interpolation points that are far from the observations at at least one side. For both types of prediction points, Model 2 have the best performance, followed by Model 1 and then the individual model. The interpolation points considered in this subsection, are $x = 1$ and $x = 2$. $x = 1$ is a pseudo extrapolation point for case 3 and 4, while for case 5, the observations are close on one side and distant on the other side. $x = 2$ is a pseudo extrapolation point for case 2 and 4, while for case 5, the observations are distant on both sides. These are the only cases considered in this subsection.

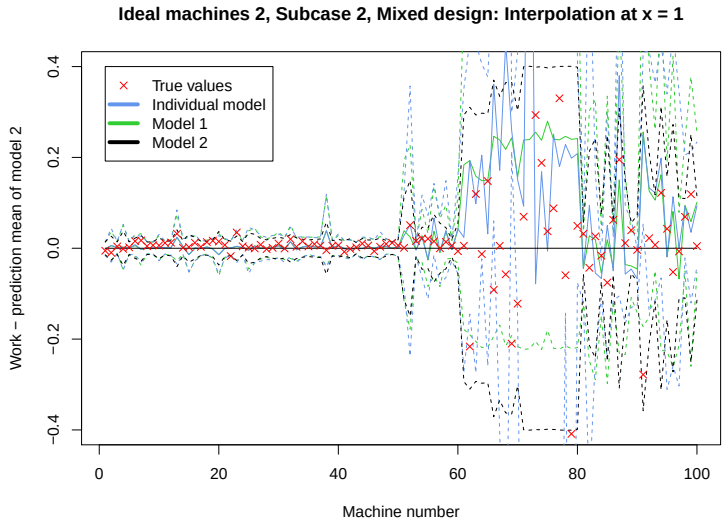
Fig. 5.15 shows the prediction plots at the interpolation points $x = 1$ and $x = 2$. There are considerable differences between the cases of the Mixed design for both the interpolation points. To look further into the details of each case, a closer look at the CRPS, MSE, CI boxplots and the coverage table is needed.

Fig. 5.16 shows the CRPS boxplots for the interpolation points $x = 1$ and $x = 2$. In general, when the prediction point is a pseudo extrapolation point and for case 5, the spread of CRPS and CRPS values are the smallest for Model 2, followed by Model 1 and then the individual model. This can be seen for $x = 1$ case 3 and case 5, and $x = 2$ case 2.

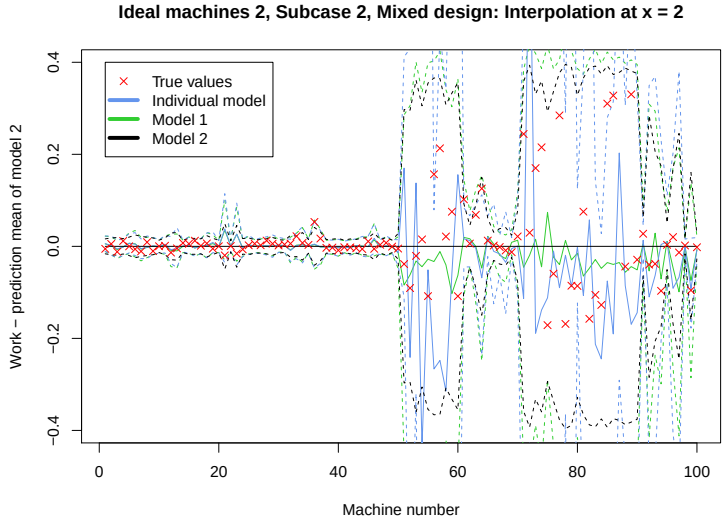
Fig. 5.17 shows the MSE boxplots for the interpolation points $x = 1$ and $x = 2$. In general, when the prediction point is a pseudo extrapolation point and for case 5, the spread of MSE and MSE values are the smallest for Model 2, followed by Model 1 and then the individual model. This can be seen for $x = 1$ case 3 and case 5, and $x = 2$ case 2. Additionally, if the prediction point is a pseudo extrapolation point, then the MSE values are more spread for all models.

Fig. 5.18 shows the CI boxplots for the interpolation points $x = 1$ and $x = 2$. In general, the further away the observations are located from the prediction point, the greater the difference between the models. When the prediction point is a pseudo estrapolation point and for case 5, Model 2 has the smallest length of CIs values and the smallest spread, followed by Model 1 and then the individual model. This is seen for $x = 1$ case 4 and 5, and $x = 2$ case 2, 4 and 5.

Table. 5.6 shows the coverage table. Note that, the relevant p-values for Model 2 are all above 5%. Thus, it is concluded that the CIs for the relevant cases of Model 2 are 95% CIs. The individual model and Model 1 have a significant p-value each, one at $x = 1$ case 3 and $x = 1$ case 5 resepctively, and it is concluded that the corresponding CIs are less than 95% CIs. Note that, even though the results for the pseudo extrapolation points and case 5 is better for Model 2 compared to the individual model and Model 1, its coverage probabilities are still large enough.

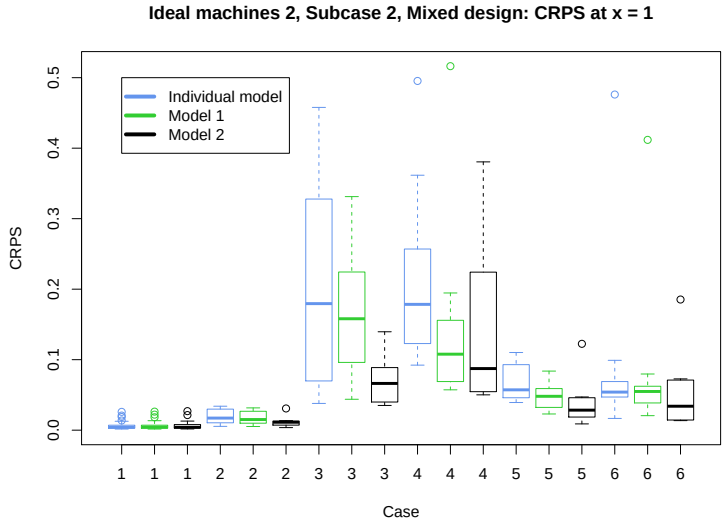


(a)

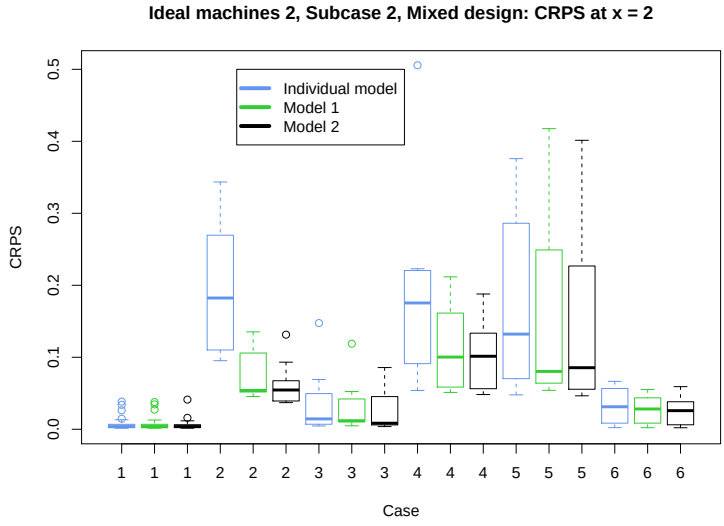


(b)

Figure 5.15: Prediction plots for the Ideal machines 2, subcase 2, Mixed design at $x = 1$ (a) and $x = 2$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

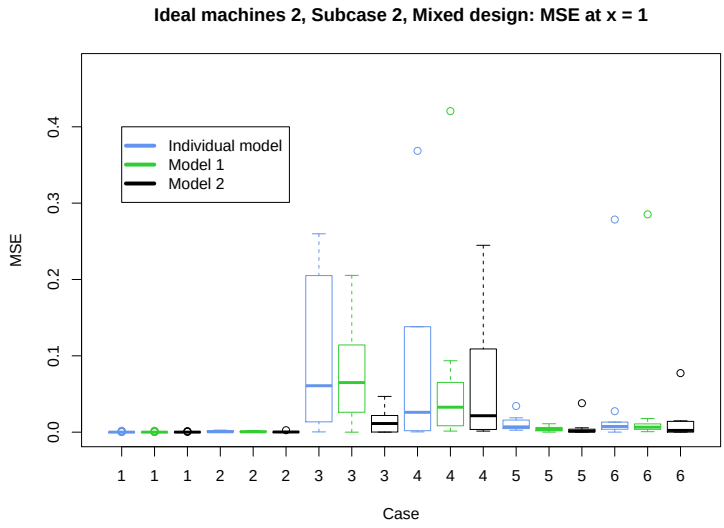


(a)

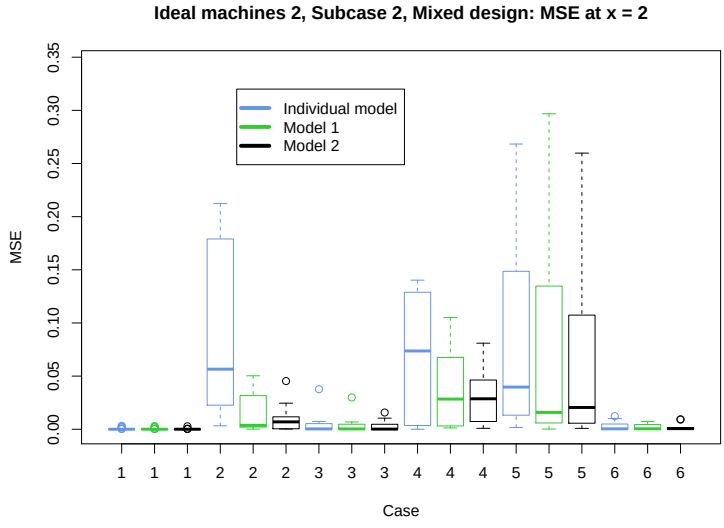


(b)

Figure 5.16: CRPS boxplots for the Ideal machines 2, subcase 2, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

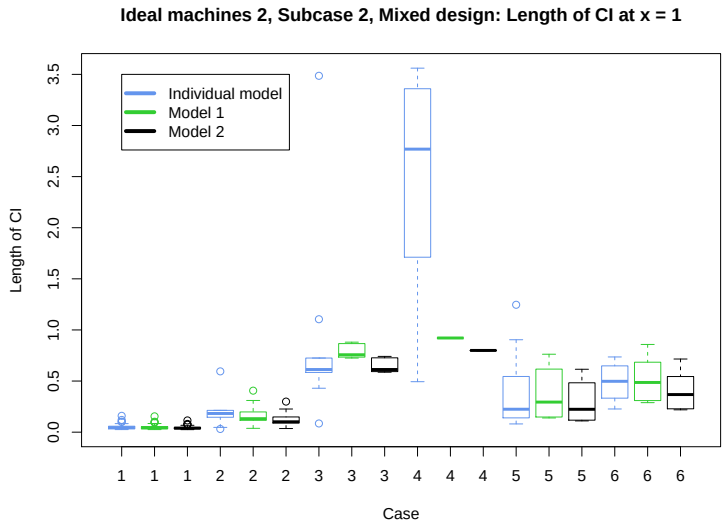


(a)

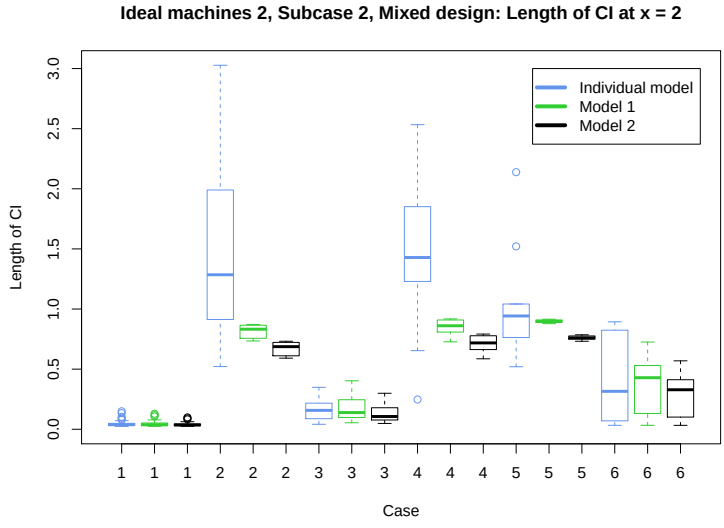


(b)

Figure 5.17: MSE boxplots for the Ideal machines 2, subcase 2, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.18: CI boxplots for the Ideal machines 2, subcase 2, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

Case	Individual model		Model 1		Model 2	
	$x = 1$	$x = 2$	$x = 1$	$x = 2$	$x = 1$	$x = 2$
1	98 (28)	100 (8)	98 (28)	100 (8)	96.5 ($<< 5$)	98.5 ($<< 5$)
2	90 (40)	**90 (40)**	90 (40)	**100 (60)**	96.9 (0.2)	**95.4 (31.2)**
3	**70 (1)**	80 (9)	**80 (9)**	100 (60)	**94.6 (30)**	94.2 (13.9)
4	**100 (60)**	**100 (60)**	**90 (40)**	**100 (60)**	**95.3 (36.6)**	**94.5 (25.3)**
5	**80 (9)**	**80 (9)**	**70 (1)**	**90 (40)**	**95.5 (26.1)**	**93.8 (5.1)**
6	90 (40)	100 (60)	90 (40)	100 (60)	96.7 (0.6)	96.5 (1.4)

Table 5.6: Coverage table for the Ideal machines 2, subcase 2, Mixed design. The values show the coverage probabilities at the interpolation points $x = 1$ and $x = 2$, and the corresponding p-values in parenthesis, all given in percentages. Percentages in bold, represent the relevant cases discussed in this subsection. The values are calculated using 1 experiment for the individual model and Model 1 each, and 100 experiments for Model 2. p-values written as ($<< 5$), are considerably smaller than 5%.

5.2.4 Subcase 3 ($c_1 = \sqrt{0.9}, c_2 = \sqrt{0.1}$), Design 2:

The motivation behind using subcase 3, Design 2, is to compare the predictive performance of the three models for interpolation and extrapolation when the common discrepancy term has a considerably larger variance than the individual discrepancy term. With 60 observation points per machine, the performance of the individual model and Model 1 are very similar when performing interpolation, but Model 2 performs better. For extrapolation however, Model 1 and Model 2 perform better than the individual model, with Model 1 performing better than Model 2. These main results are the same that is found for subcase 2, Design 2, but now that the common discrepancy term dominates considerably more, these main results are more extreme.

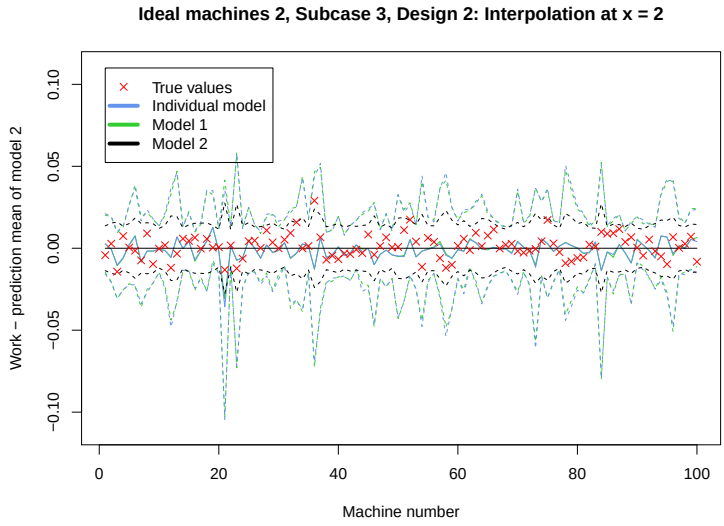
Fig. 5.19 shows the prediction plots at $x = 2$ (interpolation) and $x = 8$ (extrapolation). When interpolating at $x = 2$, there is a difference between the predictions of Model 2 compared to the individual model and Model 1. The individual model and Model 1 seem to have predictions that mostly overlap, while Model 2 have smaller length of CIs. There is a considerable difference between the three models when extrapolating at $x = 8$. The prediction means of Model 1 and Model 2 seem to overlap, but Model 1 has a consderably smaller length of CIs compared to Model 2. For the indvidual model, the length of CIs are more unstable than Model 1's and Model 2's, but the length of CIs seems to be about the same as Model 2's in average.

Fig. 5.20 shows the CRPS boxplots for the interpolation points and extrapolation points. The CRPS distributions for the interpolation points seem very similar for the individual model and Model 1, but better for Model 2 (CRPS values and spreads are smaller for Model 2). Model 1 have the best CRPS values for the extrapolation points, as the values are smaller than the individual model's and Model 2's. Model 2 has a smaller spread than the individual model, but the medians seems to be approximately the same.

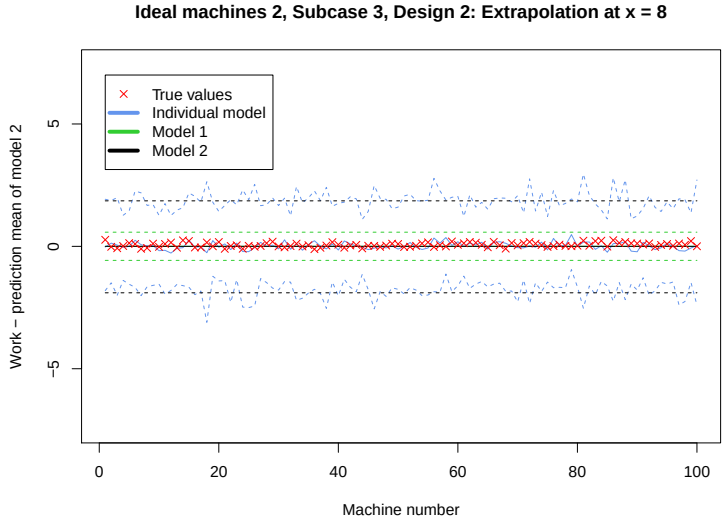
Fig. 5.21 shows the MSE boxplots for the interpolation and extrapolation points. Although it is not obvious from the prediction plots, the MSE for the extrapolation points are smaller and less spread for Model 1 and Model 2 compared to the individual model. Model 2 has better MSE values and spreads for the interpolation points compared to the individual model and Model 1.

Fig. 5.22 shows the CI boxplots for the interpolation and extrapolation points. There is no considerable difference between the length of CIs for the individual model and Model 1 at the interpolation points. However, Model 2 has smaller values and a smaller spread than the individual model and Model 1 have. When considering the extrapolation points, the length of CI distributions have a very small spread for Model 1 and Model 2 compared to the individual model, with Model 1 performing better than Model 2, as its length of CI values are smaller than Model 2's.

Table. 5.7 shows the coverage table. The p-values for all models are significant, and it is concluded that their CIs are greater than 95%. Thus, although Model 1 has better predictive performance for extrapolation and Model 2 has better predictive performance for interpolation, their coverage probabilities are still large enough.

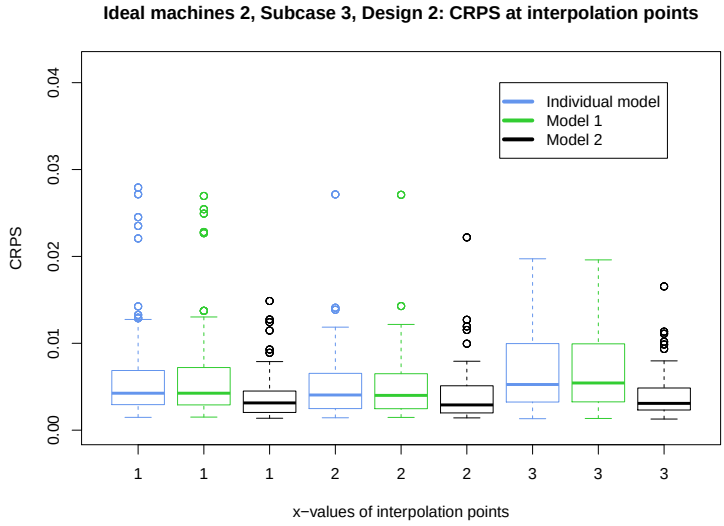


(a)

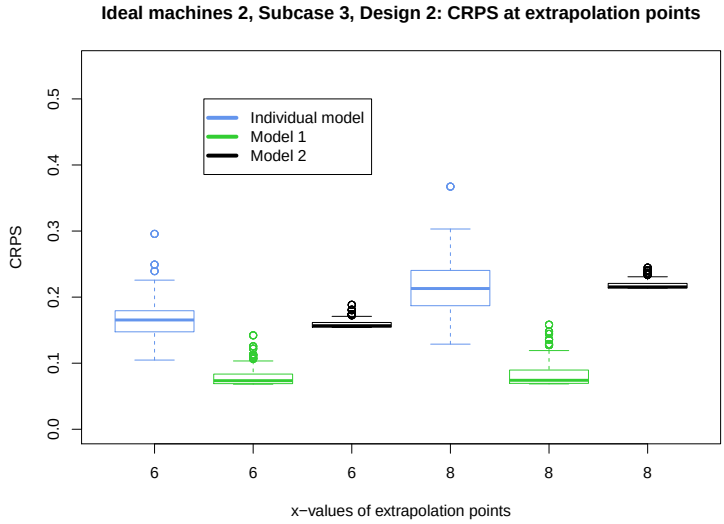


(b)

Figure 5.19: Prediction plots for the Ideal machines 2, subcase 3, Design 2 at $x = 2$ (a) and $x = 8$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

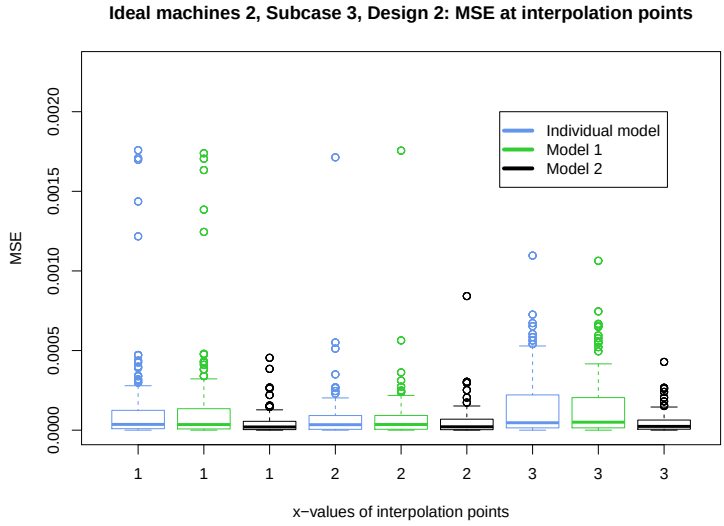


(a)

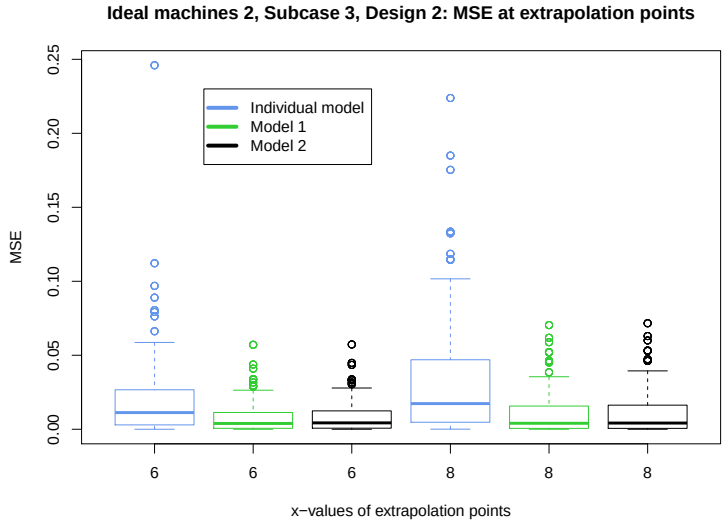


(b)

Figure 5.20: CRPS boxplots for the Ideal machines 2, subcase 3, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

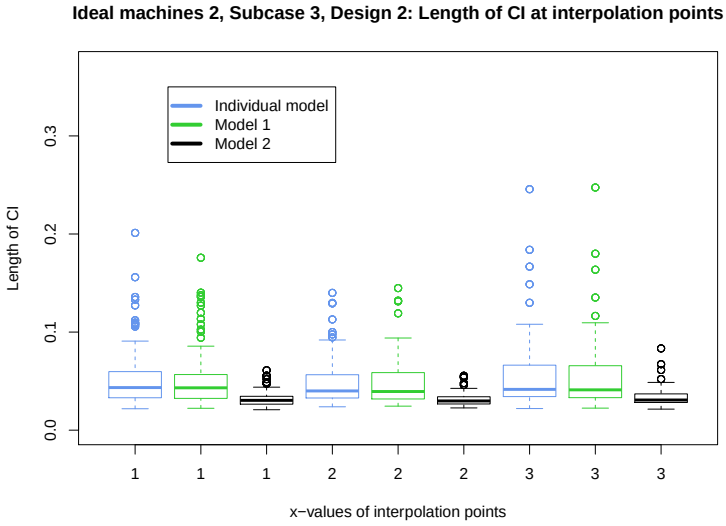


(a)

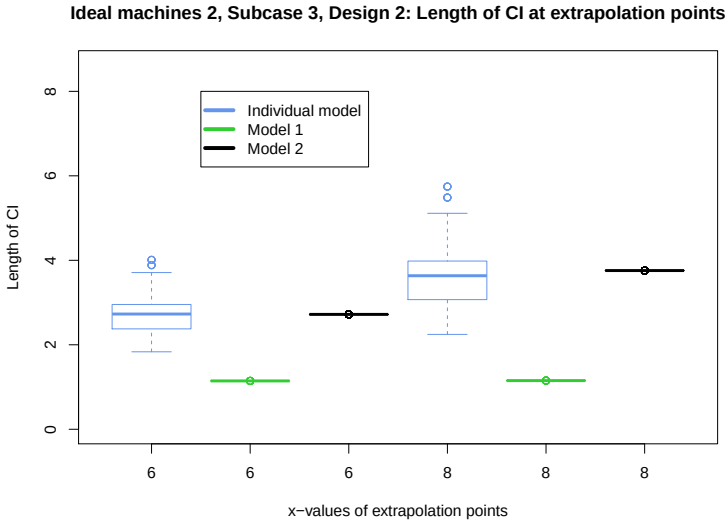


(b)

Figure 5.21: MSE boxplots for the Ideal machines 2, subcase 3, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.22: CI boxplots for the Ideal machines 2, subcase 3, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	99	100	98	100	100	99.4	$2.5 \cdot 10^{-6}$
Model 1	98	100	98	100	100	99.2	$1.7 \cdot 10^{-5}$
Model 2	99	97	99	100	100	99	$9.2 \cdot 10^{-5}$

Table 5.7: Coverage table for the Ideal machines 2, subcase 3, Design 2. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. The values are calculated using 1 experiment for each model.

5.2.5 Subcase 3 ($c_1 = \sqrt{0.9}, c_2 = \sqrt{0.1}$), Mixed design:

The motivation behind using subcase 3, Mixed design, is to compare the predictive performance of the three models at pseudo extrapolation points, and at interpolation points that are far from the observations at at least one side, when the common discrepancy term has a considerably larger variance than the individual discrepancy term. For both types of points, Model 2 have the best performance, followed by Model 1 and then the individual model. This main result is the same that is found for subcase 2, but now that the common discrepancy term dominates considerably more, this main result is more extreme. The interpolation point considered in this subsection is $x = 2$. $x = 2$ is a pseudo extrapolation point for case 2 and 4, while for case 5, the observations are distant on both sides. These are the only cases considered in this subsection.

Fig. 5.23 shows the prediction plot at the interpolation point $x = 2$. There are considerable differences between the cases of the Mixed design for all models. It is very clearly from this plot, that Model 2 has the smallest length of CIs, followed by Model 1 and then the individual model for the relevant cases. To look further into the details of each case, a closer look at the CRPS, MSE, CI boxplots and the coverage table is needed.

Fig. 5.24 shows the CRPS boxplot for the interpolation point $x = 2$. In the cases of interest (case 2,4 and 5), the spread of CRPS and CRPS values are the smallest for Model 2, followed by Model 1 and then the individual model.

Fig. 5.25 shows the MSE boxplot for the interpolation point $x = 2$. In the cases of interest (case 2,4 and 5), the spread of MSE and MSE values are the smallest for Model 2, followed by Model 1 and then the individual model.

Fig. 5.26 shows the CI boxplot for the interpolation point $x = 2$. In the cases of interest (case 2,4 and 5), the spread of length of CIs and length of CI values are the smallest for Model 2, followed by Model 1 and then the individual model.

Table. 5.8 shows the coverage table. Note that, the relevant p-values are all above 5%, and thus, it is concluded that the CIs for the relevant cases are 95% CIs. Hence, even though Model 2 has better predictive performance for pseudo extrapolation points and prediction points that are located far from the observations compared to the individual model and Model 1, its coverage probabilities are still close enough to 95%.

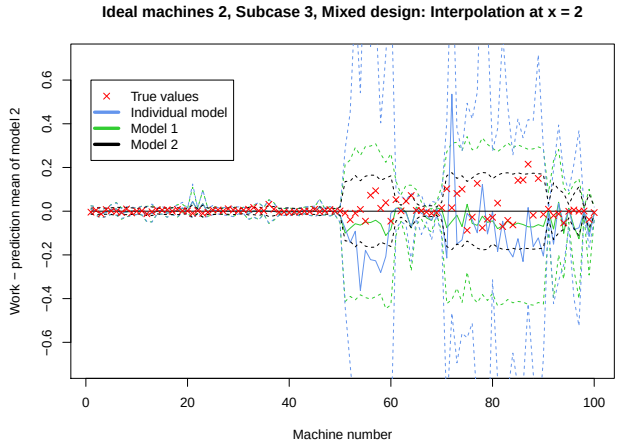


Figure 5.23: Prediction plot for the Ideal machines 2, subcase 3, Mixed design at $x = 2$. Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

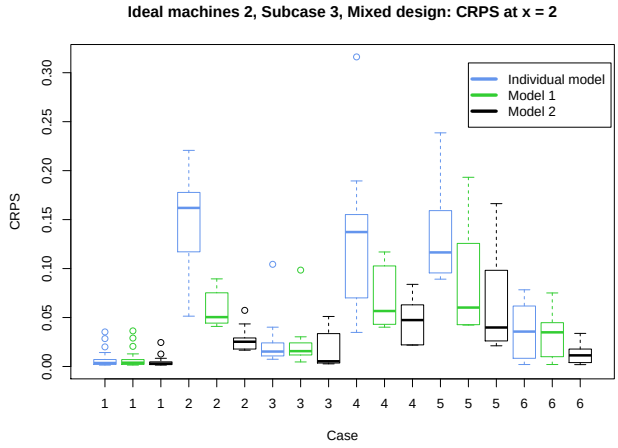


Figure 5.24: CRPS boxplot for the Ideal machines 2, subcase 3, Mixed design at the interpolation point $x = 2$. Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

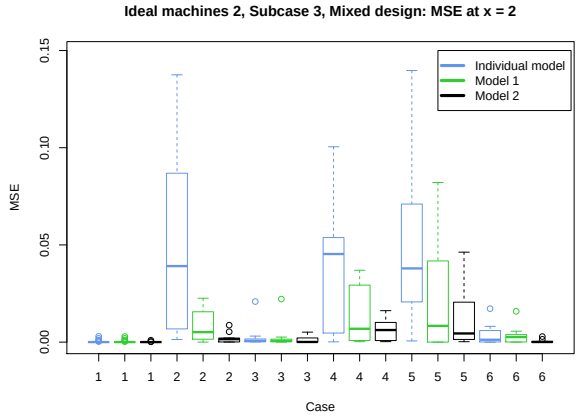


Figure 5.25: MSE boxplot for the Ideal machines 2, subcase 3, Mixed design at the interpolation point $x = 2$. Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

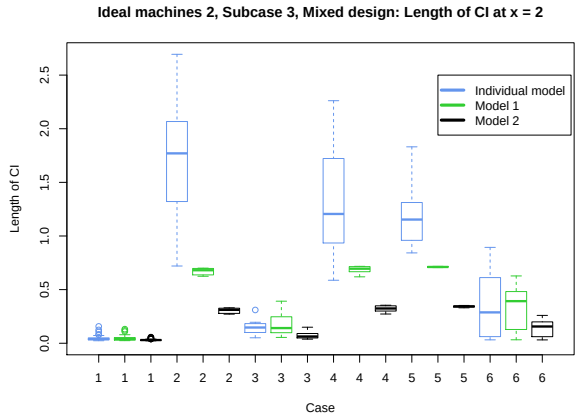


Figure 5.26: CI boxplot for the Ideal machines 2, subcase 3, Mixed design at the interpolation point $x = 2$. Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

$x = 2$	**Individual model**	**Model 1**	**Model 2**
Case 1	100 (8)	100 (8)	96 (54)
Case 2	**100 (60)**	**100 (60)**	**100 (60)**
Case 3	100 (60)	100 (60)	80 (9)
Case 4	**100 (60)**	**100 (60)**	**100 (60)**
Case 5	**100 (60)**	**100 (60)**	**90 (40)**
Case 6	90 (40)	100 (60)	100 (60)

Table 5.8: Coverage table for the Ideal machines 2, subcase 3, Mixed design. The values show the coverage probabilities at the interpolation point $x = 2$, and the corresponding p-values in parenthesis, all given in percentages. Percentages in bold, represent the relevant cases discussed in this subsection. The values are calculated using 1 experiment for each model.

5.2.6 Subcase 1 ($c_1 = \sqrt{0.1}, c_2 = \sqrt{0.9}$), Design 2:

The motivation behind using subcase 1, Design 2, is to compare the predictive performance of the three models for interpolation and extrapolation when the individual discrepancy term has a considerably larger variance than the common discrepancy term. With 60 observation points per machine, the performance of the three models are very similar, but Model 2 seem to possibly perform better. For extrapolation however, Model 1 and Model 2 performs better than the individual model, and Model 1 seems to possibly perform slightly better than Model 2. It is difficult to tell, as the differences in the boxplots are barely visible. However, if this is true, then the main conclusion of this subsection is something between the main conclusion from the Ideal machines 1 and the Ideal machines 2 subcase 2. That is, Model 1 and Model 2 performing very similarly (similar results to the Ideal machines 1), but slightly better for Model 1 and Model 2 for extrapolation and interpolation respectively (a less extreme version of the main results of the Ideal machines 2 subcase 2).

Fig. 5.27 shows the prediction plots at $x = 2$ (interpolation) and $x = 8$ (extrapolation). It is difficult to see the difference between the three models when interpolating at $x = 2$ in this plot. When extrapolating at $x = 8$, the prediction means of Model 1 and Model 2 seem to overlap, but Model 1 has a slightly smaller length of CIs compared to Model 2. For the indvidual model, the length of CIs seems to be larger than Model 1's and Model 2's.

Fig. 5.28 shows the CRPS boxplots for the interpolation points and extrapolation points. The CRPS distributions for the interpolation points seem very similar for the three models, but possibly better for Model 2 (better largest values and 75th percentile for $x = 1$ and $x = 2$, and smaller median for $x = 3$). Model 1 and Model 2 have a smaller spread and CRPS values than the individual model, but it is difficult to tell which of the two has the better performance.

Fig. 5.29 shows the MSE boxplots for the interpolation and extrapolation points. The MSE distributions for the interpolation points seem very similar for the three models, but possibly better for Model 2 (75th percentile and largest values are in general smaller for Model 2). For extrapolation, Model 1 and Model 2 have a smaller spread and MSE values than the individual model, but it is difficult to tell which of the two has the better performance.

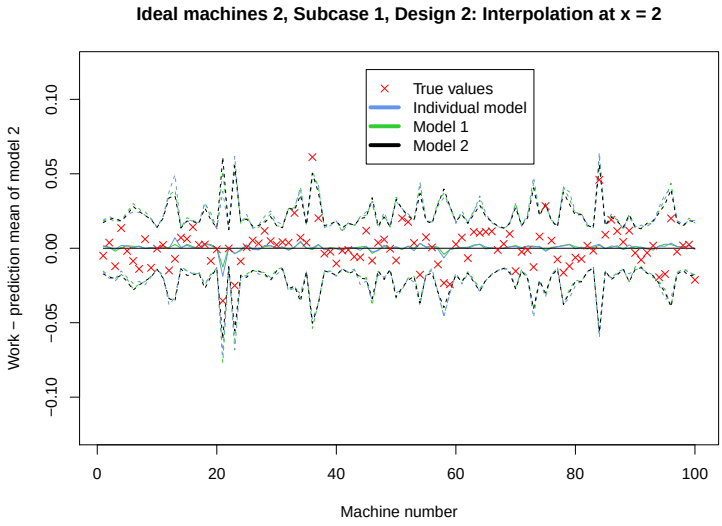
Fig. 5.30 shows the CI boxplots for the interpolation and extrapolation points. The length of CI distributions for the interpolation points seem very similar for the three models, but possibly better for Model 2 (75th percentile and largest values are in general smaller for Model 2). When considering the extrapolation points, the length of CI distributions have a very small spread for Model 1 and Model 2 compared to the individual model, with Model 1 performing slightly better than Model 2, as its length of CIs are smaller than Model 2's.

Table. 5.9 shows the coverage table. The p-value for model 2 is significant, and it is concluded that its CIs are greater than 95%. For the individual model and Model 1, the credible intervals are concluded to be 95%. Thus, although Model 2 and Model 1 have possibly better performance for interpolation points and extrapolation points respectively,

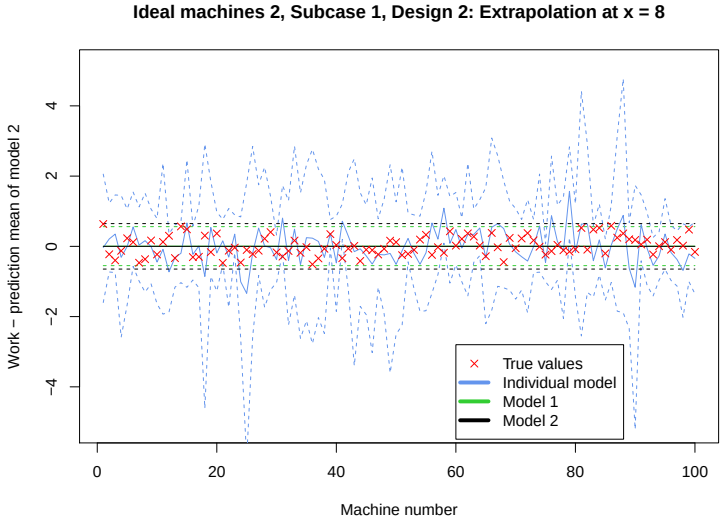
their CIs are still large enough.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	94	98	95	95	96	95.6	31.2
Model 1	94	98	95	96	97	96	17.9
Model 2	95	98	96	98	100	97.4	0.6

Table 5.9: Coverage table for the Ideal machines 2, subcase 1, Design 2. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. The values are calculated using 1 experiment for each model.

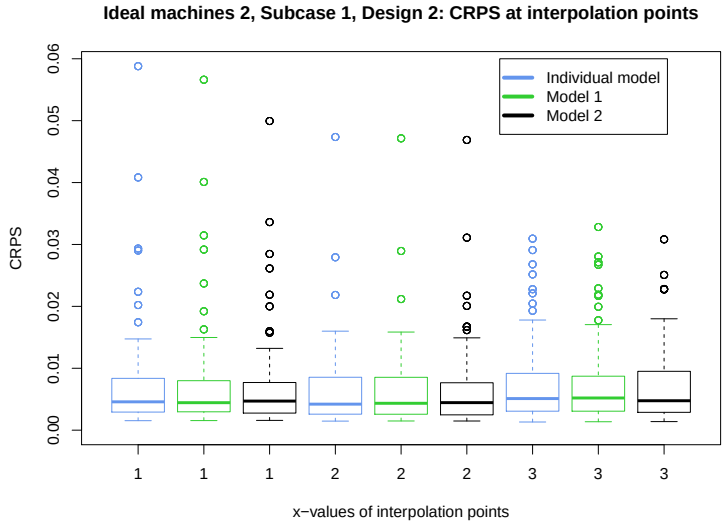


(a)

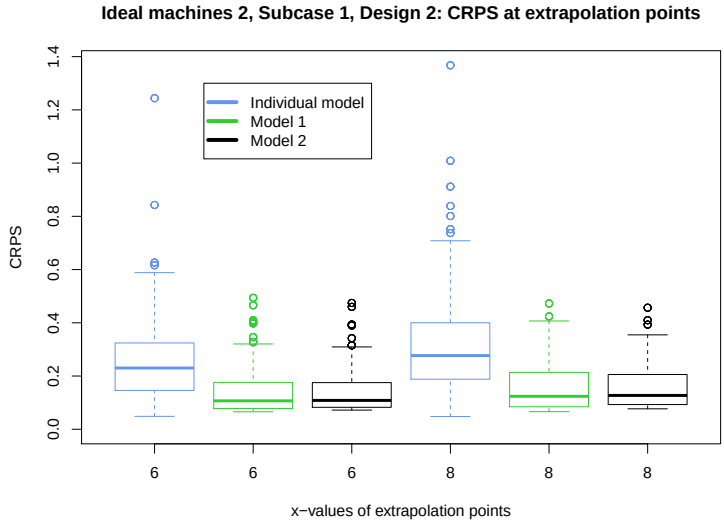


(b)

Figure 5.27: Prediction plots for the Ideal machines 2, subcase 1, Design 2 at $x = 2$ (a) and $x = 8$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

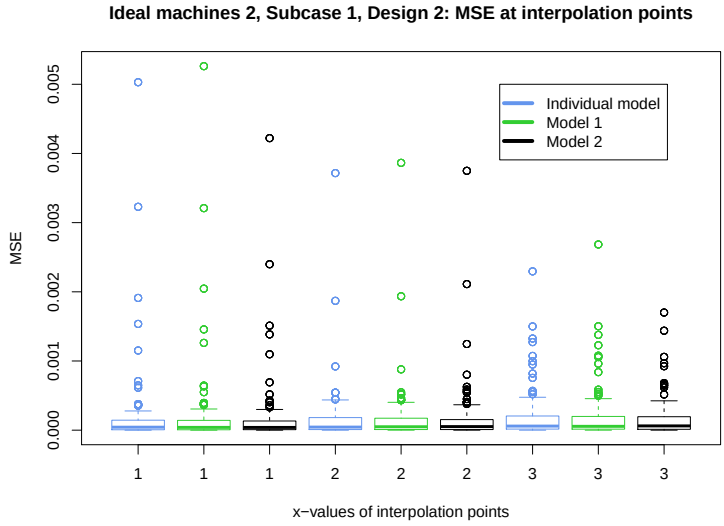


(a)

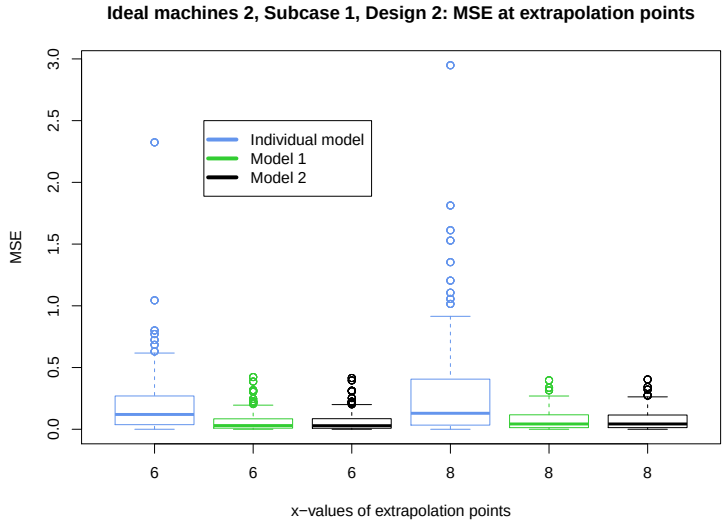


(b)

Figure 5.28: CRPS boxplots for the Ideal machines 2, subcase 1, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

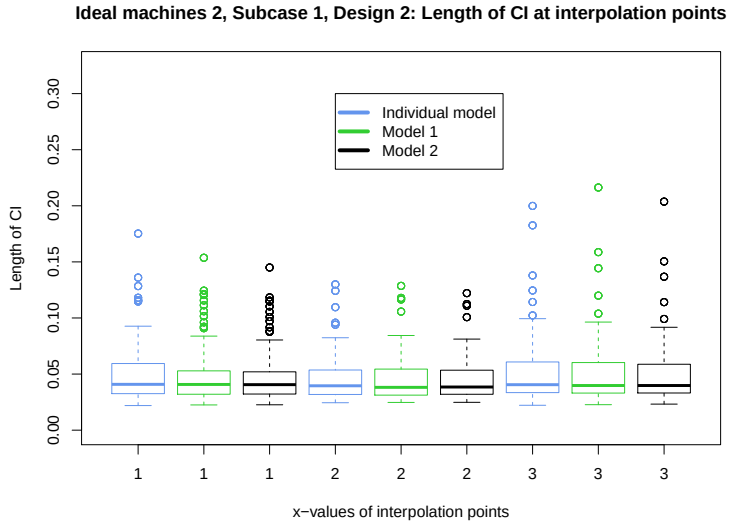


(a)

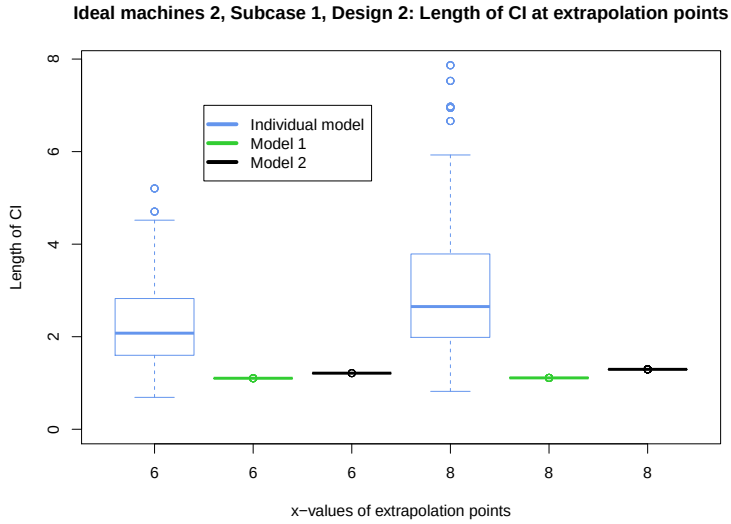


(b)

Figure 5.29: MSE boxplots for the Ideal machines 2, subcase 1, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.30: CI boxplots for the Ideal machines 2, subcase 1, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

5.2.7 Subcase 1 ($c_1 = \sqrt{0.1}, c_2 = \sqrt{0.9}$), Mixed design:

The motivation behind using subcase 1, Mixed design, is to compare the predictive performance of the three models at pseudo extrapolation points, and at interpolation points that are far from the observations at at least one side, when the individual discrepancy term has a considerably larger variance than the common discrepancy term. For both types of points, Model 1 and Model 2 have better performance compared to the individual model, but it is difficult to tell which of the 2 that has the better performance. The interpolation point considered in this subsection is $x = 2$. $x = 2$ is a pseudo extrapolation point for case 2 and 4, while for case 5, the observations are distant on both sides. These are the only cases considered in this subsection.

Fig. 5.31 shows the prediction plot at the interpolation point $x = 2$. There are considerable differences between the cases of the Mixed design. To look further into the details of each case, a closer look at the CRPS, MSE, CI boxplots and the coverage table is needed.

Fig. 5.32 shows the CRPS boxplot for the interpolation point $x = 2$. In the cases of interest (case 2,4 and 5), the spread and CRPS values are smaller for Model 1 and Model 2 compared to the individual model, but it is difficult to tell which of the two that has the better performance.

Fig. 5.33 shows the MSE boxplot for the interpolation point $x = 2$. In the cases of interest (case 2,4 and 5), the spread and MSE values are smaller for Model 1 and Model 2 compared to the individual model, but it is difficult to tell which of the two that has the better performance.

Fig. 5.34 shows the CI boxplot for the interpolation point $x = 2$. In the cases of interest (case 2,4 and 5), the spread of length of CIs and length of CI values are smaller for Model 1 and Model 2 compared to the individual model, with Model 2 being slightly better.

Table. 5.10 shows the coverage table. Note that, the relevant p-values are all above 5%, and thus, it is concluded that the CIs for the relevant cases are 95% CIs. Thus, although Model 1 and Model 2 have smaller lengths of CI than the individual model has, their CIs are still large enough.

$x = 2$	**Individual model**	**Model 1**	**Model 2**
Case 1	92 (24)	96 (54)	96 (54)
Case 2	**90 (40)**	**90 (40)**	**90 (40)**
Case 3	80 (9)	100 (60)	90 (40)
Case 4	**80 (9)**	**100 (60)**	**100 (60)**
Case 5	**90 (40)**	**100 (60)**	**100 (60)**
Case 6	100 (60)	100 (60)	100 (60)

Table 5.10: Coverage table for the Ideal machines 2, subcase 1, Mixed design. The values show the coverage probabilities at the interpolation point $x = 2$, and the corresponding p-values in parenthesis, all given in percentages. Percentages in bold, represent the relevant cases discussed in this subsection. The values are calculated using 1 experiment for each model.

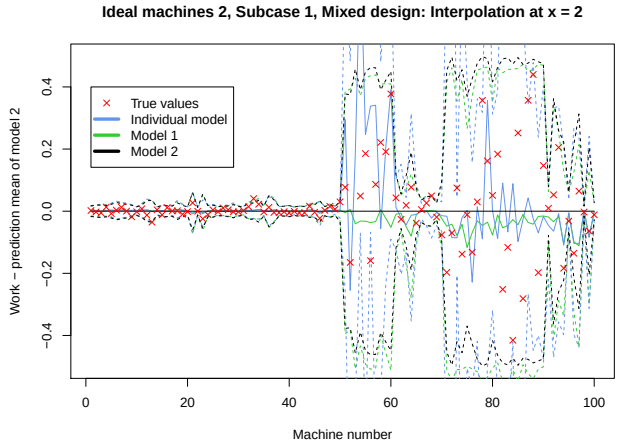


Figure 5.31: Prediction plot for the Ideal machines 2, subcase 1, Mixed design at $x = 2$. Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

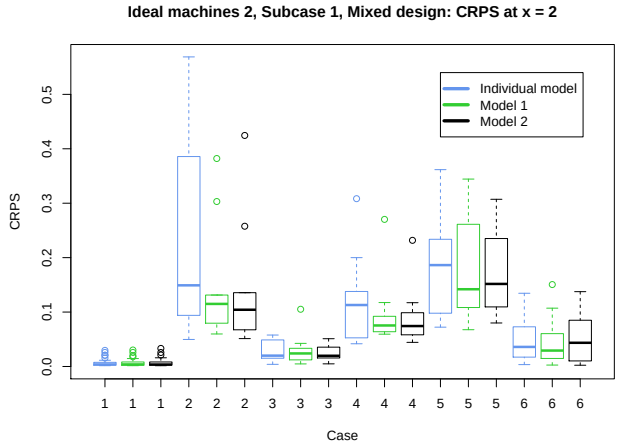


Figure 5.32: CRPS boxplot for the Ideal machines 2, subcase 1, Mixed design at the interpolation point $x = 2$. Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

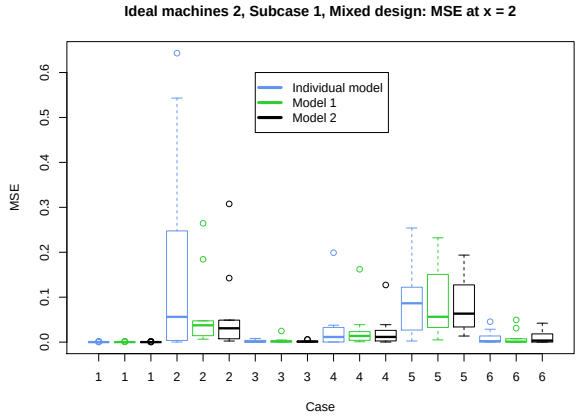


Figure 5.33: MSE boxplot for the Ideal machines 2, subcase 1, Mixed design at the interpolation point $x = 2$. Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

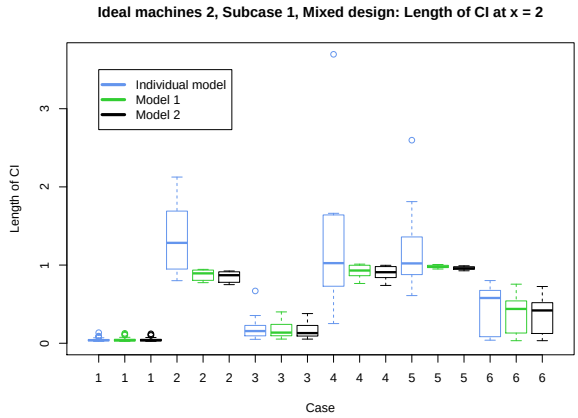


Figure 5.34: CI boxplot for the Ideal machines 2, subcase 1, Mixed design at the interpolation point $x = 2$. Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

5.2.8 Summary, Ideal machines 2:

The Ideal machines 2 are machines that have common parameters, an individual discrepancy term and a common discrepancy term. Their true processes are created using the framework of Model 2. Below follows a summary from the Ideal machines 2.

- The CI of the predictions for the individual model do not cover 95% of the true values when there are few observations per machine for subcase 2.
- When there are many observations per machine, Model 1 and Model 2 performs better for interpolation compared to the individual model, with Model 2 performing the best. It has the best CRPS, MSE and lengths of CI. How much better Model 2 performs compared to Model 1, depends on how the variance of the common discrepancy term dominates the variance of the individual discrepancy term. The more dominating the variance of the common discrepancy term, the larger the difference between the performances of Model 1 and Model 2.
- When there are many observations per machine, Model 1 and Model 2 performs better for extrapolation compared to the individual model, with Model 1 performing the best. It has the best CRPS, MSE and lengths of CI. How much better Model 1 performs compared to Model 2, depends on how the variance of the common discrepancy term dominates the variance of the individual discrepancy term. The more dominant the variance of the common discrepancy term, the larger the difference between the performance of Model 1 and Model 2. This is only in the cases where Model 1 does not have a significant p-value for coverage probability. **Fig. 5.35** shows a coverage plot for the Ideal machines 2, subcase 3, Mixed design. In this plot, Model 1 only has a coverage of 33%. When this is the case, Model 2 is considerably better at extrapolation. Hence, Model 2 is considered as the overall better model for extrapolation.
- Model 1 and Model 2 perform better pseudo extrapolation and interpolation with points that are distant from the observations at at least one side, compared to the individual model. Model 2 performs better than Model 1 in these cases, but it depends on how the variance of the common discrepancy term dominates the variance of the individual discrepancy term. The more dominant the variance of the common discrepancy term, the larger the difference between the performance of Model 1 and Model 2. Model 2 has the best CRPS, MSE and lengths of CI.

For the models considered to have better predictive performance based on the CRPS, MSE or lengths of CI, it is the spread and the values of the distributions of CRPS, MSE or lengths of CI that is considered. For example, if two models have the same median for the distributions of CRPS values, the one with the smaller spread is considered the best. This is because it has more stable results, and thus, it is considered better than a model that has really small CRPS values for some machines and really large CRPS values for other machines. If two models have approximately the same spreads on the distributions of the CRPS, then the one with the smaller values is considered the best. If a model has a larger spread, but smaller median than another model, then none of the two models are considered the better than the other.

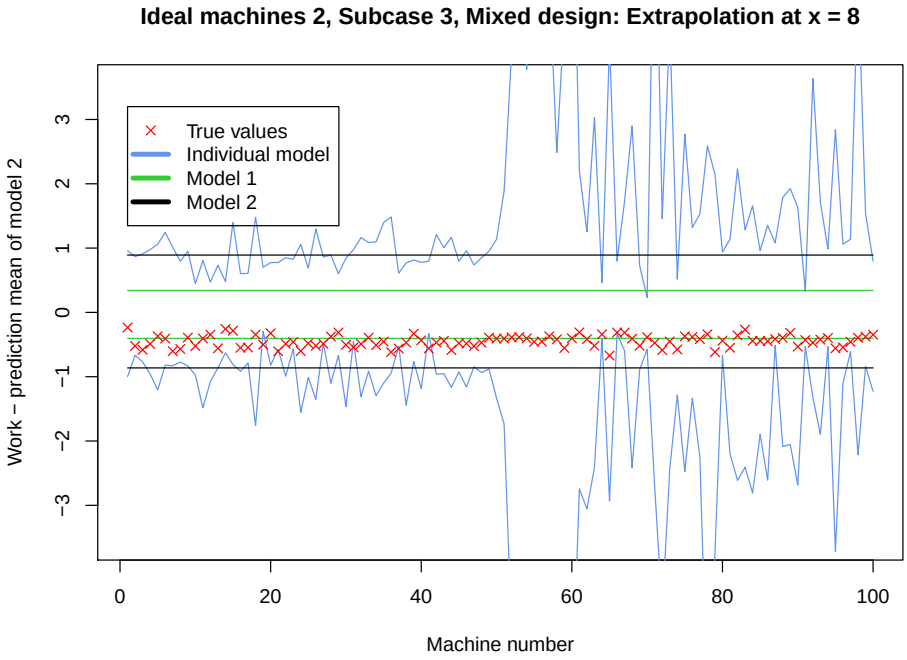


Figure 5.35: Coverage plot for the Ideal machines 2, subcase 3, Mixed design at $x = 8$. The blue, green and black curves represent the 95% credible intervals for the individual model, Model 1 and Model 2 respectively, and the red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

5.3 Multiple Simple Machines

5.3.1 Design 1:

The motivation behind using Design 1, is to check the coverage probability for the three models with few observations per machine. Model 1 seems to have the worst coverage probability, followed by the individidual model and then Model 2. All 3 models have a good coverage probability for the interpolation points, but the CIs do not cover enough true values for the extrapolation points.

Fig. 5.36 shows the coverage plot for extrapolation at $x = 6$. In this plot, Model 2 covers 72% of the points, the individual model covers 43% and Model 1 does not cover any points.

Table. 5.11 shows the coverage table. As can be seen from this table, the coverage probability is the worst for Model 1, as it does not cover any extrapolation points. The individual model performs better, as it manages to cover 43% and 13% for the extrapolation points $x = 6$ and $x = 8$ respectively. Model 2 has the best coverage probability, as it covers 72% and 15% for the extrapolation points $x = 6$ and $x = 8$ respectively. Note that all p-values in the table are significant. It is concluded that the CIs are not 95% CIs for the extrapolation points.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	97	100	95	43	13	69.6	$3.2 \cdot 10^{-72}$
Model 1	99	99	98	0	0	59.2	$2.6 \cdot 10^{-125}$
Model 2	99	99	99	72	15	76.8	$7.2 \cdot 10^{-42}$

Table 5.11: Coverage table for multiple simple machines, Design 1. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. The values are calculated using 1 experiment for each model.

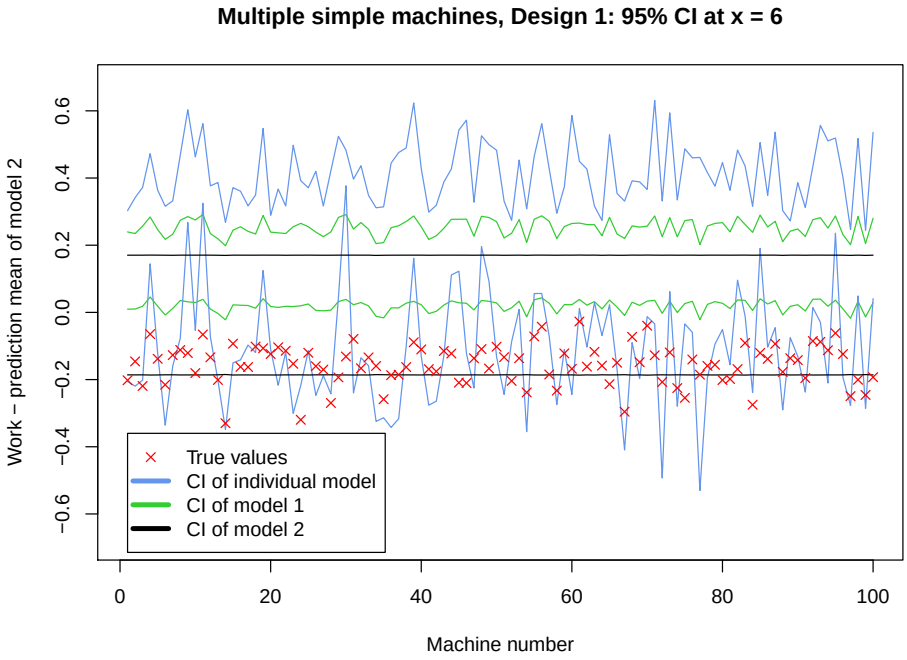


Figure 5.36: Coverage plot for multiple simple machines, Design 1 at $x = 6$. The blue, green and black curves represent the 95% credible intervals for the individual model, Model 1 and Model 2 respectively, and the red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

5.3.2 Design 2:

The motivation behind using Design 2, is to compare the predictive performance of the three models for interpolation and extrapolation. With 60 observation points per machine, there is no considerable difference between the individual model and Model 1 when performing interpolation, but Model 2 performs better. For extrapolation, Model 2 performs the best, followed by the individual model and then Model 1.

Fig. 5.37 shows the prediction plots at $x = 2$ (interpolation) and $x = 6$ (extrapolation). The prediction means and CIs of the individual model and Model 1 seem to overlap when interpolating at $x = 2$. When extrapolating at $x = 6$, the individual model and Model 1 seem to predict larger values than Model 2, and miss the true values. Model 2 also seems to miss the true values, but the prediction mean is closer to the true values compared to the individual model's and model's.

Fig. 5.38 shows the CRPS boxplots for the interpolation points and extrapolation points. For the interpolation points, the CRPS distributions seem similar for the individual model and Model 1, but for Model 2 they are smaller and have a smaller spread. For the extrapolation points, the spreads seems similar for all 3 models, but Model 2 have smaller values, followed by the individual model and then Model 1.

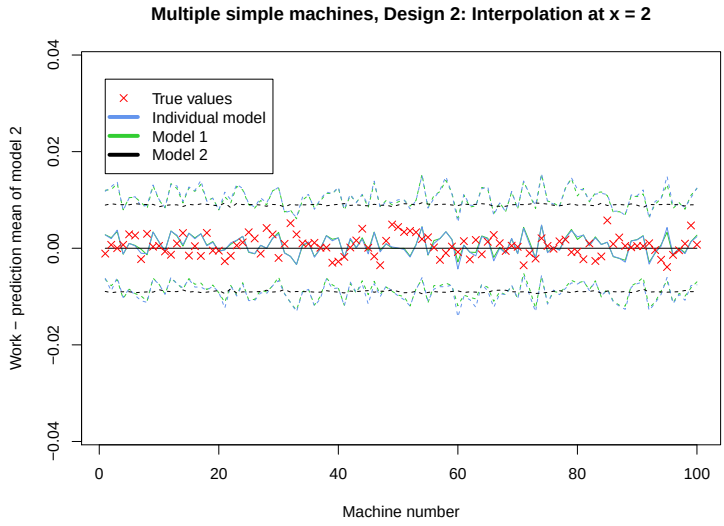
Fig. 5.39 shows the MSE boxplots for the interpolation and extrapolation points. For the interpolation points, the MSE distributions seem similar for the individual model and Model 1, but for Model 2 they are smaller and have a smaller spread. For the extrapolation points, the spreads seems similar for all 3 models, but Model 2 have smaller values, followed by the individual model and then Model 1.

Fig. 5.40 shows the CI boxplots for the interpolation and extrapolation points. For the interpolation points, the MSE values have the smallest spread and values for Model 2, followed by Model 1 and then the individual model. For the extrapolation points, the spread and values are smaller for Model 1 and Model 2 compared to the individual model, with the values of Model 1 being the smallest.

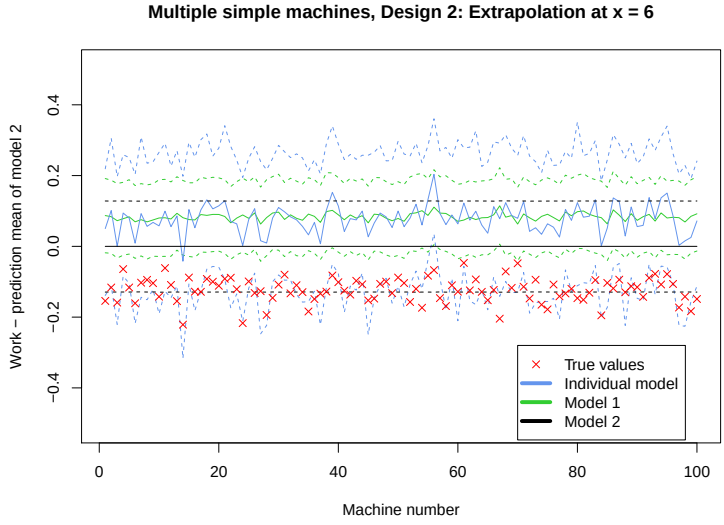
Table. 5.12 shows the coverage table. All the p-values are significant, and it is concluded that the CIs for the three models are smaller than 95% CIs for extrapolation. Note that Model 1 fails to cover any true values, and thus, even though Model 1's length of CIs are the smallest, its performance is still considered worse than the individual model and Model 2 for extrapolation.

	x = 1	**x = 2**	**x = 3**	**x = 6**	**x = 8**	**Total**	**p-value**
Individual model	100	100	99	51	3	70.6	$1.1 \cdot 10^{-67}$
Model 1	100	100	99	0	0	59.8	$5.6 \cdot 10^{-122}$
Model 2	100	100	100	60	7	73.4	$1.7 \cdot 10^{-55}$

Table 5.12: Coverage table for multiple simple machines, Design 2. The values show the coverage probabilities at the prediction points, the total coverage probabilites calculated from the 5 prediction points, and the p-values of the total coverage probabilites, all given in percentages. The values are calculated using 1 experiment for each model.

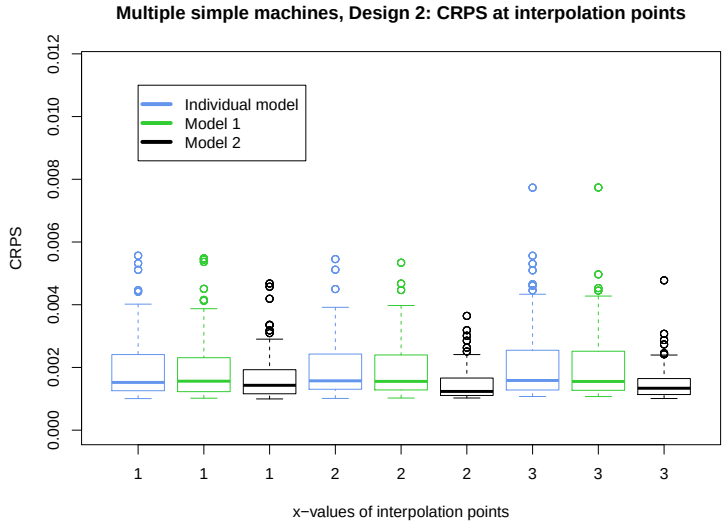


(a)

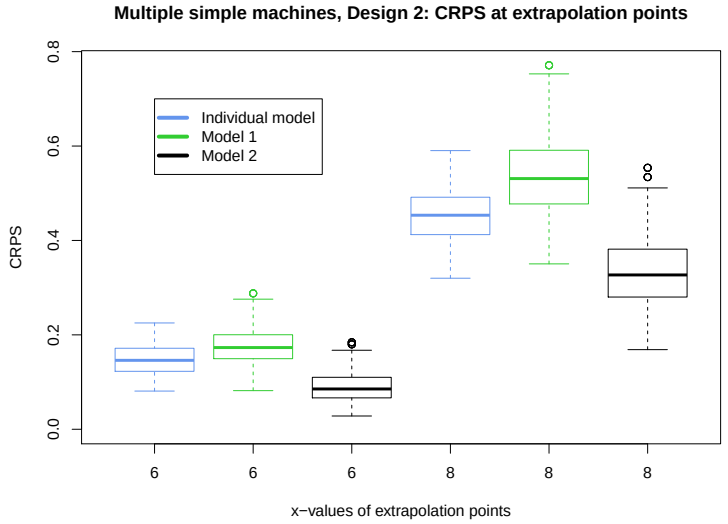


(b)

Figure 5.37: Prediction plots for multiple simple machines, Design 2 at $x = 2$ (a) and $x = 6$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

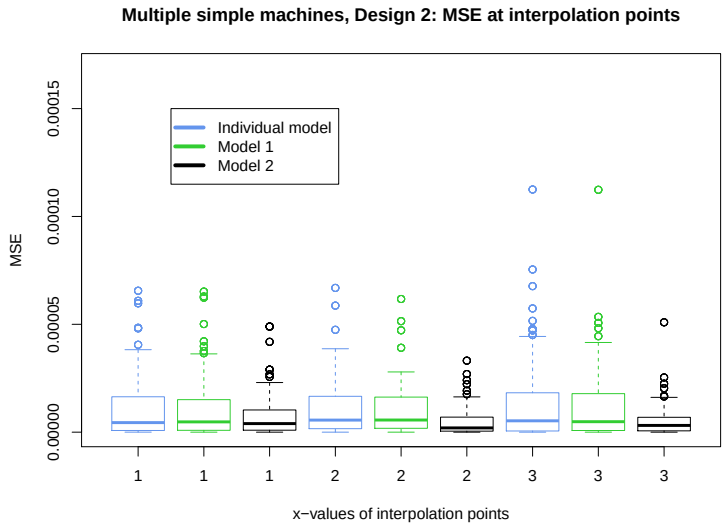


(a)

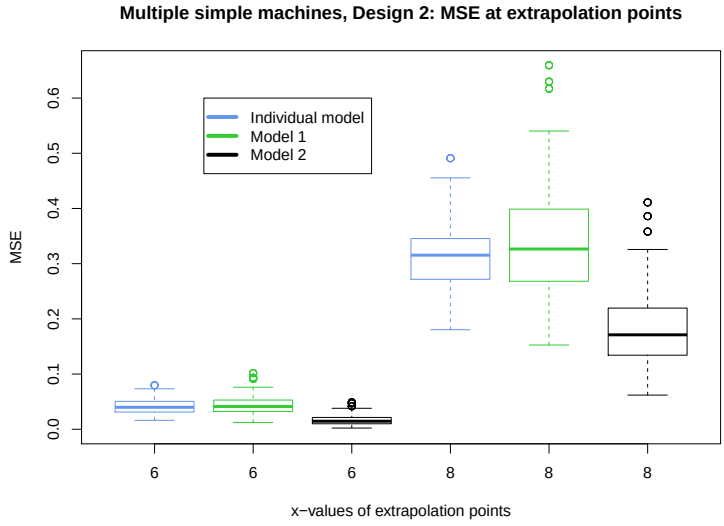


(b)

Figure 5.38: CRPS boxplots for multiple simple machines, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

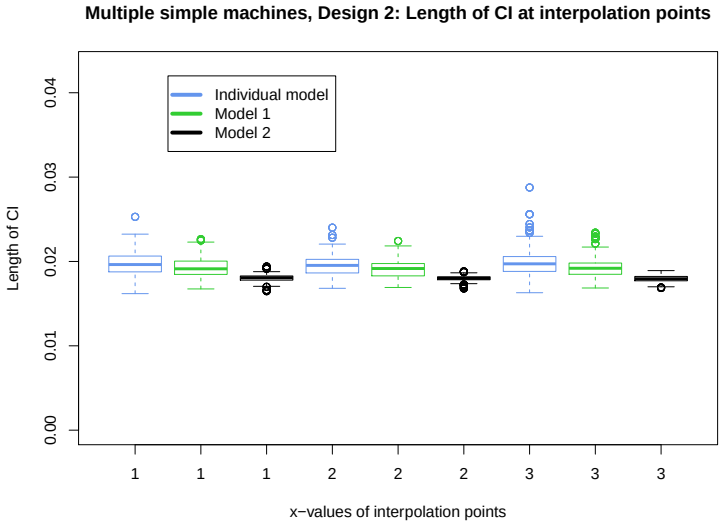


(a)

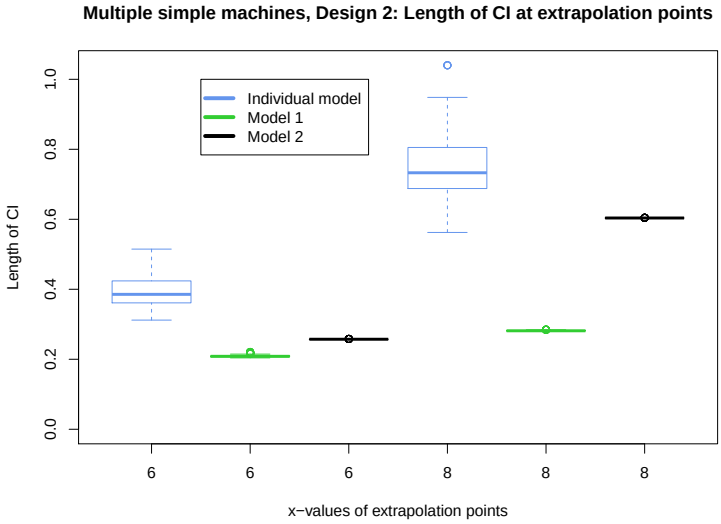


(b)

Figure 5.39: MSE boxplots for the multiple simple machines, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.40: CI boxplots for multiple simple machines, Design 2 at the interpolation points (a) and the extrapolation points (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the prediction points, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

5.3.3 Mixed design:

The motivation behind using Mixed design, is to compare the predictive performance of the three models at pseudo extrapolation points, and at interpolation points that are far from the observations at at least one side. For both types of points, Model 2 performs the best, followed by Model 1 and then the individual model. The interpolation points considered in this subsection, are $x = 1$ and $x = 2$. $x = 1$ is a pseudo extrapolation point for case 3 and 4, while for case 5, the observations are close on one side and distant on the other side. $x = 2$ is a pseudo extrapolation point for case 2 and 4, while for case 5, the observations are distant on both sides. These are the only cases considered in this subsection.

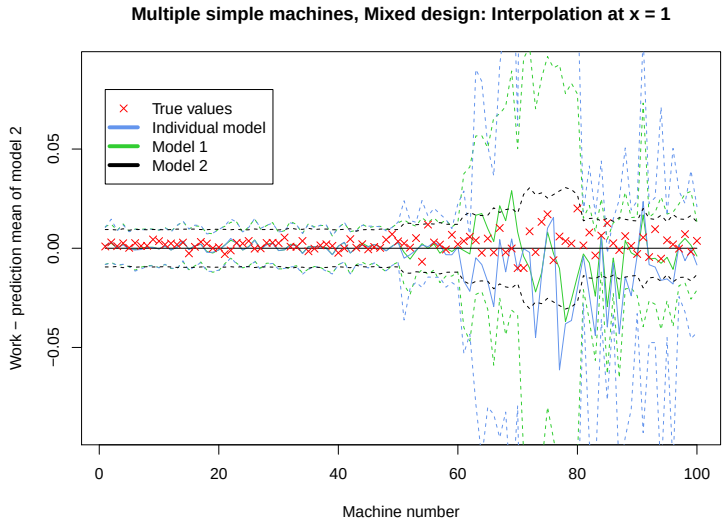
Fig. 5.41 shows the prediction plots at the interpolation points $x = 1$ and $x = 2$. There are considerable differences between the cases of the Mixed design for both the interpolation points. To look further into the details of each case, a closer look at the CRPS, MSE, CI boxplots and the coverage table is needed.

Fig. 5.42 shows the CRPS boxplots for the interpolation points $x = 1$ and $x = 2$. In general, the spread of CRPS and the CRPS values are the smallest for Model 2, followed by Model 1 and then the individual model when the prediction point is a pseudo extrapolation point and for case 5. This can be seen for $x = 1$ case 3, 4 and 5, and $x = 2$ case 4 and 5.

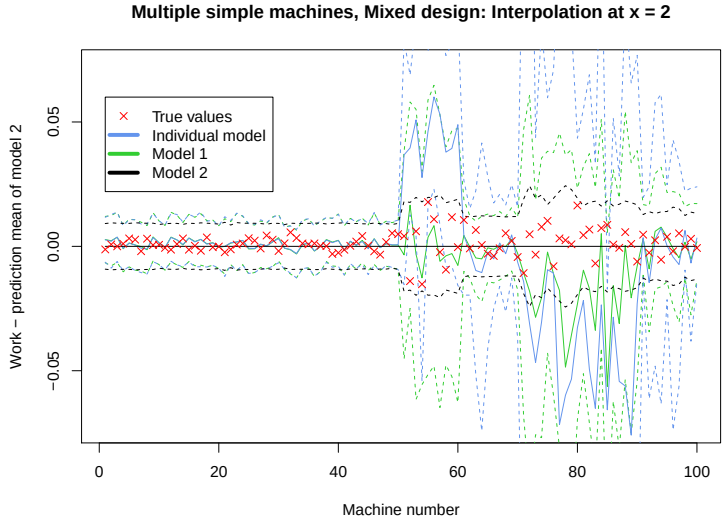
Fig. 5.43 shows the MSE boxplots for the interpolation points $x = 1$ and $x = 2$. In general, the spread of MSE and the MSE values are the smallest for Model 2, followed by Model 1 and then the individual model when the prediction point is a pseudo extrapolation point and for case 5. This can be seen for $x = 1$ case 3, 4 and 5, and $x = 2$ case 4 and 5.

Fig. 5.44 shows the CI boxplots for the interpolation points $x = 1$ and $x = 2$. In general, the spread of the length of CIs and the length of CIs values are the smallest for Model 2, followed by Model 1 and then the individual model when the prediction point is a pseudo extrapolation point and for case 5. This can be seen for $x = 1$ case 3, 4 and 5, and $x = 2$ case 4 and 5.

Table. 5.13 shows the coverage table. The p-value for the coverage probability of the individual model at $x = 2$ case 2 is significant, and it is concluded that its CIs are smaller than 95% CIs for this point and this case. All the other p-values are not significant, and thus, corresponding CIs are concluded to be 95% CIs. Although Model 1 and Model 2 have better predictive performance than the individual model for pseudo extrapolation points and interpolation points that are far from the observations, their coverage probabilities are still large enough.

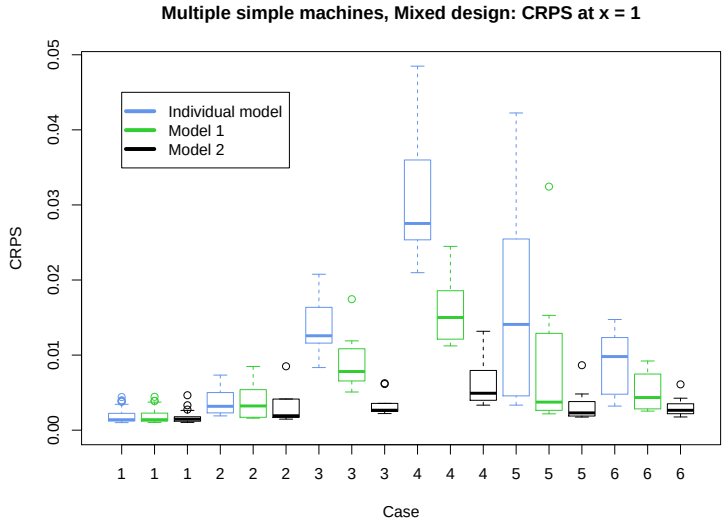


(a)

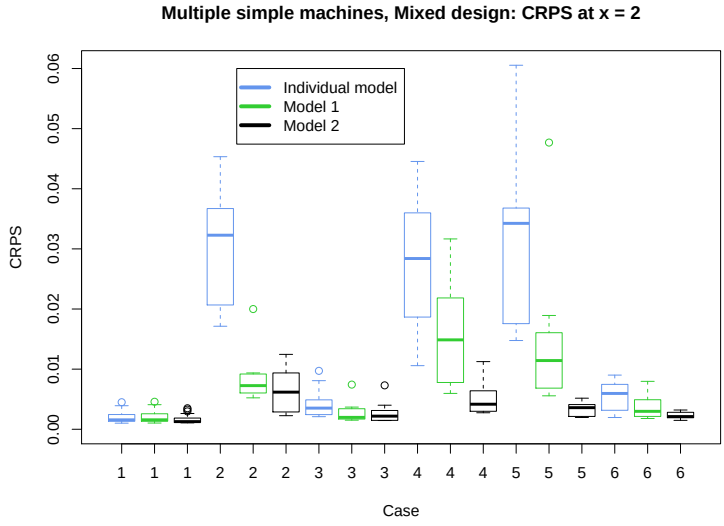


(b)

Figure 5.41: Prediction plots for multiple simple machines, Mixed design at $x = 1$ (a) and $x = 2$ (b). Blue, green and black curves represent the individual model, Model 1 and Model 2 respectively, with the solid curves being the prediction means and the dashed curves being the 95% credible intervals. The red points represent the true values. The x-axis represents machine number, and the y-axis represents the work minus the prediction mean of Model 2. Hence, $x = 0$ is the prediction mean of Model 2.

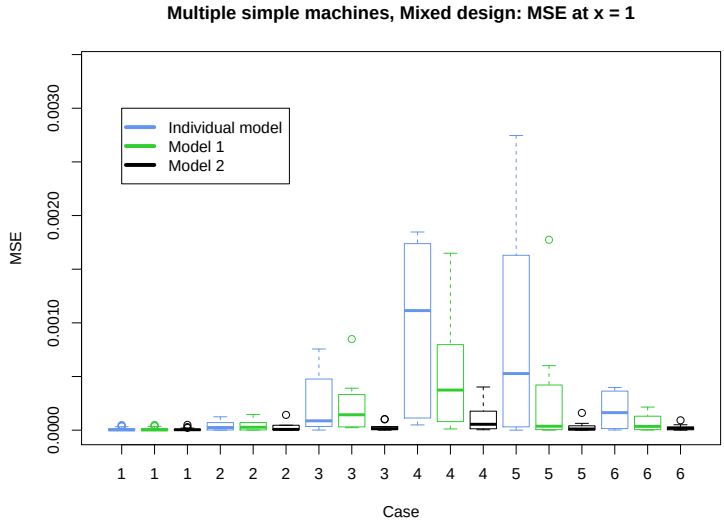


(a)

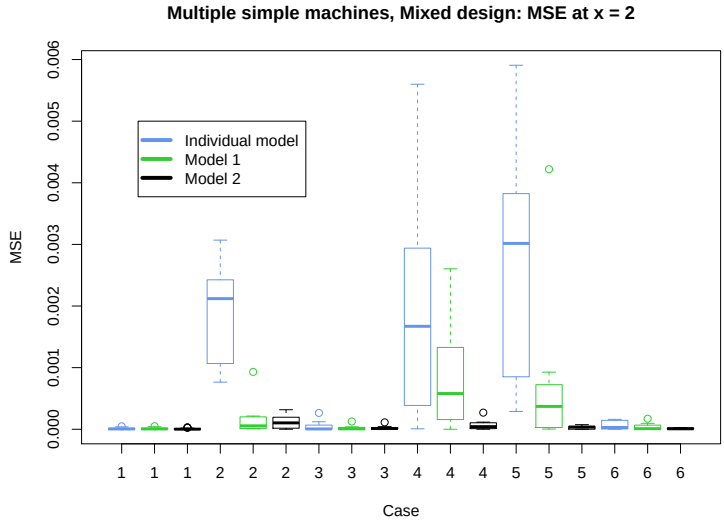


(b)

Figure 5.42: CRPS boxplots for multiple simple machines, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the CRPS. Each box represents the interquartile range (25th to 75th percentile) of the CRPS distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

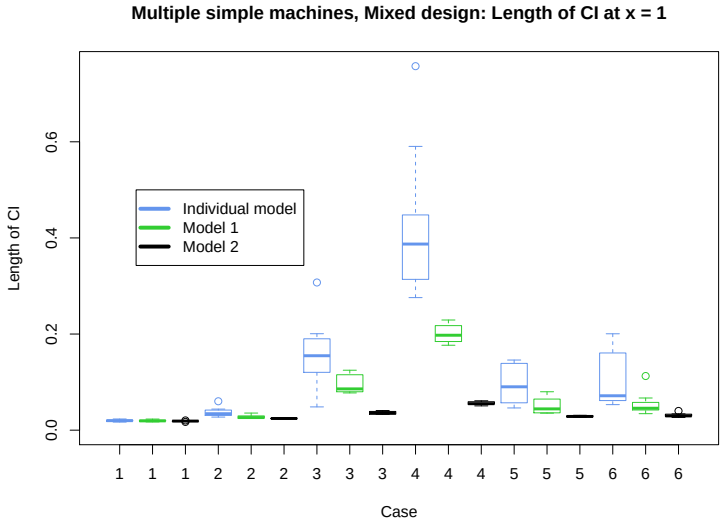


(a)

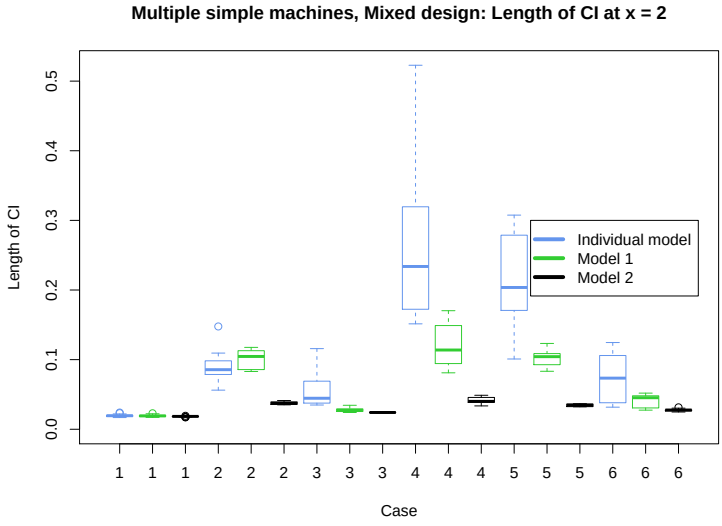


(b)

Figure 5.43: MSE boxplots for multiple simple machines, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the MSE. Each box represents the interquartile range (25th to 75th percentile) of the MSE distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.



(a)



(b)

Figure 5.44: CI boxplots for multiple simple machines, Mixed design at the interpolation points $x = 1$ (a) and $x = 2$ (b). Blue, green and black boxes represent the individual model, Model 1 and Model 2 respectively. The x-axis represents the cases of Mixed design, and the y-axis represents the length of CI. Each box represents the interquartile range (25th to 75th percentile) of the length of CI distribution, with the whiskers extending to the most extreme data point that is no more extreme than 1.5 times the interquartile range. If any points are more extreme than this, then they are represented by points. The line in the middle of each box represents the median.

Case	**Individual model**		**Model 1**		**Model 2**	
	$x = 1$	$x = 2$	$x = 1$	$x = 2$	$x = 1$	$x = 2$
1	100 (8)	100 (8)	100 (8)	100 (8)	100 (8)	100 (8)
2	100 (60)	**60 (0.1)**	100 (60)	**100 (60)**	100 (60)	**100 (60)**
3	**90 (40)**	100 (60)	**100 (60)**	100 (60)	**100 (60)**	100 (60)
4	**100 (60)**	**100 (60)**	**100 (60)**	**100 (60)**	**100 (60)**	**100 (60)**
5	**90 (40)**	**90 (40)**	**90 (40)**	**90 (40)**	**100 (60)**	**100 (60)**
6	100 (60)	100 (60)	100 (60)	100 (60)	100 (60)	100 (60)

Table 5.13: Coverage table for multiple simple machines, Mixed design. The values show the coverage probabilities at the interpolation points $x = 1$ and $x = 2$, and the corresponding p-values in parenthesis, all given in percentages. Percentages in bold, represent the relevant cases discussed in this subsection. The values are calculated using 1 experiment for each model.

5.3.4 Summary, multiple simple machines:

The multiple simple machines are machines that have common value for efficiency θ, and an a-value generated from a normal distribution. The true processes of the multiple simple machines can be found in equation (2.4) and are not based on the framework of any of the three models used to perform predictions. Below follows a summary of the results from the multiple simple machines.

- The CIs for the three models fail to cover enough true values for extrapolation. Model 1 has the worst coverage probability (it does not cover any prediction points), followed by the individual model and then Model 2. The same results are found for both Design 1 (few observations per machine) and Design 2 (many observations per machine).
- Interpolation is very similar for the individual model and Model 1, while Model 2 have the best performance. It has the best CRPS, MSE and length of CIs for interpolation points.
- For extrapolation, Model 2 has the best performance, followed by the individual model and then Model 1. Model 2 has the best CRPS, MSE, lenght of CIs and coverage probability for extrapolation points.
- For pseudo extrapolation points and interpolation points that are distant from the observations at at least one side, Model 2 performs the best, followed by Model 1 and then the individual model. Model 2 has the best CRPS, MSE and length of CIs for these points.

For the models considered to have better predictive performance based on the CRPS, MSE or lengths of CI, it is the spread and the values of the distributions of CRPS, MSE or lengths of CI that is considered. For example, if two models have the same median for the distributions of CRPS values, the one with the smaller spread is considered the best. This is because it has more stable results, and thus, it is considered better than a model that has really small CRPS values for some machines and really large CRPS values for other machines. If two models have approximately the same spreads on the distributions of the CRPS, then the one with the smaller values is considered the best. If a model has a larger spread, but smaller median than another model, then none of the two models are considered the better than the other.

Chapter 6

Discussion and Conclusion

Discussion
For the three types of machines (the Ideal machines 1, the Ideal machines 2 and multiple simple machines), and for the four types of prediction points studied in this thesis (interpolation points, extrapolation points, pseudo extrapolation points and points that are located far from the observations at least on one side), either Model 1 or Model 2 has better predictive performance than the individual model. That is, evaluating the multiple machines simultaneously gives better predicitve performance than evaluating the multiple machines individually.

It is, however, important to be aware of the limitations of the study in this thesis. For the Ideal machines 1 and the Ideal machines 2, Model 1 and Model 2 have the same mathematical structure as the machines respectively. That is, they assume that the parameters are common for all the machines, which is exactly how the observations are generated from the Ideal machines 1 and 2. It is thus, not very surprising that they both perform better than the individual model for all 4 types of prediction points. When the models are applied to multiple machines with a different mathematical structure, namely the multiple simple machines, they both fail at covering enough true values when extrapolating, expecially Model 1. Where Model 1 performs significantly better than the individual model is for the pseudo extrapolation points and the points that are located far from the observations at at least on one side. At these points, Model 1 seems to be able to learn well from the other machines that have observations located around the prediction point.

What might come as a surprise at first, is how well Model 2 performs on the multiple simple machines. It has a better predictive performance than Model 1 and the individual model for all 4 types of points. However, it is important to note that Model 2 use a prerequisite knowledge about the multiple simple machines that the other models do not. It assumes a common model discrepancy term: a common difference between the simulator and the true process. In addition, it also assumes what the individual model and Model 1 assume. That is, the individual model discrepancy term: the difference between each

machine. Brynjarsdóttir and O'Hagan (2014) have shown that, by putting contraints on the discrepancy term using prerequisite knowledge, better results are obtained. However, the improved results they obtained, was in estimation of efficiency θ, while in this thesis, the improvement is in prediction for all four types of prediction points.

Having two model discrepancy terms, Model 2 becomes more flexible than Model 1. It has better performance for the Ideal machines 2 for all four types of prediction points, and there is no considerable difference between the two models when they are applied to the Ideal machines 1. For the Ideal machines 2, the difference between the predictive performances of the two models becomes larger when the variance of the common discrepancy term for the Ideal machines 2 becomes more dominant compared to the variance of the individual discrepancy term. Model 2 gets better predictvie performance for all four types of prediction points, and even though Model 1 has better predictive performance at extrapolation points, it fails to have a large enough coverage probability for extrapolation points at times.

Note that, the study performed in this thesis is practically possible due to INLA. It is challenging to perform experiments with Model 2, as it is a model with two discrepancy terms following Gaussian fields with Matérn covariance matrices. The computation time for the 100 experiments performed on the Ideal machines 2, subcase 2, Mixed design, is about 33 hours on a laptop. It is also not obvious if it is possible to perform the experiments for Model 1 and the individual model, within a reasonable time without INLA. With a Markov chain Monte Carlo algorithm, it is practically impossible to perform the experiments in this thesis. Hence, although the results from this thesis suggest that it is better to evaluate machines simultaneously, it is only considered better because INLA has a reasonable computation time for the experiments performed.

Conclusion
Extrapolation for simulation models is difficult, and Brynjarsdóttir and O'Hagan (2014) have shown that the model of Kennedy and O'Hagan (2001), does not necessarily give good results for extrapolation. The study done in this thesis, suggests that evaluating multiple machines with the same parameters and hyperparamaters simultaneously, can improve predictive performance when extrapolating. By either using machines that have data around the extrapolation point (so that it becomes a pseudo extrapolation point), or using machines that do not have observation points around the extrapolation point (so the point is still an extrapolation point) can improve the predictive preformance. Furthermore, even interpolation and prediction on points that are located far from observations can be improved by evaluating multiple machines simultaneously.

In this study, all of the machines are synthetical. Suggestions for further work, is to perform the same experiments on real machines, or synthetical machines where the true process does not have an analytical form (for example, if solving differential equations is needed to find an approximation for the true process). Another suggestion, is to vary the calibration parameters for each machine, for example, for the multiple simple machines, the efficiency θ can be generated from a distribution with common parameters for all machines.

Bibliography

Blangiardo, M., Cameletti, M., 2015. Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.

Brynjarsdóttir, J., O'Hagan, A., 2014. Learning about physical parameters: The importance of model discrepancy. Inverse Problems 30 (11), 114007.

Casella, G., Berger, R. L., 2002. Statistical inference. Vol. 2. Duxbury Pacific Grove, CA.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Kennedy, M. C., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (3), 425–464.

Lam, H.-T., 2019. Code for "bayesian calibration models and inference for multiple machines". `https://github.com/Inusagi/Master-s-thesis`, accessed: 2019-07.

Lindgren, F., Rue, H., et al., 2015. Bayesian spatial modelling with r-inla. Journal of Statistical Software 63 (19), 1–25.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Rasmussen, C. E., 2004. Gaussian processes in machine learning. In: Advanced lectures on machine learning. Springer, pp. 63–71.

Robert, C., 2007. The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

Rue, H., Martino, S., Chopin, N., 2009a. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71 (2), 319–392.

Rue, H., Martino, S., Chopin, N., 2009b. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society B 71, 319–392.

Saigal, S., Mehrotra, D., 2012. Performance comparison of time series data using predictive data mining techniques. Advances in Information Mining 4 (1), 57–66.

Statista, 2016. Internet of things (iot) connected devices installed base worldwide from 2015 to 2025 (in billions). https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/, accessed: 2019-06.

Winsberg, E., 2003. Simulated experiments: Methodology for a virtual world. Philosophy of science 70 (1), 105–125.

Appendix A

A.1 Generation of Gaussian fields

To create observations for the Ideal machines 1 and the Ideal machines 2, Gaussian fields are needed to create the values $\xi_j(x_{ij})$ and $\delta(x_{ij})$ for $i = 1, ..., n_j$ and $j = 1, ..., N$ (equation (4.2) and (4.3)).

Generation of $\xi_j(\cdot)$
The $\xi_j(\cdot)$'s are different for all the N machines and are generated for both the Ideal machines 1 and the Ideal machines 2. For an ideal machine j, with the observation points at locations $x_{1j}, x_{2j}, ..., x_{n_j j}$, $\boldsymbol{\xi}_j = (\xi_j(x_{1j}), ..., \xi_j(x_{n_j j}))$ (a vector with the ξ_j values at the locations of the observations points) are generated from a multivariate normal distribution

$$\boldsymbol{\xi}_j \sim \mathcal{N}_{n_j}(0, \Sigma_{n_j}).$$

Here, Σ_{n_j} is a $n_j \times n_j$ matrix, and element (i, j) takes the value

$$\Sigma_{n_j}^{ij} = \sigma_0^2 \cdot \left(1 + \frac{2\sqrt{3}d}{r}\right) \exp\left(-\frac{2\sqrt{3}d}{r}\right),$$

where $d = |x_i - x_j|$, and r and σ_0^2 are the SPDE parameters found in (3.10). The spatial range is set to $r = \sqrt{2}$ and the marginal variance is set to $\sigma_0^2 = 0.09/\sqrt{0.5}$. **Alg. 1** shows the pseudocode for how the $\boldsymbol{\xi}_j$'s are generated for all the N machines.

Data: Location of observation points for all machines:
$x_{1,1}, x_{2,1}, ..., x_{n_1,1}, x_{1,2} ..., x_{n_N,N}$
Result: $\boldsymbol{\xi}_j$ for $j = 1, ..., N$
$r = \sqrt{2}$;
$\sigma_0^2 = 0.09/\sqrt{0.5}$;
for $j \leftarrow 1$ **to** N **do**
    D = distance matrix of locations of observations for machine j;
    sigma = $\sigma_0^2 \cdot \left(1 + \frac{2\sqrt{3}D}{r}\right) \exp\left(-\frac{2\sqrt{3}D}{r}\right)$;
    xi[index j] = generate n_j-variate normal distribution with mean **0** and variance sigma;
end

Algorithm 1: Pseudocode for generating $\boldsymbol{\xi}_j$ for $j = 1, ..., N$.

Generation of $\delta(\cdot)$

$\delta(\cdot)$ is identical for all the N machines and is only generated for the Ideal machines 2. Unlike with the case of $\xi_j(\cdot)$, the multivariate normal distribution used to generate $\delta(\cdot)$, does not generate values for each observation. This is due to the fact that $\delta(\cdot)$ is generated once for all the machines, and thus, the vector containing the values of $\delta(\cdot)$ is about a hundred times longer than $\boldsymbol{\xi}_j$. To get around this, the multivariate distribution generates $\delta(\cdot)$-values only at fixed locations. If any of the locations of the observations do not overlap with the fixed points, linear regression is used to approximate the $\delta(\cdot)$-value at this particular location. **Fig. A.1** shows an example of generated $\delta(\cdot)$-values at fixed locations. The red point represents an observation with a location between two of the fixed points. These two neighbor points are used to perform linear regression to find the $\delta(\cdot)$-value of the red point. This is done for all the observations for all the N machines using the exact same $\delta(x)$-function. More details on how $\boldsymbol{\delta}_j = (\delta(x_{1j}), ..., \delta(x_{n_j j}))$ is generated for all the N machines, are shown in the pseudocode **Alg. 2**.

Data: Location of observation points for all machines:
$x_{1,1}, x_{2,1}, ..., x_{n_1,1}, x_{1,2}..., x_{n_N,N}$
Result: $\boldsymbol{\delta}_j$ for $j = 1, ..., N$
$r = \sqrt{2}$;
$\sigma_0^2 = 0.09/\sqrt{0.5}$;
delta.x = evenly spaced points between $[0.2, 4]$ with step 0.01 (the fixed points);
D = distance matrix of delta.x;
sigma = $\sigma_0^2 \cdot \left(1 + \frac{2\sqrt{3}D}{r}\right) \exp\left(-\frac{2\sqrt{3}D}{r}\right)$;
delta = generated from a multivariate normal distribution with mean **0** and variance sigma (length of delta is the length of delta.x);
slopes = diff(delta)/diff(delta.x), (slopes of all pairs of neighboring points of delta.x);
for $j \leftarrow 1$ **to** N **do**
    map.x = floor locations of observations from machine j to 2 decimals;
    x.index.map = match(map.x, delta.x), (finds index of the locations in delta.x of observations floored) ;
    map.z = delta[x.index.map], (finds delta of the smallest of the two neighbors of the location of the observation);
    delta[index j] = slopes[x.index.map]*(temp.x - map.x) + map.z, (using point-slope equation of a line, to perform linear regression, and finding $\boldsymbol{\delta}_j$);
end

Algorithm 2: Pseudocode for generating $\boldsymbol{\delta}_j$ for $j = 1, ..., N$.

A.2 Variance of model discrepancy

The total variance of the model discrepancy terms for the Ideal machines 2 is

$$\begin{aligned} \text{var}(c_1\delta(x) + c_2\xi_j(x)) = \\ \text{var}(c_1\delta(x)) + \text{var}(c_2\xi_j(x)) = \\ c_1^2\text{var}(\delta(x)) + c_2^2\text{var}(\xi_j(x)) = \\ c_1^2\sigma_0^2 + c_2\sigma_0^2 = \\ (c_1^2 + c_2^2)\sigma_0^2, \end{aligned} \tag{A.1}$$

where the last equality holds because the marginal variance of $\xi_j(\cdot)$ and $\delta(\cdot)$ are equal. It is desired that the variance of the model discrepancy term of the Ideal machines 1 and the total variance of the two model discrepancy terms of the Ideal machines 2 have the same value. Thus, $(c_1^2 + c_2^2)\sigma_0^2 = \sigma_0^2 \Rightarrow c_1^2 + c_2^2 = 1$, where the first equality holds because the marginal variance used for the Ideal machines 1, have the same value as the marginal variances of the Ideal machines 2.

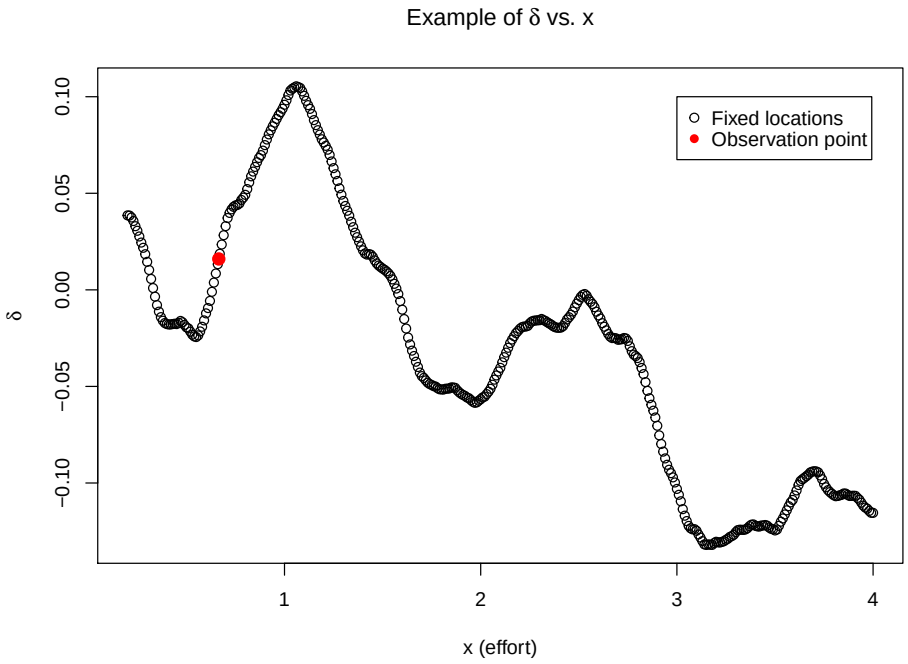


Figure A.1: $\delta(\cdot)$ vs x. at the fixed locations that are evenly spaced between $[0.2, 4]$ with increment 0.01. The red point represents an observation with a location between two fixed locations, and the corresponding delta value is found by linear regression of the two neighboring fixed locations.