Bayesian Approaches to Emulation for a Complex Computer Crop Yield Simulator with Mixed Inputs

Muhammad Mahmudul Hasan

A Thesis presented for the degree of
Doctor of Philosophy

Statistics and Probability
Department of Mathematical Sciences
University of Durham
England

Dedicated to

My lovely parents

for all their unlimited supports and prayers

My wife Suraya Binte Khurshid for her unconditional love, care and dedication and daughter Sehrish Hasan Waniya to make my days more enjoyable

Bayesian Approaches to Emulation for a Complex Computer Crop Yield Simulator with Mixed Inputs

Muhammad Mahmudul Hasan

Submitted for the degree of Doctor of Philosophy

Abstract

Agriculture is one area where the simulation of crop growth, nutrition, soil condition and pollution could be invaluable in any land management decisions. The Environmental Policy Integrated Climate Model (EPIC) is a simulation model to investigate the behaviour of crop yield in response to changes in inputs such as fertiliser levels, soil, steepness, and other environmental covariates. We build a model for crop yield around a non-linear Mitscherlich Baule growth model to make inferences about crop yield response to changes in continuous input and factor variables. A Bayesian hierarchical approach to the modelling was taken for mixed inputs, requiring Markov Chain Monte Carlo simulations to obtain samples from the posterior distributions, to validate and illustrate the results, and to carry out model selection.

The emulation of complex computer simulations has become an effective tool in exploring this high-dimensional simulated process's behaviour. Initially, we built a Bayes linear emulator to efficiently emulate crop yield as a function of the simulator's continuous inputs only. We explore emulator diagnostics and present the results from the emulation of a subset of the simulated EPIC data output. Computer models with quantitative inputs are used widely, but the challenge is incorporating the factors. We propose a framework for solving this issue considering the Bayes linear emulation approach. We explore a variety of correlation structures to represent the mixed inputs and combine this with the Bayes linear approach to construct an emulator. Finally, we developed a method to make an optimal decision for the farmers to gain maximum utility considering yield and pollutants, accounting for weather factors, land characteristics and fertiliser use.

Declaration

The work in this thesis is based on research carried out in the Department of Mathematical Sciences at Durham University. No part of this thesis has been submitted elsewhere for any degree or qualification, and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

My first and foremost acknowledgement should go to almighty Allah for inexhaustible grace and opportunities to mount from a countryside area to one of the top Russell Group universities. I am in debt for blessing me with the Durham Doctoral Scholarship out of many magnificent Candidates.

Most importantly, I would like to thank my exquisite, pleasant, and bountiful supervisor Dr Jonathan Cumming for his valuable time, proper guidance, unconditional support, and patience throughout my PhD. He has constantly stimulated me about all aspects of my life, academic or non-academic. The afford and abutment provided for me cannot be conveyed with some dictionary words. I appreciate his excellent deed, and he will always be my prayer.

I would also like to show my gratitude to my second supervisor Professor Ian Vernon. Although his role was official, he always encouraged me about this work and provided fruitful feedback. I also acknowledge Dr Ashar and Lioba for their help with the simulation part of this thesis.

It is impossible to describe the unconditional support from my family members. During COVID, I could not support my parents, but surely, they will be delighted to be part of my PhD journey. My wholehearted love towards them for their support since birth, and I am gratified to my mother for her calls every morning. My grave gratitude also goes to my parents-in-law for their stanchion; they treated me as a child of their own.

I hugely cherish my beloved wife, Suraya, for sacrificing her career to accompany me in the UK. I am grateful to her for this noble deed; without her, this work should be incomplete. The most valuable gift for me during this thesis work was the birth of my princess Sehrish. I am indebted to her for giving me joy and strength

to go forward. I also thank my brothers (Khairul, Arafat, Mostafiz, Mir) and sisters (Swarna, Sathi and Tithy) for praying and supporting me.

I want to thank my friend Asikunnaby, who has encouraged me to apply to Durham University. I am sharing my tribute to the late Professor Taslim Sazzad Mallick; special thanks to Professor Jafar Ahmed Khan and Professor Wasimul Bari for their support and recommendations for applying to this PhD study. I am grateful to Raisul for his suggestions and support during my PhD days and to Farhad for their accompany during my lonely days in Newcastle. I am acknowledging Qasem for his continuous encouragement throughout this PhD work. I am giving my inner thanks to Saidur, Shanu, Topu and Reid for supporting my family and me.

I am obliged to Durham University for awarding me the DDS to pursue this PhD. I am also grateful to the Department of Mathematical Sciences and Ustinov College, Durham University, for providing every assistance during the PhD period and the Charles Wallace Trust for a PhD Bursary. Finally, I acknowledge the University of Dhaka for the travel award from Bangladesh to the United Kingdom and all other fruitful support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Goal of the Thesis

The main objective of this thesis is to build a general framework to seek the maximum expected utility considering the proper use of fertilisers, land characteristics (soil type and steepness), and weather by combining crop yield and pollutants. We explore several vital techniques to satisfy the primary goal of this thesis. The analysis of this research uses the Environmental Policy Integrated Climate (EPIC) simulator data, a complex computer crop simulation model. The study begins by reviewing crop yield models to identify the best fit for the simulated data. We construct a hierarchical Bayesian model structure with mixed data inputs and implementing several diagnostic tools. We perform emulation using a Bayes linear approach and propose a framework for qualitative and quantitative variables. Finally, we use utility and history matching methods to seek the maximum expected utility decision concerning the inputs. Some of the achievements of this thesis can be highlighted as follows:

1. Bayes linear emulation technique for qualitative and quantitative inputs using a factor input correlation matrix (Chapters 5 and 6)

2. Using utility and implausibility to seek maximum expected utility for mixed inputs (Chapters 5 to 7)

1.2 Background and Context

The agricultural industry is a priority-based research sector that needs to analyse crop yield growth to maintain food adequacy worldwide. It is natural for farmers to seek greater profitability from agricultural production; however, it is challenging to predict crop yield in certain uncertain circumstances, such as the type of crops to cultivate, land characteristics, weather effects and fertilisation. Understandably, to make a good amount of money, our farmers must be fully aware of the proper use of fertilisers, the optimal level to be used, land characteristics and exogenous factors like weather.

Proper fertilisation is crucial for crop yield growth as an essential component of our models and to maintain the required levels of food. Nitrogen and Phosphorus are the primary two nutrients for crop productivity [101], and healthy plants use Nitrogen to preserve protein and chlorophyll for photosynthesis. According to one study [117], plants with excessive use of Nitrogen might produce lower yields, exceed the production time-bound, and damage human beings and the environment. So, proper timing and balanced Nitrogen fertilisation are essential to achieve the maximum yield [147]. Phosphorus is the second most crucial nutrient for plant growth, and optimum Phosphorus fertilisation is indispensable to achieve maximum yields [44]. Excessive use of Phosphorus fertiliser also gives way to excessive losses of the crop yield [65, 70, 109].

In addition to being a significant part of agricultural productivity, soil type is another source of nutrients for crop yield growth [88, 106]. About 95% of our total food production is over-reliant on soils. Due to the increasing population over the last five decades, we have introduced extreme pressures on soils [104], which causes lower yields, especially in grass yield [19]. Another critical factor in crop growth and yield is the weather effect due to its consequences on crop outcomes [115]. Extreme hot temperature hampers the crop's development and production, but sometimes low temperatures negatively affect yield. A study [118] illustrated that an increase in temperature led to a decrease in crops such as Wheat, Rice, Maize, and Soybean production. The low temperatures devastate crop yield, especially for vegetables [20]. From the scientific mechanism of the agricultural output, over-fertilisation and

under-fertilisation are detrimental for achieving maximum yield. The well-planned soil type choice and consideration of volatile weather effects are also crucial to crop yield.

1.3 EPIC Crop Simulation

In the recent past, conventional field experiments were considered the only data source for the global agriculture sector. However, this traditional agricultural research is losing popularity due to the increasing need for data within a limited time frame, high expense, need for intensive labour and failure to provide site-specific and complete seasonal information.

Alternatively, crop simulation computer models are promising options to solve these challenges of field experiments. Crop models are now pivotal in predicting agriculture productivity, considering weather volatility, soil erosion, and pollution [94, 119, 123]. Agricultural field experiments running for a couple of years generate the same data within minutes of formulation on a laptop, or desktop [69, 86]. These simulations are getting wider attention and can successfully give optimal decisions considering climate change, land management and so on [57, 105].

We will use one such crop simulation or computer model, the Environmental Policy Integrated Climate (EPIC) model, for this thesis. The EPIC simulator is a complex computer model [12, 24, 89] updated with weather conditions, soil type, and steepness to generate the crop yield and pollutants as time series outputs. Initially, it was used to quantify soil erosion and later changed to a biophysical system to generate crop yields, pollutants, etc. This biophysical system represents the whole physical system using a set of differential equations. The practical use of EPIC has three broad aspects: improvements of the crop growth model and assess the impact of climate change and pollution [143]. This thesis will cover all of these aspects. In Chapter 2, we discuss the EPIC model in more detail and the entire simulation process to generate the time series data based on a catchment in the UK.

1.4 Crop Yield Models

Combining crop simulation models with statistical models makes it possible to get answers to many research questions and decision support before harvesting any crop. The behaviour of the yield of a particular crop in response to endogenous variables such as fertilisation levels and exogenous variables such as weather has been extensively studied in the literature's [37, 50, 87, 110]. These literature's are providing an analytical way to gather quantitative information about the growth of crops in terms of inputs by using a mathematical equation [96]. Some linear and non-linear models are extensively used to assess crop yields. Typically, yield models comprise a non-linear relationship between the observed yield and the fertiliser levels for a particular crop. The model is also expected to respect the number of intuitive features of the relationship between these quantities, such as the yield should increase monotonically in response to additional fertiliser and the presence of a plateau effect beyond which additional fertilisation will have no further benefit. A challenge that needs to be better addressed is adapting such models to account for discrete-valued inputs, such as land characteristics or management scenarios [138]. In Chapter 3, we reveal some background studies of crop yield models and then extract the crop models to use in conjunction with EPIC data with basic principle features.

1.5 Bayesian Inference

Bayesian inference is a statistical approach which uses data to update beliefs about uncertain quantities of interest and combines prior information and data into a meaningful joint probability distribution through the Bayes rule. Bayesian inference has proven to be a very effective and helpful way to model the trend of agricultural productivity [49, 131], which is usually efficient for crop models with continuous inputs, but the presence of factor variables is challenging to formulate. However, formulation within a fully Bayesian method encounters difficulties while analysing multivariate problems, such as specifying meaningful priors and dealing with high-dimensionality. Chapter 4 forms a Bayesian hierarchical model for mixed inputs. We use Hamiltonian Monte Carlo within the No-U-Turn Sampler algorithm to generate

posterior samples and to validate it using the best-fitted crop yield growth function from Chapter 3.

1.6 Computer Models and Emulation

A computer model, f, represents a physical system and can simulate the outputs (such as crop yield and pollutants) in terms of the interest of the study [21, 56]. This model runs with a set of inputs (such as fertilisers, soil, weather for the EPIC) $x \in \mathbb{R}^p$ to a physical system and hence provides the outputs of interest. Computer models are used to explore large and complex physical systems, and these models are widely used in different fields of science, technology [21], politics [47], and business [41]. This thesis focuses on an application within agriculture - specifically, the simulation of EPIC and crop yield modelling.

1.6.1 Emulation and Bayes Linear Approach

Computer models are often complex, with many parameters, and high-dimensional, which require many more evaluations [121] to compute. Statistical modelling, or emulation, of the output of computer simulation, has become an increasingly helpful tool for analysing such complex systems within the sciences [72, 129, 134]. While a computer simulation can be constructed to capture our best understanding of the mathematical and scientific processes within the system, we are typically left with substantial uncertainties regarding the precise operation of the system. These uncertainties can range from simple uncertainties on the values of the parameters to more complex uncertainties surrounding our understanding of the science represented by the simulation [67]. Consequently, this motivates a statistical treatment of such data and model analysis.

An alternative option to understand a complex system is to use the emulation technique, which mimics the behaviour of these complex computer experiments. Emulation is an effective tool for modelling computer simulations, where a parametric model of the simulator's response to input changes may not be known a priori. A fully Bayesian approach would require distributional specifications for each of the

parameters in such an emulator and simulation-based methods for any subsequent inference, which becomes challenging when dealing with computer models with large numbers of outputs. So we need an approach which does not require detailed specifications like specifying the distributions.

A fast statistical approximation of a complex computer model uses the Bayes linear emulation approach [74]. This emulation [59] technique considers the specifications based on partial belief for all uncertain quantities rather than considering fully probabilistic specifications of prior and data observation. Two equations govern the Bayes linear emulation approach to calculate the adjusted mean and variance. This approach is proven to be effective due to its simple mathematical formulation. Chapter 5 starts with the context of emulation with the basic structure and introduces the Bayes linear and Gaussian process emulation approaches. This chapter shows the general Bayes linear emulation set-up to calculate adjusted mean and variance with some diagnostic tools to check the validity. Finally, Chapter 5 demonstrates a 1-D example and EPIC simulation data for continuous inputs over some crops.

1.6.2 Emulation with Mixed Inputs

The use of complex computer modelling with qualitative and quantitative inputs is an issue in various fields of study. But the problem is finding a suitable way to model the mixed inputs in the context of complex computer experiments. The quantitative inputs problem has a particular structure of correlation functions, but the mixed input problem still needs the proper form of this covariance function. This thesis will review existing procedures and propose a framework for mixed inputs using the Bayes linear emulation method [59], which has only been used for quantitative inputs. Chapter 6 generalises the idea of the primary emulation function for qualitative and quantitative inputs with some existing methods to construct the factor input correlation matrix. We apply our proposed model to one of the EPIC crop yields and check the performance measures of the factor input approaches using different validation tools. We combine factor modelling with the Bayes linear emulation technique to calculate the adjusted expectations and variances.

1.7 Utility and Decision

The utility is a number which measures the desirability of any items or events. We can then compare the utility of different items or events to determine the specific preference between them. Agriculture needs to balance crop yield and pollutants to seek maximum utility. Using the Bayes linear emulation technique, utility measurement, and history matching motivated us to seek an input space for maximum utility by combining crop yield and pollutants. We have to consider a utility measure by minimising the pollutants from crop yield, which arises to use a linear utility function to assess the linear effect between them. To find the region in the input space of maximum utility, we need to apply history matching [71, 74], which determines how far the emulator expectation is from a maximum expected utility. History matching requires an implausibility measure [74] to identify the mismatch between the emulator expectation and maximum expected utility. As a part of decision-making, it is considered that the smaller values of implausibility are good representations of the input space. Chapter 7 is about seeking the best input space to obtain the maximum expected utility by combining yield and pollutants for both inputs, which starts with the utility measures and the history matching technique to calculate the implausibility. A sensitivity analysis is performed to obtain the maximum expected utility and build a framework considering the history-matching approach to obtain the non-implausible region.

1.8 Organisation of the Chapters

This thesis consists of eight chapters, including this introduction. Figure 1.1 shows a graphical overview of the organisation of the thesis. Chapter 8 contains concluding remarks with limitations, further research and research achievements from Chapter 2 to 7.

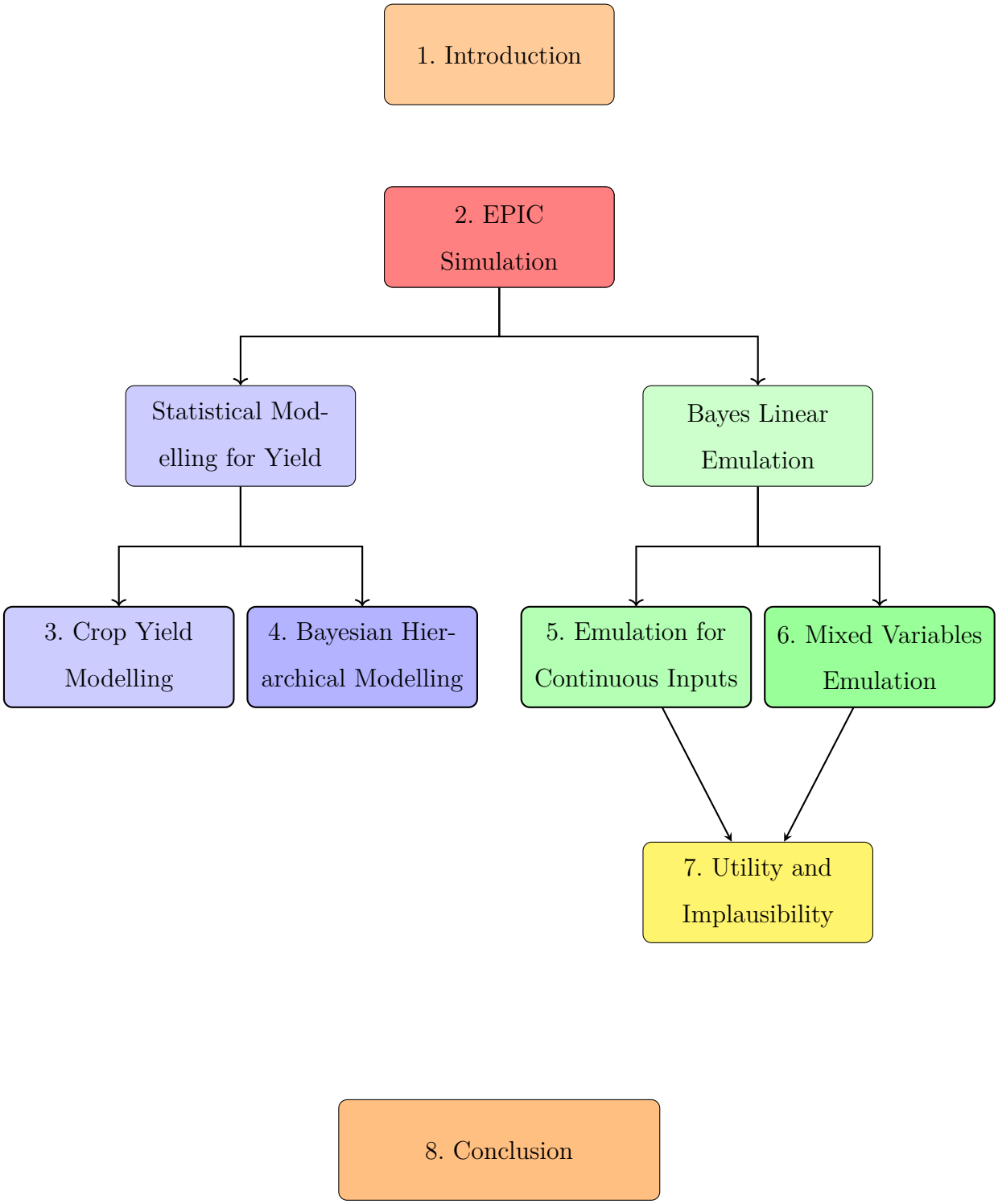


Figure 1.1: Organisation of the Chapters

Chapter 2

Environmental Policy Integrated Climate (EPIC) Simulation & Data

2.1 Introduction

A crop simulation model describes crop growth procedures as a function of soil, weather, nutrients, and crop management [54]. Crop simulation models have been used effectively for decision-making [102, 105], which is crucial for agriculture production. One such model is named the Environmental Policy Integrated Climate (EPIC) Model, a cropping system model designed with mathematical equations to generate the crop yield and pollutants [89] concerning a set of inputs. The complete analysis of this thesis is based on this EPIC simulator's outputs. The details about the EPIC model and its simulation are illustrated in the following sections.

Section 2.2 discusses the EPIC and its detailed process procedure with a flowchart. In Section 2.3, we provide the inputs and outputs for the EPIC simulation for a catchment. In the penultimate Section 2.4, we initially display a simulated data set from the EPIC simulator and then show some basic plots after processing. Finally, the concluding remarks are in Section 2.5.

2.2 EPIC Simulator

The EPIC model is a simulation-based model widely used in the agricultural industry for simulating crop yield, water use, Nitrogen and Phosphorous levels, emissions of carbon dioxide, and land management systems [13, 24, 73, 125]. The EPIC model is also used for assessing the cost of agricultural production and making optimal decisions [13]. The model was developed in 1985 in the United States, and it considered unique parameter values for eighty crops under one growth model [13, 16, 89]. The details of this growth model and other mathematical forms are extensively discussed in the EPIC manual [22].

The main components of the EPIC model are weather year simulation, hydrology, nutrient cycling, soil erosion, yield growth, tillage, the temperature of soils, economic and environmental condition [13, 16, 24]. The model can generate data for 100 years, and can be used to investigate the relationship between yield, growth and the impacts of the environment [125]. Due to its ability to generate data for 100 years, it is considered as an ideal source for projecting trends of food demand [125]. The whole process of the EPIC model is shown in Figure 2.1 and can be described as;

1. Initially, EPIC reads the starting data and computes the first day of simulation.

2. Using simulation data of day 1, it computes the daily weather effect with soil and steepness combinations.

3. EPIC models land management techniques and fertilisation for crops accounting for weather effects, and hence simulates the crop growth.

4. Storing the Day 1 outcome, EPIC simulates again for Day 2 and continues this procedure for 365 days.

5. Finally, a summary of the whole year's output is saved, and the process continues for the following year.

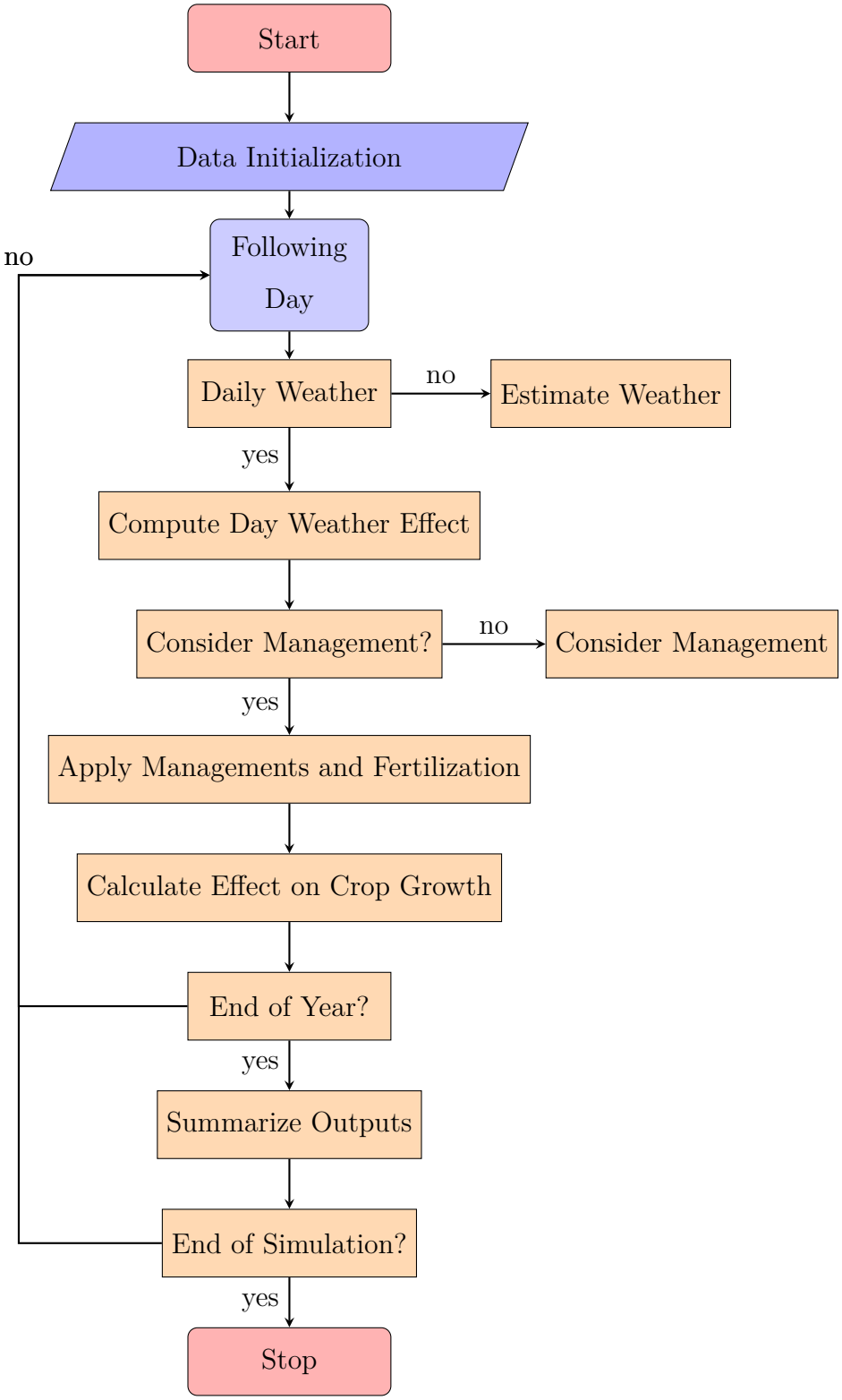


Figure 2.1: Flowchart of EPIC Model Simulation

2.3 EPIC Simulation Setup and Data

This EPIC simulation that will be investigated in this thesis is part of the Economic and Social Research Council Project (2019) [124]. In this project EPIC works on a 1-hectare area with crop yields and pollutants. It is based on two geographical catchments: (1) Eden, located in the North-East of England [128], and (2) Wensum is located in the South-East of England [132]. This research explores the data concerning the Wensum catchment.

The data for our analysis comprises a large-scale simulation from the EPIC simulator over a fully-factorial design in crop rotations, fertilisers, land characteristics and weather. The EPIC simulator for our study consists of four stages: input, simulator, output and post-process. The overall simulation run provides a simulated large ($\gg$ 20 terabyte (TB)) amount of time series of annual crop yields and pollutants for all crops over 58 years subject to the various input conditions of crop rotations, fertilisers, land characteristics and weather. These yields and pollutants were then post-processed and aggregated into data sets for each 'unique' crop. The whole process of the EPIC simulation that we consider is given as follows.

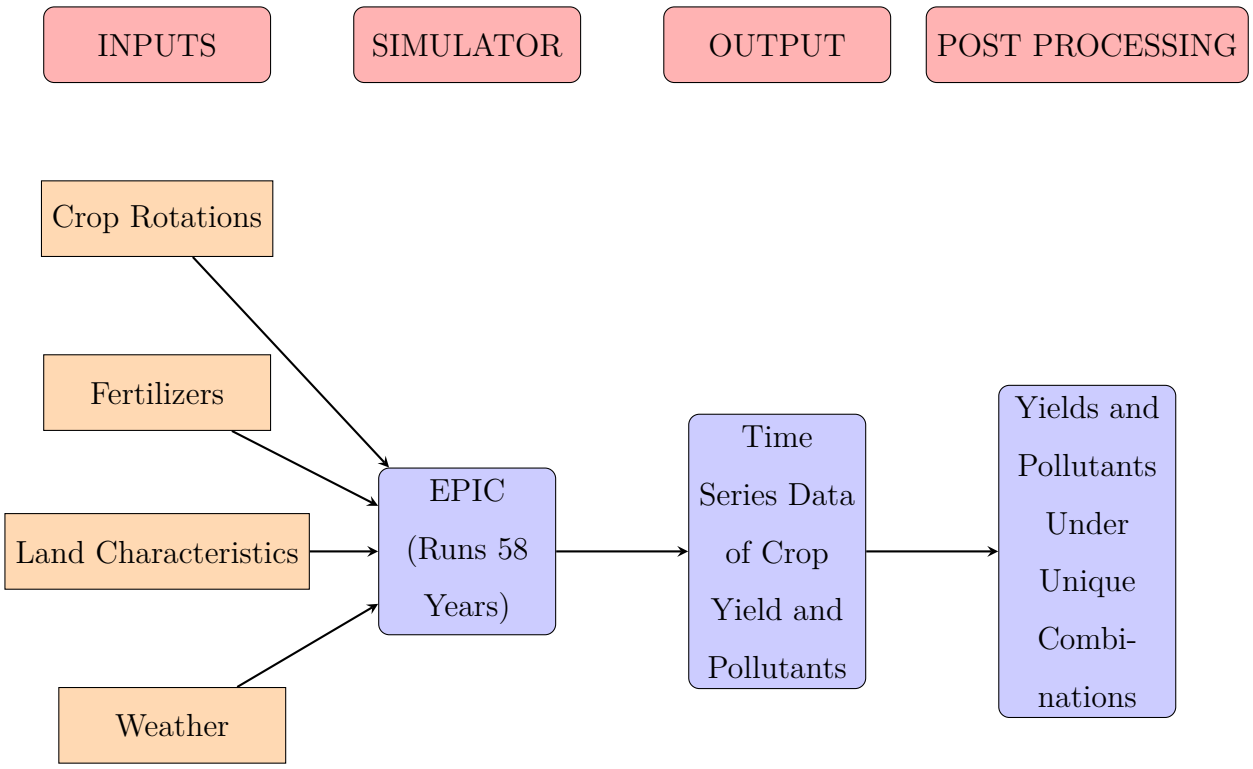


Figure 2.2: Flowchart of EPIC model Simulation to Generate Unique Yield and Pollutants

2.3.1 Crop Rotations

A crop rotation is a sequence of different crops planted and harvested in sequence over time, after which the entire sequence is repeated for as long as necessary. Typically, crop rotations are used to replenish nutrients in the soil that may have been depleted by previously harvested crops. In the case of the EPIC simulation, a total of 30 distinct crop rotations were considered.

For illustration, Rotation 6 comprises three distinct crops over a 5-year period: Winter Wheat (WW), Maize (MAIZ), and Spring Barley (SBAR). The sequence of crops in this rotation is: Winter Wheat, Maize, Spring Barley, Winter Wheat, Winter Wheat. As EPIC simulates the crops over a 58-year period, this sequence is then cycled for the duration of the simulation.

Additionally, as the sequence of crops is all which identifies a crop rotation, we can offset the start of the simulation and plant, for example: Maize, Spring Barley, followed by three instances of Winter Wheat. This is still the same rotation of the same crops, though given the sequential nature of the simulation and historical weather data we would get different simulation results for each of the crops. Repeating this process we can obtain five different sequences of the same crop rotation, as shown in Table 2.1, giving rise to five different simulations and sets of outputs for this crop rotation under the same set of simulator inputs. A subset of the other crop rotations is shown in Table 2.2, noting that the other rotations include a variety of different crops and are not all the same length.

Across all the crop rotations, several crops of particular interest were identified to be the focus of our analysis - in the case of our example rotation, these were Maize, Spring Barley, and the final instance of Winter Wheat. These crops were given labels (MAIZ1, SBAR1, WW1) to identify these unique instances of these crops. Note that while, for example, Winter Wheat will feature in many rotations, this is the only instance in conjunction with maize and spring barley and so the behaviour here is unique to this combination. Each 58-year simulation of EPIC will then yield multiple occurrences of each of these unique crops under slightly different conditions, thus one simulation run yields many simulated observations of, say, WW1. These simulated values are then aggregated along with the similar instances of WW1 from simulations of the offset rotations to form the data set for this unique crop for a given set of input parameters.

Table 2.1: Illustration of Simulation use for Weather data (Wensum - Rotation 6)

Weather Year	1954	1955	1956	1957	1958	...	2011
Year of Simulation	1	2	3	4	5	...	58
Simulation 1	WW	Maize	SBAR	WW	WW	...	SBAR
Simulation 2	WW	WW	Maize	SBAR	WW	...	Maize
Simulation 3	WW	WW	WW	Maize	SBAR	...	WW
Simulation 4	SBAR	WW	WW	WW	Maize	...	WW
Simulation 5	Maize	SBAR	WW	WW	WW	...	WW

Table 2.2: Subset of Wensum Rotations No. $1-8$; Crop codes: WOSR: Winter Oil Seed Rape, SBEET: Sugar Beet, WBAR: Winter Barley, VPE: Vining Peas, GR(R): Grazing Grass Reseed, CAR: Carrot.

Rota	1	2	3	4	5	6	7	8	9
1	WW	WW	WOSR	WW	WW				
2	WW	WW	WOSR	WW	WW	SBEET	SBAR		
3	WW	WOSR	SBEET	WW	WBAR	WOSR	WW		
4	WW	WW	WOSR	WW	WW	GR(R)	GRAZ1		
5	WW	VPE	WW	SBAR	WW	WOSR	WW		
6	WW	Maize	SBAR	WW	WW				
7	WW	WW	CAR	WW	WW	WOSR	WW		
8	VPE	WW	SBAR	WBAR	WOSR	WW	SBEET	WW	WW

2.3.2 Inputs of EPIC Simulation

The inputs are considered an essential part of a simulator. In this section, we illustrate the detailed description of the four different types of inputs processed for each simulation, which is a fully factorial design, and they are (*i*) Fertilisers, (*ii*) Land Characteristics (steepness and soil type), and (*iii*) Weather.

2.3.2.1 Fertiliser

Two fertilisers, Nitrogen (N) and Phosphorus (P), were used as inputs to our simulations. Each fertiliser took one of 13 different values 0, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100 and was treated as a continuous variable. The ranges $0 - 100$ correspond to a standardized range for each rotation, however, the actual fertilizer values have different maxima depending on the rotation.

2.3.2.2 Land Characteristics

We have two categorical variables representing land characteristics in the simulation: soil type (So) and the steepness or slope (St). The data for soil was collected based on NSRI NATMAP soil mapping data [146], giving three different soil types for the catchment. In Table 2.3 we present the name of the soil, and the categorical level used in the simulation.

Table 2.3: Name of the Soil, its Descriptions and Levels for Wensum Catchment

So level	Name	Description
6	Beccles	medium loamy over clayey chalky drift
7	Burlingham	medium loamy chalky drift
8	Barrow	light loamy over chalky drift

The steepness (St) variable is ordinal, with four different levels. The values of the variable steepness are in degrees, and for the Wensum catchment, they are presented in Table 2.4.

Table 2.4: Steepness Levels and Degrees for Wensum Catchment

St level	Steepness (Degrees)
5	0.4
6	1
7	1.6
8	2.7

2.3.2.3 Weather

Weather input to the simulation takes the form of observed weather data from 1954 - 2011 from the UK meteorological department data archive [130]. This data comprised five indicators used for the simulation purpose: (i) Daily maximum temperature, (ii) Daily minimum temperature, (iii) Precipitation, (iv) Humidity, and (v) Wind speed. However, these individual attributes were not treated as distinct inputs to the model to be varied separately, as these quantities are in the form of complex time series. Therefore, a similar strategy was adopted to that of the crop rotations whereby the time series of weather data was offset and cycled to fill the simulation period. This effectively results in a series of distinct weather scenarios which can be applied to each simulation, and thus each unique crop.

2.3.3 EPIC Simulation Outputs

From the EPIC simulations, yield is the primary output variable of the study. There are two types of yield; GYLD, which is grain yield relevant to grain crops and forage/foliage/ fodder yield (FYLD) applicable to leafy crops, both of them are expressed in terms of tonnes in a hectare (t/ha). In Table 2.5, we see grain crops Spring Barley and Maize yield output as GYLD; leafy crops Silage generate yield output as FYLD. In addition to yield output, EPIC can generate another two outputs total biomass (BIOM) and below-ground biomass (BGBM). Table 2.5 presents the yield variables used for some EPIC unique crops in terms of GYLD and FYLD.

Table 2.5: Yield Variables Used for Each Crop for Wensum Catchment

Crop	Yield Output Variable
Spring Barley	GYLD
Maize	GYLD
Silage	FYLD
Hay	FYLD+GYLD

Another output variable of the study is the pollutants for respective crops. For our analysis, we will explore pollutants: Nitrogen to the river (NRLOAD) and Phosphorus to

the river (PRLOAD). Nitrogen to the river (NRLOAD) is the load/kg per day and can be expressed as [89],

$$NRLOAD = QNO3 + DRNN + SSFN, \tag{2.1}$$

where $QNO3$ is the nitrate loss in runoff (surface runoff), $DRNN$ is the soluble Nitrogen in drainage outflow, and $SSFN$ is the Nitrogen in subsurface flow. Phosphorus to the river (PRLOAD) is the load/kg per day and can be expressed as [89],

$$PRLOAD = QAP + DRNP + SSFP, \tag{2.2}$$

where QAP is the Phosphorus loss in runoff (surface runoff), $DRNP$ is the soluble Phosphorus in drainage outflow and $SSFP$ is the Phosphorus in subsurface flow.

2.4 Simulated Data Set From EPIC

A subset of the simulated output data from EPIC simulator based on our design is shown in Table 2.6 for rotation 16 of the Wensum catchment for the crop Spring Barley, which is a grain crop. Here, N and P indicate the Nitrogen and Phosphorus fertiliser, respectively, St means the steepness, So means the soil, Wy and Sy tell the weather and simulation year factor variables depending on the crop rotation.

Table 2.6: Simulated Data Set for Unique Crop Spring Barley (Rotation-16)

Inputs						Outputs		
N	P	St	So	Wy	Sy	Yield	NRLOAD	PRLOAD
0	0	5	6	1	4	6.861	12.955	0.259
0	0	5	6	1	12	3.445	11.014	0.081
...	...	...	...	...	...	...	...	...
5	5	6	6	3	36	3.561	7.260	0.047
...	...	...	...	...	...	...	...	...
10	20	8	7	8	44	2.751	23.452	5.911
...	...	...	...	...	...	...	...	...
100	100	8	8	8	60	2.872	71.914	42.484
...	...	...	...	...	...	...	...	...

Finally, Yield, NRLOAD and PRLOAD are three outputs from the EPIC simulator which interest this study. Each row of Table 2.6 indicates one unique combination of the inputs. Columns 1 to 6 show the input data, and the remaining columns indicate our three outputs yield and two pollutants. Columns 1 and 2 are about N and P fertilizers standard values, St is steepness with four levels in column 3, column 4 reveals soil (So) levels, Wy is the weather variable with eight different levels, and Sy is the simulation runs in 8 different years. For our analysis, we will focus on the continuous fertiliser inputs of Nitrogen (N) and phosphorus (P) levels, which were simulated over a discrete grid of values, each with 13 values over [0, 100]. Thus, for a given crop and fixed combination of land and weather variables, we obtain a grid of 169 simulated yields, Y, and pollutants NRLOAD and PRLOAD in response to N and P.

Figure 2.3 presents the yield for the crop Spring Barley, and two pollutants NRLOAD and PRLOAD. The crop Spring Barley simulated at 8 different years based on crop rotations such that simulation year, $Sy = 8$ with 3 levels of soil, 4 levels of steepness and 8 levels of weather makes a total combination of $3 \times 4 \times 8 \times 8 = 768$. We present the plots for a subset of 15 different input combination values to assess the general trend. To demonstrate the basic trends of our outputs, we consider the inputs Nitrogen and Phosphorus.

From the upper panel (left) of Figure 2.3 showing the response crop yield of Spring Barley to Nitrogen, we can see that yield shows an increasing trend for all of the simulations initially; however, some of the combinations show a plateau after a certain level of Nitrogen input. On the other hand, we can see an increasing trend for the low level of Phosphorus input for a few of the unique combinations and all different simulations shows a flat and weak response. Chapters 3 and 4 will further explore the mathematical relationship and statistical modelling of the yield concerning N and P. We will also use advanced statistical techniques to assess the impact of N and P on crop yield in Chapters 5 to 7.

The lower panel of Figure 2.3 shows the line plots for the pollutants PRLOAD (left) and NRLOAD (right). The pollution for PRLOAD shows a linear relationship with respect to Phosphorus such that PRLOAD is higher for high Phosphorus values and lower for low P values. And, for the pollutant NRLOAD, we can see a non-linear relationship for the input Nitrogen. We will explore extensively in Chapter 7 for these pollutants, such as to assess the impact of both inputs on the pollutants.

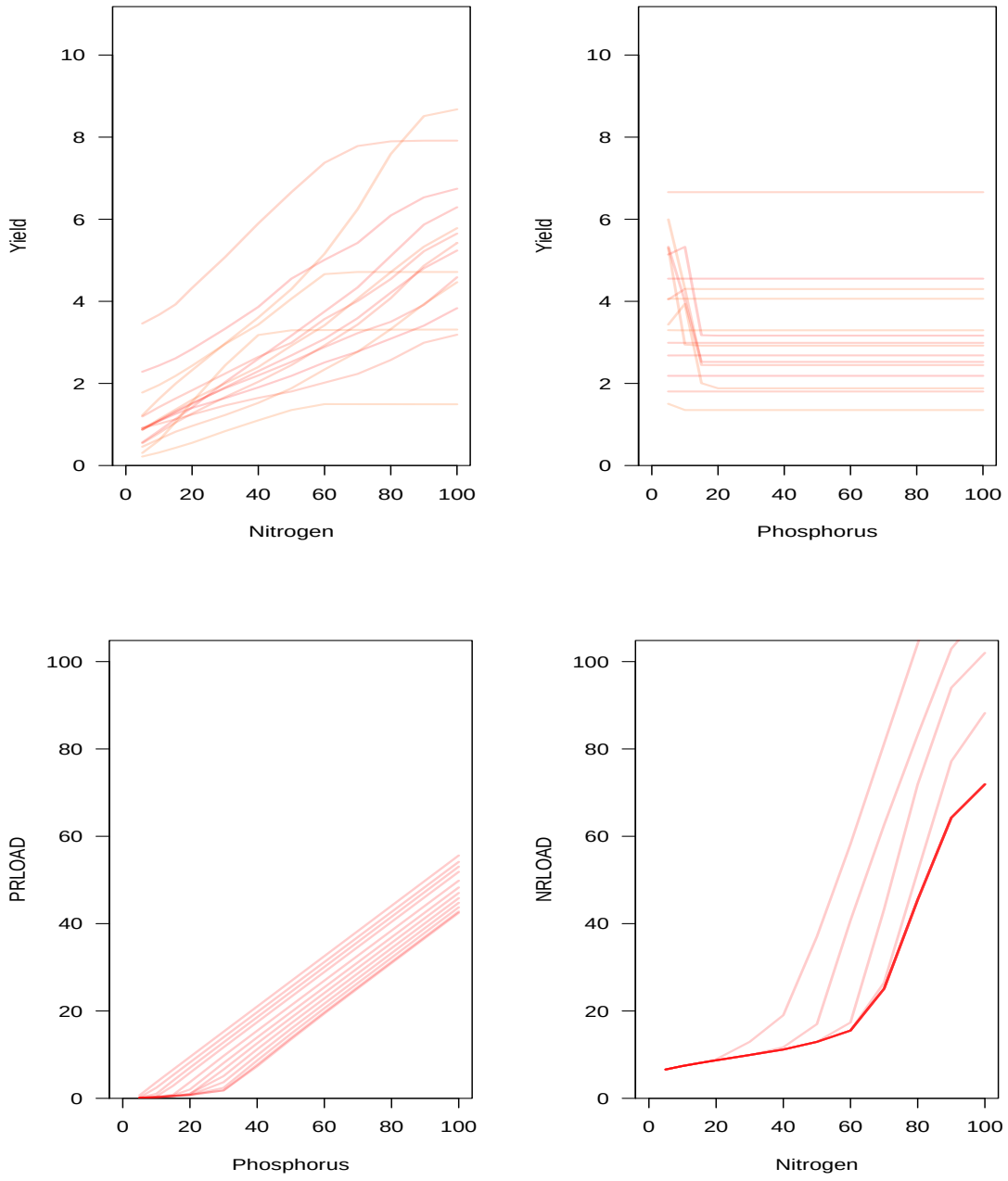


Figure 2.3: Upper left and right panel: Spring Barley crop yield with the response of Nitrogen and the Response of Phosphorus. Lower left panel: Pollutant PRLOAD for the response of Phosphorus input, Lower right panel: Plot for NRLOAD with Respect to Nitrogen.

2.5 Conclusion

In this Chapter 2, we describe the sequential stages of the EPIC simulator and outline the simulation study. We discuss the quantitative and qualitative input variables with the outputs yields and the pollutants for the catchment Wensum. Finally, we explored the general trend of the EPIC simulator outputs for the Spring Barley crop.

From Chapter 2, we have acquired knowledge about handling data from the EPIC simulator. We have gained an understanding of the crop rotations, real-life steepness, soil type, fertilisation for the Wensum catchment and the weather within the implementation of the EPIC simulator. We have also learned a general idea about the catchment, the definition of the pollutant outputs, and how to extract crop yield and different pollutants. This Chapter also provided an understanding of the yield output type, such as the grain or foliage. Finally, this Chapter taught us how to extract a subset of data and summarise the data as general plots.

Chapter 3

Crop Yield Modelling

3.1 Introduction

Identifying a proper mathematical relationship is essential for assessing crop growth in response to the inputs of a study [48, 77, 98]. The appropriate form of crop modelling is essential and enables us to predict the yield output for short to long-run periods [80]. After checking the best cropping model, decision-making becomes possible for governments to import any food or not to maintain the demand for food [61]. This Chapter explores and fits the crop yield models from existing literature, eventually using them in conjunction with the EPIC simulations data to assess the relationship with the inputs. The main goal of this Chapter is to find an appropriate mathematical relationship between the input fertilisers and the output yield.

In Section 3.2, we survey some important crop yield modelling literature to identify the crop yield models to fit to the EPIC simulated data with the basic standard features for crop yield models and summarise the relevant characteristics in Section 3.2.1. We perform some preliminary data analysis for three different crops in Section 3.3, fit the crop yield models and identify the best fitting model for the EPIC data in Section 3.4, and finally draw some concluding remarks with our understanding in Section 3.5.

3.2 Background of the Crop Yield Modelling

The response of crop yield to inputs such as fertilisation has been extensively studied in the literature [14, 17, 23, 31, 38, 37, 42, 58, 62, 87] providing many valuable insights

into the expected or desirable behaviour of any yield response model. It is expected that a yield may show monotonicity, growth plateau, and increasing return to scale, and these are treated as critical features of a crop model. Without these features, a crop response model will not respect the expected behaviour of a yield response, and a detailed discussion of these features is given as follows;

1. Linearity: Linearity is one of the basic features of the crop yield model. The mean crop yield is expected to show a linear trend with the increase in the input. So for a simple linear regression for the input N and P with output yield Y and error term e we can write;

$$\begin{aligned} Y &= \beta_0 + \beta_1 N + \beta_2 P + e, \\ E(Y) &= \beta_0 + \beta_1 N + \beta_2 P. \end{aligned} \tag{3.1}$$

2. Monotonicity: Crop yield should be monotonically increasing in response to fertilisers. For example, with the increase in Nitrogen levels, the crop yield should increase.

3. Growth Plateau: For the feature of growth plateau, crop yield will increase initially and then become constant in response to increasing fertiliser levels. For example, in Figure 2.3(h), the crop yield increases monotonically to $N = 70$, and after this adding more fertiliser has no effect such that it shows a flat trend until $N = 100$.

4. Input Substitution: Input substitution is where one input fertilizer can be replaced with another for a specific crop [111]. For example, a crop can give maximum output for either N or P and a combination of the two, so an increase in N substitutes for an increase in P and vice versa, rather than affecting the crop yield. In other words, N and P are equivalent in their effect on yield.

5. Returns to Scale: This feature measures the proportional change of the yield concerning inputs N and P. If the yield increases more than the effect of both fertiliser inputs, it is called increasing return to scale such that the sum of slopes for N and P is greater than 1. On the other hand, if yield output decreases due to both inputs' effects, it is treated as decreasing return to scale (such that the sum of regression coefficients corresponding to inputs N and P is less than 1). Finally, suppose the return of yield shows the same growth for the effect of the inputs; it is called constant return to scale (the sum of regression coefficients corresponding to inputs is 1). For

our study, increasing return to scale is very much expected because of the increasing nature of yield, such that the sum of slopes should be positive and greater than 1. The following relations clarify the feature of the returns to scale by considering β_1, and β_2 are the slopes for the inputs from Equation 3.1.

- If $\beta_1 + \beta_2 = 1$ constant returns to scale of the output yield for increasing inputs N or P.
- If $\beta_1 + \beta_2 < 1$ decreases the output for an increasing input. However, the rate still increases when both slopes are positive and within $0 < \beta_1 + \beta_2 < 1$.
- If $\beta_1 + \beta_2 > 1$ an increase of the output for an increasing input.

3.2.1 Crop Yield Models

Various functional forms of yield, Y, in response to Nitrogen, N, and Phosphorous, P, have been proposed in the literature [14, 17, 23, 31, 38, 37, 42, 58, 62, 87] and can be loosely categorised into three groups. We summarise these yield models in Table 3.1:

1. **Linear Models:** Multiple Regression, Quadratic, and Square Root function.
2. **Non-linear Models:** Power function, Gompertz function, Logistic function, and Mitscherlich-Baule function.
3. **Threshold Models:** Linear Von Liebig function, and Non-linear Von Liebig function.

The linear models are ubiquitous in statistical modelling, and the form of these models is shown in Table 3.1 with their standard features.

We illustrate a brief mathematical structure of the non-linear and threshold models with their properties.

Power Function or Cobb Douglas Model: The power function [87], also known as Cobb Douglas Production function in Economics [145], is given as follows:

$$Y = \beta_0 N^{\beta_1} P^{\beta_2}, \tag{3.2}$$

where Y equals yield, N and P are the inputs, β_0 is the intercept of the power model, and β_1, β_2 are the slopes for the inputs N and P. It is expected that the power model will show growth plateaus and monotonicity with the main features of the power function given as follows:

- If $\beta_1 + \beta_2 = 1$ constant returns to scale of the output for increasing input.
- If $\beta_1 + \beta_2 < 1$ decrease of the output for increasing input.
- If $\beta_1 + \beta_2 > 1$ increase of the output for increasing input.

Table 3.1: Summary of Models of Y, in Response to N, and P.

Name	Function	Features
Linear	$Y = \beta_0 + \beta_1 N + \beta_2 P$	Linearity
Quadratic	$Y = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 N^2 + \beta_4 P^2 + \beta_5 NP$	Input substitution
Square Root	$Y = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 (N)^{1/2} + \beta_4 (P)^{1/2} + \beta_5 (NP)^{1/2}$	Input substitution, no plateau
Power Model	$Y = \beta_0 N^{\beta_1} P^{\beta_2}$	Returns to scale, Monotonicity, Plateau
Gompertz	$Y = \beta_0 \exp\Big(-\beta_1 e^{-\beta_2 N}\Big) \exp\Big(-\beta_3 e^{-\beta_4 P}\Big)$	Increasing return to scale, Monotonicity, Plateau
Logistic	$Y = \frac{\beta_0}{\left(1+\beta_1 e^{-\beta_2 N}+\beta_3 e^{-\beta_4 P}\right)}$	Increasing return to scale, Monotonicity, Plateau
Mitscherlich-Baule (M-B)	$Y = \beta_0\Big[1 - \exp(-\beta_1(\beta_2 - \beta_3 N))\Big] \times\Big[1 - \exp(-\beta_4(\beta_5 - \beta_6 P))\Big]$	Monotonicity, Increasing return to scale, Positivity, Plateau
Linear Von Liebig	$Y = \min\Big[\beta_0,\ \beta_1 + \beta_2 N,\ \beta_3 + \beta_4 P\Big]$	Plateau, Linearity, Monotonicity
Nonlinear Von Liebig	$Y = \min\Big[\beta_0(1 - \beta_1 \exp(-\beta_2 N)), \beta_0(1 - \beta_3 \exp(-\beta_4 \times P))\Big]$	Plateau, Monotonicity, Positivity

Gompertz Growth Model: Gompertz growth model is one of the commonly [2, 87] used

model to assess the yield data and can be expressed as,

$$Y = \beta_0 \exp\Big(- \beta_1 e^{-\beta_2 N}\Big) \exp\Big(- \beta_3 e^{-\beta_4 P}\Big), \tag{3.3}$$

where there are five parameters to fit; β_0 is the maximum yield, β_1 is the slope parameter corresponding to fertiliser N, and β_2 is the slope parameter. The parameters β_3 is the parameter corresponding to fertiliser P and β_4 is the slope parameter for the P input. This model's main features are:

- The increasing return to scale such that $\beta_1 + \beta_2 + \beta_3 + \beta_4 > 1$.
- The monotonicity and growth plateau when the yield is at least β_0.

Logistic Regression Model: The logistic regression model uses a logit transformation of the response variable Y [11, 87], and can be expressed as:

$$Y = \frac{\beta_0}{\Big(1 + \beta_1 e^{-\beta_2 N} + \beta_3 e^{-\beta_4 P}\Big)}, \tag{3.4}$$

where β_0 is the maximum yield, β_2 is the corresponding growth rate for N, and β_4 is the slope for P. Like the Gompertz model, the Logistic model has the same increasing trend, monotonicity and growth plateau features.

Mitscherlich-Baule Model: The Mitscherlich-Baule (M-B) Model is generally used to predict crop yield and is considered as the most appropriate to assess the growth of crop yield [23, 31, 37, 42]. The Mitscherlich-Baule Model can be represented as,

$$\begin{aligned} Y &= \beta_0 \Big[1 - \exp(-\beta_1(\beta_2 - \beta_3 N))\Big] \times \Big[1 - \exp(-\beta_4(\beta_5 - \beta_6 P))\Big], \\ &= \beta_0 \Big[1 - \exp(-\beta_1 - \beta_2 N)\Big]\Big[1 - \exp(-\beta_3 - \beta_4 P)\Big]. \end{aligned} \tag{3.5}$$

Some properties of the Mitscherlich-Baule model are given as follows:

- M-B model coefficients should be positive [62] as this model only allows positive total production.
- Mitscherlich-Baule function allows for input substitution.
- The M-B function is monotonically increasing.
- It also allows for increasing return to scale such that $\beta_1 + \beta_2 + \beta_3 + \beta_4 > 1$.

Linear Von Liebig Model: The Von Liebig linear model is different to the usual linear regression specifications due to its growth plateau [23]. The Linear Von Liebig model can be written as follows,

$$Y_i = \min\Big[\beta_0,\ \beta_1 + \beta_2 N,\ \beta_3 + \beta_4 P\Big], \tag{3.6}$$

where β_0 is the maximum crop yield, and the other parameters are similar to previous models. The common features of the linear Von Liebig model are given as follows:

- The model equation shows linearity between yield and inputs.
- This model shows the plateau when the yield is at least β_0 [14].
- This model also allow the linear trend until the growth plateau.

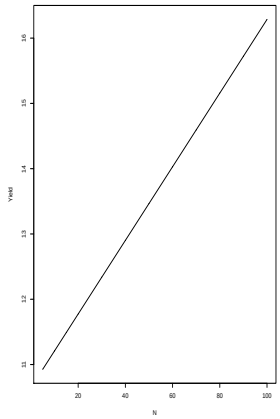
Nonlinear Von Liebig Model Another form of the model [28, 37] is known as the Non-linear Von Liebig model, and the form of this model is constructed under the Von Liebig linear model specifications, and Mistcherlich-Baule model frame work, which can be written as,

$$Y = \min\Big[\beta_0 \times (1 - \beta_1 \exp(-\beta_2 N)),\ \beta_0 \times (1 - \beta_3 \exp(-\beta_4 P))\Big]. \tag{3.7}$$

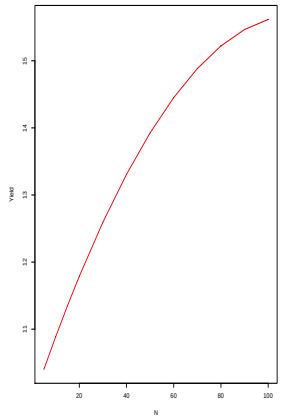
The common features of the linear Von Liebig model are given as follows:

- All the parameters $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \beta_4$ of the response function are expected to be positive [28, 37] .
- The non-linear Von Liebig model considers a yield plateau Figure 3.1 and non-substitution among N and P [37].
- This model also allows the monotonic increasing trend until the growth plateau.
- This model doesn't allow linearity but shows the increasing return to scale.
- This model doesn't allow input substitutions [62].

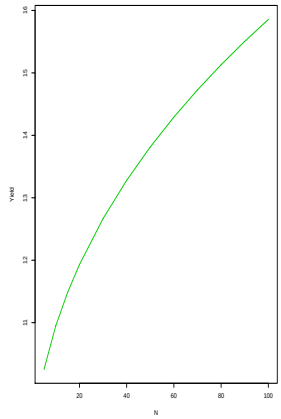
Figure 3.1 shows an illustrative plot for all nine crop yield models. These plots are the general trend plot to show the basic features of the crop yield models of interest. Overall, we can see the linearity feature from the linear model and a monotonic increasing trend feature for the quadratic, square root, Logistic, Gompertz and Mitscherlich-Baule models. The power model function shows a power shape, and finally we can see the growth plateaus from the linear and non-linear Von-Liebig models.



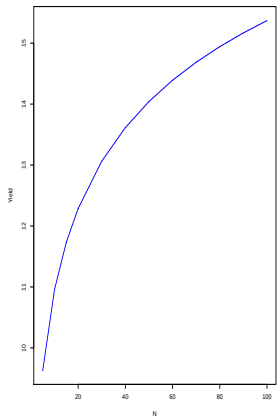
(a) Linear



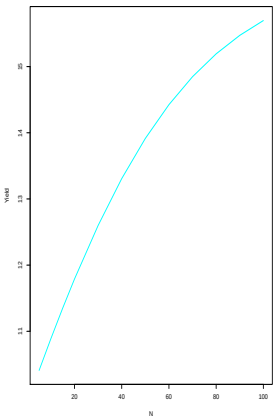
(b) Quadratic



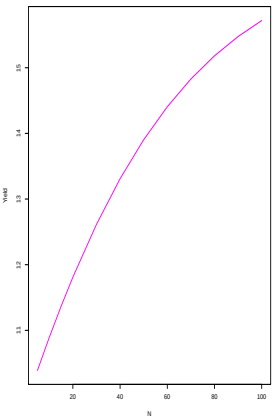
(c) Square Root



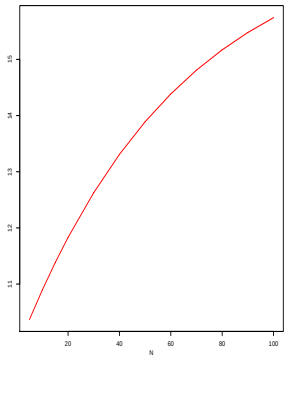
(d) Power



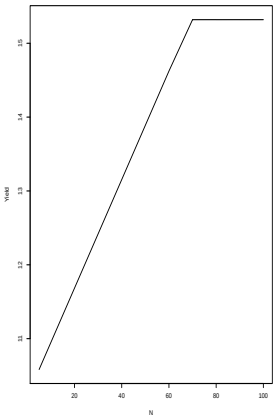
(e) Logistic



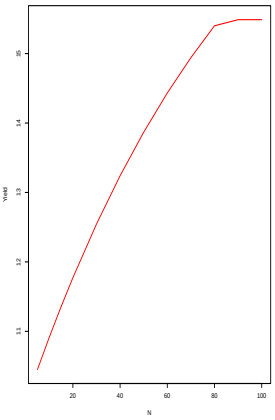
(f) Gompertz



(g) M-B



(h) Lin Von



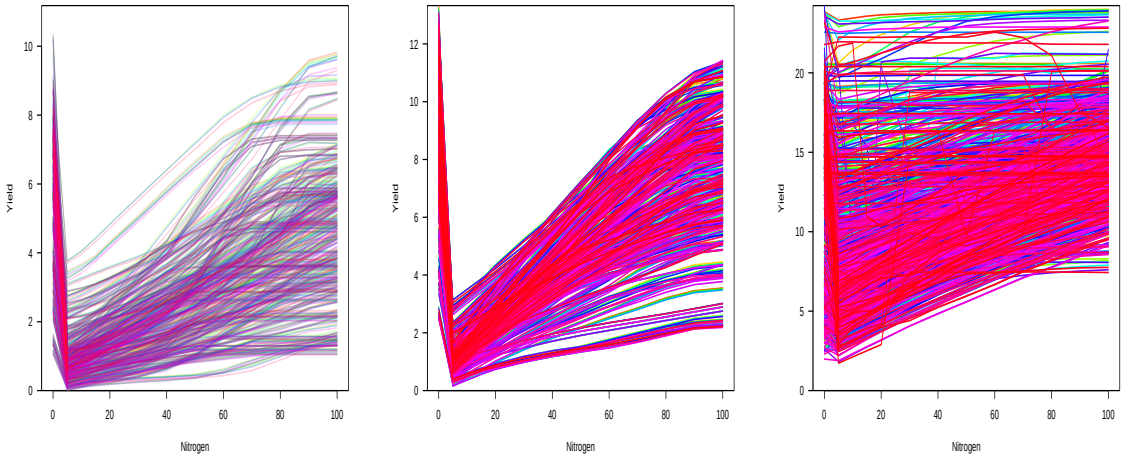
(i) Nonlin Von

Figure 3.1: Simple plot of the Crop Models

3.3 Preliminary Data Analysis for Crops Yield

In this section, we explore three sets of yield simulations from EPIC for Spring Barley, Winter Barley, and Silage. Initially, we plot the raw simulated yield for all three crops in the form of a line diagram versus Nitrogen and Phosphorus.

In Figure 3.2, we plot a line for simulated yield in response to Nitrogen for Spring Barley, Winter Barley and Silage, respectively. Colour here has no intrinsic meaning and is simply used to help differentiate different lines. Each line corresponds to a simulated yield obtained with the same combination of steepness, soil, weather scenario, and simulation year. For example, with Spring Barley we have 768 such lines obtained from 3 levels of soil, 4 levels of steepness, 8 levels of weather, and 8 levels of simulation year (this version of Spring Barley occurred 8 times within the 58-year simulation) giving this total of $3 \times 4 \times 8 \times 8 = 768$. Similarly, Winter Barley was occurred at 7 different simulation years with the same weather levels of Spring Barley, making the total combinations 672; and Silage occurred at 7 simulated years with 9 different weather levels giving the total number of combinations as 756.



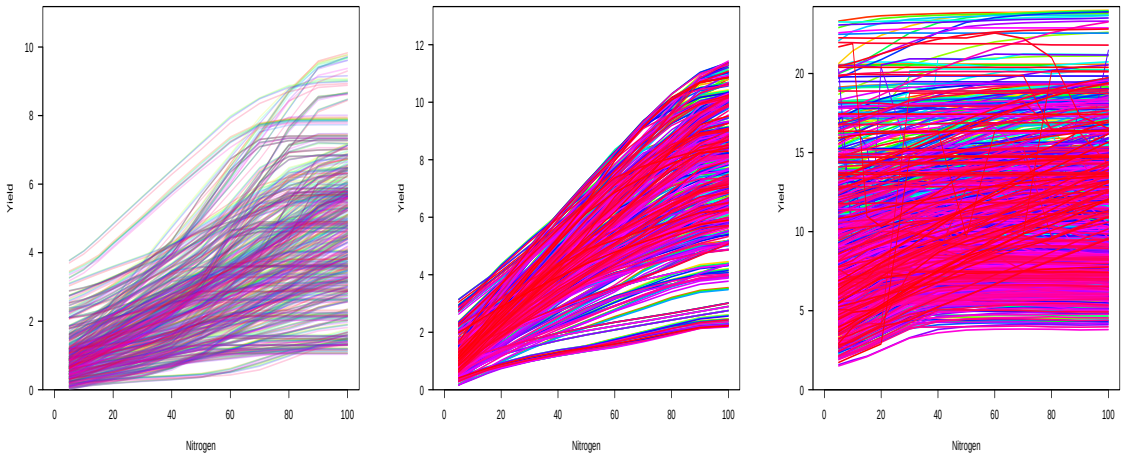
(a) Spring Barley (b) Winter Barley (c) Silage

Figure 3.2: Line Graphs for All Input Simulations Response to Nitrogen

Figure 3.2 shows all inputs simulations with respect to all 13 levels of Nitrogen inputs, including the 0 level for the crops Spring Barley, Winter Barley and Silage. Simulator outputs for $N = 0,\ P = 0$ are in-feasibly large, and clearly nonphysical as a crop response should not exhibit such discontinuity. As this looks likely due to simulation error, we

remove all data for $N = 0,\ P = 0$ from future analysis.

It is expected that yield should increase with fertiliser levels, up to a point and respond to one, other, or both fertilisers depending on the crop behaviour. From Figure 3.3, it can be said that (i) most of the simulations are showing a monotonic increasing trend as well as growth plateaus for some of the simulations concerning input Nitrogen for the crops Spring Barley and Winter Barley, (ii) Silage is showing an increasing trend for most of the simulations with some flat trends and apparent noise.



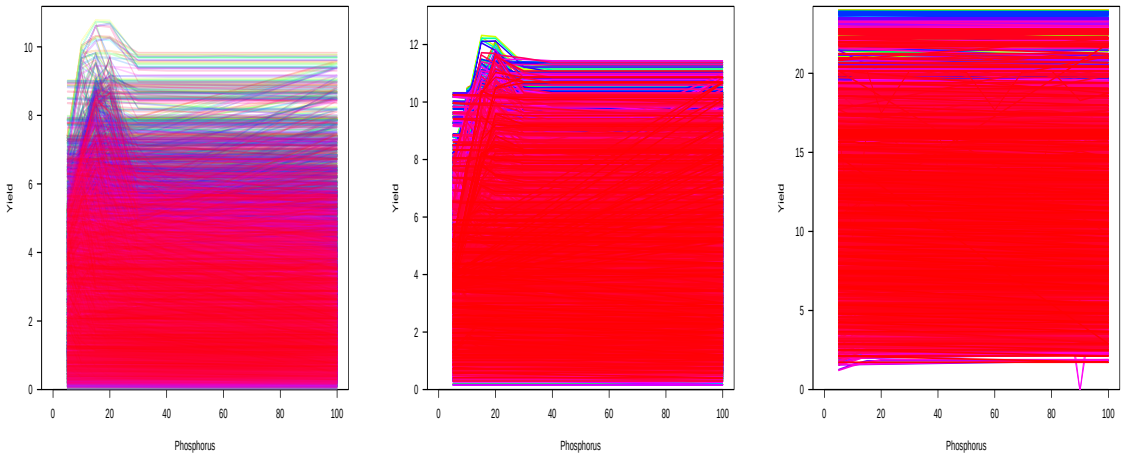
(a) Spring Barley (b) Winter Barley (c) Silage

Figure 3.3: Line Graphs for all Unique Combinations Response to Nitrogen Excluding "0" Level

In Figure 3.4, we re-plotted crop yield data of Spring Barley, Winter Barley, and Silage in response to the input Phosphorus. Figure 3.4 shows a flat yield response concerning Phosphorus (P) for most simulations except for some increasing and unusual decreasing trends for the low levels of P for Spring Barley and Winter Barley. The crop yield Silage shows a clear flat response for all levels of Phosphorus. Figure 3.5 shows 100 random simulations for P to determine the visible flat trends.

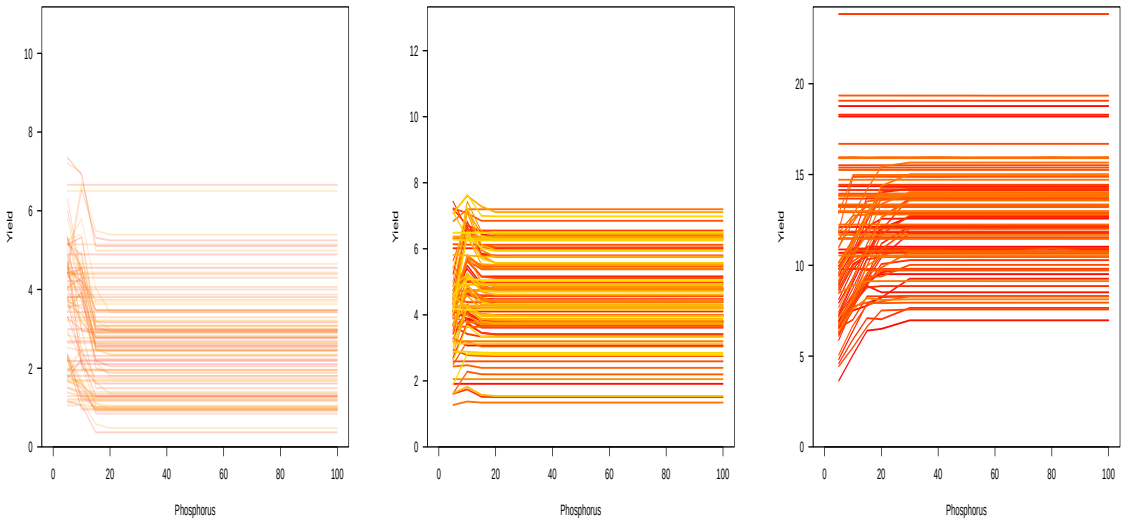
Figure 3.6 shows a subset of 15 random unique combinations to clarify the apparent trends of our three data sets. From figure 3.6, for N, it can be said that yield shows monotonicity and a growth plateau for Spring Barley and Winter Barley crops only. We can see some flat trends for a few of the unique input simulations for Silage in response to input Nitrogen with a monotonic increasing trend. However, some simulations show unusual increasing and decreasing trends for all crops at low P levels.

Overall, the yield shows a strong response to Nitrogen input for all three yield data. We can see an overall weak and flat response to the input Phosphorus except for some increasing trend for the low levels of P. The contribution of the input Phosphorus to yield will be clarified after fitting the models.



(a) Spring Barley (b) Winter Barley (c) Silage

Figure 3.4: Line Graphs for all Unique Combinations Response to the Phosphorus



(a) Spring Barley (b) Winter Barley (c) Silage

Figure 3.5: Line Graphs for 100 Combinations Response to the Input Phosphorus

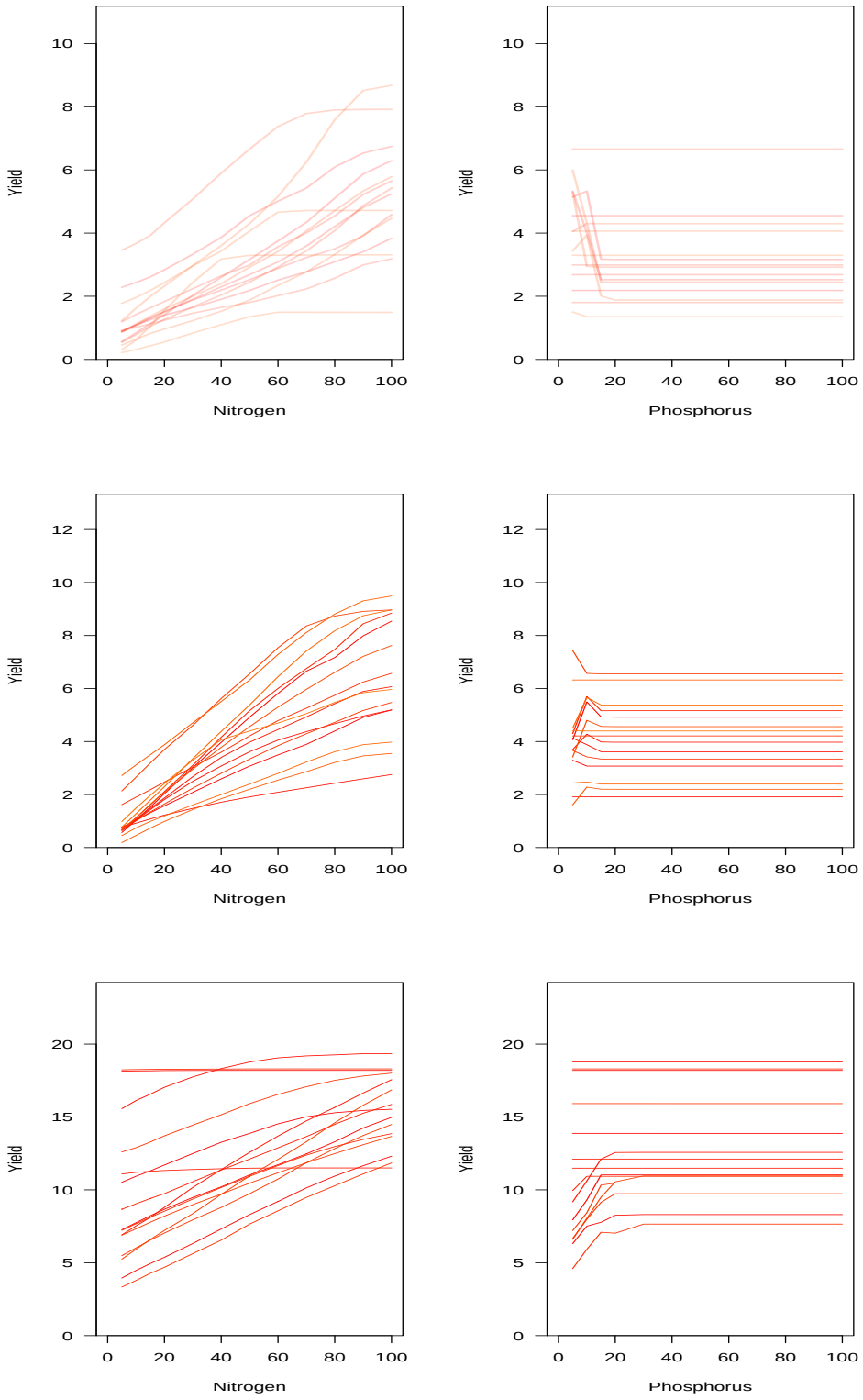


Figure 3.6: Plots of Simulated Yield for a Sample of 15 Simulations in Response to Nitrogen (Upper left panel) and Phosphorus (Upper right panel) for Spring Barley. Middle panel and lower panel Plots are for the Crops Winter Barley and Silage.

3.4 Fitting Crop Yield Models

In this section for our model-fitting, we have explored nine yield response models for three crops concerning Nitrogen and Phosphorus from Section 3.2.1. The Linear models: Simple Linear, Square Root, and Quadratic regression, are fitted using the default linear regression function $lm()$ from R-language. We assigned the starting values for the model parameters to fit the non-linear models and used the $nlsLM()$ [108] function.

We compare the nine fitted models to the data (red line) in Figure 3.7 for Spring Barley, Winter Barley, and Silage crops, respectively considering one unique simulation. From Figure 3.7, we can see that all models can achieve reasonably close fits to the data, apart from the power model (black line) for the crop Spring Barley and Winter Barley. Silage is slightly different from Spring Barley and Winter Barley, where we can see more spread and variation between the other fitted models approaching a visible plateau for both linear and non-linear Von Liebig's models. The power curve also shows a close fit to the data for the Silage crop compared to the other two crops' trends.

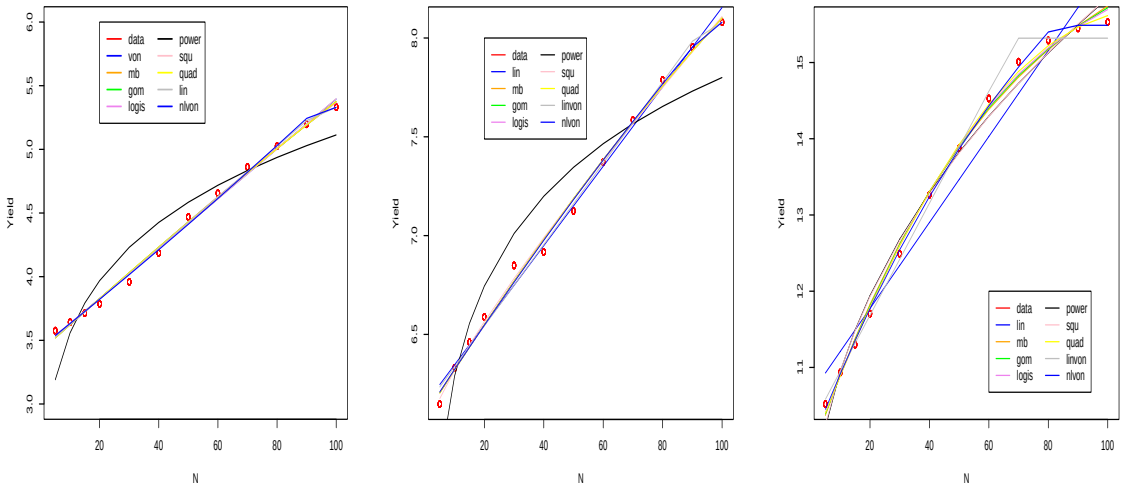


Figure 3.7: Plot of the Fitted Yield Response to Nine Models concerning Nitrogen and Phosphorus for the Spring Barley Crop (left panel); Winter Barley Crop (middle panel); Silage (right panel)

Table 3.2 summarises the nine fitted crop yield models with the residual standard error for Spring Barley and Silage crops for the subset of the data of one unique simulation. We have only presented the result of the crops Spring Barley as Winter Barley has shown the

same results. The values of the model coefficients satisfy the properties of the yield models such that the show linearity, increased return to scale such that the sum of β_i, $i = 1, 2, 3, 4$ are greater than 1, and positivity of the parameters. The models that satisfy these criteria can be considered the best-fitted models.

Table 3.2: Summary Statistics of the Fitted Nine-Crop Yield Models for the Crop Spring Barley (SBAR) and Silage

Crop(s)	Models	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	RSE
	Linear	3.44	0.019	0.0				0.05
	Quadratic	3.41	0.02	0.0	0.0001	0.0	0.0	0.04
	Square Root	3.45	0.02	0.0	0.004	0.0	0.0	0.04
	Power Model	2.48	0.16	0.0				0.19
SBAR	Logistic Model	7.70	1.25	0.01	0.0001	0.0		0.03
	Gompertz Model	9.35	1.0	0.01	0.19	0.0		0.03
	M-B Model	5.82	0.80	0.02	7.70	0.0		0.03
	Linear Von Liebig	5.33	3.42	0.02	5.33	0.0		0.04
	Non-linear Von Liebig	5.61	0.43	0.02	0.002	0.0		0.04
	Linear	10.65	0.05	0.0				0.38
	Quadratic	9.89	0.01	0.0	0.0004	0.0	0.0	0.15
	Square Root	8.53	0.01	0.0	0.78	0.0001	0.0	0.20
	Power Model	7.18	0.44	0.0001				0.42
Silage	Logistic Model	16.52	0.66	0.02	0.0004	0.0		0.12
	Gompertz Model	16.83	0.45	0.02	0.0057	0.0		0.12
	M-B Model	17.28	0.89	0.02	6.29	0.0		0.12
	Linear Von Liebig	15.32	10.21	0.07	15.32	0.0		0.14
	Non-linear Von Liebig	15.59	0.44	0.05	0.006	0.0		0.14

From the results of all the nine fitted models, the estimates corresponding to N input are significant such that p values are less than 0.05, whereas P have no effect, such that the estimated values are zero or close to zero for all the crops. Since P has no impact, we can

not substitute Phosphorus in the place of Nitrogen, meaning using both Square Root and Quadratic models is infeasible for these crop yield data. The Power model shows decreasing return to scale for both the crops such that $\hat{\beta}_1 + \hat{\beta}_2 < 1$. In all Gompertz, Logistic and Mitscherlich-Baule crop yield models, we have seen a growing return to scale such that the sum of $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 > 1$, for the crop Spring Barley only but unfortunately a decreasing trend for the Silage for Logistic and Gompertz. All of the coefficients for the Mistcherlich-Baule and non-linear Von Liebig models are positive, which is one of the main features of these models.

For the Residual Standard Errors (RSE), we have seen a higher variation among the model residuals for the Silage crop. The Power model has the highest deviation, which validates the unusual trend of Figure 3.7. The Logistic, Gompertz, and M-B models show the minimum RSE estimate for Spring Barley and Silage crops. The RSE estimates for all models are close to each other for Spring Barley, indicating a very close trend for all the models concerning simulated data apart from the power model. For the crop Silage, we can see very close RSE estimates for non-linear and threshold models but a slight variation for the linear models. So in terms of RSE, we can say the Mitscherlich-Baule, Logistic, and Gompertz models are the best-fitted yield models for the selected three crops.

The growth plateau for the Gompertz and Logistic models is not revealed for these data, which is the main feature of these models; however, they have lower RSE values and an increasing return to scale for Spring Barley. The Mitscherlich-Baule model shows a monotonic increasing trend, increasing return to scale and the positivity of the coefficients for all crops with a smaller RSE estimate. We can see the visible plateaus for both linear and non-linear Von Liebig's model, but only for the Silage crop; unfortunately, the EPIC simulated data does not reveal this attribute. The simple linear model is straightforward to use but we can not fully capture the variation of the data, and our result also showed a slightly higher RSE value. Overall, we can say that Mitscherlich-Baule (MB) model can be considered the best fit for our data for several reasons, such as (*i*) a substantial similarity between the fitted yield responses with the relative RSE values; (*ii*) only model to achieve a monotonic increasing trend, increasing return to scale and the positivity of all the coefficients; and (*iii*) is has been seen to be the best fitting model in previous studies for crop simulation models [37, 62] over others.

3.5 Conclusion

In Chapter 3, we have demonstrated nine different crop models from previous studies with their features. We have shown yield plots considering all input simulations and a 15 subset of unique input simulations. The results of these plots showed a strong response for Nitrogen for the explored crops but a weak and flat response to Phosphorus in Section 3.3. In Section 3.4, we fitted the nine crop models, and the results demonstrated that our models are very close to each other except for the power model. The Mitscherlich-Baule model showed the best fit to the data, exhibited a monotonic increasing trend, positivity and smaller RSE.

This chapter aimed to assess the best crop yield model for our crop yield data and provided a deeper understanding of the mathematical form of crop yield models and how to fit them. We also gained competence in the literature review and extracting primary information, such as model selection and reparameterisation. Finally, we have learned how to visualise a large volume of data.

The crop models are generally used to assess crop growth trends. However, linear models must fulfil some assumptions, such as linearity and independence, which makes them impracticable. Dealing with non-linear models are quite complicated due to the uncertainty of the model parameters, boundary constraint, and initialisation problems.

Chapter 4

Bayesian Hierarchical Framework for Crop Yield

4.1 Introduction

The frequentist and Bayesian approaches are two different views or statistical inference procedures used by statistical modellers. The frequentist approach estimates the parameters from the data only, like the crop yield modelling in Chapter 3. But the Bayesian approach needs to deal with both data and prior information. Bayesian inference is a well-known method to infer the posterior distribution based on prior specifications and data. A prior needs to be specified initially and then combined with the evidence using the Bayes rule [1, 18] to calculate the posterior for the parameters. This approach is extensively used in different fields of study, such as science, sports, law, medicine, etc. The Bayesian framework has been applied in agricultural research to predict crop yield growth, especially for decision making [90, 133, 136].

Several studies have been conducted from different perspectives of crop yield modelling, and some of them are highlighted in the following passage. A Bayesian framework with a normal likelihood for yield used the Michalis-Menten modified equation [97] as a mean function to estimate the effect of crop yield considering the inputs of fertiliser and farmyard manure. The crop yield was positively impacted by the input fertiliser, and the diagnostic plots showed trace plot chains are mixing well. This study illustrated the advantage of using a non-linear growth model as a mean function. A Bayesian model was used to analyse a crop simulation model (CSM) [79], and this study used Gibbs sampling with random

walk Metropolis and a satisfying convergence was diagnosed after 10000 iterations. This Bayesian model gives us an indication of the optimal number of iterations required in the MCMC. A hierarchical non-linear model was introduced for Nitrogen and soil nutrients [49] using three possible mean functions: linear plus plateau, quadratic, inverse linear, and linear methods failed to detect the unusual trend of the data. These authors used Gibbs sampling to sample from the posterior with 100000 iterations. The Bayesian framework has been used to assess the consequences of weather on yield [131] using linear and non-linear growth response mean functions. This work used the two-dimensional Gaussian regression model as the yield response, which works well to assess the effect of weather on yield. This work recommended to evaluate the variable weather in details to assess the impact on yield. A Bayesian approach was used for the analysis of yield response to inputs N fertiliser on soil type and precipitation in Africa using the Hamiltonian No-U-Turn sampler (HMC-NUTS) [133]. This study used a default four chains with 7000 iterations and found the effect of N on soils.

These previous studies mostly model yield as a function of a single fertiliser, weather or soil as a continuous input. So we must look forward to developing a more general model for our purposes with factor effects of our data. The motivation for using Bayesian inference for this research is to set up a Bayesian hierarchical framework using the non-linear Mistcherlich-Baule growth model (from Chapter 3) for both continuous and factor inputs.

We start this Chapter with the general Bayes rule, the steps needed for Bayesian modelling and the selection of priors in Section 4.2. In Section 4.2.1, we construct a Bayesian hierarchical framework for continuous inputs of N and P and incorporate the factor variable in Section 4.2.2. Section 4.2.3 discusses MCMC with some basic terms and the algorithms needed to sample from the posterior distribution. We discuss the model selection and validation tools in Section 4.2.4. We demonstrate the result of the Bayesian framework concerning EPIC data in Section 4.3 using continuous inputs. Section 4.4 provides the results of the factor incorporation, and, finally, we draw some concluding remarks in Section 4.5.

4.2 Bayesian Hierarchical Modelling Framework

In this section, we set up a Bayesian hierarchical framework for quantitative and factor inputs based on the best-fitting Mitscherlich-Baule (MB) model mean function, then draw posterior samples using the MCMC algorithm. Finally, we compare the models and validate them using different diagnostic based on the MCMC samples.

While the frequentist modelling approach is centred on using a likelihood for both observed and unobserved data, the Bayesian approach combines prior beliefs with the likelihood to obtain the posterior probability distribution. The Bayesian approach updates the posterior distribution via the Bayes rule: [1, 18]

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}, \tag{4.1}$$

where y is the real data, $\pi(\theta|y)$ is the posterior distribution of the parameters θ after gathering the data, $\pi(y|\theta)$ is the likelihood and quantifies the information from observing the data, $\pi(\theta)$ is prior distribution for θ, and $\pi(y)$ is the marginal distribution of the data, and can be expressed as;

$$\pi(y) = \int_{-\infty}^{\infty} \pi(y|\theta)\pi(\theta)\, d\theta. \tag{4.2}$$

It is common to have a non-conjugate prior distribution, thus requiring the Monte Carlo Markov Chain algorithms to simulate posterior samples. Markov Chain Monte Carlo (MCMC) methods [26, 34] are effective Monte Carlo techniques in Bayesian inference used to sample from the posterior distribution. MCMC simulations allow posterior estimates such as expectations, standard deviation, and credible intervals of Bayesian models.

There are some specific steps [51, 91, 138] to follow for our Bayesian modelling, and they are given as follows;

1. Stage I: Bayesian Modelling Framework: Specify the data model and prior without observing the data. Collect data and update the model parameters to form the posterior based on prior and likelihood.

2. Stage II: Posterior Sampling from MCMC: Draw and check the quality of the posterior MCMC samples. Finally, summarise and interpret the results obtained by MCMC sampling.

3. Stage III: Model comparison and validation: Use different diagnostic tools to validate the Bayesian model.

The prior distribution expresses a belief about an uncertain quantity before seeing data [91]. It is a key part of Bayesian analysis and can be assessed in different ways for a study. The different types of priors [10, 35, 91] are given as follows,

1. Informative Prior: A prior is said to be informative if it contains information about the parameter. We can use a Normal prior if the parameter lies on the entire line such that $N(\mu, \sigma^2)$, and Gamma prior if it lies on the positive line such that $Ga(\alpha, \lambda)$.

2. Weakly informative prior: A prior is said to be weakly informative if it contains partial information about the parameter. The main point in the use of the weakly informative prior for the stabilisation of the parameter within a reasonable range.

3. Uninformative prior: A prior is considered uninformative or diffuse if it contains only vague information about the parameters.

A prior can only be one of these three types, which are mutually exclusive. In addition to these three types of priors, a prior could also be improper if its density function can not be integrated such that CDF is less than 1 and conjugate if it results in a posterior distribution of the same type. For our problem, we will use informative and weakly informative priors.

4.2.1 Stage I: Bayesian Modelling Framework

Definition 4.2.1 (Normal Distribution)**.** *For two parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R} > 0$, the probability density function for the normal distribution with yield $Y \in \mathbb{R}$ is defined as:*

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big[-\frac{1}{2\sigma}(Y-\mu)^2 \Big]. \tag{4.3}$$

which we can write as $Y \sim N(\mu, \sigma^2)$

For each unique crop, we assume the yield Y_i, $i = 1, 2, ..., n$ follow the Normal distribution with mean $\mu(N_i, P_i|\mathcal{B})$, where $\mathcal{B} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ is the vector of regression parameters, and variance $Var[Y_i] = \sigma$. So we can say that,

$$Y_i \sim N(\mu(N_i, P_i|\mathcal{B}), \sigma). \tag{4.4}$$

Initially, we consider modelling yield as a function of the continuous covariates Nitrogen (N) and Phosphorus (P) for the study and the Mitscherlich-Baule model as the mean

response function. Thus, the MB mean yield function from Table 3.1 can be parameterised in the following way,

$$\mu_i = \mu(N_i, P_i|\mathcal{B}) = \beta_0\Big[1 - \exp(-\beta_1 - \beta_2 N_i)\Big]\Big[1 - \exp(-\beta_3 - \beta_4 P_i)\Big], \tag{4.5}$$

where parameter $\beta_0 > 0$ is the maximum yield, β_1, β_2 are the intercept and slope for the Nitrogen input, and β_3, β_4 are the intercept and slope for the Phosphorus input.

Definition 4.2.2 (Gamma Distribution)**.** *The density function of the Gamma distribution for the random variable X with shape parameter α and rate parameter λ can be written as,*

$$f(Y|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-y\lambda), \tag{4.6}$$

for $y > 0$ and zero otherwise. So we can write that $Y \sim Gamma(\alpha, \lambda)$

Parameter β_0 is the maximum yield, which is a positive finite value. We require positive estimates from the parameters β_1, β_2 for the Nitrogen variable and β_3, β_4 for the Phosphorus variable from the nonlinear MB model based on the increasing positive trend in the graphical analysis in Chapter 3. Due to these reasons and the positive nature of the MB model [37], we have considered the Gamma distribution as our priors for all five parameters of the Mistcherlich-Baule model. We have also considered the Gamma prior for our variance parameter since $\sigma > 0$. Thus we can write,

$$\begin{aligned} &\beta_k \sim Gamma(\alpha_k, \lambda_k);\ k = 0, 1, 2, 3, 4 \\ &\sigma \sim Gamma(u, v). \end{aligned} \tag{4.7}$$

Considering the term $\frac{\lambda^\alpha}{\Gamma(\alpha)}$ as constant and under the assumption that all of the parameters are independent, the general form of the joint prior distribution of all parameters can be presented as follows,

$$\begin{aligned} \pi(\beta_k; \sigma) &= \pi(\sigma) \prod_{k=0}^{4} \pi(\beta_k), \\ &\propto (\sigma)^{u-1} \exp(-\sigma v) \prod_{k=0}^{4} \beta_k^{\alpha_k - 1} \exp(-\beta_k \lambda_k), \\ &\propto (\sigma)^{u-1} \exp(-\sigma v) \exp\Big(- \sum_{k=0}^{4} \beta_k \lambda_k \Big) \prod_{k=0}^{4} \beta_k^{\alpha_k - 1}, \\ &\propto (\sigma)^{u-1} \exp\Big(- \sigma v - \sum_{k=0}^{4} \beta_k \lambda_k \Big) \prod_{k=0}^{4} \beta_k^{\alpha_k - 1}, \end{aligned} \tag{4.8}$$

where α_j and λ_k are the shape parameters and rate parameters for the β_k; u and v are the shape parameters and rate parameters for the variance σ.

Considering the yield observations Y_i as a vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ and vector parameters $\mathcal{B} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$, and assuming conditional independence of Y given $\mathcal{B}$, σ we can write the likelihood function;

$$L(\mathbf{Y}|\mathcal{B}, \sigma) = \prod_{i=1}^{n} \pi(\mathbf{Y}|\mathcal{B}, \sigma). \tag{4.9}$$

We have already assumed that our yield model follows a Normal distribution in Equation (4.4), so the likelihood function can be written as,

$$\begin{aligned} L(\mathbf{Y}|\mathcal{B}, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma\pi}} \exp\left[-\frac{1}{2\sigma}(Y_i - \mu_i)^2 \right], \\ &\propto \sigma^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma} \sum_{i=0}^{n} (Y_i - \mu_i)^2 \right]. \end{aligned} \tag{4.10}$$

Using the likelihood for an MB model for mean yield in Equation (4.10) and prior distribution in Equation (4.8), we can write the form of our posterior distribution as,

$$\begin{aligned} \pi(\mathcal{B}, \sigma|\mathbf{Y}) &\propto \sigma^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma} \sum_{i=0}^{n} (Y_i - \mu_i)^2 \right] \times (\sigma)^{u-1} \exp(-\sigma v) \exp\left(-\sum_{k=0}^{4} \beta_k \lambda_k \right) \prod_{k=0}^{4} \beta_k^{\alpha_k - 1}, \\ &\propto \sigma^{u-1-\frac{n}{2}} \exp\left[-\frac{1}{2W} \sum_{i=0}^{n} (Y_i - \mu_i)^2 \right] \times \exp(-\sigma v) \exp\left(-\sum_{k=0}^{4} \beta_k \lambda_k \right) \prod_{k=0}^{4} \beta_k^{\alpha_k - 1}, \\ &\propto \sigma^{u-1-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma} \sum_{i=1}^{n} (Y_i - \mu_i)^2 - \sigma v - \sum_{k=0}^{4} \beta_k \lambda_k \right] \prod_{k=0}^{4} \beta_k^{\alpha_k - 1}, \\ &\propto \sigma^{u-1-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma} \sum_{i=1}^{n} (Y_i - \beta_0[1 - \exp(-\beta_1 - \beta_2 N_i)][1 - \exp(-\beta_3 - \beta_4 P_i)])^2 \right. \\ &\quad \left. - \sigma v - \sum_{k=0}^{4} \beta_j \lambda_k \right] \prod_{k=0}^{4} \beta_k^{\alpha_k - 1}. \end{aligned} \tag{4.11}$$

Equation (4.11) is not recognisable as a similar form as Equation (4.8), and so our problem is non-conjugate. The structure of posterior distribution requires MCMC to draw posterior samples.

Another possible option is to use the log-normal distribution or truncated Normal, i.e. $Y_i \sim LN(\mu(N_i, P_i|\mathcal{B}), \sigma)$ due to the strictly positive nature of the yield but this would completely change the interpretation of our model. For the log-normal distribution, the mean of yield is $E(Y_i) = \exp^{\mu + \frac{\sigma^2}{2}}$, which creates the complexity due to $E(Y_i) \neq MB$ and we would need to re-parameterise the model. The yield variance is generally low, so that

negative yield values will be highly unlikely. So to maintain the same structure of the MB model, we base our analysis on the Normal distribution to consider the MB model as the mean function of the Bayesian inference.

After this, we can adapt the mean function for the factor effect considering the design matrix with $0-1$ encoding and follow the stages mentioned above to execute the Bayesian framework for mixed inputs.

4.2.2 Incorporating Factor Effects

The structure of the MB model must be generalised, such that it can be used as a mean function for a fully Bayesian model with mixed inputs. There are several ways to formulate this, and we use a pragmatic approach to keep the problem simple by introducing factor inputs using a 0 - 1 encoding. For a factor variable with levels c_j with an observed value of f_i, we write that;

$$Z_{i,j} = \begin{cases} 1 & \text{if } f_i = c_j, \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

such that if the ith observation of the factor has the jth level c_j then we have $Z_{i,j} = 1$, and $Z_{i,j} = 0$ otherwise.

We assume that the factors only influence maximum yield β_0, and do so by shifting the value. Hence, we use a simple hierarchical structure of the following form for a single factor input;

$$\begin{aligned} \beta_0 &= \gamma_0 + \gamma_1 Z_{i,1} + ... + \gamma_j Z_{i,j}, \\ &= \gamma_0 + \sum_{k=1}^{j} \gamma_k Z_{i,k}, \\ &= \gamma_0 + Z_i^T \gamma, \end{aligned} \tag{4.13}$$

where Z is the model matrix of the factor variables, and γ is the vector of factor effect parameters;; j is the length of the factor levels and γ_0 is the coefficient similar to the maximum yield of β_0. The prior for the γ_0 is considered as $Gamma(\alpha, \lambda)$; for γ_k, we consider the standard normal prior. Section 4.4 reveals more details about choosing these priors. In each model, we take the first-factor level as the baseline value, for example, for the first level of steepness factor, $Z_1^{St} = 0$ in Equation (4.13). Thus the N input MB model

can be written as follows;

$$\mu_i^N = (\gamma_0 + Z_i^T \gamma)[1 - \exp(-\beta_1 - \beta_2 N_i)]. \tag{4.14}$$

Using Equation (4.13) in Equations (4.10) and (4.11) makes it possible to calculate the likelihood and hence posterior distribution for the effect of the factors. The form of the posterior distribution is similar to Equation 4.11 and thus needs to use MCMC to generate posterior samples.

4.2.3 Stage II: Posterior Sampling via MCMC

This section initially discusses MCMC with some basic terms and gives a brief overview of the Gibbs sampling and Metropolis-Hastings algorithms. We then introduce Hamiltonian Monte Carlo within a No-U-Turn sampler.

4.2.3.1 Markov Chain Monte Carlo

MCMC is one of the key ideas in Bayesian analysis. MCMC consists of two parts; the first part is the Markov Chain (MC), and the second part is called Monte Carlo sampling. MCMC is a method applied through a set of sampling algorithms used to sample the posterior parameters using MC [113].

A MC is the basis of MCMC to generate the posterior samples from a potentially complex, high-dimensional and non-conjugate probability distribution. With a large number of iterations, the distribution of the samples converges to the desired distribution. A set of algorithms are used to build Markov chains, such as (*i*) Metropolis-Hastings, (*ii*) Gibbs Sampling, (*iii*) Hamiltonian Monte Carlo and so on. Building a MC with the target distribution makes it possible to sample from the posterior as samples from the Markov chain [113].

4.2.3.2 Metropolis-Hastings (MH) Algorithm

The Metropolis-Hastings (MH) algorithm [7] generates candidates based on the full joint density distribution of priors. MH is used to generate the sequence of random samples from a target distribution (a distribution depends on the current sample to draw the next sample) with a Markov chain. This algorithm works well if the proposal distribution matches the target distribution/ posterior distribution to generate posterior samples. The details of the Metropolis-Hastings algorithm are discussed in [33].

The Metropolis-Hastings algorithm is infeasible to use for the Bayesian set-up discussed in Section 4.2.1 because of choosing the appropriate initial value is quite challenging, the low acceptance rate and slow mixing. For a non-linear mean function, it is inefficient to compute the conditional distribution for each parameter.

4.2.3.3 Gibbs Sampling

Gibbs sampling [43] is the special case of the Metropolis-Hastings algorithm for which the acceptance probability is always one. This sampling is based on the conditional distributions of priors rather than joint distribution. Gibbs sampling works at the two-conditional distribution of $\pi(x|y)$ and $\pi(y|x)$ rather than the joint distribution of $\pi(x, y)$. A detailed explanation of Gibbs sampling is discussed in [25, 43].

The correlated nature among the parameters of the Mitscherlich-Baule model also makes the use of Gibbs sampling impractical for our particular hierarchical Bayesian setup. So we need to look at another possible option to generate samples considering the parameters high dimensional and correlated nature.

4.2.3.4 Hamiltonian Monte Carlo within No-U-Turn Sampler

Hamiltonian Monte Carlo (HMC) [83, 126] is becoming a popular MCMC method due to its unique behaviour, which use a differential equation system to produce marginal variance to discard the random walk behavior and sensitivity to correlated parameters. It uses algorithm differentiation to get derivatives of every parameter. HMC calculates the gradient of the posterior distribution and can also simulate samples with a high acceptance rate and fewer iterations for convergence over broader range of parameter space. However, sometimes a poor parameter value selection drastically reduces the effectiveness of MCMC. For this problem, Hoffman and Gelman [100] introduced a No-U-Turn sampler within HMC using a recursive algorithm for the efficiency of HMC and implemented by the Stan language [112]. The details of this method is extensively discussed in [83, 100, 126].

4.2.4 Stage III: Model Selection and Validation

In this section, we introduce multiple model selection criteria to select the best fitting model. For this, we considered the Expected Log point-wise Predictive Density(ELPD) [91], Leave-One-Out cross-validation criterion (LOOIC) [116], and, Widely Applicable In-

formation Criterion (WAIC) [116] validation tools to evaluate and compare the Bayesian models.

Let us consider for our yield $Y_i = (Y_1, Y_2, ..., Y_n)$ given parameters $\mathcal{B} = (\beta_0, ..., \beta_4)$, the prior distribution $\pi(\mathcal{B})$ and the posterior distribution $\pi(\mathcal{B}|Y_i)$. The Expected Log point wise Predictive Density (ELPD) for n data points [116] can be expressed as ;

$$ELPD = \sum_{i=1}^{n} \int \pi(y^*) \log \pi(y^*|Y_i) dy^*, \tag{4.15}$$

where $\int \pi(y^*) \log \pi(y^*|Y_i) dy^*$ is the log predictive density for a new observation y^*, which can not be calculated directly due to its unknown feature but can be estimated using LOOIC and WAIC; the details are discussed as follows. In general, we would seek $\mathcal{B}$ parameters that would maximise the predictive density, thus the maximum value of ELPD is the best selection criterion. The log score $\log \pi(y^*|Y_i)$ used for determining the predictive density can be written as follows;

$$LPD = \log \pi(y^*|Y_i) = \int \pi(y^*|\mathcal{B}) \log \pi(\mathcal{B}|Y_i) d\mathcal{B}. \tag{4.16}$$

So, in the practice, the log point-wise density (LPD) can be calculated from the samples of the posterior distribution $\pi(\mathcal{B}|Y_i)$, and these samples can be denoted as $\mathcal{B}^k; k = 1, ..., K$. This too can be estimated by.

$$\widehat{LPD} = \sum_{i=1}^{n} \log \left(\frac{1}{K} \sum_{k=1}^{K} \pi(Y_i|\mathcal{B}^k) \right). \tag{4.17}$$

The Leave-One-Out (LOO) approach is one of the method to estimate ELPD. So the LOO estimate for the Bayesian model selection [116] is expressed as;

$$\widehat{ELPD}_{LOO} = \sum_{i=1}^{n} \log \pi(Y_i|Y_{-i}), \tag{4.18}$$

where $\pi(Y_i|Y_{-i}) = \int \pi(Y_i|\mathcal{B}) \pi(\mathcal{B}|Y_{-i}) d\mathcal{B}$ is the LOOIC density for given yields considering ith yields with $\pi(Y_i|\mathcal{B})$ is the likelihood and $\pi(\mathcal{B}|Y_{-i})$ is the posterior for the parameter vector $\mathcal{B}$.

WAIC [75] is another approach for estimating expected log point-wise predictive density and can be written as,

$$\widehat{ELPD}_{WAIC} = \widehat{LPD} - \widehat{\pi}_{WAIC}, \tag{4.19}$$

where $\widehat{\pi}_{WAIC}$ can be calculated from $\sum_{i=1}^{n} Var(\log \pi(Y_i|\mathcal{B}))$. So for WAIC and LOOIC, we can express them as follows;

$$\begin{aligned} WAIC &= -2\widehat{LPD} + 2\widehat{\pi}_{WAIC}, \\ LOOIC &= -2\widehat{ELPD}. \end{aligned} \tag{4.20}$$

The LOOIC and WAIC behave like the AIC criterion to minimise the information loss and select the best fitting model with minimum information loss. Due to this reason, the minimum value of WAIC and LOOIC are considered the best selection criterion.

4.3 Results of Bayesian Analysis

In this section, we perform the Bayesian analysis of the EPIC data corresponding to three crops Spring Barley, Winter Barley and Silage. For the continuous set-up, we have fixed one combination of factor variables steepness, soil and weather such that $St = 5,\ So = 6$, $Wy = 1$, and $Sy = 4$ which generate the data for the yield with the inputs N and P. Later on, we also demonstrate the Bayesian framework for all the factor combinations incorporating the soil, steepness and weather.

The Mistcherlich-Baule response function, upon which we base our model, has several characteristics which can be incorporated into the prior distributions for its parameters [62]. Namely, we know that: (i) β_0 is the maximum yield which should be positive and finite; (ii) the intercept, β_1, and slope, β_2, for the Nitrogen (N) response should have positive parameters; (iii) we expect similar properties from the Phosphorus (P) parameters β_3, and β_4. As all parameters are required to be positive, we adopt weak Gamma priors to enforce this positivity for the parameters $\beta_1, \beta_2, \beta_3$ and β_4. Specifically, we used the $\beta_0 \sim Ga(Y_{max}, 1)$ for the maximum yield parameter where Y_{max} is the observed maximum observed yield value for Spring Barley, which also means that $E(\beta_0) = Y_{max} = 4.68$.

We consider the priors of the remaining parameters subjectively. For our analysis, we choose $\beta_2 \sim Ga(1, 10)$ the slope prior corresponding to N. Based on the basic yield Figures in Chapter 3, we have seen an increasing trend for all the crops corresponding to N and a flat trend for P which motivates to assume the prior $\beta_4 \sim Ga(1, 500)$, where the mean is close to zero. For the intercepts of N and P, we assume that $\beta_1,\ \beta_3 \sim G(5, 1)$ as a weak priors around a mean of 5 to give same weight to the slopes. Finally, our error term variance parameter σ was given a $Ga(1, 100)$ relatively weak prior.

To sample the posterior distributions of these parameters, we used Hamiltonian Monte Carlo-NUTS using RStan [112]. For the Hamiltonian Monte Carlo-NUTS, we perform 10000 iterations with four different chains and discard 5000 burn-in samples for each chain as a default set-up of RStan.

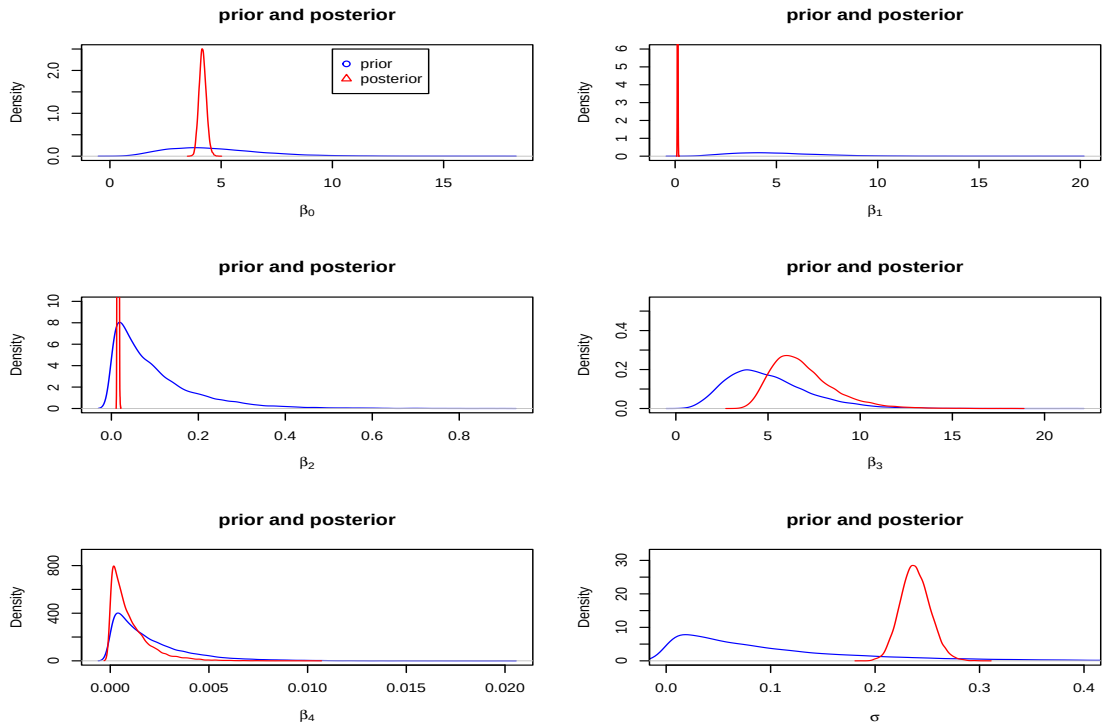


Figure 4.1: Prior and Posterior Plots for the Hyperparameters.

Figure 4.1 shows the prior and posterior density plot for the parameters of the Bayesian hierarchical model. The prior and posterior density plots show strong learning from the data concerning the input N for the parameters β_0 and β_2. However, we have seen reasonable learning for the β_1 such that the posterior mean is not close to prior mean. We can not see any substantial change for the parameters β_3 and β_4 concerning Phosphorus input such that prior and posterior are very close to each other. The error variance σ is also showing strong learning and we can reveal a good understanding from the data for the parameters β_0, β_2 and σ.

In Table 4.1, we present the summary of these posterior simulations, giving the mean values for the fitted coefficients with 2.5%, and 97.5% posterior credible intervals from the Bayesian model fitting for Spring Barley, Winter Barley and Silage simulations.

The N coefficients for all crops indicate a positive response to N of yield. Still, the Phosphorus coefficient β_4 estimate is close to zero for all crops showing a negligible reaction to P. We also tabulate the effective sample size, n_{eff}, which varies considerably with a good portion of accepted posterior samples. We also evaluate $\hat{R}$, the potential scale reduction factor, also known as Gelman-Rubin statistic [91] to summarise each chain in the sampler. This statistic has the property that values between 1.00 to 1.01, indicating

that our chains are essentially indistinguishable from one another, suggesting no evidence of lack of convergence. For the chains corresponding to the results in Table 4.1, all the $\hat{R}$ values lie between 1.00 to 1.01, suggesting a suitable level of convergence had been achieved. For four parallel chains, it is recommended that n_{eff} should be at least 400 to make $\hat{R}$ useful for the HMC-NUTS [135]. So, we can see that our effective sample sizes are far higher than the recommended size, which indicates the MCMC simulation is working effectively.

Table 4.1: Posterior Sample Summary Statistics ($1.0 \leqslant \hat{R} < 1.02$)

Crops	Coefficients	Mean	2.5%	97.5%	n_{eff}
Spring Barley	β_0	4.172	3.871	4.502	7472
	β_1	0.125	0.101	0.153	14163
	β_2	0.015	0.013	0.018	7713
	β_3	6.900	4.412	11.213	11925
	β_4	0.001	0.00002	0.004	17539
	σ	0.238	0.212	0.267	9571
Winter Barley	β_0	4.824	4.539	5.117	6371
	β_1	0.031	0.013	0.053	15204
	β_2	0.022	0.019	0.025	6426
	β_3	7.005	4.857	10.080	11495
	β_4	0.0002	0.000004	0.0007	18526
	σ	0.336	0.294	0.384	7538
Silage	β_0	17.269	17.057	17.499	6417
	β_1	0.838	0.825	0.853	7462
	β_2	0.016	0.014	0.017	6804
	β_3	13.009	6.954	20.917	13449
	β_4	0.005	0.0001	0.018	17839
	σ	0.127	0.113	0.144	13390

Diagnostic trace plots for the crops Spring Barley, Winter Barley and Silage with four different chains are shown in Figure 4.2, 4.5 and 4.6, indicating generally good mixing and no signs of lack of convergence of the chains. From the pairs plot in Figure 4.3 for the crop

Spring Barley, we can see some parameters are weakly correlated. Others are negatively correlated, such as β_0 with β_2. For the autocorrelation plot in Figure 4.4, all parameters show a decreasing trend with the increase of the lags, indicating good mixing of chains. The parameter β_4 showed a weakly correlated relationship with all other parameters. In further diagnostics of auto-correlations, pairs plots for all the crops showed no features of concern.

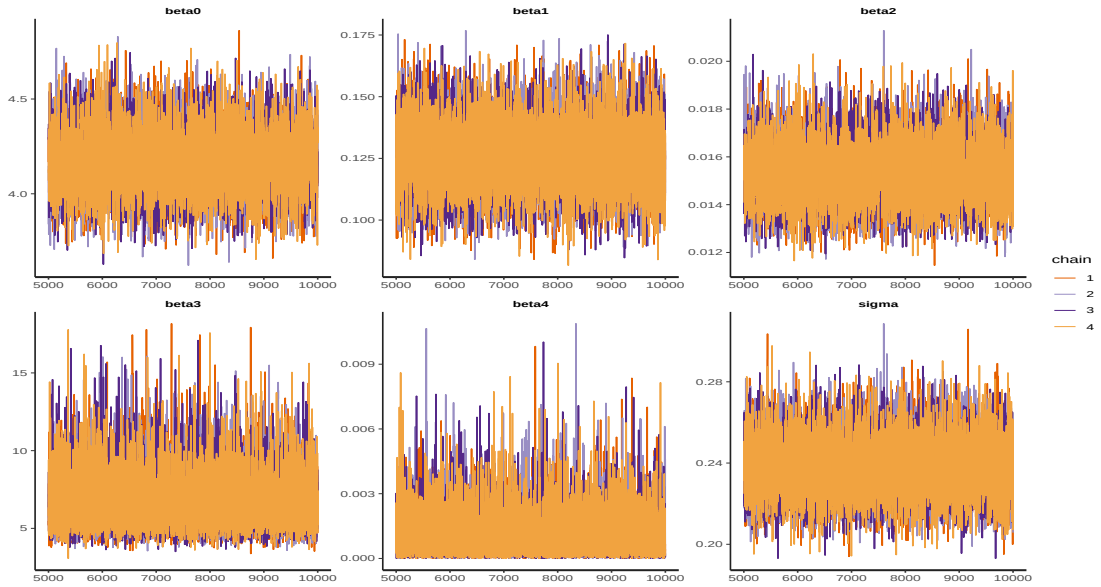


Figure 4.2: Trace plot for the Crop Spring Barley.

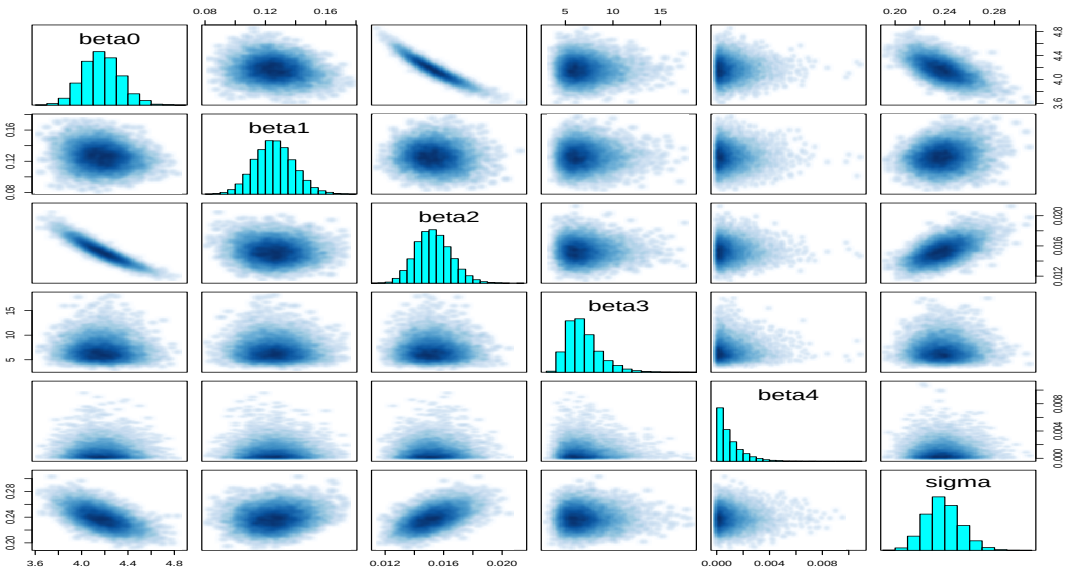


Figure 4.3: Pairs plot for the crop Spring Barley.

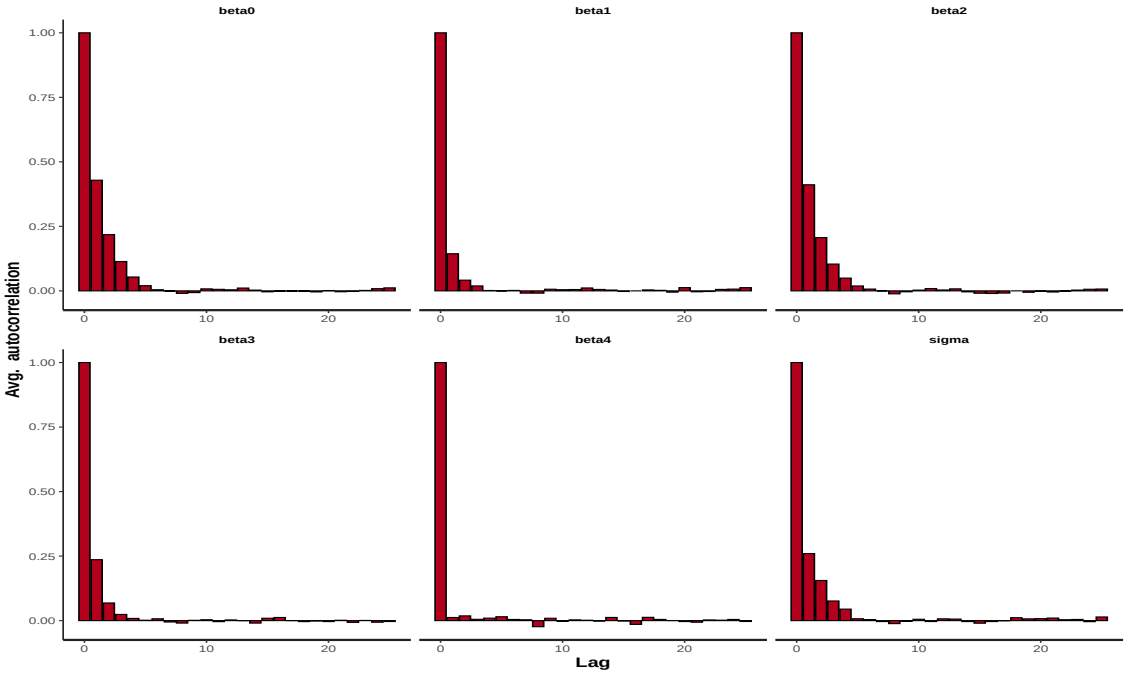


Figure 4.4: Autocorrelation Diagnostic Plot for the Crop Spring Barley.

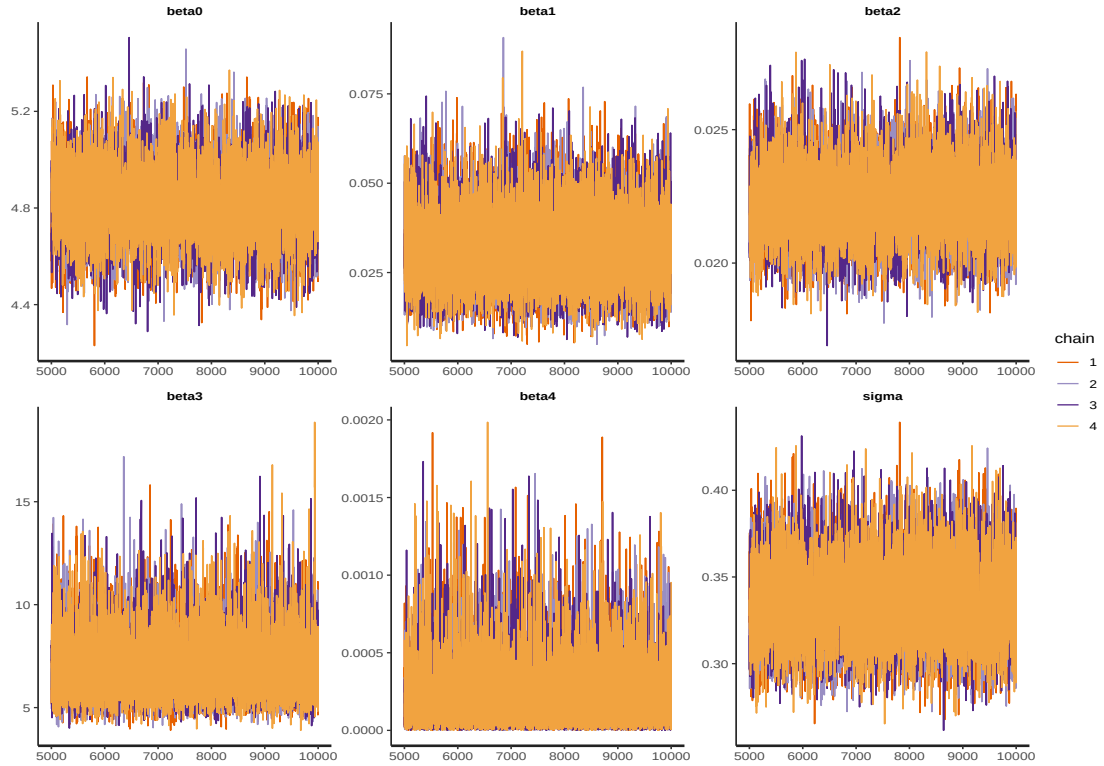


Figure 4.5: Trace Plot for the Crop Winter Barley.

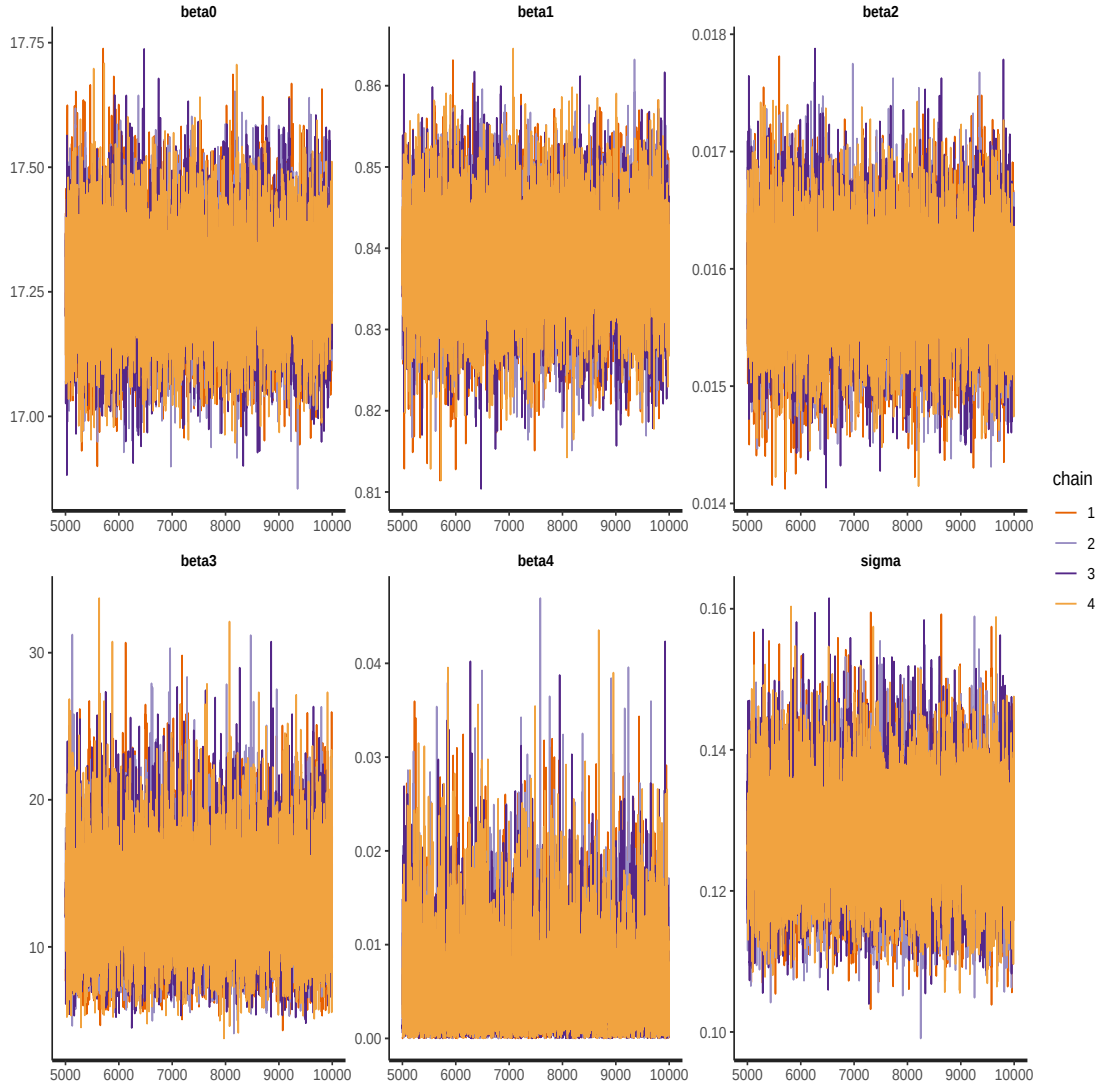


Figure 4.6: Trace Plot for the Crop Silage.

Synthesising these results, in Figure 4.7, we plot the yield data (black circles) alongside the point predictions (green circles), predicted yield curve (black line), 95% credible intervals (blue line) for the mean function only and 95% predicted credible intervals (red line) shows the credible intervals corresponding to the yield Y. We note that the simulated data lie within the predicted credible intervals with narrow uncertainty.

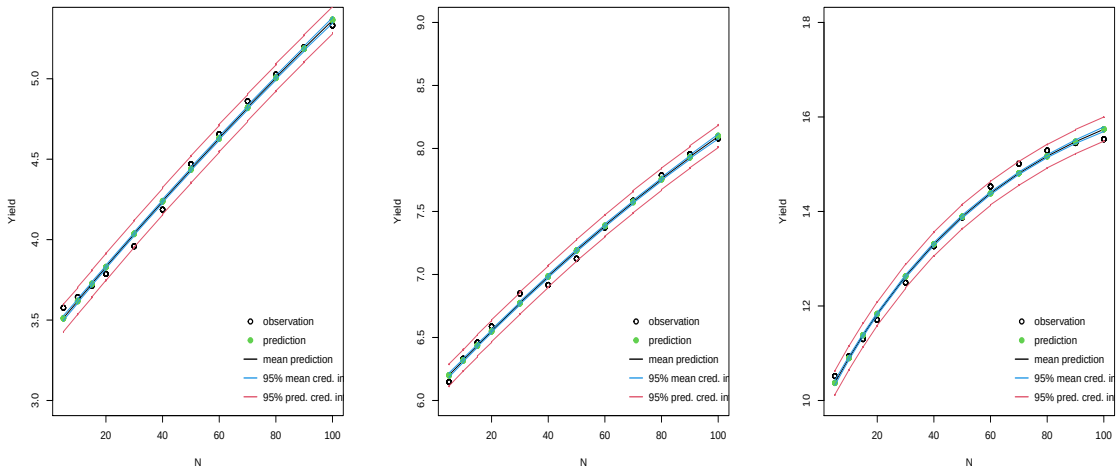


Figure 4.7: Non-linear Bayesian Model Fitting for the Crops Spring Barley (right); Winter barley (middle); Silage (right).

4.3.1 Model Comparison and Validation

Not all crops are expected to respond to both inputs, so we use model selection tools to identify the appropriate model in terms of the fertiliser response. It is reasonable to consider whether retaining Phosphorous in the yield model is meaningful after observing such weak dependency. Given the results observed above, where the influence of Phosphorous in the model appeared negligible, we now move to model comparison. We considered three possible models for yield response: (i) Nitrogen only, (ii) Phosphorus only, and (iii) both Nitrogen and Phosphorus giving the following mean functions.

$$\begin{aligned} \mu_i^N &= \beta_0[1 - \exp(-\beta_1 - \beta_2 N_i)], \\ \mu_i^P &= \beta_0[1 - \exp(-\beta_3 - \beta_4 P_i)], \\ \mu_i^{NP} &= \beta_0[1 - \exp(-\beta_1 - \beta_2 N_i)][1 - \exp(-\beta_3 - \beta_4 P_i)]. \end{aligned} \tag{4.21}$$

To select between the models, we apply the criteria of Subsection 4.2.4. Evaluating the ELPD, LOOIC, and WAIC statistics gives the results in Table 4.2, where we see that every crop favours the N only model highlighted in red. We observe that the values of ELPD are maximised for the N-only model and minimised for the LOOIC and WAIC as desired. So we remove P from the model for further analysis.

Table 4.2: Bayesian Model Comparison Results

Crops	Models	ELPD	LOOIC	WAIC
Spring Barley	N	251.1	-502.2	-502.3
	P	-135.6	271.2	271
	$N+P$	225.4	-450.8	-497.5
Winter Barley	N	246.1	-492.2	-492.3
	P	-140.6	281.2	280
	$N+P$	219.7	-439.4	-481.5
Silage	N	91.2	-182.4	-182.34
	P	-290.9	581.8	581.71
	$N+P$	89.1	-178.2	-180.7

4.4 Results of Incorporation of a Factor Variable

We now apply the approach described in Section 4.2.2 with an N-only model as suggested by the model selection above in Table 4.2. Our data set has three different factor inputs: steepness, soil, and weather. We have seen the same result pattern of a solid response to Nitrogen and a weak response to Phosphorus for all the crops for the Bayesian model with continuous inputs. Now, we explore including categorical variables in the model for the Spring Barley crop only.

We consider the effect of including a single factor variable into the model, as in Equation (4.14). In each model, we take the first-factor level as the baseline value giving up to seven additional parameters depending on the factor variable denoted by the vector Z_i^T representing the deviation of the maximum yield from this baseline value. We note that the three-factor variables are such that we have four levels of steepness, three levels of soil, and eight levels of weather as discussed in Chapter 2. The priors for γ_0, β_1 and β_2 the same as the priors of β_0, β_1, β_2 from the continuous set up. The factor effect Z_i^T in Equation (4.13) can be positive or negative (to allow the negative effect of factor effects). Without precise information on the model's behaviour under the different factor levels, Normal independent priors were adopted for each component of Z_i^T with mean 0

and variance 1.

In Table 4.3, we present the summary statistics for the Spring Barley crop using N input only without the factors. The Nitrogen coefficients indicate a strong positive response to Nitrogen of yield. The $\hat{R}$ statistic means that our chains are mixing well such that there is no evidence of lack of convergence. For the chains corresponding to the results in Table 4.3, all the $\hat{R}$ values lie between 1.00 to 1.01, suggesting a suitable level of convergence had been achieved.

Table 4.3: Posterior Sample Summary Statistics for Spring Barley N Only Model ($1.0 \leqslant \hat{R} < 1.01$)

Crops	Coefficients	Mean	2.5%	97.5%	n_{eff}
Spring Barley	β_0	4.164	3.869	4.483	6092
	β_1	0.126	0.102	3.152	9261
	β_2	0.015	0.013	0.018	6224
	σ	0.238	0.212	0.267	7119

Now, we consider the models of Spring Barley yield which include each one of the categorical variables. Applying the same Hamiltonian Monte Carlo-NUTS approach used previously for the continuous inputs, we obtained the summary statistics given in Table 4.4 from our factor posterior simulations. Introducing the factors to the model has increased the overall acceptance rate compared to Table 4.3. In diagnostics, all of the $\hat{R}$ values lie between 1.00 to 1.01, suggesting no evidence of lack of convergence of our posterior samples. The sum of γ_0, and $\gamma_{k,i}$ is greater than zero for all three factors such that maximum yield coefficient β_0 is positive.

In general, most of the estimated values for the effects of different levels of the factors are far from zero except the weather factor level Z_6^W, indicating a negative departure from the baseline maximum yield level. Additionally, we observe that the posterior estimates of the error variance parameter have increased from 0.238 to 0.26 for the steepness factor, 0.34 for soil factor, and 0.36 for weather factor, after adding the factors into our modelling. This indicates a broader uncertainty due to the additional variability and increasing number of parameters through the expanded model. We have seen a broader uncertainty for the factor of weather due to more variation of the simulations.

Table 4.4: Posterior Sample Summary Statistics for N-only Spring Barley Model, each Including a Single Factor Input ($1.0 \leqslant \hat{R} < 1.01$)

Factor	Coefficients	Mean	2.5%	97.5%	n_{eff}
Steepness Only	γ_0	4.18	3.97	4.40	7137
	β_1	0.12	0.10	0.13	14104
	β_2	0.014	0.012	0.016	8017
	Z_2^{St}	0.21	0.10	0.33	10154
	Z_3^{St}	0.18	0.07	0.30	10587
	Z_4^{St}	0.14	0.03	0.26	10282
	σ	0.26	0.24	0.28	8862
Soil Only	γ_0	4.15	3.90	4.42	7707
	β_1	0.15	0.14	0.17	13300
	β_2	0.013	0.011	0.015	7280
	Z_2^{So}	0.94	0.79	1.11	10543
	Z_3^{So}	1.91	1.73	2.09	9674
	σ	0.34	0.33	0.38	9810
Weather Only	γ_0	4.15	3.96	4.35	6388
	β_1	0.05	0.04	0.06	9756
	β_2	0.013	0.012	0.014	5461
	Z_2^{W}	2.0021	1.82	2.19	5603
	Z_3^{W}	2.70	2.51	2.89	5236
	Z_4^{W}	3.71	3.50	3.92	4906
	Z_5^{W}	0.46	0.40	0.64	6309
	Z_6^{W}	2.48	2.29	2.67	5599
	Z_7^{W}	-0.28	-0.46	-0.10	6472
	Z_8^{W}	4.86	4.63	5.10	4985
	σ	0.36	0.35	0.39	9044

The estimated value of the coefficients β_1 is similar for the factors steepness and soil, which is expected. However, introducing the factor weather has decreased the estimated value of β_1 from the baseline model values from Table 4.3. The estimate for β_2 remained

moreover similar after factoring compared to the base result in Table 4.3. Compared to the baseline N-only model, the coefficient for the maximum yield γ_0 is approximately equal to β_0.

The trace plots in Figures 4.8, 4.10, and 4.9 of the model fitting for the factor soil, weather, and steepness still show good mixing of the chains for all the coefficients. Additionally, an inspection of auto-correlation plots 4.12 showed that not much information is lost because the factors' thinning and posterior density plots 4.11 also show normality, which is the expected attribute for the density plot diagnostic. The equivalent figures for the other factors are showing similar results, see Appendix A.3.

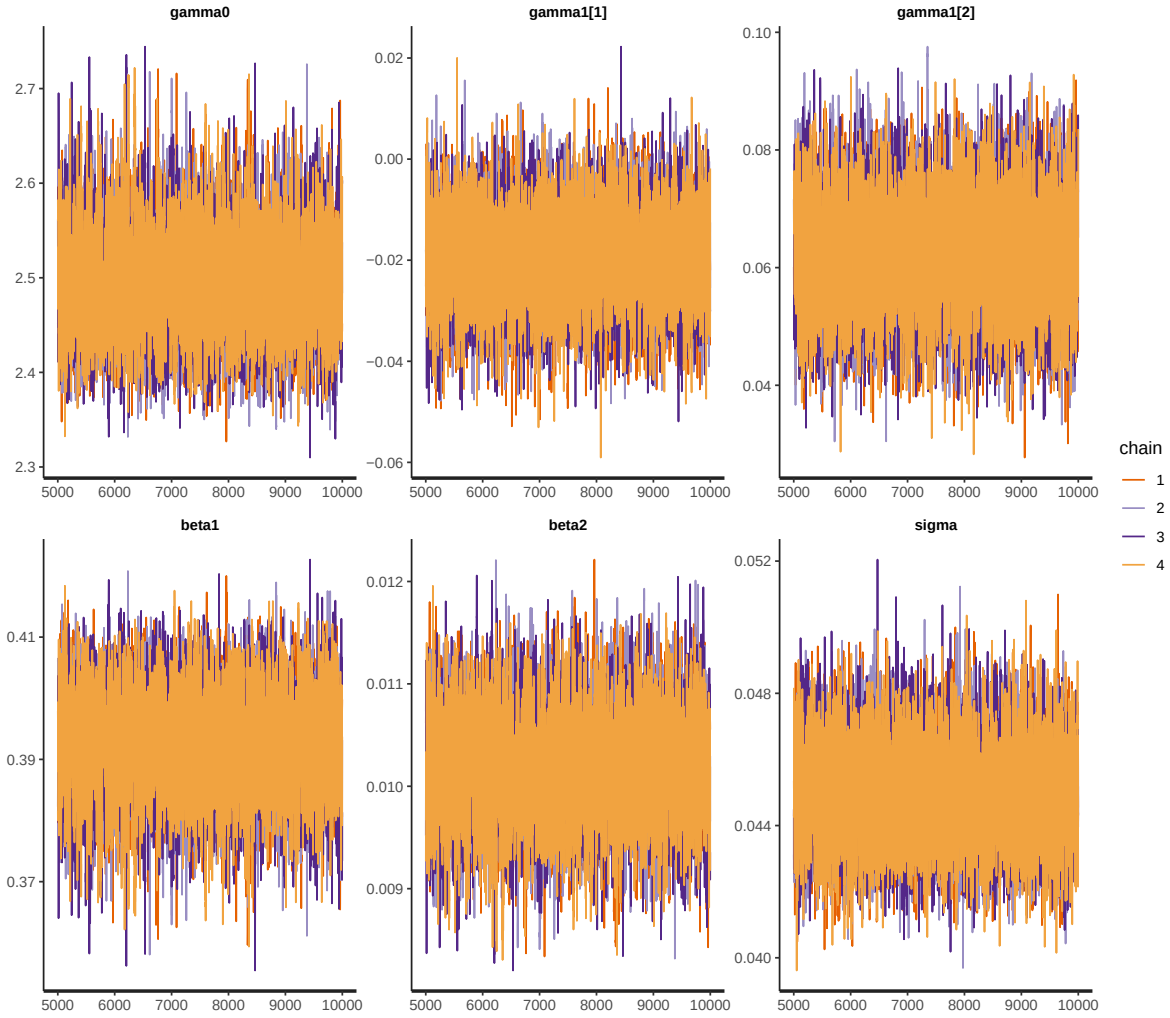


Figure 4.8: Trace plot for the Factor Effect Considering Soil

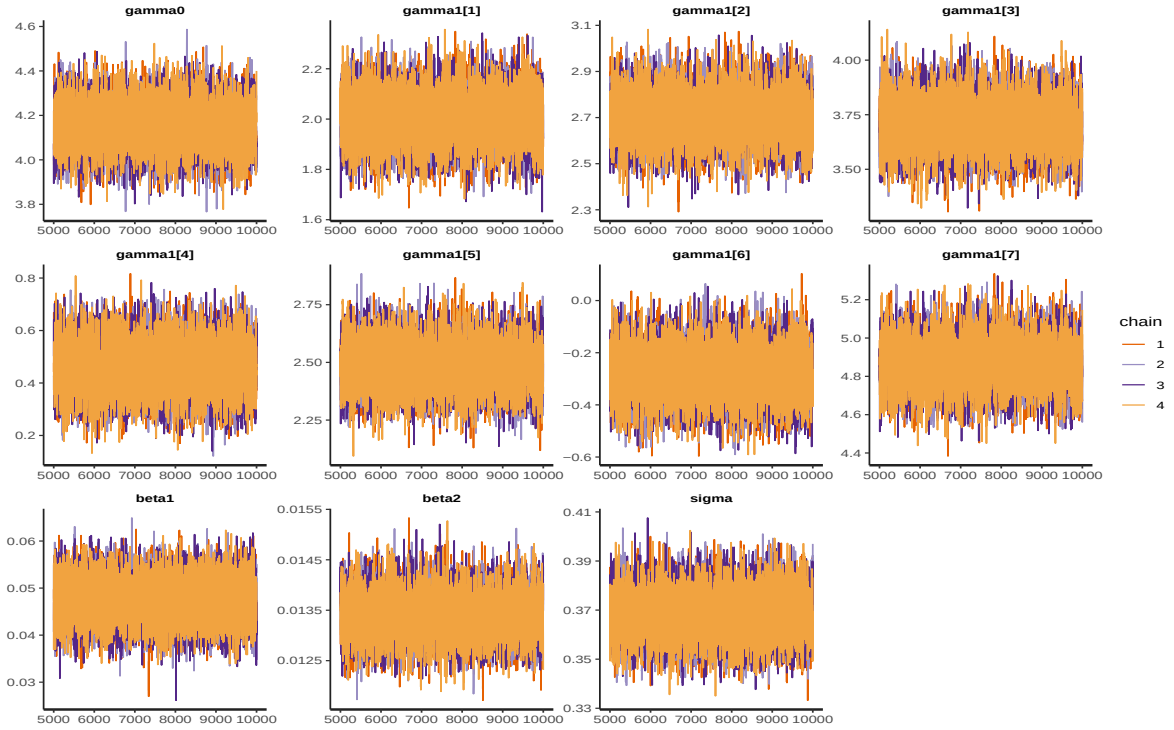


Figure 4.9: Trace plot for the Factor Effect Considering Weather

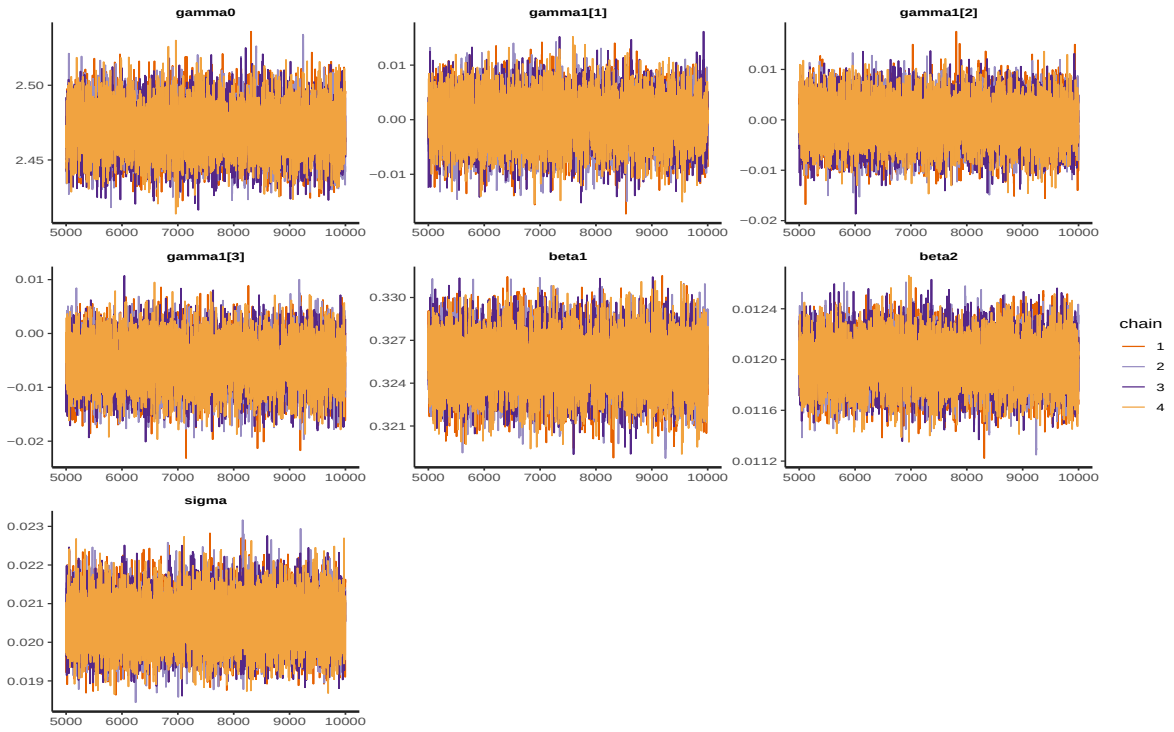


Figure 4.10: Trace plot for the Factor Effect Considering Steepness

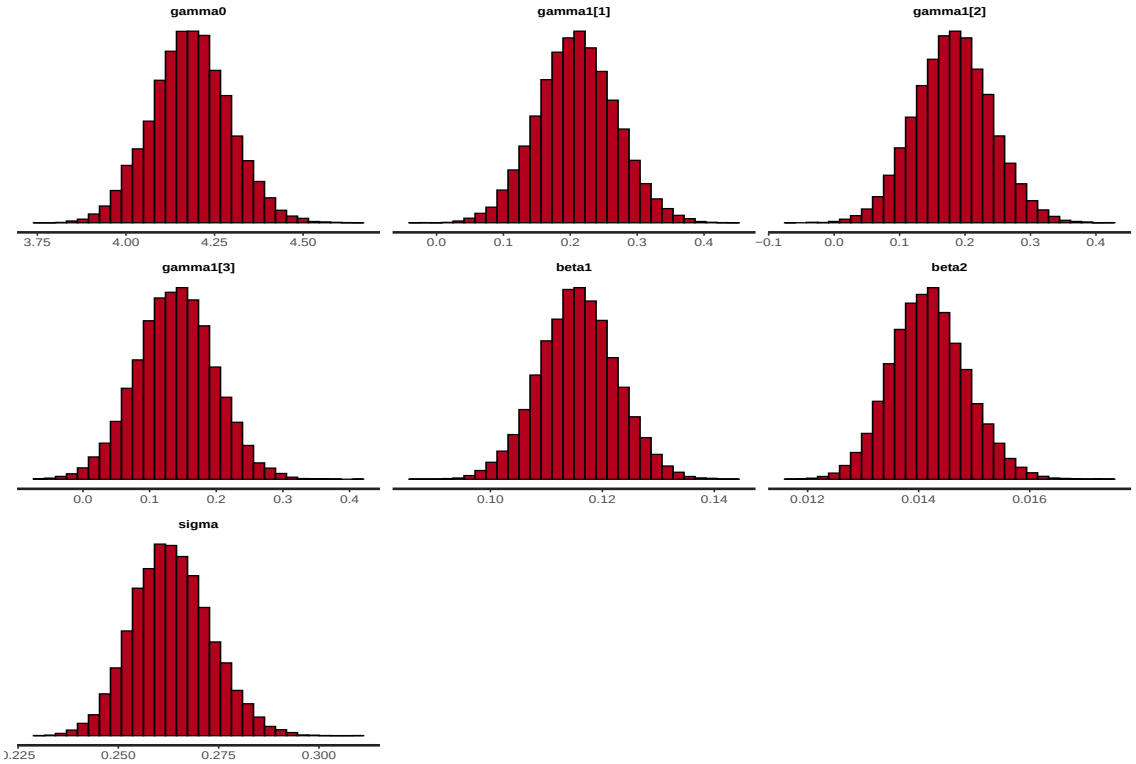


Figure 4.11: Posterior density plot using the Single Factor Steepness.

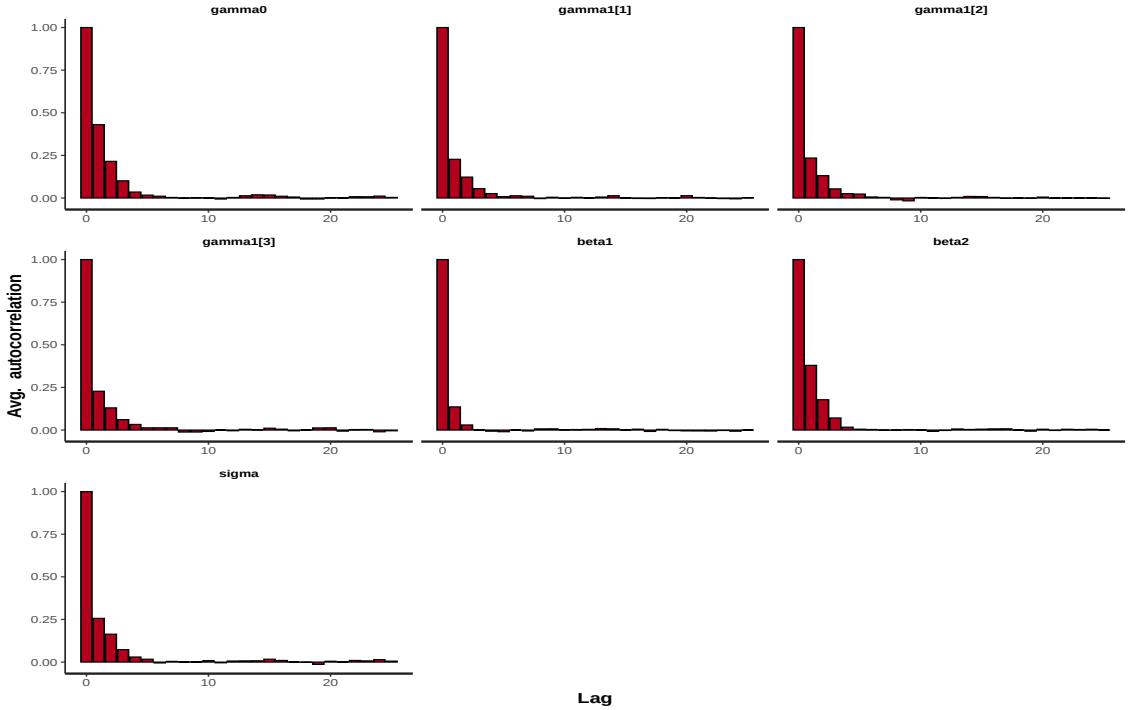
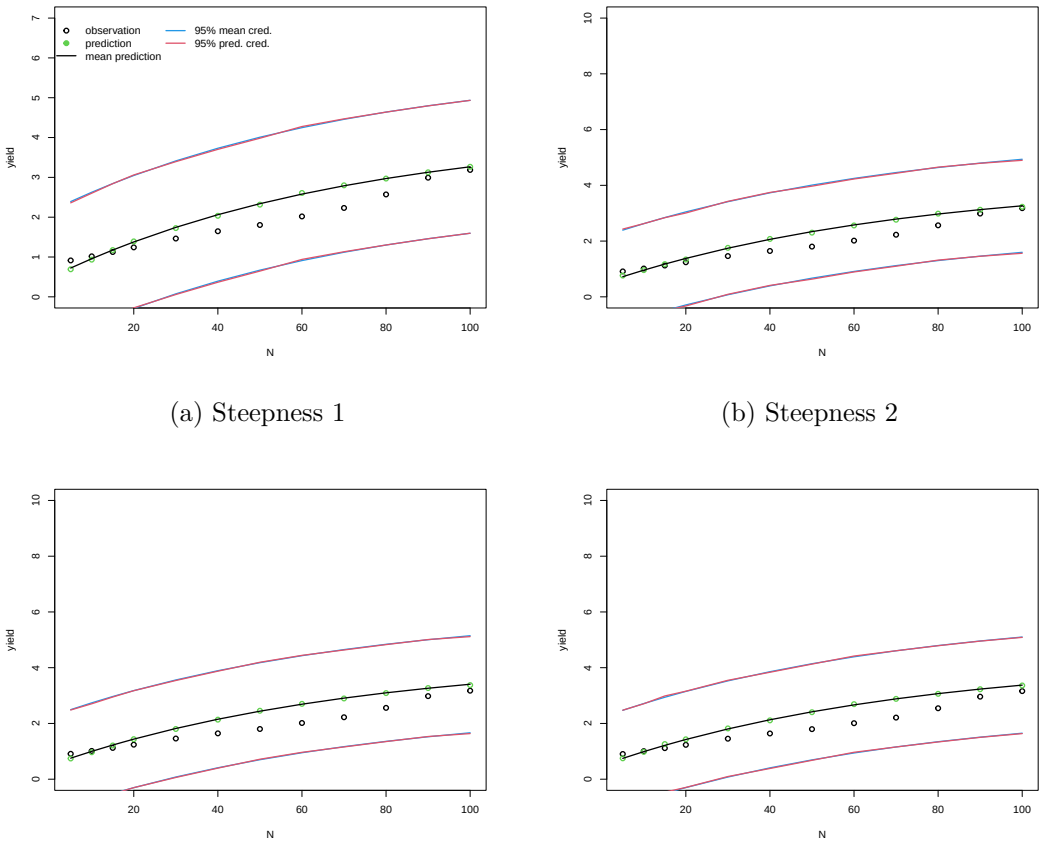


Figure 4.12: Autocorrelation Diagnostic Plot for the Factor Steepness.

We plot the observed yield data (black circles) alongside the point predictions (green circles), predicted yield curve (black line), 95% mean credible intervals (blue line), and 95% predicted credible intervals (red line). All the simulated data lie within the mean and predicted credible intervals.

Generating plots of the predicted yield as a function of N in Figure 4.13 including the steepness factor, we see the baseline (Steepness 1), level 2 (Steepness 2), level 3 (Steepness 3) and level 4 (Steepness 4). All the steepness factor levels shows a wider uncertainty and we can also see that the models with factor effects all show the same trend, which is fixed by the model, and the estimated posterior values for β_1, β_2, and γ_0 are approximately similar. The last three levels of the steepness factor are showing the same shape and uncertainty, indicates little variation among the factor levels.



(a) Steepness 1

(b) Steepness 2

(c) Steepness 3

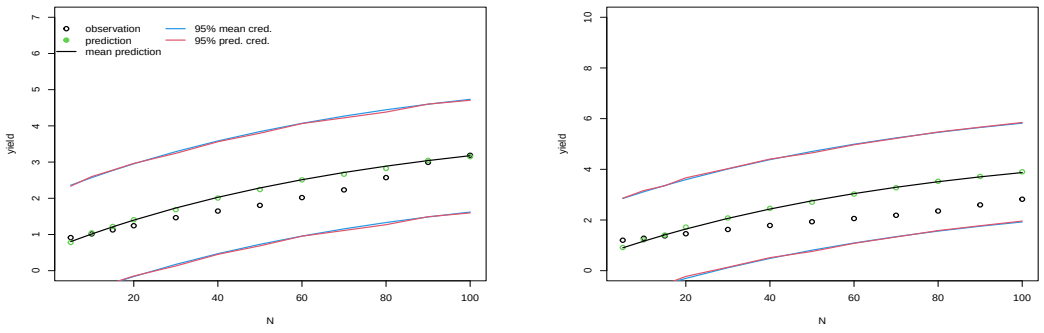
(d) Steepness 4

Figure 4.13: Prediction Plot for the Factor Steepness

Figure 4.14 also shows the prediction plot for model including the factor soil. This figure shows a similar trend for the baseline and soil level 2 with broad uncertainty. For the

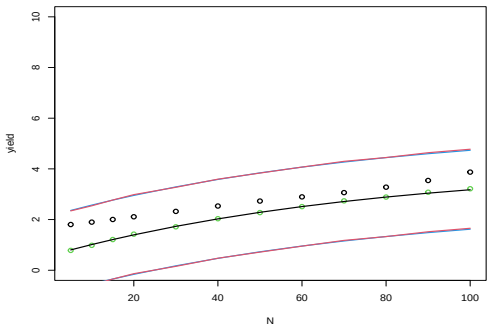
third level, we can see a comparatively a narrow uncertainty boundary and the simulated values and predicted values lie within the intervals indicating well-behaved and consistent predictions. Figure 4.15 shows the prediction plot for the factor weather, which also yields good and consistent predictions. But compared to the other factor inputs, it offers more variability among the factor levels. We can see that factor levels weather 4, 5 and weather 7 fit closely with slightly less uncertainty compared to the others levels. The shape of the observations for factor levels 6 and 8 deviates from the MB model. For the large N values baseline level is showing the growth plateau, as well as for levels 2 and 3.

The figures are showing under-confidence and bias. The failure of MB model to capture the trend of the simulated data creates the bias and under-confidence. However, for the factor inputs model, we have seen generally good and consistent fitting for all the curves with an increase in variance. However, a potential improvement could be made by introducing a further factor effect to modify the β_1 parameter to permit different strengths of the relationship between yield and N at the different categorical levels.



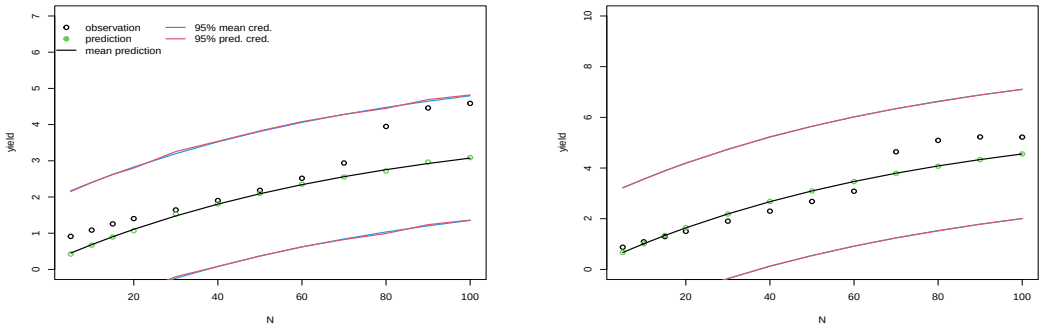
(a) Soil 1

(b) Soil 2



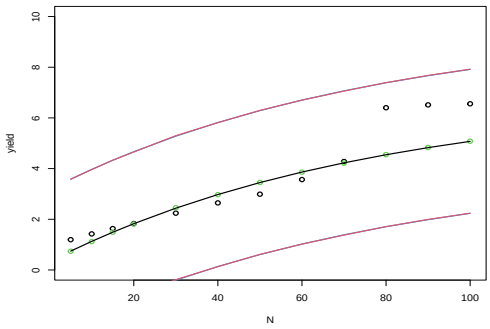
(c) Soil 3

Figure 4.14: Prediction Plot for the Factor Soil

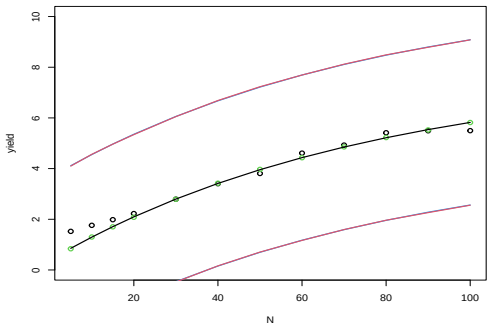


(a) Weather 1

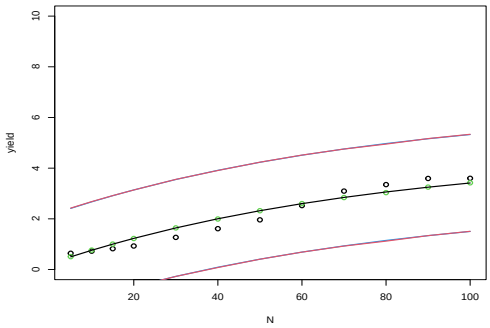
(b) Weather 2



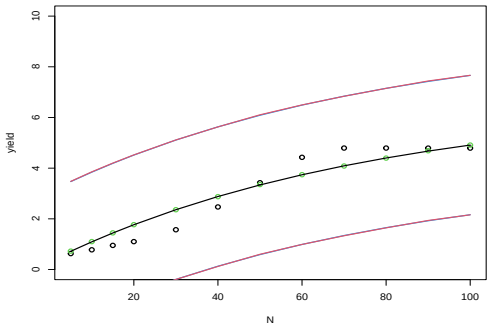
(c) Weather 3



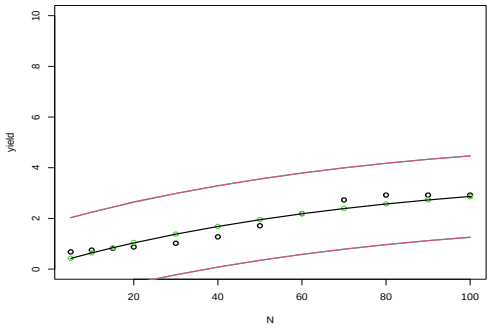
(d) Weather 4



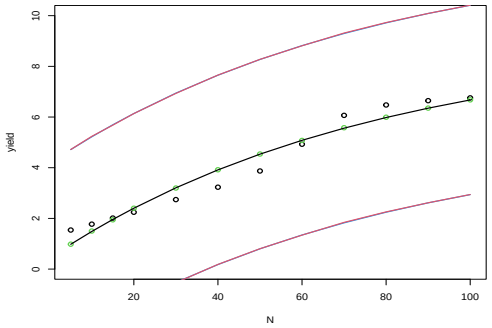
(e) Weather 5



(f) Weather 6



(g) Weather 7



(h) Weather 8

Figure 4.15: Prediction Plot for the Factor Weather

4.5 Concluding Remarks

The purpose of chapter 4 is to demonstrate Bayesian hierarchical modelling for the non-linear Mistcherlic Baule model to crop yield. This chapter also illustrated this analysis by comparing and validating the model using different modern diagnostic tools. Finally, this chapter also incorporated factor variables into the Bayesian analysis.

From this chapter, we have learned some significant features of modern Bayesian hierarchical modelling. We learned to use the non-linear function as a mean and about the mathematical expression of the prior, posterior and likelihood functions. We utilised the modern MCMC algorithm of HMC-NUTS. We gathered knowledge about the diagnostics tools and their use in Bayesian inference. Finally, we experienced the challenge of factor incorporation into Bayesian inference such as under-confidence and biased behaviour motivated us to explore different method.

A Bayesian hierarchical model has proven to be a valuable and effective tool for modelling the behaviour of the critical quantity of interest in our agricultural application. The fully-Bayesian approach allowed for appropriate modelling and capture of the uncertainties in the problem. However, without solid prior information to inform the process, we prefer a less computationally intensive approach that could deliver results of similar quality. For example, a Bayes linear [59] emulator would capture much of the uncertainty without requiring intensive simulation.

Chapter 5

Emulation Approaches For Quantitative Inputs

5.1 Introduction

Complex computer models are usually used to represent physical systems. The evaluation of this whole process is highly complex, contains too many parameters and requires a lot of time [121, 139] to complete. So we often need a simple surrogate of this complex system to assess efficiently. An alternative option to understand real-life applications is to use the emulation technique, which mimics the behaviour of these complex computer experiments.

This chapter starts with the context of emulation by introducing the concept of a simulator, emulator, and the basic idea of emulation in Section 5.2. In Section 5.3, we provide the general structure of an emulator with variance and correlation specifications and finally show the general outline of emulation approaches such as Gaussian process emulation and Bayes linear emulation. The construction of the emulator using the Bayes linear emulation technique for continuous inputs is illustrated in Sections 5.4 and 5.5 with maximum likelihood inference and emulator diagnostics. The Bayes linear method is demonstrated initially using a 1-D example and then using the EPIC simulator data of 2-D continuous inputs shown in Section 5.6. Finally, the chapter ends with concluding remarks in Section 5.7.

5.2 Context of Emulation

In this section, we discuss the context of emulation starting with the basic definition of a simulator and emulator. We also discuss the general idea of emulation with its necessity for the complex physical system.

5.2.1 Simulator

A simulator is the computer code used to describe a complex physical system or process. A simulator aims to generalise the physical process, such as which input settings can be used to model the real-life application to get the maximum output. A simulator tends to have many combinations of inputs and outputs, which lie in a high dimensional space. This calculation can require many thousands of lines of code to simulate the whole physical system. The simulator is also used to compare observed data to simulated data, and the formation for this naturally requires a fully Bayesian approach [120]. Let us consider the following equation of a simulation process,

$$y = F(x), \tag{5.1}$$

where y is the simulator output; x is the input(s) and F is the simulator.

Let us consider an one-dimensional $(1 - D)$ example;

$$F(x) = \frac{1}{2}x + \cos(x) + \log(3x).$$

We consider $F(x)$ as the simulator's function for the input x. We evaluate our function $F(x)$ for $n = 5$ data points such that $x = (1, 2, 3, 4, 5)$. The simulator output for the inputs x is; $F(x) = (1.09, 0.25, -0.68, 0.38, 3.27)$. Figure 5.1 is the basic 1-dimensional simulator plot for five input points. We have five different simulator outputs for five different inputs from 1 to 5. This problem is simple to generate, but most real simulators of physical systems are high-dimensional, with thousands of data points, making them infeasible and challenging to formulate.

The simulator $F(x)$ in Figure 5.1 has only limited evaluations for those five data points (blue). However, without evaluating the model, we do not know the behaviour of $F(x)$ between those five data points. So we must introduce an approach representing the unknown behaviour of $F(x)$ away from the evaluations.

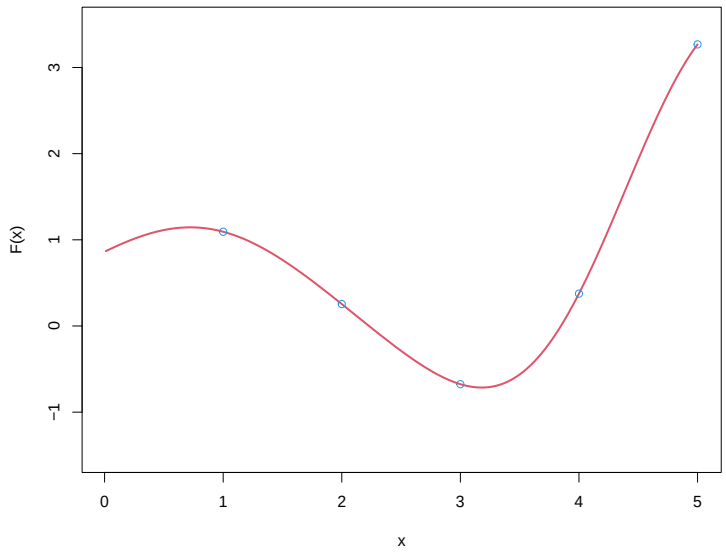


Figure 5.1: Plot for a Simple 1-D Simulator

5.2.2 Emulator

An emulator is a stochastic representation of a simulator that is used to represent the uncertainty in the behaviour of $F(x)$ for any space of input x [56]. The emulation process can be written as the following Equation;

$$y^{'} = f(x), \tag{5.2}$$

where $y^{'}$ is the emulator output; x is the input and f is the emulator, which is different to the simulator function F. Emulators are generally fast approximations, generating the predicted value for any given input with the related uncertainty. The simulator is certain (only for deterministic simulators) but the emulator represents uncertain knowledge about the simulator. The mean of $f(x)$ is usually interpolated between F evaluations, and the variance of $f(x)$ is used to represent the uncertainty in the simulator values for inputs x.

5.2.3 General Idea of Emulation

The main concept of emulation is to use the small number of simulations to gain an understanding of the simulator behaviour and to use this learning to predict over input space without re-running the model [59]. Statistical emulation is able to catch the complexity of the simulator without knowing the underlying details of the physical system, and can be

used to formulate model simplification, optimisation, and calibration [99]. Emulation has been used in many fields such as climate change [95], system biology [121], hydrocarbons reservoir [66], ocean carbon cycle [140], energy system [144], galaxy formation [71], COVID model [137] and so on. This study explores the use of emulation in agricultural research from an environmental complex computer simulator [134].

5.2.4 Necessity of Emulation

The use of a complex computer model or simulator is becoming more popular to model and understand the behaviour of real-life applications or physcial systems. However, it is not a simple task to model that physcial system due to some obstacles. Statistical emulation is a solution to those obstacles, which is highlighted as follows;

- The simulator represents the behaviour of y over the entire input space, but it is not easy to explore the whole range of input variables due to the model's complexity and computational expanse. Emulation techniques can represent the behaviour of the entire input space of the computer model, and are simple and feasible to specify.
- For a complex model, the simulator can be defined over a high dimensional space, requiring substantial computational resources to complete the simulation. However, the emulation technique is a fast approximation for the same input space.
- With the complex nature of a simulator, we also need to deal with the uncertainties such as input uncertainty, observational uncertainty, model discrepancy and so on. This uncertainty analysis requires many simulator evaluations [127]; thus, we need emulators to include these uncertainties.

5.3 Basic Structure and Approaches of Emulation

This section is about the basic structure of an emulator, different correlation functions and emulation approaches. This section also discusses the various forms of correlation functions and illustrates the two emulation techniques: Gaussian Process Emulation and Bayes Linear Emulation. Finally, we suggest some reasons to choose one for the rest of the analysis.

5.3.1 General Structure of an Emulator

An emulator, meta model, or surrogate model is a fast approximation of the complex computer model [56, 59, 74]. It is the approximation of the simulator, F, and constructed by a set of training points say $f(X_p) = \{f(x^{(1)}), ..., f(x^{(n)})\}$ over the input space of X given by $X_p = \{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$, also known as design points.
The general form of an emulator $f(x)$ for inputs x can be expressed as;

$$\begin{aligned} f(x) &= g(x)^T\beta + u(x), \\ &= \sum_{j=1}^{p} g_j(x)^T\beta_j + u(x). \end{aligned} \tag{5.3}$$

Equation (5.3) consists of two terms and the first term $\sum_{j=1}^{p} g_j(x)^T\beta_j$ represents the mean function in regression form expressed in terms of input variables. This is the product of basis functions $g_j(x) = \Big(g_1(x), ..., g_p(x)\Big)$, a vector of known functions, and the parameters $\beta_j = (\beta_1, ..., \beta_p)$ are unknown regression coefficients where $j = 1, 2, ..., p$.

The second term $u(x)$ is the residual process, which is a zero mean weakly stationary process to explain the additional variation around the mean function in terms of input x. The emulator residual process $u(x)$ often follows the Gaussian form of covariance, with $Cov[x, x^{'}] = \sigma^2 R(x - x^{'})$, where R is the function of covariance, $sigma^2$ is error variance and x, $x^{'}$ are two distinct inputs.

5.3.2 Active Variables and Nugget

The form of Equation (5.3) can be modified by restricting the input variables x to be the active variables, x_A, which are those inputs which are very influential for the emulation process [66, 74]. Most complex computer simulators require a high dimensional space with many parameters; under this situation, a portion of input variables x is used as active inputs x_A, which can explain the majority of the simulator variability. By considering the problem in this way, high dimensional complex simulators can be modelled with substantially simpler models, increasing the emulator's performance [56, 66, 120, 121]. We can modify the general form of the emulator as follows;

$$f(x) = g(x_A)^T\beta + u(x_A) + \nu(x), \tag{5.4}$$

where $g(x_A)^T\beta$ is the mean function in regression form for the active inputs x_A only; $u(x_A)$ is the residual variance with zero mean and Gaussian covariance for the active

inputs x_A. The final term $\nu(x)$ is called the nugget, has a zero mean and variance σ_ν^2 such that $Cov[\nu(x), \nu(x^{'})] = \sigma_\nu^2$ for $x = x^{'}$ and 0 otherwise.

In emulation, an active variable has the most impact on the simulator. The active variable only affects the mean function $g(x_A)^T\beta$, and the residual process $u(x_A)$. There are many ways to select the active variables, and model selection criteria are mostly used to determine the best inputs. On the other hand, the inactive variables, considered less necessary and sometimes referred to as the noise variables, are not used to model fit as the training points.

5.3.3 Variance Specification and Correlation Functions

We require a covariance function for the residual process of $u(x)$. A common choice of covariance function for the inputs over the output components is the following structure;

$$Cov\Big[u(x), u(x^{'})\Big] = \sigma^2 R(x, x^{'}), \tag{5.5}$$

where $R(x, x^{'})$ is the correlation between the inputs x and $x^{'}$ and σ^2 is the variance of the residual between simulator outputs. Suppose the input values are close to each other, then we would expect the values of u to show a high correlation over the input space, and if the inputs are far apart then we would expect a low correlation between the process values at inputs x and $x^{'}$. There are many forms of correlations for $R(x, x^{'})$, including: Exponential Kernel, Squared Exponential Kernel, and Matérn Kernel.

One of the simple forms of the covariance function for the residual process $u(x)$ is the exponential kernel which can be expressed as;

$$R(x, x^{'}) = \exp\Big[-\theta|x - x^{'}|\Big], \tag{5.6}$$

where $\theta > 0$ is the correlation length. The size of the correlation length determines the magnitude of the correlation for points of a fixed distance $|x - x^{'}|$. The exponential kernel is positive and semi-definite and for the smoothing feature, this kernel is suitable due to its infinitely differentiable properties. This method is sensitive to the correlation length such that it will produce large values for a smaller length choice.

A common choice of correlation for the residual process $u(x)$ is the squared exponential kernel which can be expressed as;

$$R(x, x^{'}) = \exp\Big[-\theta(x - x^{'})^2\Big]. \tag{5.7}$$

The properties of the exponential kernel and the squared exponential kernel are moreover similar but the latter kernel shows more smoothness. This kernel is also sensitive to correlation length, θ.

The Matérn covariance is a covariance function between the inputs when they are stationary only, and the basic form can be expressed as;

$$R(x, x^{'}) = \frac{2^{1-s}}{\Gamma(s)}\left(\frac{\sqrt{2s}|x - x^{'}|^{s}}{\theta}\right)K_s\left(\frac{\sqrt{2s}|x - x^{'}|}{\theta}\right), \tag{5.8}$$

where θ is the length parameter, s is a positive power parameter, and K_s is the modified 2nd-order Bassel function. The smoothness of the process while using the Matérn function depends on s, and it has three forms of set-up for $s = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$, which are as follows;

$$R^{\frac{1}{2}}(x, x^{'}) = \exp\left(-\frac{|x - x^{'}|}{\theta}\right), \tag{5.9a}$$

$$R^{\frac{3}{2}}(x, x^{'}) = \left(1 + \frac{\sqrt{3}|x - x^{'}|}{\theta}\right)\exp\left(-\frac{\sqrt{3}|x - x^{'}|}{\theta}\right), \tag{5.9b}$$

$$R^{\frac{5}{2}}(x, x^{'}) = \left(1 + \frac{\sqrt{5}|x - x^{'}|}{\theta} + \frac{5(x - x^{'})^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}|x - x^{'}|}{\theta}\right), \tag{5.9c}$$

For Matérn $s = \frac{1}{2}$ the Equation (5.8) shows a similar form to the exponential kernel. This kernel is very effective for unknown smoothness. But this method is very sensitive to the parameters s, and θ, so proper selection of these parameters is crucial.

5.3.4 Approaches for Emulation

The general emulator form given by (5.3) can be used for all emulation problems. For representing uncertainty, there are two main approaches: the Bayes linear emulation approach and the Gaussian process emulation approach.

5.3.4.1 Gaussian Process Emulation

A fully Bayesian emulation process requires the joint probability distribution for all uncertain quantities. One common emulation preference is to use a Gaussian process emulator (GPE). A Gaussian process (GP) [4, 55] is a stochastic process, say X_a, of a collection of random variables indexed by time or space a, such that for every finite set of elements $a_1, a_2, ..., a_n$ we can write as;

$$X_{a_1,...,a_n} = \left[X_{a_1}, ..., X_{a_n}\right], \tag{5.10}$$

where each collection $X_{a_1}, ..., X_{a_n}$ follows the normal distribution and together have a joint distribution which follows the multivariate normal distribution.

Let us assume that a Gaussian process emulator f takes inputs of $x_1, x_2, ..., x_n$ and then the GP can be expressed as;

$$f \sim GP\Big(\mu(x), \Sigma(x, x^{'})\Big), \tag{5.11}$$

where $\mu(x)$ is the mean function, which could be a first-order polynomial function of input variables $\mu(x) = g(x)^T\beta$, and the term Σ is the covariance function and can be expressed as follows;

$$\Sigma(x, x^{'}) = \sigma^2[R(x, x^{'}, \theta)], \tag{5.12}$$

where σ^2 is the error variance, $R(x, x^{'})$ is the correlation function, and θ is the input's correlation length. The GP in Equation (5.11) can be expressed as the following compact matrix form as $G = (g_1(x), ..., g_p(x))$ and $\beta = (\beta_1, ..., \beta_p)$;

$$f \sim GP\Big(G\beta, \sigma^2[R(x, x^{'}, \theta)]\Big), \tag{5.13}$$

where G is the matrix of basis function for regressors and β is the vector of regression coefficients. So we need to model these parameters of $\beta,\ \sigma^2,\ \theta$ by using a fully Bayesian approach. To model these hyperparameters, we need to consider a suitable prior and then combine with the multivariate Gaussian likelihood to obtain the posterior distribution. To determine the posterior distribution, we often require MCMC methods.

So, for the output of design points $f(x) = \{f(x^{(1)}), ..., f(x^{(n)})\}$ for a univariate emulator f, the Gaussian process $f|f(x), \mu(x), \Sigma(x, x^{'}) \sim GP\Big(\mu^*(x), \Sigma^*(x, x^{'})\Big)$, where $\mu^*(x)$ is the posterior mean and $\Sigma^*(x, x^{'})$ is the posterior variance. So the posterior mean and variance can be expressed as [142];

$$\mu^*(x) = \mu(x) + Cov[f(x^{'}), f(x)]Var[f(x)]^{-1}\Big(f(x) - E[f(x)]\Big), \tag{5.14a}$$

$$\Sigma^*(x, x^{'}) = \Sigma(x, x^{'}) - Cov[f(x^{'}), f(x)]Var[f(x)]^{-1}Cov[f(x),\ f(x^{'})], \tag{5.14b}$$

where $Cov[f(x),\ f(x^{'})]$ is the covariance; $E(f(x^{'}))$ is the mean for testing design points $f(x^{'})$ and $E(f(x))$ is the mean for training points.

5.3.4.2 Bayes Linear Emulation Approach

Bayes linear emulation is a different emulation approach which can be used to assess the relationship between the input and output of the simulator in terms of the adjusted mean

and variance instead of the probability distribution [36, 59]. For the uncertain, complex computer output, Bayes linear emulation can be viewed as a simple belief specification and a fast approximation, which only needs the specifications for the expectations and variances rather than the full probability distribution.

Bayes linear emulation is usually applied for complex physical systems when it requires fewer model evaluations and precise specifications. It is widely used in different fields of study. This approach has been used in the system biology model of hormonal cross-talk in Arabidopsis Thailana [121, 139], human skin risk assessment [92], hydrocarbon reservoirs [36, 66], energy system models [122], galaxy formation [71], and water research [27].

Bayes linear inference depends on expectations and measures of dispersion, which are primitive, and its update is performed by orthogonal projection [59]. Bayes linear emulation [59] provides a new form of diagnostic tools and removes the need of full Bayesian probability distributions. For example, if we have two sets of random quantities, say $P = (P_1, P_2, ..., P_i)$ and $Q = (Q_1, ..., Q_j)$, Bayes linear analysis updates the subjective belief of P, given the observation Q. At first, we need to specify the prior mean $E(P)$ and variance $Var(P)$, and the prior mean $E(Q)$ and variance $Var(Q)$. The covariance for the vectors P and Q can be expressed as $Cov(P, Q)$. The Bayes linear update for the vector of observations P given Q can be expressed as,

$$E_Q[P] = E[P] + Cov[P, Q]Var[Q]^{-1}(Q - E(Q)), \tag{5.15a}$$

$$Var_Q[P] = Var[P] - Cov[P, Q]Var[Q]^{-1}Cov[Q, P], \tag{5.15b}$$

$$Cov_Q[P_1, P_2] = Cov[P_1, P_2] - Cov[P_1, Q]Var[Q]^{-1}Cov[Q, P_2], \tag{5.15c}$$

where $E_Q[P]$ and $Var_Q[P]$ are the adjusted expectation and variance for the observation of P given Q. The $Cov_Q[P_1, P_2]$ is the covariance of the sub-collection P_1 and P_2 of the observation P. So, considering the concept of training $f(x) = \{f(x^{(1)}), ..., f(x^{(n)})\}$ and testing data, $f(x^{'})$ for a univariate emulator, the Bayes linear update can be written similarly as (5.15);

$$E_{f(x)}[f(x^{'})] = E[f(x^{'})] + Cov[f(x^{'}), f(x)]Var[f(x)]^{-1}(f(x) - E[f(x)]) \tag{5.16a}$$

$$Var_{f(x)}[f(x^{'})] = Var[f(x^{'})] - Cov[f(x^{'}), f(x)]Var[f(x)]^{-1}Cov[f(x), f(x^{'})]. \tag{5.16b}$$

5.3.4.3 Choice of Bayes Linear or Gaussian Process Emulation

The updated emulator equations for both the Gaussian process (5.14) and Bayes linear emulation (5.16) are the same except for the probabilistic assumptions of the Gaussian

process. In this thesis, we prefer the Bayes linear emulation approach to the fully Bayesian Gaussian process approach for reasons as follows:

1. We must specify the full joint distribution for all random quantities in a fully Bayesian approach. It needs the entire probability distribution for all the random variables for the GP emulation; on the other hand, Bayes linear emulation approach does not require any distribution for the random variables, but it does need all expectations and variances.

2. In a fully Bayesian approach, one of the concerns is to specify the meaningful form of the likelihood and prior distribution. However, the Bayes linear approach does not require taking any form of the likelihood, which makes this process simpler to evaluate.

3. For the higher dimensional problems, the likelihood function for fully Bayesian framework calculations can be challenging to work with, for example, due to the surface having many peaks, needing to optimise and being high dimensional. This causes problems in finding the posterior, but Bayes linear approach doesn't have these issues.

4. If the fully Bayesian approach doesn't yield an analytic joint posterior distribution, then MCMC methods are required to get the posterior samples. MCMC methods are sometimes problematic due to a lengthy process to execute due to a high number of iterations to converge, failure to mix of the chains, problems with the correlated inputs, high dimensional parameters space and unsatisfactory autocorrelation plots. However, the Bayes linear approach is free from this computational problem.

While the Bayes linear (BL) method has advantages over the GP method, the GP emulator can give credible intervals due to its probabilistic nature, as well as the mean and variance. Often, the choice between GP and BL emulator depends on the form of the problem and philosophical view of analysis [68, 74, 78]. Suppose a problem needs a fully probabilistic distribution; in that case, the natural choice is the GP emulator. For our situation, we will use the Bayes linear emulation approach to avoid the burden of computational issues, the absence of meaningful prior knowledge, and the non-conjugate form of the posterior distribution created in Chapter 4.

5.4 Construction of Emulators for Continuous Inputs

From the general emulation function, we have seen the first part is the multiplication of basis function $g(x)^T$ and parameters β, and the second part is about the residual process $u(x)$, which needs the covariance structure from the kernels involving two hyperparameters σ^2 and θ. Given a lack of useful information about β, σ^2, θ, and the non-linear way in which (σ^2, θ) appear in the emulator, we will seek to estimate these quantities by maximum likelihood. Given that large volume of available data, the MLE and Bayes estimates will be sufficiently close that the impact will be negligible.

5.4.1 Maximum Likelihood Inference for the Parameters

The form of the emulator for the single model can be written as;

$$f(x) = g(x)^T\beta + u(x). \tag{5.17}$$

The first part of the above equation can be expressed in matrix form $G = (g_1(x), g_2(x), ..., g_p(x))^T$ and $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$; p is the total number of regression parameters. In the emulator, most of the response variables' structural behaviour is captured by the trend component, with the reminder explained by the residual GP. Assume that the emulator takes the form, $f(x) \sim (G\beta, \sigma^2 R_\theta)$. Then the likelihood function can be written as,

$$\begin{aligned} L(\beta, \sigma^2, \theta | f(x)) &= \frac{1}{\sqrt{(2\pi\sigma^2)^n |R_\theta|}} \exp\left[-\frac{1}{2\sigma^2}(f(x) - G\beta)^T R_\theta^{-1}(f(x) - G\beta)\right], \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}|R_\theta|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(f(x) - G\beta)^T R_\theta^{-1}(f(x) - G\beta)\right]. \end{aligned} \tag{5.18}$$

Now taking logarithmic transformation on Equation (5.18) gives;

$$LL(\beta, \sigma^2, \theta | f(x)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln|R_\theta| - \frac{1}{2\sigma^2}(f(x) - G\beta)^T R_\theta^{-1}(f(x) - G\beta). \tag{5.19}$$

With negative log likelihood:

$$\begin{aligned} NL(\beta, \sigma^2, \theta | f(x)) &= -1 \times LL(\beta, \sigma^2, \theta | f(x)) \\ &= c + \frac{n}{2}\ln(\sigma^2) + \frac{1}{2}\ln|R_\theta| \\ &\quad + \frac{1}{2\sigma^2}(f(x) - G\beta)^T R_\theta^{-1}(f(x) - G\beta), \end{aligned} \tag{5.20}$$

where $c = \frac{n}{2}\ln(2\pi)$ is a constant.

Theorem 5.4.1. *If $f(x) \sim MVN(G\beta, \sigma^2 R_\theta)$, then the estimates of the parameters are $\hat{\beta} = (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} f(x)$ and $\widehat{\sigma}^2 = \frac{1}{n}(f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta})$ respectively where θ is assumed known, which is generalized least squares problem [3, 8].*

Proof. Ignoring the common factor of $-\frac{1}{2}$ from Equation (5.19) and considering the first term as constant gives us the above-simplified likelihood equation in Equation (5.20). Now the work is to estimate the parameters β and σ^2 using the generalized least squares method. So differentiating Equation (5.20) for β and equating zero gives us the estimate for the β coefficient.

From Equation (5.20) and using the gradient of quadratic matrix equation, we can write that;

$$\begin{aligned}
\frac{\partial NL}{\partial \beta} &= (f(x) - G\beta)^T R^{-1}(f(x) - G\beta) = 0, \\
&= -(f(x) - G\beta)^T R_\theta^{-1} G - (f(x) - G\beta)^T R_\theta^{-1} G, \\
&= -2(f(x) - G\hat{\beta})^T R_\theta^{-1} G.
\end{aligned}$$

For two invertible matrices of same order we can write $(G\beta)^T = \beta^T G^T$ and setting $\frac{\partial NL}{\partial \beta} = 0$ at $\beta = \hat{\beta}$:

$$\begin{aligned}
-2(f(x) - G\hat{\beta})^T R_\theta^{-1} G &= 0, \\
-(f(x) - G\hat{\beta})^T R_\theta^{-1} G &= 0, \\
-f(x)^T R_\theta^{-1} G + G^T \hat{\beta}^T R_\theta^{-1} G &= 0, \\
G^T \hat{\beta}^T R_\theta^{-1} G &= f(x)^T R_\theta^{-1} G, \\
(G\hat{\beta})^T R_\theta^{-1} G &= f(x)^T R_\theta^{-1} G, \\
\hat{\beta}^T G^T R_\theta^{-1} G &= f(x)^T R_\theta^{-1} G, \\
\hat{\beta}^T &= f(x)^T R_\theta^{-1} G (G^T R_\theta^{-1} G)^{-1}, \\
\hat{\beta} &= \left[(G^T R_\theta^{-1} G)^{-1}\right]^T (f(x)^T R_\theta^{-1} G)^T, \\
\hat{\beta} &= (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} f(x).
\end{aligned}$$

Now again, for Equation (5.20) make the substitution for the precision $\phi = \frac{1}{\sigma^2}$, differentiate with respect to ϕ and equate to zero, which gives us the estimate for ϕ and hence σ^2. From Equation (5.20) we can write that;

$$NL(\beta, \phi, \theta) = -c - \frac{n}{2}\ln\Big(\frac{1}{\phi}\Big) - \frac{1}{2}\ln|R_\theta| - \frac{1}{2}\phi(f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta}),$$

$$= c + \frac{n}{2}\ln(\phi) - \frac{1}{2}\ln|R_\theta| - \frac{1}{2}\ln|R_\theta| - \frac{1}{2}\phi(f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta}).$$

Now setting $\frac{\partial NL(\beta,\phi,\theta|f(x))}{\partial\phi} = 0$ at $\phi = \hat{\phi}$ and then substituting in variance of the MLE where $\widehat{\phi} = \frac{1}{\widehat{\sigma}^2}$:

$$\frac{n}{\widehat{\phi}} - (f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta}) = 0,$$

$$n\widehat{\sigma}^2 = (f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta}),$$

$$\widehat{\sigma}^2 = \frac{1}{n}(f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta}).$$

It is noted that both $\hat{\beta}$ and $\widehat{\sigma}^2$ depends on θ. □

Theorem 5.4.2. *If the emulator $f(x) = G\beta + V$, where $V = u(x_1, ..., x_p)^T$ with residual variance V is $\sigma^2 R_\theta$ then $Var(\hat{\beta}) = \widehat{\sigma}^2(G^T R_\theta^{-1} G)^{-1}$ [3, 8] also depend on θ.*

Proof. The variance $Var(\hat{\beta})$ can be calculated from the following equation;

$$Var(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]. \tag{5.21}$$

Considering the estimate of $\hat{\beta}$ from Theorem 5.4.1;

$$\begin{aligned}
\hat{\beta} &= (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} f(x), \\
&= (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1}(G\beta + V), \\
&= (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} G\beta + (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} V, \\
&= \beta + (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} V.
\end{aligned} \tag{5.22}$$

Now we put the value of $\hat{\beta}$ from Equation (5.22) into Equation (5.21). Since $E(V) = 0$ then $E(VV^T) = \sigma^2 R_\theta$, and also by symmetry $E(VV^T)^T = E(VV^T)$. Now fixing $\sigma^2 = \widehat{\sigma}^2$, and also using the symmetry of R_θ to get $(R_\theta^{-1})^T = R_\theta^T$ we have;

$$\begin{aligned}
Var(\hat{\beta}) &= E\Big[(\beta + (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} V - \beta)(\beta + (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} V - \beta)^T\Big], \\
&= E\Big[\{(G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} V\}\{(G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} V\}^T\Big], \\
&= (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} E(VV^T) R_\theta^{-1} G(G^T R_\theta^{-1} G)^{-1}, \\
&= E(VV^T) R_\theta^{-1}(G^T R_\theta^{-1} G)^{-1}(G^T R_\theta^{-1} G)(G^T R_\theta^{-1} G)^{-1}, \\
&= \widehat{\sigma}^2 R_\theta R_\theta^{-1}(G^T R_\theta^{-1} G)^{-1}, \\
&= \widehat{\sigma}^2(G^T R_\theta^{-1} G)^{-1}.
\end{aligned}$$

□

The overall negative log-likelihood function can be written in a simplified way by introducing the estimate of $\widehat{\sigma}^2$ into Equation (5.19) which gives;

$$\begin{aligned} NL &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\widehat{\sigma}^2) - \frac{1}{2}\ln|R_\theta| - \frac{1}{2\sigma^2}(f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta}), \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\widehat{\sigma}^2) - \frac{1}{2}\ln|R_\theta| - \frac{1}{2\frac{1}{n}(f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta})} \\ &\quad (f(x) - G\hat{\beta})^T R_\theta^{-1}(f(x) - G\hat{\beta}), \\ &= -\frac{n}{2}\ln(2\pi\widehat{\sigma}^2) - \frac{1}{2}\ln|R_\theta| - \frac{n}{2}. \end{aligned} \tag{5.23}$$

where $\widehat{\sigma}^2$ and R_θ depends on the correlation parameters θ. Equation (5.23) can be used as the objective function in the minimisation problem and can be solved to find the maximum likelihood estimates for $\widehat{\sigma}^2,\ \hat{\beta},\ \hat{\theta}$.

$$\begin{aligned} \hat{\theta}\left[\widehat{\sigma}^2, \hat{\beta}\right] = \min_{\theta} \quad & \left[-\frac{n}{2}\ln(2\pi\hat{\sigma}^2) - \frac{1}{2}\ln|R_\theta| - \frac{n}{2}\right]. \\ \text{Subject to} \quad & \theta_i \geqslant 0; i = 1, ..., p \end{aligned} \tag{5.24}$$

5.4.2 Introducing Nugget Effect on Optimisation

Introducing the nugget term affects the emulator interpolation for noisy data by imposing a small positive quantity as a diagonal covariance matrix element. Given that for a single x, $Var[u(x)] = \sigma^2$, the incorporation of an independent nugget term $\nu(x)$ of size σ_v^2 would mean that,

$$\begin{aligned} \sigma^2 = Var(u(x)) = \Sigma &= Var[u(x_A)] + Var[\nu(x)], \\ &= \sigma_A^2 + \sigma_\nu^2. \end{aligned} \tag{5.25}$$

Thus a convenient parameterisation is to consider σ_ν^2 as a function of δ and σ^2, so from Equation (5.25) we can write;

$$\sigma^2 = (1 - \delta)\sigma^2 + \delta\sigma^2, \tag{5.26}$$

Wherever $x \neq x^{'}$, we would also have $Var[u(x), u(x^{'})] = (1 - \delta)\sigma^2 R_\theta(x, x^{'})$ So, now our emulator of $f(x)$ will follow a multivariate normal distribution with mean $G\beta$ and variance-covariance $\sigma^2 R_{[\theta,\delta]}$. So all the maximum likelihood estimate equations will be changed from R_θ to $R_{[\theta,\delta]}$ to include the nugget effect. And the new objective function with the nugget effect can be expressed as similar way as in Equation (5.24).

5.4.3 Algorithm to Estimate Correlation Parameters

We have showed the Equations to estimate the coefficients and the objective function with the nugget effect in the previous Sections. Now, we need to follow certain steps to estimate the correlation parameters and nuggets. Initially, we need to set up a non-linear function like Equation (5.24) and then use it as an optimisation function with the relevant algorithms to estimate the parameters. A nice way to present this is to use an algorithm which is given as follows;

Algorithm 1 Estimation process for the MLE and correlation parameters

1: Calculate the Gaussian correlation matrix $R_{[\theta,\delta]}$ with the nugget effect.
2: Calculate the inverse of the $R_{[\theta,\delta]}$.
3: Calculate the determinant $R_{[\theta,\delta]}$.
4: Calculate $\hat{\beta} = (G^T R^{-1}_{[\theta,\delta]} G)^{-1} G^T R^{-1}_{[\theta,\delta]} f(x)$.
5: Calculate $\widehat{\sigma}^2 = \frac{1}{n}(f(x) - G\hat{\beta})^T R^{-1}_{[\theta,\delta]}(f(x) - G\hat{\beta})$
6: Calculate $Var(\hat{\beta}) = \widehat{\sigma}^2 (G^T R^{-1}_{[\theta,\delta]} G)^{-1}$.
7: Use an optimisation technique within the constraints for θ and δ, to give us the estimate for the correlation parameters and nugget term.

5.5 Construction of the Bayes Linear Emulator

5.5.1 Emulator Prior Specifications

To fit the emulator $f(x)$ for complex computer experiments, we need some specifications, known as prior specifications, before fitting the emulator. From the updated Equations (5.16), we have the terms $E[f(x)]$, $E[f(x^{'})]$, $Var[f(x^{'})]^{-1}$, $Var[f(x)], Cov[f(x^{'}), f(x)]$, to specify before the evaluation of the emulator. These terms are known as the prior specifications. These specifications are needed to update the adjusted mean and variance of the Bayes linear emulation. The important ingredients are as follows;

- Basis functions $[g(x)]$: The basis functions are crucial to represent the behaviour of the simulator mean. There are many ways to select the basis functions, such as model selection criterion, or to assess them from emulator diagnostic tools, such as standard prediction errors, resolution, etc. These can also be selected [82] by expert

knowledge and elicitation, or for extensive model evaluations $g(x)$ may follow the form of polynomials, Fourier functions and so on. One can also explore the data to suggest the appropriate functions to use as the basis function.

- The prior mean, variance and covariance (between each pair) for the regression parameters $(\beta_1, ..., \beta_p)$. The MLE estimate for these parameters can be estimated from Algorithm 1 using the training data. After estimating these parameters, we can multiply them by the basis functions to calculate the prior mean for the testing and training data. It is noted that $\widehat{\beta}, \widehat{\sigma}^2, \widehat{\theta}$ are fixed and known at MLE values. More details about the process are illustrated in Sections 5.5.2 and 5.5.3.

- The prior expectation and variance of the residual covariance $u(x)$. It is the multiplication of residual variance σ^2 and covariance function $R(x, x^{'})$ with correlation parameter θ and nugget δ. In addition, residual variance σ^2 and the correlation function parameter θ require specification [120] which can also be calculated using the Algorithm 1.

5.5.2 Calculation of Bayes Linear Emulation

The general form of a univariate emulator is based on a combination of a regression surface with correlated errors for observed and unobserved data can be written as follows:

$$f(x) = g(x)^T\beta + u(x), \tag{5.27a}$$

$$f(x^{'}) = g(x^{'})^T\beta + u(x^{'}), \tag{5.27b}$$

where x and $x^{'}$ are two distinct points treated as the emulator training and testing data points. Based on the univariate emulator with training points x and testing points $x^{'}$, the Bayes linear update can be written on the basis of general forms of the emulator;

$$E_{f(x)}[f(x^{'})] = E[f(x^{'})] + Cov[f(x^{'}), f(x)]Var[f(x)]^{-1}(f(x) - E[f(x)]), \tag{5.28a}$$

$$Var_{f(x)}[f(x^{'})] = Var[f(x^{'})] - Cov[f(x^{'}), f(x))]Var[f(x)]^{-1}Cov[f(x), f(x^{'})]. \tag{5.28b}$$

Due to property of weakly stationary process we can write, $E(u(x)) = 0 \text{and} E(u(x^{'})) = 0$. We know that from general covariance formula, $Cov[\beta, u(x)] = E[\beta\ u(x)] - E(\beta)E[u(x)]$ such that $Cov(\beta, u(x^{'})) = 0$; $Cov(u(x), \beta) = 0$ as $E(u(x)) = 0$. Using Equation (5.27a) and

(5.27b) into the Equation (5.28a), we can write for the expectation part of the emulator:

$$\begin{aligned}
E_{f(x)}[f(x^{'})] &= E[f(x^{'})] + Cov[f(x^{'}), f(x)]Var[f(x)]^{-1}(f(x) - E[f(x)]),\\
&= E[g(x^{'})^T\beta + u(x^{'})] + Cov[g(x^{'})^T\beta + u(x^{'}), g(x)^T\beta + u(x)]Var[f(x)]^{-1},\\
&\quad \times \Big(f(x) - E(g(x)^T\beta + u(x)\Big),\\
&= E[g(x^{'})^T\beta] + Cov[g(x^{'})^T\beta + u(x^{'}), g(x)^T\beta + u(x)]Var[f(x)]^{-1}\\
&\quad \times \Big(f(x) - E(g(x)^T\beta)\Big),\\
&= g(x)^T E(\beta) + \Big[Cov[g(x)^T\beta, g(x)^T\beta] + Cov[g(x^{'})^T\beta, u(x)] + Cov[u(x^{'}), g(x)^T\beta],\\
&\quad + Cov[u(x^{'}), u(x)]Var\Big[f(x)\Big]^{-1}\Big(f(x) - E(g(x)^T\beta)\Big),\\
&= g(x^{'})^T E(\beta) + \Big[Cov[g(x^{'})^T\beta, g(x)^T\beta] + Cov[u(x^{'}), u(x)]\Big],\\
&\quad \times Var\Big[f(x)\Big]^{-1}\Big(f(x) - g(x)^T E(\beta)\Big),\\
&= g(x^{'})^T E(\beta) + \Big[g(x^{'})^T Cov[\beta, \beta]g(x) + Cov[u(x^{'}), u(x)]\Big],\\
&\quad \times Var[f(x)]^{-1}\Big(f(x) - g(x)^T E(\beta)\Big),\\
&= g(x^{'})^T E(\beta) + \Big[g(x^{'})^T Var[\beta]g(x) + Cov[u(x^{'}), u(x)]\Big]Var[f(x)]^{-1},\\
&\quad \times \Big(f(x) - g(x)^T E(\beta)\Big).
\end{aligned} \tag{5.29}$$

The Bayes linear adjusted mean formulation of Equation (5.29) is similar to the posterior mean function of Equation (5.14) of the Gaussian process emulation. For the variance part of the Bayes linear update, we know that:

$$\begin{aligned}
Var_{f(x)}[f(x^{'})] &= Var[f(x^{'})] - Cov[f(x^{'}), f(x)]Var[f(x)]^{-1}Cov[f(x), f(x^{'})],\\
&= Var[f(x^{'})] - \Big[g(x^{'})^T Var[\beta]g(x) + Cov[u(x^{'}), u(x)]\Big]Var[f(x)]^{-1},\\
&\quad \times \Big[g(x)^T Var[\beta]g(x^{'}) + Cov[u(x), u(x^{'})]\Big].
\end{aligned} \tag{5.30}$$

We need to expand the $Var[f(x)]$ and $Var[f(x^{'})]$ such that;

$$\begin{aligned}
Var[f(x)] &= Var\Big[g(x)^T\beta + u(x)\Big],\\
&= g(x)^T Var[\beta]g(x) + Var[u(x)],\\
&= g(x)^T Var[\beta]g(x) + Cov[u(x), u(x)].
\end{aligned} \tag{5.31}$$

$$\begin{aligned} Var[f(x^{'})] &= Var\Big[g(x^{'})^T\beta + u(x^{'})\Big], \\ &= g(x^{'})^T Var[\beta]g(x^{'}) + Var[u(x^{'})], \\ &= g(x^{'})^T Var[\beta]g(x^{'}) + Cov[u(x^{'}), u(x^{'})]. \end{aligned} \tag{5.32}$$

So Equation (5.30) can be written as;

$$\begin{aligned} Var_{f(x)}[f(x^{'})] = \Big[g(x^{'})^T Var[\beta]g(^{'}) + Cov[u(x^{'}), u(x^{'})]\Big] - \Big[g(x^{'})^T Var[\beta]g(x) + Cov[u(x^{'}), u(x)]\Big], \\ \times \Big[g(x)^T Var[\beta]g(x) + Cov[u(x), u(x)]\Big]^{-1}\Big[g(x)^T Var[\beta]g(x^{'}) + Cov[u(x), u(x^{'})]\Big]. \end{aligned} \tag{5.33}$$

The Bayes linear adjusted variance formulation of Equation (5.33) is also similar to the Gaussian process emulation posterior variance function of Equation (5.14).

5.5.3 Formulation of Bayes Linear Emulation

The formulation of the Bayes linear emulation has been illustrated extensively in the previous Sections. So for simplification, we have formulated the whole procedure step by step, described as follows;

1. Step 1: Construct the data frame for the basis of continuous inputs, denoted by G such that $G = (g_1(x), g_2(x), , ..., g_p(x))$.

2. Step 2: Find estimate the maximum likelihood estimates of the correlation parameters and all other parameters considering the objective function (5.24).

3. Step 3: Calculate the mean function values by using $\hat{\beta}$: Find the estimates of $\hat{\beta} = (G^T R_\theta^{-1} G)^{-1} G^T R_\theta^{-1} f(x)$, then multiplying and summing with basis terms gives the prior mean for training $E[f(x)]$ and testing $E[f(x^{'})]$ data points. This part is considered as emulation's regression part.

4. Step 4: Estimate $Var(\hat{\beta})$ and calculate the variance part of the emulation using $g(x)^T Var[\beta]g(x)$ for training points and same way for the testing points.

5. Step 5: Residual process $u(x)$: This part is known as the Gaussian part can be calculated using kernels such that $Cov[u(x^{'}), u(x)] = \widehat{\sigma}^2 R_{[\theta,\delta]}$. The first part $\widehat{\sigma}^2$ and the second part $R_{[\theta,\delta]}$ depends on the correlation length θ and nugget δ, which are possible to estimate using Algorithm 1.

6. Step 6: Now add the variance and Gaussian parts from Step 4 and Step 5 to calculate the variance of training and testing data from Equations (5.31) and (5.32).

7. Step 7: Finally, we needed to add every element from Steps 1-6 to give us the adjusted expectation and variance for the Bayes linear emulation for the continuous inputs.

5.5.4 Diagnostics of Bayes Linear Emulation

To assess the validity of our emulator, we can compute various diagnostics such as (i) resolution, (ii) standardised prediction errors (iii) Mahalanobis distance.

Firstly, the resolution [59] of the Bayes linear update can be expressed as,

$$R_{f(x)}[f(x^{'})] = 1 - \frac{Var_{f(x)}[f(x^{'})]}{Var[f(x)]}. \tag{5.34}$$

The resolution lies between 0 and 1 and functions much like a classical R^2 where resolution values close to 1 indicate a high proportion of the variation has been explained. Values close to 0 indicate the emulator is unable to explain the variation in the simulator. The resolution is related to the emulator adjusted variance but they are not reciprocal.

Secondly, the standardised prediction errors (SPE) or squared standardised changes [59, 64] are the differences between simulator output and emulator expectation for the same input. For a simulation value of y_s with corresponding inputs the SPE can be expressed as,

$$SPE = \frac{y_s - E_{f(x)}[f(x^{'})]}{\sqrt{Var_{f(x)}[f(x^{'})]}}. \tag{5.35}$$

Large values of SPE indicate an apparent conflict between the emulator and simulator, indicative of deficiencies in the fit of the emulator or surprising simulator output values. Generally, the SPE threshold is considered as $+2$ and -2 [64]. An enormous value or outlier on the edge of input space can be discounted due to the difficulty of emulating on that space. Values larger than ± 2 are considered poor matches between the emulator and the simulator. If most values lie within ± 2, it is regarded as a good match between the emulator and the simulator.

Finally, the Mahalanobis distance (MD) between the simulator outputs and the emulator outputs at testing points can be expressed as;

$$MD = (y_s - E_{f(x)}[f(x^{'})])^T Var_{f(x)}[f(x^{'})](y_s - E_{f(x)}[f(x^{'})]), \tag{5.36}$$

where very large or small values outside the range of -2 to $+2$ for the observed Mahalanobis distance (MD) indicate clear conflict between the emulator and simulator [64].

5.5.5 One Dimensional Example

The simple one-dimensional (1-D) example can be considered;

$$F(x) = \frac{1}{2}x + \cos(x) + \log(3x). \tag{5.37}$$

Recalling the data points and the simulator outputs from Section 5.2.1, we construct an emulator to get the simulator output $F(x)$ at new testing points from 0.01 to 5 with an evenly spaced distance of 0.005, which creates 999 points. The emulator adjusted mean and variance are calculated using the Equations (5.29) and (5.33), respectively. For our analysis, we used the maximum likelihood estimate technique algorithm and the L-BFGS-B [39] method to estimate the value of correlation parameters. We also assume $E[f(x)] = 0$ and $E[f(x^{'})] = 0$ for all x considering the values of β are zero. This problem has no nugget such that $\delta = 0$. We have estimated $\sigma^2 = 0.75$ and the correlation parameter $\theta = 0.45$ using MLE. So the prior covariance for inputs x and $x^{'}$ can be written as;

$$Cov[f(x), f(x^{'})] = 0.75 \times \exp\{-0.45(x - x^{'})^2\}. \tag{5.38}$$

Considering all our prior belief specifications, we updated the Bayes linear emulation adjusted expectation and variance. Figure 5.2 shows the result of the fitted emulator.

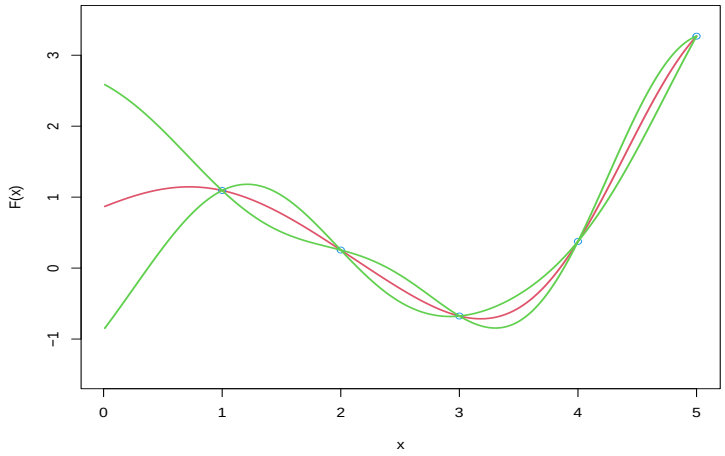


Figure 5.2: x-axis: x is the inputs, y-axis: $F(x)$ is the simulator outputs. Emulator Adjusted expectation $E_{f(x)}[f(x^{'})]$ (red colour) with ± 3 Emulator Standard Deviation $Var_{f(x)}[f(x^{'})]$ (green colour)

The smooth red line represents the emulator expectation $E_{f(x)}[f(x^{'})]$, and the two green lines are at ± 3 of the emulator standard deviation. The blue circle points are our training points of $F(x)$.

So from the result, we can see that our emulator adjusted expected line passed through our five simulator outputs of $f(x)$ with a narrow uncertainty and some uncertainty for the initial points. In the next Section, we demonstrate emulation techniques over the EPIC simulator data of continuous inputs in 2-D space. We also check the validity of our emulator using the diagnostic tools and fit the emulator with more training points on a grid.

5.6 Application to EPIC Simulator Data

In this Section, we extend our problem to a two-dimensional problem of the EPIC simulator data set for the continuous inputs of Nitrogen and Phosphorus.

5.6.1 Emulator Fitting for Crop Spring Barley and Winter Barley

To construct our Bayes linear emulator, we structure the mean function as a simple regression in terms of the simple basis $[1, N, P]$. Thus, for the prior expectation of the simulator $f(x)$ we can write as $E[f(x)] = E[\beta_0] + E[\beta_1]N + E[\beta_2]P$, in terms of three regression coefficients $\beta_0, \beta_1, \beta_2$. Values for these three coefficients, nuggets δ and σ^2, were estimated by maximum likelihood over the training data set. The correlation length parameters for N and P were estimated at $\widehat{\theta}_N = 0.02$ and $\widehat{\theta}_P = 0.03$, respectively. The nugget parameter is estimated as $\widehat{\delta} = 0.05$, and the error variance is estimated as $\widehat{\sigma}^2 = 1$ for the emulation process. So only the Gaussian process is updated from data via Bayes linear emulation approach.

For our analysis, we present results from a subset of crops, namely Spring and Winter Barley. For both of the crops, we have considered one input simulation combination by fixing $St = 5$, $So = 6$, $Wy = 1$, $Sy = 4$ for Spring Barley and $Sy = 6$ for Winter Barley, respectively, which explores a 12×12 grid of combinations of the two fertilizer inputs, N and P. Focusing on this single grid of simulated yield, we construct a Bayes linear emulator based on a simple linear regression and a correlated error with squared exponential covariance function using 60% of the available data for fitting, reserving the

remaining 40% for testing and diagnostics.

The emulator is finally updated for the new training data a 100×100 grid via the Bayes linear formulae in (5.29) and (5.33) for the functions of N and P, which are used to calculate the adjusted emulator mean, standard deviations and the diagnostic resolution. For the diagnostic of the standardized prediction errors, we used the 40% data from the EPIC simulator, as we can not generate the yield for this extended grid.

Figures 5.3 and 5.4 show the result of emulator adjusted mean and standard deviations with resolution diagnostics for the crop Spring Barley and Winter Barley, respectively.

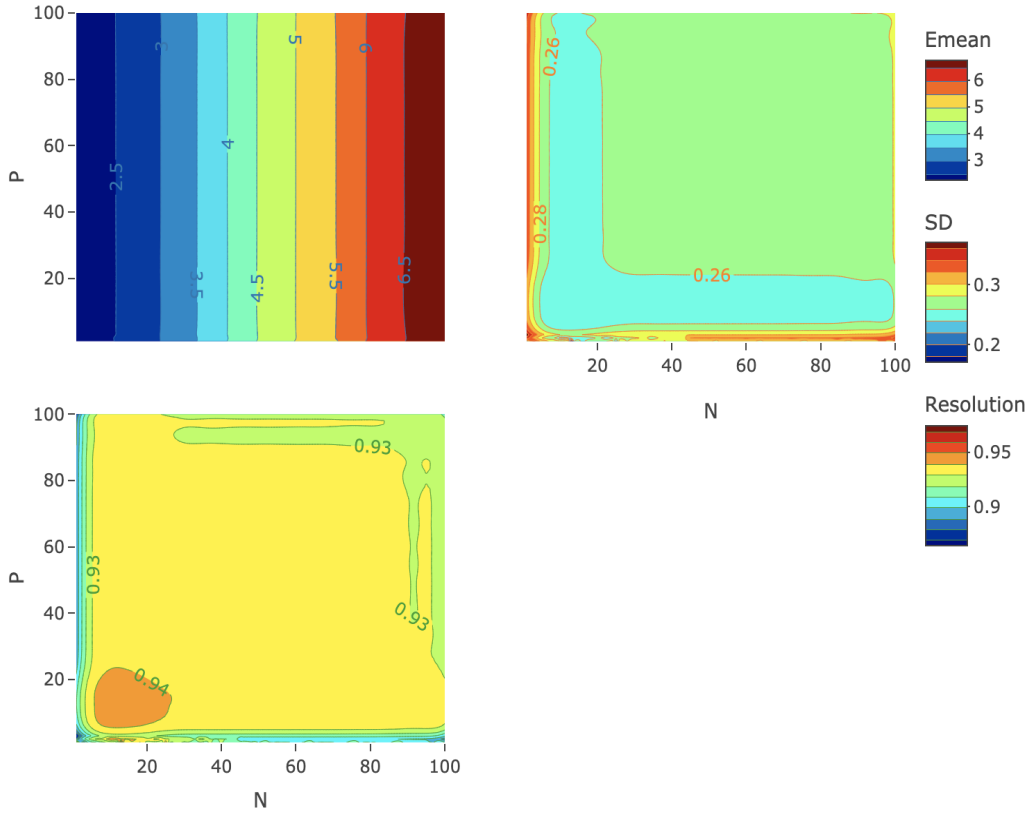


Figure 5.3: Upper Panel: Adjusted Emulator Mean and Standard Deviations for Spring Barley; Lower Panel: Resolution Diagnostic.

The left upper and right upper panels show the adjusted emulated mean and its associated standard deviations as functions of N and P. We note that the crop yield is monotonically increasing with increasing Nitrogen levels and higher for high N values. The crop yield appears insensitive to values of P for both crops. The weak effect of P is much more apparent showing a flat trend such that the dependency on Phosphorous has disappeared entirely due to more prediction points. This trend is exactly showing the

same outcome as in Chapters 3 and 4. The adjusted standard deviations plot illustrates low-level uncertainty around the locations of simulations and flat all over the grid space and an increasing trend as we move away from those simulation points for the low levels of N.

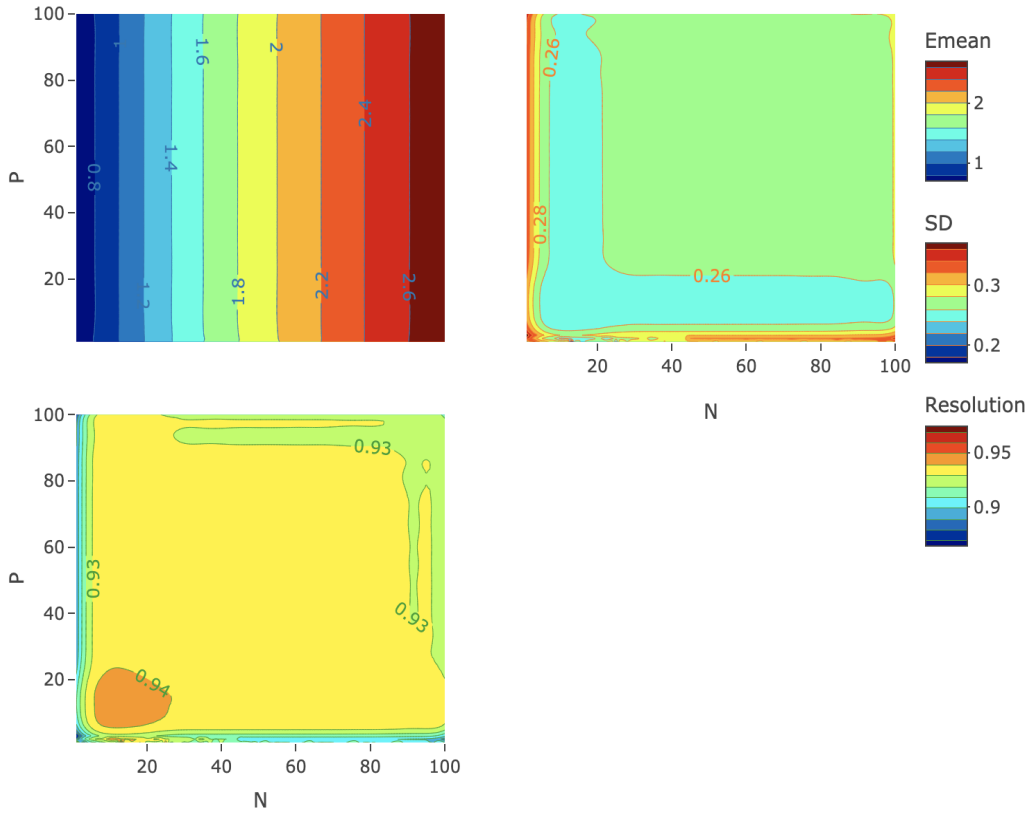


Figure 5.4: Upper Panel: Adjusted Emulator Mean and Standard Deviations for Winter Barley; Lower Panel: Resolution Diagnostic.

From diagnostics of the resolution, we can see high flat resolution all over the space of greater than 0.9, indicating the emulator is very confident in explaining the variability of the simulator. The simple linear regression fitting using $lm()$ function in R-language for N and P inputs shows $R^2 = 0.9482$ that 94.82% of the variation of the dependent variable yield can be explained by the independent variables fertilisers. Clearly, the emulator and simulator have no conflict, hence the valid emulators for both crop yields.

Diagnostic plots of standardised prediction errors for emulating Spring Barley yield and Winter Barley are given in Figures 5.5 and 5.6. All points lie within ± 2, suggesting high consistency and agreement between the emulator and the simulator. The SPE are all smaller than 1 for both crops, indicating the under-confidence of the emulators. The

problem of under-confidence depends on the specific application, such as the concern of sufficient reflection of beliefs about the emulators [120]. It is noted that overconfidence is treated as more problematic than under-confidence if the emulators are used for history matching [120]. For this problem, we only use 40% of the 144 observations, which is a small number of prediction points, and they are also consistent with error variance $\sigma^2 = 1$. All of these prediction points are very close to emulator-adjusted expectations. SPE will exhibit more variations when we explore the factor effect.

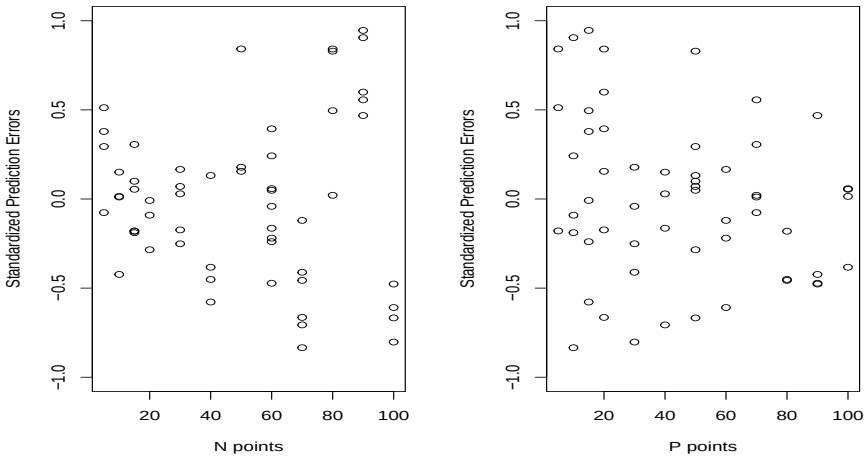


Figure 5.5: Left Panel: Standardised Prediction Errors Corresponding to N; Right Panel: Standardised Prediction Errors Corresponding to P for Spring Barley

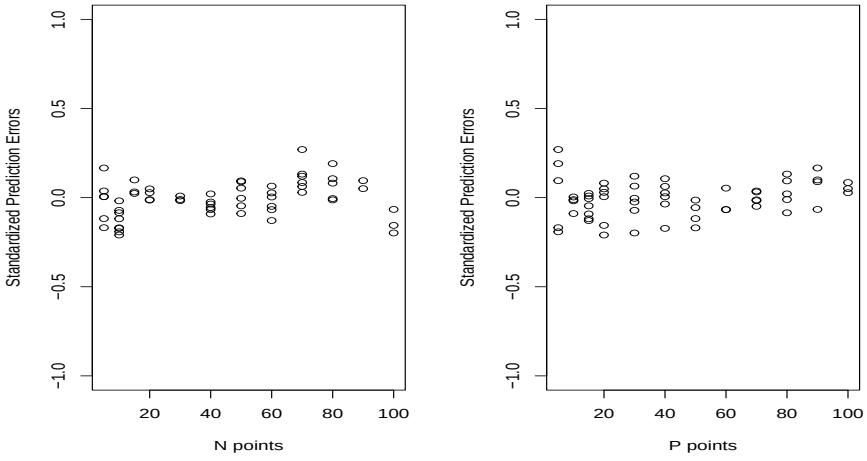


Figure 5.6: Left Panel: Standardised Prediction Errors Corresponding to N; Right Panel: Standardised Prediction Errors Corresponding to P for Winter Barley

5.7 Conclusion

This chapter initially discussed the concept of an emulator and simulator. We set up a general structure of an emulator and illustrated the idea of both Bayes linear and Gaussian process emulation approaches with relevant correlation functions; we have preference for Bayes linear emulation and discussed the various steps to construct the emulator for continuous inputs. Finally, we have shown some tools as a diagnostic to validate our emulator.

Initially, we applied Bayes linear emulation techniques to a one-dimensional example. We then used emulation techniques for EPIC simulator data for the continuous inputs of N and P. Spring and Winter Barley crops have a strong response to N but a weak response to P. The emulator uncertainties for both crops are generally low and only higher near the boundaries of the parameter space. The diagnostics of the emulators illustrated the validity and no conflict with the simulator. The high resolution for both crops indicates the emulators explain most of the simulator uncertainties.

We adopted a Bayes linear approach for our analysis, which substantially simplified the computation and complexity in fitting the model while providing a powerful tool for modelling and analysing the computer model output. The emulator's quality and performance can be readily assessed and monitored through appropriate diagnostics. A natural progression is broadening the input space and considering the effects of the entire collection of simulator inputs, including continuous and categorical variables.

Chapter 6

Mixed Variable Bayes Linear Emulation

6.1 Introduction

The use of complex computer models with qualitative and quantitative inputs are increasing in various field of study. But a key issue is finding a suitable way of modelling the mixed inputs in the context of complex computer experiments. The Gaussian process model is widely used [52, 63, 107] for solving the computer experiments problem for qualitative and quantitative inputs. The continuous input problem can be addressed using the techniques in Chapter 5. However, the main challenge is constructing the covariance between the factor levels. The challenge with incorporation of the factor inputs, as we don't have any particular type of correlation structure to account for the factor correlations. This chapter aims to propose a possible solution to this challenging problem.

Different approaches have been developed to analyze the mixed inputs problem using a Gaussian process. McMillan [40], Joseph and Delaney [60] used the Gaussian process model and proposed a restrictive correction function for assessing the correlation between the levels of factor effects. Qian et al. (2008) [63] illustrated a layout for the nonrestrictive nature of the correlation structure to calculate the correlation between the factor levels in two steps using an optimis ation technique. Furthermore, Zhou et al. (2011) [84] proposed a hypersphere decomposition technique. All the literature used the Gaussian process approach to emulation to assess the mixed input effect for the complex computer problem. An alternative way of handling complex computer experiments is using the Bayes lin-

ear emulation method [59], which has only been used for quantitative inputs so far. Our current analysis will explore the Bayes linear method for mixed factor inputs.

One of the limitations of the existing methods is the use of a constant mean for their analysis procedure [63, 40], and often only considering a single factor effect. Most of the previous studies exclude the nugget effect. Our current study will explore the impact of the mixed inputs using Bayes linear emulation. In Section 6.2, we generalise the idea of the emulation function with the factor effect layout. Section 6.3 discusses three approaches used for mixed inputs computer modelling. In Section 6.4, we use the maximum likelihood theory for mixed inputs with an objective function to estimate the optimum values. Section 6.5 shows the steps to calculate the emulator mean and variance using a Bayes linear emulation approach. We apply the Bayes linear emulation framework techniques to the EPIC simulator in agricultural research demonstrated in Sections 6.6 to 6.8. In Section 6.6, we construct the correlation matrix and identify the best option for the EPIC simulations, and build the emulator for factor steepness and soil in Section 6.7. In Section 6.8, we create the emulator for all inputs, including weather and finally, draw concluding remarks in Section 6.9.

6.2 General Model and Factor Effect Layout

Suppose that a computer experiment involves two distinct types of inputs $v = (x, w)^T$, where $x = (x_1, ..., x_I)^T$ are quantitative variables and $w = (w_1, ..., w_J)^T$ are qualitative variables. So the general emulation function can be written as;

$$f(v) = g(v)^T\beta + u(v),$$

where $g(v)^T\beta$ represents the mean function in a regression form, expressed in terms of the mixed input variables, x and w. The parameters β are unknown scalar regression coefficients corresponding to the regression matrix of basis functions, say G, for the active inputs. The final component, $u(v)$, is the residual process for mixed inputs, which denotes a zero mean weakly stationary process to explain additional variation around the mean function in terms of inputs. We assume that, $G_{p\times k} = (g_1(v), g_2(v), ..., g_p(v))^T$ and $\beta_{k\times 1} = (\beta_1, \beta_2, ..., \beta_k)^T$ and $f(v)$ follows multivariate normal distribution with mean $G\beta$ and variance-covariance of $\sigma^2 R_{[\theta,\tau]}$; where θ is the correlation length for continuous variables and τ is the vector of correlation parameters for factor variables. $R_{[\theta,\tau]}$ is the com-

bination of Gaussian covariance function $R_{[\theta]}$ for continuous inputs with factor correlation matrix T.

Consider a factor variable, w, with levels $c_1, c_2, ..., c_j$ and two distinct factor inputs w and $w^{'}$ such that $w, w^{'} \in \{c_1, c_2, ..., c_j\}$. So for a single factor variable w and a single quantitative input x, the correlation between $v = (x, w)$ and $v^{'} = (x^{'}, w^{'})$ can be expressed as follows (Quin et al. (2008) [63]);

$$Cor(u(v), u(v^{'})) = R(x, x^{'}; \theta) \times T(w, w^{'}; \tau). \tag{6.1}$$

For $x, x^{'} \in \mathbb{R}$, $R(x, x^{'}, \theta)$ is the Gaussian correlation function with correlation length parameter θ such that Equation (6.1) can be written as;

$$Cor(u(v), u(v^{'})) = \exp\Big\{-\sum_{i=1}^{I} \theta_i (x_i - x_i^{'})^2\Big\} \times T(w, w^{'}, \tau). \tag{6.2}$$

where $x = (x_1, ..., x_I)^T$ is quantitative input for $i = 1, 2, ..., I$.

If $w = c_i$ and $w^{'} = c_k$, then we write $T(w, w^{'}) = T(c_i, c_k) = t_{c_i, c_k} = t_{i,k}$, where $t_{i,k} \in [-1, 1]$ is the correlation between levels i and k. We can summarise these correlations for all levels of the factor in the matrix T,

$$T = \begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,j} \\ t_{1,2} & t_{2,2} & \cdots & t_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ t_{j,1} & t_{j,2} & \cdots & t_{j,j} \end{bmatrix} \tag{6.3}$$

For multiple categorical variables such that $v = (x, w)^T$ and $v^{'} = (x^{'}, w^{'})^T$, where $w, w^{'}$ is a vector of factors of length j, and $w_a, w_a^{'} \in \{c_{a,1}, ..., c_{a,j_a}\}$, for the a^{th} factor with j_a levels when $a = 1, 2, ..., j$. Then, using separability, the correlation can be written as;

$$\begin{aligned} Cor(u(v), u(v^{'})) &= R(x, x^{'}, \theta) \times \ T_1(w_1, w_1^{'}; \tau) \times ... \times \ T_j(w_j, w_j^{'}; \tau), \\ &= R(x, x^{'}, \theta) \prod_{a=1}^{j} T_a(w_a, w_a^{'}; \tau). \end{aligned} \tag{6.4}$$

The correlations associated with each factor can be summarised in a $j_a \times j_a$ matrix, one for each factor of $a = 1, ..., j$. The term T_a is the correlation matrix for the factor variables with two distinct inputs of w_a and $w_a^{'}$ which can be expressed as follows;

$$T_a = \begin{bmatrix} t^a_{1,1} & t^a_{1,2} & t^a_{1,3} & \cdots & t^a_{1,j_a} \\ t^a_{1,2} & t^a_{2,2} & t^a_{2,3} & \cdots & t^a_{2,j_a} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t^a_{j_a,1} & t^a_{j_a,2} & t^a_{j_a,3} & \cdots & t^a_{j_a,j_a} \end{bmatrix} \tag{6.5}$$

Matrix T_a will be a valid factor correlation matrix if T_a is symmetric, and positive definite matrix with diagonal elements $t^a_{b,c} = 1$ for all $b = c$ [63].

6.3 Approaches to Model Factor Input Correlation

There are several approaches to determine the correlation values to assign to matrix T and its elements. We consider the (*i*) McMillan approach [40], (*ii*) Zhou approach [84], which was also used by Zhang et al. [107], (*iii*) Exchangeable approach by Quin et al. [63] and Joseph et al. [60] and a modified version of this correlation approach. The details of these approaches are given as follows;

1. The McMillan Approach (MC): The McMillan approach [40] is a multiplicative way of estimating the correlation between the factor levels of $t_{i,j}$ as follows:

$$t_{i,j} = \begin{cases} \exp\{-\tau_i\}\exp\{-\tau_j\} = \exp\{-(\tau_i + \tau_j)\}, & \text{if } c_i \neq c_j \\ 1 & c_i = c_j \end{cases} \tag{6.6}$$

   where the parameters are $\tau_{i,j} > 0$ and the diagonal elements of the factor correlation matrix T will equal 1, and the matrix will be symmetric. So for a factor matrix T with 3 levels $t_{1,1}, t_{1,2}, t_{1,3}$ the correlation matrix between the factor levels can be expressed as follows considering the MC approach;

$$\mathbf{T} = \begin{bmatrix} 1 & & \\ \exp\{-(\tau_1 + \tau_2)\} & 1 & \\ \exp\{-(\tau_1 + \tau_3)\} & \exp\{-(\tau_2 + \tau_3)\} & 1 \end{bmatrix} \tag{6.7}$$

   So for a three-factor levels correlation matrix, the MC method needs three parameters $\tau_1,\ \tau_2,\ \tau_3$.

2. Exchangeable Correlation (EC) approach: An exchangeable correlation function was proposed by Joseph and Delaney [60] initially and later used by Qian [63]. This EC approach can be expressed as;

$$t_{j,i} = t_{i,j} = \begin{cases} \tau, & \text{if } i \neq j \\ 1 & i = j \end{cases} \tag{6.8}$$

   where, τ parameter lies between -1 to +1. This method is considering as a simplification of (6.6) where all $\tau_i = \tau_j$. This model has only one parameter, τ.

- General Correlation (GC) Approach: The Exchangeable approach makes all the factor correlation values as the same, which can be highly restrictive. To get different correlation values, we can modify Equation (6.8) as follows:

$$t_{j,i} = t_{i,j} = \begin{cases} \tau_{i,j} & \text{if } i \neq j \\ 1 & i = j \end{cases} \tag{6.9}$$

For a factor matrix T with 3 levels $\tau_{1,1},\ \tau_{1,2},$ and $\tau_{1,3}$ the correlation matrix between the factor levels can be expressed as follows considering the GC approach;

$$\mathbf{T} = \begin{bmatrix} 1 & & \\ \tau_{1,2} & 1 & \\ \tau_{1,3} & \tau_{2,3} & 1 \end{bmatrix} \tag{6.10}$$

where $\tau_{1,2}$ is the correlation between the factor levels of 1 and 2. This method also gives correlation matrices that are symmetric, have unit diagonal, and positive definiteness such that the eigen values are positive. For this approach, three different parameters are needed to estimate for a 3×3 matrix.

3. Zhou Method of Hypersphere Decomposition: This method [84] differs from the previous techniques as it allows the correlation between factor levels to be positive and negative. This approach uses a hypersphere decomposition (which ensures both negative and positive correlation and also ensures the diagonal element of the correlation matrix are 1) to model the factor levels correlation matrix T. The approach has two different steps as follows:

   (a) Step 1: Applying a Cholesky decomposition to T,

$$T = LL^T, \tag{6.11}$$

   where $L = [l_{d,s}]$ is a strictly positive diagonal lower triangle matrix.

   (b) Step 2: L can be treated as a spherical coordinate system of equations for each row vector in L of the surface point on a d dimensional unit hypersphere such that dimension of the matrix (2×2 dimension for a 2 levels factor variable) and can be expressed as follows;

$$\begin{aligned} l_{1,1} &= 1 \\ l_{d,1} &= \cos(\tau_{d,1}) \\ l_{d,s} &= \sin(\tau_{d,1}), ..., \sin(\tau_{d,s-1}) \cos(\tau_{d,s}) \text{ for } s = 2, ..., d-1 \\ l_{d,d} &= \sin(\tau_{d,1}), ..., \sin(\tau_{d,d-2}) \sin(\tau_{d,d-1}) \end{aligned} \tag{6.12}$$

where the $\tau_{i,j}$ values lies between 0 and π. The nature of the equations makes the unit diagonal element equal to 1. For the three-factor input $d = 3$, the correlation matrix for the factor input levels can be written as,

$$\mathbf{T_3} = \begin{bmatrix} 1 & t_{1,2} & t_{1,3} \\ t_{1,2} & 1 & t_{2,3} \\ t_{1,3} & t_{2,3} & 1 \end{bmatrix} \tag{6.13}$$

Following the step 1, matrix T_3 can be expressed as follows,

$$\begin{aligned} \mathbf{T_3} &= L_3 L_3^T \\ &= \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & l_{12} & l_{13} \\ 0 & l_{22} & l_{23} \\ 0 & 0 & l_{33} \end{bmatrix} \\ &= \begin{bmatrix} 1 & l_{12} & l_{13} \\ l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix} \end{aligned} \tag{6.14}$$

For two factor levels we can write;

$$\begin{aligned} l_{21} &= \cos(\tau_{21}), \\ l_{22} &= \sin(\tau_{21}). \end{aligned} \tag{6.15}$$

From Equation (6.15), we can compute the correlation of two factor level as follows. The coordinates of l_{21} and l_{22} are shown on Figure 6.1 as a half unit circle such that $l_{21}^2 + l_{22}^2 = 1$,

$$\begin{aligned} \mathbf{T_2} &= L_2 L_2^T \\ &= \begin{bmatrix} 1 & l_{12} \\ l_{21} & l_{21}^2 + l_{22}^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \cos(\tau_{1,2}) \\ \cos(\tau_{2,1}) & \cos^2(\tau_{2,1}) + \sin^2(\tau_{2,1}) \end{bmatrix} \\ &= \begin{bmatrix} 1 & \cos(\tau_{1,2}) \\ \cos(\tau_{2,1}) & 1 \end{bmatrix} \end{aligned} \tag{6.16}$$

Similarly, for factor input level $d = 3$, can be written as

$$\begin{aligned} l_{31} &= \cos(\tau_{3,1}), \\ l_{32} &= \sin(\tau_{3,1})\cos(\tau_{3,2}), \\ l_{33} &= \sin(\tau_{3,1})\sin(\tau_{3,2}). \end{aligned} \tag{6.17}$$

and using the above values of l_{31}, l_{32}, l_{33} and the coordinates from Figure 6.1(b), we can calculate the correlation between factor levels for T_3 in Equation (6.13) as;

$$\mathbf{T_3} = \begin{bmatrix} 1 & & \\ \cos(\tau_{2,1}) & 1 & \\ \cos(\tau_{3,1}) & \cos(\tau_{2,1})\cos(\tau_{3,1}) + \sin(\tau_{2,1})\sin(\tau_{3,1})\cos(\tau_{3,2}) & 1 \end{bmatrix} \tag{6.18}$$

Thus, we need to estimate the parameters τ, which it the parameter of the factor correlations to evaluate the correlation matrix. So for Equation (6.18), we need three parameters to assess the correlation matrix T_3; we also need six parameters to estimate for a 4-factor levels variable and 28 parameters for an 8-factor levels variable. So all three methods required the same number of parameters as there are unique correlations, except the simple Exchangeable approach. The nature of the McMillan and General approach are such that they are mathematically amenable to estimate, but, the more involved nature of the Zhou method makes it hard to handle the large number of factor levels.

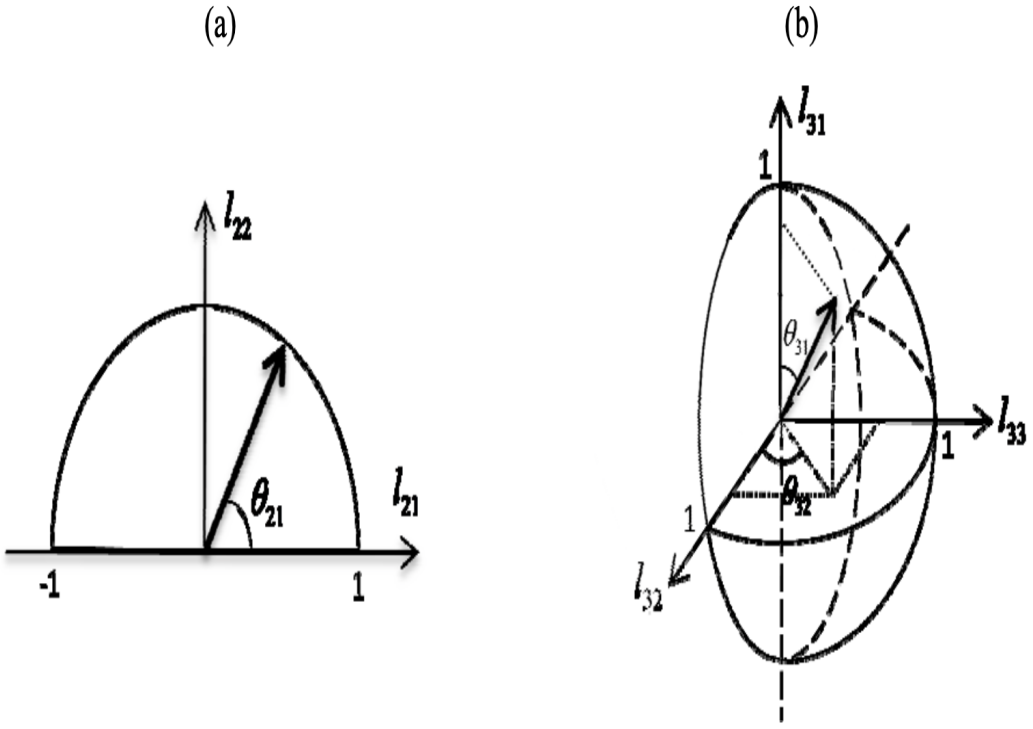


Figure 6.1: Zhou Method of Hypersphere Decomposition for 2-Factor Input (a) and 3-factor input (b) Extracted from the Paper [84]

6.4 Maximum Likelihood Inference for Correlation Parameters

By recalling the assumptions from Section 5.4 and Equation (5.18), the likelihood function of the $f(v)$ parameters can be written, with factor inputs as follows:

$$\begin{aligned}
L(\beta,\sigma^2,\theta,\tau|f(v)) &= \frac{1}{\sqrt{(2\pi\sigma^2)^n|R_{\left[\theta,\tau\right]}|}}\exp\left[-\frac{1}{2\sigma^2}(f(v)-G\beta)^T R^{-1}_{\left[\theta,\tau\right]}(f(v)-G\beta)\right],\\
&= (2\pi)^{-n/2}(\sigma^2)^{-n/2}|R_{\left[\theta,\tau\right]}|^{-1/2}\exp\Bigg[-\frac{1}{2\sigma^2}(f(v)-G\beta)^T R^{-1}_{\left[\theta,\tau\right]}\\
&\quad\times(f(v)-G\beta)\Bigg],\\
\log L(\beta,\sigma^2,\theta,\tau|f(v)) &= -\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{1}{2}\ln|R_{\left[\theta,T\tau\right]}|-\frac{1}{2\sigma^2}(f(v)-G\beta)^T R^{-1}_{\left[\theta,\tau\right]},\\
&\quad\times(f(v)-G\beta),\\
&= c+\frac{n}{2}\log(\sigma^2)+\frac{1}{2}\log|R_{\left[\theta,\tau\right]}|+\frac{1}{2\sigma^2}(f(v)-G\beta)^T R^{-1}_{\left[\theta,\tau\right]}(f(v)-G\beta).
\end{aligned} \tag{6.19}$$

Following the results from Chapter 5 and changing the R_θ to $R_{\left[\theta,\tau\right]}$ with the addition of factor effect gives us the following estimates for our parameters for mixed inputs:

$$\begin{aligned}
\widehat{\beta} &= (G^T R^{-1}_{\left[\theta,T\right]}G)^{-1}G^T R^{-1}_{\left[\theta,\tau\right]}f(v)\\
\widehat{\sigma}^2 &= \frac{1}{n}(f(v)-G\beta)^T R^{-1}_{\left[\theta,\tau\right]}(f(v)-G\beta)\\
Var(\widehat{\beta}) &= \widehat{\sigma}^2\left(G^T R^{-1}_{\left[\theta,\tau\right]}G\right)^{-1}
\end{aligned}$$

6.4.1 Objective Function for Mixed Inputs

In Chapter 5, we estimated the correlation lengths for the Gaussian correlation function by maximum likelihood. In this Chapter, we need to apply an optimisation technique to estimate the factor correlation values. We need the estimates for τ to build the matrix T, so an optimisation technique must be applied.

For our problem, the overall negative log likelihood function can be written in a sim-

plified way by substituting the estimate of $\widehat{\sigma}^2$ into the Equation (6.19);

$$\begin{aligned} NL(\theta,\tau|f(v)) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\widehat{\sigma}^2) - \frac{1}{2}\log|R_{\left[\theta,\tau\right]}| - \frac{1}{2\widehat{\sigma}^2}(f(v)-G\widehat{\beta})^T \\ &\quad \times R^{-1}_{\left[\theta,\tau\right]}(f(v)-G\hat{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\widehat{\sigma}^2) - \frac{1}{2}\log|R_{\left[\theta,\tau\right]}| \\ &\quad - \frac{1}{2\frac{1}{n}(f(v)-G\widehat{\beta})^T R^{-1}_{\left[\theta,\tau\right]}(f(v)-G\widehat{\beta})}(f(v)-G\hat{\beta})^T R^{-1}_{\left[\theta,\tau\right]}(f(v)-G\widehat{\beta}), \\ NL(\theta,\tau|f(v)) &= -\frac{n}{2}\log(2\pi\widehat{\sigma}^2) - \frac{1}{2}\log|R_{\left[\theta,\tau\right]}| - \frac{n}{2}, \end{aligned} \tag{6.20}$$

where $\widehat{\beta}$, $\widehat{\sigma}^2$ and $R_{\left[\theta,\tau\right]}$ depends on the correlation parameters θ and the factor correlation levels τ. To estimate the parameters, we must minimise Equation (6.20).

$$\begin{aligned} (\hat{\theta},\hat{\tau}) = \min_{\theta,\tau} \quad & \left[-\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2}\log|R_{[\theta,\tau]}| - \frac{n}{2}\right]. \\ \text{Subject to} \quad & \theta_i \geqslant 0; i = 1,...,I \\ & t_{j,k} = 1,\ t_{j,k} \in [-1,1]; j,k = 1,...,J \end{aligned} \tag{6.21}$$

The constraint $\theta_i \geqslant 0$ is the correlation length for the continuous inputs, and $diag(T) = 1$ ensures the diagonal element of the factor correlation matrix is 1. We have another constraint to ensure the positive definite of the correlation matrix, which is not easy to express in terms of the parameter τ. We can confirm these properties after building the correlation matrix using the estimated values from the optimisation techniques mentioned in Chapter 5.

6.4.1.1 Introducing Nugget Effect on Mixed Inputs

To introduce a nugget term from Chapter to $f(v)$, with factor inputs can be expressed as,

$$\begin{aligned} \sigma^2 = Var(u(v)) = \Sigma &= Var[u(v_A)] + Var[\nu(v)], \\ &= \sigma_A^2 + \sigma_\nu^2. \end{aligned} \tag{6.22}$$

Thus a convenient parameterization is to consider σ_ν^2 as a function of δ of σ^2, so from Equation (6.22) we can write;

$$\sigma^2 = (1-\delta)\sigma^2 + \delta\sigma^2. \tag{6.23}$$

wherever $v \neq v^{'}$. We would also have $Var[u(v), u(v^{'}] = (1-\delta)\sigma^2 R_{[\theta,\tau]}(v, v^{'})$

With the nugget parameter the emulator of $f(v)$ will now follow a multivariate normal distribution with mean $G\beta$ and variance-covariance of $\widehat{\sigma}^2 R_{\left[\theta,\tau,\delta\right]}$. The maximum likelihood estimates will be;

$$\widehat{\beta}(\theta,\tau,\delta) = (G^T R_{\left[\theta,\tau,\delta\right]}{}^{-1} G)^{-1} G^T R_{\left[\theta,\tau,\delta\right]}{}^{-1} f(v), \tag{6.24}$$

$$\widehat{\sigma}^2(\theta,\tau,\delta) = \frac{1}{n}(f(v) - G\widehat{\beta})^T (R_{\left[\theta,\tau,\delta\right]})^{-1} (f(v) - G\widehat{\beta}), \tag{6.25}$$

$$Var(\widehat{\beta}) = \widehat{\sigma}^2 (G^T (R_{\left[\theta,\tau,\delta\right]})^{-1} G)^{-1}. \tag{6.26}$$

And the new objective function with the nugget effect is;

$$\begin{aligned}(\hat{\theta},\hat{\tau},\hat{\delta}) = \min_{\theta,\tau,\delta} \quad & \left[-\frac{n}{2}\log(2\pi\widehat{\sigma}^2) - \frac{1}{2}\log|R_{\left[\theta,\tau,\delta\right]}| - \frac{n}{2}\right] \\ \text{Subject to} \quad & \theta_i \geqslant 0;\ i = 1,...,I \\ & \delta \in [0,1] \\ & diag(T_j) = 1;\ j = 1,...,J \\ & \tau_{i,j} \in [-1,1]\end{aligned} \tag{6.27}$$

We need to estimate the maximum likelihood and correlation parameters considering the objective function of Equation (6.27). For this, we set an algorithm for the whole procedure as follows;

Algorithm 2 Estimation Process for the MLE and Correlation Parameters

1: Calculate the Correlation matrix $R'_{\left[\theta,T,\delta\right]}$.
2: Calculate the factor effect correlation matrix T.
3: Matrix multiplication of steps (1) and (2) gives the correlation $Cor(v, v')$.
4: Calculate inverse and determinant of $R'_{\left[\theta,\tau,\delta\right]}$.
5: Calculate $\widehat{\beta}$ from Equation (6.24), which gives the estimate for $\widehat{\sigma}^2$ from Equation (6.25).
6: Plugging $\widehat{\sigma}^2$ estimate gives $Var(\widehat{\beta})$ from Equation (6.26).
7: Use optimisation technique within the constraints gives us the estimate for the correlation parameters and lengths.

6.5 Bayes Linear Emulation for Mixed Inputs

The objective of this section is to use the Bayes linear emulation technique for mixed inputs. We apply mixed inputs general layout with the optimisation function to estimate the values and hence use the Bayes linear update to calculate emulator adjusted expectations and variances.

The general form of an emulator with two distinct inputs v and $v^{'}$ and nugget effect δ can be written as follows:

$$f(v) = g(v)^T\beta + u(v) + \delta,$$
$$f(v^{'}) = g(v^{'})^T\beta + u(v^{'}) + \delta.$$

The Bayes linear update can be written based on general forms of a univariate emulator for the mixed inputs as follows;

$$E_{f(v)}[f(v^{'})] = E[f(v^{'})] + Cov[f(v^{'}), f(v)]Var[f(v)]^{-1}(f(v) - E[f(v)]), \tag{6.28}$$
$$Var_{f(v)}[f(v^{'})] = Var[f(v^{'})] - Cov[f(v^{'}), f(v)]Var[f(v)]^{-1}Cov[f(v), f(v^{'})]. \tag{6.29}$$

The formulation of the Bayes linear emulation for mixed inputs is more complex to compute. For simplification, we have formulated the whole procedure step by step. The reason to simplify the whole process in to different steps is to make it more understandable, which can be described as follows;

1. Step 1: We consider the design matrix for the factor input $w = (w_1, ..., w_J)$, say $Z = (Z_{1,2}, Z_{1,3}, ..., Z_{1,J})^T$.

2. Step 2: Construct the data frame for a combination of factor inputs w and continuous inputs $x = (x_1, x_2, ..., x_I)^T$, such that $G = (g_1(x, w), g_2(x, w), ..., g_p(x, w))^T$.

3. Step 3 (Expectation and Variance): Calculate the prior expectation and variance similar to Chapter 5 using Algorithm 2.

4. Step 4: Calculate the Gaussian residual process part with the factor inputs and error variance;

$$Cor(u(v), u(v^{'})) = \widehat{\sigma}^2 R_{[\theta,\tau,\delta]} \times \mathbf{T} \tag{6.30}$$

Now adding the variance part from Step 3 and the Gaussian part from Step 4 gives the covariance $Cov[f(v), f(v^{'})]$ of training and testing data of (x, w) and $(x^{'}, w^{'})$.

5. Step 5: Finally, using each part from Steps 1 to 4 in Equations (6.28) and (6.29) gives the emulator adjusted expectation and variance for the Bayes linear mixed inputs emulation. Now we can use the adjusted expectation and variance for the emulator diagnostic process, similar to Chapter 5.

6.6 Application to EPIC Simulator Data

We applied our methodology to the EPIC simulator data for the Spring Barley crop and initially measured the performance of three approaches: McMillan approach, Zhou approach and the general correlation approach. Firstly, two factors, steepness and soil, were used to fit the emulator to assess the efficiency of our methods. We estimated the MLE for the correlation parameters and then constructed the correlation matrix. Finally, we considered another factor input, weather, assessing our approach works nicely for many mixed inputs.

6.6.1 Correlation Matrix and Performance Measures of Approaches

In this subsection, we illustrate the use of Algorithm 2 to estimate the optimum values for the correlation of the factor levels. Considering $f(v)$ we use the objective function of (6.27) with a nugget effect δ to optimize the factor-level correlations. We fixed the same hyperparameters from Chapter 5 to optimise the factor level correlations for steepness, soil and weather subject to boundary constraints on the parameters $\tau,\ \theta$.

For the mixed inputs problem, initially, we consider only the two-factor inputs of steepness and soil, plus the quantitative inputs of Nitrogen and Phosphorus. We explore three possible bases $g(v)$ for our preliminary analysis, namely linear, second order and third-order polynomials in $N,\ P$ summarized in Table 6.1.

Let us denote the three levels of the soil factor as $w_1 \in \{c_{1,1},\ c_{1,2},\ c_{1,3}\}$ and the four levels of steepness factor as $w_2 \in \{c_{2,1},\ c_{2,2},\ c_{2,3},\ c_{2,4}\}$. Then we can construct a set of dummy variables based on $0-1$ as a factor encoding as in Section 4.2.2.

Table 6.1: Mean Function and Basis for Factor Steepness and Soil

Mean Function	Simple Basis $[g(v)^T]$	No. of Coefficients
Linear	$1,\ N,\ P,\ Z_2^{So},\ Z_3^{So},\ Z_2^{St},\ Z_3^{St},\ Z_4^{St}$	8
2nd Order	$1,\ N,\ P,\ NP,\ N^2,\ P^2,\ Z_2^{So},\ Z_3^{So},\ Z_2^{St},\ Z_3^{St},\ Z_4^{St}$	11
3rd Order	$1,\ N,\ P,\ NP,\ N^2,\ P^2,\ N^3,\ P^3,\ N^2P,\ NP^2,$ $Z_2^{So},\ Z_3^{So},\ Z_2^{St},\ Z_3^{St},\ Z_4^{St}$	15

For the inputs we explore three mean functions as a simple basis in terms of 8, 11 and 15 regression coefficients shown in Table 6.1, with a correlated error with squared exponential covariance function. We fit the model on 60% of the available data, reserving the remaining 40% for testing and diagnostics. At first, we constructed the correlation matrix using the McMillan, Zhou and General correlation approaches. The results of the estimated factor level correlations using the General correlation (GC) in Equation (6.9), McMillan in Equation (6.6) & Zhou approach in Equation (6.18) are shown in Figures 6.2 and 6.3 respectively. The upper panel correlations show the model with a linear mean function. The middle panel correlations are for the model with the 2nd-order polynomial mean function, and the lower panel correlations are the 3rd-order polynomial mean function.

The steepness factor level correlations get higher with the increase in complexity of the mean function. We observed a weak (> 0.31) to strong (> 0.7) correlation between the majority of the levels of the factor steepness in the right column. The correlations between steepness levels 2 and 4 and steepness levels 3 and 4 are high, indicating a moderate degree of association for all three polynomial mean functions. A high correlation between all the steepness factor levels is observable for the third order mean function. On the other hand, soil factor level correlations remain low across all three cases of the mean function so that we can potentially treat them as approximately independent. Checking the eigen-values of the factor correlation matrix, we find all are greater than zero which fulfils the positive definite matrix's requirement for our correlation matrix.

Figure 6.3 shows the correlation matrix of the McMillan approach and Zhou approach using the third-order polynomial mean function. From McMillan's and Zhou's approaches, we can see a high positive correlation between the factor levels for the factor steepness compared to GC approach. We can also see very low correlations, which are close to zero for the factor soil for the MC approach.

We consider the standardised prediction errors (SPE) and mean squared errors (MSE) diagnostics. The box plot of the standardised prediction errors (SPE) for three methods is shown in Figure 6.4. The Zhou and Exchangeable methods satisfy SPE's diagnostic threshold, falling between -2 to 2 but, the McMillan approach indicates a comparatively poorer performance due to SPE values falling outside of the threshold, which possibly indicates conflict between the emulator and the simulator. The box plot of MSE for the three methods is shown in Figure 6.4. The GC method shows the smaller MSE compared to the other two methods. However, all three methods looks very close to each other, but GC method is still performing better. For both SPE and MSE selection criteria, the GC method performs better among the approaches. However, the Zhou method yields a good SPE but is outperformed by GC regarding MSE.

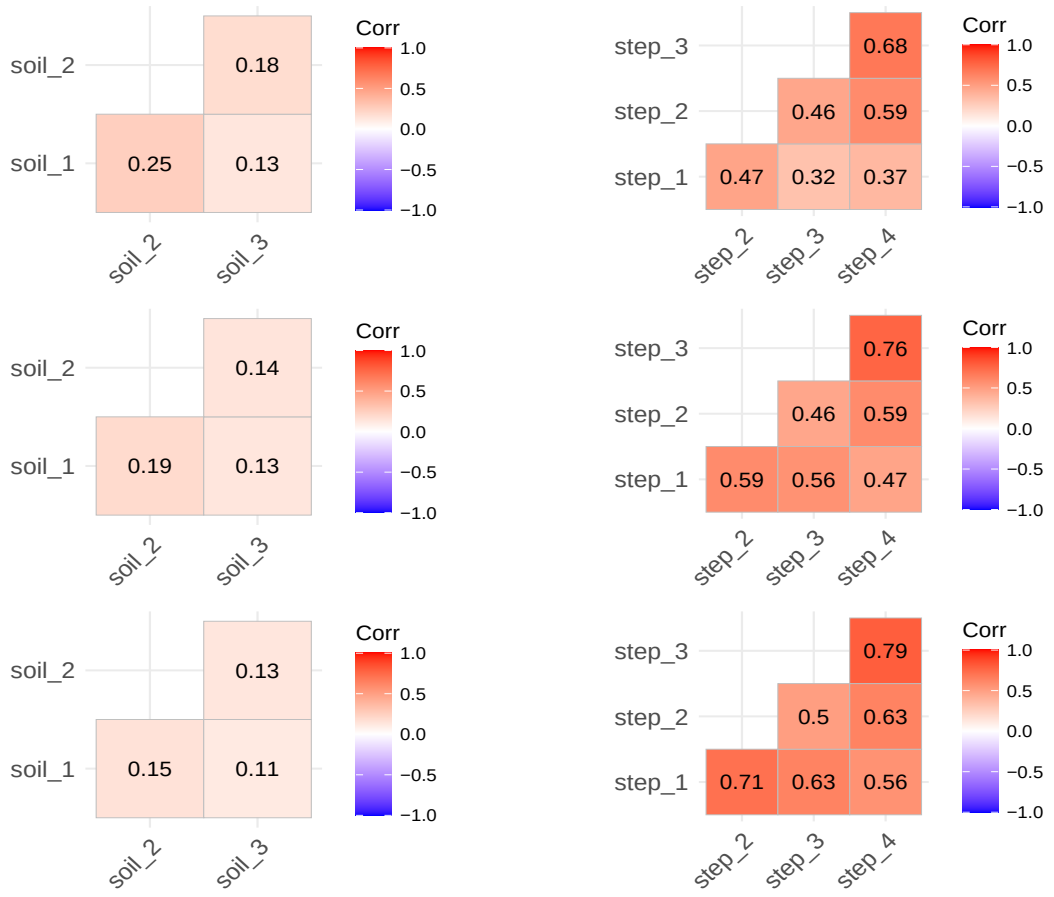


Figure 6.2: Estimated Factor Correlations Using GC. Left Column: Soil, Right Column: Steepness. Top Row: Linear, Middle Row: 2nd order, Bottom Row: 3rd order mean function

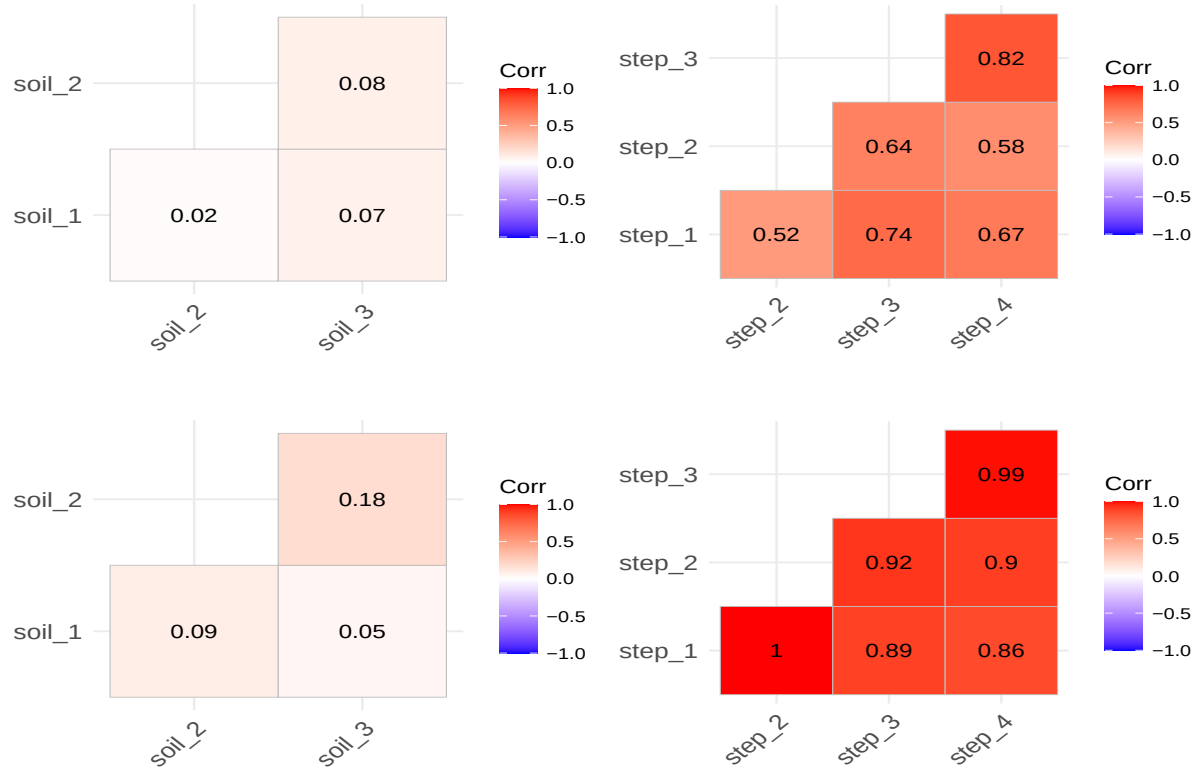


Figure 6.3: Estimated Factor Correlations Using McMillan and Zhou approaches with 3rd Order Mean Function. Left Column: Soil, Right Column: Steepness. Top Row: McMillan Approach, Bottom Row: Zhou Approach

The structure of the Zhou method is complicated due to the hypersphere decomposition. Suppose we include the weather factor variable with 8 levels and the other two factors of steepness and soil. In that case, using the Zhou method becomes very complex due to the hypersphere decomposition required to construct a correlation matrix of 96 factor combinations, which will create difficulties to optimize over so many parameters. The McMillan method is comparatively less flexible than the other two methods because of the nature of the correlation function. For example, if we have four-factor levels, say $w_2 \in \{c_1, c_2, c_3, c_4\}$. Consider the case where we have a factor with four levels. Levels 1 and 2 are strongly correlated, giving large values for τ_1 and τ_2. Similarly, levels 3 and 4 are also highly correlated with τ_3 and τ_4 large. Under the McMillan model, this will force, for example, levels 1 and levels 3 to also be highly correlated as the correlation is given by exp(-(τ_1+τ_3)), which will be large. Thus it would be impossible to represent a situation where these pairs of factor levels were uncorrelated in this model.

In summary, we can see the largest SPE, and MSE for MC and Zhou compared to the GC method. So, considering all the drawbacks and the diagnostics, we will proceed with our further analysis using the GC method correlation matrix as our preference.

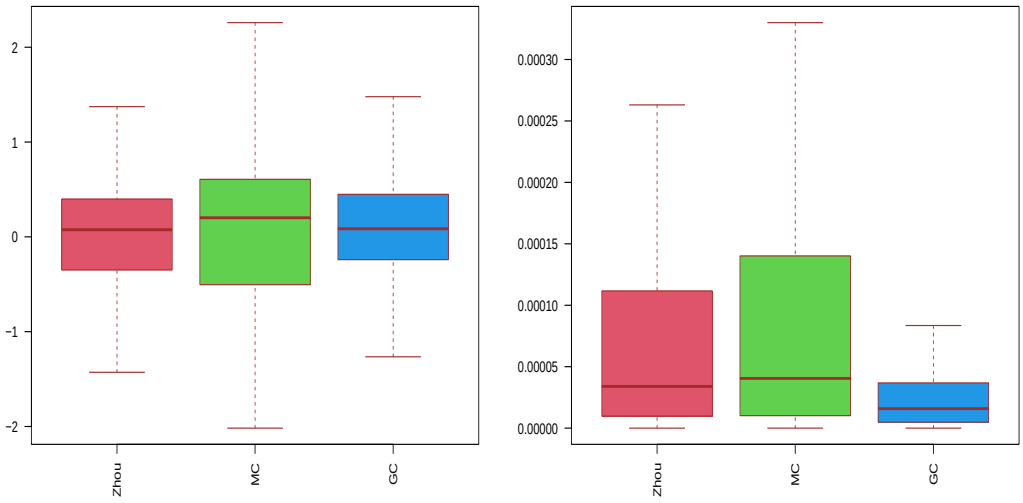


Figure 6.4: Left Panel: Box plot for the Standardised Prediction Error; Right Panel: Box Plot for the MSE using Zhou method, McMillan (MC) Method and General Approach (GC) Method

6.7 Emulation with Steepness and Soil Factors

Using the mean function and basis from Table 6.1 with the GC approach to represent the correlations, the emulators are updated from the training data using the formulae (6.28) and (6.29) for mixed inputs to calculate the adjusted mean and standard deviations. Figures $6.5 - 6.7$ shows the result of adjusted emulator means and standard deviations with companion diagnostics of resolution and standardized prediction errors considering the simple linear, second order polynomial and third order polynomial mean function respectively as a function of N and P.

From all three orders of mean functions emulator expectations from Figures $6.5 - 6.7$, it is noticeable that crop yield is increasing with increasing N levels. However, the effect of P is weak. From Figures 6.5 to 6.7, we can say standard deviations display comparatively low levels of uncertainty around the locations where we have simulations and show an increasing trend as we move away from those simulations points. However, the uncertainty

is lower for the linear mean function, going to around 0.11, with slight higher standard deviations for the second-order mean function around 0.25 and slightly higher for the third-order polynomial compared to second order, which is around 0.30.

From diagnostics of the resolution from Figures 6.5 and 6.6, we notice the low resolution all over the space, whose values are less than 0.4 for first order and less than 0.4 to 0.65 for the second order mean function, indicating the emulator's failure to explain most of the variability of the simulator. Finally, from Figure 6.7, the resolution plot for third order mean function shows high resolution values, which indicates the emulator mostly catches the variation of most of the simulator behaviour.

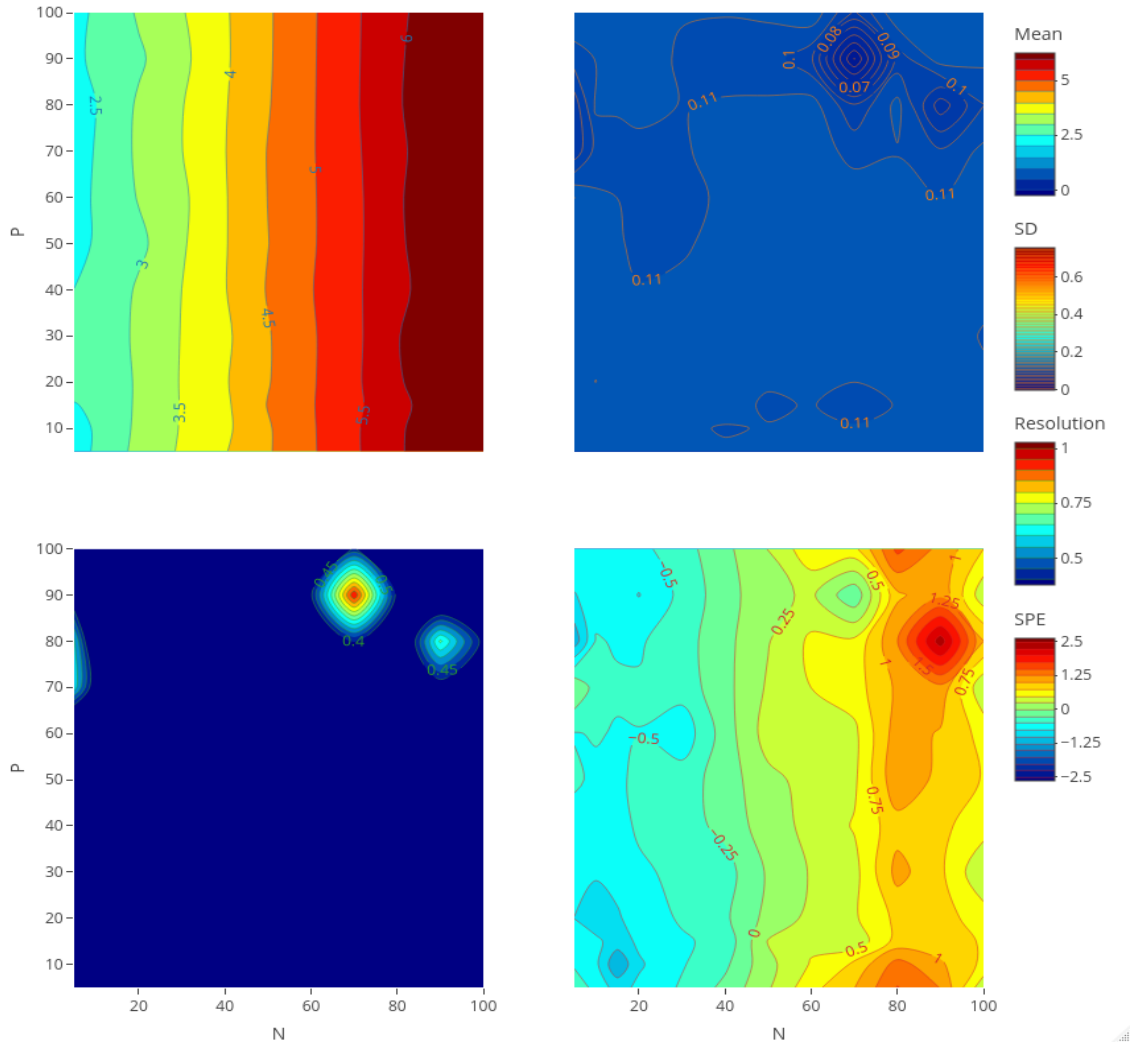


Figure 6.5: Linear mean function when steepness $St = 6$, soil $So = 5$, Weather $Wy = 1$, and $Sy = 4$. Top left: Emulator expectation; Top right: Emulator Standard Deviations; Bottom left: Emulator Resolution; Bottom right: SPE

Standardised prediction error values outside ±2 standard deviation indicate a conflict between the simulator and emulator. The linear mean function in Figure 6.5 shows several values outside of the threshold for the high N values, which is the indication of under-fitting of the mean. The 2nd order polynomial function in Figure 6.6 shows a few values outside of the threshold, close to the edge and there exist no testing points to emulate and can be ignored [64]. Figure 6.7 shows standardised prediction errors lying between ±2 of the standard deviation highlighting no conflict between the emulator and simulator.

In summary, we see clear evidence in favour of the third-order polynomial means functions over the other two possible mean functions based on the results of our diagnostics. Due to the lower SPE and higher resolution, we will proceed with the third-order polynomial mean function as a basis of $g(v)^T$ for further analysis with the GC approach.

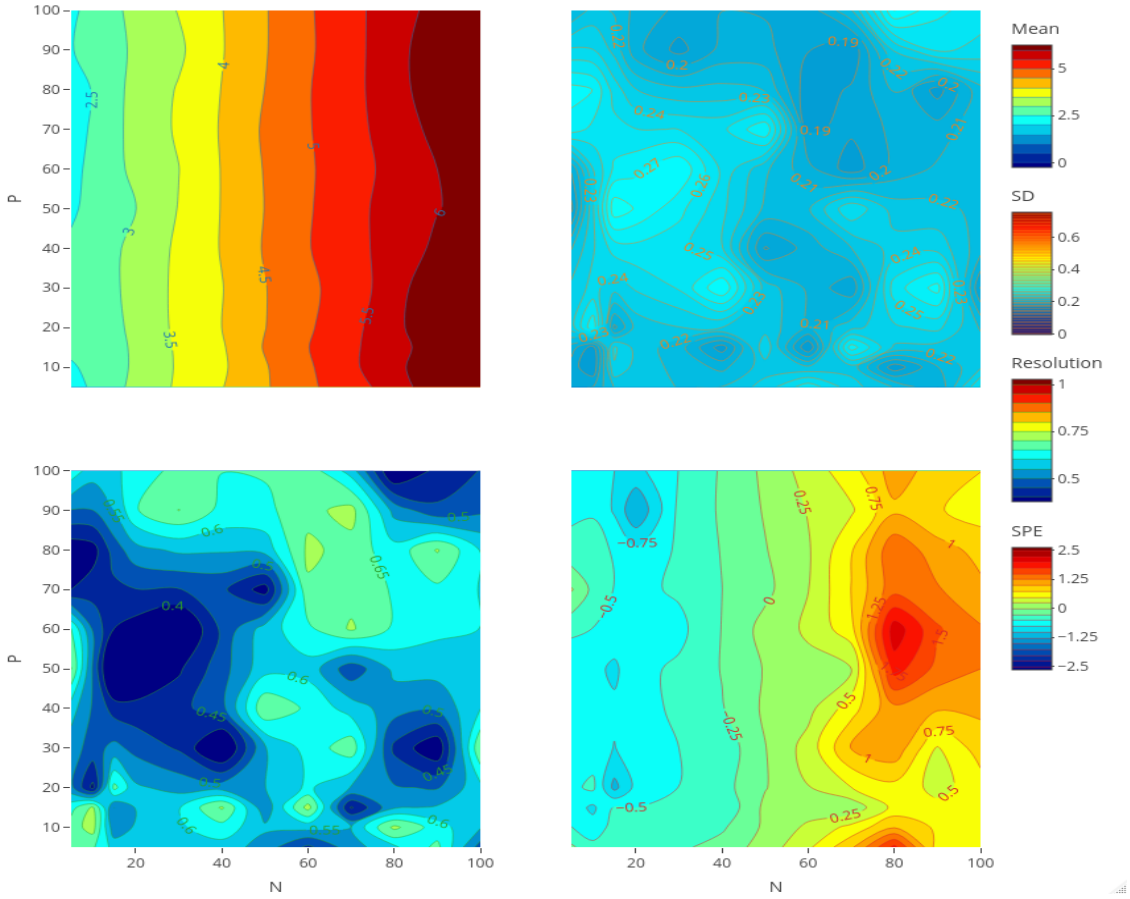


Figure 6.6: 2nd Order mean function when $St = 6$, $So = 5$. Top left: Emulator Expectation; Top right: Emulator Standard Deviations; Bottom left: Emulator Resolution; Bottom right: SPE

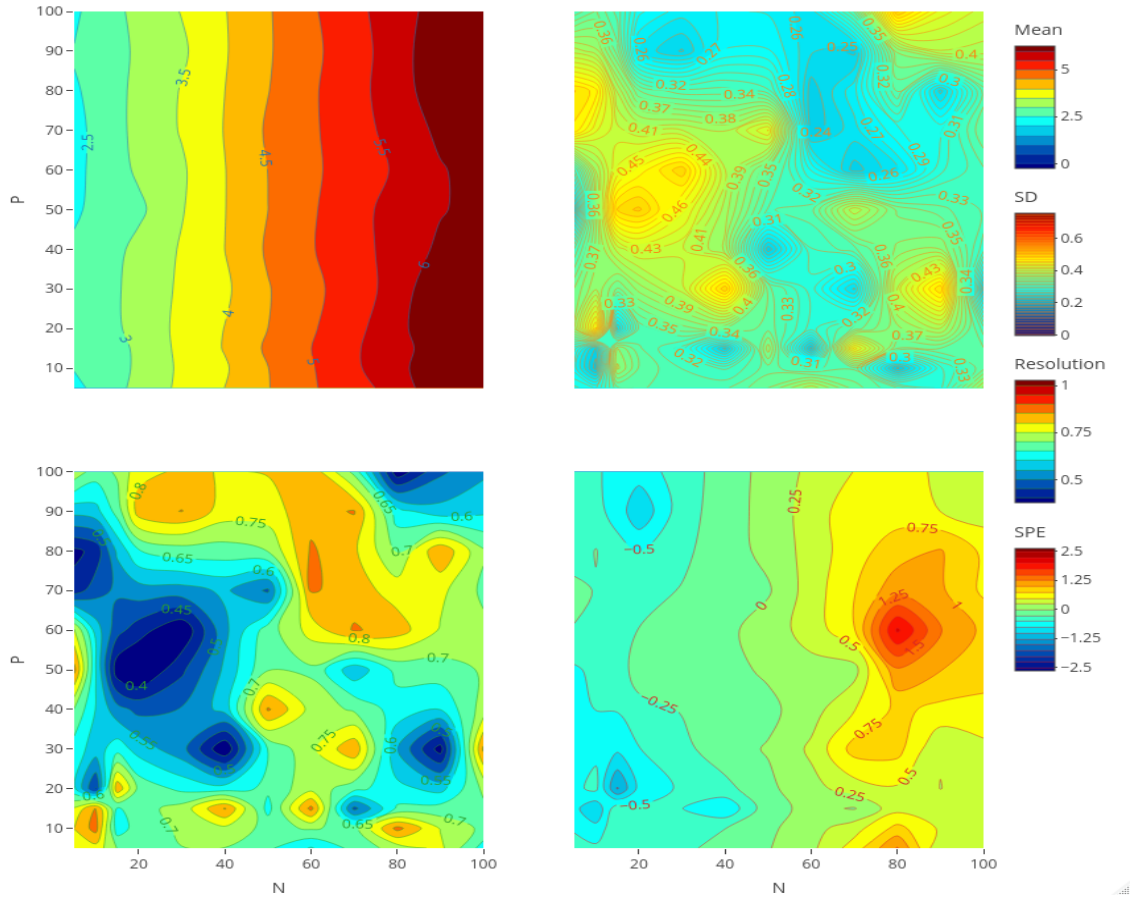


Figure 6.7: 3rd Order mean function when $St = 6$, $So = 5$, $Wy = 1$, and $Sy = 4$. Top left: Adjusted Emulator Expectation; Top right: Emulator Standard Deviations; Bottom left: Emulator Resolution; Bottom right: SPE

6.8 Emulation for Factors Weather, Steepness and Soil

In this section, we apply the mixed input emulation approach to all factors including the additional variable weather with 8 levels. We now need to set up an emulator which can evaluate the 96 possible factor combinations of steepness, soil and weather ($3 \times 4 \times 8 = 96$).

We use the General approach from Equation (6.9) to parameterize the factor correlations. Initially, we use Algorithm 2 to estimate the 28 values for the correlation parameters for weather factor levels. These optimisations were carried out using the L-BFGS-B optimization method, which was implemented via the *optim*() function in R-language adapting from Algorithm 2. We use the objective function of (6.27) with the nugget effect to op-

timize the factor level correlation assuming the same hyper-parameter values such that $\widehat{\theta}_N = 0.02$, $\widehat{\theta}_P = 0.03$, $\widehat{\delta} = 0.05$ similar to continuous input emulation.

We consider the third-order polynomial mean function for the mixed inputs problem involving weather. The estimated factor level correlation matrix for factor weather is shown in Figure 6.8. We observe some weak (< 0.30) to moderate correlations ($0.40 < m < 0.7$) between the levels of the factor weather. Factor level 1 has moderate correlations with factor levels $2, 6,$ and 7. And, factor level 2 shows the same moderate corrections trend with factor levels 6 and 7.

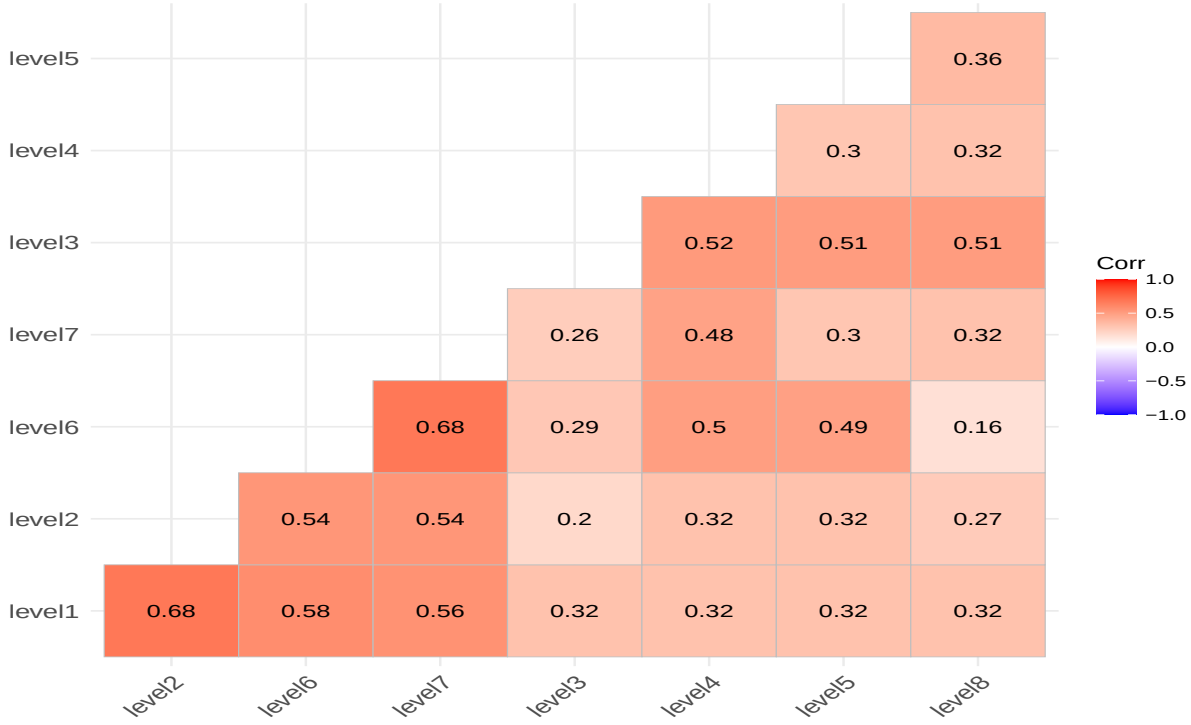


Figure 6.8: Correlation Matrix for Factor Weather using General Correlation Approach

For mixed inputs of weather, steepness and soil analysis with an extended grid, we construct a Bayes emulator based on a 3rd-order polynomial regression and a correlated error with covariance function and nugget effect using all simulator data as training points and a grid of 15×15 with 21600 observations as testing points and resolution diagnostic in terms of the simple basis of $g(v)^T = [1,\ N,\ P,\ NP,\ N^2,\ P^2,\ N^3,\ P^3,\ N^2P,\ NP^2,\ Z_2^{So},\ Z_3^{So},\ Z_2^{St}, Z_3^{St}, Z_4^{St},\ Z_2^W,\ Z_3^W,\ Z_4^W,\ Z_5^W,\ Z_6^W,\ Z_7^W,\ Z_8^W]$ and $E[f(v)] = g(v)^T E(\beta)$ with 22 regression coefficients. Expectations $E[f(v)]$ and variances $Var[f(v)]$ for these coefficients are calculate by using Algorithm 2 for all mixed inputs.

Figure 6.9 offers the result of the adjusted emulator mean and standard deviations

with companion diagnostics of resolution considering the 3rd-order polynomial regression mean function over a 15×15 grid, with standardised prediction errors for the testing data.

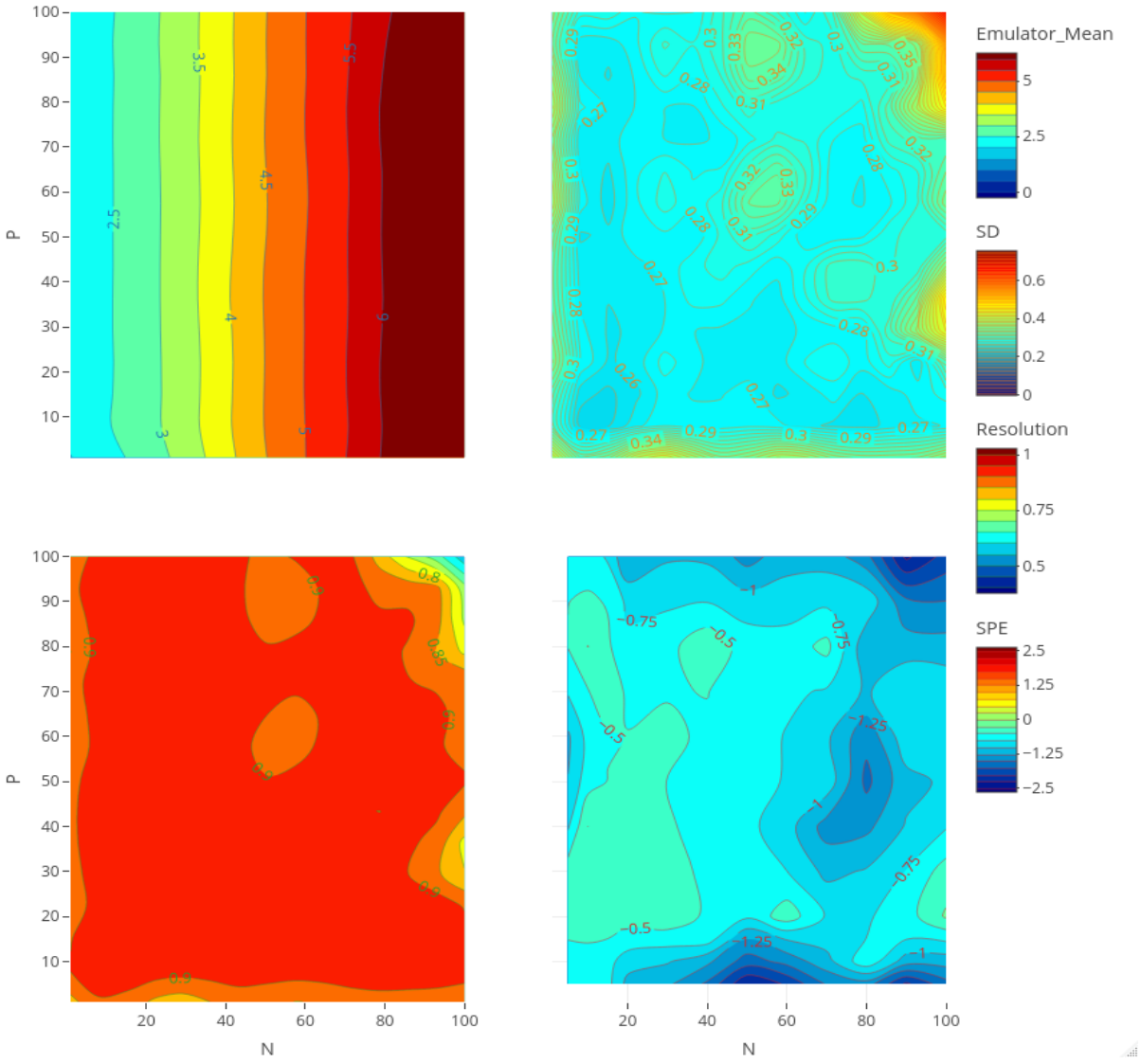


Figure 6.9: 3rd Order mean function when $St = 6$, $So = 5$, $Wy = 1$, $Sy = 4$. Top left: Emulator Adjusted Expectation; Top right: Emulator Standard Deviations; Bottom left: Emulator Resolution; Bottom right: SPE

From Figure 6.9, the crop yield is showing same response with respect to inputs Nitrogen and Phosphorus. The adjusted standard deviations plot is showing slightly narrower uncertainty compared to the steepness and soil only problem after adding the factor weather. From diagnostics of the resolution, we can see high flat resolution all over the space of greater than 0.9 except edges, indicating the same trend as the steepness and soil only problem. The standardised prediction error values lying clearly inside of ± 2 standard

deviations indicate no evidence of conflict and a high degree of consistency and agreement between the emulator and the simulator.

6.9 Conclusion

In this chapter, we have proposed an approach for emulating complex computer models with quantitative and categorical inputs using Bayes linear emulation. We demonstrated the methods by applying them to the EPIC crop yield simulator.

We have shown the extension of Bayes continuous emulation inputs to Bayes mixed input emulation in Section 6.2. Section 6.3 has featured the different approaches to model the factor input correlations. Section 6.4 found the MLE estimates of the correlation parameters, and set-up the Bayes linear mixed input emulation. We used the EPIC simulator data for the Spring Barley crop, initially estimated the MLE of the correlation parameters, and constructed the correlation matrices. In Section 6.7, we updated the emulators for two factors, steepness and soil, using linear and polynomial mean functions. We extended our problem by adding weather for the third-order polynomial mean function and updated this complex emulator. The diagnostics revealed no conflict between the emulator and the simulator.

We learned some valuable tools for the emulation approach, which will be a handy addition to the emulation methodology. Firstly, we have acquired skills in the mathematical formulation of an emulator considering quantitative and qualitative inputs together. We used direct maximum likelihood estimation to estimate the parameter values. This chapter reviewed three approaches to model the mixed inputs with a modified general version of the correlation function. Finally, we gained understanding how to apply the correlation matrix to construct the emulators in the context of the Bayes linear approach .

Chapter 7

Bayes Linear Emulation Approach for Utility and Implausibility

7.1 Introduction

The main goal of this chapter is to balance the crop yield and the pollutant outputs of the simulator to find the best choice of Nitrogen and Phosphorus fertilisers levels in the input space. We will use utility to quantify the combined value of yield and pollution to determine which combinations of inputs give better overall results than others. So this chapter will initially explore utility over the input space and seek the maximum expected value to find the best input choice.

The final part of this chapter is about identifying an area using the mixed input emulators for the combined yield and pollutants by using history matching and implausibility techniques. The approach we will use to determine the optimal input space for maximum utility by emulation will be history matching. History matching is a technique for finding a set of model inputs for the simulator such that its outputs most closely resemble the observed data accounting for model uncertainty [74, 121, 129]. The method works by identifying and excluding the input space where the simulator fails to match the observed data (the history); this region is called the implausible [74, 121, 129] region. For this research, we build an emulator of the expected utility and apply the history-matching technique to seek input values that best match the maximum utility value.

This chapter starts with the basic concept of utility in Section 7.2. We then introduce the idea of history matching (HM) with implausibility in Section 7.3. We set up a utility

function considering yield and two pollutants with the respective coefficients in Section 7.4. In this Chapter, we initially build the emulators for the pollutants NRLOAD (N_p)and PRLOAD (P_p) considering both continuous and mixed inputs like we did with crop yield of Spring Barley in Chapters 5 and 6. Section 7.5 applies the concept of history matching and implausibility to our utility function to find the optimum region for the yield in terms of the fertilisers Nitrogen and Phosphorus. Thus, we must emulate the pollutants N_p and P_p in Section 7.6. We perform a sensitivity analysis for the utility parameters in Section 7.7. Section 7.8 is about utility emulation and implausibility, where the utility and implausibility for the continuous inputs are calculate. Section 7.8.2 covers building the emulators for the pollutants with the steepness and soil factors. Variable selection is applied in Section 7.8.3 to remove the unnecessary basis terms, and we build the final emulators for the pollutants in Section 7.8.3.2 with the selected inputs, including the weather factor. Then we calculated the utility and implausibility for all inputs. Finally, some concluding remarks are in Section 7.9.

7.2 Utility Measures and Functions

In general, utility is a numeric value measuring the preference for an event by the decision maker. The importance of the event increases with the increase of the utility. So a decision maker makes a decision based on the maximum principle such that they choose the best one which provides higher utility. However, sometimes it becomes complicated to describe an event with a numeric value which is not generally measurable. So it will be simpler if we start the utility estimation procedure by considering the presence of a numerical value [15].

A utility function is required to identify the optimal combination of crop yield and pollutants to quantify the overall benefit. The utility is a measure of preference which depends on the nature of the study. A cardinal utility [5] is the preference of the individuals based on a countable number. For example, a farmer harvested two crops, Maize and Barley, in different crop rotations, and the Maize crop provided a 500 kg yield, but the Barley crop produced a 450 kg yield. The farmer's utility lies on the numeric scale, such as the monetary value of the crops, and the farmer will prefer the crop Maize for the next year due to its higher production if both crops have same price.

For this thesis, the utility will be calculated based on a simple linear combination of

the variables and can be written as:

$$U(x_1, x_2, ..., x_n) = b_0x_1 + b_1x_2 + ... + b_nx_n, \tag{7.1}$$

where $x_1, x_2, ..., x_n$ are the variables of the interest and $b_0,\ b_1, ..., b_n$ are the parameters which quantify the value of a single unit of each variable. Our interest is to obtain the maximum utility from the yield, so we need to consider the linear effect yield by subtracting the impact of pollutants in terms of cost coefficients. So the Equation (7.1) can be written as;

$$U(x_1, x_2, ..., x_n) = b_0x_1 - b_1x_2 - ... - b_nx_n, \tag{7.2}$$

where x_1 is the yield, $x_2, ..., x_n$ are the pollutants.

7.3 History Matching

History matching (HM) is a method used to calibrate complex computer models like EPIC. In general, these models consist of a set of inputs with an actual value at which it is assumed the simulator will replicate the real-world system. However, these actual values have yet to be discovered, and the real-world values are observed with error. So history matching is a technique used to identify sets of input parameters which give acceptable matches between model output and physical experiments or real data, accounting for the effects of model discrepancy and observational error [36, 45, 74, 81, 121]. This method follows an iterative procedure that eliminates the unmatched model parameter regions between the model output and real observations. History matching requires extensive exploration of the model parameter space. So it's most commonly used with a Bayes linear approximation to emulation due to its speedy evaluations, and computational simplicity [141]. Suppose history matching finds that the parameters are not matched with the real data. In that case, the model will not accurately represent reality, suggesting a need to consider model discrepancy. The use of history matching is as widespread as emulation and has been used in different backgrounds such as climate [76, 103, 114], epidemiology [85], and reservoir modelling [36, 45, 72]. This research is an effort to use the history-matching technique in agricultural research.

In order to analyze the concept of history matching, a statistical relationship between the computer model $F(.)$ and related system y [46] can be expressed by the best input assumption [120];

$$y = F(x^*) + \epsilon, \tag{7.3}$$

where ϵ_i is the model discrepancy between the related system and computer model such that $\epsilon = y - F(x^*)$, and x^* is the best input. Under the best input assumption, ϵ is assumed to be independent of the model output $F(x^*)$ with variance $Var(\epsilon) = \sigma_\epsilon^2$; in other words we have that $F(x) \perp\!\!\!\perp \epsilon$ [120].

History matching requires a system observation z of the value of real data subject to unknown observational error e. Assuming that the observation is obtained from the system value y combined with additive observational error, we obtain the observation equation [141]:

$$z = y + e, \tag{7.4}$$

where e is the observational error such that $E(e) = 0$ and $Var(e) = \sigma_e^2$ which are judged independent of y. So from Equations (7.3) and (7.4), we can write the observed history as

$$z = F(x^*) + \epsilon + e. \tag{7.5}$$

Thus we can relate the observed values z to the simulator F, evaluated at its best input x^* under model discrepancy ϵ and observational error e.

7.3.1 Implausibility Measure

History matching requires a means to identify the region of acceptable matches between simulator output and reality, for this we use the implausibility measure. The implausibility measure, $I(x)$, is defined over $x \in \chi$, which is a parameter space and quantifies the mismatch between the simulator output and the actu al observation. The input values for which $I(x)$ is large form the implausible region, and those with $I(x)$ small are considered non-implausible regions.

For an individual simulator output, $F(x)$, implausibility measures the deviation between $F(x)$ and y with specific tolerances to identify model discrepancy. In an ideal situation, where we are given the true value of the system y without the error, we could quantify the implausibility as:

$$I_1(x) = \frac{\Big[F(x) - y\Big]^2}{Var(\epsilon)}. \tag{7.6}$$

where the main goal of the implausibility measure is to identify values of x for which $F(x)$ and y are close. But, in reality, it is impossible to calculate the value of y, only being

observed as a noisy observation z, which also includes the observational errors e. Thus we obtain:

$$I_2(x) = \frac{\Big[F(x) - z\Big]^2}{Var(\epsilon) + Var(e)}. \tag{7.7}$$

where ϵ and e are assumed to be uncorrelated, $I^2(x)$ now measures how far the observed value is from the output $F(x)$. Given the computationally expensive nature of the simulator we need to replace $F(x)$ with the emulator. Considering the independence of ϵ and e and the emulator adjusted mean and variance $E_{f(x')}[f(x)]$ and $Var_{f(x')}[f(x)]$, the implausibility $I(x)$ for any input parameter x can be given by [74, 121];

$$I(x) = \frac{\Big[E_{f(x')}[f(x)] - z\Big]^2}{\Big[Var_{f(x')}[f(x)] + Var(\epsilon) + Var(e)\Big]}. \tag{7.8}$$

The numerator is the deviation between the observed value and the emulator expectation and the denominator is the sum of all the uncertainties of emulator variance, discrepancy and observed error. To obtain large values of implausibility, we need large differences for the numerator and smaller values for the variances such that we are confident of a bad representation of reality. On the other hand, small differences and large variances give small implausible values such as a good match between the simulator output and reality or sufficient uncertainty to be unable to determine. So now, a threshold value, c, can be introduced to define a region of poorly matched input space, using $I(x) > c$. A common choice is to use $c = 3$ using Pukelsheim's 3-sigma rule [32].

Naturally, complex computer simulators are used to produce multiple outputs. So considering this fact, we can consider a collection of O univariate responses. For a multivariate output, it is a desirable to combine the individual implausibilities of each output into a single implausible measure. We combine the implausibility measures over the different outputs, $I_i(x)$, for each output $i = 1, ..., O$ by finding the maximum implausibility. Which can be written as;

$$I_M(x) = \max_{i \in O} \; I_i(x). \tag{7.9}$$

To be small, we need a good match on all outputs for $I_M(x)$. For example, combining smaller implausibilities for one output and larger for the other will create a bad match with reality.

7.4 Utility Function for Yield and Pollutants

To get the utility value of a particular simulation, we need to combine the yield and pollutants. For illustration purpose, we use a linear utility function in the pollutants and yields. Naturally, one might want to maximise yield but whilst minimising pollution, so we subtract the effect of the pollutants from the yield. Thus the function for utility is the difference between yield and pollutants up to some multiplicative coefficients. Alternative utilities could also consider by deducting fertiliser costs, in addition to pollution contributions. For simplicity, we will not consider this at this stage. The linear utility function $U(Y, N_p, P_p)$ considering the yield Y and pollutants N_p, P_p can be expressed as;

$$U(Y, N_p, P_p) = b_0 Y - b_1 N_p - b_2 P_p, \tag{7.10}$$

where b_0, b_1 and b_2 are the coefficients corresponding to yield, nitrogen pollutant and phosphorus pollutant, respectively. These coefficients are treated as a gain, per unit yield or unit pollution such that the value per unit, b_0, can be determined from the market price for a given crop. However, there is no equivalent way to determine the pollutants coefficients. A sensitivity analysis is performed to explore the different values of these coefficients in Section 7.7.

Now considering the mean and variance for the yield and pollutants, we can calculate the expected utility and variance of the utility using Equation (7.10). The expected utility is expressed as;

$$\begin{aligned} E\big[U(Y, N_p, P_p)\big] &= E(b_0 Y - b_1 N_p - b_2 P_p), \\ &= b_0 E(Y) - b_1 E(N_p) - b_2 E(P_p). \end{aligned} \tag{7.11}$$

where $E(Y)$ is the emulator mean for the yield, constructed in Chapters 5 and 6 for continuous inputs and mixed variables, respectively. The emulator means of $E(N_p)$ and $E(P_p)$ will be constructed in this Chapter and then applied to this utility function. Assuming the yield Y and two pollutants N_p and P_p are independent such that $Cov[Y, N_p] = Cov[Y, P_p] = Cov[N_p, P_p] = 0$, then the variance of the utility function can be expressed as follows;

$$\begin{aligned} Var\big[U(Y, N_p, P_p)\big] &= Var(b_0 Y - b_1 N_p - b_2 P_p), \\ &= b_0^2 Var[Y] + b_1^2 Var[N_p] + b_2^2 Var[P_p] - 2b_0 b_1 Cov[Y, N_p] \\ &\quad - 2b_0 b_1 Cov[Y, P_p] - 2b_1 b_2 Cov[N_p, P_p], \end{aligned} \tag{7.12a}$$

$$= b_0^2 Var[Y] + b_1^2 Var[N_p] + b_2^2 Var[P_p], \tag{7.12b}$$

where $Var(Y)$ is the emulator variance for the yield, which was already constructed in Chapters 5 and 6. The emulator variance for the pollutants $Var(N_p)$ and $Var(P_p)$ will be built in this Chapter and then applied to this utility function to create an emulator. If we were not to assume that Y, N_p, P_p as independent, we would need a multivariate emulator for $[Y, N_p, P_p]$ such that $U = b^T \begin{bmatrix} Y \\ N_p \\ P_p \end{bmatrix}$ from which we would obtain the necessary co-variances to evaluate (7.12a).

7.5 Implausibility for Utility Function

As our goal is to find the maximum value of utility, not to match the observation z, we need to replace the z with a suitable maximum expected utility value. So, we modify the basic implausibility Equation (7.8) in terms of the utility function to compare to a maximum utility value such that;

$$I(x) = \frac{\Big[E\{U(Y, N_p, P_p)\} - U^*\Big]^2}{Var\{U(Y, N_p, P_p)\} + Var[U^*]}, \tag{7.13}$$

where $E\{U(Y, N_p, P_p)\}$ is the expected utility, which can be estimated using Equation (7.11) and $Var\{U(Y, N_p, P_p)\}$ the variance of the utility, which can be calculated using Equation (7.12*b*). The term U^* is the maximum expected utility value of $E\{U(Y, N_p, P_p)\}$ such that $U^* = \max[E\{U(Y, N_p, P_p)]$. And also, note that we remove the terms $Var(\epsilon)$ and $Var(e)$, as we don't have the real observations for comparison and we will treat U^* as fixed and known. So Equation (7.13) can be written as

$$I(x) = \frac{\Big[E\{U(Y, N_p, P_p)\} - \max\left[E\{U(Y, N_p, P_p)\}\right]\Big]^2}{Var\Big[U(Y, N_p, P_p)\Big]}. \tag{7.14}$$

To calculate the implausibility from Equation (7.14) we need the expected utility, maximum expected utility and the utility variance by combining yield and pollutants. To estimate the value of U^*, firstly, we need to calculate the expected utility from Equation (7.11) for which we required the emulators adjusted expected values with the values for

b_0, b_1, b_2 coefficients and then to take the maximum of the calculated expected utility values. The smaller the difference between the expected utility and the maximum expected utility with large utility variance, yields the lower implausibility and vice versa. The low non-implausible region is the matched region for the expected and maximum expected utility. If the expected utility values are close to the maximum expected value, this yields a good match and hence the lower implausible values. On the other hand, if the expected utility values show a departure from the maximum expected utility value indicates a bad match between them and high implausibility.

This chapter intends to determine a region for which the expected and maximum expected utility are close to each other and the variance of the utility is small. We split the problem into two parts, like the continuous case in Chapter 5 and mixed inputs problem in Chapter 6. So for the continuous case, we used one combination of the factors, where we can only calculate the implausibility $I(x)$ concerning the inputs Nitrogen and Phosphorus. This case intends to find the region of low $I(x)$ and discard the larger values. This region can be calculated by using $I(x) < 3$.

For mixed inputs, we have many possible combinations of the factor variables to consider. Our optimisation will be for the continuous inputs, N and P only as categorical factors can not be controlled by the farmer here. So we will apply a similar maximization technique to Equation (7.9), where we consider each of the $i = 1, 2, ..., n_c$ factor combinations separately, generating implausibilities, which are combined into

$$I_M(x) = \max_i \ I_i(x). \tag{7.15}$$

Suppose we have two different outputs from the simulator, then $I_M(x)$ is used to combine them to get a single measurement which is its main property. We use the same concept to combine $I_i(x)$ across the different categorical factors. So for a given continuous input x, we have one set of implausibility measures for a given factor combination w and another for $w^{'}$ until we have all n_c unique factor combinations. The maximum overall factor combinations thus provides the value of I_M for each input setting x.

7.6 Emulation for Continuous Inputs Pollutants

In this section, we fit the emulators for the pollutants Nitrogen to the river (N_p), and Phosphorus to the river (P_p) for the continuous inputs N and P. The definitions of these variables are already defined in Chapter 2.

To construct the Bayes linear emulator, we need to structure the mean function as the basis function of $g^T(x)$ in terms of the regression parameters. For the pollutants, P_p and N_p, to construct the Bayes linear emulator, we also structure the mean function as a simple regression in terms of the simple basis $[1, N, P]$ as for yield. Thus, for the prior expectation of the simulator $f^N(x)$ we can write as $E[f^N(x)] = E[\beta_0^N] + E[\beta_1^N]N + E[\beta_2^N]P$, in terms of three regression coefficients β_0^N, β_1^N, β_2^N. The MLE estimate values of the nugget $\hat{\delta} = 0.054$ and the estimated correlation length parameters for N and P are $\hat{\theta}_N = 0.017$ and $\hat{\theta}_P = 0.024$ respectively, and we used the same optimised values for the pollutant P_p. These estimated values moves slightly and return the estimates close to starting values for the pollutant N_p, which may happen for a flat gradient.

We explore Bayes linear emulation techniques for more prediction points and the emulator updates from the training data over the 100×100 grid for the functions of N and P for crop yield Spring Barley 5. Figure 7.1 shows the result of adjusted emulator mean and standard deviations with relevant diagnostics considering the 1st-order polynomial regression mean function for the pollutant N_p.

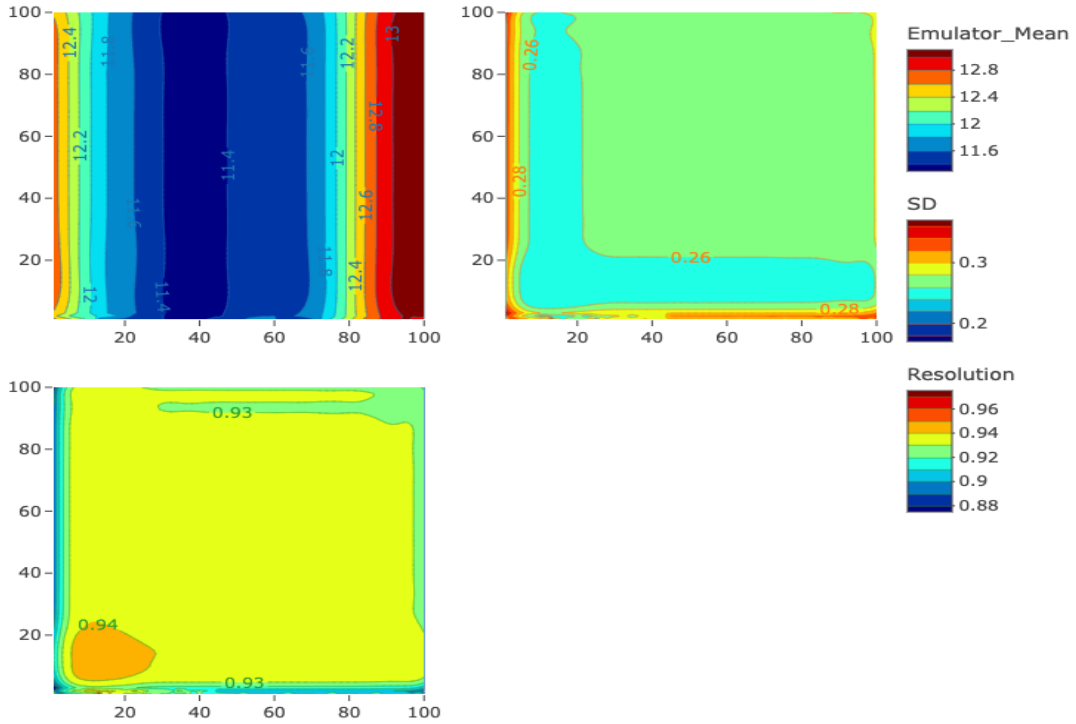


Figure 7.1: Upper Panel: Adjusted Emulator Expectation (Left) And Standard Deviations(Right) for N_p as a function of Nitrogen (N) on the x-axis and Phosphorus (P) on the y-axis using $St = 5, So = 6, Wy = 1, Sy = 4$; Lower Panel: Resolution for N_p.

The left upper panel of Figure 7.1 shows adjusted emulator expectation for N_p; it is noticeable that pollution is increasing with increasing N and getting flat for $N > 85$ and also high for very low values of N. The effect of P appears negligible, as with crop yield Spring Barley. Figure 7.2 left upper panel shows the plot of emulator adjusted expectation for P_p; it is noticeable that pollution is increasing substantially with increasing P and getting higher for high levels of P, and shows higher pollution for low N values indicating an interaction with P. So the overall trend of the adjusted emulator expectation plots shows the increasing trend for N_p pollution with the increase of N values and increasing P_p pollution with the increase of P values. The emulator standard deviations plots from Figures 7.1 and 7.2 illustrate narrow low-level uncertainty around the locations and moreover flat all over the parameter space similar to the emulator adjusted variance plot for crop yield Spring Barley.

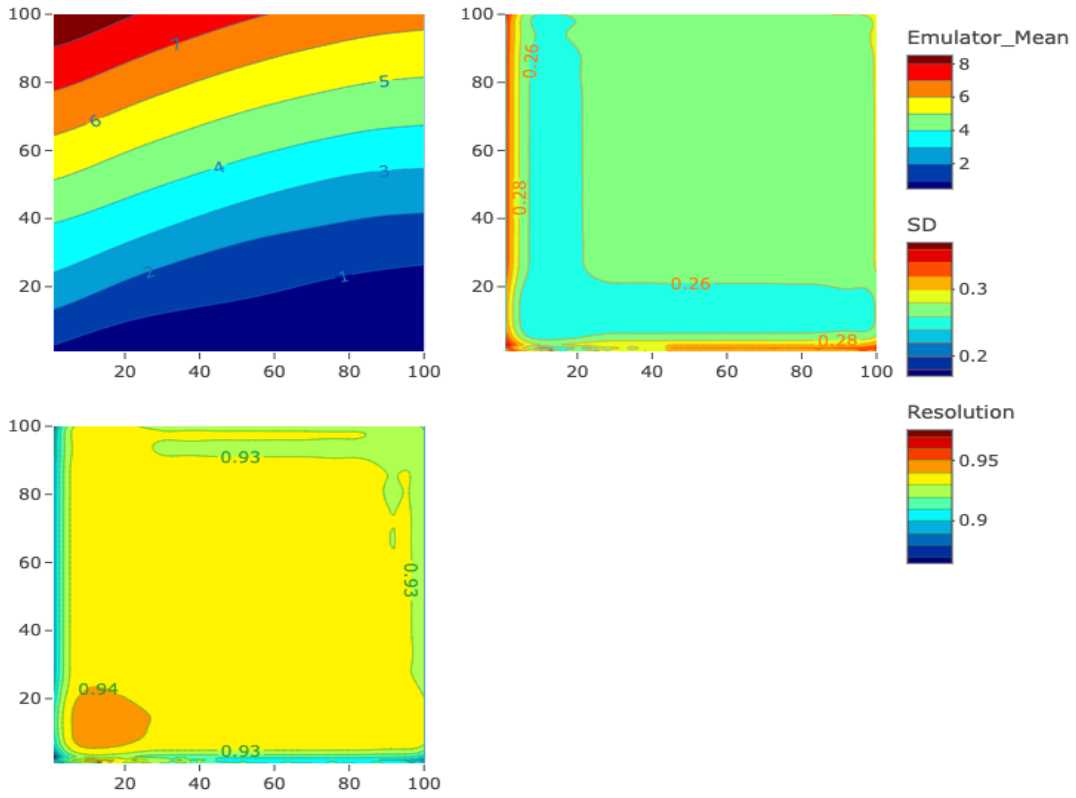


Figure 7.2: Upper Panel: Adjusted Emulator Expectation (Left) And Standard Deviations (Right) for P_p as a function of Nitrogen (N) on the x-axis and Phosphorus (P) on the y-axis using $St = 5, So = 6, Wy = 1$; Lower Panel: Resolution for P_p.

From diagnostics in the resolution plot from Figures 7.1 and 7.2, it shows high flat

resolution ($> 90\%$) all over the input space for both pollutants. So the emulator is very confident in explaining the variability of the simulator for more prediction points. Figure 7.3 shows the result of standardized prediction errors (SPE) for the two pollutants N_p (Right) and P_p (Left). From the result for P_p, we can see two points are outside the band for the low values of P, which is negligible in practice. The right panel shows the result of SPE for N_p, and no points are outside the band ± 2. The values for the N_p are scattered all over the input space indicating the emulator is not under-confident like crop yield. Considering the two plots, it is clear that the emulator and simulator have no conflict, hence the valid emulators for both pollutants for further analysis.

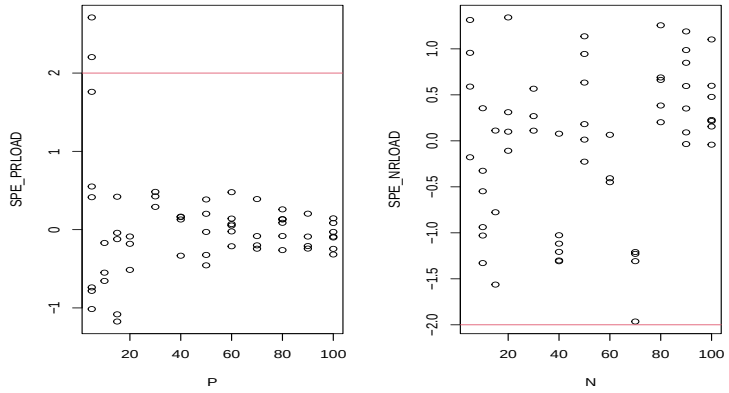


Figure 7.3: Standardized Prediction Errors for P_p (Left) and P_p (Right)

Overall, the adjusted emulator standard deviations plot results of pollutants show the same trend as the yield plot in Chapter 5. The resolution plots can explain most of the simulators' variability for yield and pollutants outputs. Finally, the plots of SPE for all output variables of interest also showed no evidence of poor fits indicating the validity of the emulators.

7.7 Sensitivity Analysis of the Utility Parameters

This section explores the values of b_1 and b_2 suitable for the utility and implausibility measurement. To calculate the expected utility and variance from Equations (7.11) and (7.12b) respectively, we need to specify values for the utility coefficients b_0, b_1 and b_2. For an emulator of Spring Barley, the value for the coefficient b_0 corresponding to yield is chosen to be $b_0 = 0.15$, which is the wholesale barley price in British pounds sterling

per kg for the year 2022 [148]. However, it is impossible to determine the values for the pollutants coefficients b_1 and b_2 similarly; instead, we must make a subjective specification. We need to gain prior knowledge and perform a sensitivity analysis to assess the impact of the possible options.

Using the emulator for the crop yield of Spring Barley from Chapter 5 and the pollutant emulator from Section 7.6, we can calculate the expected utility and implausibility for the continuous inputs from Equation (7.13). We fix $b_0 = 0.15$ and assume a domain from 0 to 0.3 for the coefficients b_1 and b_2 with an equal space of 0.02 distance, which generates 225 evenly spaced grid points as a part of explorations. It is noted that the mean for N_p is 13.625, approximately three times higher than the yield mean of 4.21, and the average for pollutant P_p is 4.89. Also, the standard deviation of N_p is 6.42, about 3.5 times the standard deviation of 1.99 for P_p and 1.82 for yield. In Table 7.1, we present the results of different combinations as a subset of 225 combinations of $b_0,\ b_1,\ b_2$. Column five of Table 7.1 shows the range of the implausibility values $I_i(x)$, column six shows the results for a total number of implausibility observations less than cut point 3, and the last two columns reveals the regions for the N and P values when $I_i(x) < 3$.

From the output of Table 7.1 when we penalize N_p such that the value of b_1 is higher than b_0 and b_2 in rows 8, we get lower maximum utility and higher implausibility values because substantial N_p penalty pushes many values further from the maximum. On the other hand, when we penalize P_p in row 7 such that the value of b_2 is higher than b_0 and b_1, we also get negative maximum utility values but a broader region of implausibility for N with high values. If we penalise both pollutants in row 6, we can also see a departure from the maximum such that a high negative maximum utility with a wider implausibility range for N. When we give the same weight to all coefficients in row 3, we can still see a departure from the maximum, as two of the three emulators don't depend on P. Providing very minimal penalties to both pollutants in row 2 causes a high positive maximum utility value. However, considering the region for all P values is not an expected feature as P has no effect. If we give the same weight to the yield and N_p with a minimal penalty for P_p in row 1, we still see a negative utility with a broad range of P implausibility values. Finally, when we give the same weight for yield and P_p coefficient and minimal penalties for N_p in rows 4 and 5, we can see positive maximum utility values with a narrow range for both N and P, which is the desired feature of the emulators.

Table 7.1: Sensitivity Analysis for the Coefficients b_0, b_1 and b_2

Row	b_0	b_1	b_2	U^*	$I_i(x)$	$I_i(x) < 3$	Region of N	Region of P
1	0.15	0.15	0.01	-0.44	$[0, 13.79]$	5143	$[46, 100]$	$[1, 100]$
2	0.15	0.01	0.01	0.85	$[0, 9.47]$	4824	$[47, 100]$	$[1, 100]$
3	0.15	0.15	0.15	-0.44	$[0, 19.10]$	1892	$[43, 100]$	$[1, 49]$
4	0.15	0.01	0.15	0.85	$[0, 17.55]$	1760	$[40, 100]$	$[1, 44]$
5	0.15	0.07	0.15	0.30	$[0, 19.03]$	1727	$[42, 100]$	$[1, 45]$
6	0.15	0.29	0.29	-1.74	$[0, 19.48]$	1860	$[38, 100]$	$[1, 45]$
7	0.15	0.15	0.27	-0.45	$[0, 20.18]$	1642	$[38, 100]$	$[1, 41]$
8	0.15	0.27	0.15	-1.55	$[0, 17.63]$	2365	$[42, 100]$	$[1, 59]$

Figure 7.4 illustrates the expected utility surface for the eight rows in Table 7.1. The general trend of all eight combinations of $b_0,\ b_1$, and b_2 reveals high utility for the high N and low P values. However, the first plot of column 1 for row 1 shows trapezium shape for the region of maximum utility and gives the region of high maximum utility for all the P input space, which is unrealistic as the input P has no effect on crop yield. For the first plot of column 2 corresponding to row 2, which is equal weight for the pollutants of no penalties the maximum utility area covers approximately 60% of the region for P response, which is also not the desired feature.

For all other plots from rows 3 to 8, we can see a similar shape such that maximum utility area covers the high N values and low P values. We can see a very narrower area for P for the equal weight of pollutant P_p and yield in row 4 (second plot of column 2), which is mostly desired. As mentioned earlier, we can see the very close mean and standard deviation for yield and pollutant P_p, and we are also getting the desired feature by considering the equal weight for both. So for further analysis, we will use the values of $b_0 = 0.15,\ b_1 = 0.01,\ b_2 = 0.15$.

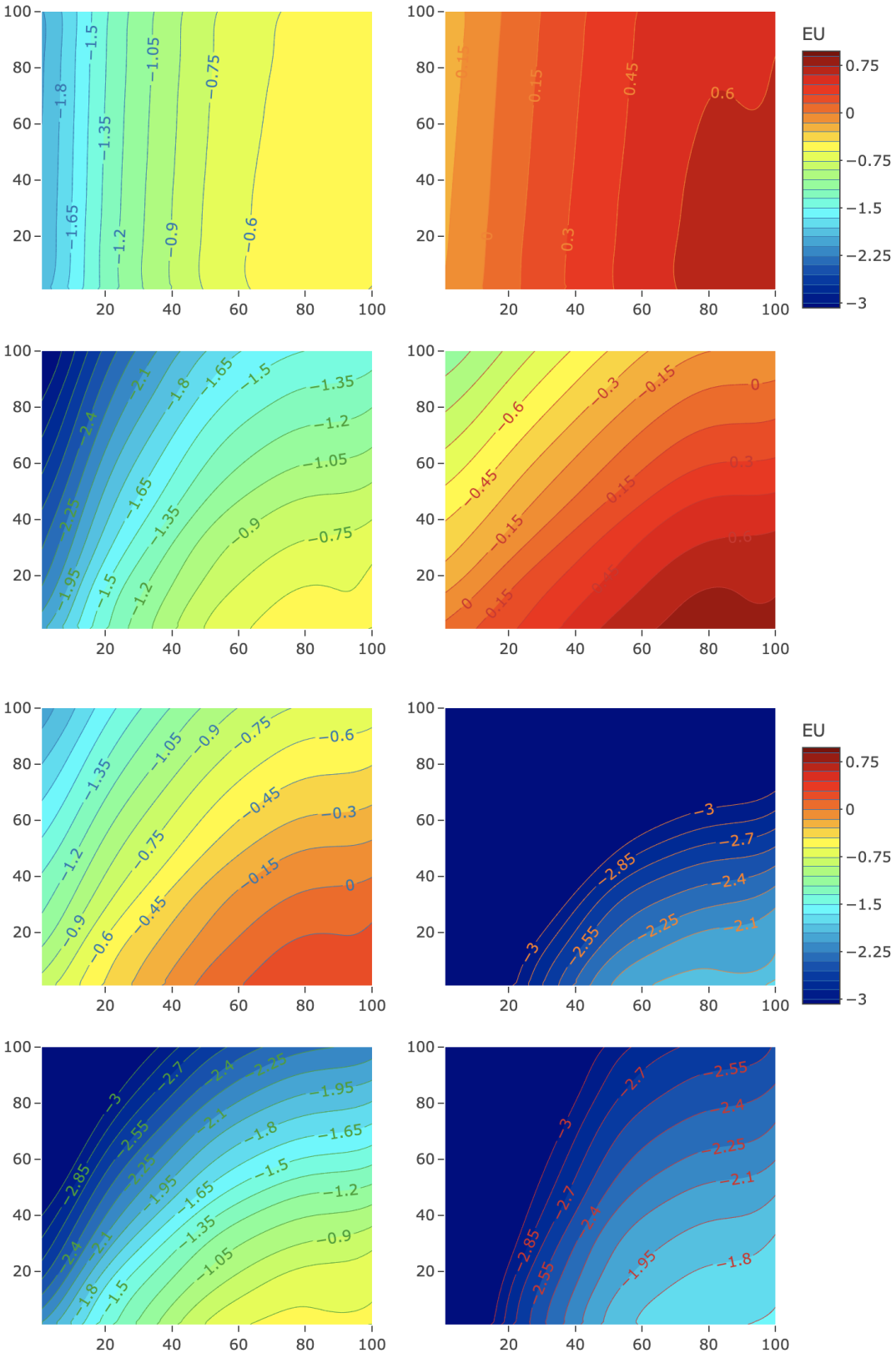


Figure 7.4: Expected Utility Plot for the Subset of Eight Unique Combinations; Column 1: Plots for Odd Rows; Column 2: Plots for Even Rows

7.8 Utility Emulation and Implausibility

This section concerns the utility emulation and implausibility for the continuous and mixed inputs. Initially, we show the result for the continuous input and then built the emulators for mixed inputs.

7.8.1 Continuous Inputs Only

In this section we fit the emulated utility for Spring Barley, including the two pollutants N_p and P_p for continuous inputs by fixing $St = 5$, $So = 6$,$Wy = 1$ and $Sy = 4$. We then evaluate the implausibility for a 100×100 of points using $b_0 = 0.15,\ b_1 = 0.01,\ b_2 = 0.15$. Figure 7.5 presents the results of expected utility and variance with the implausibility and the region for the implausibility using the threshold $(I(x) < 3)$.

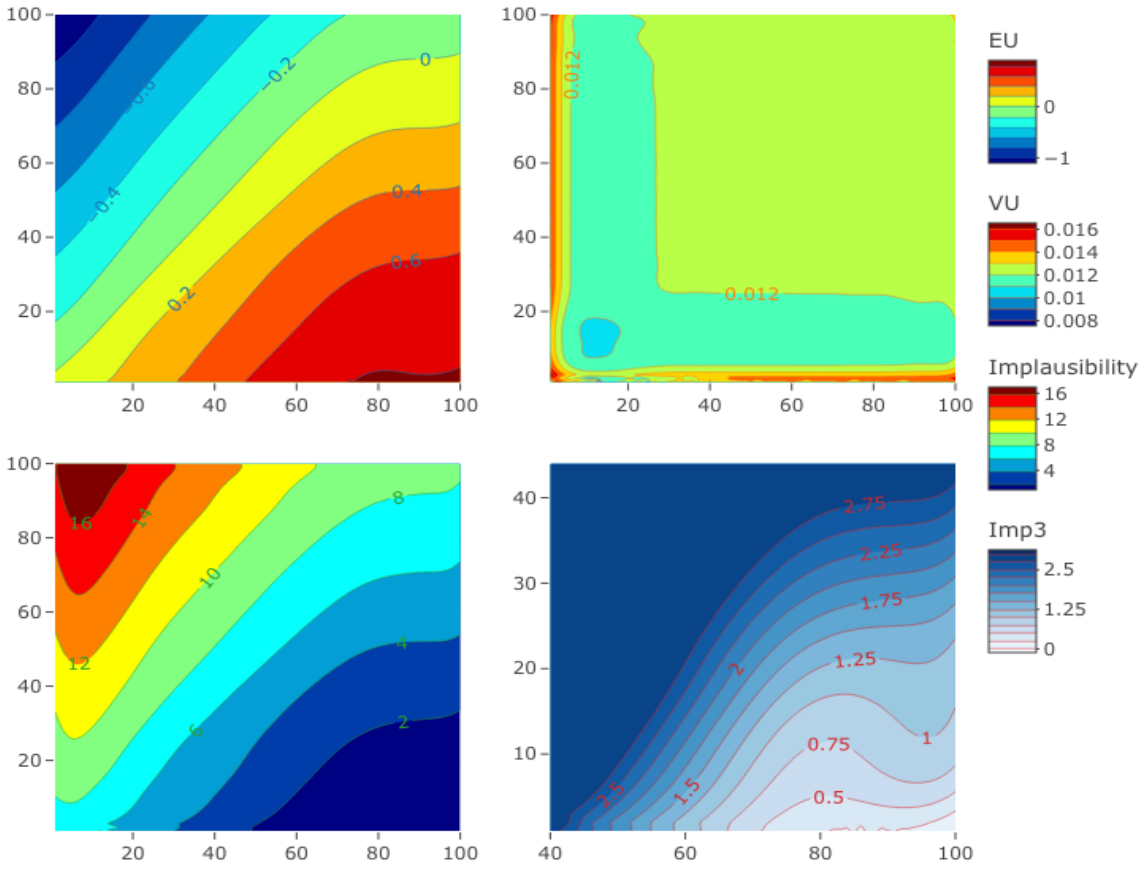


Figure 7.5: Upper panel: Utility Expectation (left) and Variance as a function of Nitrogen (N) on x-axis and Phosphorus (P) on y-axis; Lower panel: Implausibility (left) and the Region of Implausibility for $(I_x < 3)$ with $b_0 = 0.15,\ b_1 = 0.01$ and $b_2 = 0.15$.

From Figure 7.5 we can see the region of maximum expected utility for high values of Nitrogen and very low values of Phosphorus, which is the desired feature of the crop Spring Barley. As we have seen a flat response for P, so increasing P has no benefit, only penalties. We can see a narrow uncertainty for the variance of the utility over the whole input space. The lower panel of Figure 7.5 of the implausibility confirms that the implausibility is higher for high P values and lower N. The reason behind this is that we have a low yield on the lower values of N and very high pollution for P_p in that region. The right lower panel shows the implausibility refocused on the region for which $(I(x) < 3)$ denoting the non-implausible region with respect to N and P. For $b_0 = 0.15,\ b_1 = 0.01$ and $b_2 = 0.15$ we can approximate the non-implausible region by the rectangle $N \in [40, 100]$ and $P \in [1, 44]$.

7.8.2 Mixed Inputs Including Steepness and Soil

In this section, we will construct the correlation matrices and then build the Bayes linear mixed inputs emulators for the pollutants N_p and P_p considering the soil and steepness factors. Finally, we check the robustness of the emulators based on the diagnostic resolution and SPE.

7.8.2.1 Emulating Pollutants

We consider the linear mean function for the pollutant P_p with the interaction of N and P, which is shown in Figure 7.2. For the pollutant N_p, we assume the linear regression terms of N and P and the interaction effect of N with the factors of soil and steepness as we have seen that it has a solid response to N as an increasing trend for low values and then monotonic increase for large N values.

Table 7.2 summarises the mean function for N_p and P_p with inputs steepness, and soil with quantitative variables Nitrogen, Phosphorus and simple basis in terms of 9 and 14 regression coefficients for the two pollutants.

Table 7.2: Basis for Pollutants Considering Inputs Nitrogen and Phosphorus With Factors Steepness and Soil

Pollutant	Simple Basis $[g(v)^T]$	No. of Coefficients
P_p	$1, N, P, NP, Z_2^{So}, Z_3^{So}, Z_2^{St}, Z_3^{St}, Z_4^{St}$	9
N_p	$1, N, P, NP, Z_2^{So}, Z_3^{So}, Z_2^{St}, Z_3^{St}, Z_4^{St}$ $NZ_2^{So}, NZ_3^{So}, NZ_2^{St}, NZ_3^{St}, NZ_4^{St}$	14

Considering a correlated error with squared exponential covariance function for the continuous variables, we construct the factor level correlation matrix for factor soil T_{So} and for factor steepness T_{St} using the general correlation approach in Equation (6.9). The estimated matrices T_{So} and T_{St} are shown as follows, following the methods of Chapter 6;

$$T_{So} = \begin{bmatrix} \tau_{1,1} & \tau_{1,2} & \tau_{1,3} \\ \tau_{1,2} & \tau_{2,2} & \tau_{2,3} \\ \tau_{1,3} & \tau_{2,3} & \tau_{3,3} \end{bmatrix} = \begin{bmatrix} 1 & 0.31 & 0.12 \\ 0.31 & 1 & 0.22 \\ 0.12 & 0.22 & 1 \end{bmatrix} \tag{7.16}$$

$$T_{St} = \begin{bmatrix} \tau_{1,1} & \tau_{1,2} & \tau_{1,3} & \tau_{1,4} \\ \tau_{1,2} & \tau_{2,2} & \tau_{2,3} & \tau_{2,4} \\ \tau_{1,3} & \tau_{2,3} & \tau_{3,3} & \tau_{3,4} \\ \tau_{1,4} & \tau_{2,4} & \tau_{3,4} & \tau_{4,4} \end{bmatrix} = \begin{bmatrix} 1 & 0.85 & 0.76 & 0.62 \\ 0.85 & 1 & 0.78 & 0.80 \\ 0.76 & 0.78 & 1 & 0.93 \\ 0.62 & 0.80 & 0.93 & 1 \end{bmatrix} \tag{7.17}$$

We can see a weak correlation between the factor levels from the correlation matrix T_{So} for factor soil. On the other hand, in the correlation matrix T_{St} for factor steepness, we can see a strong positive correlation among the factor levels for most cases.

To build the Bayes linear emulators for both pollutants, we assume the same hyper-parameters of $\widehat{\theta}_N = 0.017$, $\widehat{\theta}_P = 0.024$ and nugget $\widehat{\delta} = 0.054$ similar to the continuous input problem with the above-constructed correlation matrices of T_{So} and T_{St} for the factors steepness and soil only. Expectations $E[f(v)]$ and variances $Var[f(v)]$ for the emulators used in Table 7.2 are calculated by using Algorithm 2 for mixed inputs from Chapter 6. The emulator is updated by using the 15×15 simulation grid for functions of N and P. Figure 7.6 shows the adjusted emulator mean and standard deviations with resolution considering the basis functions from Table 7.2 for the pollutants.

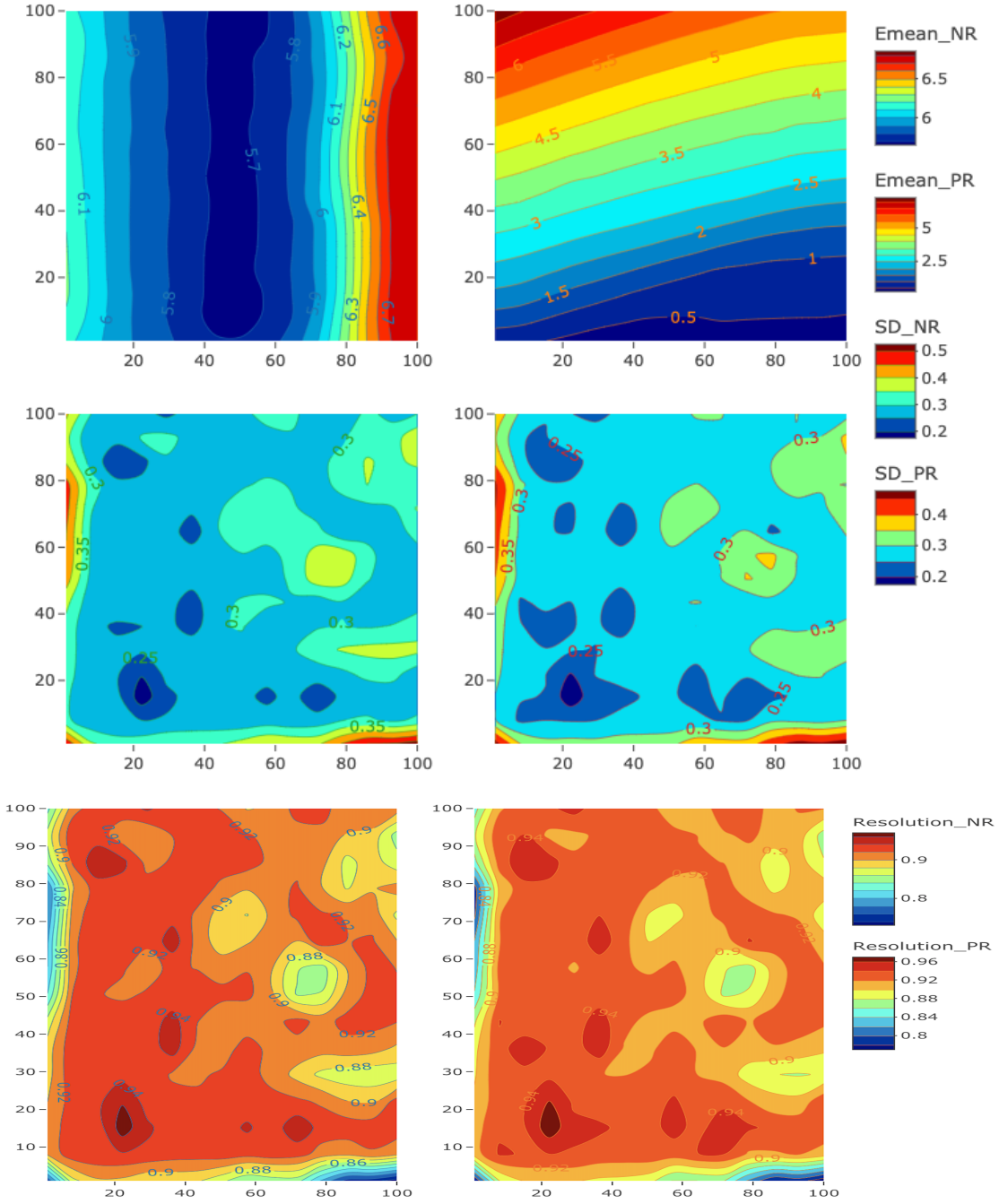


Figure 7.6: Upper panel: Emulator expectations for N_p (left) and P_p (right); Middle panel: Standard deviations of the pollutants; Lower panel: Resolution Considering the Factor Steepness and Soil only fixing $St = 5$, $So = 6$, $Wy = 1$ and $Sy = 4$.

Figure 7.6 shows that N_p (upper left panel) increases with increasing N levels as in the continuous input problem. However, the trend appears to be decreasing from a spuriously high level of N_p at very low levels of N before behaving the desired increasing feature. Further exploration is needed with this input space to catch behaviour. The upper right

panel shows the adjusted emulator expectation for P_p, which shows the same trend for continuous inputs. The standard deviation plots in the middle panel show entirely different features from the continuous input results, with slightly higher uncertainty around the locations for both pollutants over the input space after introducing the factor effects into the model. From diagnostics of the resolution for both pollutants, we can see a different feature than the continuous problem. However, it's not flat but shows high resolutions greater than 0.8 all over the input space indicating the emulators are very confident in explaining the variability of the simulator.

7.8.2.2 Utility and Implausibility for Steepness and Soil

In this section, we calculate the expected utility and variance corresponding to the emulators for crop yield of Spring Barley from Chapter 6 and the two pollutants considering factors of steepness and soil from Section 7.8.2.1. Figure 7.7 shows the expected utility and variance in the upper panel using the steepness and soil factors fixing $Wy = 1$ and $Sy = 4$. The lower panel shows the maximum implausibility and the region of maximum implausibility for which $I_M(x) < 3$.

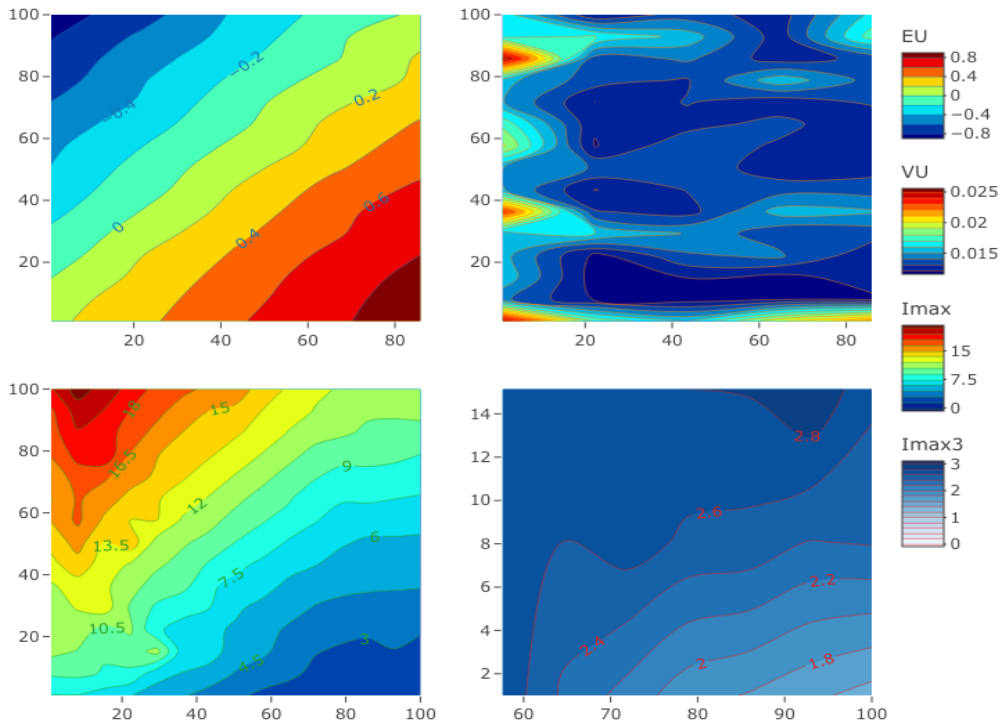


Figure 7.7: Upper panel: Expected utility (Left) and Variance (Right); Lower panel: Maximum Implausibility (Left) and the Region of Maximum Implausibility (Right).

From Figure 7.7, we can see the expected utility is maximised for high values of N and low values of P, which is the same trend as the continuous input problem. The variance of the utility plot shows in Figure 7.7, which is showing quite different trend from the continuous input problem with slightly higher uncertainty values. This unusual variations may happen due to the factor effect, where more data points were used to emulate and the parameter space are not flat like as continuous-only problem.

The left lower panel in Figure 7.7 shows the maximum implausibility over the factors. The result confirms that the maximum implausibility is higher for high P values and for low values of N. The region with $I_M(x) < 3$ suggests an approximate non-implausible region of $N \in [60, 100]$ and $P \in [1, 15]$. This result gives us the same evidence as the previous Chapters that the Spring Barley crop has the highest yield for high N and low P values. Figure 7.8 shows the plot of implausibility for three different unique combinations (i) $St = 5,\ So = 6$, (ii) $St = 5,\ So = 7$, and (iii) $St = 6,\ So = 6$, with the lower right panel the maximum implausibility $I_M(x)$ for all steepness and soil. These plots are similar to the continuous only problem in Section 7.6.

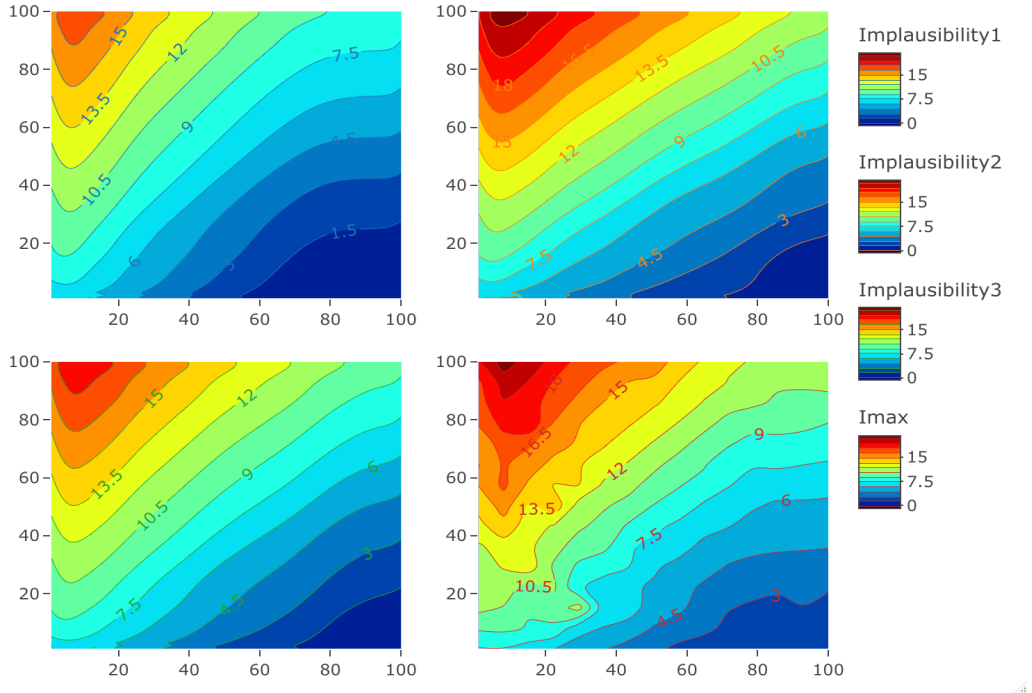


Figure 7.8: Upper Panel: Implausibility Plots for the Combinations $St = 5,\ So = 6,\ Wy = 1$ (Left) and $St = 5,\ So = 7, Wy = 1$ (Right); Lower Panel: Implausibility Plot (Left) for the Combination $St = 6,\ So = 6,\ Wy = 1,\ Sy = 4$ and Maximum Implausibility (Right) $(I_M(v))$.

The upper left panel plot for implausibility when $St = 5,\ So = 6,\ Wy = 1,\ Sy = 4$ shows a similar trend and shape as the continuous-only problem. From the plots in the upper panels for the different factors, we can see a similar shape and trend of implausibility plots, which suggests implausibility is always higher for high P values across the factors. The right lower panel is the plot of the maximum implausibility, which exhibits the same general shape and trend as the other unique combinations. So similar results for different factor combinations indicates that some factors have no effect, which motivated us to perform variable selection to remove non-essential terms.

7.8.3 Mixed Inputs Including Weather

7.8.3.1 Selection of Inputs Using Stepwise Regression

In this section, we will apply variable selection methods to remove some basis terms irrelevant to the emulation process, thus reducing $g(v)$ and selecting the essential regression parameters. For this, we will use stepwise regression.

Stepwise regression is a system to construct a model of essential inputs using an inclusion and exclusion process in a systematic way [9]. Three different ways are available to identify the important inputs:

- Forward Stepwise Selection: This selection method starts with no variables in the model, then adds each of the variables using the model selection criterion and repeats the procedure until there is no improvement of the model.
- Backward Stepwise Selection: This method starts with all variables, deletes each term, and repeats until there is no improvement in the model.
- Both-Direction Stepwise Regression: This method combines the above two selection approaches to eliminate and add variables simultaneously.

Table 7.3 shows the result for the forward stepwise regression of the essential variables considered first, and the relevant AIC values corresponding to the three outputs N_p, P_p and Yield with the inputs steepness, soil, weather, Nitrogen and Phosphorus. We have considered the intercept term as the null model and all other inputs as the alternative.

The forward selection method and both-way stepwise regression give the same results presented in Table 7.3. The results for all outcome variables are showing in Table 7.3, but we only discuss the input P_p.

Table 7.3: Inputs Selection using Forward Stepwise Regression for N_p, P_p and Yield

N_p		P_p		Yield	
$g(v)$	AIC	$g(v)$	AIC	$g(v)$	AIC
Null	22802.27	Null	22802.27	Null	9755.25
Z_8^W	22336.98	P	13895.25	N	3120.54
Z_7^W	21782.51	Z_8^W	12479.89	Z_7^W	−1576.87
Z_6^W	21306.43	Z_7^W	10597.76	Z_3^W	−6715.84
N	21003.17	Z_6^W	8531.70	Z_8^W	−8358.41
Z_2^{So}	20810.91	N	6738.89	Z_6^W	−9260.72
Z_4^W	20744.88	Z_2^{So}	5492.25	N^3	−10095.47
Z_3^{So}	20702.79	Z_4^W	5030.50	N^2	−10893.16
Z_3^W	20688.86	Z_3^{So}	4612.48	Z_5^W	−11761.68
Z_5^W	20688.05	Z_3^W	4468.19	Z_3^{So}	−12220.62
NZ_7^W	20659.76	Z_5^W	4348.39	N^2P	−12255.30
Z_2^W	20655.33	NP	4240.04	Z_4^W	−12271.71
NZ_3^{So}	20653.04	Z_2^W	4158.25	Z_2^W	−12277.21
NZ_6^W	20652.79	Z_4^{St}	4154.62	Z_2^{So}	−12291.07
NZ_8^W	20652.32			NP	−13001.49
NZ_4^W	20648.10			NP^2	−13059.00
NZ_2^{So}	20647.81			Z_2^{St}	−13090.01
NZ_2^W	20647.37			P	−13151.23
P	20643.24			Z_4^{St}	−13155.14
				Z_3^{St}	−13165.04
				P^2	−13227.75
				P^3	−13240.21

Firstly we fit the intercept-only model, which has an AIC value of 22802.27. Secondly, we identify the model with the lowest AIC compared to the null model, so input P with the AIC value of 13895.25 is added. After this, we need to fit every viable two-term model,

including P and select the model that produced the lowest AIC compared to the one-term AIC value. Then we need to consider the three-term models and so on with the comparison to the previous AIC value, and the procedure stops when the model fails to reduce the AIC such that there is no improvement in the model. The result shows the expected feature of P_p that the variable was P have added first as well N and its interaction term NP.

For the pollutant variable N_p, we can see that input steepness is never added by the forward selection method, indicating that the input steepness is unimportant for this outcome variable. The levels of the factor variable steepness shows high dependency among them for N_p, which is similar to result in Chapter 4. The interaction N with weather levels $Z^W_{3,3}$, $Z^W_{3,5}$ are the same as baseline and hence not included in the final model. From the result, we can also see that the variable P is included by the selection method as the last choice, which is the same conclusion as for N_p, such that there is a weak effect of P.

For the outcome P_p, we can see that the selection method has not included all input steepness levels except Z^{St}_4. So we discount the steepness variable for the emulation, and the result demonstrated that the input N and NP are included by the model for outcome P_p, which shows the same feature from the baseline figures.

For the outcome yield, steepness has some significant effects in the context of the other two outcome variables. The result illustrated that the model selection for the yield included inputs P as the last choice and N as the first choice, which is the main contributor to the yield. So our results show that steepness does not affect both pollutants, but does effect yield. Now we will fit the emulators using the selected inputs for the pollutants only.

7.8.3.2 Emulation of Pollutants

In this section, we fit the emulators for both P_p and N_p using the selected inputs from Table 7.3. A simple basis of 13 and 19 regression coefficients in terms of two pollutants are shown in Table 7.4 for the selected inputs. The design matrix is calculated in the same way as Chapter 6.

Table 7.4: Basis for Pollutants Considering Inputs N and P with Selected Inputs

Pollutant(s)	Simple Basis $[g(v)^T]$	No. of Coefficients
P_p	$1, N, P, NP, Z_2^{So}, Z_3^{So}, Z_2^W, Z_3^W, Z_4^W, Z_5^W, Z_6^W, Z_7^W, Z_8^W$	13
N_p	$1, N, P, Z_{1,2}^{So}, Z_{1,3}^{So}, Z_{3,2}^W, Z_{3,3}^W, Z_{3,4}^W, Z_{3,5}^W, Z_{3,6}^W, Z_{3,7}^W, Z_{3,8}^W$ $NZ_{1,2}^{So}, NZ_{1,3}^{So}, NZ_{3,2}^W, NZ_{3,4}^W, NZ_{3,6}^W, NZ_{3,7}^W, NZ_{3,8}^W$	19

Using the general correlation approach, we construct the factor level correlation matrix by considering a correlated error with a squared exponential covariance function. We use the same correlation matrix from Equation (7.16) for factor soil. For factor weather, the correlation matrices $T_W(N_p)$ and $T_W(P_p)$ are constructed using the general correlation approach as in Chapter 6. Equations (7.18) and (7.19) show the correlation matrices for the pollutants N_p and P_p with factor weather, respectively, and both are symmetric. The matrices $T_W^{(N_p)}$ and $T_W^{(P_p)}$ are as follows;

$$T_W^{(N_p)} = \begin{array}{c} \begin{matrix} \tau_{3,1} & \tau_{3,2} & \tau_{3,3} & \tau_{3,4} & \tau_{3,5} & \tau_{3,6} & \tau_{3,7} & \tau_{3,8} \end{matrix} \\ \begin{pmatrix} 1 & & & & & & & \\ 0.33 & 1 & & & & & & \\ 0.19 & 0.23 & 1 & & & & & \\ 0.32 & 0.12 & 0.25 & 1 & & & & \\ 0.17 & 0.15 & 0.30 & 0.23 & 1 & & & \\ 0.42 & 0.28 & 0.28 & 0.4 & 0.16 & 1 & & \\ 0.32 & 0.19 & 0.26 & 0.19 & 0.30 & 0.14 & 1 & \\ 0.22 & 0.26 & 0.32 & 0.14 & 0.19 & 0.16 & 0.13 & 1 \end{pmatrix} \end{array} \begin{matrix} \tau_{3,1} \\ \tau_{3,2} \\ \tau_{3,3} \\ \tau_{3,4} \\ \tau_{3,5} \\ \tau_{3,6} \\ \tau_{3,7} \\ \tau_{3,8} \end{matrix} \tag{7.18}$$

$$T_W^{(P_p)} = \begin{array}{c} \begin{matrix} \tau_{3,1} & \tau_{3,2} & \tau_{3,3} & \tau_{3,4} & \tau_{3,5} & \tau_{3,6} & \tau_{3,7} & \tau_{3,8} \end{matrix} \\ \begin{pmatrix} 1 & & & & & & & \\ 0.26 & 1 & & & & & & \\ 0.68 & 0.37 & 1 & & & & & \\ 0.30 & 0.32 & 0.29 & 1 & & & & \\ 0.31 & 0.49 & 0.51 & 0.35 & 1 & & & \\ 0.43 & 0.45 & 0.29 & 0.25 & 0.31 & 1 & & \\ 0.31 & 0.60 & 0.21 & 0.48 & 0.37 & 0.26 & 1 & \\ 0.35 & 0.27 & 0.51 & 0.22 & 0.20 & 0.36 & 0.32 & 1 \end{pmatrix} \end{array} \begin{matrix} \tau_{3,1} \\ \tau_{3,2} \\ \tau_{3,3} \\ \tau_{3,4} \\ \tau_{3,5} \\ \tau_{3,6} \\ \tau_{3,7} \\ \tau_{3,8} \end{matrix} \tag{7.19}$$

From the correlation matrix (7.18) for the factor input weather concerning N_p, we can see that all values are below 0.43, indicating a weak correlation between the factor levels. On the other hand, in the correlation matrix $T_W^{(P_p)}$ for pollutant P_p, we can see moderate to weak correlations among the factor levels most of the time. However, we can perhaps see a slightly stronger correlation between the factor level 1 and 3 and levels 2 and 7.

To build the Bayes linear emulators for both pollutants, we consider correlation matrices for the factor weather and soil constructed with the same hyper-parameters as for Section 7.6 .

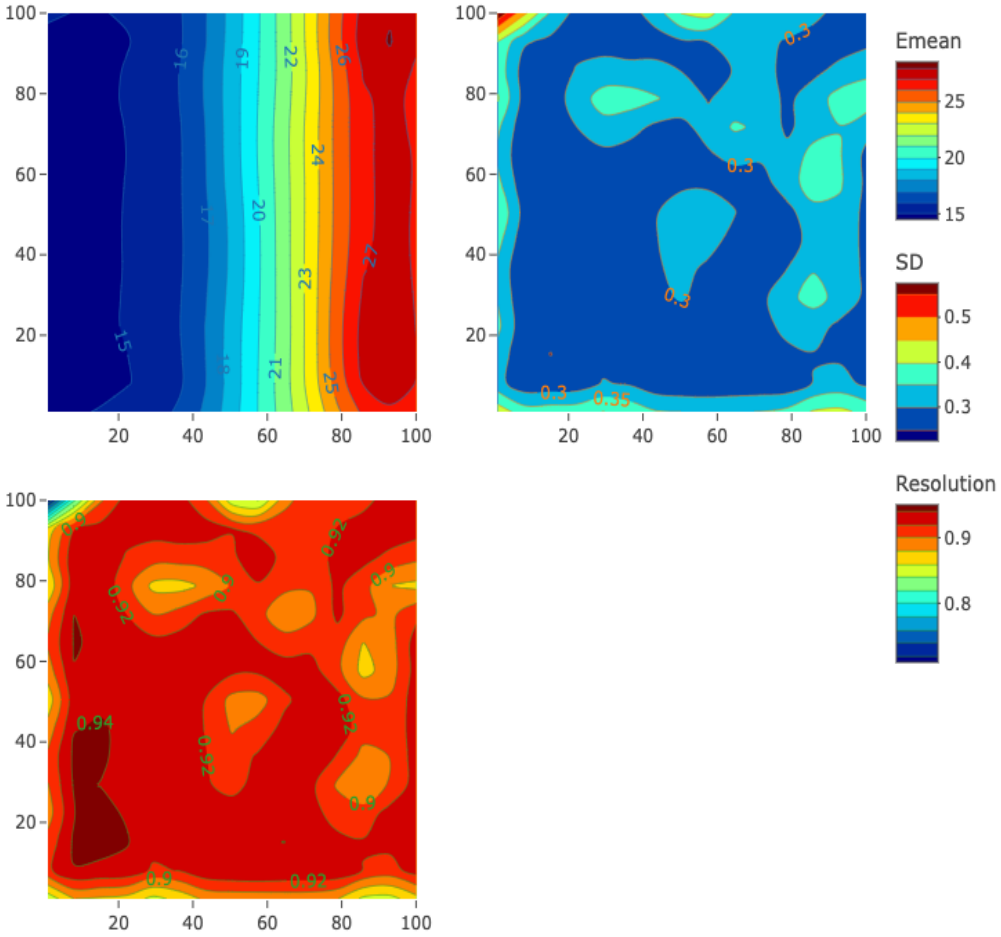


Figure 7.9: Upper panel: Adjusted Emulator expectation for N_p (left) as a function of Nitrogen (N) on the x-axis and Phosphorus (P) on the y-axis; Emulator Standard Deviations (right) of the N_p; Lower panel: Resolution of the N_p.

From Figure 7.9, we can see the adjusted emulator mean (upper left panel) for N_p, which shows the desired feature of an increasing trend with an increase of Nitrogen values

and the unusual trend for steepness and the soil-only problem has now disappeared. The upper left panel in Figure 7.10 shows the adjusted emulator expectation for P_p. The plot shows a clear increasing trend concerning P values such that high pollution for higher P and for lower N values. The trend of P_p shows the same direction and interaction with N, similar to the problem of continuous input, and steepness and soil only problem. The effect of P on N_p shows a weak dependency, which is an expected pattern.

The adjusted emulator standard deviations plots from the upper right panel of Figures 7.9 and 7.10 illustrate a high uncertainty around the locations for both pollutants over the input space after considering the weather factor. From diagnostics of the resolution for both pollutants, we can see high flat resolution over the input pace, greater than 0.9 for most of the predictions. It indicates that the emulators can explain most of the variability of the simulator.

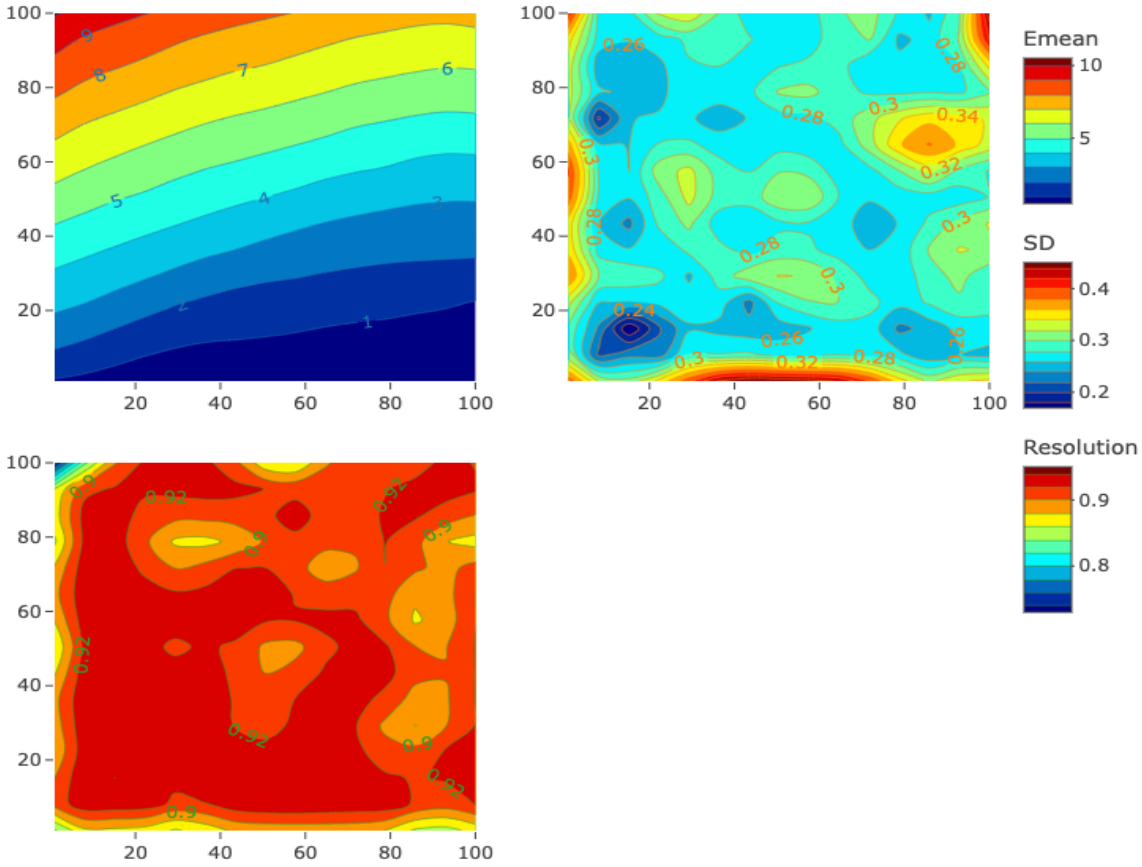


Figure 7.10: Upper panel: Adjusted Emulator Expectation for P_p (Left) as a function of Nitrogen (N) on the x-axis and Phosphorus (P) on the y-axis; Emulator Standard Deviations (Right) of the P_p; Lower panel: Resolution of the P_p (Left).

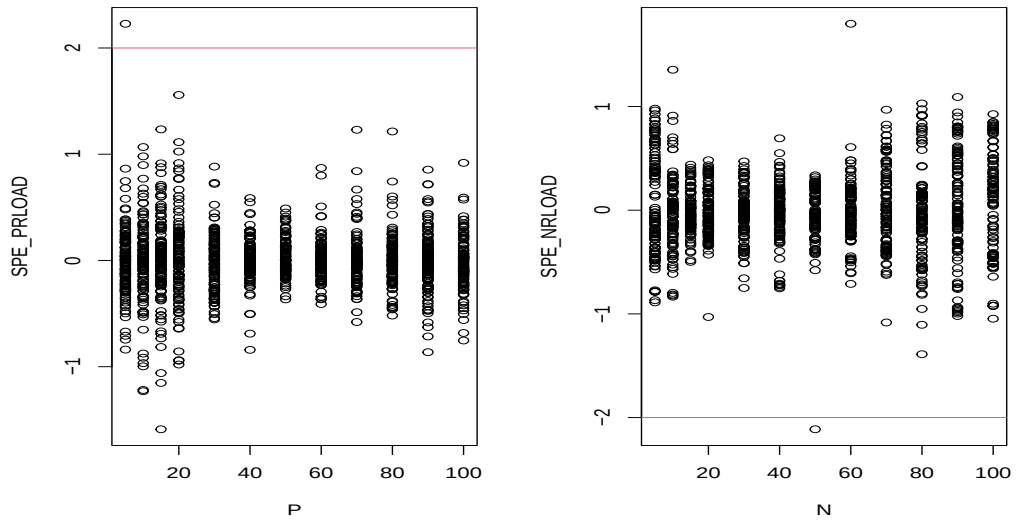


Figure 7.11: SPE Plot for Selected Inputs of P_p (Left) and N_p (Right)

Figure 7.11 shows the plot of SPE for both pollutants considering the 60% of the observations as training data and the rest of the data as testing. Both figures shows no concern about the conflict of the emulator and simulator. So we can say that our emulators for the selected input pollutants are valid and can be used for implausibility and utility.

7.8.3.3 Utility and Implausibility

We now construct the expected utility of crop yield of Spring Barley and the maximum implausibilities by using the values of $b_0 = 0.15$, $b_1 = 0.01$ and $b_2 = 0.15$. Figure 7.12 presents the result of implausibilities considering eight different weather levels.

The implausibility plots corresponding to weather levels show the same trends and shapes. From the figure we can see the non-implausible region for N values varying between $N \in [1, 100]$ but P values are within $P \in [1, 20]$. We see a smooth trend for all weather levels. However, we can see different ranges of the implausibilities for different weather levels, such as higher for weather levels 2 (row 1 right), 4 (row 2 right), 6 (row 3 right), and 7 (row 4 left). The pattern and trend for weather level 1 are similar to $St = 5,\ So = 6,\ Wy = 1,\ Sy = 4$ from Figure 7.5.

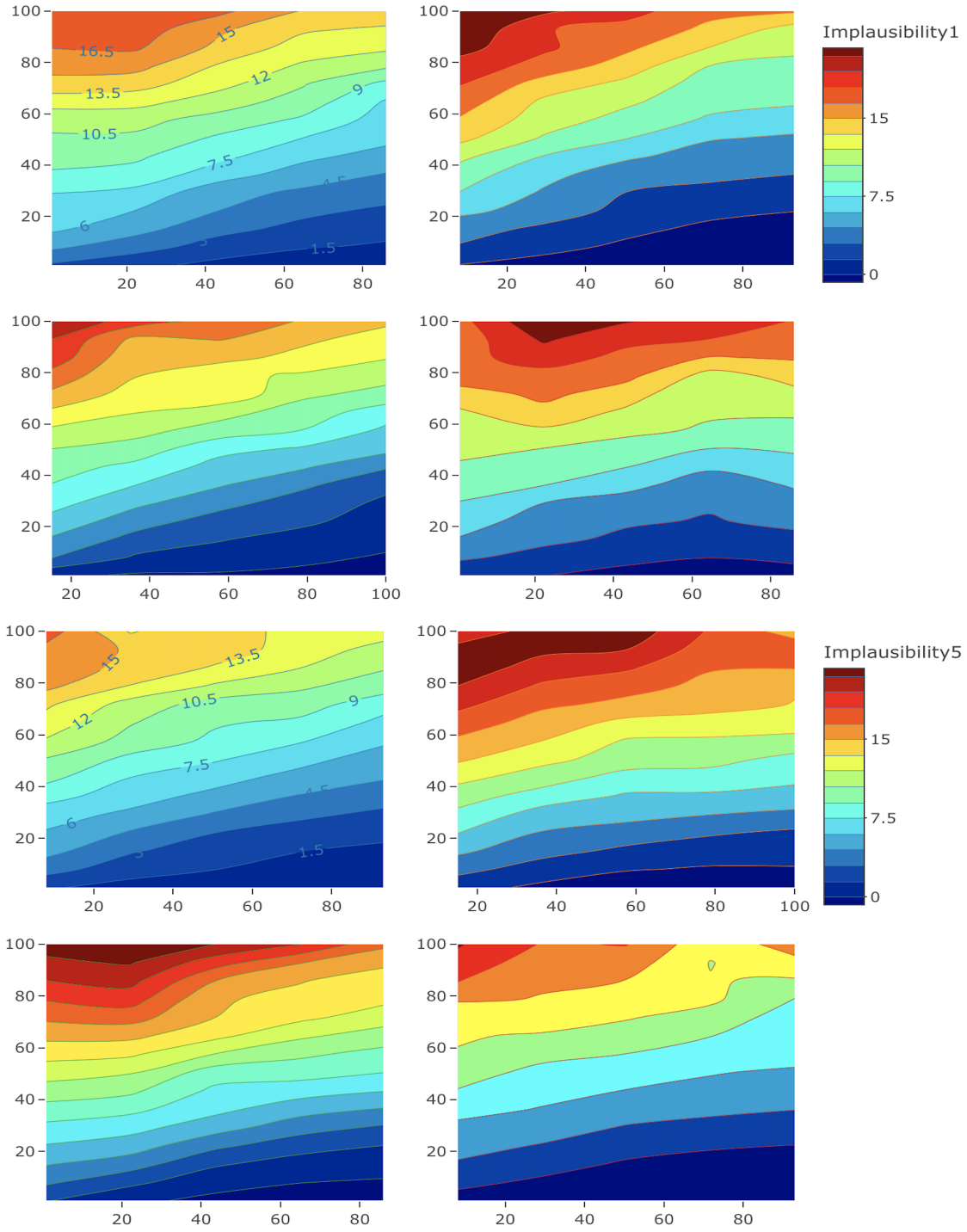


Figure 7.12: Implausibility ($I(x)$) plots for Weather levels 1-8 (Row 1 is for weather level 1 (Left), and 2 (Right)), Row 2 is for levels 3 and 4, Row 3 is for levels 5 and 6, Row 4 is for levels 7 and 8.

Both expected utility and variance show the same shape and trend, similar to the mixed inputs problem for steepness and soil only. In Figure, 7.13 upper panel shows the results

of the expectation and variance of the utility by fixing $St = 5,\ So = 6,\ Wy = 1,\ Sy = 4$. The lower panel shows plots of maximum implausibility and maximum implausibility for the threshold ($I_M(x) < 3$). From Figure 7.13, we can see the expected utility is highest for the region of $N \in [45, 100]$ and $P \in [1, 10]$. The uncertainty is slightly higher for the utility variance in Figure 7.13 over the input space of prediction points. The uncertainty is higher on the edges of the region but can be ignored due to minimal prediction points to emulate.

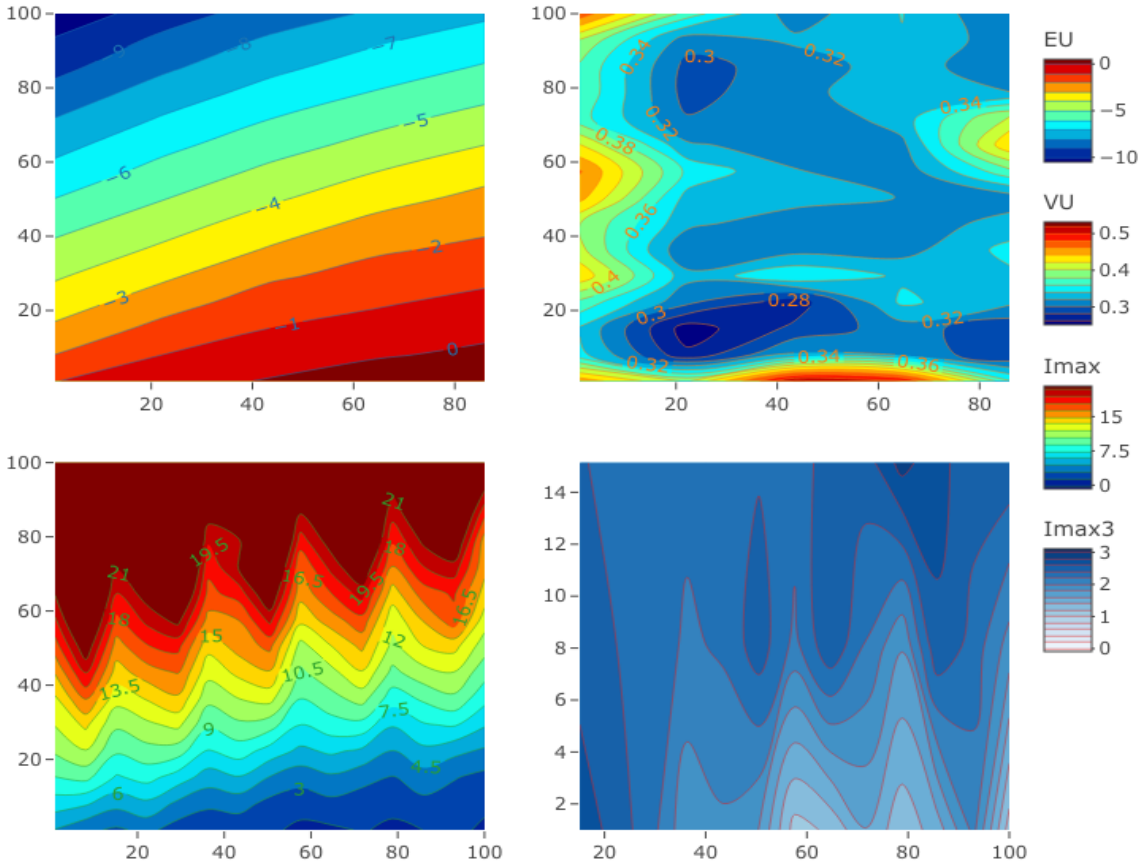


Figure 7.13: Upper panel: Expected utility (Left) and Variance of the Utility (Right); Lower panel: Maximum Implausibility and Zooming Region of Maximum Implausibility (Left).

The left lower panel from Figure 7.13 shows the maximum implausibility. The result confirms that the maximum implausibility is higher for high P and low N values. The shape of the plot is different and non-smooth, as we have seen that the weather levels show a diverse range of implausibilities and fluctuations. The right lower panel shows the region of maximum implausibility considering the threshold value of 3, confirming the

non-implausible area to be approximately $N \in [20, 100]$ and $P \in [1, 15]$. The region for N is wider, which causes high uncertainty. Weather is the main contributor to uncertainty, so further analysis is needed. If the factor steepness and soil values are known, it could be conditioned to reduce the uncertainty. One possible option is to look at the region for the threshold $I(x) < 2$.

This thesis aimed to make a framework for determining the best input values for the Spring Barley crop considering all the associated factors. Spring Barley gives the non-implausible region approximately at $N \in [20, 100]$ and $P \in [1, 15]$, which provides us with the best combinations of fertilisers to apply after accounting for all land characteristics and weather factors.

7.9 Conclusion

This chapter created a framework to seek maximum utility considering the input space of Nitrogen and Phosphorus for the crop Spring Barley.

We introduced the utility and history matching with implausibility measures and used a linear utility function. The difference between expected utility value and maximum expected utility value with variance of the expected utility used to calculate the maximum implausibility and hence to construct the emulators. The emulator's diagnostic plots illustrated narrow uncertainty around the prediction points and no conflict with the simulator. We assessed a sensitivity analysis for the coefficients related to yield and pollutants by fixing the value of b_0. Then we calculated the utility and implausibility for the continuous case and mixed inputs. We performed variable selection to identify the essential basis terms, including the factor weather, to reduce the extra complexity and to remove the unessential inputs. Finally, we built the emulators for the reduced inputs, hence finding the utility and implausibility.

From this chapter, we constructed a utility function to represent the value of a given output, and used it to find the region of N and P for mixed and continuous inputs. We learned the basics of implausibility and the maximum implausibility to determine the optimal area. We also understand how to perform the sensitivity analysis for the extended grid points and, finally, the stepwise regression method to select the best contributors. The further scope of this chapter is use different yield function to include the cost of N and P.

Chapter 8

Conclusion

This thesis aimed to develop a method to make an optimal decision for the farmers to gain maximum utility considering yield and pollutants, accounting for weather factors, land characteristics and fertilizer use. This research focuses on dealing with EPIC simulator outputs generated by both categorical and continuous inputs. In this thesis, we have explored three different views of the problem: a frequentist approach of crop modelling in Chapter 3, Bayesian hierarchical inference in Chapter 4, and subjective Bayesian emulation techniques in Chapters 5 to 7. Some of the significant achievements of this thesis are:

1. Extension of the Bayes linear emulation approach for mixed inputs with the nugget effect.

2. Use the history matching technique with implausibility measures in terms of the utility theory to find an optimal region of the expected maximum utility for the qualitative and quantitative inputs.

3. Application of the Bayes linear emulation techniques with mixed inputs and history matching in the agricultural sector for the first time.

One of this thesis's main limitations is using real-life agricultural data; future research can apply the same technique to real-life data. Overall, this thesis provided an optimal decision for the region to get the expected maximum utility for the farmers, considering the effects of weather, soil type, steepness, and fertilisation to yield and pollutants. We aimed to build general framework for the farmers and implement them through the EPIC crop simulation model data. This thesis developed and illustrated the methodology using Spring Barley crop data, and future researchers can follow our steps to apply it to real-life data. Our

farmers are the core for feeding the whole world, and they need more agricultural outputs; this thesis's general set-up is just one of the contributions of this world's challenge for food.

8.1 Summary of the Chapters

In Chapter 2, we presented the detailed procedure of the EPIC simulator, which is the base of the analysis of this thesis. We performed the EPIC simulation study for 58 years with different types of inputs and outputs for the Wensum catchment. A subset of simulated data was demonstrated in this chapter and explored to assess the crop yield trend for inputs Nitrogen and Phosphorus. We only considered two pollutants and a few crops because our interest was to develop a general framework rather than extensive data analysis.

Chapter 3 mainly reviewed the basic crop yield models with their features and fit them to the EPIC simulation data. These models were fitted for three different crops Spring Barley, Winter Barley and Silage. The fitted curve and estimates of these crops were shown, whose were only satisfying all features of the Mitscherlich-Baule (MB) model with lower RSE estimates and also favoured by most of the pieces of literature. However, the crop models must fulfil the standard features, and violating these makes them impractical to use. Despite the typical features, the boundary constraints and initialization for non-linear fitting models are complicated.

Setting up a non-linear Bayesian hierarchical framework in terms of mixed inputs was the main objective of Chapter 4 with the MB model. This chapter set up a Bayesian hierarchical model for the MB model for mixed inputs. For the non-conjugate form, we used the HMC-NUTS MCMC to generate the posterior samples. The results showed strong evidence of Nitrogen but a weak response to Phosphorus. A model comparison was performed using some diagnostic tools, and the diagnostics showed no concern to worry. Finally, factors incorporation were explored using N only reduced model from model comparison. We then reran the MCMC with factor variables, which also showed no concern in diagnostics, and the simulated data lie within the mean and predicted credible intervals for all factors. However, we have seen a broader uncertainty for all the factors, and some failed to capture the data trend. The possible reason for this failure is that the MB model can not capture the direction of the monotonic nature of the yield, and another reason is to assess the effect on the maximum crop yield parameter only.

The Bayes linear emulation approach is introduced with application to EPIC data for continuous input only in Chapter 5. The chapter started by discussing the context of emulation, the basic structure and approaches. We introduced the maximum likelihood inference to estimate the parameters for constructing the emulator. The detailed calculation of the Bayes linear emulation was shown with the formulations to generate adjusted emulator expectations and variance with some diagnostics tools. One dimensional example was demonstrated, and finally, EPIC simulator data was used to build the emulator for the crops Spring Barley and Winter Barley. The adjusted emulator expectation revealed a solid response for Nitrogen only with narrow uncertainties for both crops. We also assessed the validity of the emulators using some diagnostic tools and found no evidence of conflict with the simulator.

In Chapter 6, we focused on developing a framework for the Bayes linear mixed inputs emulation using approaches to build the factor correlation matrix. We presented three approaches to assessing the correlation matrix for factor levels. We provided a general set-up to estimate the correlation matrix for qualitative and quantitative inputs. Finally, we combined the correlation matrix with the Bayes linear emulation technique to calculate the emulator-adjusted expectations and variances. We have applied our new method to the EPIC simulator data for the Spring Barley crop and identified the general approach as the best correlation approach using the SPE and MSE diagnostic tools. We then built the emulators of mixed inputs combining the correlation matrix for the chosen correlation approach. The results also revealed a solid response to Nitrogen and no response to Phosphorus with little uncertainty over the input space.

Chapter 7 started by introducing utility, history matching, and implausibility measures. A linear utility function was considered by combining yield and pollutants with the respective coefficients. We applied the concept of history matching by using the expected maximum utility as the actual observation to calculate the implausibility. Initially, we constructed the emulators for the continuous and steepness soil only for the pollutants. Then, a sensitivity analysis for coefficients related to yield and pollutants was assessed by fixing the value of b_0 and hence calculated the utility and implausibility for the continuous case and mixed inputs soil-steepness only. We performed a variable selection method, including the weather factor, and finally, built the emulators for the reduced inputs. We calculated adjusted emulator expected mean and variance for the utility with the maximum implausibility with the reduced mean function. Our result showed the area of expected maximum

utility for the lower values of P and higher values of N.

8.2 Future Work

The possible future works of this thesis are presented as follows:

1. To explore more crop yields, and pollutants, especially the outputs of total biomass and below-ground biomass general trends concerning fertilisers for the Wensum Catchment.

2. To analyze and compare the Eden catchment data and the eight management scenarios, such as Tillage, Cover crops, etc. Future work may consider these scenarios to analyze the trend of crop yields under different scenarios for both catchments.

3. The future extension of this work is to consider the factor effects for the Phosphorus inputs while setting the Bayesian framework.

4. To extend this work to other factor effects on coefficients β_1 and β_2 and assess the overall factor effects in Chapter 4 for the Bayesian modelling framework.

5. To consider the multivariate emulation technique for the output pollutants to see the overall pollution effect in Chapter 7.

8.3 Research Achievements and Awards

One of our research works has already been published in Springer conference proceedings (Chapter 5 [134]) of Data Analysis. The Bayesian hierarchical framework of Chapter 4 with the yield model selection from Chapter 3 is under revision in the Journal of Statistical Modelling. It has already been presented at the 13th International Conference of the ERCIM WG on Computational and Methodological Statistics (CM Statistics, 2020). Chapters 6 and 7 are in the submission process. The work for the continuous part of the Bayes linear emulation approach from Chapters 5 to 7 already submitted to the Environmental Modelling and Software journal. Our work on Chapter 6 has already been presented in a contributed session at the Royal Statistical Society Conference, 2022 and the SIAM student chapter conference. These research works helped me to achieve "The Euan Squires Memorial Prize," awarded to an overseas Postgraduate Research (PGR)

student with an excellent academic performance from Durham University Mathematical Sciences Department and the reputed "Charles Wallace Trust PhD Bursary" for final year students. Despite these, I have been awarded travel awards from Ustinov College for the RSS conference, SIAM UKIE National Student Chapter Conference, and the conference registration fee award from the 6th Canadian Conference in Applied Statistics. Moreover, I have also been awarded the most competitive Durham Doctoral Scholarship (DDS) for my ongoing PhD work.

Appendix A

Bayesian Hierarchical Framework for Crop Yield

A.1 MCMC Algorithms

A.1.1 Metropolis-Hastings Algorithm

Definition A.1.1 (Proposal Distribution)**.** *A distribution $\pi(a^{'})$ is said to be a proposal distribution if it depends on the current sample a_i to draw the following sample a_{i+1}.*

Definition A.1.2 (Target Distribution)**.** *A distribution $\pi(\theta|y)$ is said to be a target distribution or posterior distribution if it uses to generate the posterior samples.*

1. Initialize the starting parameter a_0
2. For $i = 1, 2, ..., N$
    - Generate a sample from the proposal distribution such that $a^{'} \sim \pi(a^{'}|a_i)$
    - Use $u \sim U(0, 1)$ to generate sample from the uniform distribution
    - Compute $H = \frac{\pi(a^{'})\pi(a_i|a^{'})}{\pi(a_i)\pi(a^{'}|a_i)}$, which is known as Hastings ratio
    - Now, if $u < \min(1 + H)$ then accept the proposal and set $a_{i+1} = a^{'}$
    - Else $a_{i+1} = a_i$
3. Repeat step 2.

A.1.2 Gibbs Sampling

1. Set the initial value $a^{(0)} = \left(a_1^{(0)}, ..., a_n^{(0)}\right)$,
2. Repeat for $k = 1, 2, ..., M$
   - Generate $a_1^{(k+1)}$ from $\pi\left(a_1 | a_2^{(k)}, ..., a_n^{(k)}\right)$,

     $\vdots$
   - Generate $a_n^{(k+1)}$ from $\pi\left(a_n | a_1^{(k+1)}, ..., a_{n-1}^{(k+1)}\right)$,
3. Repeat the values of $\left[a^{(1)}, a^{(2)}, ..., a^{(M)}\right]$.

A.1.3 Hamiltonian Monte Carlo Within No-U-Turn Sampler

- Supplementary Variable: A variable is said to be called supplementary due to its indirect effect such that it cannot use as an outcome or explanatory variable.
- Hamiltonian Dynamics: A Hamiltonian dynamics is a dynamical system of the scalar function $H(a, b, t)$, which is generalized by two coordinates a and b known as momentum and position, respectively. So with respect to Hamilton's Equation [93] we can write;

$$\begin{aligned} \frac{\delta a}{\delta t} &= -\frac{\delta H}{\delta b}, \\ \frac{\delta b}{\delta t} &= \frac{\delta H}{\delta a}. \end{aligned} \tag{A.1.1}$$

- Leapfrog Integration: Leapfrog integration is the numerical integration of the differential equations of the second order; the more details are explained in [30] and can be expressed as;

$$y^{'} = \frac{d^2 y}{dt} = A(y). \tag{A.1.2}$$

  In leapfrog integration, the simultaneous equations are as follows [30] :

$$\begin{aligned} a_j &= A(y_j), \\ v_j + \frac{1}{2} &= v_{j-\frac{1}{2}} + a_j \ \Delta t, \\ y_{j+1} &= y_j + v_{j+\frac{1}{2}} \ \Delta t. \end{aligned}$$

Here, y_j is the position at step j; for step $j + \frac{1}{2}$, the first derivative of y is $v_j + \frac{1}{2}$; the second derivative of y for the step j is $a_j = A(y_j)$; and Δt is called the size of every time step.

- Tuning Parameter: A parameter Λ is called the tuning parameter, which is used to control the strength of the penalty term.

HMC efficiency strongly relies on the tuning parameters of momentum covariance, step size, and the step number corresponding to several iterations [100]. So the introduction of NUTS within HMC adaptively tunes the parameters during the warming period and adjusts the step size number during the iteration process. HMC can be written as [100, 126];

$$H(\lambda, k) = v(\lambda) + p(k), \tag{A.1.3}$$

where $v(\lambda)$ and $p(k)$ are the potential and dynamic energies. In HMC, for estimating λ with $f(\lambda)$, we can introduce a supplementary variable [126] k such that $f(k) \sim N(0, D)$ with zero mean and covariance matrix D. So the joint density of $f(\lambda, k)$ can be written as;

$$\begin{aligned} f(\lambda, k) = \exp \log f(\lambda) + \log f(k) &\propto \exp(\log f(\lambda) - \frac{1}{2}\lambda' D^{-1}\lambda), \\ &= \exp(-v(\lambda) - p(k), \\ &= \exp(-H(\lambda, k)), \end{aligned} \tag{A.1.4}$$

where $v(\lambda) = -\log f(\lambda)$, and $p(k) = \frac{1}{2}\lambda' D^{-1}\lambda$. So HMC generate samples from the joint distribution of (λ, k) and Hamiltonian dynamics can be written as following two differential equations;

$$\begin{aligned} \frac{\delta\lambda}{\delta t} &= -\frac{\delta H}{\delta k}, \\ \frac{\delta k}{\delta t} &= \frac{\delta H}{\delta \lambda}. \end{aligned} \tag{A.1.5}$$

It is also known that [83] Hamiltonian equation has no potential solution but can be approximated for the discrete setting using the leapfrog method of integration. So using the leapfrog integration method, we can write that;

$$\begin{aligned} k(a + \frac{1}{2}b) &= p(a) - \frac{1}{2}b\frac{\delta f(\lambda)}{\delta \lambda}(\lambda(a)), \\ \lambda(a + b) &= \lambda(a) + bk(a + \frac{1}{2}b), \\ k(a + b) &= k(a + \frac{1}{2}b) - \frac{1}{2}b\frac{\delta f(\lambda)}{\delta \lambda}(\lambda(a + b)), \end{aligned} \tag{A.1.6}$$

where b is the integration step size and a is the time within the range $1 \leqslant a \leqslant Q_t$; Q_t is the total number of integration steps. The complexity arises for HMC due to the large value of b leading to a low acceptance rate, and a smaller value of b increases the time length; also, smaller values of Q increase the auto-correlation.

For NO-U-Turn, Sampler (NUTS) selects an appropriate length for Q for every iteration, maximizing the distance between each step and working efficiently via the doubling method. NUTS started from a supplementary variable k i.e. $k(t|\lambda) \sim U(0, \exp(\log f(\lambda) - \frac{1}{2}\lambda' D^{-1}\lambda))$. By doubling the size, NUTS generates a finite set for (λ, k) in repeated nature. Let us consider M a subset for the candidate states (λ, k) and M is considered from (λ, k) by doubling process to justify $t \leqslant \exp(\log f(\lambda) - \frac{1}{2}\lambda' D^{-1}\lambda)$. If we consider (θ, p) is our initial values and the following values of (θ^*, p^*) are sampled from M. For that, Hoffman and Gelman [100] proposed a Kernel for double sampling steps and also determined 0.6 as the optimal acceptance probability. For j^{th} iteration of MC in NUTS of the tuning b can be expressed as;

$$\log(b_{j+1}) \longleftarrow \Delta - \frac{\sqrt{j}}{\beta}\frac{1}{j + j_0}\sum_{i=1}^{j}(p_t - \alpha_{ja}),$$

$$\log(\overline{\mathrm{b_{j+1}}}) \longleftarrow \sigma_j \log(b_{j+1}) + (1 - \sigma_j)\log(\overline{\mathrm{b_j}}),$$

$$(b_{j+1}) \longleftarrow \overline{\mathrm{b_{j+1}}},$$

where $\sigma_j = j^{-w}$ set by [100]; α_{ja} is the actual acceptance probability, and p_t is the desired acceptance probability; Δ is the arbitrary choosing point; j_0 used to initial exploration. The acceptance probability is calculated as follows;

$$\alpha_{ja} = \frac{1}{|\gamma_j|}\sum_{\lambda, k \epsilon \gamma_{ja}} \min\left[1, \frac{\pi(\lambda^j, k^j)}{\pi(\lambda^{j-1}, k^{j,0})}\right],$$

where λ_j, k_j are candidates, $\lambda^{j-1}, k^{j,0}$ are initial values, γ_{ja} are all possible sets explored during MCMC. So the algorithm of NUTS can be briefly described as follows [126];

1. Set the initial value for λ, b and values of $p_t, \Delta, \beta, j_0, w$.
2. Generate $k \sim N(0, I)$.
3. Generate $t \sim U(0, \exp(\log f(\lambda) - \frac{1}{2}k' M^{-1}k)$
4. Set M from double sampling within kernel mentioned in Hoffman and Gelman [100].
5. The proposed (θ^*, p^*) is accepted with probability α_{ja} for j^{th} iteration.

6. Renew b_j through dual averaging.

7. Repeat steps 2 to 6.

Using the above discussed set-up of Hamiltonian Monte Carlo within a No-U-Turn sampler, we update the posterior probability distribution in Equation (4.11) to generate the posterior samples.

A.2 Diagnostics of the Bayesian Analysis

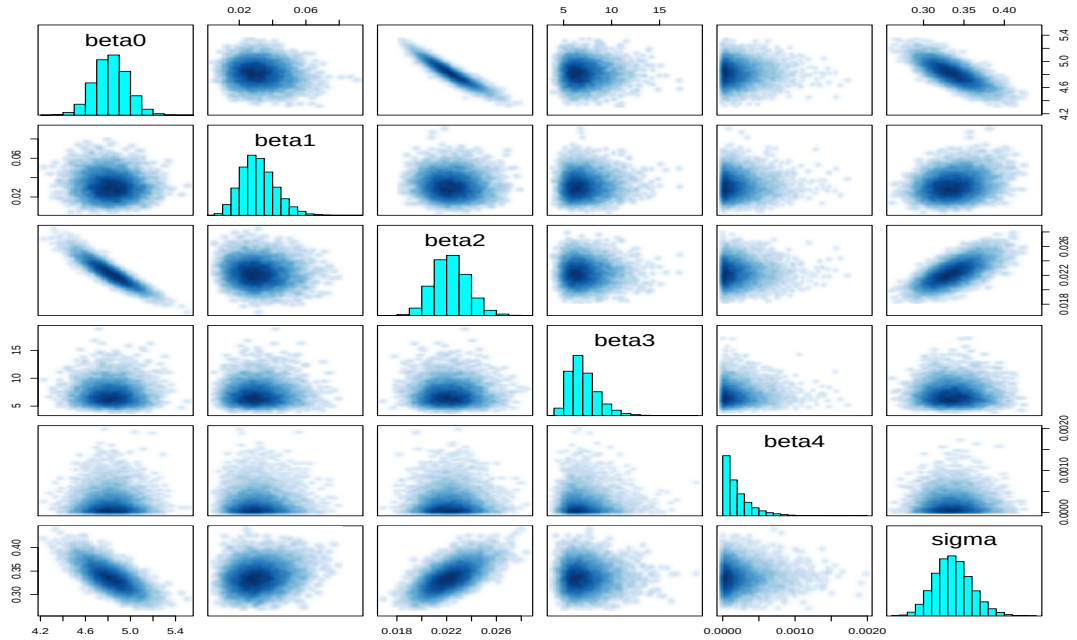


Figure A.1: Pairs plot for the crop Winter Barley. Like the crop Spring Barley, most are independent such that no trend, except β_0, is positively correlated with β_2 with a linear trend.

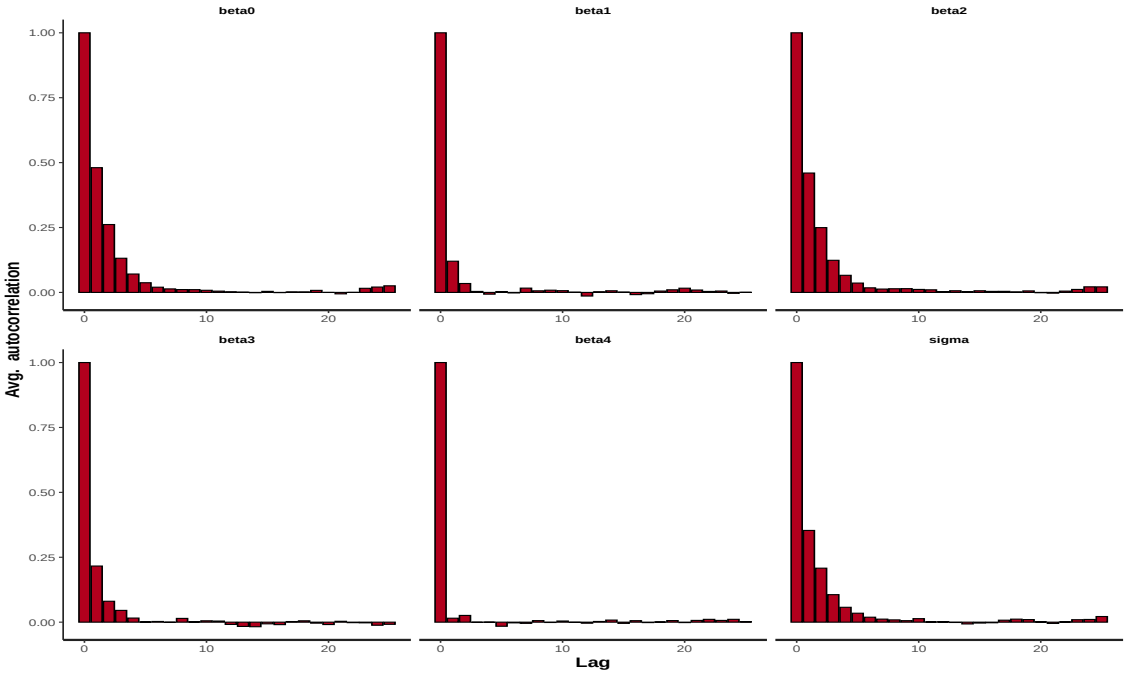


Figure A.2: Autocorrelation diagnostic plot for the crop Winter Barley and from this plot can be shown a decreasing trend of bins with the increase of the lags, indicating good mixing of chains like as Spring Barley with high effective samples for β_4.

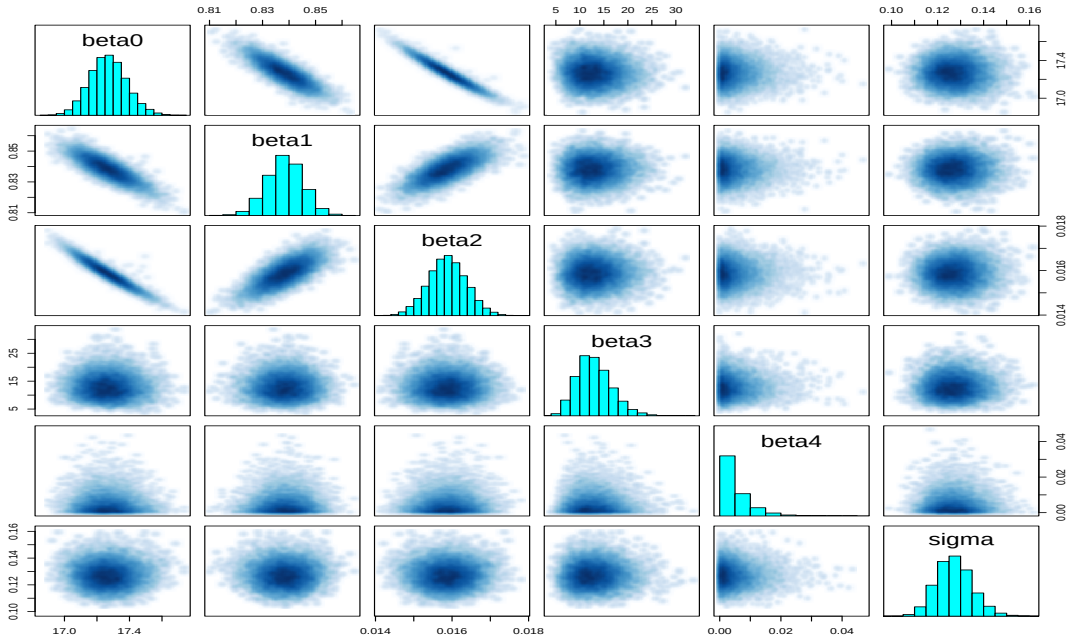


Figure A.3: Pairs plot for the crop Silage, and from this figure, we can see the parameters β_3, β_4, and sigma show no visible trend but a linear trend among the parameters β_0, β_1 and β_2.

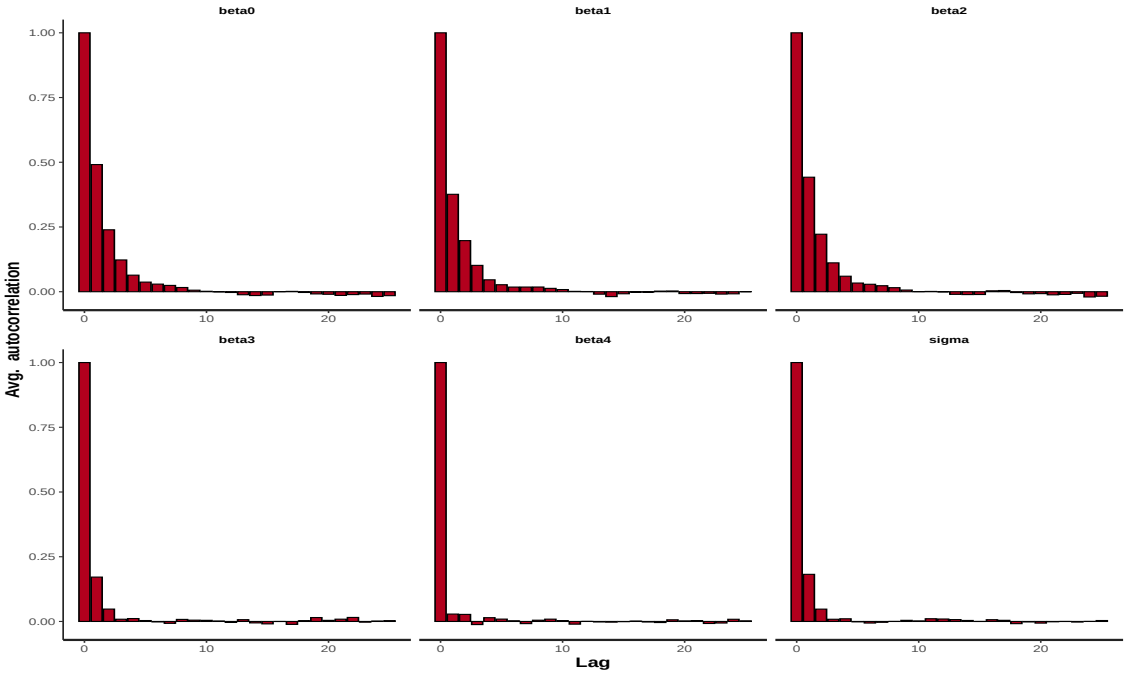


Figure A.4: Autocorrelation diagnostic plot for the crop Silage and from this plot shows a significant decreasing trend of bins with the increase of the lags indicates a high effective sample for β_3, β_4, σ but a gradual decrease for other parameters.

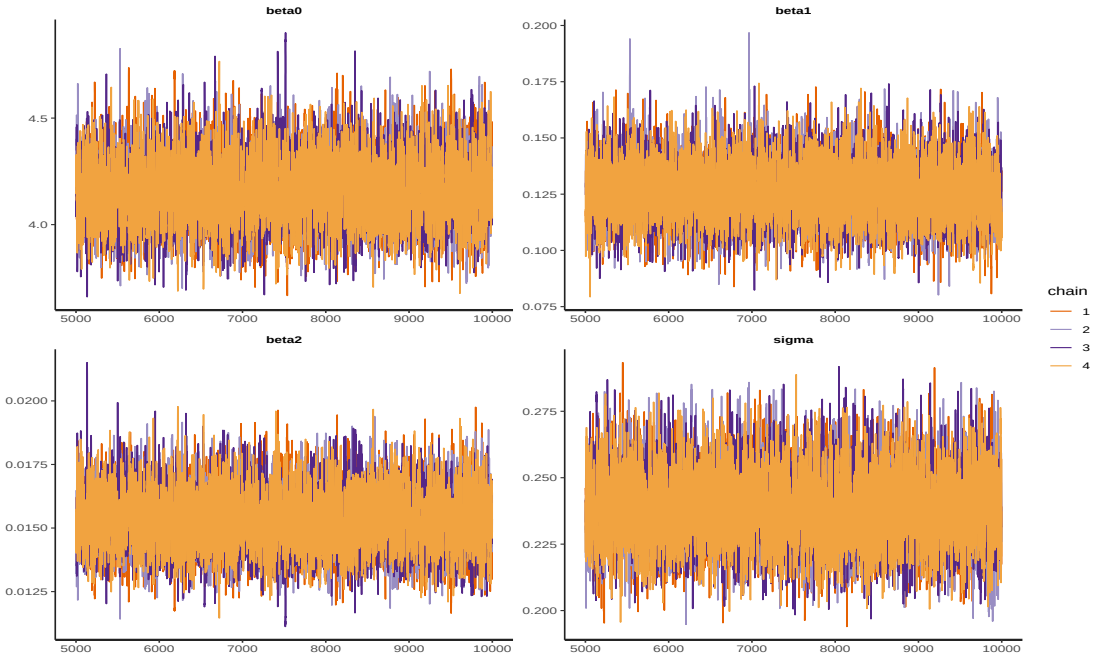


Figure A.5: The trace plot for the crop Spring Barley for the input N only with four different chains and figure reveals a good mixing.

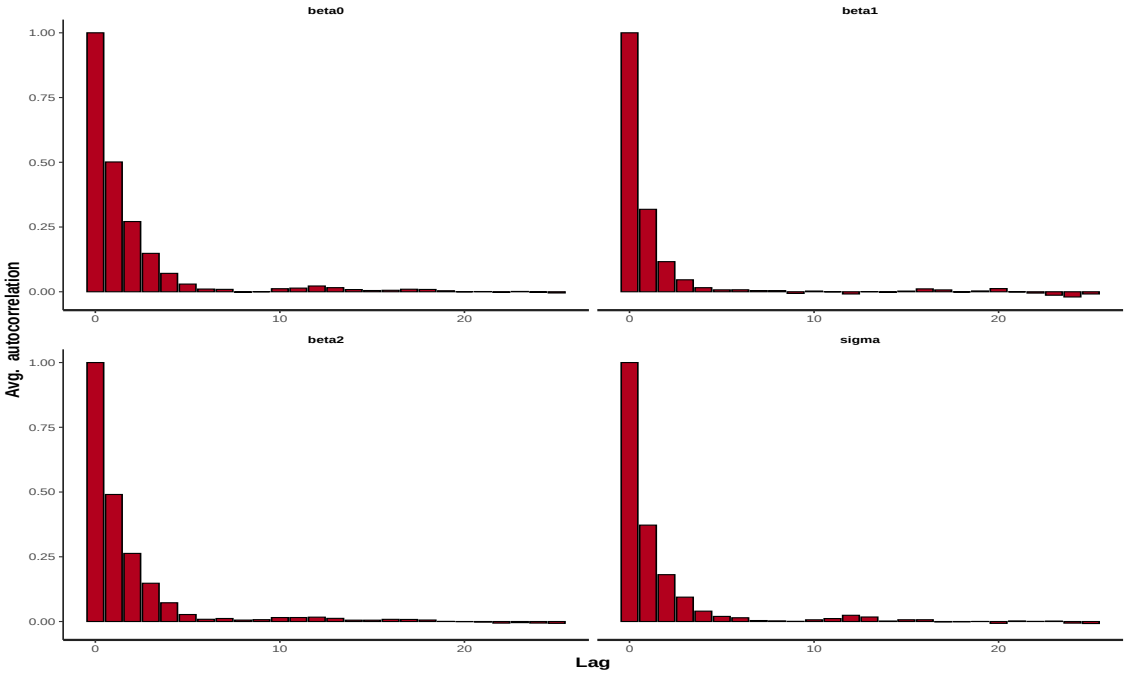


Figure A.6: Autocorrelation diagnostic plot for the crop Spring Barley using the response N and bins are decreasing gradually with the increase of lags, suggesting no concern to worry about the chain mixing.

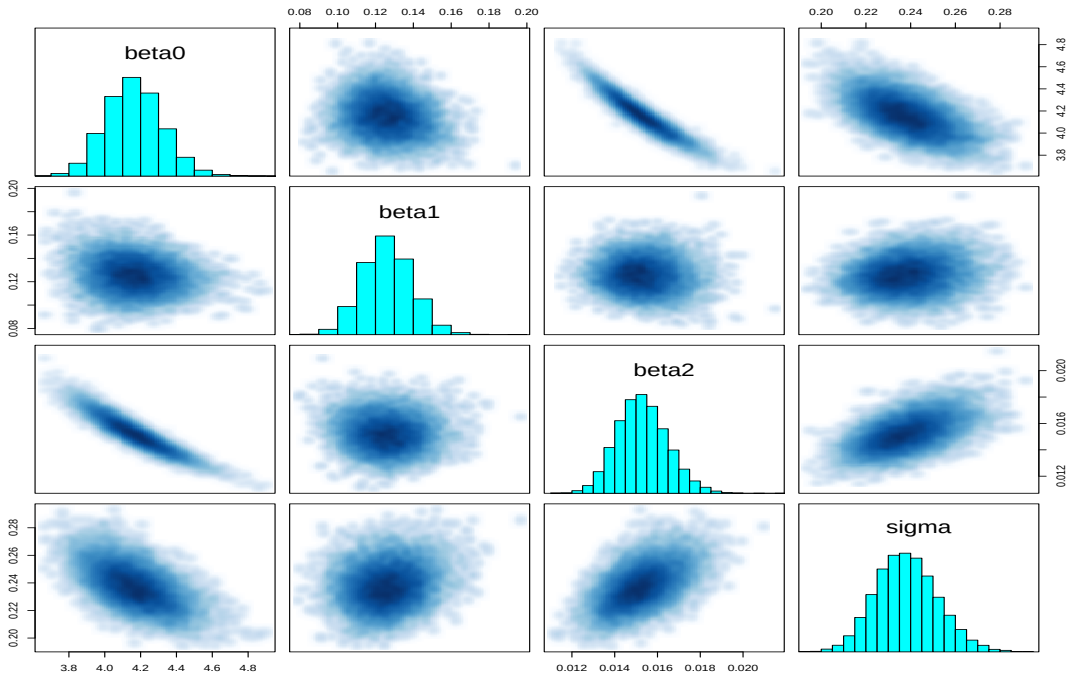


Figure A.7: Pairs plot for the crop Spring Barley using N response and like the entire model with P, most are showing no trend, except β_0, is showing a linear trend with β_2.

A.3 Diagnostics of Incorporating Bayesian Factor Inputs

In this Section, we have demonstrated the diagnostics of the factor input steepness, soil and weather. We have presented the density plot, autocorrelation plot and pairs plot for all three inputs. Our diagnostics reveal no concern to worry about the framework; hence, our Bayesian mixed inputs framework is valid.

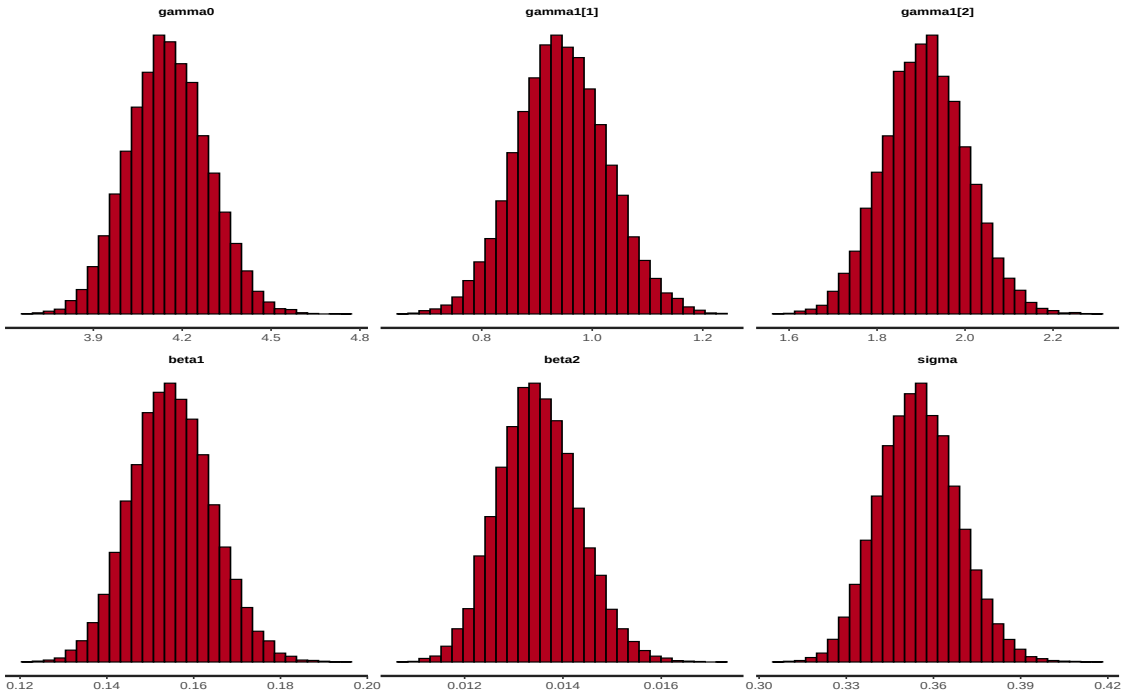


Figure A.8: The figure for the posterior density of the parameters using soil and all of them are normally distributed.

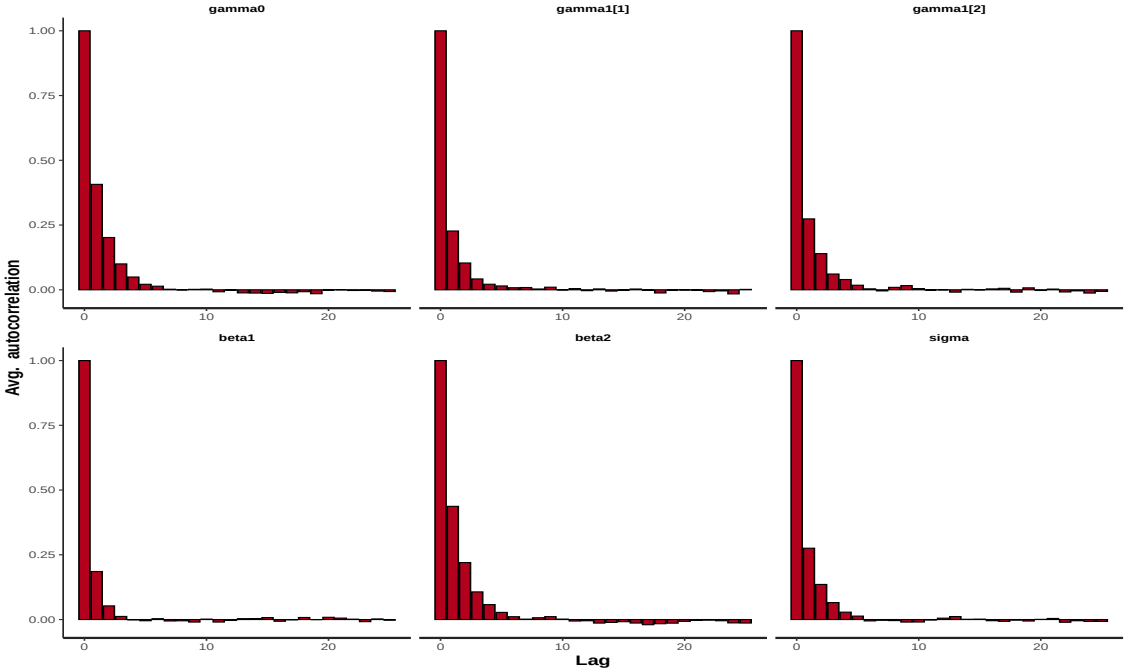


Figure A.9: An autocorrelation plot using the factor soil shows a gradual decrease of the bins with the increase of lags, which is the expectation of good mixing chains with adequate, effective samples.

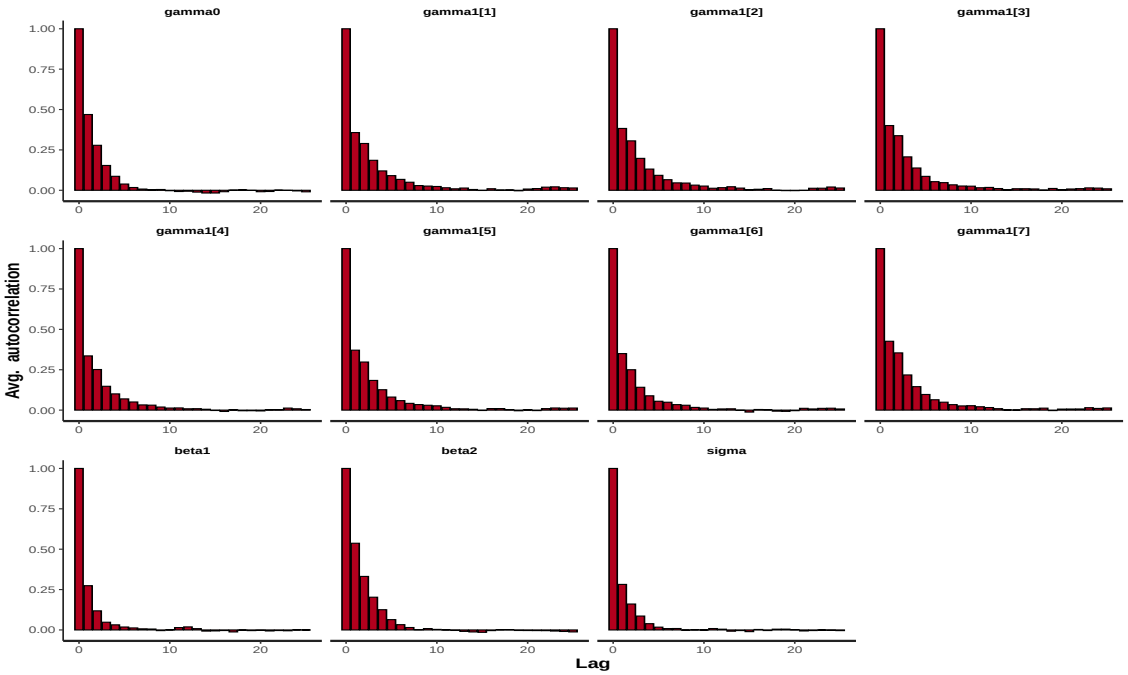


Figure A.10: The plot for autocorrelation diagnostic of the factor weather illustrates a decreasing trend with increasing lags such that the chains mixing are justified and hence the Bayesian framework.

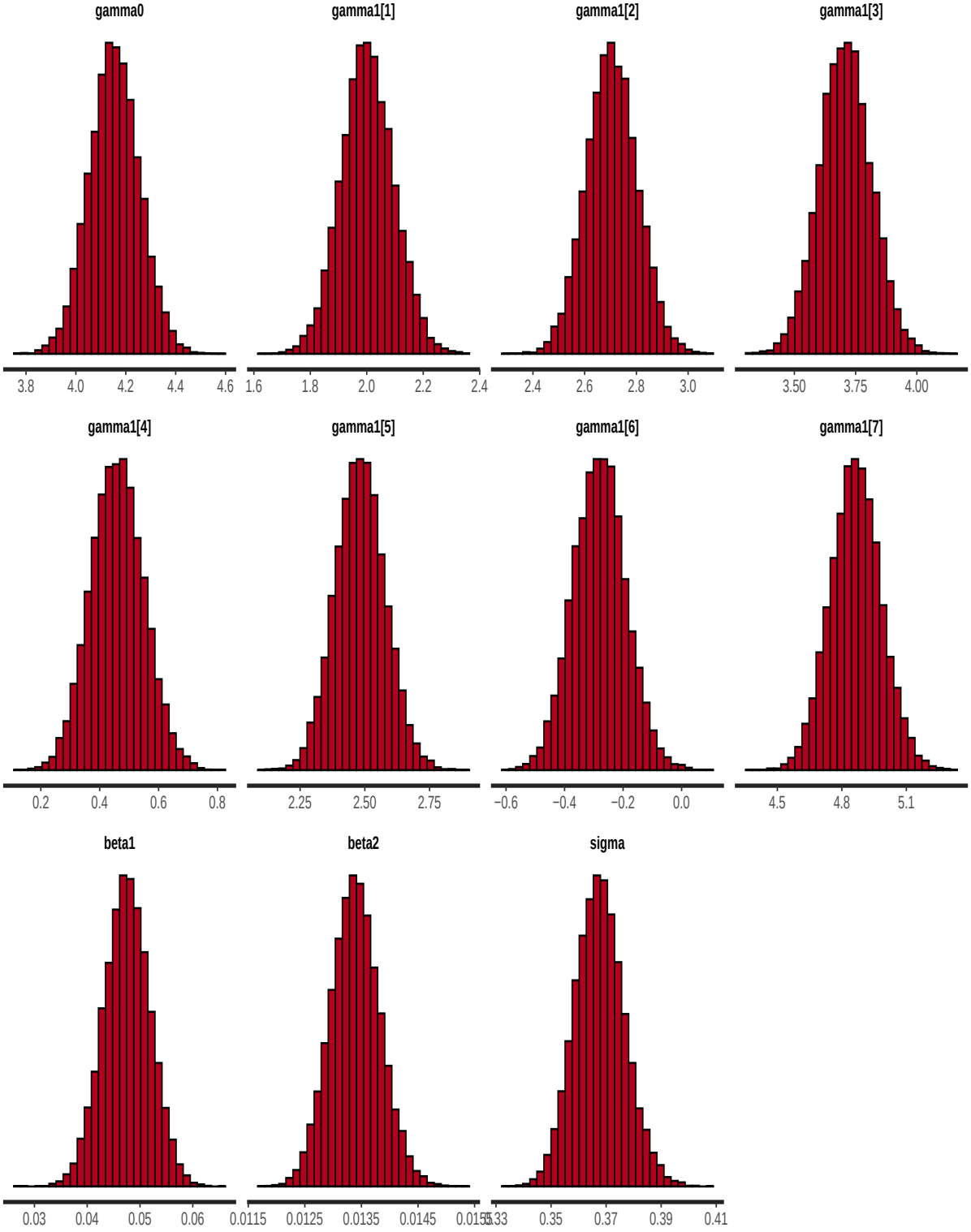


Figure A.11: The posterior density diagnostic plot figure using the factor weather and all eleven parameters are normally distributed. So, the factor weather in the Bayesian inference framework is valid.

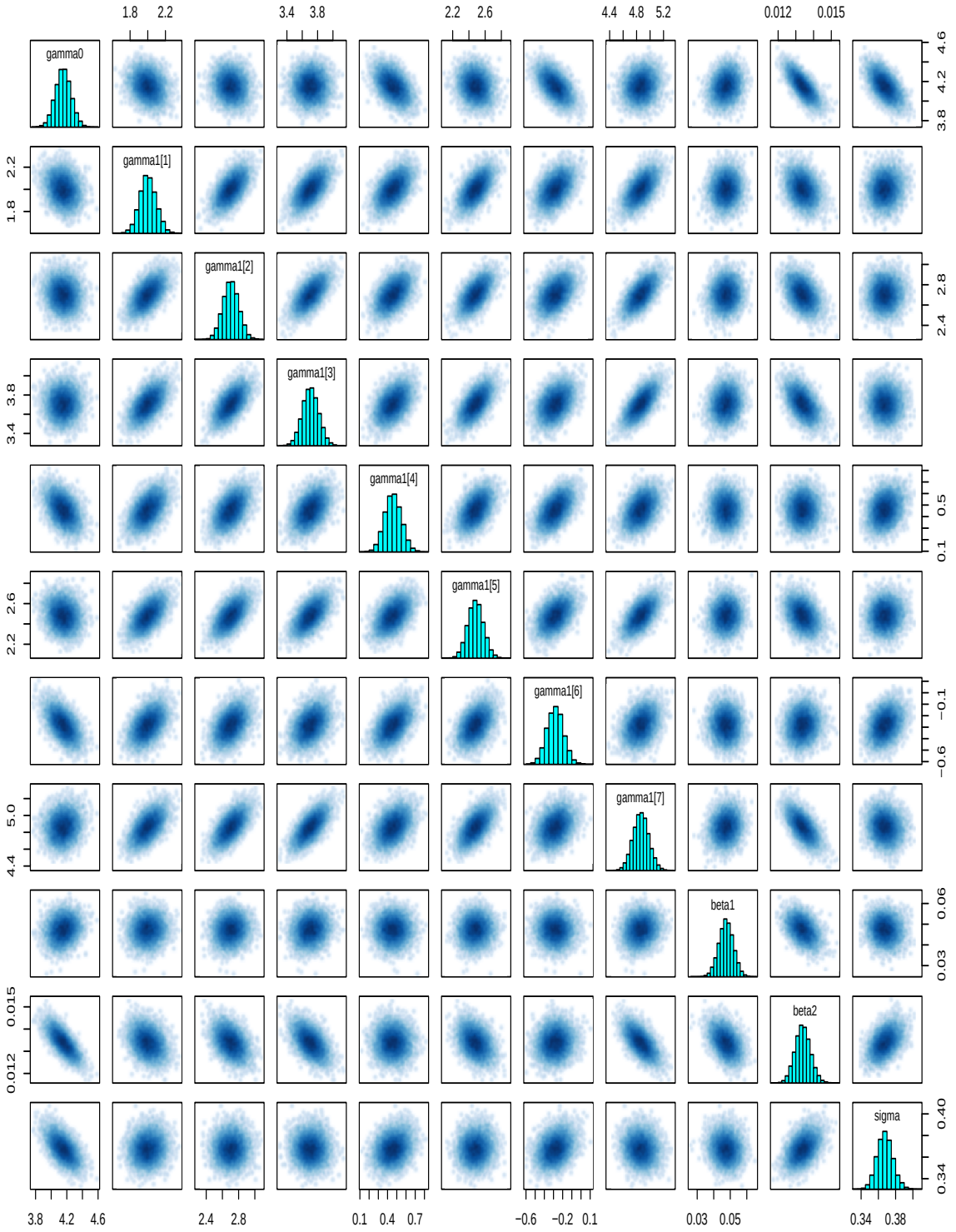


Figure A.12: The diagnostic of pairs plot among the hyperparameters for the factor incorporation weather shows no clear pattern indicating weak correlations, which is desired.

Appendix B

Bayes Linear Emulation Approach For Quantitative Inputs

B.1 Objective Function and Optimization Techniques

Optimization is the problem of finding optimum values from every set of possible solutions. The general form of optimization can be expressed as follows [53];

$$\begin{aligned} O \ \min_{y} f(y) \\ \text{Subject to} \quad & m_i(y) \leqslant 0; i = 1, ..., p \\ & k_j(y) = 0; j = 1, ..., q \end{aligned} \tag{B.1.1}$$

where; $f : \mathbb{R}^n \to \mathbb{R}$ is set to be minimizing objective function over the vector of y for the n^{th} variables. And, also $m_i(y) \leqslant 0$ is the inequality constraint and $k_j(y) = 0$ is the equal constraint for $p, q \geqslant 0$.

Definition B.1.1 (Objective Function)**.** *An objective function is the real-valued function of mathematical optimization of the problem O to be minimized or maximized, such that $f(y)$ is the objective function.*

B.1.0.1 Optimizations Techniques

There are two widely used optimization techniques available to solve the non-linear objective function with box constraints such as;

- Limited memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) Box Constrained (L-BFGS-B) method
- Nelder Mead method

B.1.1 Broyden–Fletcher–Goldfarb–Shanno (BFGS) Method

Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is the local search optimization algorithm. It usually uses the quasi-Newton method, which minimizes $f(y)$ concerning y and second order derivative, which requires a gradient of f to be known and uses the Hessian matrix [39, 53].
Like the BFGS method, BFGS-L (Limited Memory) [39, 53] method use the approximation of the Hessian matrix. Still, this algorithm store only a few vectors, say a, compared to $n \times n$ dense such that $a \ll n$ and then update the position of y and gradient $\bigtriangledown f(y)$.
The L-BFGS-B method handles the box constraints of $p_i \leqslant y \leqslant q_i$, where p_i and q_i are the lower and upper bounds. This method is used to figure out the free and fixed variable at every step of iteration from a simple gradient method and then use the L-BFGS technique to those free variables for identifying the highest precision and repeat the procedure until the required iterations of sample converge.

B.1.2 Nelder Mead Method

The Nelder Mead optimization technique is used for non-linear optimization, for which direct derivatives are hard to formulate. This method was developed by Nelder and Mead in 1965 [6]. This method uses the simplex idea and approximates the local optimum of n variables for the uni-modal and smooth objective function. For the n dimension, Nelder-Mead uses $n + 1$ testing points and extrapolates the nature of the objective function for each testing point to identify new test points [29]. The method also performs well for noisy and discontinuity objective functions.

B.2 Emulation for Continuous Inputs Spring and Winter Barley

Figure $B.1$ shows the result of Bayes linear update for the crop Spring Barley and the resolution diagnostic for 12×12 grid points. The left upper and right upper panels show the emulated posterior mean for the Spring Barley crop and its associated standard deviations as functions of N and P. We note that the crop yield is increasing with increasing Nitrogen levels. However, the effect of Phosphorous is much less pronounced and arguably only significant when Nitrogen levels are low for Spring Barley.

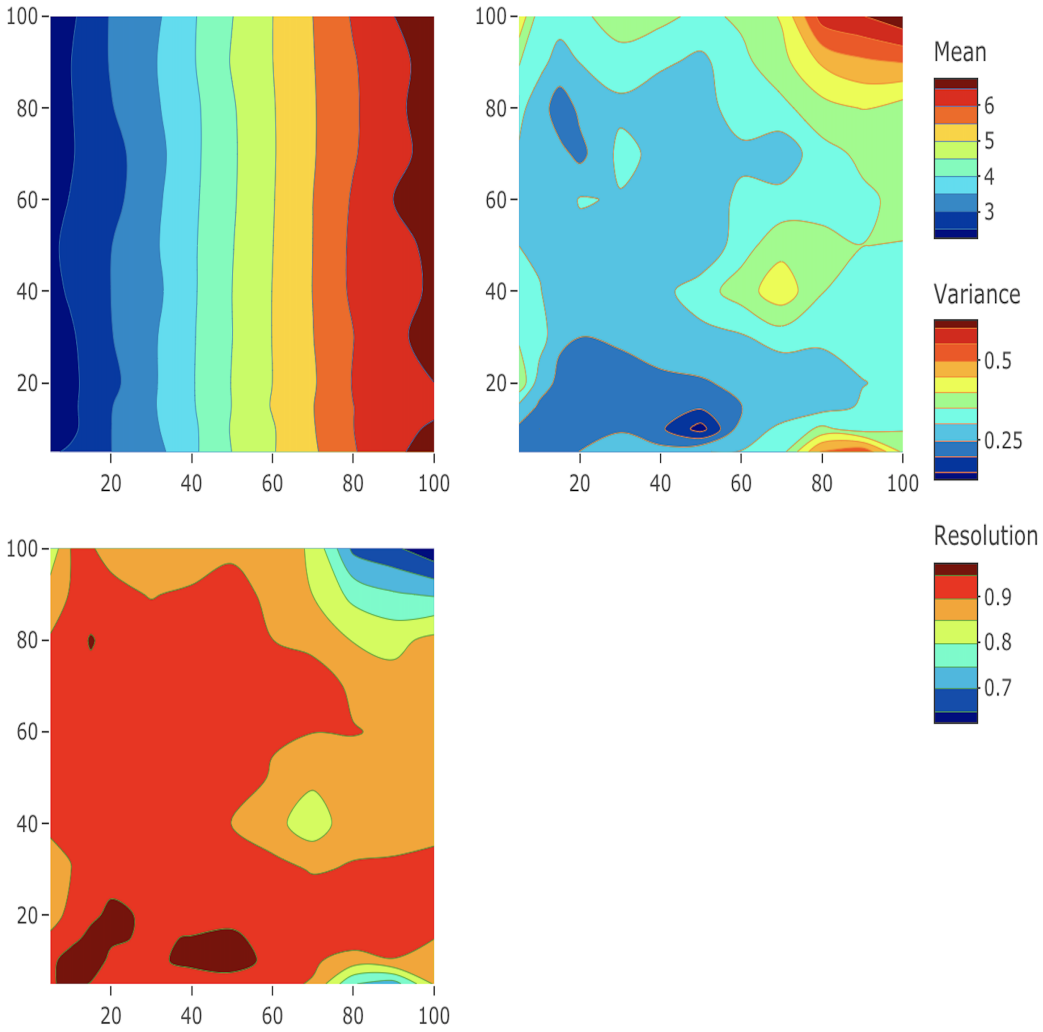


Figure B.1: Left Upper Panel: Adjusted Emulated Mean for Spring Barley Yield as a Function of Nitrogen (N) on x axis and Phosphorus (P) on y axis; Right Upper Panel: Emulator Standard Deviations ($\sqrt{Var}$); Left Lower Panel: Resolution Diagnostic.

The standard deviations plot highlights low levels of uncertainty in Spring Barley yield around the locations for which we have simulations, with uncertainty increasing as we move away from these points. It is noticeable that higher uncertainties are on edge for the higher levels of N and P, which need more training points to emulate. The left lower shows the emulated resolution diagnostic for Spring Barley, indicates most values lie above 0.70 over much of the space. So, an indication of high-resolution values means that our emulator can explain most of the variation of that complex simulator, and the estimated values are valid to use for further analysis. The blue regions of low resolution indicate locations corresponding to the test data, which were not used for emulator fitting; hence little data was available to reduce the variance in these locations.

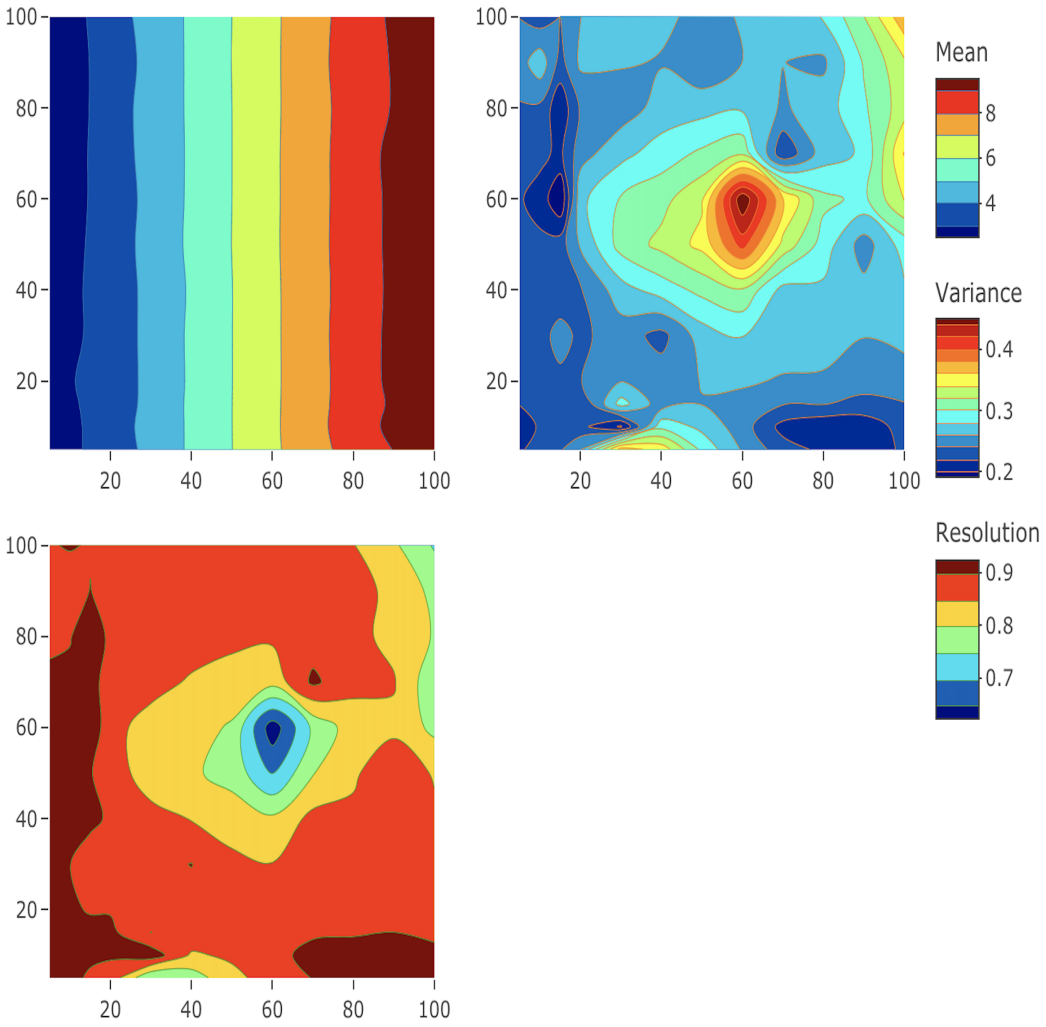


Figure B.2: Left Upper Panel: Adjusted Emulated Mean for Winter Barley Yield; Right Upper Panel: Emulator Standard Deviations ($\sqrt{Var}$); Left Lower Panel: Resolution Diagnostic.

Figure $B.2$ shows the result of Bayes linear update for the crop Winter Barley and the resolution diagnostic. The left upper and right upper panels show the emulated mean Winter Barley yield and its associated standard deviations as functions of N and P. We note that the weak dependency on Phosphorous has disappeared entirely, and the crop yield appears insensitive to values of P. The standard deviations plot highlights low levels of uncertainty all over the space but high uncertainty around 60 levels for both N and P. The left lower panel shows the emulated resolution diagnostic, where most values lie above 0.70 over the space. So, an indication of high-resolution values means that our emulator can quite explain most of the variation of the EPIC complex simulator. The blue regions of low resolution 60 levels for both N and P indicate locations corresponding to the test data, which were not used for emulator fitting. There needed to be more data available to reduce the variance in these locations.

B.3 Building the Covariance Matrix

Building the block structure is one of the essential parts of mixed inputs modelling. Its focus is on the multiplication of each block with the corresponding factor effect. The layout of the block structure for the single factor is easy to implement by using the product of covariance matrix ($R'_{[\theta,T,g]}$) with the correlation matrix T. For the block structure of the single factor, we needed to make inferences about the variance part of the single model emulator of $f(y)$ and the covariance part of $f(x), f(y)$. The block structure for the variance part can be expressed as follows for the 2-level single factor input, and quantitative inputs can be expressed as follows;

$$Var[f(v)] = \left(\begin{array}{c|c} C_{1,1} & C_{1,2} \\ \hline C_{2,1} & C_{2,2} \end{array}\right) \odot \left(\begin{array}{c|c} 1 & t_{1,2} \\ \hline t_{2,1} & 1 \end{array}\right)$$

So we can write the above matrix in terms of the general model layout;

$$\begin{aligned} C_{1,1} &= \sigma^2 R\Big[(x,w)_1, (x,w)_1^{'}\Big], \\ C_{1,2} &= t_{1,2} \times \ \sigma^2 R\Big[(x,w)_1, (x,w)_2\Big], \\ C_{2,1} &= t_{2,1} \times \ \sigma^2 R\Big[(x,w)_2, (x,w)_1\Big], \\ C_{2,2} &= \sigma^2 R\Big[(x,w)_2, (x,w)_2^{'})\Big], \end{aligned}$$

where, x and w are the continuous inputs, and factor inputs respectively, which are treated as testing data set; $t_{1,2}$ is the correlation between factor level 1 and factor level 2. The error variance σ^2 can be estimated using the maximum likelihood estimate of the theorem 2. Similarly, the variance part of the testing data can be expressed for training data for the 2-level factor input;

$$Var[f(v^{'})] = \left(\begin{array}{c|c} C_{1,1} & C_{1,2} \\ \hline C_{2,1} & C_{2,2} \end{array}\right) \odot \left(\begin{array}{c|c} 1 & t_{1,2} \\ \hline t_{2,1} & 1 \end{array}\right)$$

The variance of $C_{1,1}$ is also calculated by using the same general model layout. The covariance between the training and testing data set for the emulator $f(x)$ and $f(y)$, which can be expressed in the following way;

$$Cov[f(v), f(v^{'})] = \left(\begin{array}{c|c} Cov_{1,1} & Cov_{1,2} \\ \hline Cov_{2,1} & Cov_{2,2} \end{array}\right) \odot \left(\begin{array}{c|c} 1 & t_{1,2} \\ \hline t_{2,1} & 1 \end{array}\right)$$

Here, the $Cov_{1,1}$ is the variance between the training data for the factor level-1 and the testing data for the factor level-1. So the general layout of the block structure of the variance part for the J factor levels and I continuous inputs can be written as,

$$V[f(v)] = \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,I} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ C_{J,1} & C_{J,2} & \cdots & C_{J,I} \end{bmatrix} \odot \begin{bmatrix} 1 & t_{1,2} & \cdots & t_{1,I} \\ t_{2,1} & 1 & \cdots & t_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ t_{J,1} & t_{J,2} & \cdots & 1 \end{bmatrix} \tag{B.3.2}$$

B.3.1 Formulation of Block Structure

Recalling the concept of section 6.2, let us consider the factor variable $w = (w_1, w_2, ..., w_m)$. Again considering $w_1 \in \{c_{1,1}, c_{1,2}, ..., c_{1,j_a}\}$; $w_2 \in \{c_{2,1}, c_{2,2}, ..., c_{2,j_a}\}$ and $w_m \in \{c_{m,1}, c_{m,2}, ..., c_{m,j_a}\}$. Now consider these factors levels as a column of the matrix W_F, we can write as:

$$W_F = \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,j_a} \\ C_{1,2} & C_{2,2} & \cdots & C_{2,j_a} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1,j_a} & C_{2,j_a} & \cdots & C_{m,j_a} \end{bmatrix} \tag{B.3.3}$$

Now we have to calculate the factor correlation matrix $(T_1, T_2, ..., T_m)$ corresponding to $w_1, w_2, ...w_m$ by using the approaches discussed in Section 6.3. Let e_l is the l^{th} column of the matrix W_F and it can be presented as,

$$f_l = W_F \; e_l, \tag{B.3.4}$$

where, $l = 1, 2, ...m$. Let again $T_1^{'}$ is the subset of the factor correlation matrix W_F on the basis of Equation B.3.4 when $l = 1$. And, e_l is the vector of 0 with 1 in position l. Similarly for $T_2^{'}$, when $l = 2$ and so on for $T_m^{'}$ when $l = m$ can be expressed as following Equations.

$$\begin{aligned}
\left[T_1^{'}\right]_{ij} &= \left[T_1\right]_{f_{1,i}, f_{1^{'},j}} \\
\left[T_2^{'}\right]_{ij} &= \left[T_2\right]_{f_{2,i}, f_{2^{'},j}} \\
&\vdots \\
\left[T_m^{'}\right]_{ij} &= \left[T_m\right]_{f_{m,i}, f_{m^{'},j}}
\end{aligned} \tag{B.3.5}$$

The whole procedure is presented as follows by an Algorithm in a compact way.

Algorithm 3 Algorithm for Block Structure Formulation

1: Calculate $R^{'}_{\left[\theta,T,g\right]}$ for continuous inputs.
2: Calculate the Factor correlation matrix W_F, using approaches.
3: Extract the column of the factor matrix $f_l; l = 1, 2, ..., m$
4: Calculate $T_1^{'}$ using the concept from Equation (B.3.4).
5: Repeat the same procedure for $T_2^{'}, ..., T_m^{'}$.
6: Calculate $T = T_1^{'} \times T_2{'}\times, ..., \times T_m{'}$
7: Calculate $Cor(v, v^{'}) = T \times R^{'}_{\left[\theta,T,g\right]}$.

B.4 Results of Factors Weather, Steepness and Soil

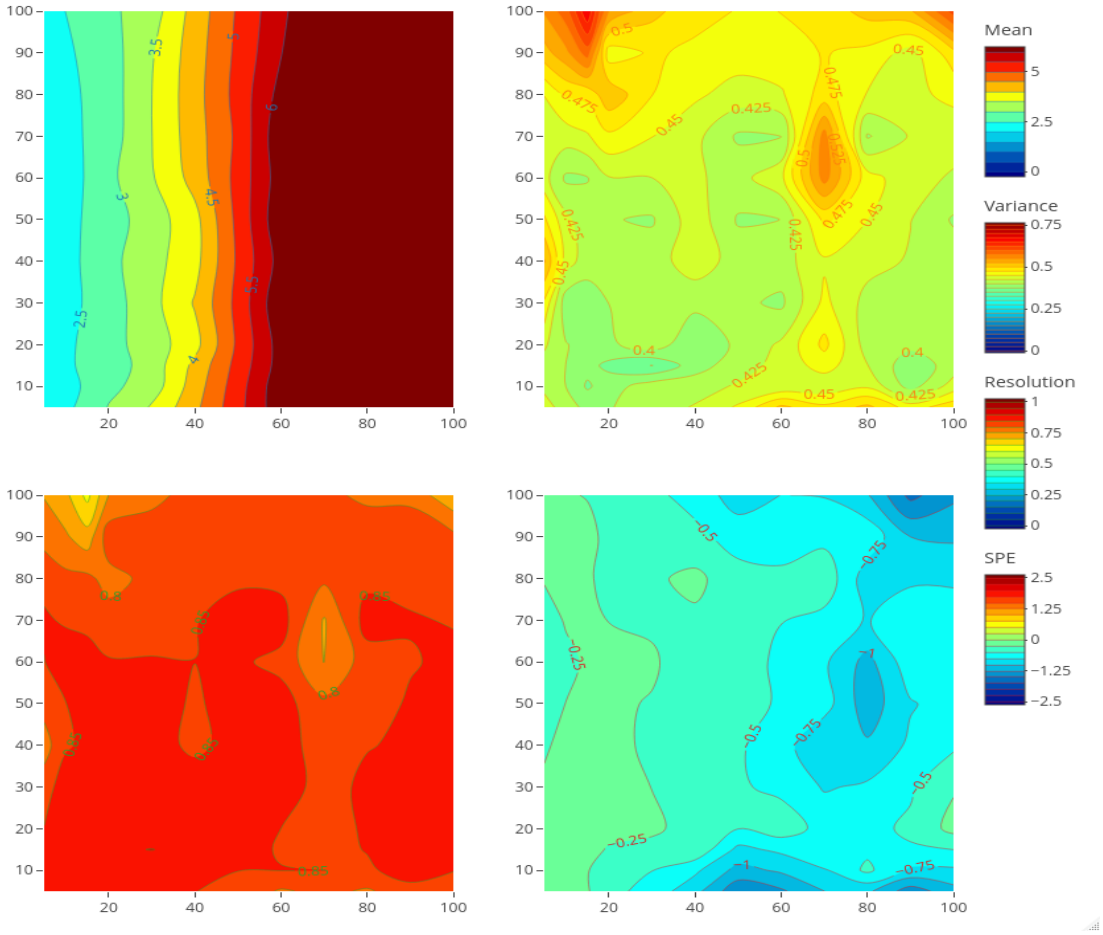


Figure B.3: 3rd Order mean function when $Sy = 6$, $So = 5$, $Wy = 1$, $Sy = 4$. Top left: Emulator Adjusted Expectation; Top right: Emulator adjusted Standard Deviations ($\sqrt{Var}$); Bottom left: Emulator Resolution; Bottom right: SPE

Figure B.3 shows the results of weather, steepness and soil factors using the 60 % training and 40 % testing data points with the same hyperparameters. The posterior mean shows the increasing trend concerning N but no effect on P response. The adjusted standard deviations plot shows the narrower uncertainty for most input space and a growing trend in the edge, where the N is very low and P are very high. The resolution plot shows that most values are around 0.85, indicating the emulator can explain most of the variability. The diagnostic of the SPE also shows no points lying outside the threshold, indicating no conflict between the emulator and the simulator.

Nomenclature

μ_i^N N only MB model, ... page 42

μ_i^P P only MB model, ... page 52

$1-D$ One Dimensional, ... page 63

$2-D$ Two Dimensional, ... page 63

α_k Shape parameter for $k = 0, 1, 2, 3, 4$, ... page 41

α_{ja} Actual acceptance probability, ... page 149

β_0 Coefficient for maximum yield, ... page 25

β_1 Parameter corresponding to fertilizer N, ... page 25

β_2 Slope for N, ... page 25

β_3 Parameter corresponding to fertilizer for P, ... page 25

β_4 Slope for P, ... page 25

β_k Set of regression coefficients for $k = 0, 1, 2, 3, 4$, ... page 41

β_j Vector of unknown regression coefficients , ... page 67

Δ Arbitrary choosing point, ... page 149

δ Nugget effect, ... page 76

ϵ Model discrepancy, ... page 114

γ Vector of factor effect parameters, ... page 42

γ_0 Coefficient similar to the maximum yield of β_0, ... page 42

γ_{ja} All possible sets explored during MCMC, ... page 149

$\hat{R}$ Potential scale reduction factor, ... page 47

Λ Tuning parameter, ... page 148

$\lambda^{j-1}, k^{j,0}$ Initial values, ... page 149

λ_k Rate parameter for $k = 0, 1, 2, 3, 4$, ... page 41

$\mathbb{R}^p$ Generic set of outcome with dimension p, ... page 5

$\mathcal{B}$ Set of $\beta_i;\ i = 0, 1, 2, 3, 4$, ... page 39

$\mathcal{B}^k$ Simulations for the parameters $\mathcal{B}; k = 1, ..., K$, ... page 45

$\mu(.)$ Mean function of Normal distribution, ... page 39

μ_i^{NP} M-B model with both N and P, ... page 52

$\nu(x)$ Nugget term for the active inputs, ... page 68

ϕ Precision, ... page 74

$\pi(\beta_k; \sigma)$ Prior joint distribution , ... page 40

$\pi(\theta)$ Prior distribution for parameter θ, ... page 38

$\pi(y)$ Marginal distribution of y, ... page 38

σ Variance parameter for the Normal distribution, ... page 40

σ^2 Variance of the residual between simulator outputs, ... page 68

σ_ν^2 Variance of the nugget term for active inputs, ... page 68

θ Correlation length, ... page 68

a Integration time, ... page 149

b integration step size, ... page 149

b_0 Coefficient related to yield for the utility, ... page 115

b_1 Coefficient related to the N_p pollution, ... page 115

b_2 Coefficient related to the P_p pollution, ... page 115

c Threshold value to determine optimal region, ... page 114

$Cov(P, Q)$ Covariance between P and Q, ... page 71

$Cov_Q[P_1, P_2]$ Covariance of the sub-collection P_1 and P_2 of the observation P, .. page 71

D Covariance matrix for $f(k)$, ... page 148

d Dimension of the surface unit, ... page 92

$DRNN$ Soluble Nitrogen in drainage outflow, ... page 17

$DRNP$ Soluble Phosphorus in drainage outflow, ... page 17

e Unknown errors, ... page 113

$E(P)$ Prior mean for training data, ... page 71

$E_Q[P]$ Adjusted expectation for P and Q, ... page 71

F Simulator, ... page 64

f Emulator, ... page 65

$f(.)$ Density function of the Gamma distribution, ... page 40

$f(v)$ Emulator for the mixed inputs v, ... page 89

$f(X_p)$ Set of training points over the input space X, ... page 67

G Matrix of basis function, ... page 70

$g(v)$ Basis for the mixed input, ... page 89

$g_j(x)$ Vector of basis function, ... page 67

$Ga(\alpha, \lambda)$ Gamma distribution with rate parameter α and shape parameter λ, .. page 39

H Hastings ratio, ... page 146

$H(\lambda, k)$ Hamiltonian dynamics, ... page 148

$I(x)$ Implausibility measure, ... page 113

R_θ Correlation function containing θ, ... page 76

$R_{[\theta,\tau]}$ Gaussian kernel for continuous inputs correlation $R_{[\theta]}$ with the factor correlation matrix T, ... page 90

$R_{f(x)}[f(x^{'})]$ Resolution, ... page 81

s Positive power parameter, ... page 69

So Soil, ... page 17

$SSFN$ Nitrogen in subsurface flow, ... page 17

$SSFP$ Phosphorus in subsurface flow, ... page 17

St Steepness, ... page 17

Sy Simulation Year, ... page 17

t Elements of factor correlation matrix, ... page 91

t/ha Tonnes per hectare, ... page 16

T_a Correlation matrix for the factor variables, ... page 90

$t_{i,k}$ Correlation between levels i and k, ... page 90

T_{So} Correlation matrix for factor steepness, ... page 126

T_{St} Correlation matrix for factor soil, ... page 126

u Shape parameter for σ, ... page 41

$u(v)$ Residual process for mixed inputs, ... page 89

$u(x)$ Residual process for input x, ... page 67

$u(x_A)$ Residual variance with active inputs, ... page 67

U^* Maximum Utility Value, ... page 116

V Matrix of residual process, ... page 75

v Mixed input consisting of factor input w and quantitative input x, ... page 89

$v(\lambda)$ Potential energies, ... page 148

$Var(P)$ Posterior mean for training data, ... page 71

$Var_Q[P]$ Adjusted variance for P and Q, ... page 71

w Qualitative inputs, ... page 89

Wy Weather, ... page 17

x Inputs used for simulation, ... page 64

X_a Collection of random variables indexed by time or space a, ... page 69

x_A Active inputs, ... page 67

X_p Design points for $p = 1, 2, ..., n$, ... page 67

Y Yield, ... page 23

y Real observation, ... page 113

$y^{'}$ Emulator output, ... page 65

Y_i Yield for the ith observations, ... page 39

y_s Simulation value for SPE, ... page 81

Y_{max} Maximum Yield, ... page 46

Z Model matrix for factors, ... page 43

z Observed values to the simulator F, ... page 113

$Z_{i,j}$ Design matrix for the factor inputs, ... page 42

BL Bayes Linear, ... page 72

EC Exchangeable Correlation, ... page 91

ELPD Expected Log point-wise Predictive Density, ... page 45

EPIC Environmental Policy Integrated Climate, ... page 9

FYLD Forage/Foliage/ Fodder Yield, ... page 16

GC General Correlation, ... page 92

GP Gaussian process, ... page 72

GYLD Grain Yield, ... page 16

HMC Hamiltonian Monte Carlo, ... page 148

LOOIC Leave-One-Out cross-validation criterion, ... page 44

LPD Log point-wise density, ... page 45

M-B Mitscherlich-Baule, ... page 25

MC McMillan approach, ... page 91

MCMC Markov Chain Monte Carlo, ... page 38

MD Mahalanobis distance, ... page 81

MH Metropolis-Hastings, ... page 44

NATMAP National Soil Map of England and Wales, ... page 15

NRLOAD Nitrogen to the river, ... page 16

NSRI National Soil Resources Institute, ... page 15

NUTS No-U-Turn sampler, ... page 37

PRLOAD Phosphorus to the river, ... page 17

SPE Standardised prediction errors, ... page 81

U(.) Utility function, ... page 112

WAIC Widely Applicable Information Criterion, ... page 44

WOSR Winter Oil Seed Rape, ... page 14

Bibliography

[1] T. Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S". *Philosophical transactions of the Royal Society of London* 53 (1763), pp. 370–418.

[2] C. P. Winsor. "The Gompertz curve as a growth curve". *Proceedings of the National Academy of Sciences of the United States of America* 18.1 (1932), p. 1.

[3] A. C. Aitken. "IV.—On least squares and linear combination of observations". *Proceedings of the Royal Society of Edinburgh* 55 (1936), pp. 42–48.

[4] M. Kac, A. Siegert. "An explicit representation of a stationary Gaussian process". *The Annals of Mathematical Statistics* 18.3 (1947), pp. 438–442.

[5] R. H. Strotz. "Cardinal utility". *The American Economic Review* 43.2 (1953), pp. 384–397.

[6] J. A. Nelder, R. Mead. "A simplex method for function minimization". *The computer journal* 7.4 (1965), pp. 308–313.

[7] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika* (1970).

[8] J. Kmenta, L. R. Klein. *Elements of econometrics*. Vol. 655. Macmillan New York, 1971.

[9] R. Hocking. "Theanalysis and selection of variables in linear regression". *Biometrica* 32 (1976), pp. 1–49.

[10] J. M. Bernardo. "Reference posterior distributions for Bayesian inference". *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979), pp. 113–128.

[11] D. W. Hosmer, S. Lemesbow. "Goodness of fit tests for the multiple logistic regression model". *Communications in statistics-Theory and Methods* 9.10 (1980), pp. 1043–1069.

[12] C. Jones et al. "A simplified soil and plant phosphorus model: I. Documentation 1". *Soil Science Society of America Journal* 48.4 (1984), pp. 800–805.

[13] J. Williams, C. Jones, P. Dyke. "The EPIC model and its application". *Proc. Int. Symp. on minimum data sets for agrotechnology transfer*. 1984, pp. 111–121.

[14] C. Ackello-Ogutu, Q. Paris, W. A. Williams. "Testing a von Liebig crop response function against polynomial specifications". *American Journal of Agricultural Economics* 67.4 (1985), pp. 873–880.

[15] D. V. Lindley. *Making Decisions*. John Wiley & Sons, 1985.

[16] A. Sharpley. "The selection erosion of plant nutrients in runoff". *Soil Science Society of America Journal* 49.6 (1985), pp. 1527–1534.

[17] S. S. Grimm, Q. Paris, W. A. Williams. "A von Liebig model for water and nitrogen crop response". *Western Journal of Agricultural Economics* 12.1836-2016-150922 (1987), pp. 182–192.

[18] P. Davies. *Kendall's Advanced Theory of Statistics. Volume 1. Distribution Theory*. 1988.

[19] J. Frost. "Effects on crop yields of machinery traffic and soil loosening Part 1. Effects on grass yield of traffic frequency and date of loosening". *Journal of Agricultural Engineering Research* 39.4 (1988), pp. 301–312.

[20] C. Pollock, C. Eagles. "Low temperature and the growth of plants." *Symposia of the Society for Experimental Biology*. Vol. 42. 1988, pp. 157–180.

[21] J. Sacks et al. "Design and analysis of computer experiments". *Statistical science* 4.4 (1989), pp. 409–423.

[22] J. Williams. "EPIC: The erosion-productivity impact calculator" (1989).

[23] M. D. Frank, B. R. Beattie, M. E. Embleton. "A comparison of alternative crop response models". *American Journal of Agricultural Economics* 72.3 (1990), pp. 597–603.

[24] C. Jones et al. "EPIC: an operational model for evaluation of agricultural sustainability". *Agricultural Systems* 37.4 (1991), pp. 341–350.

[25] G. Casella, E. I. George. "Explaining the Gibbs sampler". *The American Statistician* 46.3 (1992), pp. 167–174.

[26] C. J. Geyer. "Practical markov chain monte carlo". *Statistical science* (1992), pp. 473–483.

[27] A. O'Hagan, E. Glennie, R. Beardsall. "Subjective modelling and Bayes linear estimation in the UK water industry". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41.3 (1992), pp. 563–577.

[28] Q. Paris. "The von Liebig hypothesis". *American Journal of Agricultural Economics* 74.4 (1992), pp. 1019–1028.

[29] W. H. Press. "Downhill simplex method in multidimensions". *Numerical recipes in C* (1992).

[30] R. D. Skeel. "Variable step size destabilizes the Stormer/leapfrog/Verlet method". *BIT Numerical Mathematics* 33.1 (1993), pp. 172–175.

[31] J. Sumelius. "A response analysis of wheat and barley to nitrogen in Finland". *Agricultural and Food Science* 2.6 (1993), pp. 465–479.

[32] F. Pukelsheim. "The three sigma rule". *The American Statistician* 48.2 (1994), pp. 88–91.

[33] S. Chib, E. Greenberg. "Understanding the metropolis-hastings algorithm". *The american statistician* 49.4 (1995), pp. 327–335.

[34] W. R. Gilks, S. Richardson, D. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

[35] J. O. Berger, W. E. Strawderman, et al. "Choice of hierarchical priors: admissibility in estimation of normal means". *The Annals of Statistics* 24.3 (1996), pp. 931–951.

[36] P. S. Craig et al. "Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments". *Case studies in Bayesian statistics*. Springer, 1997, pp. 37–93.

[37] R. V. Llewelyn, A. M. Featherstone. "A comparison of crop production functions using simulated data for irrigated corn in western Kansas". *Agricultural Systems* 54.4 (1997), pp. 521–538.

[38] A. Werker, K. Jaggard. "Modelling asymmetrical growth curves that rise and then fall: applications to foliage dynamics of sugar beet (Beta vulgaris L.)" *Annals of Botany* 79.6 (1997), pp. 657–665.

[39] C. Zhu et al. "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". *ACM Transactions on mathematical software (TOMS)* 23.4 (1997), pp. 550–560.

[40] N. J. McMillan et al. "Analysis of protein activity data by Gaussian stochastic process models". *Journal of Biopharmaceutical Statistics* 9.1 (1999), pp. 145–160.

[41] M. P. Meuwissen et al. "A model to estimate the financial consequences of classical swine fever outbreaks: principles and outcomes". *Preventive Veterinary Medicine* 42.3-4 (1999), pp. 249–270.

[42] C. Rosenzweig et al. "Wheat yield functions for analysis of land-use change in China". *Environmental Modeling & Assessment* 4.2-3 (1999), pp. 115–132.

[43] A. E. Gelfand. "Gibbs sampling". *Journal of the American statistical Association* 95.452 (2000), pp. 1300–1304.

[44] V. Smil. "Phosphorus in the environment: natural flows and human interferences". *Annual review of energy and the environment* 25.1 (2000), pp. 53–88.

[45] P. S. Craig et al. "Bayesian forecasting for complex systems using computer simulators". *Journal of the American Statistical Association* 96.454 (2001), pp. 717–729.

[46] M. C. Kennedy, A. O'Hagan. "Bayesian calibration of computer models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464.

[47] J. M. Epstein. "Modeling civil violence: An agent-based computational approach". *Proceedings of the National Academy of Sciences* 99.suppl_3 (2002), pp. 7243–7250.

[48] G. Hammer et al. "Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement". *European Journal of Agronomy* 18.1-2 (2002), pp. 15–31.

[49] C. M. Theobald, M. Talbot, F. Nabugoomu. "A Bayesian approach to regional and local-area prediction from crop variety trials". *Journal of agricultural, biological, and environmental statistics* 7.3 (2002), pp. 403–419.

[50] K. Alivelu et al. "Comparison of modified Mitscherlich and response plateau models for calibrating soil test based nitrogen recommendations for rice on Typic Ustropept". *Communications in soil science and plant analysis* 34.17-18 (2003), pp. 2633–2643.

[51] A. Blasco, M. Piles, L. Varona. "A Bayesian analysis of the effect of selection for growth rate on growth curves in rabbits". *Genetics Selection Evolution* 35.1 (2003), pp. 21–41.

[52] T. J. Santner et al. *The design and analysis of computer experiments*. Vol. 1. Springer, 2003.

[53] S. Boyd, S. P. Boyd, L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[54] G. Hoogenboom, J. W. White, C. D. Messina. "From genome to crop: integration through simulation modeling". *Field Crops Research* 90.1 (2004), pp. 145–163.

[55] M. Seeger. "Gaussian processes for machine learning". *International journal of neural systems* 14.02 (2004), pp. 69–106.

[56] A. O'Hagan. "Bayesian analysis of computer code outputs: A tutorial". *Reliability Engineering & System Safety* 91.10-11 (2006), pp. 1290–1300.

[57] M. J. H. Amanullah, K. Nawab, A. Ali. "Response of specific leaf area (SLA), leaf area index (LAI) and leaf area ratio (LAR) of maize (Zea mays L.) to plant density, rate and timing of nitrogen application". *World Applied Sciences Journal* 2.3 (2007), pp. 235–243.

[58] R. Finger, S. Schmid. *Modeling agricultural production risk and the adaptation to climate change*. Tech. rep. 2007.

[59] M. Goldstein, D. Wooff. *Bayes linear statistics: Theory and methods*. Vol. 716. John Wiley & Sons, 2007.

[60] V. R. Joseph, J. D. Delaney. "Functionally induced priors for the analysis of experiments". *Technometrics* 49.1 (2007), pp. 1–11.

[61] N. R. Council et al. *Global Challenges and Directions for Agricultural Biotechnology: Workshop Report*. National Academies Press, 2008.

[62] R. Finger, W. Hediger. "The application of robust regression to a production function comparison". *The Open Agriculture Journal* 2.1 (2008).

[63] P. Z. G. Qian, H. Wu, C. J. Wu. "Gaussian process models for computer experiments with qualitative and quantitative factors". *Technometrics* 50.3 (2008), pp. 383–396.

[64] L. S. Bastos, A. O'Hagan. "Diagnostics for Gaussian process emulators". *Technometrics* 51.4 (2009), pp. 425–438.

[65] D. Cordell, J.-O. Drangert, S. White. "The story of phosphorus: global food security and food for thought". *Global environmental change* 19.2 (2009), pp. 292–305.

[66] J. A. Cumming, M. Goldstein. "Small sample bayesian designs for complex high-dimensional models based on information gained using fast approximations". *Technometrics* 51.4 (2009), pp. 377–388.

[67] M. Goldstein, J. Rougier. "Reified Bayesian modelling and inference for physical systems". *Journal of statistical planning and inference* 139.3 (2009), pp. 1221–1239.

[68] L. House, M. Goldstein, I. Vernon. "Exchangeable computer models." (2009).

[69] P. Steduto et al. "Concepts and applications of AquaCrop: The FAO crop water productivity model". *Crop modeling and decision support*. Springer, 2009, pp. 175–191.

[70] D. A. Vaccari. "Phosphorus: a looming crisis". *Scientific American* 300.6 (2009), pp. 54–59.

[71] R. G. Bower, M. Goldstein, I. Vernon. "Galaxy formation: a Bayesian uncertainty analysis". *Bayesian analysis* 5.4 (2010), pp. 619–669.

[72] J. A. Cumming, M. Goldstein. "Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments". *The Oxford handbook of applied Bayesian analysis* (2010), pp. 241–270.

[73] T. Gaiser et al. "Validation and reliability of the EPIC model to simulate maize production in small-holder farming systems in tropical sub-humid West Africa and semi-arid Brazil". *Agriculture, ecosystems & environment* 135.4 (2010), pp. 318–327.

[74] I. Vernon, M. Goldstein, et al. "A Bayes Linear approach to systems biology." (2010).

[75] S. Watanabe, M. Opper. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of machine learning research* 11.12 (2010).

[76] D. Williamson. "Policy making using computer simulators for complex physical systems; Bayesian decision support for the development of adaptive strategies". PhD thesis. Durham University, 2010.

[77] S. Asseng, I. Foster, N. C. Turner. "The impact of temperature variability on wheat yields". *Global Change Biology* 17.2 (2011), pp. 997–1012.

[78] P. Challenor. "Using emulators to estimate uncertainty in complex models". *IFIP Working Conference on Uncertainty Quantification*. Springer. 2011, pp. 151–164.

[79] A. O. Finley, S. Banerjee, B. Basso. "Improving crop model inference through Bayesian melding with spatially varying parameters". *Journal of agricultural, biological, and environmental statistics* 16.4 (2011), pp. 453–474.

[80] J. A. Foley et al. "Solutions for a cultivated planet". *Nature* 478.7369 (2011), pp. 337–342.

[81] M. Goldstein. "Bayes linear analysis for complex physical systems modeled by computer simulators". *IFIP Working Conference on Uncertainty Quantification*. Springer. 2011, pp. 78–94.

[82] MUCM toolkit. *Discussion: Theoretical Aspects of Bayes linear*. 2011.

[83] R. M. Neal et al. "MCMC using Hamiltonian dynamics". *Handbook of Markov chain monte carlo* 2.11 (2011), p. 2.

[84] Q. Zhou, P. Z. Qian, S. Zhou. "A simple approach to emulation for computer models with qualitative and quantitative factors". *Technometrics* 53.3 (2011), pp. 266–273.

[85] I. Andrianakis, P. G. Challenor. "The effect of the nugget on Gaussian process emulators of computer models". *Computational Statistics & Data Analysis* 56.12 (2012), pp. 4215–4228.

[86] A. Bationo et al. "Knowing the African soils to improve fertilizer recommendations". *Improving soil fertility recommendations in Africa using the decision support system for agrotechnology transfer (DSSAT)*. Springer, 2012, pp. 19–42.

[87] C. T. Paine et al. "How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists". *Methods in Ecology and Evolution* 3.2 (2012), pp. 245–256.

[88] S. J. Parikh, B. R. James. "Soil: the foundation of agriculture". *Nature Education Knowledge* 3.10 (2012), p. 2.

[89] Texas A & M Agrilife Research. *EPIC & APEX Models*. Epic User Guide. 2012.

[90] J. C. Wang et al. "A Bayesian approach to estimating agricultural yield based on multiple repeated surveys". *Journal of agricultural, biological, and environmental statistics* 17.1 (2012), pp. 84–106.

[91] A. Gelman et al. *Bayesian data analysis*. CRC press, 2013.

[92] J. P. Gosling et al. "A Bayes Linear approach to weight-of-evidence risk assessment for skin allergy". *Bayesian Analysis* 8.1 (2013), pp. 169–186.

[93] L. D. Landau, E. M. Lifshitz. *Course of theoretical physics*. Elsevier, 2013.

[94] P. Oteng-Darko et al. "Crop modeling: A tool for agricultural research–A review" (2013).

[95] D. Williamson et al. "History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble". *Climate dynamics* 41.7 (2013), pp. 1703–1729.

[96] S. Asseng et al. "Simulation modeling: applications in cropping systems" (2014).

[97] H. Chen, J. Yamagishi, H. Kishino. "Bayesian inference of baseline fertility and treatment effects via a crop yield-fertility model". *PloS one* 9.11 (2014).

[98] D. Deryng et al. "Global crop yield response to extreme heat stress under multiple climate change futures". *Environmental Research Letters* 9.3 (2014), p. 034011.

[99] A. Grow, J. Hilton. "Statistical emulation". *Wiley StatsRef: Statistics Reference Online* (2014), pp. 1–8.

[100] M. D. Hoffman, A. Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[101] The Fertilizer Institute. *Fertilizer 101: The Big 3 - Nitrogen, Phosphorus and Potassium*. Online; accessed 07 May 2014. 2014.

[102] W. Xiong et al. "A calibration procedure to improve global rice yield simulations with EPIC". *Ecological modelling* 273 (2014), pp. 128–139.

[103] C. C. Caiado, M. Goldstein. "Bayesian uncertainty analysis for complex physical systems modelled by computer simulators with applications to tipping points". *Communications in Nonlinear Science and Numerical Simulation* 26.1-3 (2015), pp. 123–136.

[104] I. Fao. "WFP. 2015". *The state of food insecurity in the world* (2015), pp. 1–62.

[105] M. Kadiyala et al. "An integrated crop model and GIS decision support system for assisting agronomic decision making under climate change". *Science of the Total Environment* 521 (2015), pp. 123–134.

[106] A. J. Sindelar, J. A. Lamb, J. A. Coulter. "Short-term stover, tillage, and nitrogen management affect near-surface soil organic matter". *Soil Science Society of America Journal* 79.1 (2015), pp. 251–260.

[107] Y. Zhang, W. I. Notz. "Computer experiments with qualitative and quantitative variables: A review and reexamination". *Quality Engineering* 27.1 (2015), pp. 2–13.

[108] T. V. Elzhov et al. "Package 'minpack. lm'". *Title R Interface Levenberg-Marquardt Nonlinear Least-Sq. Algorithm Found MINPACK Plus Support Bounds* (2016).

[109] M. Mew. "Phosphate rock costs, prices and resources interaction". *Science of the Total Environment* 542 (2016), pp. 1008–1012.

[110] T. J. Salo et al. "Comparing the performance of 11 crop simulation models in predicting yield response to nitrogen fertilization". *The Journal of Agricultural Science* 154.7 (2016), pp. 1218–1240.

[111] Y. Sheng et al. "Input substitution, productivity performance and farm size". *australian Journal of agricultural and Resource Economics* 60.3 (2016), pp. 327–347.

[112] S. D. Team et al. "RStan: the R interface to Stan". *R package version* 2.1 (2016), p. 522.

[113] P. A. Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

[114] M. Goldstein, N. Huntley. "Bayes linear emulation, history matching and forecasting for complex computer simulators". *Handbook of Uncertainty Quantification, Springer, Cham, Switzerland* (2017).

[115] United States Environmental Protection Agency. *Crop Damage Complaints Related to Dicamba Herbicides Raising Concerns*. Online; accessed July 2017. 2017.

[116] A. Vehtari, A. Gelman, J. Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". *Statistics and computing* 27.5 (2017), pp. 1413–1432.

[117] L. York, G. Lobet. "The American Society of Plant Biologists". *The Plant Cell* 2017 (2017).

[118] C. Zhao et al. "Temperature increase reduces global yields of major crops in four independent estimates". *Proceedings of the National Academy of Sciences* 114.35 (2017), pp. 9326–9331.

[119] M. P. Hoffmann et al. "Exploring adaptations of groundnut cropping to prevailing climate variability and extremes in Limpopo Province, South Africa". *Field Crops Research* 219 (2018), pp. 1–13.

[120] S. Jackson, I. Vernon. "Design of physical system experiments using bayes linear emulation and history matching methodology with application to arabidopsis Thaliana". PhD thesis. Durham University, 2018.

[121] I. Vernon et al. "Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions". *BMC systems biology* 12.1 (2018), pp. 1–29.

[122] A. Wilson, C. Dent, M. Goldstein. "Quantifying uncertainty in wholesale electricity price projections using Bayesian emulation of a generation investment model". *Sustainable Energy, Grids and Networks* 13 (2018), pp. 42–55.

[123] D. J. Choruma, J. Balkovic, O. N. Odume. "Calibration and validation of the EPIC model for maize production in the Eastern Cape, South Africa". *Agronomy* 9.9 (2019), p. 494.

[124] Economic and Social Research Council. *Spatially targeted and coordinated regulation of agricultural externalities: an economic perspective*. 2019.

[125] International Institute for Applied Systems Analysis. *The Environmental Policy Integrated Model (EPIC)*. Online; accessed 30 July 2019. 2019.

[126] M. Nishio, A. Arakawa. "Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for estimating genetic parameters and breeding values". *Genetics Selection Evolution* 51.1 (2019), pp. 1–12.

[127] I. Vernon, S. E. Jackson, J. A. Cumming. "Known boundary emulation of complex computer models". *SIAM/ASA Journal on Uncertainty Quantification* 7.3 (2019), pp. 838–876.

[128] Eden DTC. *Eden DTC A National Demonstration Test Catchment*. 2020.

[129] S. E. Jackson et al. "Understanding hormonal crosstalk in Arabidopsis root development via emulation and history matching". *Statistical Applications in Genetics and Molecular Biology* 19.2 (2020).

[130] Met Office. *Weather and climate data*. 2020.

[131] R. Shirley et al. "An empirical, Bayesian approach to modelling crop yield: Maize in USA". *Environmental Research Communications* (2020).

[132] University of East Anglia and Wensum Alliance. *River Wensum Demonstration Test Catchment Project*. 2020.

[133] H. Asai, K. Saito, K. Kawamura. "Application of a Bayesian approach to quantify the impact of nitrogen fertilizer on upland rice yield in sub-Saharan Africa". *Field Crops Research* 272 (2021), p. 108284.

[134] M. M. Hasan, J. A. Cumming. "Bayes Linear Emulation of Simulated Crop Yield". *Canadian Conference in Applied Statistics*. Springer. 2021, pp. 145–151.

[135] A. Vehtari et al. "Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion)". *Bayesian analysis* 16.2 (2021), pp. 667–718.

[136] O. Bazrafshan et al. "Predicting crop yields using a new robust Bayesian averaging model based on multiple hybrid ANFIS and MLP models". *Ain Shams Engineering Journal* 13.5 (2022), p. 101724.

[137] M. Dunne et al. "Complex model calibration through emulation, a worked example for a stochastic epidemic model". *Epidemics* (2022), p. 100574.

[138] M. M. Hasan, J. A. Cumming. "A Bayesian hierarchical framework for emulating a complex crop yield simulator". *arXiv preprint arXiv:2207.12999* (2022).

[139] S. E. Jackson, I. Vernon. "Efficient Emulation of Computer Models Utilising Multiple Known Boundaries of Differing Dimension". *Bayesian Analysis* 1.1 (2022), pp. 1–27.

[140] R. H. Oughton, M. Goldstein, J. C. Hemmings. "Intermediate Variable Emulation: using internal processes in simulators to build more informative emulators". *SIAM/ASA Journal on Uncertainty Quantification* 10.1 (2022), pp. 268–293.

[141] J. Owen, I. Vernon. "Bayesian Uncertainty Analysis and Decision Support for Complex Models of Physical Systems with Application to Production Optimisation of Subsurface Energy Resources". PhD thesis. Durham University, 2022.

[142] I. Vernon, J. P. Gosling. "A Bayesian Computer Model Analysis of Robust Bayesian Analyses". *Bayesian Analysis* (2022).

[143] Z. Wang et al. "Review of application of EPIC crop growth model". *Ecological Modelling* 467 (2022), p. 109952.

[144] A. L. Wilson, M. Goldstein, C. J. Dent. "Varying coefficient models and design choice for Bayes linear emulation of complex computer models with limited model evaluations". *SIAM/ASA Journal on Uncertainty Quantification* 10.1 (2022), pp. 350–378.

[145] N. Gujrati. *Damodar,(1995)," Basic Econometrics"*.

[146] lendIS. *Soil Series Properties*. Online accessed 29/4/2020.

[147] Mosaic CropNutrition. *Nitrogen in Plants*. Online accessed.

[148] Selina Wamucii. *United Kingdom (UK) Barley Prices*. Online; accessed 29 June 2022.