Investigating uncertainty and emulating process-based models with multivariate outputs, applied to aquaculture

Michael Currie

This thesis is submitted in fulfilment of the requirements
of the degree of Doctor of Philosophy

School of Mathematics & Statistics
College of Science and Engineering
University of Glasgow

Abstract

There remain environmental challenges which can only accurately be assessed by process-based modelling. An example of this is the monitoring of the environmental impacts of aquaculture, where the logistical difficulty and cost of collecting data over large areas make mathematical modelling the more effective approach. Such approaches are computationally intensive and do not account for uncertainty. NewDEPOMOD is an example of a process-based model that is used within aquaculture to model the environmental impacts of aquaculture. This thesis provides an in-depth investigation of uncertainty in such a model using sensitivity analysis, and proposes a novel statistical emulation framework to approximate the output from NewDEPOMOD, reducing the computational cost.

NewDEPOMOD is a complex mathematical model that was developed in order to estimate and predict the transportation of waste particles from fish farms to their deposition on the seabed. It features a number of different types of input, representing features such as the fish farm physical structure, flow speeds and waste transportation properties. In addition, the output produced by NewDEPOMOD provides a measure known as Solids Flux in grid cells across the domain, representing the environmental impact. This can be visualised as either a univariate or multivariate output.

The univariate outputs produced by NewDEPOMOD are the Total Area Impacted and 99th Percentile of Solids Flux which provide a measure of the size and intensity of the impact on the seabed. In collaboration with the Scottish Environment Protection agency ('SEPA'), with application to fish farm sites around the coast of Scotland, a set of inputs were identified as being of most importance for investigating the effect of their uncertainty on the NewDEPOMOD outputs. In this thesis, sensitivity analyses are conducted at multiple fish farm sites, classed as high and low energy based on their flow speeds, using random forest models. Random forest models are proposed as they are flexible, efficient, and the importance values produced by the models can be used to rank the inputs based on their influence on the output data.

To assess the impact of changing the inputs values on the output maps produced by NewDEPOMOD, traditional univariate sensitivity analysis techniques are expanded here to develop novel sensitivity analysis methods for considering multivariate model outputs. Three different approaches to investigating the output maps are considered: 1) shape analysis based on a landmark approach for identifying the main shape of the impact, 2) bivariate functional analysis where the output maps are considered as smooth surfaces, and 3) grid cell approach where the Solids Flux in each grid cell is considered individually. The performance of each approach was considered individually before developing a framework, using a subset of the approaches, that could be applied to multiple sites to assess parameter uncertainty, and hence the impact of altering the inputs on the output maps.

The application of statistical emulation to model the univariate outputs from NewDEPOMOD reducing the computational cost is a novel approach. The methods proposed for the emulation are random forests and Gaussian processes which both provide flexibility and allow for fast predictions for new data in comparison to the time taken to run NewDEPOMOD. For each site, training data will be used to fit the emulation models for each approach before using a test set of data to assess their predictive performance. Root Mean Squared Error ('RMSE') and the Mean Absolute Error ('MAE') are both considered as measures of how well the emulators perform and allow for comparisons to be made between the approaches. Further investigation assesses the suitability of a single emulator to be used at all sites, or whether the emulators should be built individually for each site.

In practice, correlated outputs are more realistic in such a scenario and hence the emulation framework for the univariate outputs is expanded to consider the univariate outputs together as a correlated multivariate output. Extensions to the random forest and Gaussian process models are proposed which account for correlation between the outputs. The predictive performance for both approaches can be reviewed using RMSE and MAE to determine if there are improvements when modelling the univariate outputs together as a correlated output.

This research provides a deeper understanding of NewDEPOMOD through the development of novel sensitivity analysis and emulation tools for computationally efficient analyses of data on the impact of fish farms. Remarks on the approaches used and their results are provided throughout this thesis, along with potential future extensions to the research.

Contents

List of Tables

List of Figures

Declaration of Authorship

I, Michael Currie, declare that this thesis titled, 'Investigating uncertainty and emulating process-based models with multivariate output, applied to aquaculture' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 13 September 2022

Acknowledgements

Firstly, I would like to thank my supervisors, Prof Claire Miller and Prof Marian Scott. You have both been extremely supportive throughout my PhD, boosting my confidence when it was low and encouraging me when I was finding it tough. You were always able to help me see the positives at every stage, and I am not sure how I am going to cope without our weekly Thursday meetings. I couldn't have done this without your supervision and guidance, and I will always be grateful.

I would also like to thank EPSRC and SEPA for their generous funding. Special thanks go to Dr Alan Hills for your suggestions and Dr Andrew Berkeley, for your help and advice with NewDEPOMOD.

Thank you to my fellow PhD students, especially Ivona, Florence, Alan and Kamol, you helped me to solve my problems when things weren't going to plan and always put a smile on my face when I was finding work difficult.

I would like to thank my friends and family for all of your encouragement and support over my academic career. You have learned more about fish farming in the last four years than you ever would have expected to and were always a welcome distraction when I needed to take my mind off work. In particular, I would like to thank my mum and dad, for everything you have done over the years. You have encouraged me to succeed since I was young, and your belief in me has helped keep me going through the highs and lows.

My final thanks goes to my wife, Kelly, you have been there every step of the way. Not only did you give me the confidence to consider a PhD in the first place, you have been incredibly supportive the whole way through. I know there have been some surprises along the way (including three wedding dates, two dogs and a pandemic), but I wouldn't have wanted to experience it with anyone else.

Chapter 1

Introduction & Background

1.1 Background to the research

Aquaculture accounts for nearly half of the global fish supply and it is anticipated that this will continue to grow as global demand for seafood continues to rise (Campbell & Pauly 2013). The United Nations Food and Agriculture Organisation ('FAO') have identified aquaculture as being one of the faster growing global food production industries. As a result, regulatory bodies across the world are now having to further scrutinise the environmental impacts of these farms. The industry in Scotland is regulated by the Scottish Environment Protection Agency ('SEPA') and current regulations rely on a mathematical model (NewDEPOMOD), to assess the spatial extent of the environmental impact in Scottish marine waters. Due to a lack of real data collected from the seabed, the model has not been validated accordingly and the uncertainty of some of the model inputs has not been considered in detail. In addition to the lack of validation of the model, the model can be computationally expensive, with single runs taking minutes. In addition, NewDEPOMOD contains a random walk element, which has to be factored into any analysis. This can then be problematic when considering multiple different scenarios to be tested, where potentially thousands of runs are required.

With plans for future expansion of the industry in Scotland, SEPA require more knowledge of this mathematical model to improve regulations and avoid potentially irreparable damage to the seabed. This thesis aims to investigate the properties of NewDEPOMOD and the influence on model predictions of uncertainty in some of the default values for the inputs through sensitivity analyses. This will provide the foundation for creating a statistical emulator of NewDEPOMOD that will allow various scenarios to be tested at fish farm sites, without the computational cost of running NewDEPOMOD.

1.1.1 Aquaculture Background

1.1.1.1 The Aquaculture Industry

Aquaculture as an industry has grown significantly throughout the world since 1980 - with farmed salmon production in particular taking place on all continents excluding Africa (Asche & Bjorndal 1996). Salmon farming focuses on three different species - Atlantic salmon, coho salmon and salmon trout - with Atlantic salmon accounting for more than half of the total output of farmed salmon. With Atlantic salmon being the main species produced in Scotland, this will be the main focus of this project. The aquaculture industry as a whole has grown significantly and in 1980, aquaculture accounted for only 6.5% of the total world fish production, but by 2018, it accounted for 46%, according to the FAO's 'The State of World Fisheries and Aquaculture 2020' (this will be abbreviated in future as 'according to the FAO'). It has grown from a total production of 4.7 million tonnes in 1980, to 82.1 million tonnes in 2020, according to the FAO. It is anticipated that the reliance on aquaculture will continue to grow as global demand for seafood continues to rise (Campbell & Pauly 2013). As expected, with the significant increases in the production of farmed salmon, this has been associated with reductions in the value of salmon, with the price in 2008 less than one third of the price in the early 1980s (adjusted for inflation) (Asche & Bjorndal 1996). The increased production has been a result of improvements in technology, with respect to cage manufacturing, feed, health, and research into the best methods for creating a farm in terms of location and cage setup.

1.1.1.2 Atlantic Salmon Farming

As a result of the declining wild fisheries in the late 1960's in rural Norway, caged salmon farming began in these rural fishing communities with the help of significant governmental support and investment in a bid to rebuild these communities (Willoughby 1999). After the success of floating cage farms in Norway, and due to license restrictions on farm size, many companies decided to invest in creating farms overseas (Willoughby 1999). In the early 1970's, Scotland had taken over as the major producer of Atlantic salmon, before Norway regained its place in 1974, and continues to be the major producer to date according to the FAO. Chile began its rapid rise to become one of the major producers of farmed Atlantic salmon in the late 1980's. Since 1999, it has been the second largest producer of farmed Atlantic salmon according to the FAO. In Norway in 1980, farmed fish accounted for approximately 6.9%

of total fish exports, and by 1990, this had increased to 40.5% due to the various tax and financial incentives throughout the 1980's (Willoughby 1999). The FAO's '2015 Fisheries and Aquaculture Statistics' provide the most recent breakdown of production by country and indicate that Norway continues to be the major producer of farmed salmon, accounting for approximately 54.7% of world Atlantic salmon production. The other two major producers of farmed Atlantic salmon are Chile (25.55%) and Scotland (7.23%).

1.1.2 Structure and Layout of a Fish Farm

A major factor in the success of fish farming in Norway is the location of farms in fjords with depths of up to 300m (Taranger et al. 2015). Many farms in Norway that are situated in sheltered, coastal waters will produce between $3,000$ and $5,000$ tonnes of farmed salmon annually, and in dynamic coastal sites the production can be $14,000$ tonnes in an 18 month period (Taranger et al. 2015).

Currently in Scotland, fish farming has been debated and SEPA have reviewed the regulatory framework, enabling an expansion of the industry. Previously, there was a maximum annual biomass of $2,500$ tonnes for each farm (total weight of fish stocked at a farm). However, SEPA have removed this in the new regulatory framework to allow biomass limits to be matched to the capacity at specific sites - allowing for larger farms to be approved in better locations. Farms located in sustainable locations can now be approved with larger biomass limits - which would have been rejected under the previous framework.

Fish farms in Scotland all follow a similar layout, consisting of groups of six or eight circular cages grouped together. An example of the typical layout of a fish farm can be seen in Figure 1.1. Figure 1.1 illustrates a typical farm in Scotland, which contains circular cages set up in pairs, and in this case they are split up into two groups. In most instances, the cages will be set up in pairs, as one group of potentially 8 cages. However, at certain farms, such as the one in Figure 1.1, they are split up into two groups. In addition to this, the set up at some of the smaller farms consists of rectangular cages that are situated much closer together. The industry in Scotland is evolving, with the new regulatory framework and guidelines for farms in place since 2019. SEPA have also sanctioned licenses to one company for two farms in nearby locations that are operated as one farm - similar to the farm in Figure 1.1, but with a larger gap between the groups of cages. This combined farm was licensed in coordination with SEPA in order to explore the effects of operating two

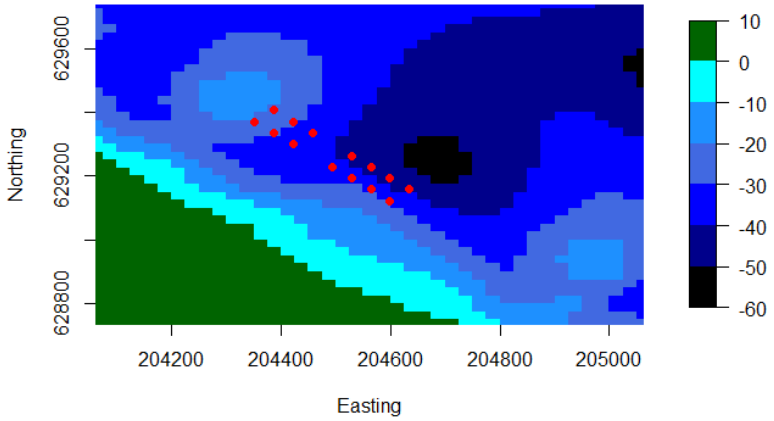


Figure 1.1: Plot of the bathymetry and cage location at a farm in Scotland, with the cages represented by the red circles, land represented by the green grid cells and the light to dark blue representing the depth below the seabed in metres.

nearby farms together, to increase Biomass, and how this would impact the environment. The environmental impacts of fish farming are a crucial factor in SEPA's plans to ensure increased production in the industry can be done in a safe manner.

1.1.3 Environmental Impacts of Fish Farming

The success of a fish farm is dependent upon good environmental conditions (Willoughby 1999), which means that farm operators are as keen as SEPA to monitor and protect the environment around their farm. Intensive fish farming is often criticised for having a negative impact on the environment, and following a 2017 consultation, two Scottish parliamentary committees, and extensive work by SEPA, the new regulatory framework was developed and came into place in 2019. The regulation of the industry has changed since the beginning of the project, and the environmental impacts of fish farming are the driving force for the work, so these are now described in more detail.

At fish farms, faeces and waste feed are released into the water column below the farms and transported by the water current flow and turbulence. Over time the particles will drift towards the seabed, and if the current speeds fall below a critical deposition speed, they will be deposited on the seabed.

After being deposited on the seabed, these particles may be resuspended from the seabed if the current speed increases above a critical value, and released back into the water column and transported by the current. So depending on the current flow in the area surrounding the farm, the particles may be spread out more evenly in faster current speeds from the initial transportation or from resuspension from the seabed. At farms with slower current speeds, the initial transportation of particles will not be as far spread and resuspension may not occur which could result in more intense deposition in the area closer to the farm. The bathymetry of the site also plays a part in the intensity of the depositions. So, even with slower current speeds, if the waters are deeper the intensity of the deposition in the area below the farm may not be as severe, the particles will be in the water column longer and therefore transported further in the initial settling stage.

Fish health is another key factor in the operation of a fish farm as it affects the quality and speed of growth. In addition to this, farmed salmon are prone to outbreaks of sea lice which can cause mortality or also restrict the infected fish from being sold. To protect the fish from sealice there are two common approaches used in Scotland: 1) chemical treatment (either a bath for the fish or by incorporating medicine into the fish food) or 2) using a cleaner fish (wrasse) which feed on the lice. Obviously, using chemicals and medicines mean that the chemicals are released into the environment and can have a negative impact, and so the use of cleaner fish has been increasing.

With increasing levels of aquaculture taking place, it is becoming more important to monitor the environment and try to minimise any negative effects. Concern among governments and the public for the environment around farms has increased following the rapid rise of aquaculture. The main cause for concern has been that irreparable damage will be caused to the environment if future expansion of the industry continues on its current path. In Scotland, the new regulatory framework produced by SEPA aims to manage the different environmental issues faced by the industry, but also provide scope to grow and expand safely. Due to the difficulty faced with trying to collect real data over such a large area, much of the work is completed using computer models that replicate the operation of a fish farm and track the transportation of waste. The model that is used predominantly by SEPA is NewDEPOMOD which will be described in more detail.

1.1.4 Introduction to fish farm sites of interest

Throughout the thesis, a number of different sites will be considered for the analyses. Fish farms throughout Scotland are located in areas with varying site characteristics such as flow speed and direction and bathymetry (underwater depth), as well as different operational aspects such as Biomass and cage properties. SEPA have identified low energy sites (sites with slower flow speeds) as being less environmentally friendly, and are aiming to reduce the environmental impact of the industry by locating farms in higher energy sites. In consultation with SEPA, a number of low and high energy sites were identified for further investigation. These sites would allow the impact of the site characteristics to be considered as SEPA aim to increase production in Scotland safely.

Low energy sites refer to sites that have slower flow speeds and therefore less dispersion of waste from cages. These sites are of interest as the predicted impact is likely to be focused on the seabed below the farm, with less dispersion as waste settles through the water column due to the slower flow speeds. The slower flow speeds also restrict the amount of resuspension and the distance resuspended particles are transported. This results in greater quantities of waste consolidating in the area below the farm which could cause irreparable damage to the seabed. The sites being considered within the thesis can be seen in the map in Figure 1.2 The low energy sites that will be considered are called Ardessie, Ardentinny and West Strome. Ardessie is a relatively small farm with a licensed Biomass of 270 tonnes compared to the larger farms at Ardentinny, which has a licensed Biomass of over $2,000$ tonnes, and West Strome, which also has a licensed Biomass of $2,500$ tonnes.

With high energy sites being identified by SEPA as the preferred locations for larger farms in the future, they will be considered throughout this project. The high energy sites allow waste to be dispersed further and less intensely, and the faster current speeds result in more resuspension events taking place. This is beneficial as the result of more resuspension is that less waste material consolidates on the seabed and reduces the risk of irreparable damage to the seabed. The main high energy site that will be considered throughout the project is Muck. It is a relatively large farm with a licensed Biomass of $2,500$ tonnes, with SEPA considering the potential to increase the Biomass. An additional high energy site will be considered in some analyses, situated at Djuba Wick, which has a licensed Biomass of almost $2,500$ tonnes.

Changes in modelling guidance from SEPA since the beginning of the project have identified that using flat bathymetry when running NewDEPOMOD produces more realistic results. Previously, at the beginning of the project, variable

Figure 1.2: Map showing the locations of the sites in Scotland.

bathymetry was used for running NewDEPOMOD - this will be highlighted at the beginning of the analysis. In line with the changes in the modelling guidance from SEPA, analysis was conducted using sites with flat bathymetry following the introduction of the new guidance which suggested that the variable bathymetry did not allow particles to be transported in a realistic manner near the seabed.

1.2 Introduction to NewDEPOMOD

NewDEPOMOD is a computer particle tracking model which was developed in order to provide better predictions of the impact of large marine cage fish farms on the seabed. It is an updated version of DEPOMOD (Cromey et al. 2002), with most of the updates being related to the user experience, as well as some additional capabilities related to the updated SEPA monitoring guidance. DEPOMOD will be described in more detail to give a general idea of the model. DEPOMOD was created in order to assist the regulatory bodies in the monitoring and licensing of fish farms (Cromey et al. 2002). It is based on the BenOss model (Cromey et al. 1998) that was used for long sea sewage

outfalls. The aim of the BenOss model was to predict "the relative impacts of preliminary, primary and secondary treated sewage effluent; the long-term average of organic carbon accumulating in the near vicinity of a domestic sewage outfall; and the effects of changing carbon deposition on a benthic community" (Cromey et al. 1998). Sewage particles that have been discharged are tracked by the BenOss model as they settle through the water column to the seabed, and then using hydrodynamic data for the area, predictions of the impact on the seabed can be made based on the effects of advection, dispersion, deposition and resuspension (Cromey et al. 1998). With fish farms, the waste follows a similar pattern to that of long sea sewage outfalls, with faeces and feed waste settling through the water column and being subject to the hydrodynamics of the site. By modifying the BenOss model to account for the waste from fish farms, DEPOMOD was created to predict the solids accumulation in the area surrounding a fish farm (Cromey et al. 2002). Cromey et al. (1998) conducted a tracer study in order to validate the resuspension module. In addition to this, benthic data from five sites were compared to the carbon predictions in order to validate the benthic module (Cromey et al. 1998). Initial comparisons of the model predictions to field data collected in a case study by Cromey et al. (1998) showed general agreement. The model can be broken down into 4 main modules (Cromey et al. 1998):

- Grid generation module
- Particle tracking module
- Resuspension module
- Benthic module

This led to the development of DEPOMOD by Cromey et al. (2002) which had a similar structure to the BenOss model with the appropriate modifications. In terms of the use of DEPOMOD by the regulatory bodies, it requires input data that are easy to collect or provide for a typical fish farm, such as the flow of the water currents, the depth and shape of the seabed contours, the layout of the cages for the fish farm and the fish stocking density (Keeley et al. 2013). The flowchart in Figure 1.3, from the NewDEPOMOD user guide (https://depomod.sams.ac.uk/docs/UserGuide.pdf) illustrates the processes involved in the model. Firstly, DEPOMOD generates a regular grid pattern for the surrounding area of the fish farm cage in order to make predictions of the deposits on the seabed using the following inputs (Cromey et al. 2002):

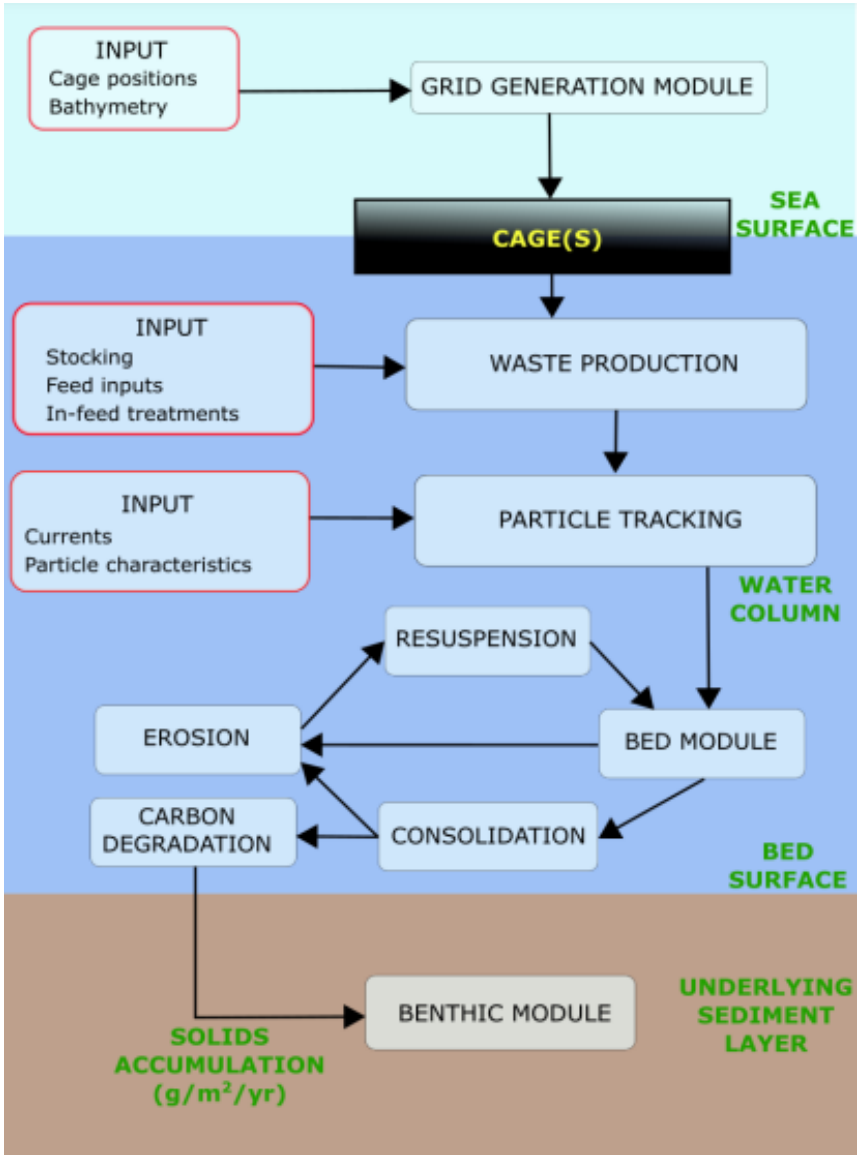


Figure 1.3: Flowchart illustrating the processes and modules within NewDEPOMOD (https://depomod.sams.ac.uk/docs/UserGuide.pdf).

- Cage positions
- Bathymetry
- Sampling station positions

Where the predicted waste depositions of the fish farm are expected to be less than 100m from the cage, a finer grid cell resolution is more appropriate (e.g. 10m) (Cromey et al. 2002). If the predicted waste depositions are expected to be larger, then it is more appropriate to use a larger grid cell resolution such as 25m (Cromey et al. 2002).

Following the generation of the grid, the particle tracking module explains how faecal and food waste particles will travel from the cage to the seabed

(Cromey et al. 2002). This model requires information as to the feed input of a farm, as well as the expected faecal and food waste (Cromey et al. 2002). The model then tracks particles as they travel through the water column. In the water column, particles are subject to movement representing the settling velocity, movement by the current, and a random walk in three dimensions, representing turbulence (Cromey et al. 2002). The water column is often divided into 3 layers, with each layer often having different current amplitude and direction. In order to obtain the current values, the current speed and direction is often measured over a suitable time period at a specific location near the farm, at 3 different depths. Regarding the random walk aspect, in the document relating to the BenOss model, the following random walk model is used to calculate the size of the random walk step used to represent turbulence (Cromey et al. 2002):

$$rw_{Step} = rw_{dir}\sqrt{(2k\delta t)}.$$

The elements of the model are described below (Cromey et al. 2002):

- rw_{Step} - size of the step
- rw_{dir} - direction of the step - this is given by a random number generator, picking either 1 or -1.
- k - dispersion coefficients (k_x, k_y, k_z).
- δt - the time the particle is in the turbulent field.

In DEPOMOD, the x and y values of the dispersion coefficient represent the East-West and North-South directions, and almost always, these are considered to be the same value. In the model, there is the option to have different values of dispersion coefficients for particles that are in the initial settling phase from the cage, and particles that have been resuspended, however, these are often kept the same. In addition, the z values of the dispersion coefficient represent the turbulence in the vertical direction, which is considered to be reduced in comparison to the x and y values. The time step within DEPOMOD can be altered, but for the purposes of this thesis, it remained at the default value used by SEPA, of $60s$.

The aim of the resuspension module is to estimate the amount of particles that will be accumulating in the grid area $(g/m^2/year)$ (Cromey et al. 1998). The resuspension model is separated into erosion, transport, deposition and consolidation components (Cromey et al. 1998). Firstly, an erosion event takes place when the shear stress near the seabed exceeds the critical shear stress for erosion (Cromey et al. 2002). After some development of the initial formula

by Cromey et al. (1998) for the rate of erosion, the formula that is used in the most recent version of DEPOMOD is:

$$M_e = M\left(\tau_{bot} - \tau_{crit}\right) \qquad \text{if } \tau_{bot} > \tau_{crit} \tag{1.1}$$

$$M_e = 0 \qquad \text{if } \tau_{bot} \leq \tau_{crit}, \tag{1.2}$$

where the elements of the model are:

- M_e - Mass of particles eroded
- M - Rate of erosion
- τ_{bot} - Shear stress at the seabed
- τ_{crit} - Critical Shear Stress for erosion

The mass of particles eroded is proportionate to the amount that the shear stress is above the critical value (Cromey et al. 1998). Both the rate of erosion (M) and critical shear stress for erosion (τ_{crit}) are physical properties that can be changed in DEPOMOD. On the other hand, if the magnitude of the shear stress falls below a critical threshold for deposition, then a deposition event will take place. In the case where a particle is resuspended, the particle will be lifted to a certain height above the seabed, transported at the surrounding current speed until it is deposited on the seabed, and after a given time period on the seabed consolidation of bed particles will occur (Cromey et al. 2002).

A Benthic response model is then used to predict the impacts on the seabed. DEPOMOD provides a prediction of the total solids flux (measured in $g/m^2/year$) being deposited on the seabed. When taking samples from the seabed, the effect on the seabed is measured as Infaunal Trophic Index (ITI) and the total abundance based on particular levels of solids accumulation (Cromey et al. 2002). In order to compare the predictions from DEPOMOD to samples taken from the seabed, a conversion between solids flux to ITI was required. This conversion was based on real data collected from the seabed and compared to model predictions of solids flux in an experiment which was carried out by SEPA. SEPA have identified this conversion as a possible area of uncertainty when comparing seabed samples to model predictions. However, the main focus of this research is understanding more about the model and the predictions rather than looking into the conversion between model predictions of solids flux to ITI.

Further research into modelling the waste transportation at fish farms has allowed DEPOMOD to be developed over the years to create NewDEPOMOD

which is more user friendly. As previously mentioned, NewDEPOMOD has the same key aspects as DEPOMOD which is described above. This thesis will aim to investigate NewDEPOMOD in detail through a sensitivity analysis and develop approaches to approximate the NewDEPOMOD output through the use of statistical modelling.

1.2.1 Input & Output Data for NewDEPOMOD

For a given site, there are multiple input files that are required to complete a NewDEPOMOD run. Firstly, there is a bathymetry input file that contains information on the location of the domain, and the depth of the seabed (or the height of land) within the domain. Following the introduction of the new regulatory framework, the updated guidelines indicated that modelling should be completed with flat bathymetry rather than variable as it produces more accurate results. Within the model, particles are deposited on the seabed as soon as they come into contact with the seabed. Using a variable bathymetry meant that particles being transported horizontally near the seabed would be deposited as soon as they came into contact with a shallower section of the seabed. This is not in line with what would be expected in reality, where the flow of the water over the shallow area would likely mean that waste particles are not deposited at a shallower section. SEPA therefore altered the guidelines as flat bathymetry was able to represent the transportation of particles near the seabed more effectively. As the guidelines came into place after the start of this project, some of the initial work is completed using a variable bathymetry input file. There are a number of different input files within NewDEPOMOD,

- **Bathymetry** - this contains the domain location data and the corresponding bathymetry data within the domain.
- **Cages** - this contains the cage location and dimension data.
- **Flowmetry** - this contains the flow data for the site, with the location and depths at which it was collected over a specific time period.
- **Inputs** - this contains data relating to the Biomass, feeding rates and the composition properties of the waste.
- **Models** - this contains the model run properties such as the duration of the run, as well as the values of the inputs based on the physical properties. This file contains more information, which will be described below.

When a run is created, a model file is produced which contains four input files that can be altered. The first of these is a configuration properties file that includes basic information about the run such as the feeding rate for the fish. The second is the model properties file that contains information relating to the time period for each run. The third is the physical properties input file that contains all of the values for the physical properties that can be changed within the model - there are over 200 elements that can be changed. Lastly, the fourth model file is the runtime properties file which links all of the input files, described above, to the run. All of these input files are required in order to complete a run for a site.

Once a run has been completed, a results file is compiled which contains information on the amount of waste particles located within each grid cell - known as Solids Flux which is measured in *grams per metre squared per year*, ('$g/m^2/y$'). An example of an output map produced by NewDEPOMOD is given in Figure 1.4. Figure 1.4 shows the output map that can be created

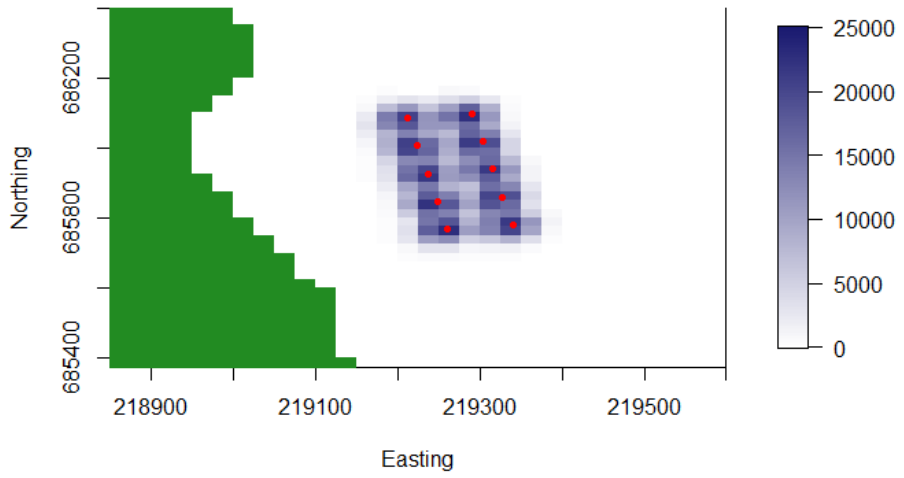


Figure 1.4: Example of an output map produced by NewDEPOMOD - land represented by green grid cells, and the cages are represented by the red points.

using the information in the results file. It shows that the deposition is highly variable, with an intense impact directly below the cages and much lower impact elsewhere. A number of other summaries that capture the impact of the farm can be calculated. Within this project, the main summary statistics that will

be of interest are the Total Area Impacted, 99th Percentile of Solids Flux, and Mass Balance.

The Total Area Impacted is of interest as it determines the overall size of the impact, in some instances it is appropriate to consider the Total Area Impacted for Solids Flux greater than a specific value.

The 99th Percentile of Solids Flux is a measure of the intensity of the impact, which is of importance in determining areas that may be subject to irreparable damage if this value is especially large. This is calculated as the 99th Percentile across the grid cells within the domain, where the grid cells without deposition are not included in the calculation. It allows for a measure of the intensity to be considered, which is not affected by potential outliers in the Solids Flux output for each grid cell.

Finally, Mass Balance is the proportion of waste particles left in the domain in comparison to the total waste particles that left the cages. It should be noted that if waste particles are transported outside of the domain within the model run timeframe, they are no longer included in a model run. The domain size for modelling is set at the beginning, and SEPA advised that this is set to a maximum of 4km^2, as NewDEPOMOD does not perform as well for larger domains. It is calculated as follows:

$$\text{Mass Balance} = \frac{\text{Total Mass in the domain}}{\text{Total Mass Released from the Cages}}. \tag{1.3}$$

From Equation 1.3, we can see that this value will lie in the interval $[0, 1]$, as the mass left in the domain cannot be larger than the total mass that was released from the cages. This value will be affected by all of the input factors being tested, as they determine how long particles remain in the water column and therefore how far they can be transported, which may result in large values of mass leaving the domain. In addition to these scalar summaries, it is of importance to consider the shape of the impact, which will be done using the information from the results file, which can be used to produce a map of the waste deposition within the domain.

1.3 Research aims

Process-based modelling is an effective tool that is used to assess environmental challenges where collecting data is not practical or cost-effective, as well as for situations where models are being used for simulating the future. As with any method for assessing environmental challenges, there are advantages and dis-

advantages. While process-based modelling is more practical and cost-effective for some challenges, there are some drawbacks - they can be computationally intensive and they do not always account for uncertainty. NewDEPOMOD is an example of a complex process-based model and will be the focus of this research, with the main aim of the research being:

- To investigate and quantify uncertainty within NewDEPOMOD inputs to assess their impact on NewDEPOMOD predictions. In addition, the aim is to develop a statistical emulator of NewDEPOMOD to overcome the computational challenge of running NewDEPOMOD when testing different model setups.

Both univariate and multivariate outputs can be produced by NewDEPOMOD, so this research will focus on both of these outputs as an application for completing the following statistical objectives:

- Investigate sensitivity analyses methods for univariate model outputs.
- Expand on the traditional sensitivity analysis techniques for univariate model outputs to develop novel methods for considering multivariate model outputs.
- Develop a novel statistical emulation framework for the environmental impacts of fish farms to approximate the univariate outputs without the computational cost.
- Expand the emulation framework to multivariate output emulation methods for correlated outputs.

Uncertainty quantification and sensitivity analyses are common approaches for assessing uncertainty in process-based models to quantify uncertainty in model outputs and attribute them to variations in the model inputs. With advancements in modelling techniques and software, it is now possible to extend traditional methods for univariate outputs to models with multivariate outputs. Therefore sensitivity analyses of process-based models with multivariate outputs is an area of interest. This research applies the methods to NewDEPOMOD, however the novel approach can be applied to any model with multivariate output.

Computational time is a common problem in process-based modelling, and statistical emulation is a novel approach to modelling the environmental impacts of aquaculture which can approximate the impact much more efficiently. The univariate outputs can be considered using common emulation approaches,

with additional challenges presenting when extending univariate output for correlated multivariate outputs. The univariate output emulation approaches are extended by considering the univariate outputs as a correlated multivariate output to investigate the possibility of information gain by introducing a correlation structure between the outputs.

1.4 Sensitivity Analysis Background

Many models in environmental science, like NewDEPOMOD, aim to replicate systems within our environment. In the creation of these models, assumptions are made, and parameter values estimated due to the complexity of the systems or the lack of resources to complete physical experiments (Saltelli et al. 2000). Sensitivity analyses are often used to investigate uncertainties within models and increase confidence in the model predictions by improving the understanding of how the model output reacts to changes in the inputs (Saltelli et al. 2000). A brief introduction to sensitivity analyses will be given here, with further details provided in Chapters 2 and 3.

Saltelli et al. (2004) defined a sensitivity analysis as, "The study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input." Sensitivity analyses are a useful exercise to confirm the consistency of the model outputs and the models robustness to uncertain model inputs (Pianosi et al. 2016). Often when completing a sensitivity analysis, an uncertainty analysis is also executed at the same time (Saltelli et al. 2008). An uncertainty analysis is a closely related topic which focuses more on quantifying uncertainty in the model output (Saltelli et al. 2008). A sensitivity analysis can provide the following information in modelling (Saltelli et al. 2000, 2008):

- determine if a model is similar to the system that is being studied,
- pinpoint critical areas in the input space,
- identify areas to focus more research,
- reduce complexity of models by identifying areas that can be simplified,
- establish possible errors in the model.

There are multiple ways in which the above areas can be investigated.

1.4.1 Sensitivity Analysis Workflow

In each case, a sensitivity analysis will follow a distinct pattern. The sensitivity analysis will rely on the model being executed multiple times for different combinations of sample values for the input factors, with the following steps being executed (Saltelli et al. 2000):

1. Determine what questions relating to the model should be answered and identify the input factors that should be involved in the analysis.
2. Establish suitable ranges of variation for each input factor and identify the relevant probability density functions.
3. Identify an appropriate design to generate the required input matrix.
4. Complete model evaluations to create the required outputs for analysis.
5. Analyse the effect of each input factor on the output variable.

Most sensitivity analyses will follow these basic steps, with different options being available depending on the aims of the analysis. Within the first step, the decisions that have to be made relate to the input factors, and whether all factors, or only a proportion of these are required to answer the relevant questions about the model (Saltelli et al. 2000). Moreover, the output variable(s) will be determined at this point. Following this, the next step requires decisions to be made about the possible ranges for each input factor within the analysis. Choices regarding the ranges can be made based on the literature relating to the input factors (where available), or expert knowledge in the area. When suitable ranges have been produced, the sampling method then has to be considered. The main choice here revolves around whether the input factors should be varied one-at-a-time, or all at once. Using the input matrix created by the relevant sampling method, the model is evaluated to create the outputs that can then be analysed to identify which input factors are causing any variations in the output.

1.4.2 Types of Sensitivity Analysis

In order to achieve the different goals of a sensitivity analysis, different methods have to be used. There are 3 main purposes/settings that are involved in determining the goal of the sensitivity analysis (Pianosi et al. 2016):

- **Ranking** - This is the process of ranking the input factors by their influence on the variability of the output.

- **Screening** - This aims to identify the input factors that have negligible contribution to the variability of the output.
- **Mapping** - The goal of mapping is to identify areas of the input space that produce extreme output values.

The choice of the appropriate sensitivity analysis method is guided by the purpose. There are other purposes that have been proposed, but the main ones are provided above and are the most common (Pianosi et al. 2016).

1.4.2.1 Local Sensitivity Analysis

The local approach is the first known application of sensitivity analysis, where the model output is assessed based on small perturbations to the model inputs. It is feasible when the input factors have a relatively small variation around their midpoint, with the relationship between input and output often assumed to be linear (Saltelli et al. 2000). By keeping the range of variation for the input factors equal for all input factors ($\pm 5\%$), it allows the effects of the input factors to be calculated. In essence, local sensitivity analysis considers varying one input factor, while the others remain constant, and inspecting the effect on the model output (Gan et al. 2014). The main shortfalls of the local approach are that they do not cope well with input factors with different levels of uncertainty, and when interactions exist between input factors (Saltelli et al. 2000). In these more complex situations, it is appropriate to use the global sensitivity analysis approach.

1.4.2.2 Global Sensitivity Analysis

Saltelli et al. (2000) described a sensitivity analysis as being global when (i) all the input factors are varied at the same time and (ii) the sensitivity of each input factor is measured over its total range, for bounded input factors. Global methods overcome the limitations of local methods by varying input factors simultaneously (Gan et al. 2014). One drawback of global methods in the past has been the computational cost of implementing it, but using a Design of Experiments approach, sampling techniques such as Latin Hypercube Sampling (McKay et al. 1979), Monte Carlo (Metropolis & Ulam 1949) and Orthogonal Array (Owen 1992) can be implemented to improve computational cost. Using appropriate sampling methods, computational costs can be reduced and the ability of global methods to deal with interactions between input factors make them the better option for completing sensitivity analyses of more complex models (Pianosi et al. 2016). With an appropriately chosen design, an

understanding of the influence of the input factors can be determined using appropriate sensitivity measures.

1.5 Emulation Background

Complex environmental systems are regularly imitated by computer models within scientific research. The complexity of these models (simulators) results in large computation times, which can cause difficulties when a large number of model runs are required for validation and calibration purposes. This has led to the development of statistical modelling techniques that allow highly efficient meta-models (emulators) to be built allowing approximations of the simulator to be made without the computational intensity (Conti & O'Hagan 2010). Developing a statistical emulator is therefore a fundamental step when looking to develop a greater understanding of a simulator (Overstall & Woods 2016). A brief introduction to the idea of statistical emulation is provided below, with further details given in Chapters 4 and 5.

The basic idea of an emulator is to create a statistical model to imitate a simulator, using a set of costly training runs generated by the simulator. Emulators are therefore an indirect approximation of the complex environmental systems. Emulators can be used in many ways and are useful when trying to gain a deeper understanding of a simulator model and how the inputs affect the output. If a suitable emulator is created, it could then be used instead of the computationally expensive simulator. Given a vector of p input variables, $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ in the p-dimensional input space $\mathcal{X}$, let the simulator be described by the black-box function, $f : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}^k$, where $\mathcal{Y}$ is in the k-dimensional output space. Essentially, at the given input combination, $\mathbf{x}$,

$$\mathbf{Y} = f(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})).$$

In this instance, $\mathbf{Y}$ is the $k \times 1$ output vector for $\mathbf{x}$. For a given input combination $\mathbf{x}_0$ which has not been evaluated by the simulator, an emulator is a prediction equation for $f(\cdot)$ that provides a substitute for $f(\mathbf{x}_0)$. An emulator is therefore an approximation of $f(\cdot)$, where

$$\hat{\mathbf{Y}} = \hat{f}(\mathbf{x}_0) = (\hat{f}_1(\mathbf{x}_0), \ldots, \hat{f}_k(\mathbf{x}_0)).$$

Different statistical methods can be used to determine $\hat{f}(\cdot)$, such as linear regression, generalized linear models, regression splines, and Gaussian processes (Grow & Hilton 2018), with further details to follow. Linear regression is of-

ten used for emulation purposes due to the simplicity and the obvious small computational cost of running these models (Kleijnen 1979, Madu 1990, Jalal et al. 2013, Grow 2016). With advances in computational power, simulators have increased in complexity, and simultaneously, emulation techniques have changed, with Gaussian process emulation becoming more popular. Gaussian processes are a more flexible approach and are able to capture non-linear patterns (Kennedy et al. 2006, Conti et al. 2009, Rajabi & Ketabchi 2017, Noè et al. 2019). Simulators do not always produce a singular, scalar output and different techniques have had to be developed to deal with multivariate outputs from simulators (Conti & O'Hagan 2010, Overstall & Woods 2016, Alvarez & Lawrence 2009, 2011). Multivariate emulators can be produced in a Bayesian framework and also using a functional approach, which is often preferred for simulators that produce a time series output (Sacks et al. 1989, Bayarri et al. 2005, 2007, Liu et al. 2009). Hence, there are a number of options available when creating an emulator, with different properties of the simulator determining the method that would be appropriate.

1.5.1 Linear Regression Emulation

Linear regression models can be used for emulation, and are often referred to as regression metamodels (Grow & Hilton 2018). Consider the output space $\mathcal{Y} \subset \mathbb{R}$, then a typical first-order polynomial can be used to approximate the simulation model:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon. \tag{1.4}$$

Here, β_0 is the intercept and β_i represent the estimated linear effects of each input, x_i, on the average value of y. This model however, does not account for any interactions between the inputs and only allows for linear relationships between the inputs and the output. Equation 1.4 can be modified to include interactions between inputs and curvature in the relationships between inputs and the output.

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \beta_{ij} x_i x_j + \sum_{i=1}^{p} \beta_{ii} x_i^2 + \epsilon. \tag{1.5}$$

In Equation 1.5, interactions between inputs x_i and x_j are included, with the estimated effect given by β_{ij}. To represent curvature in the relationship between x_i and the output, β_{ii} estimates the quadratic effect. Higher order polynomials do not tend to be used as they can be difficult to interpret and can have

problems with robustness. Ordinary Least Squares ('OLS') is often used to estimate the parameters of the regression metamodels, meaning the standard assumptions for the error terms, ϵ, must be satisfied:

1. The errors should be normally distributed.
2. The errors should have a mean value approximately equal to zero.
3. The variance of the errors should be homogenous.
4. The errors are independent.

The above assumptions can be checked in the standard way by inspecting plots of the residuals. As expected, all of the above assumptions must be met before any conclusions can be made.

As a whole, Equations 1.4 & 1.5 demonstrate that the regression metamodels can be reasonably flexible in relation to the choices of parameters, and it is therefore appropriate to fit initial models based on previous knowledge of the inputs and the expected behaviours in the real life scenario, and then assess their performance. The fit of the regression metamodel is obviously crucial, and therefore how well its predictions compare to the observed simulations is a key indicator of how well a metamodel performs (Grow 2016). A formal lack-of-fit ('LOF') test can be used in this instance to assess the performance of the regression metamodel (Grow 2016, Grow & Hilton 2018). The LOF partitions the total error (ϵ_E) into the pure error (ϵ_{PE}) and the error due to the lack of fit of the regression model (ϵ_{LOF}):

$$\epsilon_E = \epsilon_{PE} + \epsilon_{LOF}.$$

The LOF tests the null hypothesis that $\epsilon_{LOF} = 0$, meaning that $\epsilon_E = \epsilon_{PE}$. The alternative hypothesis that $\epsilon_{LOF} \neq 0$, meaning that there is an error due to the lack of fit of the model. In order to calculate ϵ_{PE} and ϵ_{LOF}, first consider $a = 1, \ldots, p$ different combinations of model input values, where $b = 1, \ldots, q_a$ simulations have been conducted for each a. Let $Q = \sum_{a=1}^{p} q_a$ be the total number of simulations that are completed, so ϵ_{PE} can be calculated as follows:

$$\epsilon_{PE} = \sum_{a=1}^{p} \sum_{b=1}^{q_a} (y_{ab} - \bar{y}_a)^2.$$

Here, $\bar{y}_a$ refers to the average output observed across the q_a simulations for the combination of model inputs, a. This calculation is suitable for Stochastic

models, as a deterministic model would result in $\epsilon_{PE} = 0$. Next, ϵ_{LOF} can be calculated as follows:

$$\epsilon_{LOF} = \sum_{a=1}^{p} q_a(\bar{y}_a - \hat{y}_a)^2.$$

Hence, ϵ_{LOF} considers the difference between $\bar{y}_a$ and the prediction made by the regression metamodel, $\hat{y}_a$. Using the above calculations, the test statistic for LOF is as follows (Grow & Hilton 2018):

$$F_{LOF} = \frac{\epsilon_{LOF}/(p-r)}{\epsilon_{PE}/(Q-p)}.$$

In this instance, r refers to the number of parameters within the regression metamodel. F_{LOF} has an F-distribution with $(p-r)$ degrees of freedom for ϵ_{LOF}, and $(Q-p)$ degrees of freedom for ϵ_{PE}. If F_{LOF} is considered to be significant when comparing to the F-distribution, $F\left((p-r),(Q-p)\right)$, then the null hypothesis that there is a lack of fit cannot be rejected (Grow 2016). If this is the case, the metamodel would therefore have to be adjusted to try and improve its performance.

The standard linear regression approach can be extended to deal with more complex relationships between the inputs and the simulator output. Generalized linear models ('GLMs') would be a simple extension to standard linear regression, where the assumption that the errors are normally distributed is not required (McCullagh & Nelder 1989). An extension to GLMs is Generalized Additive Models ('GAMs'), which are a more flexible approach where the relationship between the inputs and the output can be described by a smooth function. Given our output y and inputs x_i, as seen in Equation 1.4, an additive model can be written as:

$$g(\mathrm{E}(y)) = \beta_0 + \sum_{i=1}^{n} f_i(x_i), \tag{1.6}$$

where f_i represent smooth functions of the output against the input (Wood 2017). In order for a GAM to be appropriate, plots of the residuals should be checked, as with a standard linear regression model. A GAM can be built based on a training set of simulator runs, which can then allow predictions to be made for a test set of data and the related errors for each prediction. Within emulation research, the most common approach is using Gaussian processes (Conti et al. 2009, Bastos & O'Hagan 2009, Rajabi & Ketabchi 2017, Parker et al. 2019).

1.5.2 Gaussian Process Emulation

When an LOF test identifies that a regression metamodel is not appropriate, it may be that a more flexible, non-parametric modelling method may be required for an emulator. Gaussian processes are widely used in computer experiments and emulation due to their flexible nature. They are used in many different branches of statistics for emulating simulators when investigating complex simulators (Conti et al. 2009, Conti & O'Hagan 2010, Rajabi & Ketabchi 2017, Parker et al. 2019). Over the years, different variations of Gaussian processes have been developed, such as sparse Gaussian processes (Titsias 2009), K-nearest neighbour local Gaussian process approach (Gramacy & Apley 2015), low-rank Gaussian processes (Wood 2017), sparse convolved Gaussian processes for multi-output regression (Alvarez & Lawrence 2009, 2011).

Gaussian processes are used as emulators due to their ability to model smooth relationships between simulator inputs and outputs, while also being able to make predictions for new inputs and the associated uncertainty (Grow & Hilton 2018). Gaussian process emulators rely on the training data to model the simulator, so the more design points the smaller the uncertainty (O'Hagan 2010). Rasmussen & Williams (2006) defined a Gaussian process as:

> '... a collection of random variables, any finite number of which have a joint Gaussian distribution',

and it is defined by its mean function and covariance function (Rasmussen & Williams 2006). In comparison to a GAM (Equation 1.6), the smooth function f, essentially has prior information - the mean and covariance functions. These functions can be defined as follows, for a process, $f(\mathbf{x})$:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \end{aligned}$$

A Gaussian process can then be written as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')),$$

indicating that the function f is distributed as a Gaussian process with a mean function m, and a covariance function k (Rasmussen & Williams 2006). As Gaussian processes are defined as collections of random variables, it therefore has a consistency requirement (also known as a marginalization property, Rasmussen & Williams (2006)). Essentially, for $(a, b) \sim \mathcal{N}(\mu, \Sigma)$, then $a \sim \mathcal{N}(\mu_1, \Sigma_1)$, where Σ_1 is the relevant submatrix of Σ. Rasmussen & Williams

(2006) introduced a simple example of a Gaussian process which is summarised below:

> A Bayesian Linear Regression model given by $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$ where $\phi(\mathbf{x})$ is a set of basis functions, and $\mathbf{w}$ is a vector of weights with prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$, can be described as a Gaussian process (Rasmussen & Williams 2006). The mean and covariance functions can be described as follows:
>
> $$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] = \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = 0 \tag{1.7}$$
> $$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[f(\mathbf{x}) f(\mathbf{x}')] = \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \Sigma_p \boldsymbol{\phi}(\mathbf{x}'). \tag{1.8}$$
>
> From Equation 1.7 and Equation 1.8, it can be concluded that $f(\mathbf{x})$ and $f(\mathbf{x'})$ are jointly Gaussian with mean equal to zero and covariance given by $\boldsymbol{\phi}(\mathbf{x})^\top \Sigma_p \boldsymbol{\phi}(\mathbf{x}')$. The mean function of a Gaussian process, Equation 1.7, is often set either to zero, or takes a parametric form (Grow & Hilton 2018). The covariance function, Equation 1.8, plays a pivotal role in the production of a Gaussian process predictor and will be considered in more detail later.

Gaussian processes are an efficient and accurate tool for emulating a complex simulator, and will therefore be considered for emulating NewDEPOMOD output in this thesis. More detail and background information will be provided later in the thesis prior to fitting the Gaussian process regression models.

1.5.3 Multivariate Emulation

As emulation is expanded to account for multivariate response variables, the amount of data to be considered in the building of the emulator becomes a challenge (Rougier 2008). One approach for emulating multiple outputs is to consider building an independent univariate emulator for each of the outputs. The drawback to this approach is that it does not account for the fact that there may be relationships between the output variables and therefore information may be lost by modelling them independently (Fricker et al. 2013). Two different classes of emulator were described by Fricker et al. (2013):

- **Field output** - This refers to output 'that simulates a quantity over a continuous field, often space or time.'
- **Multiple-type output** - This refers to 'simulators that simulate different types of quantities jointly.'

For field output, each output refers to the value of a quantity at a specific location within the field. By considering the output index as a new input for the simulator, Kennedy & O'Hagan (2001) emulated field output using a univariate emulator with a stationary parametric covariance function. This method was replicated by Conti & O'Hagan (2010) and McFarland et al. (2008). Rougier (2008) developed a method that did not require the output index to be considered as a new input, by combining a parametric regression model with a correlation structure on the output index, to emulate the field output directly. The index for multi-type outputs is just a label and the outputs have different units and therefore a distance measure between the outputs cannot be calculated. As a result, the outputs are often emulated independently, or a separable covariance structure is used to emulate these jointly (Fricker et al. 2013). Conti & O'Hagan (2010) used single output Gaussian processes in conjunction with dimension reduction techniques to consider multiple outputs. In contrast, correlated outputs can also be considered in a Gaussian process framework (Alvarez & Lawrence 2009, 2011, Roberts et al. 2013). Correlated outputs will be considered in further detail in Chapter 5, when considering multiple outputs from NewDEPOMOD.

1.6 Structure of the thesis

The main aim of the thesis is to develop a deeper understanding of NewDEPOMOD through sensitivity analyses and use the information to develop emulators that allow approximations of NewDEPOMOD predictions to be made without the computational cost.

Chapter 2 will focus on presenting a sensitivity analysis for each of the univariate outputs at both a high energy site and a low energy site to compare and contrast. Design of experiments techniques are considered in order to account for correlations within the input structure, before using sensitivity analysis approaches to rank the inputs.

This will then be developed to complete a sensitivity analysis for a multivariate output in Chapter 3 to assess the impact of altering the inputs on the output maps. Part of this chapter considers how the NewDEPOMOD output maps are represented. A shape analysis, where the output map is summarised to pick out the main shape of the deposition on the seabed is one of the approaches considered. An alternative approach then considers the whole output map as a functional output - represented by a smooth surface. A further approach considered the output from individual grid cells.

These sensitivity analyses will provide the necessary information to develop a univariate and multivariate framework for statistical emulation of NewDEPOMOD predictions. Chapter 4 will initially consider the emulation of the univariate outputs, using familiar techniques such as Gaussian processes.

The next stage of the statistical emulation process will consider the multivariate outputs in Chapter 5. The multivariate outputs that are to be considered are the functional representation of the output maps, which is developed using a functional principal components analysis, as well as considering the univariate outputs together as a correlated multivariate output. Multi-output Gaussian processes are considered in this Chapter, allowing for independent and correlated outputs within the models.

Chapter 6 will then summarise the research presented in the thesis and discuss the achievements, limitations and scope for future work.

Chapter 2

Sensitivity Analysis for Scalar Outputs

2.1 Introduction

As previously mentioned in Chapter 1, NewDEPOMOD is a complex process-based model with a number of inputs. Running NewDEPOMOD produces an output map specifying the amount of waste deposition across the domain of interest. The output data contains estimations of the waste deposition (Solids Flux) within each grid cell of the domain - with an example output map given in Figure 2.1. The output maps produced can be summarised to provide relevant

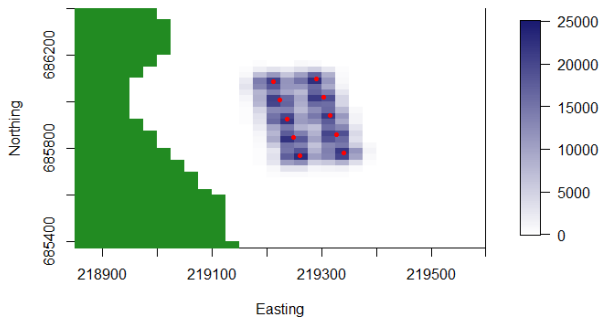


Figure 2.1: Example of a NewDEPOMOD output map showing the Solids Flux across the domain, with land specified by the green grid cells, and the cages given by the red points.

scalar outputs that are of interest to SEPA. Three summaries of the output maps that will be considered in this thesis are:

- Total Area Impacted - this is a measure of the size of the impact across the domain.
- 99th Percentile of Solids Flux - this provides a measure of the intensity of the deposition on the seabed.
- Mass Balance - this is the proportion of waste material that remains in the domain at the completion of the run.

Total Area Impacted and 99th Percentile of Solids Flux are important summaries of the output maps as these provide SEPA with an idea of the potential size and scale of the impact on the seabed. These are essential in the process of determining limits on licenses for new farms or the expansion of current farms. Mass Balance is a further important summary to be considered. Within NewDEPOMOD, when the current at a site transports waste material outwith the boundary of the domain being considered, the material is no longer a part of the simulation. Mass Balance is then a measure of how much of the original waste material remains in the domain. As Mass Balance is a proportion, it is therefore restricted to the interval $[0, 1]$.

There are a number of different inputs involved in NewDEPOMOD, which allows for different analyses to be considered. The inputs can be classified as two groups - 1) inputs based on the physical properties and 2) operational inputs. The inputs based on the physical properties relate to the physical process within the model such as the transportation, deposition and resuspension of waste material. The operational inputs refer to the farm properties that are controlled by the operator of the farm. Three different analyses are to be considered in this Chapter:

- Sensitivity analysis of the inputs based on the physical properties,
- Sensitivity analysis of the operational inputs,
- Combined sensitivity analysis of the physical properties and operational inputs.

For the sensitivity analysis of the inputs based on the physical properties, this will focus on a number of the inputs where their default values are considered to be uncertain. Multiple studies for some of the inputs, such as the Critical Shear Stress for Erosion, have discovered a variety of potential values which can be dependent on the location where the studies took place. These studies and inputs are discussed further in this Chapter. The aim of this Chapter will be to identify which of the inputs have the biggest impact on the scalar outputs

when they are altered. These inputs will be referred to as 'uncertain inputs' throughout the thesis. The sensitivity analysis of the operational inputs is considered as a way to assess the impact of increasing the capacity of a farm and how it may impact the scalar outputs. Finally, both groups of inputs will be considered together to assess the impact of altering both inputs at the same time.

The first two analyses of the input groups separately will be considered for two sites in detail to investigate the influence on the scalar outputs of changing the inputs. The sites of interest will feature a low energy site, where the current speeds are low, and a high energy site, where the current speeds are high. The sites being considered are Ardessie and Muck, with only two considered for the initial analysis. Finally, for the combined analysis of the two groups of inputs, a total of 4 sites will be considered. In particular, two low energy sites and two high energy sites are considered. Ardessie is removed as a site of interest as it has lower production, and replaced with two sites with larger production to allow for better comparisons between the large production high energy sites that SEPA identified as being the most effective for reducing the environmental impacts. This was done to investigate and compare how the environmental impacts are affected by changing the input values for the different type of sites.

2.2 Sensitivity Analysis - Inputs Based on the Physical Properties

A sensitivity analysis of the inputs based on the physical properties has been developed to assess the impact of their parameter uncertainty on NewDEPOMOD predictions. This sensitivity analysis will follow the workflow described by (Saltelli et al. 2000). Due to the complex nature of NewDEPOMOD, there exists interactions between some of the inputs, which will be considered when preparing and completing the sensitivity analysis. The sensitivity analysis has been carried out for two different sites, with different characteristics such as current speed and depth, in order to determine if the sensitivity ranking of the inputs is different.

2.2.1 Aims of the analysis and the inputs to be investigated

In collaboration with SEPA, a subset of the inputs based on the physical properties were identified for the purpose of conducting an initial sensitivity analysis. These inputs were considered to be of importance based on SEPA's previous modelling experience and the uncertainty surrounding the default parameter values for these inputs. The inputs that were chosen can be seen below:

- **Critical Shear Stress** - Threshold value for which an erosion event takes place when the shear stress on the seabed exceeds this value.
- **Rate of Erosion** - This input determines how much material is eroded when an erosion event takes place.
- **Release Height** - When an erosion event takes place and particles are resuspended into the water column, this represents the height at which the particle is lifted.
- **Settling Velocity of Faeces** - The Settling Velocity determine how long a faecal particle will stay in the water column.
- **Settling Velocity of Resuspended Material** - This Settling Velocity is typically smaller as the particles tend to be smaller in size.
- **Dispersion Coefficients of Material from the cages** - The Dispersion Coefficients are used to calculate the size of the step for the random walk element of DEPOMOD which represents turbulence. These coefficients represent turbulence in 3-dimensions for material settling from the cages.
- **Dispersion Coefficients of Resuspended Material** - These coefficients represent turbulence in 3-dimensions for material that has been resuspended from the seabed.

The aim of this sensitivity analysis is to identify which of the above inputs with uncertainty in their parameter values had the biggest influence on the scalar outputs. This therefore related to a sensitivity analysis ranking problem.

2.2.2 Establishing suitable ranges for inputs

In order to look at the sensitivity of some of the inputs of NewDEPOMOD, a suitable range of values has to be considered for them. In choosing some of

these values, they have to be considered on a site by site basis. For instance, when considering a suitable range of values for Critical Shear Stress for Erosion, (τ_{crit}), this is dependent on the flow speed for that site. We will only be able to gain useful insights if we consider plausible values for τ_{crit} that are based on the flow speeds observed at that site. For example, choosing τ_{crit} based on a flow speed that is greater than the maximum flow speed for that site, will not provide any more information than if we choose τ_{crit} based on this maximum flow speed, as no erosion would take place for either of these values. Similarly for the lower bound of τ_{crit}, this should be based on the minimum flow speed observed at that site. By choosing a suitable range of τ_{crit} based on the minimum and maximum flow speeds at each site, we will be able to see fully the effect of this input on the predictions of solids flux made by NewDEPOMOD.

When an erosion event takes place, the amount of material that is eroded is dependent on the exceedance of the critical shear stress, as well as the rate of erosion (M), seen in Equation 1.2. As a result, the level of erosion could be decreased in two ways:

1. Reducing the amount of material moved in each erosion rate - i.e. reducing M.

2. Reducing the number of erosion events that take place - i.e. increasing τ_{crit}.

Looking at the range of values to consider for M, we base this on the current default value of $0.031 kg/m^2/s$. Mitchener & Torfs (1996) found that M ranges from 2×10^{-4} to 6×10^{-4}, depending on the type of sediment on the seabed. The range of values found by Mitchener & Torfs (1996) were two orders of magnitude lower than the current default value used, and so this should be considered when choosing a suitable range of values to look at for M. For the upper bound, no literature found the rate of erosion rate to be any larger than the default value, and increase by one order of magnitude higher than the default value was considered a good starting point.

The next input to be considered is the release height of the material that has been eroded. When looking at this, there are two different implications, depending on the value that is used. First of all, it is linked to how long the particle is in the water column, as the particles have a specific value for the settling velocity, and the higher the particle is released, the longer it will be in the water column and therefore the further that it could travel. Additionally, resuspended particles will settle quicker if they have been eroded and the seabed shallows in the direction they are being transported. As a result, if the particles

have been released at a great enough height, they may be able to avoid certain shallower areas of the seabed and therefore travel a greater difference. Above the seabed there will be a cloud of particles, some that are settling, and some that have been resuspended. The default value that has been used for the release height is $0.12m$, which is considered to be the median of this cloud of particles. There is a lack of literature available relating to this value indicating that it is an area that has not been studied in detail. As a result, discussions with SEPA determined that it would be best to consider the minimum value as 0m, where the particles are not lifted from the seabed. As an initial maximum, a height of 1m was determined to be a good starting point.

The settling velocity of faeces is one input where there are varying opinions on appropriate values. Across a wide range of experiments completed by a number of researchers, different results have been seen. When initially creating the NewDEPOMOD model, Cromey et al. (2002) found that the settling velocity for both faecal waste, and resuspended material, was variable, but concluded that both of these could be modelled by a Gaussian distribution. In terms of the faecal waste, Cromey et al. (2002) centred the Gaussian distribution around a settling velocity of $0.032m/s$. Chen et al. (2003) found that the settling velocity of faecal waste ranged from $0.037 - 0.092m/s$, with the mean between $0.051 - 0.064m/s$. On the other hand, an experiment carried out by Bannister et al. (2016) found that more than half of the faecal waste, for salmon of different sizes, had settling velocities between $0.05 - 0.10m/s$. However, in this case, they found that the distribution of these settling velocities had a right, positive skew, and concluded that modelling the settling velocities of the faecal waste as a Gaussian distribution was not appropriate. Bannister et al. (2016) found that less than 10% of the faecal waste had settling velocities less than $0.01m/s$. Due to the varying results seen in different papers, it may be appropriate in this case to consider a range of values for the mean of a Gaussian distribution. For the settling velocity of the resuspended material, the default value for the centre of the Gaussian distribution is $0.0054m/s$. There also appears to be a lack of literature relating to this value, and so the range of values was determined by reducing and increasing the values by approximately an order of magnitude. For this initial sensitivity analysis, only values for the centres of the Gaussian distributions were considered and their variances and distributions can be considered in the future.

Moving on to the random walk aspect of the model which represents turbulence, the default values that are used for both sets of dispersion coefficients are $k = (0.1, 0.1, 0.001)$ (Gillibrand & Turrell 1997). This indicates that turbu-

lence has less of an effect vertically than it does horizontally. There appears to be no reason why there should be any difference between the sets of dispersion coefficients, and so the same ranges will be used. When completing model simulations, Bannister et al. (2016) used $k = (0.018, 0.018, 0.00058)$ to represent turbulence. Cromey et al. (2002) conducted a sensitivity analysis, and came to the conclusion that the model showed little sensitivity to the vertical steps of the random walk, as the steps were small in comparison to the settling velocities of the particles. Regarding the horizontal steps, Cromey et al. (2002) noted that the size of the steps in a given time period could exceed the distance a particle will be transported by the current. In terms of the bounds for these coefficients, it would appear appropriate to consider an upper bound of $0.5m/s$ for the horizontal coefficients (Cromey et al. 2002). Using this upper bound, the formula for the magnitude of the random walk step gives $\sqrt{2 \times 0.5 \times 60} = 7.7m$ for a time period of $60s$, which is larger than the distance that would be covered by a particle based on a current speed of $0.1m/s$ for $60s$ (Cromey et al. 2002). $0.1m/s$ is a current speed that is normally considered reasonably high, and in some sites the current speed rarely reaches this level, and so $0.5m/s$ is most likely an unrealistic value for horizontal dispersion coefficients, but it is likely to provide useful information with regards to the sensitivity analysis. Following discussion with SEPA, a maximum value of $1m/s$ would provide useful information. In terms of the vertical dispersion coefficient, the upper bound will increase the default value by an order of magnitude to $0.01m/s$. For the lower bounds, the effect of no random walk may provide useful information, and so having the coefficients close to zero will confirm how sensitive the model is to the random walk component.

Using all of the information above, the final parameter ranges can be seen below in Table 2.1.

For simplicity, a uniform distribution is considered for all the inputs in the analysis. There is no evidence that would suggest any alternative distributions should be considered, and so uniform distributions are considered throughout the thesis.

2.2.3 Sampling design

A key element of any sensitivity analysis is the sampling method that will be used. Depending on the aims of the sensitivity analysis and the computer power available, certain sampling methods are preferred to others.

Inputs	**Default Value**	**Lower Bound**	**Upper Bound**
Critical Shear Stress (τ_{crit})	0.02	Based on Min. flow at site	Based on Max flow at site
Rate of Erosion (M)	0.031	2×10^{-4}	0.310
Release Height	0.12	0.00	1.00
Settling Velocity (faeces)	0.032	0.005	0.100
Settling Velocity (sediment)	0.0054	0.0005	0.05
Material from cages - Dispersion coefficients (k_x, k_y, k_z)	(0.1, 0.1, 0.001)	(0, 0, 0)	(1.0, 1.0, 0.01)
Resuspended Material - Dispersion coefficients (k_x, k_y, k_z)	(0.1, 0.1, 0.001)	(0, 0, 0)	(1.0, 1.0, 0.01)

Table 2.1: Sensitivity Analysis - inputs based on the physical properties of interest and their ranges

2.2.3.1 Latin Hypercube Sampling

Latin Hypercube Sampling ('LHS') is a stratification method described by McKay et al. (1979), which can be used to create random samples from a sample space Ψ. LHS is an extension of the Latin Square method which dates back to 1624 according to Preece (1983'), allowing for samples to be taken from multiple dimensions.

The main aim of LHS is to capture as much of the sample space Ψ as possible (McKay et al. 1979). In order to extend the idea of the Latin Square, each parameter, $\boldsymbol{\beta}_i$ where $i = 1, \ldots, K$, in Ψ is given a distribution, with the parameters $\boldsymbol{\beta}_i$ being independent. Using the Cumulative Distribution Function ('CDF') of each $\boldsymbol{\beta}_i$, N strata of equal probability, $1/N$ are created, where N will be the number of sets of samples required. The strata are converted on to the parameter scale for $\boldsymbol{\beta}_i$, to create individual stratum, from which a random sample will be taken. As a result, a set of N samples have been created for each $\boldsymbol{\beta}_i$. The set of N values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are combined randomly, and without replacement, to produce a set of ordered pairs $[\beta_{1j}, \beta_{2j}]$, where $j = 1, \ldots, N$. The set of ordered pairs, $[\beta_{1j}, \beta_{2j}]$, are then combined at random with the set of values for $\boldsymbol{\beta}_3$ to produce a set of ordered triples, $[\beta_{1j}, \beta_{2j}, \beta_{3j}]$. This process is then repeated for the K parameters to produce an $N \times K$ Latin Hypercube.

Example 1. Consider the case where we have 3 parameters of interest, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$, and we are looking to create 100 sets of samples. Define the distributions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to be $\mathcal{U}(0,1)$, and the distribution of $\boldsymbol{\beta}_3$ to be $\mathcal{N}(0,1)$. Using the CDF, 100 strata with equal probability of 1/100 are created and can be seen in Figures 2.2 and 2.3. In Figures 2.2 and 2.3, 100 strata of equal probability

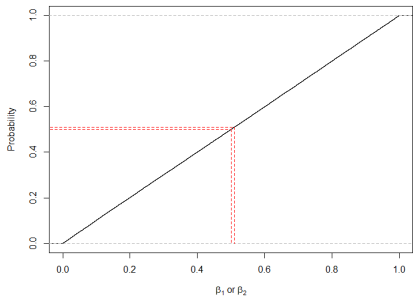


Figure 2.2: Plot showing a stratum that would be used for sampling of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

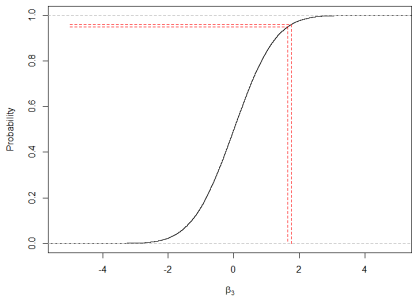


Figure 2.3: Plot showing a stratum that would be used for sampling of $\boldsymbol{\beta}_3$.

were created and then converted on to the parameter scale for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$, and 1 individual stratum can be seen in both plots. Within each stratum on the parameter scale, a sample is taken and these were randomly combined to produce 100 ordered triples $[\beta_{1j}, \beta_{2j}, \beta_{3j}]$, for $j = 1, \ldots, 100$. A property of LHS is that the parameters are required to be independent, which is a limitation of the method for sampling, however, it is one that can be overcome.

2.2.3.2 Correlated Latin Hypercube Sampling

In some cases, the parameters that are being sampled may not be independent and so the standard LHS would not be appropriate. Iman & Conover (1982) introduced a restricted pairing procedure in order to account for correlations between variables. Where correlations exist between parameters, the restricted pairing procedure creates a LHS with a rank correlation structure close to the rank correlation structure specified (Dandekar et al. 2001). The specified correlation structure is often based on previous literature relating to the parameters, or the knowledge and experience of the modeller (McKay et al. 1979).

The initial process of completing the restricted pairing procedure that was introduced by Iman & Conover (1982) is as follows:

1. Define a target correlation matrix, $\mathbf{C}^*$ (provided by the user).

2. Complete a Cholesky Decomposition of $\mathbf{C}^*$ to obtain a lower triangular

matrix $\mathbf{P}$ such that,

$$\mathbf{C}^* = \mathbf{P}\mathbf{P}^{\mathrm{T}}. \tag{2.1}$$

3. Let $\mathbf{L}$ be a LHS with k parameters and n samples, where each row contains a sample of each of the k parameters. Multiplying $\mathbf{L}$ by $\mathbf{P}^{\mathrm{T}}$, from Equation 2.1 gives a matrix $\mathbf{L}^*$, which should have a correlation matrix $\mathbf{M}$ such that,

$$\mathbf{M} \approx \mathbf{C}^*. \tag{2.2}$$

Iman & Conover (1982) noted concern in that the transformation matrix $\mathbf{P}$ was only dependent on $\mathbf{C}^*$. This meant that in certain applications, the correlation matrix from Equation 2.2, $\mathbf{M}$, calculated for the transformation $\mathbf{L}\mathbf{P}^{\mathrm{T}}$, may not be close enough to $\mathbf{C}^*$. Iman & Conover (1982) then proceeded to use a variance reduction technique that would allow the sample correlation matrix, $\mathbf{M}$, to be much closer to $\mathbf{C}^*$. The alterations of the above method are described below (Iman & Conover 1982):

- Define the sample correlation of the initial LHS as $\mathbf{T}$, and use the Cholesky Decomposition to find $\mathbf{Q}$ such that,

$$\mathbf{T} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}. \tag{2.3}$$

- Using Equation 2.3, a matrix $\mathbf{S}$ is then found such that,

$$\mathbf{C}^* = \mathbf{S}\mathbf{T}\mathbf{S}^{\mathrm{T}} \iff \mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{S}\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} \tag{2.4}$$

- The solutions for Equation 2.4 are then:

$$\mathbf{S}\mathbf{Q} = \mathbf{P} \iff \mathbf{S} = \mathbf{P}\mathbf{Q}^{-1}$$

- The improved restricted pairing procedure can then be used to calculate the correlated LHS using the following equation,

$$\mathbf{L}_{\mathrm{B}}^* = \mathbf{L}\mathbf{S}^{\mathrm{T}} \tag{2.5}$$

From Equation 2.5, $\mathbf{L}_{\mathrm{B}}^*$ should then have a sample correlation matrix, $\mathbf{M}_{\mathrm{B}}$, which is approximately equal to $\mathbf{C}^*$.

This restricted pairing procedure will therefore allow the LHS to be conducted for dependent variables. The following example will demonstrate how this works in practice, continuing on from Example 1.

Example 2. In order to demonstrate the correlated LHS, the same structure as in Example 1 will be used, with the same $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$. The following correlation matrix will be used for the restricted pairing procedure:

$$\mathbf{C}^* = \begin{pmatrix} 1.0 & 0.9 & 0.0 \\ 0.9 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}. \tag{2.6}$$

From Equation 2.6, a strong correlation of 0.9 between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ has been used to allow the restricted pairing procedure to be illustrated in the 3-dimensional plot of the LHS in Figure 2.4. The strong correlation between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ from

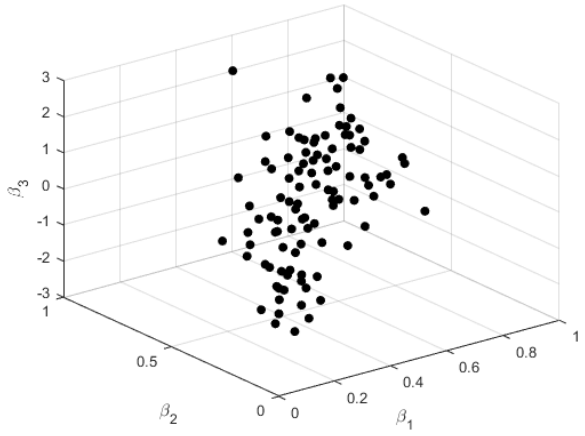


Figure 2.4: Plot of the Correlated Latin Hypercube Samples for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$.

Equation 2.6 can be seen in Figure 2.4. The actual correlation matrix that is calculated for this correlated LHS is:

$$\mathbf{M}_{\mathrm{B}} = \begin{pmatrix} 1.0000 & 0.8804 & 0.0201 \\ 0.8804 & 1.0000 & 0.0128 \\ 0.0201 & 0.0128 & 1.0000 \end{pmatrix}. \tag{2.7}$$

This correlation matrix from Equation 2.7 is close to the pre-defined correlation matrix in Equation 2.6, and if the number of samples was increased from 100, they would be even closer.

2.2.3.3 Correlated LHS for Sensitivity Analysis of NewDEPOMOD - inputs based on the physical properties

For the construction of the sampling design for NewDEPOMOD, correlations between the inputs have to be considered in order to make sure that the NewDEPOMOD runs are not producing implausible results. To do so, a target

correlation matrix, $\mathbf{C}^*$ had to be constructed. This was done with the guidance of SEPA. The order of the inputs within the target correlation matrix and throughout this section is as follows: 1) Critical Shear Stress for Erosion, 2) Rate of Erosion, 3) Release Height of Resuspended Material, 4) Settling Velocity of Faeces, 5) Settling Velocity of Sediment, 6-8) Dispersion Coefficient of Material from the Cages (X, Y, Z directions), 9-11) Dispersion Coefficient for Resuspended Material (X, Y, Z directions).

$$\mathbf{C}^* = \begin{pmatrix} 1.00 & -0.90 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.90 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & -0.90 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & -0.90 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.99 & 0.00 & 0.99 & 0.99 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99 & 1.00 & 0.00 & 0.99 & 0.99 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.99 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99 & 0.99 & 0.00 & 1.00 & 0.99 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99 & 0.99 & 0.00 & 0.99 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.99 & 0.00 & 0.00 & 1.00 \end{pmatrix}. \quad (2.8)$$

Firstly, the negative correlation between 'Critical Shear Stress for Erosion' and 'Rate of Erosion' was defined in order to keep a reasonable balance in the equation for calculating the amount of material eroded (Equation 1.2). 'Release Height' and 'Settling Velocity of Sediment' also have a negative correlation. These both relate to material that has been resuspended from the seabed, and the negative correlation is to represent the fact that the particles can travel a similar distance by reducing one input value and increasing the other. The choice of the values of -0.9 for these correlations was made with the assistance of SEPA to represent a strong relationship between the inputs, but with a small level of flexibility. The other pre-defined correlations are between the Dispersion coefficients. First of all, there is a strong, positive correlation between each 'Dispersion Coefficient of Material from the cages' and each 'Dispersion coefficient for Resuspended Material'. These are used for the initial Correlated LHS as, SEPA advised that there is no reason for different Dispersion Coefficients to be specified for the different materials based on their experience. Secondly, there is a strong, positive correlation between the X and Y Dispersion Coefficients as SEPA advised that there is no reason why material would be affected more in either of the horizontal axes by turbulence. 0.99 was chosen to represent the correlation between the Dispersion Coefficients as these values are always considered as being the same when modelling. The choices of the values for $\mathbf{C}^*$ were made in collaboration with SEPA and therefore there is

uncertainty relating to the values, but these could be altered in future analyses if more information is available. The uncertainty of the values for $\mathbf{C}^*$ could impact the results if they were changed significantly, however, with the information available and to save computational time, these are not investigated further.

This target correlation matrix could then be used to add a correlation structure to the LHS. For this analysis, 100 different input sets were created using a standard LHS approach, before using the restricted pairing procedure to implement the correlation structure to the data.

2.2.4 Setup of NewDEPOMOD runs - inputs based on the physical properties

For each set of sample values created by the Correlated LHS, 100 replicate runs were completed to account for the random walk element within NewDEPOMOD. As a result, 10,000 runs were completed for each site. Due to the complex nature of NewDEPOMOD, these runs can take some time to complete, depending on the characteristics of the site. The average time for each run between the two sites in this analysis was approximately $60s$. This meant that the 10,000 runs would take approximately one week to complete.

For each of the completed runs at a given site, calculations of the scalar outputs could be made. The outputs being considered provide different measures of the impact of a farm on the environment. The Total Area Impacted provides an idea of the overall size of the impact, and the 99th Percentile of Solids Flux then provides a measurement of the intensity of the impact.

2.2.5 Methods for analysing the effect of the inputs on the scalar outputs of NewDEPOMOD

The main aims of the sensitivity analysis are to assess the influence of the uncertain inputs on the different model outputs. By ranking the inputs, it will provide useful information to SEPA by identifying elements of the model that will need to be considered cautiously when using NewDEPOMOD for modelling purposes. Moreover, by comparing two sites with different physical properties, any similarities or differences in the effects of the inputs can be identified.

Saltelli et al. (2008) outlined that global sensitivity analysis approaches tend to consider quantitative importance indices as a measure of comparison of the uncertain inputs.

2.2.5.1 Random Forest Models

Pianosi et al. (2016) mentioned that a way to rank inputs, using non-linear regression methods, is using Random Forests. Breiman (2001) defined a random forest as follows:

> *"A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{h(\boldsymbol{x}, \Theta_k), k = 1, \ldots\}$ *where the* $\{\Theta_k\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $\boldsymbol{x}$*."*

The basic idea of random forests was to combine the ideas of *Classification and Regression Trees* ('CART') (Breiman et al. 1984) and *bootstrapping aggregation* ('bagging') (Breiman 1996). Random forests can be used for either classification data or for regression purposes. In the classification case, they are produced by creating multiple classification trees using bootstrap samples of the data (Breiman 2001, Liaw & Wiener 2002). They are a nonparametric classification method, where multiple classification and regression trees are produced using random subsets of the data. However, the random forests being considered in this work relate to the regression framework, which will be described in more detail.

In order to explain the methodology of random forests for a regression setting, consider the p-dimensional input $\mathbf{x} = (x^1, \ldots, x^p)$, and the response Y, such that $Y = f(\mathbf{x}) + \epsilon$, with $\mathbb{E}[\epsilon|\mathbf{x}] = 0$. Consider a learning set, $L = \{(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)\}$, consisting of n independent observations of the vector $(\mathbf{x}, Y)$. Bagging is an ensemble learning method that generates B bootstrap samples from L. For each bootstrap sample, $\mathbf{Z}^{*b}$, the model is fitted, providing predictions, $\hat{f}_{*b}(\cdot)$, for $b = 1, \ldots, B$ (Hastie et al. 2009). For a given input set, $\mathbf{x}$, predictions can be made, $\hat{f}_{*b}(\mathbf{x})$ for each sample, $\mathbf{Z}^{*b}$, and the bagging estimate is defined as (Hastie et al. 2009),

$$\hat{f}_{bag}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{*b}(\mathbf{x}).$$

The bagging estimate, $\hat{f}_{bag}(\mathbf{x})$, is different from $\hat{f}(\mathbf{x})$, but is considered an effective tool for improving unstable estimates, and works well for high-variance, low-bias procedures such as trees (Hastie et al. 2009). CART (Breiman et al. 1984), are a technique that estimates f with respect to the mean square risk function. The starting point for their construction is to specify splitting rules of the form $(x^j < t)$, by recursive partitioning to obtain a maximal tree (Antoniadis et al. 2021). Using the learning sample, L, greedy selection is used to

select the best split which maximizes a local decreasing of heterogeneity which is measured by the difference between the variance of Y in the parent node and the output in the childs node (Antoniadis et al. 2021). In order to avoid overfitting the learning data using the maximal trees, insignificant nodes are cut off in order to choose the right size of tree - this process is called pruning (Breiman et al. 1984). Pruning is completed through the minimization of a penalized mean square error which features a penalty term that is linear in the number of leaves. One considerable drawback to CART are stability issues which occur through small changes in the learning set, L, which can have a large affect on the structure of the tree and any prediction values.

In order to overcome the stability issues surrounding CART, Breiman (2001) introduced the idea of random forests. Random trees are built using n_{tree} samples, $L^1, \ldots, L^{n_{tree}}$, from the learning set, L, and aggregating this set, representing the bagging element described previously (Antoniadis et al. 2021). The next step is to incorporate the modified CART methods (Breiman 2001). In order to speed up computations and without reducing the performance of the model, two changes are made to the CART approach (Breiman 2001):

1. A fixed number of randomly chosen inputs are considered at each node to identify the best split.

2. All of the trees in the forest are maximal trees that are not pruned.

Using this approach, the developed learning rule is the aggregation of all of the estimators resulting from those trees, which are given as $\hat{f}_1, \ldots, \hat{f}_{n_{tree}}$. The Out-Of-Bag ('OOB') sample is important in the definition of the variable importance. For a given tree, k, the OOB sample is given as the set of observations that are excluded from the bootstrap sample used in the construction of the tree k, and is denoted by $\bar{L}_k$. The OOB sample, $\bar{L}_k$, can be used to calculate the error of tree k by using $\bar{L}_k$ as a test sample. Doing this for each tree results in the OOB error for a random forest being defined as the average value of all the trees of the forest.

Using the trees, importance can be quantified and calculations of error rates can be made for the inputs in the classification and regression cases (Breiman 2001). Breiman (2001) proposed the permutation variable importance ('PVI') in the random forest model, which is the most used measure in the literature (Antoniadis et al. 2021). The PVI for a given input is defined as the mean over the trees of the forest, of the decreasing of the OOB error of a tree, when the values of the input are randomly permuted in the OOB samples (Breiman 2001). The mean square error ('MSE') is used to measure the OOB

error of a tree for regression random forests. For each tree, $k = 1, \ldots, n_{tree}$, the prediction error of $\hat{f}_k$ is evaluated among it's OOB sample, $\bar{L}_k$, with the empirical estimator:

$$\hat{R}(\hat{f}_k, \bar{L}_k) = \frac{1}{|\bar{L}_k|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{L}_k} \left(Y_i - \hat{f}_k(\mathbf{X}_i) \right)^2 .$$

Before moving on to calculate the PVI, let $\bar{L}_k^j$ denote the permuted OOB sample obtained from $\bar{L}_k$, after random permutation of the values of the jth input. Then the PVI for a given input, X_j, can be expressed as:

$$\hat{I}(X_j) = \frac{1}{n_{tree}} \sum_{k=1}^{n_{tree}} \left[\hat{R}(\hat{f}_k, \bar{L}_k^j) - \hat{R}(\hat{f}_k, \bar{L}_k) \right]. \tag{2.9}$$

This measure of importance can be defined as the mean increase in the prediction error, which is estimated with the help of the OOB error, over all the trees (Antoniadis et al. 2021). Due to the use of bootstrap sampling, the importance values for the parameters will be different when random forests are run multiple times, but the ranking of the inputs does not tend to vary unless importance values are very similar Liaw & Wiener (2002). As it is the ranking of the inputs that is required for this data, there is no need to run the random forest multiple times, and importance values that are similar would indicate that they have a similar ranking.

Harper et al. (2011) used random forests to develop a global sensitivity analysis method that would be appropriate for ecological models. Harper et al. (2011) used the concept of random forests in the Global Sensitivity Analysis as a measure of ranking the parameters of a model by their influence on model predictions, and recommended its use to assist with prioritization of research efforts using the ranking of parameters by their importance. The benefits of using random forests for ranking are:

- their ability to deal with non-linear relationships,
- their ability to incorporate interactions between inputs,
- and the importance values produced can be interpreted easily.

As a result of the above advantages, random forests are an appropriate tool here for assessing the influence of the uncertain inputs. The relative importance values will allow for the inputs to be ranked effectively.

2.2.6 Sensitivity analysis results for the inputs based on the physical properties

Ardessie and Muck were chosen as the two sites for this analysis as they have contrasting properties, with Ardessie being a low energy site and Muck being a high energy site. This will allow comparisons to be made between the sites to determine if the site characteristics impact the influence of the inputs. Each of the scalar outputs will be considered along with the inputs based on the physical properties from Table 2.1.

2.2.6.1 Total Area Impacted

In order to calculate this value, the total number of grid cells in the domain with a Solids Flux value greater than 0 were determined. Each grid cell in the domain is 25m $\times$ 25m and has a total area of 625m^2, so the Total Area impacted can be calculated by multiplying the number of grid cells with Solids Flux greater than 0 by the area of one grid cell. As the domain size is so big, the values calculated were then converted to km^2.

First, the Ardessie site will be considered. The Correlated LHS was created for Ardessie, with the Critical Shear Stress for Erosion values calculated using the minimum and maximum flow speeds at the site. For each set of sample values generated by the Correlated LHS, 100 replicate runs were completed. The output data could then be explored through a histogram. Figure 2.5 shows

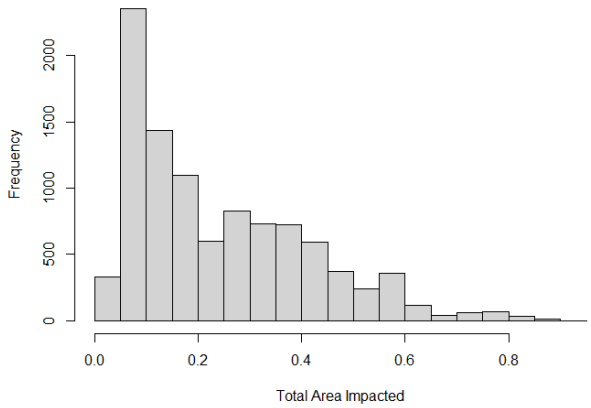


Figure 2.5: Histogram of the Total Area Impacted for Ardessie (km^2).

some skew in the data with a large proportion of the data having values between 0.05 and 0.1km^2. As there are a number of inputs being considered, as well as potential interactions between them, a random forest model will be fit-

ted first, before considering scatterplots of the Total Area Impacted against the higher ranking inputs. In this instance, 2000 trees were grown in order to fit the random forest model. The model explained approximately 99% of the variation in the data, indicating a very good model fit, but potentially some overfitting. In this instance, the aim of the analysis is to determine the inputs that have the biggest influence, therefore the overfitting isn't considered in detail. One potential reason for the large variation explained is the small influence of the random walk element within the NewDEPOMOD when considering the Total Area Impacted as the output. The importance values for the inputs were calculated using the formula in Equation 2.9, with values given in Table 2.2. Table

Inputs	**Importance**
Critical Shear Stress for Erosion	75.99%
Rate of Erosion	50.81%
Release Height	33.24%
Settling Velocity of Faeces	127.24%
Settling Velocity of Sediment	32.88%
Cage Dispersion Coefficient (X)	38.22%
Cage Dispersion Coefficient (Y)	37.96%
Cage Dispersion Coefficient (Z)	36.26%
Resuspended Material Dispersion Coefficient (X)	38.49%
Resuspended Material Dispersion Coefficient (Y)	38.03%
Resuspended Material Dispersion Coefficient (Z)	30.91%

Table 2.2: Table of Importance values from the random forest model of Total Area Impacted at Ardessie.

2.2 identifies the Settling Velocity of Faeces as being the most influential input in relation to the Total Area Impacted. The Critical Shear Stress for Erosion has the second highest Importance value, with Rate of Erosion having a slightly higher value than the remaining inputs. Scatterplots of the top two ranking inputs will now be considered. It should be noted that the importance values in Table 2.2 relate to the mean increase in the prediction error for the OOB samples. Therefore percentages greater than 100% are possible and represent an input with a great deal of influence on the output. Figure 2.7 appears to show a clear relationship between the Total Area Impacted and the Settling Velocity of Faeces. The Total Area Impacted remains constant (with a slight increase around -0.08m/s), until the Settling Velocity of Faeces increases to -0.04m/s, when it begins to increase. This would appear to make sense, as the closer the Settling Velocity is to zero, the longer the faeces remains in the water column to be transported by the currents. Figure 2.6 illustrates a weak, negative trend between Critical Shear Stress for Erosion and Total Area Im-

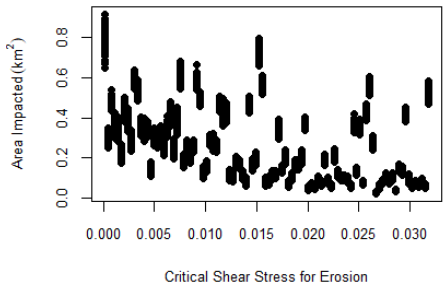


Figure 2.6: Plot of Total Area Impacted against the Critical Shear Stress for Erosion - Ardessie.

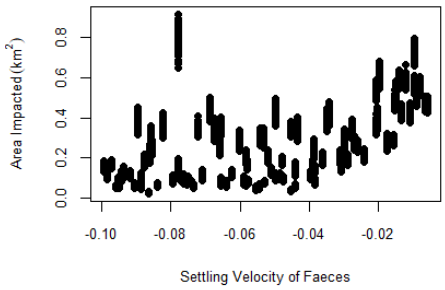


Figure 2.7: Plot of Total Area Impacted against the Settling Velocity of Faeces - Ardessie.

pacted. Again this appears reasonable, as the larger the Critical Shear Stress, the less resuspension that takes place, and therefore the less waste will travel.

Next, the high energy site, Muck, will be considered. As with Ardessie, a correlated LHS was created to create a total of 100 input sets, at which NewDEPOMOD was run 100 times to account for the random walk. Again, an initial plot of the Total Area Impacted date will be considered in the form of a histogram. Figure 2.8 shows that the values for Total Area Impacted

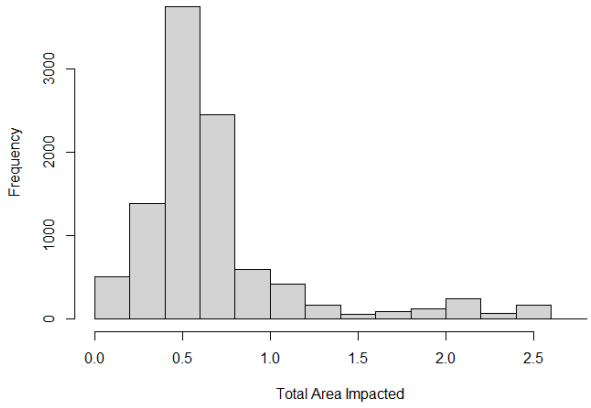


Figure 2.8: Histogram of the Total Area Impacted for Muck (km^2).

are larger than those seen at Ardessie. The majority of the data appears to be close to 0.5km^2, with a number of large values seen, going up to over 2.5km^2. The difference in values is likely down to Muck being a larger farm, and the faster current speeds producing more dispersion of the waste. Again, a random forest model with 2000 trees is fitted to the data, to allow the inputs to be ranked. The fitted model was again able to explain approximately 99%

of the variation in the data, and the importance values are given in Table 2.3. Table 2.3 identifies Settling Velocity of Faeces as having the highest importance

Inputs	**Importance**
Critical Shear Stress for Erosion	61.60%
Rate of Erosion	49.99%
Release Height	40.70%
Settling Velocity of Faeces	73.54%
Settling Velocity of Sediment	49.66%
Cage Dispersion Coefficient (X)	34.20%
Cage Dispersion Coefficient (Y)	31.73%
Cage Dispersion Coefficient (Z)	32.79%
Resuspended Material Dispersion Coefficient (X)	30.96%
Resuspended Material Dispersion Coefficient (Y)	28.99%
Resuspended Material Dispersion Coefficient (Z)	36.62%

Table 2.3: Table of Importance values from the random forest model of Total Area Impacted at Muck.

value, but the difference between this and Critical Shear Stress for Erosion is much smaller in comparison to the importance values in Table 2.2. The Rate of Erosion and Settling Velocity of Sediment both have similar importance values, and are slightly higher than the importance values for the remaining inputs. Scatterplots of the Settling Velocity of Faeces and Critical Shear stress are given below. One thing to note from Figure 2.9 is that the Total Area

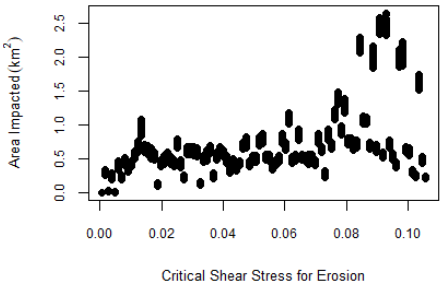


Figure 2.9: Plot of Total Area Impacted against the Critical Shear Stress for Erosion - Muck.

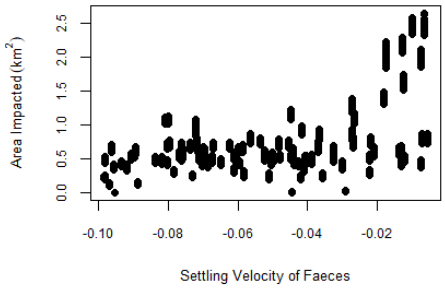


Figure 2.10: Plot of Total Area Impacted against the Settling Velocity of Faeces - Muck.

Impacted appears to be fairly consistent across the range of Critical Shear Stress, with a small positive trend, except for a number of large values seen for large values of Critical Shear Stress. One possible explanation for these large values could be that there is an interaction effect with another input that is causing these values. Next, looking at Figure 2.10, there appears to be a

similar pattern to Figure 2.9, where the values are consistent across the x-axis, with the exception of some large values for Total Area Impacted occurring at the Settling Velocity of Faeces close to zero. This could highlight that the two inputs are having an effect on the Total Area Impacted.

Looking at both Tables 2.2 and 2.3, there are similarities in the ranking of the inputs with the three largest importance values. The differences relate to the actual importance values. At Ardessie, Settling Velocity of Faeces had an importance value that is almost double that of Critical Shear Stress for Erosion, demonstrating that at the site with slower current speeds (Ardessie), Settling Velocity of Faeces plays a more dominant role. At the faster flowing site (Muck), Settling Velocity of Faeces is still the highest ranked input, but its importance value is closer to the importance value for the second highest ranked inputs - Critical Shear Stress for Erosion. This may indicate that the Settling Velocity of Faeces plays a bigger role in the Total Area Impacted in the low energy sites as it affects the amount of time particles spend in the water column, and therefore how far the particles are transported. In the faster flowing sites, it would indicate that it does not have as big an effect as the current speeds are faster and the effect of resuspension is stronger at these sites. The scatterplots for the two highest ranked inputs will be compared to identify any similarities/differences in the patterns. The first thing to notice

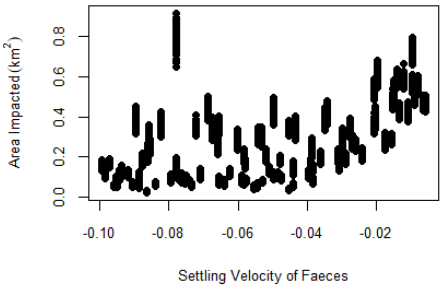


Figure 2.11: Plot of Total Area Impacted against the Settling Velocity of Faeces - Ardessie.

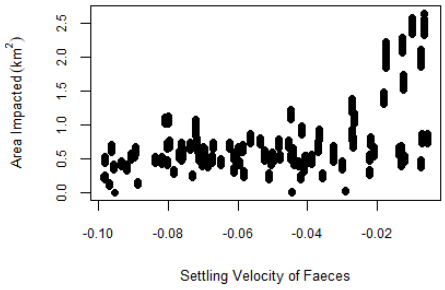


Figure 2.12: Plot of Total Area Impacted against the Settling Velocity of Faeces - Muck.

across all of the plots is the scale of the y-axis (Total Area Impacted) for both sites. The Total Area Impacted is greater for Muck which is potentially a result of the fact that it is a larger farm. Despite this, the main focus is the shape of the pattern for the two highest ranked inputs. First, the shape for both sites when looking at Settling Velocity of Faeces is similar, with the exception of some low values for Total Area Impacted seen at Settling Velocities close

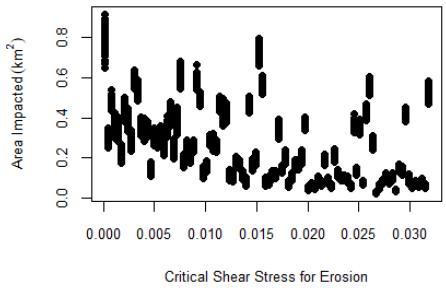


Figure 2.13: Plot of Total Area Impacted against the Critical Shear Stress for Erosion - Ardessie.

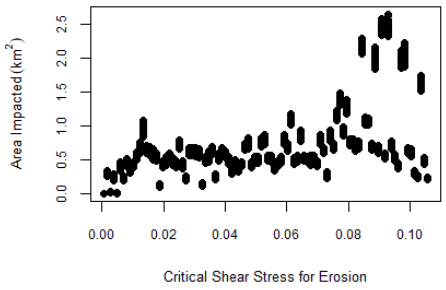


Figure 2.14: Plot of Total Area Impacted against the Critical Shear Stress for Erosion - Muck.

to zero. Excluding one set of points in Figure 2.11 when Settling Velocity of Faeces is -0.08m/s, the variance of the data remains fairly even. In Figure 2.12, the variance is consistent until Settling Velocity of Faeces is -0.04m/s, when it increases. Both Figures 2.13 and 2.14 appear to show opposing patterns. At Ardessie, Critical Shear Stress for Erosion has a slight negative trend and fairly consistent variance. Whereas for Muck, there is potentially a small positive trend, but an increase of variance is seen as Critical Shear Stress increases.

2.2.6.2 99th Percentile of Solids Flux

As the methods used for the 99th Percentile of Solids Flux are similar to the methods for Total Area Impacted, a comparison between the results for the two sites will be sufficient. The analysis will feature the importance values from the random forest models, as well as scatterplots for the highest ranking inputs against the 99th Percentile of Solids Flux. Firstly, the initial histograms of the output data are considered. The data for Ardessie appears to be slightly more skewed than the data for Muck, which appears to be bimodal, with a dip seen at 8kg/m^2/year. For the comparison of the ranking, a random forest model was fitted for each site (with both explaining approximately 99% of the variation in the data) and the importance values for the inputs can be seen in Table 2.4. For the 99th Percentile, the top two inputs are different for the two sites, and it is also different to the ranking for Total Area Impacted in Tables 2.2 and 2.3. Firstly, looking at Ardessie, Settling Velocity of Faeces is again the inputs with the biggest influence on the 99th Percentile of Solids Flux, with its importance value being much bigger than the second highest ranked inputs. The order of Critical Shear Stress for Erosion and Rate of Erosion has changed this time, but with only a small difference between their importance values.

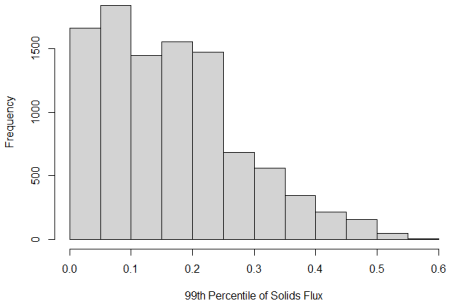


Figure 2.15: Histogram of the 99th Percentile for Solids Flux at Ardessie (kg/m^2/year).

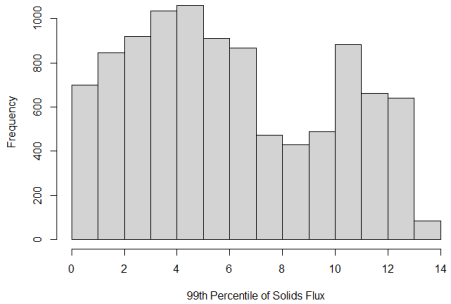


Figure 2.16: Histogram of the 99th Percentile for Solids Flux at Muck (kg/m^2/year).

	Importance	
Inputs	**Ardessie**	**Muck**
Critical Shear Stress for Erosion	51.78%	90.11%
Rate of Erosion	55.04%	50.14%
Release Height	34.59%	46.28%
Settling Velocity of Faeces	125.66%	85.41%
Settling Velocity of Sediment	42.09%	63.71%
Cage Dispersion Coefficient (X)	46.43%	29.44%
Cage Dispersion Coefficient (Y)	42.10%	34.63%
Cage Dispersion Coefficient (Z)	39.87%	33.18%
Resuspended Material Dispersion Coefficient (X)	39.63%	28.44%
Resuspended Material Dispersion Coefficient (Y)	42.91%	29.56%
Resuspended Material Dispersion Coefficient (Z)	36.25%	38.80%

Table 2.4: Table of Importance values from the random forest Model of 99th Percentile at each site.

The remaining inputs have importance values that are reasonably close to the importance values for Rate of Erosion and Critical Shear Stress for Erosion.

Now considering the ranking of the inputs for Muck, the highest ranked inputs is Critical Shear Stress for Erosion. Settling Velocity of Faeces is second, but with a small difference between it and Critical Shear Stress for Erosion.

Now comparing the importance values for the two sites, it is clear that Settling Velocity of Faeces plays a big role at both sites, but it is much more influential at the site with slower current speeds again. The Critical Shear Stress for Erosion is much more influential at the faster flowing site, which is to be expected as the faster current speeds will erode more particles from the seabed and transport them further. Again, scatterplots of the inputs with the highest importance values can be compared for the two sites. Due to the differences in the biomass for each farm, the scale of the y-axis is much larger

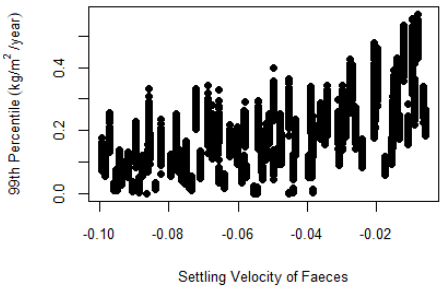


Figure 2.17: Plot of 99th Percentile for Solids Flux against the Settling Velocity of Faeces - Ardessie.

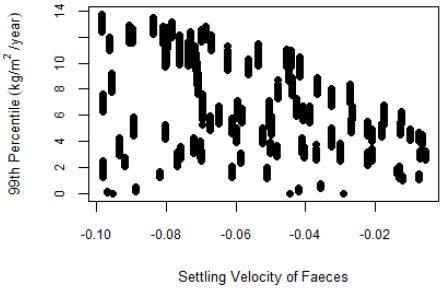


Figure 2.18: Plot of 99th Percentile for Solids Flux against the Settling Velocity of Faeces - Muck.

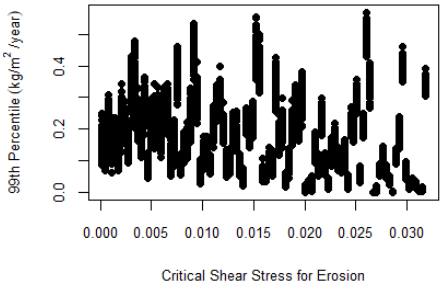


Figure 2.19: Plot of 99th Percentile for Solids Flux against the Critical Shear Stress for Erosion - Ardessie.

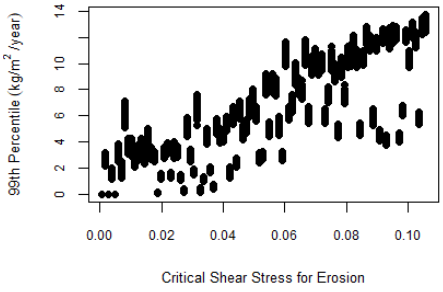


Figure 2.20: Plot of 99th Percentile for Solids Flux against the Critical Shear Stress for Erosion - Muck.

for the site at Muck, but again, the focus is on the patterns in the plots. First, considering the Settling Velocity of Faeces in Figures 2.17 and 2.18, there is an overall positive trend for Ardessie, and a possible decreasing trend for Muck, with a decrease in the variation as the Settling Velocity of Faeces increases. At Muck, the closer the Settling Velocity of Faeces gets to zero, the more time particles spend in the water column, and are therefore dispersed more across the domain, resulting in lower values for the 99th Percentile of Solids Flux. However, at Ardessie, the opposite is seen, where the 99th Percentile values increase as the Settling Velocity of Faeces approaches zero. Figure 2.19 does not demonstrate an obvious pattern, whereas there is a clear increasing trend in the 99th Percentile as Critical Shear Stress for Erosion increases in Figure 2.20. This is in line with what would be expected, as increases in Critical Shear Stress for Erosion mean that less particles are eroded from the seabed, and therefore they are not being transported as far, which would result in a

more intense impact in some grid cells.

2.2.6.3 Mass Balance

Again, comparisons will be made between the sites, with the initial output data considered before fitting random forest models, and looking in more detail at the highest ranking inputs. The Mass Balance data at Ardessie in Figure

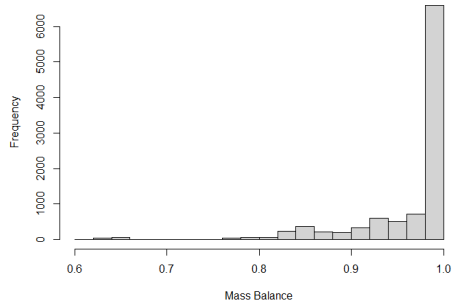


Figure 2.21: Histogram of the Mass Balance at Ardessie.

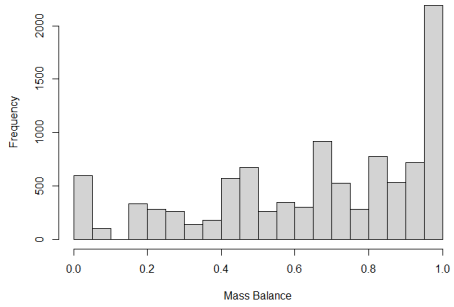


Figure 2.22: Histogram of the Mass Balance at Muck.

2.21 indicates that in most of the NewDEPOMOD runs, approximately all of the waste material remains in the domain. In comparison, at Muck in Figure 2.22, there are still a large amount of runs with approximately all of the waste material remaining in the domain, but there are more runs where waste material leaves the domain in comparison to Ardessie. Transformations of the Mass Balance data such as log, square root, cube root, were considered, but even after testing multiple transformations, the data remained heavily skewed, with only small improvements seen. One thing to note with the Mass Balance output data is that it is bounded in the unit interval. The purpose of this analysis is to determine which of the inputs have the biggest influence on Mass Balance, therefore, for the purposes of this analysis, the random forest model is fitted without restricting the output data. Upon fitting the models, the importance values will be examined to determine if they seem plausible and the unrestricted output data can be used for this purpose. Random forest models were fitted to the original data in a similar way as before in order to rank the inputs and compare sites. The calculated importance values are displayed in Table 2.5. In the first instance, the results appear to be plausible, with the Settling Velocity of Faeces and Critical Shear Stress for Erosion being most influential at both sites. This would be expected as the Settling Velocity of Faeces determines how far the waste particles are transported initially, and Critical Shear Stress for Erosion determines how much material is resuspended. Looking at the

Inputs	Importance	
	Ardessie	**Muck**
Critical Shear Stress for Erosion	62.09%	92.25%
Rate of Erosion	49.98%	48.76%
Release Height	37.89%	58.27%
Settling Velocity of Faeces	81.97%	68.86%
Settling Velocity of Sediment	44.76%	76.27%
Cage Dispersion Coefficient (X)	18.86%	33.13%
Cage Dispersion Coefficient (Y)	26.39%	36.89%
Cage Dispersion Coefficient (Z)	28.94%	33.68%
Resuspended Material Dispersion Coefficient (X)	22.30%	35.55%
Resuspended Material Dispersion Coefficient (Y)	22.19%	35.20%
Resuspended Material Dispersion Coefficient (Z)	31.07%	38.29%

Table 2.5: Table of Importance values from the random forest model of Mass Balance at each site.

sites individually, it can be seen that Settling Velocity of Faeces is again the most important input for Ardessie, with Critical Shear Stress for Erosion and Rate of Erosion having the second and third highest values. Release Height and Settling Velocity of Sediment are not far behind Rate of Erosion, with the dispersion inputs all having similar low values. For Muck, the Critical Shear Stress for Erosion is the highest ranked input, with Settling Velocity of Sediment and Settling Velocity of Faeces second and third. Release Height and Rate of Erosion are not far behind, while the Dispersion coefficients are also lower and of similar values for Muck as well.

One thing to note when looking at both sites is that Release Height and Settling Velocity of Sediment are playing a bigger role when considering Mass Balance as the output. As these inputs are involved in the process of resuspension of particles on the seabed, it would be expected that these have more of an influence on Mass Balance. Together with the erosion inputs, they influence how long resuspended particles remain in the water column and therefore how far they are transported. The further they are transported will impact whether or not they remain in the domain. As expected, they are more influential at the faster flowing site as particles will be transported further by the faster flowing currents at Muck. At Ardessie, the Settling Velocity of Faeces is more likely to play a role in the transportation of particles as resuspended particles will not be transported as far by the slower current.

As the Critical Shear Stress for Erosion and Settling Velocity of Faeces are the most important inputs at the sites, their scatterplots will be compared below. Looking at Figures 2.23 and 2.24, the scales of the y-axis differ again,

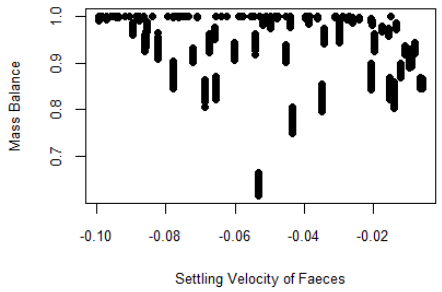


Figure 2.23: Plot of Mass Balance against the Settling Velocity of Faeces - Ardessie.

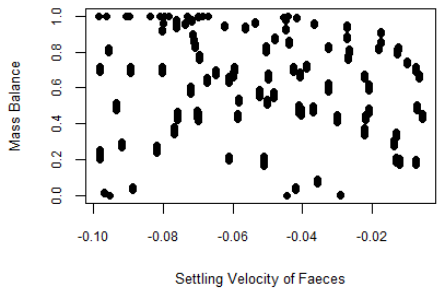


Figure 2.24: Plot of Mass Balance against the Settling Velocity of Faeces - Muck.

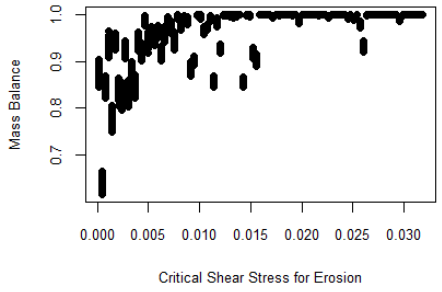


Figure 2.25: Plot of Mass Balance against the Critical Shear Stress for Erosion - Ardessie.

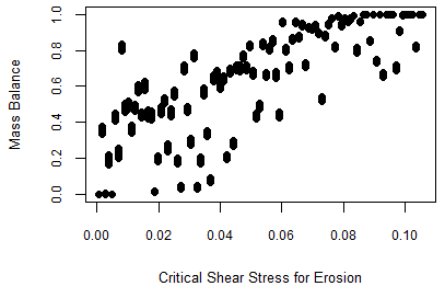


Figure 2.26: Plot of Mass Balance against the Critical Shear Stress for Erosion - Muck.

which is to be expected as at Ardessie, the slower flow speeds mean most of the material remains in the domain. There is a lot of variation in the data, but there does appear to be a negative trend when the absolute value of the Settling Velocity of Faeces is below 0.02m/s. This may indicate a possible threshold value where it is unlikely for all of the mass to remain in the domain, even for a site with slower flow speeds. Now considering Figures 2.25 and 2.26, there is a clear positive trend at both sites. This is to be expected, as lowering the Critical Shear Stress for Erosion allows more particles to be resuspended and therefore transported further, and potentially out of the domain.

2.2.6.4 Summary

The sensitivity analysis of the inputs based on the physical properties were completed at two different sites, and also considered three different scalar outputs. There were differences between the two sites in regards to the most

influential inputs, with the site characteristics being the likely cause. It was clear throughout that the Settling Velocity of Faeces appeared to be more influential at the low energy site for all outputs. At the high energy site, the Settling Velocity of Faeces and the Critical Shear Stress for erosion were identified as the highest ranking inputs for the different outputs. At the other end of the scale, it was clear that the dispersion coefficients had the least influence, with lower importance values for all of the scalar outputs. In addition, the Release Height of Resuspended Material and the Settling Velocity of Sediment had lower importance values similar to the dispersion coefficients for the low energy site. In contrast, these two inputs appeared to be more influential than the dispersion coefficients at the high energy site.

Due to the low rankings of the dispersion coefficients at both the low and high energy sites, their uncertainty does not appear to be influencing the scalar outputs calculated from the output maps. As a result, these inputs will not be considered in the further combined analysis including the physical properties and operational inputs. The remaining inputs will be considered in a further sensitivity analysis containing the inputs based on the physical properties, as well as the operational inputs.

2.3 Sensitivity Analysis - Operational Inputs

Moving on, the operational inputs will now be considered to determine the impacts of increasing the Biomass and altering the cage setup. For this analysis, a new low energy site, Ardentinny, will be considered alongside Muck, though the methods could be repeated for sites with different properties and compared. Ardentinny is a larger farm than Ardessie, with a Biomass value similar to Muck.

2.3.1 Aims of the analysis and the inputs to be investigated

SEPA identified the operational inputs of interest to be Biomass, Cage Diameter and Number of Cages. Biomass is of interest as it refers to the amount of fish being farmed at a particular site, and there are future plans to increase production in Scotland, and one way to do so is by increasing Biomass at sites. In order to increase Biomass at sites, the cage setup will have to be altered, either by making the current cages larger, or by adding more cages. These inputs will be considered to help determine the impact on NewDEPOMOD

predictions of increasing the Biomass and altering the farm setup.

2.3.2 Establishing suitable ranges for inputs

There are several ways that a fish farm can be constructed and so, before beginning the analysis, the inputs will be considered in more detail. The analysis will be based around the current setup for each licensed site. Changes to the operational setup of the farms will involve:

- increasing Biomass,
- increasing Cage Size,
- adding extra cages.

There are multiple options for altering the three inputs, and so the analysis was limited to allow more realistic scenarios to be considered at each site. Another factor that was pivotal in determining the farm setups being tested was the Stocking Density. This is a measure of how many fish are kept in the cages - measured in kilograms per cubic metre (kg/m^3). The stocking density is calculated based on the size of the cages and the Biomass and is a crucial component for maximising fish growth but not compromising fish welfare. Turnbull et al. (2005) found that stocking densities above $22kg/m^3$ resulted in lower levels of fish welfare, and Canon Jones et al. (2011) discovered that salmon became more aggressive in more densely stocked cages. Another drawback of densely stocked cages is the increased risk of pathogens evolving at rapid rates as the densely packed cages provide perfect conditions (Sundberg et al. 2016). RSPCA standards specify a maximum Stocking Density of $22kg/m^3$, however, fish farms are not required to be certified by the RSPCA standards in Scotland - in 2018 approximately 78% of farms were certified. SEPA specified that a plausible maximum would be $25kg/m^3$, but across all farms in Scotland, it is rare for a farm to operate at this Stocking Density. As a result, 4 different maximum Stocking Densities were considered in the analysis: $\{16.3kg/m^3$ (median across farms in Scotland), $18.4kg/m^3$ (average of the median and 95th percentile), $20.5kg/m^3$ (95th percentile), $25kg/m^3$ (SEPA recommended maximum)$\}$.

In order to test the effects of increasing Biomass, the analysis is focused around the default Biomass value provided for each site's DEPOMOD inputs - consider this as 100%. In order to allow for considerable expansions to the industry over time, a maximum Biomass level of 300% of the default value was determined. Focus was placed on smaller increases in Biomass to investigate more plausible increases that could be seen in the industry, and the following

Biomass percentages of the default were chosen: {100% (default), 110%, 120%, 150%, 200%, 300%}.

Across Scotland, the largest diameter of cage that is used is approximately 38.2m and can be seen at the largest sites. AKVA Group is the leading supplier of plastic and steel cages in aquaculture, with the largest cage in production having a diameter of 83m. The largest cage size to be considered in this analysis is double the current largest size (38.2m) that is used at the largest sites - 76.4m. As with the Biomass values, there will be smaller differences between the values closer to the default.

The final operational input that will be altered in the analysis is the number of cages in the farm setup. The standard setup of a fish farm in Scotland has the cages laid out in pairs, therefore when cages are added to a farm, it will be done in pairs to keep the same pattern. In order to stop the analysis from becoming unrealistic by adding multiple new cages, the number of additional cages is restricted to eight (4 pairs). This will allow for two additional pairs of cages at either end of the current cage layout.

2.3.3 Setup of NewDEPOMOD runs - operational inputs

As previously mentioned, the scenarios being tested had to be considered as realistic. As a result, specific values were chosen to be tested for each of the different inputs being considered in the analysis. These values were chosen to allow a reasonable range of different operational scenarios to be considered, that take into account future advancements in the industry. The method used to manage the number of runs required is based around the maximum Stocking Density, and only altering the farm setup when the maximum Stocking Density is exceeded - as seen in Figure 2.27. The flowchart will determine the number of runs required for each site, and the different scenarios to be tested. In each case where the Stocking Density is exceeded, there are two scenarios to be tested - 1) increasing the cage diameter, 2) adding more cages to the farm setup. One key point to note is the maximum of 4 additional pairs of cages being added to the default farm setup. If the maximum stocking density is exceeded after adding the four pairs of cages, then the new cages are removed, the cage diameter is increased and the process of adding pairs of cages begins again. One key thing to note is that due to the structure of the analysis, the inputs are not independent of each other, therefore interaction terms will be required in any modelling approaches.

Using Figure 2.27, the different scenarios to be tested were identified and

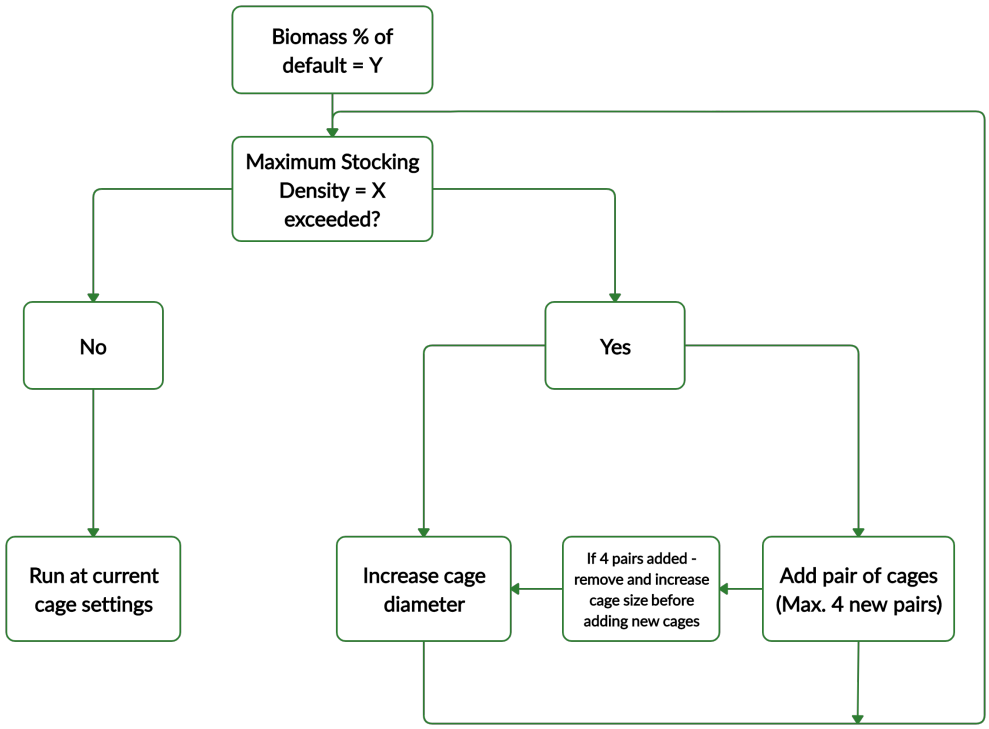


Figure 2.27: Flow chart illustrating the process for creating runs at each site.

are displayed in Tables 2.6 and 2.7 for each site. For a given Biomass level, if the maximum stocking density is not exceeded for 1 or more of the values, then only 1 scenario is required at the default settings. If one of the maximum stocking densities is exceeded for a Biomass value, then 2 scenarios are required to represent increased cage size and additional cages. This approach was used to identify the potential different scenarios to be considered for the low energy site, Ardentinny, and the high energy site, Muck. The following two tables illustrate the potential scenarios. Consider a Biomass level of 120% of the

Biomass % of Default	**Stocking Density Exceeds Maximum? (Y or N)**				**No. of Scenarios**
	Max = 16.3	**Max = 18.4**	**Max = 20.5**	**Max = 25.0**	
100	N	N	N	N	1
110	N	N	N	N	1
120	Y	N	N	N	3
150	Y	Y	N	N	5
200	Y	Y	Y	Y	8
300	Y	Y	Y	Y	8
Total					**26**

Table 2.6: Scenarios to be tested for analysing the effect of altering the operational inputs of NewDEPOMOD at Muck.

default, then from Table 2.6, the maximum stocking density of 16.3kg/m^3 is exceeded, and so the cage setup will have to be altered. Therefore there will be 2

Biomass % of Default	**Stocking Density Exceeds Maximum? (Y or N)**				**No. of Scenarios**
	Max = 16.3	**Max = 18.4**	**Max = 20.5**	**Max = 25.0**	
100	N	N	N	N	1
110	N	N	N	N	1
120	N	N	N	N	1
150	Y	N	N	N	3
200	Y	Y	Y	N	7
300	Y	Y	Y	Y	8
Total					**21**

Table 2.7: Scenarios to be tested for analysing the effect of altering the operational inputs of NewDEPOMOD at Ardentinny.

scenarios that are constructed to reduce the stocking density for this particular maximum. The stocking density is not exceeded for any of the other maximum values, and so one scenario with the default settings is suitable for the other three maximum values. This approach was used to determine the number of different scenarios to be set up and run. For the two scenarios to be tested when the maximum stocking density is exceeded, there are potentially cage setups that overlap for a given Biomass value. Therefore the maximum number of scenarios to be tested will be 26 for this site. In addition, for Ardentinny, Table 2.7 highlights that a maximum number of 21 different scenarios are to be tested.

2.3.4 Results for operational inputs

The operational inputs are ones that can be controlled by the fish farm operators, so the effects of altering these inputs on the scalar summaries will be considered for Ardentinny (low energy site) and Muck (high energy site). For this analysis, it will be considered in two parts - 1) Total Area Impacted will be considered as the output at both sites, then 2) 99th Percentile of Solids Flux will be considered as the output at both sites. In order to avoid repetition, the analysis for Total Area Impacted will be considered in more detail for Ardentinny, and then a summary of the results will be provided for the remaining site, and also for the consideration of 99th Percentile of Solids Flux as the output. As Mass Balance does not provide an indication of the environmental impact of the fish farm, it will not be considered in this, or any further analyses.

2.3.4.1 Total Area Impacted

The first stage of the analysis will consider Total Area Impacted as the scalar output, and aim to identify the effects of altering the operational inputs at Ardentinny. Before considering the sensitivity analysis, initial plots of the Total Area Impacted against each input individually will be considered. As each of the operational inputs are considered as discrete variables in the setup of the analysis, the initial plots are boxplots. From Figure 2.28 there is a clear

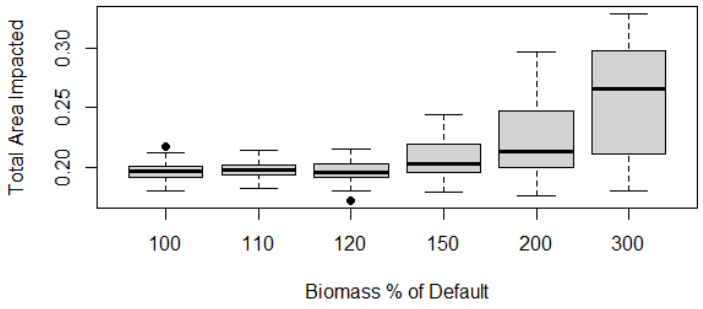


Figure 2.28: Box plot of the Total Area Impacted (km^2) against the Biomass - Ardentinny.

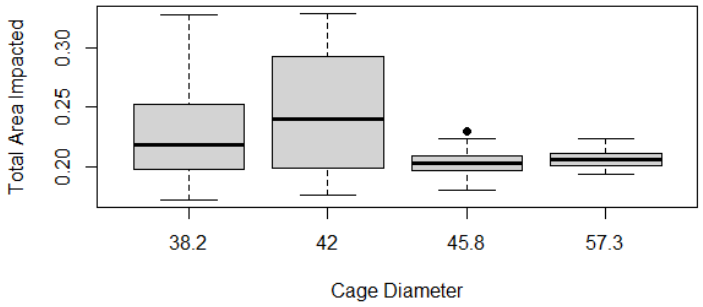


Figure 2.29: Box plot of the Total Area Impacted (km^2) against the Cage Diameter - Ardentinny.

increasing trend in the average Total Area Impacted as Biomass increases, with an additional increase in the variance as Biomass increases. Figure 2.29 shows larger variance for the smaller Cage Diameters, likely caused by the different number of Biomass values tested at the smaller cage sizes. In contrast to the other two plots, Figure 2.30 has constant variance across all values, and a positive, linear trend. The trend seen in Figure 2.30 suggests that it is the

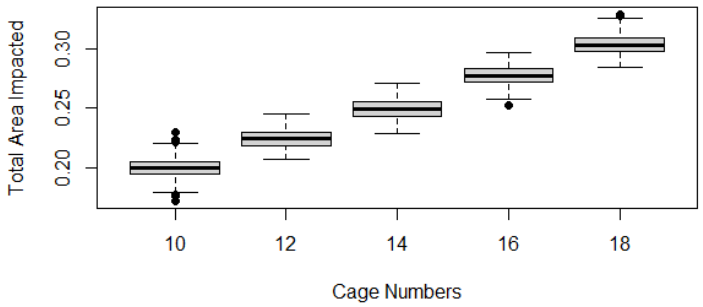


Figure 2.30: Box plot of the Total Area Impacted (km^2) against the Number of Cages - Ardentinny.

dominant input at this site. To explore potential interactions, scatterplots for Total Area Impacted against Biomass and Cage Diameter are produced, with the points coloured based on the corresponding Number of Cages for a give run. Figures 2.31 and 2.32 highlight the fact that the lower values

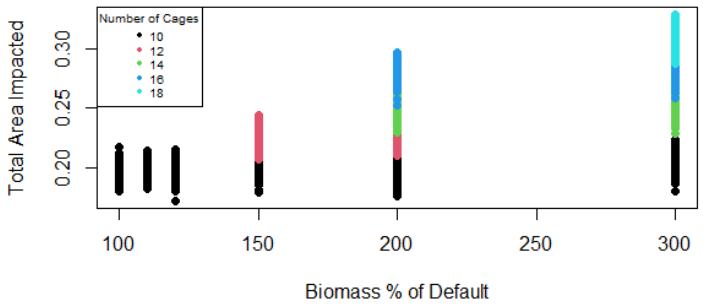


Figure 2.31: Initial plot of the Total Area Impacted (km^2) against the Biomass, coloured by the relative Number of Cages - Ardentinny.

of Total Area Impacted are a result of the smaller Number of Cages in the operational setup. Figure 2.31 reveals that, where multiple Numbers of Cages are considered for a given Biomass (300% of Default) or Cage Diameter in Figure 2.32 (e.g. 38.2m), the Total Area Impacted appears to be ordered based on the Number of Cages. This enforces the idea that the Number of Cages is the driving force at Ardentinny. The next step is to assess the sensitivity of Total Area Impacted to altering these operational inputs, in order to confirm the most influential inputs. A similar approach to the method used for the inputs based on the physical properties will be considered. A random forest model is fitted, which produces a ranking of the inputs based on an importance value.

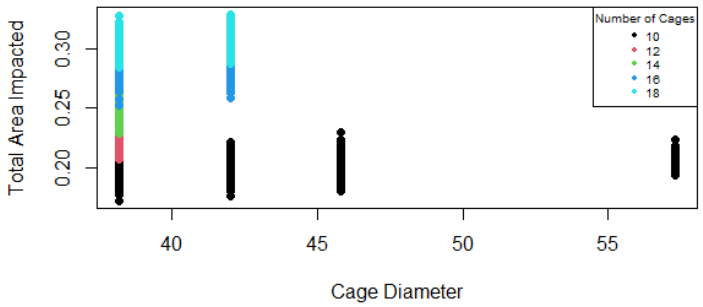


Figure 2.32: Initial plot of the Total Area Impacted (km^2) against the Number of Cages, coloured by the relative Biomass values - Ardentinny.

The random forest model was fitted, with 2000 trees grown, which produced a model that explained approximately 91% of the variation in the Total Area Impacted. From the model, importance values were able to be determined based on the increase in MSE, and can be seen in Table 2.8. As the random

Inputs	**Importance**
Biomass % of Default value	44.6%
Cage Diameter	40.5%
Number of Cages	68.5%

Table 2.8: Table of Importance values from the random forest model corresponding to the Total Area Impacted modelled by the operational inputs - Ardentinny.

forest model was able to explain a large amount of the variation in the Total Area Impacted, suitable conclusions can be drawn from the importance values in Table 2.8. The Number of Cages was ranked as the most important predictor, which could have been predicted based on the initial plots. It appeared to be the dominant input, and played a role in the increased variance seen for certain Biomass and Cage Diameters.

Following the analysis of the runs from Ardentinny, the results from the runs at Muck will be considered and summarised. As with Ardentinny, box plots were considered initially to help identify a potentially dominant input. Biomass was identified as the dominant input from the box plots, where it demonstrated a positive linear trend, and similar variance across all Biomass values. To confirm this, scatterplots for Total Area Impacted against Cage Diameter and Number of Cages were produced, with the points coloured based on the corresponding Biomass value. Figures 2.33 and 2.34 highlight the fact that the lower values for Cage Diameter and Number of Cages feature a

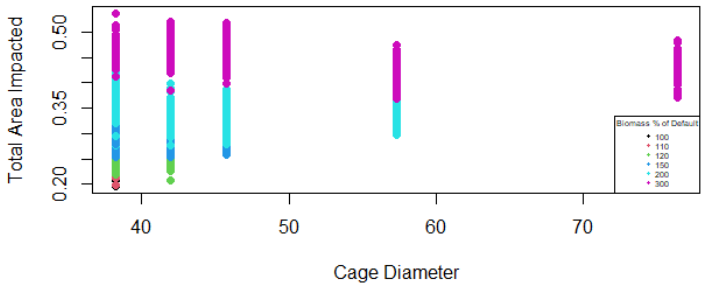


Figure 2.33: Initial plot of the Total Area Impacted (km^2) against the Cage Diameter, coloured by the relative Biomass values - Muck.

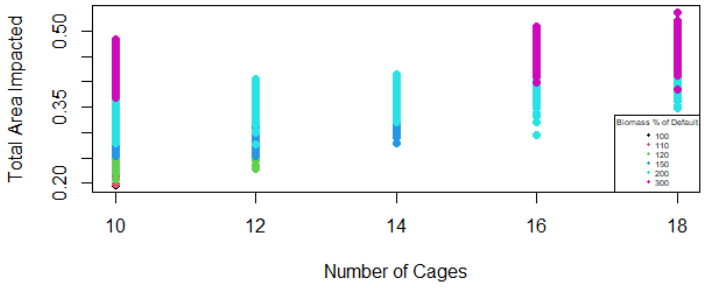


Figure 2.34: Initial plot of the Total Area Impacted (km^2) against the Number of Cages, coloured by the relative Biomass values - Muck.

variety of Biomass values which appear to be the cause for the larger amounts of variation. Where a number of Biomass values are considered for a given Cage Diameter or Number of cages, they appear to be stacked, with lower Total Area Impacted for lower Biomass values. As with Ardentinny, a random forest model is fitted to rank the inputs. The model for this site explained approximately 90.6% of the variation in the data, and the importance values are given in Table 2.9. From Table 2.9, the Biomass was ranked as the most important predictor,

Inputs	**Importance**
Biomass % of Default value	73.5%
Cage Diameter	55.6%
Number of Cages	60.9%

Table 2.9: Table of Importance values from the random forest model corresponding to the Total Area Impacted modelled by the operational inputs.

which was expected based on the initial plots. In comparison to the results for Ardentinny in Table 2.8, different inputs are identified as being the most

influential, but Cage Diameter is identified as the least influential at both sites when considering Total Area Impacted as the output.

2.3.4.2 99th Percentile of Solids Flux

Following the analysis when considering the Total Area Impacted as the single, scalar input, the 99th Percentile of Solids Flux will be considered. As mentioned previously, this is used as a measure of the intensity of the impact on the seabed. A similar approach will be used as the one used for the Total Area Impacted. Again, Ardentinny was considered first, with box plots viewed to help identify a potentially dominant input. From the initial box plots, Biomass appeared to be the dominant input, with a positive linear trend as Biomass increased, but with a slight increase in variance for larger Biomass values. Using Biomass as the dominant input, scatterplots of 99th Percentile of Solids Flux against the remaining inputs are produced, with the observations coloured based on the corresponding Biomass value. In a similar manner to the plots

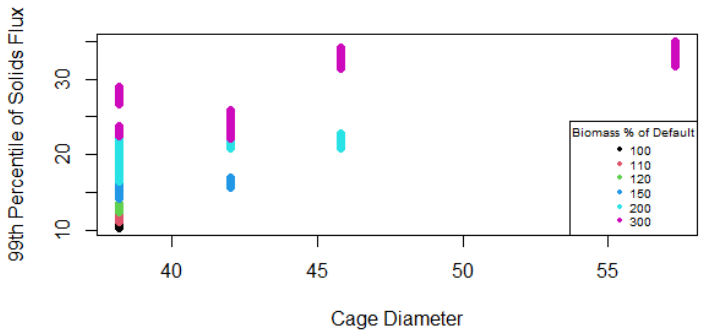


Figure 2.35: Initial plot of the 99th Percentile of Solids Flux against the Cage Diameter, coloured by the relative Biomass - Ardentinny.

for the Total Area Impacted, here, the Biomass appears to be the dominant input. Figure 2.35 shows a clear stacking of the Biomass values for the lowest Cage Diameter size. The same pattern can be seen in Figure 2.36 when the farm features 10 cages. Next, the random forest model was fitted to the data, and it described approximately 93.6% of the variation, with the corresponding importance values for the inputs given in the following table. Table 2.10 shows that Biomass was identified as the most important input, as expected. The remaining inputs then have similar importance values, with the Number of Cages slightly larger.

Next, the same analysis will be completed for the data at Muck. The initial boxplots also identified Biomass as the dominant input for this site, and the

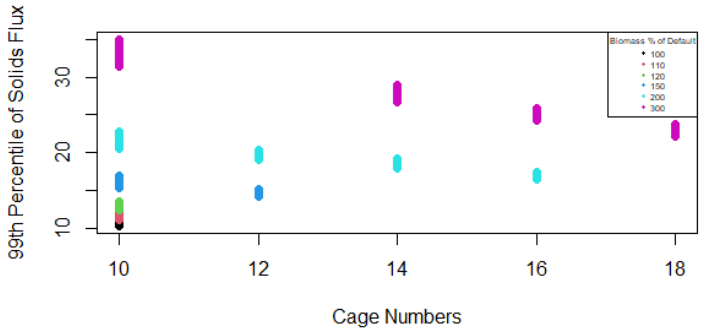


Figure 2.36: Initial plot of the 99th Percentile of Solids Flux against the Number of Cages, coloured by the relative Biomass - Ardentinny.

Inputs	**Importance**
Biomass % of Default value	74.4%
Cage Diameter	50.1%
Number of Cages	55.5%

Table 2.10: Table of Importance values from the random forest model corresponding to the 99th Percentile of Solids Flux modelled by the operational inputs - Ardentinny.

following scatterplots show the 99th Percentile of Solids Flux plotted against the remaining inputs, with the points coloured based on the Biomass values.

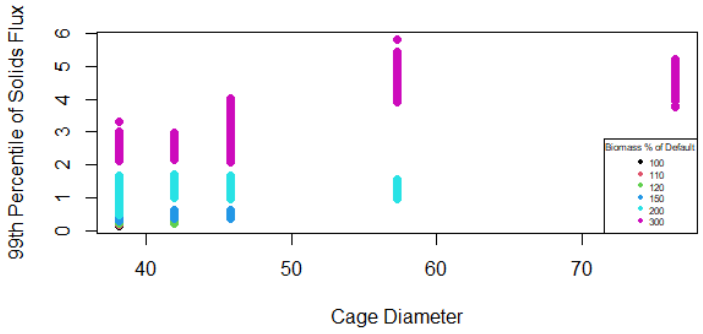


Figure 2.37: Initial plot of the 99th Percentile of Solids Flux against the Cage Diameter, coloured by the relative Biomass - Muck.

Again, the plots appear to show a stacking pattern, with the larger values for 99th Percentile of Solids Flux corresponding to the larger values of Biomass. In order to confirm if Biomass is again the dominant input, a random forest model was fitted, which described approximately 95% of the variation in the data, and the resulting importance values are given in the table below. As

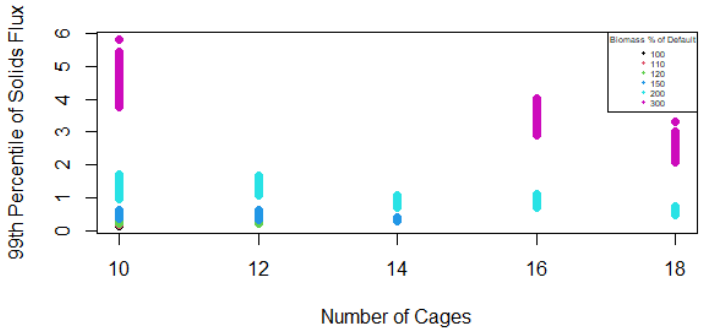


Figure 2.38: Initial plot of the 99th Percentile of Solids Flux against the Number of Cages, coloured by the relative Biomass - Muck.

Inputs	Importance
Biomass % of Default value	71.2%
Cage Diameter	45.6%
Number of Cages	57.4%

Table 2.11: Table of Importance values from the random forest model corresponding to the 99th Percentile of Solids Flux modelled by the operational inputs - Muck.

expected the Biomass is identified as the most influential input, and a larger gap between the remaining inputs in comparison to the results at Ardentinny. Tables 2.11 and 2.10 both identify the Biomass as having the most influence on the 99th Percentile of Solids Flux. This seems reasonable as the larger Biomass results in more waste leaving the cages, and therefore, the deposition on the seabed will be more intense.

2.3.4.3 Summary

This process has identified the influence of the operational inputs on the scalar outputs. Different inputs were identified as being the most important for the low and high energy sites when considering the Total Area Impacted as the output. The low current speeds at Ardentinny mean that most of the deposition occurs directly below the cages, so adding cages to the farm will increase the area on the seabed directly below the farm where deposition occurs. In addition, when considering the 99th Percentile of Solids Flux as the output, the Biomass was identified as being the most influential at both sites. This makes sense, as increasing the Biomass produces more waste, meaning the deposition will be more intense. As increasing Cage Diameter and the Number of Cages are two alternative approaches to allow for increased Biomass, they will both be

considered in the next steps, where the inputs based on the physical properties and the operational inputs are combined for a further sensitivity analysis.

By creating the sampling design using the approach described previously, this creates an imbalance. As a result, there is the possibility of the relationships between the inputs affecting the results. Therefore, the results are considered with caution and future work may consider a single input relating to the operational setup of the farm, rather than the three inputs described here.

2.4 Sensitivity Analysis - Inputs Based on the Physical Properties and Operational Inputs

Within DEPOMOD, altering the inputs based on the physical properties and the operational inputs individually impacts the calculations of the scalar outputs. Combining the effects of the two input sets will be important for simulating farms with extended production capabilities in the future.

2.4.1 Inputs and their ranges

The previous sensitivity analyses provided useful information that will be used to reduce the number of inputs being used. From the sensitivity analysis of the inputs based on the physical properties, the dispersion coefficients were consistently ranked as the least important inputs for all of the scalar summary outputs. Six of the eleven inputs based on the physical properties are the dispersion coefficients, so removing these will help improve efficiency of this combined analysis. For the inputs based on the physical properties that remain, the range of values they can take will remain the same.

The discrete operational inputs from the previous analysis will all be used, however, the number of options for each will be reduced. In the analysis of the operational inputs, a total of 18 and 23 combinations were considered for Ardentinny and Muck. These totals included different cage setups based on varying levels for the maximum Stocking Density. In order to reduce the number of slices (and therefore the total number of runs), only one maximum Stocking Density will be considered - 16.3kg/m^3, which is the most restrictive. Following the process set out in the flowchart in Figure 2.27, no alterations were made to the farm set up until Biomass was increased by 50%, therefore, the 10% increase of Biomass was removed from this analysis to also reduce the number of

runs required. By removing the 10% increase in Biomass from the operational setups, this meant that savings could be made in the computational time. As with the previous analysis, the three operational inputs are considered, but this will be reviewed later in the thesis to determine if a single input representing the operational setup is more suitable.

2.4.2 Sampling Design

This sensitivity analysis includes a combination of continuous and discrete inputs. Therefore, the sampling design will have to be chosen to reflect this. Qian & Wu (2009) introduced methods for creating space-filling designs for quantitative and qualitative inputs in two steps:

1. For quantitative inputs, LHS is generated based on a sliced orthogonal array. It is then partitioned into different groups, where points in each group achieve good space-filling properties in low dimensions.

2. Different level combinations of the qualitative inputs are then associated with the groups.

This approach essentially creates multiple LHS for the continuous inputs and assigns each a combination of the discrete inputs to one of the LHS. The method created by Qian & Wu (2009) was studied further and Ba et al. (2015) expanded it to create an optimal sliced LHS using the maximin-distance approach, which was used to create a package and implement it in R.

Ba et al. (2015) describes how the construction of the sliced LHS can be completed in two steps. To construct a sliced LHS, the total number of samples, n, is calculated based on the number of slices (equal to the number of discrete inputs, d), and the number of samples within each slice, s, with $n = sd$. The final element involved in the construction of the sliced LHS is the number of continuous outputs, c, with the two steps for constructing it described as follows:

1. Construct d independent LHS for each discrete input, $\mathbf{D}_1, \ldots, \mathbf{D}_d$, containing s points for c inputs. Denote their factor levels by $1, \ldots, s$, and the samples should then be stacked to produce an $(n \times c)$ matrix, $\mathbf{D} = \cup_{i=1}^{d} \mathbf{D}_i$.

2. Independently, in each column of the matrix, $\mathbf{D}$, replace the d entries of level $l = 1, \ldots, s$, with a random permutation, $\mathbf{\Pi}_d$,of elements $\{(l-1)d + 1, \ldots, ld\}$.

The above method described by Ba et al. (2015) differs from the original approach by Qian & Wu (2009) which generated the whole design using a column by column approach. Ba et al. (2015)'s next step was to improve the space-filling qualities of the design using the maximin-distance criteria (Johnson et al. 1990). The sliced LHS approach produces an overall LHS, $\mathbf{S}$, and the smaller LHS for each slice, $\mathbf{S}_1, \ldots, \mathbf{S}_d$, therefore the maximin distance criteria will have to satisfy all LHS. The work of Johnson et al. (1990) was extended, with the aim being to minimize the average reciprocal interpoint distance of the design $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ (Morris & Mitchell 1995, Jin et al. 2005):

$$\phi_r(\mathbf{X}) = \left(\frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} \frac{1}{d(\mathbf{x}_i, \mathbf{x}_j)^r} \right)^{1/r}. \tag{2.10}$$

Here, $d(\mathbf{x}_i, \mathbf{x}_j)^r$ is some distance measurement such as Euclidean distance. Minimizing ϕ_r when $r \to \infty$, is equivalent to maximizing the minimum distance between the design points (Ba et al. 2015). Extending Equation 2.10 for evaluating the space-filling qualities of the design requires $\phi_r(\mathbf{D})$ to be minimized for all design points, as well as minimizing $\phi_r(\mathbf{D}_i)$ for each slice $(i = 1, \ldots, d)$ (Ba et al. 2015). A single objective function was proposed by Ba et al. (2015) to solve the optimization problem:

$$\phi_{Mm}(D) = \frac{1}{2} \left(\phi_r(\mathbf{D}) + \frac{1}{d} \sum_{i=1}^{d} \phi_r(\mathbf{D}_i) \right). \tag{2.11}$$

The optimal sliced LHS is therefore the design that minimizes $\phi_{Mm}(D)$. This method can be applied in R using the package 'SLHD' (Ba 2015).

The sampling design is then setup for the analysis to be completed at the fish farm sites. To create the sampling design, multiple choices have to be made to allow accurate conclusions to be made and to keep the total runtime to a minimum. First, determining the number of slices required, it was previously mentioned that for this analysis, the most restrictive maximum Stocking Density (16.3kg/m^3) will be used to determine the operational inputs required. After removing the 10% increase of Biomass, for Ardentinny, there are a total of eight different farm setups to be considered, and therefore, a total of eight slices in the sampling design. The next choice for the analysis is the number of samples for each slice of the design. In order to cover the sample space for the five continuous inputs, a total of 50 samples will be taken for each slice, resulting in 400 input sets within the analysis. In addition, NewDEPOMOD is run 50 times for each input set to account for the random walk element within

NewDEPOMOD, resulting in 20,000 runs in total. The number of replicate runs was reduced from 100 in the previous sensitivity analyses to 50 to reduce the computational time, but still account for the random walk element of NewDEPOMOD. Running NewDEPOMOD this many times is computationally expensive, taking approximately 10 days to complete the runs, which is why the number of samples within each slice was limited to 50. Following the creation of the optimal sliced LHS, the restricted pairing procedure that was used previously for the correlated LHS, was applied to account for the relationships between the continuous inputs.

The other low energy site being considered is West Strome, which had a total of 9 different farm setups to be considered, when following the flowchart in Figure 2.27. This resulted in a total of 450 different input sets, and 22,500 runs when considering the replicate runs. For the high energy sites, Muck and Djuba Wick, there are 9 and 8 different farm setups to be considered. One difference for these sites though, is that the computational time for the runs can be up to five times longer, due to the more complex waste transportation within the model due to the faster flow speed. To reduce the computational time required, the number of replicate runs at these sites was reduced from 50 to 10.

2.4.3 Results for combined analysis

As with the previous analyses, the scalar summary outputs from the NewDEPOMOD runs will be considered to assess the size and intensity of the environmental impact. This will allow the combined effects of the operational inputs and the inputs based on the physical properties on the Total Area Impacted and the 99th Percentile of Solids Flux to be considered. For this analysis, the output that will be discussed will be from simulations at the low energy sites Ardentinny and West Strome. Additionally, the high energy sites Muck and Djuba Wick will be considered for comparison. A detailed analysis will be considered for Ardentinny, before a summary of the results for the remaining sites are reviewed.

2.4.3.1 Total Area Impacted

For the NewDEPOMOD runs at Ardentinny, the sampling design resulted in a total of eight different operational setups, and therefore eight slices in the sliced LHS. Within each slice, there were 50 input sets produced, resulting in a total of 400 different input sets across the eight slices. Considering the Total

Area Impacted as the output, the range of values will be considered through a histogram. It is clear from Figure 2.39 that almost all of the data lies in the

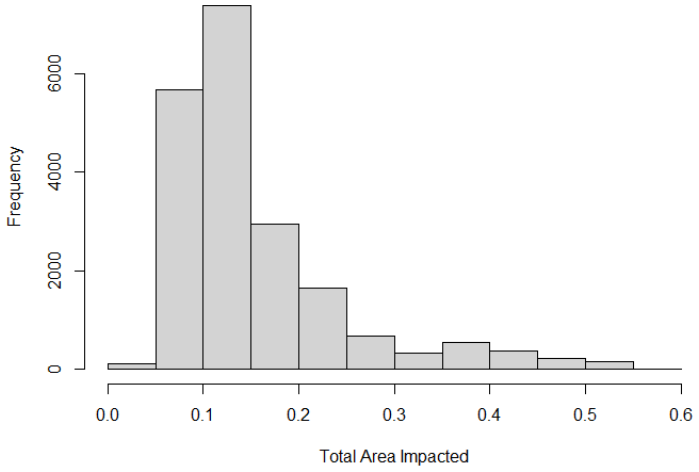


Figure 2.39: Histogram of Total Area Impacted - Ardentinny.

interval $0.05 - 0.20\text{km}^2$. This is confirmed when calculating the interquartile range, which lies between 0.097 and 0.172. The next step is to fit a random forest model for the Total Area Impacted. The fitted model is able to explain approximately 94% of the variation in the data, and the top five ranking inputs are given in Table 2.12. Table 2.12 highlights that in the combined sensitivity

Inputs	**Importance**
Settling Velocity of Faeces	217.8%
Settling Velocity of Sediment	108.7%
Critical Shear Stress for Erosion	86.4%
Number of Cages	85.4%
Biomass % of Default	83.7%

Table 2.12: Table of Importance values from the random forest model of Total Area Impacted at Ardentinny.

analysis, the Settling Velocity of Faeces is the most important input, with an importance value more than double that of the second most important input. The Number of Cages is the most important of the operational inputs, which fits in with the previous analysis of the operational inputs, however, the difference between the importance value for the Number of Cages and the Biomass is low. In order to explore the relationship between the Total Area Impacted and the Settling Velocity of Faeces further, a scatterplot is produced. Figure 2.40 does not appear to show any clear relationship between the Total Area Impacted

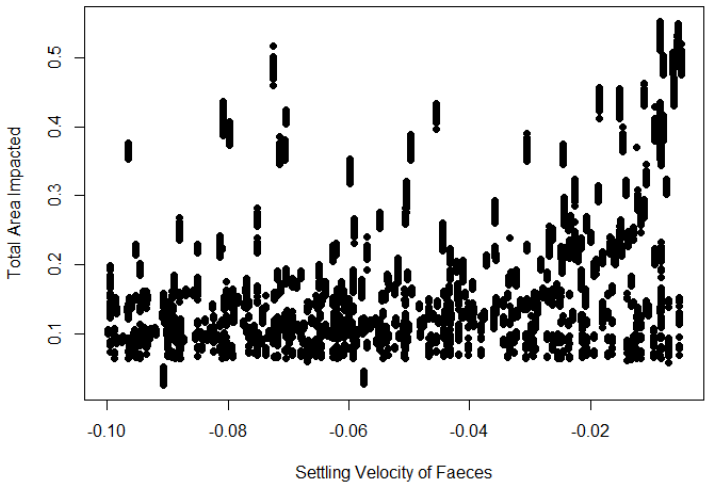


Figure 2.40: Plot of the Total Area Impacted against the Settling Velocity of Faeces - Ardentinny.

and the Settling Velocity of Faeces. It can be seen that some large values for Total Area Impacted are present across the range of values for Settling Velocity of Faeces. This could indicate that the importance of this input is influenced by it's interactions which are considered in the model.

In order to explore the similarities/differences between the low and high energy sites, random forest models were fitted for the Total Area Impacted for each of the remaining sites. These were all able to explain over 90% of the variation in the data, and the resulting importance values for each of the inputs are given below for the four sites. First, considering the two low energy sites, Ardentinny and West Strome, the Settling Velocity of Faeces is the most important input at both sites. However, the effect of the Settling Velocity of Sediment on the Total Area Impacted for West Strome is ranked lower. For the operational inputs, the Biomass and Number of Cages have similar importance values at Ardentinny, whereas the Number of Cages plays a more dominant role at West Strome.

Next, the high energy sites, Muck and Djuba Wick, will be considered. First, looking at the importance values for Muck, the importance values for the operational inputs are the lowest ranking. The Settling Velocity of Faeces and the Critical Shear Stress for Erosion are the top ranking inputs. At Djuba Wick, the Settling Velocity of Faeces is also the highest ranking input, but with the Critical Shear Stress for Erosion not playing as big a role, with a similar importance value as the other inputs. The importance values for Djuba Wick are much lower than the values for the other sites, indicating that the inputs

Inputs	Importance			
	Ardentinny	**West Strome**	**Muck**	**Djuba Wick**
Critical Shear Stress for Erosion	86.4%	58.2%	76.3%	35.6%
Rate of Erosion	80.9%	47.3%	59.6%	30.7%
Release Height	80.0%	54.8%	50.7%	24.8%
Settling Velocity of Faeces	217.8%	248.9%	92.9%	94.6%
Settling Velocity of Sediment	108.7%	58.1%	68.7%	25.8%
Biomass % of Default	83.8%	71.1%	40.3%	26.7%
Cage Diameter	64.6%	51.7%	42.1%	19.2%
Number of Cages	85.4%	87.4%	37.6%	31.5%

Table 2.13: Table of Importance values from the random forest model for the Total Area Impacted at each site.

do not have as big an influence at this site. Djuba Wick has very high current speeds in comparison to the others and it is likely that the explanation for this is that the current speeds play such a big role that the inputs are not as influential.

Comparing the results from the low and the high energy sites, the Settling Velocity of Faeces plays a dominant role across all sites, with the highest importance values. However, at Muck, the importance value for the Settling Velocity of Faeces is closer to the next ranked input. Djuba Wick appears to have similar patterns to the low energy sites, with Settling Velocity of Faeces having an importance value more than two times the size of the second highest ranking input. It would be expected that with low current speeds, resuspended waste will not be transported as far, meaning that the initial deposition is likely to have a bigger influence at the low energy sites. As a result, the Total Area Impacted will be influenced more by the Settling Velocity of Faeces at the low energy sites. However, at Djuba Wick, the inputs appear to have a similar pat-

tern in their importance values. At Muck, the operational inputs have lower importance values than the inputs based on the physical properties. As mentioned before, there are no similarities between the importance values for the high energy sites.

2.4.3.2 99th Percentile of Solids Flux

Next, the 99th Percentile of Solids Flux is considered and plotted as a histogram for data from the runs at Ardentinny. Figure 2.41 has a more even spread of the

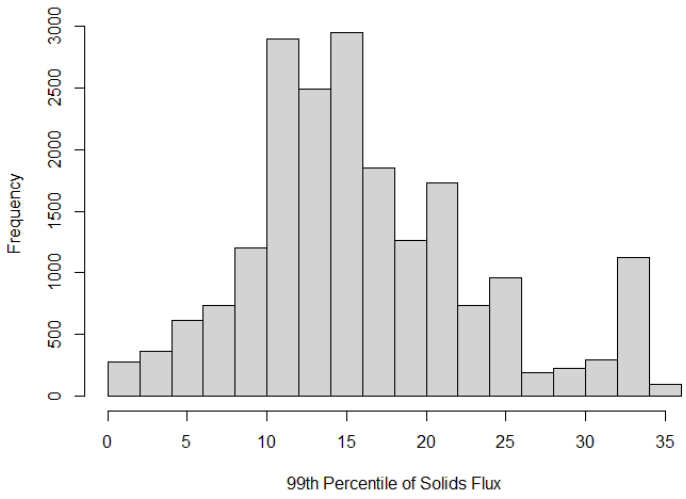


Figure 2.41: Histogram of 99th Percentile of Solids Flux - Ardentinny.

data than Figure 2.39. Random forest models are fitted for each of the sites, which all explain over 90% of the variation in the data. Table 2.14 provides the importance values for the at each of the sites. Considering the two low energy sites, Ardentinny and West Strome, the highest ranking input at both is the Settling Velocity of Faeces. This is where the similarities then end, with the inputs linked to resuspension at Ardentinny (Critical Shear Stress for Erosion, Release Height, Rate of Erosion, and Settling Velocity of Sediment) being the next highest ranking inputs, before the operational inputs. In contrast, at West Strome, the second and third ranked inputs are the Number of Cages and the Biomass, both of which have similar roles.

For the high energy sites, Muck and Djuba Wick, the Settling Velocity of Faeces is the highest ranking input at Muck, but at Djuba Wick, the Critical Shear Stress for Erosion has a slightly higher importance value than the Settling Velocity of Faeces. For the 99th Percentile of Solids Flux, there appears to be some more similarities between the two high energy sites, with the Crit-

Inputs	Importance			
	Ardentinny	**West Strome**	**Muck**	**Djuba Wick**
Critical Shear Stress for Erosion	97.2%	77.7%	61.8%	58.2%
Rate of Erosion	82.7%	73.0%	42.0%	32.9%
Release Height	83.1%	70.1%	51.1%	30.5%
Settling Velocity of Faeces	129.0%	171.5%	73.5%	53.8%
Settling Velocity of Sediment	78.7%	71.7%	53.4%	31.5%
Biomass % of Default	73.4%	85.3%	45.2%	32.2%
Cage Diameter	61.5%	56.8%	35.7%	31.1%
Number of Cages	73.1%	86.1%	37.3%	33.8%

Table 2.14: Table of Importance values from the random forest model for the 99th Percentile of Solids Flux at each site.

ical Shear Stress for Erosion and Settling Velocity of Faeces being the highest ranking inputs with importance values slightly bigger than the other inputs.

Comparing the low and the high energy sites, the Settling Velocity of Faeces appears to play a more dominant role at the low energy sites, where the importance values are much larger than for the other inputs. At the high energy sites, the Settling Velocity of Faeces and the Critical Shear Stress for Erosion are the highest ranking inputs, with similar importance values. The other inputs at the high energy sites have slightly lower importance values than the values for the Critical Shear Stress for Erosion and the Settling Velocity of Faeces. At West Strome, the Settling Velocity of Faeces has an importance value more than two times the size of the next highest ranking input. The difference between the importance values at Ardentinny is not as big as at West Strome. In addition, the operational inputs appear to play a bigger role at West Strome, with the Biomass and Number of Cages having the second and third highest importance values.

2.5 Discussion

The main aim of this Chapter was to consider the inputs based on the physical properties that were identified as having uncertain default values, as well as the operational inputs that could be altered to allow for expansion at farms, in order to identify the inputs with the most influence on the scalar outputs used to measure the environmental impact of a fish farm. The first analysis featured only two sites, and only considered the inputs based on the physical properties. From this analysis, differences were identified between the two sites relating to the influence of the resuspension inputs, such as the Critical Shear Stress for Erosion, which played a bigger role at the high energy site compared to the low energy site across all of the outputs being considered.

Moving on, two sites were again considered to focus on altering the operational inputs which relate to the farm setup and can be altered by the fish farm operator. Differences were identified between the two sites when considering the Total Area Impacted as the output. Moreover, the 99th Percentile of Solids Flux identified the Biomass as the most important operational input at both sites.

Both the previous analyses were used as a starting point to identify inputs to be investigated, before considering the two sets of inputs together in a combined analysis, looking at 4 sites in total - containing two low energy and two high energy sites. Two low energy and two high energy sites were considered, with the aim of comparing between sites with similar properties and contrasting properties. For both of the outputs, the two low energy sites identified the Settling Velocity of Faeces as being the most important, but at Ardentinny, the influence of the resuspension inputs was larger in comparison to West Strome. For the 99th Percentile of Solids Flux, at West Strome, the Settling Velocity of Faeces had a much bigger importance value in comparison to the others, and the operational inputs also played a bigger role. At the high energy sites, for the Total Area Impacted, the Settling Velocity of Faeces is also the top ranked input, with the Critical Shear Stress for Erosion playing a bigger role at Muck. In addition, the structure of the ranking at Djuba Wick is similar to the low energy sites. For the 99th Percentile at the high energy sites, the Critical Shear Stress for Erosion and the Settling Velocity of Faeces are the top ranking inputs, with similar importance values that are bigger than the other inputs. For the Total Area Impacted, the low energy sites and Djuba Wick identified similar patterns from the sensitivity analysis, with differences seen in the results for Muck, where the Critical Shear Stress for Erosion playing a bigger role. Considering the 99th Percentile, the results for the high energy sites and

Ardentinny have a similar structure, with West Strome having different results with the operational inputs playing a bigger role. Differences were identified between the sites with similar characteristics for each output. This suggests that the influence of the inputs differs for sites with the same characteristics, but the inputs that were identified as being most important across the sites were the Critical Shear Stress for Erosion and the Settling Velocity of Faeces.

Chapter 3

Sensitivity Analysis for Output Maps

3.1 Introduction

Having considered scalar summaries of the NewDEPOMOD output, the next steps will consider the multivariate output, i.e. the maps. The sensitivity analysis of these output maps is an important aspect in determining how influential the inputs are over the domain, therefore the development of methodology to assess the influence is essential. Throughout the chapter, three different approaches to investigating the multivariate output will be developed: 1) shape analysis of the main impact area, 2) a bivariate functional approach, where the output map is considered as a surface and 3) considering the output from individual grid cells.

The focus of this Chapter is to extend the work of Chapter 2 in order to investigate the effect of the inputs on the NewDEPOMOD output maps. The three individual approaches to how the output maps will be investigated provide useful information, but a framework is required in order to investigate the features of the map and the variation in the data. This framework will be presented in this Chapter, along with an application of the framework to multiple sites.

3.1.1 NewDEPOMOD output maps

As was mentioned in the introduction to this chapter, different representations of the output maps will be considered. The maps represent the NewDEPOMOD output showing the Solids Flux in each grid cell, and therefore approaches to analyse these maps are essential to gaining a better understanding

of NewDEPOMOD. It was mentioned in Chapter 1 that Solids Flux is a measure of the deposition on the seabed, and NewDEPOMOD output maps provide measures of Solids Flux across the domain. The data featured in this Chapter is the same data that were used in the previous Chapter, but with a focus on the output maps instead of the scalar outputs. In order to highlight the types of variation seen in the output maps, two examples of output maps from the runs at Ardentinny are given in Figures 3.1 and 3.2. Differences can be

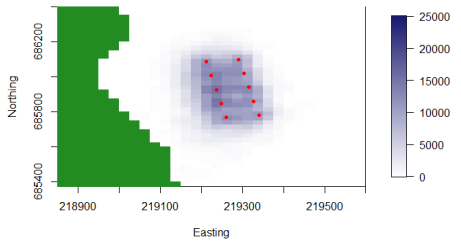


Figure 3.1: NewDEPOMOD output map of the Solids Flux $(g/m^2/y)$, Example 1 from Ardentinny- land indicated by green grid cells and cages indicated by red points.

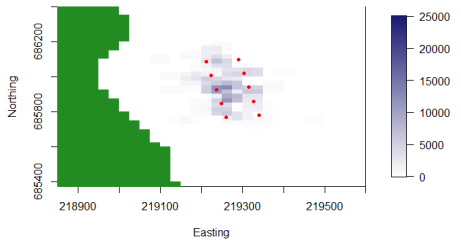


Figure 3.2: NewDEPOMOD output map of the Solids Flux $(g/m^2/y)$, Example 2 from Ardentinny - land indicated by green grid cells and cages indicated by red points.

seen here in the shape of the main impact on the seabed. In Figure 3.1, the main area of deposition is directly below the cages and appears to be in an ordered fashion. In comparison, Figure 3.2 shows less structure in the pattern of deposition, with some deposition spreading to the West of the cages. This Chapter will therefore investigate ways to attribute the variation in the maps to changes in the inputs. An additional point to be highlighted from Figures 3.1 and 3.2 is that a lot of the area in the domain features no deposition.

3.2 Shape analysis approach for investigating NewDEPOMOD output maps

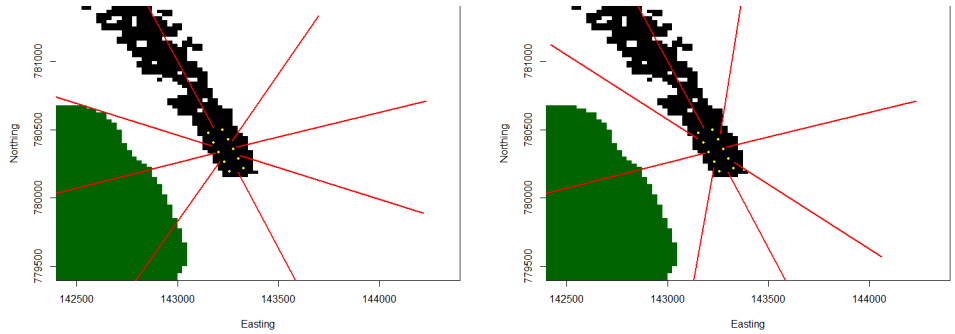
As a starting point for analysing the effect of altering the inputs on the output maps, a shape analysis will be considered. This approach does not consider the full domain, and only looks at the main shape of the impact on the seabed by identifying landmarks using transects from the farm centre. A shape Principal Components Analysis ('PCA') (Dryden & Mardia 2016) will be used to identify

the main areas of variation in the shape, and the PC scores used to identify which inputs are driving the variations.

3.2.1 Landmark approach for identifying predicted size and shape of the impact area on the seabed

A finite number of points on an object can be used to describe a shape, and are commonly known as landmarks. These are points of correspondence on each object that can be matched between and within interested populations (Dryden & Mardia 2016). Landmarks are often labelled to allow comparisons to be made between different shapes within a dataset. Shape analysis has long been used in Biology by comparing distances between landmarks. Pearson (1926) looked at similarities between skulls using the distances between landmarks.

To pick the landmarks for these NewDEPOMOD maps, 8 transects were taken from the centre of the farm, illustrated in Figure 3.3. The 8 transects can be taken using two different methods: (i) where the angles between the transects are equal (Figure 3.3a) and (ii) where the distance between the transects on the perimeter of the farm are equal (Figure 3.3b). Figure 3.3 illustrates how



(a) Transects based on equal angles between each transect.

(b) Transects based on equal distances around farm perimeter.

Figure 3.3: Plots of the possible transect options for calculating landmarks in the shape analysis.

the different methods of producing the transects will produce different landmarks. In this case, the majority of impact appears to be in a more narrow column, and so the equal distances around the farm perimeter in Figure 3.3b appears to capture this more effectively. As a result, these transects will be used to calculate the landmarks, and the next decision is how to determine what grid cells are landmarks on each transect.

The aim of the shape analysis is to capture the main shape of the impact,

and so, only grid cells with a Solids Flux value greater than 192g/m^2/year are considered - it was mentioned in Chapter 1 that this value is considered as a threshold, above which damage to the seabed can occur. In order to try and obtain landmarks that lie on the outline of the main shape, a nearest neighbour approach was considered. Each grid cell is bordered by 8 other grid cells. Using a nearest neighbour approach, a grid cell must be surrounded by at least five grid cells with Solids Flux values greater than the threshold of 192g/m^2/year, to be considered part of the main shape. Five grid cells were used as a cut-off point for the main shape as this meant that for a given grid cell, most of the surrounding grid cells are above the threshold, indicating that it is part of a larger shape.

Code that was created to calculate the landmarks for each run automatically was computationally expensive, with the time taken for one run being approximately $3s$. In order to simplify this analysis, the data being used is the NewDEPOMOD output maps from the sensitivity analysis of the inputs based on the physical properties, completed at the sites Ardessie and Muck. These analyses contained a total of 10,000 runs, made up of 100 different input sets with 100 replicate runs for each of them. Running the code to automatically calculate the landmarks for each run could take approximately 8 hours, so this was reduced by only considering 1 run from each set of 100 replicates for an initial analysis.

3.2.2 Procrustes and principal component analysis for analysing shape variation

Procrustes Analysis is often used as a tool for comparing shapes as it removes the effects of scale, rotation and translation (location). Within Procrustes Analysis, there are two different types that can be used in different scenarios (Dryden & Mardia 2016):

- **Ordinary Procrustes Analysis ('OPA')** - to be used in the case where one shape is compared to another shape, or where an arbitrary reference shape is being used, a set of shapes are compared to this reference shape.

- **Generalised Procrustes Analysis ('GPA')** - to be used when a set of objects are to be compared simultaneously, with a so-called 'mean shape' being produced as a reference shape.

GPA is defined as the translation, rescaling and rotation of the shape configurations $(X_1, X_2, \ldots, X_n)$ relative to each other, to minimize a total sum of

squares (Dryden & Mardia 2016):

$$G(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \parallel (\beta_i X_i \Gamma_i + 1_k \gamma_i^\top) - \mu \parallel^2,$$

with respect to $\beta_i, \Gamma_i, \gamma_i$, for $i = 1, \ldots, n$ and μ, subject to an overall size constraint that is chosen. $\beta_i > 0$ refers to a scale parameter, Γ_i is a rotation matrix, γ_i is a location vector and μ is the population mean shape. One measure of the shape for Procrustes analysis is the 'centroid size', which is a measure of the size of the shape, calculated using the landmark coordinates. For a matrix of landmark coordinates X, with $k \times m$ dimensions (where k is the number of landmarks in m real dimensions), the centroid size is given by (Dryden & Mardia 2016):

$$S(X) = \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{m} (X_{ij} - \bar{X}_j)^2}, X \in \mathbb{R}^{km},$$

where X_{ij} is the (i, j)th element of X, and $\bar{X} = \frac{1}{k} \sum_{i=1}^{k} X_{ij}$ is the arithmetic mean in the jth dimension (Dryden & Mardia 2016). The centroid size is a commonly used measure of size in geometrical shape analysis.

In addition to the measures of centroid size, it is also of interest to consider the structure of size and shape variability (Dryden & Mardia 2016).

3.2.3 Results from shape analysis

For the sensitivity analysis of the physical properties inputs, the landmarks and shape analysis approach was considered. This would produce PC scores, which could be modelled to determine which inputs are potentially driving the variations in the shape.

3.2.3.1 Ardessie

The Generalised Procrustes analysis calculated the centroid size and the PCs. The first 7 PCs were able to explain approximately 89% of the variation, with the breakdown shown in Table 3.1. Table 3.1 identifies the first PC as explaining 36.0% of the variation in the shape of the impact. Figure 3.4 illustrates the shape variation that is described by the first three PCs. It is difficult to tell from Figure 3.4 the variation that the PCs are describing. Each PC appears to describe some information relating to the width of the shape, which can be seen when comparing the shapes in the first column to the shapes in the third column. In addition, the second PC appears to be describing the variation of

Principal Component	Percentage of Variability Captured
PC 1	36.0%
PC 2	23.9%
PC 3	10.1%
PC 4	7.3%
PC 5	4.9%
PC 6	4.1%
PC 7	3.1%
Total	89.4%

Table 3.1: Table of the Principal Component percentages for Solids Flux 192g/m^2/year - Ardessie.

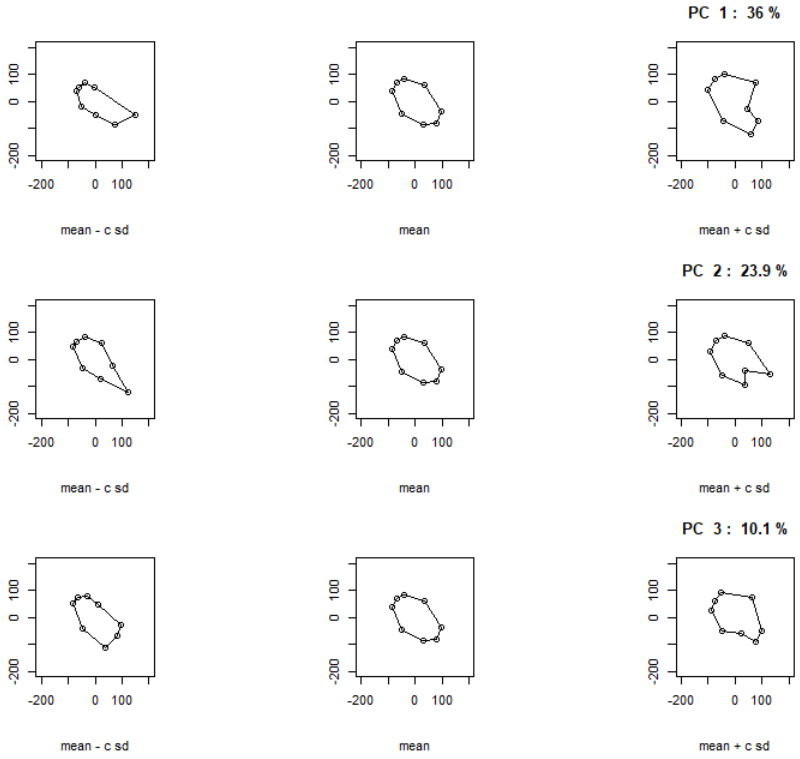


Figure 3.4: Plots of the shape variation described by the first 3 PC's - Ardessie.

the shape in the transect that runs in the South-East direction, with the plot in the first column having this point located further away from the centre.

Using the PCs that have been calculated, and also the centroid sizes, models will be fitted to determine which, if any, input factors are related to the variation in the shapes. As a starting point, individual linear models were fitted to the centroid size and the PC's to determine the effects of the input factors on the shape. A standard multiple regression formula with interactions is given in

Equation 3.2.

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \gamma_1 x_{1i} x_{2i} + \ldots \tag{3.1}$$

$$+ \gamma_Q x_{(p-1)i} x_{pi} + \epsilon_i \qquad \text{for } i = 1, \ldots, n. \tag{3.2}$$

Here, there are interaction terms between each of the p inputs, and $Q = p(p-1)/n$ interaction terms when including all possible two-way interactions. In this case, the prior knowledge of the correlation structure reduces the number of two-way interactions included in the general formula in Equation 3.2. Two-way interaction terms were only included for the inputs that were considered to be correlated when setting up the design matrix using the correlated LHS in Chapter 2. Table 3.2 provides information about the fit of the models, and the input factors that have a significant effect on the outputs. Table 3.2

Model Output	**R-Squared**	**Significant Input Factors**
Centroid size	67.8%	Intercept RH RH:SS SF DispCageZ
PC 1 (36.0%)	29.4%	RoE
PC 2 (23.9%)	15.5%	CSS
PC 3 (10.1%)	23.2%	DispCageX DispCageX:DispCageY
PC 4 (7.3%)	16.7%	CSS:RoE SF
PC 5 (4.9%)	22.6%	SF
PC 6 (4.1%)	3.8%	RoE CSS:RoE
PC 7 (3.1%)	5.1%	SF

Table 3.2: Linear model output for the shape analysis where Solids Flux 192g/m^2/year - Ardessie. *(CSS - Critical Shear Stress for Erosion, RoE - Rate of Erosion, RH - Release Height, SS - Settling Velocity of Sediment, SF - Settling Velocity of Faeces, DispCageX/Y/Z - Dispersion Coefficient for Material from cages (X/Y/Z directions)*

appears to show a reasonably good fit for the linear model of Centroid size, and identified input factors that are significant in the linear model. Settling Velocity of Faeces is significant, which would be expected given it's influence on the Total Area Impacted in the sensitivity analysis for the physical properties inputs in Chapter 2. One thing to note is that some of the Dispersion coefficients are significant, when they did not appear to be ranked highly in the random forest models for Ardessie in Chapter 2. The reason they may affect the shape of

the impact may be down to the fact that flow speeds are slower at this site and therefore, the random walk element of the model is playing a role in the shape of the impact. The linear models fitted to the PCs do not appear to fit the data well, but have identified significant input factors, and more flexible models may be more appropriate. As with the fitting of any linear models, residual plots were checked, with some deviances from the line of equality in the Normal Q-Q plot at the tails, but there was generally agreement with the standard assumptions.

In order to investigate possible non-linear patterns, Gaussian GAMs were fitted to the Centroid sizes and the PCs. An example of a GAM with interaction terms is given in Equation 3.4.

$$y_i = \beta_0 + f_1(x_{1i}) + \ldots + f_p(x_{pi}) + g_1(x_{1i}, x_{2i}) + \ldots \tag{3.3}$$

$$+ g_Q(x_{(p-1)i}, x_{pi}) + \epsilon_i \qquad \text{for } i = 1, \ldots, n \tag{3.4}$$

As with the linear model approach, the number of interaction terms within the GAM were reduced based on the prior knowledge of the correlations between the inputs. The results from the GAMs are given in Table 3.3. Table 3.3

Model Output	**Deviance Explained**	**Significant Input Factors**
Centroid size	73.3%	(CSS, RoE) - $edf = 2$ SF
PC 1 (36.0%)	41.1%	(CSS, RoE)
PC 2 (23.9%)	47.5%	(CSS, RoE) (RH, SS) - $edf = 2$
PC 3 (10.1%)	63.0%	(CSS, RoE) SF
PC 4 (7.3%)	36.9%	SF
PC 5 (4.9%)	66.9%	(CSS, RoE) SF - $edf = 1$ (DispCageX, DispCageY) DispCageZ
PC 6 (4.1%)	16.2%	No input factors with significant p-value
PC 7 (3.1%)	39.1%	SF (DispCageX, DispCageY)

Table 3.3: GAM model output for the shape analysis where Solids Flux 192g/m^2/year - Ardessie. *(CSS - Critical Shear Stress for Erosion, RoE - Rate of Erosion, RH - Release Height, SS - Settling Velocity of Sediment, SF - Settling Velocity of Faeces, DispCageX/Y/Z - Dispersion Coefficient for Material from cages (X/Y/Z directions)*

confirms that the added flexibility of the models has improved the model fit,

which is to be expected. Critical Shear Stress for Erosion, Rate of Erosion and Settling Velocity of Faeces appear to be the ones that are significant most often in the model. In some instances, the estimated degrees of freedom (edf) = 1 or 2, indicating that a flexible function in the model is not be required. Where the edf is not quoted in Table 3.3, this suggests that a flexible function was appropriate.

From fitting models to the PCs and Centroid size, the input factors that appear to have the biggest influence on the shape of the impact are Critical Shear Stress for Erosion, Rate of Erosion and Settling Velocity of Faeces. They were also ranked highest when considering the Total Area Impacted, 99th Percentile of Solids Flux and Mass Balance as the outputs at Ardessie in Chapter 2.

3.2.3.2 Muck

As with Ardessie, the Procrustes analysis was used to calculate centroid size and PCs. At this site, the first four PCs explain 93.5% of the variation, with a breakdown shown in Table 3.4. The larger amount of variation described by

Principal Component	**Percentage of Variability Captured**
PC 1	59.0%
PC 2	22.0%
PC 3	7.5%
PC 4	5.0%
Total	93.5%

Table 3.4: Table of the Principal Component percentages for Solids Flux 192g/m^2/year - Muck.

the first 4 PCs indicate that the shapes are more consistent in their variation. The majority of the variation in the shapes is described by the first PC. The variation described by the first three PCs is illustrated in Figure 3.5. From Figure 3.5, it is clear that the first PC describes the variation in the length of the impact shape in the North-West direction, and also some variation in the width of the shape. The other two PCs appear to also describe some variation in the width of the shape.

Following the same method as was used for Ardessie, individual linear models were fitted to the centroid size and the first four PCs, with information about the model output provided in Table 3.5. The fit of the models are poor, with relatively low R-squared values, especially for PC 2. In terms of the significant input factors, none of the Dispersion Coefficients appear to have an effect on the shape, with the other input factors being significant on more than one

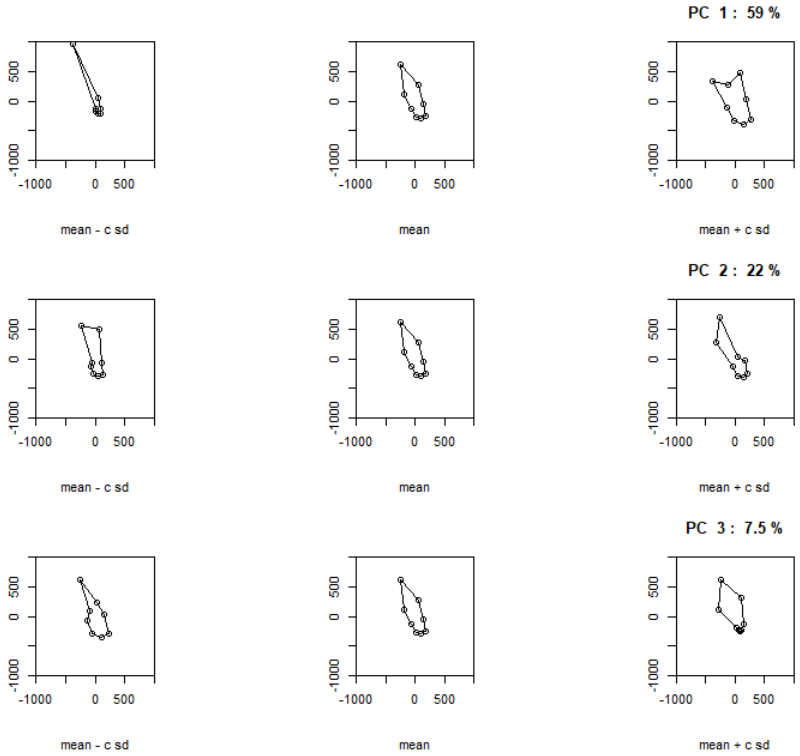


Figure 3.5: Plots of the shape variation described by the first 3 PC's - Muck.

Model Output	**R-Squared**	**Significant Input Factors**
Centroid size	32.6%	Intercept SF
PC 1 (59.0%)	33.0%	Intercept CSS RoE CSS:RoE SF
PC 2 (22.0%)	14.0%	No input factors with significant p-value
PC 3 (7.5%)	36.9%	SS RH:SS
PC 4 (5.0%)	32.4%	CSS:RoE SS RH:SS SF

Table 3.5: Linear model output for the shape analysis for Solids Flux 192g/m^2/year - Muck. *(CSS - Critical Shear Stress for Erosion, RoE - Rate of Erosion, RH - Release Height, SS - Settling Velocity of Sediment, SF - Settling Velocity of Faeces)*

occasion. Considering the poor R-squared values, it may be appropriate to consider more flexible models.

As with Ardessie, GAMs were fitted to the centroid size and PCs, with

the output information provided in Table 3.6. As expected, the more flexible

Model Output	Deviance Explained	Significant Input Factors
Centroid size	48.1%	(CSS, RoE) - $edf = 2$ SF
PC 1 (59.0%)	60.6%	(CSS, RoE) (RH, SS) SF - $edf = 1$
PC 2 (22.0%)	41.4%	(CSS, RoE)
PC 3 (7.5%)	60.0%	(RH,SS)
PC 4 (5.0%)	59.2%	(CSS, RoE) (RH, SS) SF

Table 3.6: GAM model output for the shape analysis for Solids Flux 192g/m^2/year. *(CSS - Critical Shear Stress for Erosion, RoE - Rate of Erosion, RH - Release Height, SS - Settling Velocity of Sediment, SF - Settling Velocity of Faeces)*

models appear to fit the data better. Again, none of the Dispersion Coefficients are significant in any of the models. the combined term of Critical Shear Stress for Erosion and Rate of Erosion is significant in 4 of the 5 models, suggesting that they have a big influence on the shape of the impact. The Release Height and Settling Velocity of Sediment feature as significant inputs in Table 3.6, whereas they only feature as significant for the first PC in Table 3.3. This indicates that they have a bigger influence on the shape of the impact at Muck compared to Ardessie, which is most likely down to the fact that resuspension plays a bigger role at the faster flowing site.

3.2.4 Review

For the physical properties inputs, the shape analysis was able to identify which parameters were likely to be causing variations in the shapes of the impact. There were some common results between the two sites, with the erosion parameters appearing to be influential, and the dispersion coefficients not featuring a great deal. The main difference between the high and low energy sites was that the Release Height and Settling Velocity of Sediment appeared to be more influential at the high energy site, which could be a result of the larger amounts of resuspension taking place.

The shape analysis approach is a worthwhile consideration in the case where the impact on the seabed has a main shape. This is not always the case, and the landmark approach may not always produce a good representation of the

impact. Therefore, the next stage of the analysis will consider the output maps in more detail using different approaches.

3.3 Bivariate functional analysis approach for investigating NewDEPOMD output maps

Functional data analysis ('FDA') is a method used to analyse data providing information over a curve or a surface. For $i = 1, \ldots, n$, and $t \in T$, where T is a real interval, FDA relates to data where the ith observation is areal function $x_i(p)$, where each x_i is a point in some function space, P (Ramsay & Dalzell 1991). Therefore, a functional data approach can be considered for the analysis of the output maps, where they are represented as a function over space.

The idea of FDA can be explained further in the following equations, where discrete data are to be converted to functional data using basis functions (Ramsay & Silverman 2005).

$$\mathbf{y}_i(\mathbf{p}) = \mathbf{x}_i(\mathbf{p}) + \epsilon_i,$$

where $\mathbf{y}_i(\mathbf{p})$ are the values of the discrete data measured over a continuum, ϵ_i are measurement errors, $\mathbf{x}_i(\mathbf{p})$ are the linear combinations of the basis functions, $\phi_{ij}(\mathbf{p})$, with the coefficients $\mathbf{c}_{ij}$:

$$\mathbf{x}_i(\mathbf{p}) = \sum_{j=1}^{J} \mathbf{c}_{ij}\phi_{ij}(\mathbf{p}).$$

Here, $i = 1, \ldots, N$ is the number of observations and $j = 1, \ldots, J$ is the number of basis functions. B-spline basis functions are commonly used in FDA, along with Fourier basis functions (Ramsay & Silverman 2005, Abraham et al. 2003, Serban & Wasserman 2005, Dannenmaier et al. 2020). Fourier basis functions are suitable for data exhibiting a cyclic trend and are less appropriate for the NewDEPOMOD data. B-spline basis functions will therefore be used throughout. Their setup requires several choices to be made:

- The range over which the function is to be evaluated.
- The number of basis functions to be used.
- The order of the b-splines.
- The number of knots.

The number of basis functions, the order and the number of knots are all linked by the following equation:

$$\text{No. of knots} = (\text{No. of basis functions}) - (\text{order}) + 2.$$

The choices for the basis functions contribute to how flexible the functions describing the output maps will be. When smoothing the functional data, the amount of smoothing required is specified by the smoothing parameter, λ (Ramsay & Silverman 2005). The smoothing parameter λ can be tuned to improve the performance of the functions, with Generalised Cross Validation ('GCV') being an appropriate approach for identifying the best value. To do this, smoothing is done using different values for λ, and the average GCV values are calculated, with the lowest GCV value identifying the optimal choice for λ. Commonly, FDA is used for data measured over time, where univariate basis functions are used, but to model the output maps, the FDA is used to model the data which is measured over space, with bivariate basis functions being used.

3.3.1 Fitting a surface using functional data approach

Fitting surfaces to the NewDEPOMOD output maps can be challenging due to the vast areas of the domain where zero deposition occurs. Due to the considerable differences in variability across the domain, it can be difficult to produce an accurate representation of the surface using a functional approach. The output maps were reduced by removing sections to the East of the farm where no deposition occurs across all of te runs. However, there is still a large area where there is a very small amount of deposition, which occurs in some of the runs, so these grid cells remained in the analysis. An example of an output map illustrating the variability is given in Figure 3.6. This output map will then be used an example for testing different smoothing approaches. An adaptive smoothing approach is proposed to allow the level of smoothness to adapt to the variability in certain regions. The adaptive penalty matrix will be used to allow areas with larger variability to be penalised less, to allow the modelled surface to capture this variability.

To describe the method used, it will be described for the univariate functional representation. Given a response, $Y(t)$ measured over time, it can be represented using a functional approach as:

$$Y(t) = \boldsymbol{\Phi}(t)\boldsymbol{\beta} + \boldsymbol{\epsilon}(t).$$

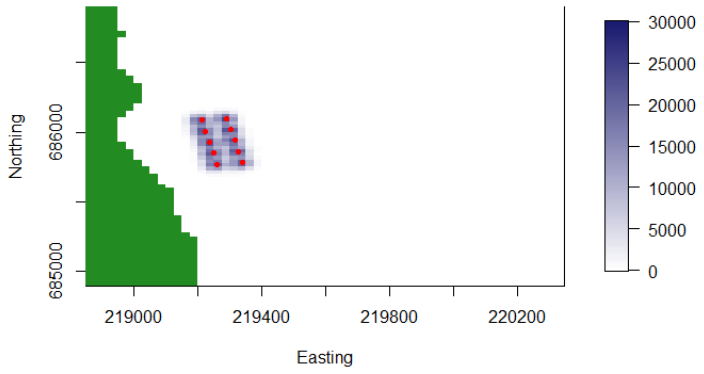


Figure 3.6: NewDEPOMOD output illustrating the high levels of variability and large proportion of areas with zero deposition within the domain.

Here, $\mathbf{\Phi}(t)$ represents a univariate B-spline basis function, and $\boldsymbol{\beta}$ represents the coefficients for the basis functions. When fitting a standard P-spline model, a penalty term is used to control the smoothness of $\hat{Y}(t)$. The coefficient vector, $\hat{\boldsymbol{\beta}}$, is found by minimizing the penalised least squares criterion, with the second difference matrix, $\mathbf{D}$:

$$\|\mathbf{Y} - \mathbf{\Phi}\boldsymbol{\beta}\|^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|^2 \tag{3.5}$$

Criterion 3.5 applies the same penalty across each of the elements in $\boldsymbol{\beta}$. This method is altered to allow for adaptive smoothing to take place by replacing $\mathbf{D}$ with a second difference matrix of a diagonal matrix that has elements that represent the different levels of variability across the time points. The variability levels are denoted by α_j, with $j = 1, \ldots, J$, where J represents the dimension of the B-spline basis, and therefore the number of coefficients in $\boldsymbol{\beta}$. The matrix $\mathbf{D}$ in criterion 3.5 is replaced by the second difference matrix of the diagonal matrix with elements α_j on the diagonal, and will be denoted as $\mathbf{D}_\alpha$. With this approach, the calculation of α_j is crucial and may result in over adjustment, with some time points being over penalised in comparison to the rest. Different modification techniques can be applied to account for this, such as a lower cap, a square root transformation or a log transformation. These will be considered in more detail later.

The above method will have to be altered to account for bivariate functional data in order to apply this to the output data for NewDEPOMOD. Xiao et al. (2013) introduced the 'sandwich smoother' as a way to implement a fast pe-

nalised spline method for bivariate smoothing. This approach will therefore be used to expand the adaptive smoothing method for bivariate functional data. If we now consider $\mathbf{Y}$ to be an $r \times s$ matrix representing the output on a regular grid, then Xiao et al. (2013) proposed smoothing across the rows and down the columns of $\mathbf{Y}$ such that:

$$\hat{\mathbf{Y}} = \boldsymbol{\mathcal{S}}_e \mathbf{Y} \boldsymbol{\mathcal{S}}_n,$$

where $\boldsymbol{\mathcal{S}}_e$ and $\boldsymbol{\mathcal{S}}_n$ are the smoother matrices for each dimension. If we stack the columns of $\mathbf{Y}$ into a vector, then by the properties of the tensor product (Seber 2007),

$$\hat{\mathbf{y}} = (\boldsymbol{\mathcal{S}}_n \otimes \boldsymbol{\mathcal{S}}_e)\mathbf{y}. \tag{3.6}$$

By calculating the tensor product of the two univariate smoother matrices, an overall smoother matrix can be calculated. Each smoother matrix can be calculated as follows using P-splines for $l = e, n$:

$$\boldsymbol{\mathcal{S}}_l = \boldsymbol{\Phi}_l(\boldsymbol{\Phi}_l^\top \boldsymbol{\Phi}_l + \lambda \mathbf{D}_l^\top \mathbf{D}_l)^{-1} \boldsymbol{\Phi}_l^\top, \tag{3.7}$$

with $\boldsymbol{\Phi}_l$ representing the model matrix for each dimension using a B-spline basis, and $\mathbf{D}_l$ denoting the difference matrices. The adaptive approach will be applied to $\boldsymbol{\mathcal{S}}_e$ and $\boldsymbol{\mathcal{S}}_n$ by substituting in $\mathbf{D}_{l\alpha}$ to implement an adaptive penalty. Using the adaptive smoothing approach, the calculation of the smooth surface, $\hat{\mathbf{y}}$ can be expressed as follows:

$$\hat{\mathbf{y}} = \boldsymbol{\mathcal{S}}\mathbf{y}. \tag{3.8}$$

Here the tensor product is defined as $\boldsymbol{\mathcal{S}} = (\boldsymbol{\mathcal{S}}_n \otimes \boldsymbol{\mathcal{S}}_e)$. One choice to be made when applying the adaptive smoothing method is the smoothing parameter, λ. This can be done by fitting surfaces using a range of values for λ and calculating the GCV value for each, with the lowest value indicating the optimal λ. The other choice to be made is the calculation of α, which denotes the variability level for calculating the penalty, on the scale $[0, 1]$. For applying this to the output maps from NewDEPOMOD, the calculation of α will be dependent on the variation in the Easting and Northing directions, V_e, V_n. If there are $j = 1, \ldots, r$ coordinates in the Easting direction and $\{\mathbf{V}_e = (V_{e,1}, \ldots, V_{e,r})\}$, then the calculation is given as follows:

$$\alpha_{e,j} = \frac{|V_{e,j} - \max(\mathbf{V}_e)|}{\max(\mathbf{V}_e) - \min(\mathbf{V}_e)}. \tag{3.9}$$

This equation will allow areas with lower variability to be penalised more, to allow the fitted surface to capture the areas with higher variability more

accurately. The above equation can be altered to account for the variation in the Northing direction, where there are $j = 1, \ldots, s$ coordinates, by changing the subscripts e to n.

For the method described above, there is a choice to be made regarding the setup of the B-spline basis. It has been mentioned previously that there are large areas of the domain where zero deposition occurs, therefore an irregular basis setup could be used to improve the accuracy in the areas with larger variance in the deposition. This was done through a dropped knots approach, where a saturated basis function is set up initially for Easting and Northing, with knots removed in the areas where the variance of Solids Flux is equal to zero.

This method can be compared to the standard smoothing method with no penalty term as well as the adaptive smoothing where knots are placed regularly at every second grid cell for both methods. Table 3.7 highlights the improvement that can be seen by using an adaptive smoothing approach with irregular knots to fit a smooth surface to a sample output map. Figures

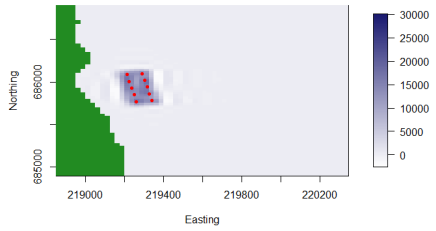


Figure 3.7: Plot of the fitted surface for Solids Flux across the domain with no penalty term and knots placed at regular intervals, every second grid cell.

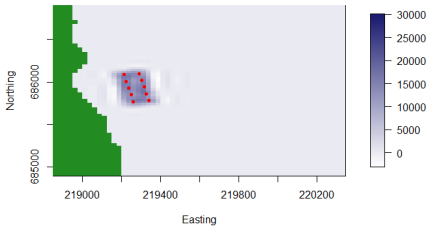


Figure 3.8: Plot of the fitted surface for Solids Flux across the domain with adaptive penalty term and knots placed at regular intervals, every second grid cell.

3.7 and 3.8 slightly overpredict Solids Flux in the areas where the original output map (Figure 3.10) suggest that there is zero deposition. One potential reason for the overprediction is that the deposition over the rest of the runs is higher, and so these areas have a predicted deposition slightly greater than zero. In comparison, using irregular knots and the adaptive penalty results in a much better representation of the surface. This is further highlighted by considering the MSE for each of the fitted surfaces in Table 3.7. Comparing the two models with the knots placed every second grid cell, adding the adaptive penalty produces a slight increase in the MSE. Moving on to using the dropped

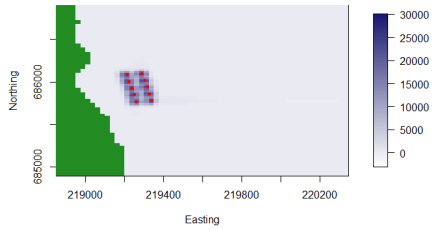


Figure 3.9: Plot of the fitted surface for Solids Flux across the domain with adaptive penalty term and knots placed at irregular intervals, using a dropped knots approach.

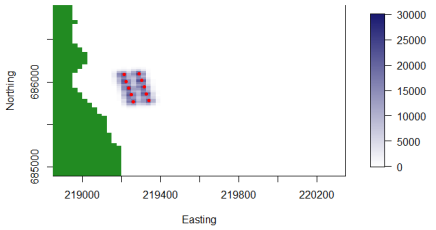


Figure 3.10: Plot of the original output map from NewDEPOMOD.

Method	**MSE**
No penalty (Regular knots) - Figure 3.7	417013
Adaptive Smoothing (Regular knots) - Figure 3.8	419611
Adaptive Smoothing (Irregular knots) - Figure 3.9	14109

Table 3.7: Table of MSE for the different approaches to fitting the smooth surfaces.

knots approach produced a much improved MSE. The adaptive smoothing approach with dropped knots demonstrates an ability to create a surface that represents the output maps effectively in comparison to the other approaches.

After determining that an adaptive smoothing approach is best, consistency in the way these surfaces are produced will allow more robust comparisons to be made. The consistency refers to the choice of smoothing parameter, λ, and the values of $\boldsymbol{\alpha}_{e(n)}$. A quick exploration will therefore consider surfaces created using an optimal λ for each surface, and $\boldsymbol{\alpha}_{e(n)}$ calculated for each output map. A sample of output maps were chosen based on a set of percentiles of the total Solids Flux for each output map. The percentiles that were chosen were: $\{5\text{th}, 25\text{th}, 50\text{th}, 75\text{th}, 95\text{th}\}$. Using the associated output maps for each percentile, smooth surfaces were then fitted, where the optimal value for λ was chosen using GCV, and $\boldsymbol{\alpha}_{e(n)}$ was calculated individually for each output map, where $\boldsymbol{\alpha}_{e(n)}$ is an abbreviation rather than writing both $\boldsymbol{\alpha}_e$ and $\boldsymbol{\alpha}_n$. In addition, surfaces were also fitted with an overall $\boldsymbol{\alpha}_{e(n)}$ used for each output map, and the optimal value for λ also chosen using GCV. The overall $\boldsymbol{\alpha}_{e(n)}$ is calculated using Equation 3.9 as before, however, the calculations for variance are made using the data for each Easting (Northing) coordinate across all of the output

maps being considered. Figure 3.11 highlights that the fitted surfaces have a

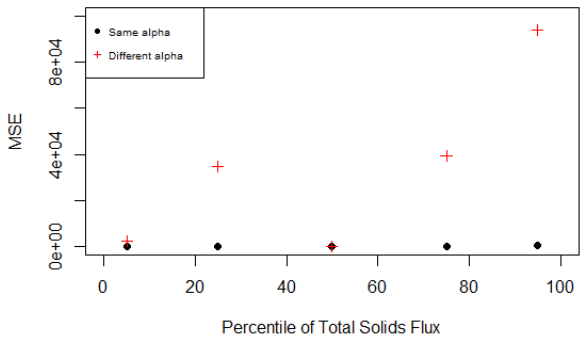


Figure 3.11: Plot of the MSE for surfaces fitted using the same $\boldsymbol{\alpha}_{e(n)}$ and surfaces fitted using different $\boldsymbol{\alpha}_{e(n)}$ for each surface.

lower value for MSE when using the same $\boldsymbol{\alpha}_{e(n)}$ for all of the surfaces, indicating the simpler approach is more effective. For the surfaces fitted using the same $\boldsymbol{\alpha}_{e(n)}$, the optimal λ value was considered for each surface. From Figure 3.12,

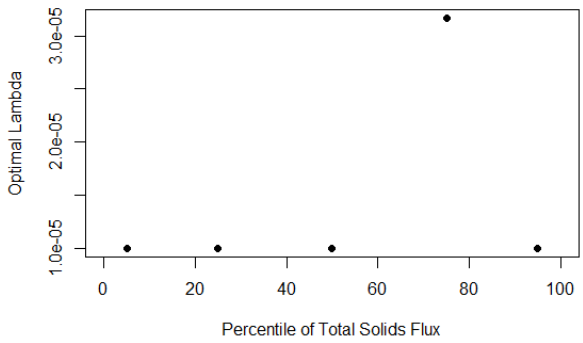


Figure 3.12: Plot of the optimal values for λ for surfaces fitted using the same $\boldsymbol{\alpha}_{e(n)}$ and surfaces fitted using different $\boldsymbol{\alpha}_{e(n)}$ for each surface.

four of the five optimal λ values were the same. This would suggest that it is appropriate to consider an overall λ value to be used to fit all surfaces rather than adding additional complexity and computational time.

In order to confirm the approach of using a single value for λ, a set of the 50 replicate runs for a given set of inputs will be considered rather than an average of these runs that was considered for Figures 3.11 and **??**. Therefore, surfaces were fitted for each of the replicate runs using the optimal value for λ. Figure 3.13 identifies that the majority of the values are similar, with 38 of the

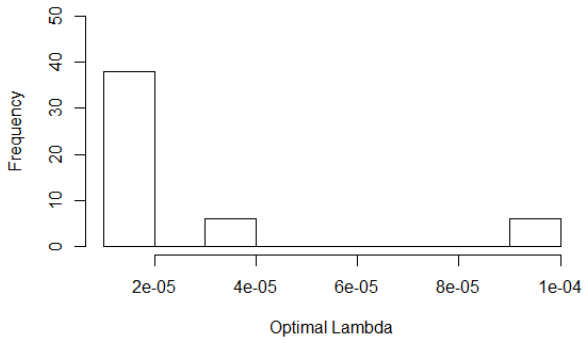


Figure 3.13: Histogram of the optimal values for λ for each of the surfaces for the replicate runs.

50 replicates having an optimal λ value of 1×10^{-5}. This indicates again the idea of a single value for λ being appropriate. The next step of the process will involve checking the optimal values for λ when considering the output maps for another site. Doing so for an additional site resulted in 100% of the values for λ being equal. Therefore, the investigation shows that using a common value for λ to create each surface is appropriate.

When dealing with replicate runs, the method described above will be applied for each set of replicates, producing an individual smoothing matrix, $\boldsymbol{\mathcal{S}}_m$, for each set of NewDEPOMOD inputs, $m = 1, \ldots, M$.

After fitting a surface to the NewDEPOMOD output map, it is important to consider the variance and standard error of the fitted surface. Xiao (2012), Xiao et al. (2013) considered the variance-covariance matrix for fitting functional data in the univariate case, where the output is measured over time, $\{t_1, \ldots, t_m\}$. Each row in the output matrix, $\mathbf{Y}$, corresponds to an output measured over the m timepoints. To calculate the sample variance-covariance matrix, each element, $K(t_j, t_l)$, corresponds to the covariance between the output at each pair of sampling points t_j and t_l (Xiao 2012, Xiao et al. 2013).

This method will have to be altered for bivariate functional data, dealing

with an output measured over two dimensions. To calculate each element of the sample variance-covariance, rather than a sampling point being an individual timepoint, t_j, it will instead be considered as a set of given coordinates, (e, n). Consider a set of N output maps, $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$, each of which are $(r \times s)$ matrices. Within each matrix, each element corresponds to a grid cell in the domain, with coordinates (e_i, n_j), for $i = 1, \ldots, r$ and $j = 1, \ldots, s$. Let $\mathbf{y}_{i,j}$ be a vector of length N, containing the i, jth element of each matrix $\mathbf{Y}_k$, for $k = 1, \ldots, N$. In other words, $\mathbf{y}_{i,j} = ((\mathbf{Y}_1)_{i,j}, \ldots, (\mathbf{Y}_N)_{i,j})$. Then, the sample variance-covariance matrix, $\mathbf{S}$ has the following dimensions: $(rs \times rs)$. The diagonal and off-diagonal elements of $\mathbf{S}$ can be expressed as:

$$\mathbf{S} = \begin{pmatrix} \text{Var}(\mathbf{y}_{1,1}) & \text{Cov}(\mathbf{y}_{2,1}, \mathbf{y}_{1,1}) & \cdots & \text{Cov}(\mathbf{y}_{r,s}, \mathbf{y}_{1,1}) \\ \text{Cov}(\mathbf{y}_{1,1}, \mathbf{y}_{2,1}) & \text{Var}(\mathbf{y}_{2,1}) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \text{Cov}(\mathbf{y}_{1,1}, \mathbf{y}_{r,1}) & \text{Cov}(\mathbf{y}_{2,1}, \mathbf{y}_{r,1}) & \cdots & \text{Cov}(\mathbf{y}_{r,s}, \mathbf{y}_{r,1}) \\ \vdots & \vdots & \cdots & \vdots \\ \text{Cov}(\mathbf{y}_{1,1}, \mathbf{y}_{1,s}) & \text{Cov}(\mathbf{y}_{2,1}, \mathbf{y}_{1,s}) & \cdots & \text{Cov}(\mathbf{y}_{r,s}, \mathbf{y}_{1,s}) \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \text{Cov}(\mathbf{y}_{r,s}, \mathbf{y}_{(r-1),s}) \\ \text{Cov}(\mathbf{y}_{1,1}, \mathbf{y}_{r,s}) & \text{Cov}(\mathbf{y}_{2,1}, \mathbf{y}_{r,s}) & \cdots & \text{Var}(\mathbf{y}_{r,s}) \end{pmatrix}.$$

The diagonal elements of $\mathbf{S}$ refer to the variance of an individual grid cell. The remaining elements in a given column are the covariance between the given grid cell and each of the remaining grid cells. The next step is to use the sample variance-covariance matrix from the original output maps to calculate an estimate of the fitted variance-covariance matrix for the smooth surfaces. To estimate the variance-covariance matrix for a fitted surface, the smoothing matrix, $\boldsymbol{\mathcal{S}}$, used to fit the surface will be considered. For a fitted surface, $\hat{\mathbf{Y}}_k$, the estimated variance-covariance matrix can be calculated as follows, using the sample variance-covariance matrix, $\mathbf{S}$:

$$\text{Var}(\hat{\mathbf{Y}}) = \text{Var}(\boldsymbol{\mathcal{S}}\mathbf{Y}) \tag{3.10}$$

$$= \boldsymbol{\mathcal{S}}\mathbf{S}\boldsymbol{\mathcal{S}}^\top. \tag{3.11}$$

Equation 3.11 will produce an $(rs \times rs)$ matrix corresponding to the estimated variance-covariance matrix for the fitted surface, which will be denoted, $\boldsymbol{\Sigma}$. The diagonal elements of $\boldsymbol{\Sigma}$ correspond to the estimated variance for an individual grid cell, so the standard error can be calculated by taking the square root of

these values.

Due to the random walk element of NewDEPOMOD, replicate runs are often completed for each set of NewDEPOMOD inputs. Therefore, the above method is suitable for each set of replicate runs to estimate the standard error map for a given set of NewDEPOMOD inputs. Let M be the number of NewDEPOMOD input sets being considered for a given analysis. For a given set of inputs, $m = 1, \ldots, M$, there are $\mathbf{Y}_{m,1}, \ldots, \mathbf{Y}_{m,N}$ output maps, where N represents the number of replicate runs completed. Using the above method, a sample variance-covariance matrix can be calculated for the input set m, and is denoted as $\mathbf{S}_m$. Using Equation 3.11, an estimated variance-covariance matrix, $\boldsymbol{\Sigma}_m$ can be calculated for each set of inputs, m.

To illustrate this in practice two sets of replicate runs for two different input sets were considered in order to compare the estimated standard error maps. The above methods were applied to estimate two variance-covariance matrices, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. By taking the square root of the diagonals, the standard error maps can be produced. There are some differences between both maps,

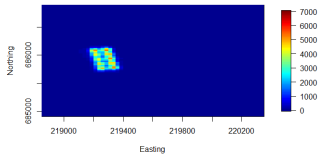


Figure 3.14: Map of the estimated standard errors of the fitted surfaces for a given set of replicate runs - Example 1.

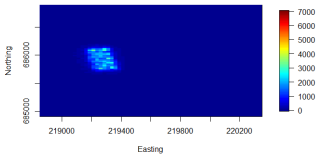


Figure 3.15: Map of the estimated standard errors of the fitted surfaces for a given set of replicate runs - Example 2.

which is what we would expect. In Figure 3.15, there is an area below the cages with standard error values greater than zero, which is not seen in Figure 3.14. Computationally, these standard error maps are expensive to run. The calculation of each $\boldsymbol{\Sigma}_m$ requires the matrix multiplication of three $(rs \times rs)$ matrices, with the matrix multiplication of two of these matrices having a complexity of somewhere between $O((rs)^{2.37})$ and $O((rs)^3)$. Therefore will not be calculated for every surface and the plots above are included to illustrate how the standard error maps can be calculated and how they would look.

3.3.2 Bivariate functional PCA approach for identifying areas of variation

A functional PCA ('FPCA') approach will be considered first. This will identify the main areas of variation within the domain, with the aim of determining which inputs are driving these variations using the PC scores.

To begin the FPCA approach, the functional representations of the output maps, $\hat{\mathbf{Y}}_i$, will be expressed as a linear combination of the set of basis functions. To do so, the fitted surfaces, $\hat{\mathbf{Y}}_i$ are converted into vectors by stacking the columns, $\hat{\mathbf{y}}_i = vec(\hat{\mathbf{Y}}_i)$, and can be expressed as follows:

$$\hat{\mathbf{y}}_i = \boldsymbol{\Phi}\hat{\boldsymbol{\beta}}_i.$$

Here, $\boldsymbol{\Phi}$, refers to the model matrix for the bivariate B-spline basis, where $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_n \otimes \boldsymbol{\Phi}_e)$. Next, $\hat{\boldsymbol{\beta}}_i$ is the vector of estimated coefficients for the relative basis functions, used to produce the ith fitted surface, $\hat{\mathbf{Y}}_i$. Each element of the matrix, $\hat{\mathbf{Y}}_i$, corresponds to the estimated value of Solids Flux for a grid cell with coordinates, $\{(e_a, n_b) : a = 1, \ldots, A \,\&\, b = 1, \ldots, B\}$. To proceed with the next step, there is an assumption that the set of outputs have zero mean. For univariate functional data measured over time, this requires the data to have zero mean for each time point. Extending this to the bivariate case, the data will require each grid cell to have zero mean over all of the observations. This can be done by subtracting a mean map, $\bar{\mathbf{Y}}$, from each fitted surface, $\hat{\mathbf{Y}}_i$. From now on, the fitted surfaces, $\hat{\mathbf{Y}}_i$, will correspond to the centered surfaces. The covariance functions for each grid cell can then be expressed as:

$$V\left(e_a, n_a, e_b, n_b\right) = \frac{1}{N}\boldsymbol{\Phi}\left(e_a, n_a\right)^\top \hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}}\boldsymbol{\Phi}\left(e_b, n_b\right).$$

The respective eigenproblem to be solved for the functional PCA is:

$$\int\int V\left(e_a, n_a, e, n\right)\xi\left(e, n\right) dedn = \lambda\xi\left(e_a, n_a\right). \tag{3.12}$$

The following two orthonormal conditions must be satisfied to solve the eigenproblem:

1. $\int\int \xi_p(e,n)^2 dedn = 1$,
2. $\int\int xi_p(e,n)\xi_q(e,n) dedn = 0$,

where $p \neq q$ are indices for eigenfunctions. To solve Equation 3.12, a further basis expansion is required, $\xi(e,n) = \boldsymbol{\Phi}(e,n)^\top \mathbf{c}$. Then defining the matrix,

$W(e,n) = \mathbf{\Phi}(e,n)\mathbf{\Phi}(e,n)^\top$. In order to solve Equation 3.12, the trapezoidal rule is used to approximate the double integral $\mathbf{W} = \int\int W(e,n)\, dedn$ over the range of e and n values for the domain, which is essential to solving the eigenproblem (Gong et al. 2015). The left hand side of equation 3.12 can be expressed as,

$$\int\int V\left(e_a, n_a, e, n\right)\xi\left(e,n\right) dedn \tag{3.13}$$

$$= \frac{1}{N}\int\int \mathbf{\Phi}\left(e_a,n_a\right)^\top \hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}} \mathbf{\Phi}\left(e,n\right)\mathbf{\Phi}(e,n)^\top \mathbf{c}\, dedn \tag{3.14}$$

$$= \frac{1}{N}\mathbf{\Phi}\left(e_a,n_a\right)^\top \hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}} \int\int \mathbf{\Phi}\left(e,n\right)\mathbf{\Phi}(e,n)^\top \mathbf{c}\, dedn \tag{3.15}$$

$$= \frac{1}{N}\mathbf{\Phi}\left(e_a,n_a\right)^\top \hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}}\mathbf{W}\mathbf{c}. \tag{3.16}$$

$$\tag{3.17}$$

Therefore, using Equation 3.17 and the basis expansion of $\xi(e,n)$, the approximated eigenproblem can then be shown as

$$\frac{1}{N}\mathbf{\Phi}\left(e_a,n_a\right)^\top \hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}}\mathbf{W}\mathbf{c} = \lambda\mathbf{\Phi}(e,n)^\top\mathbf{c}. \tag{3.18}$$

The next step of the process involves substitution to convert Equation 3.18 to a symmetric eigenproblem. The required substitution is $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{c}$, which produces the following eigenproblem:

$$\frac{1}{N}\mathbf{W}^{1/2}\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}}\mathbf{W}^{1/2}\mathbf{u} = \lambda\mathbf{u}. \tag{3.19}$$

Equation 3.19 will then be solved for λ and $\mathbf{u}$. Following this, the reverse problem, $\mathbf{c} = \mathbf{W}^{-1/2}\mathbf{u}$, will be solved. This will allow the eigenfunction $\xi(e,n) = \mathbf{\Phi}(e,n)^\top\mathbf{c}$ to be calculated. The eigenvalues, λ, indicate the proportion of the variation explained by each of the principal components. The corresponding principal component scores are then calculated by:

$$z_i = \int\int \xi(e,n)\hat{\mathbf{Y}}_i\, dedn,\ i = 1,\ldots,N. \tag{3.20}$$

The eigenfunctions, $\xi(e,n)$ will provide information as to the sources of variation within the domain. Principal component scores are calculate for each principal component being considered, and can be used, along with the eigenfunctions, to reconstruct the original output, $\hat{\mathbf{Y}}_i$.

3.3.3 Results from bivariate functional analysis approach

For the analysis of the output maps using the bivariate functional analysis approach, the data from the combined sensitivity analysis of the inputs based on the physical properties and the operational inputs at Ardentinny will be considered. This approach will allow the whole output map to be considered, instead of just the main shape of the impact. The output maps for this site contain (60×80) pixels in the domain.

The approach used to fit the 2-dimensional functional data for the NewDEPOMOD output maps required a large number of basis functions to capture the variability of the Solids Flux over the whole domain. As a result, the above approach which approximates the double integrals becomes an infeasible calculation. One consideration to allow the calculations to be completed is to use a sample of the basis functions that were used to calculate the smooth surfaces.

As mentioned previously, to solve the eigenproblem, the output maps were centred. Due to the replication of runs at each set of input values, the mean value for each grid cell over the set of replicate runs was removed. This means that over the set of replicate runs, the mean value for each grid cell will be zero. When investigating the number of basis functions required to calculate the smooth surfaces, the computation time had to be considered. Using the dropped knots approach, described earlier in the Chapter, it was determined that 2900 basis functions were able to capture the high levels of variability in some areas of the domain. Computing the functional PCA is computationally demanding, and reducing the number of basis functions was explored to improve efficiency without sacrificing accuracy. It was determined that reducing the number of basis functions to a sample of 105, evenly spaced over the original basis matrix, allowed the trapezoidal approximation of the integrals to be completed efficiently without a detrimental effect on the performance, which was seen by the similar values for MSE when comparing to the original output map. This then allowed the eigenproblem to be solved and the PCA to be completed.

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	2.22×10^9	1.93×10^9	1.63×10^9	1.42×10^9	1.30×10^9
Variance Proportion	22.9%	19.9%	16.8%	14.7%	13.4%

Table 3.8: Eigenvalues and Variance Proportion for the first five PCs.

The first five PCs are able to describe approximately 88% of the variation in the data, with all of them playing a similar role, with a difference of less than 10% between the first and the fifth PC. There is then a drop in the % of variance explained by the sixth PC to 4.5%. The eigenfunctions for the first two PCs are given in Figures 3.16 and 3.17, and highlight the different variation patterns over the domain that are described by each PC.

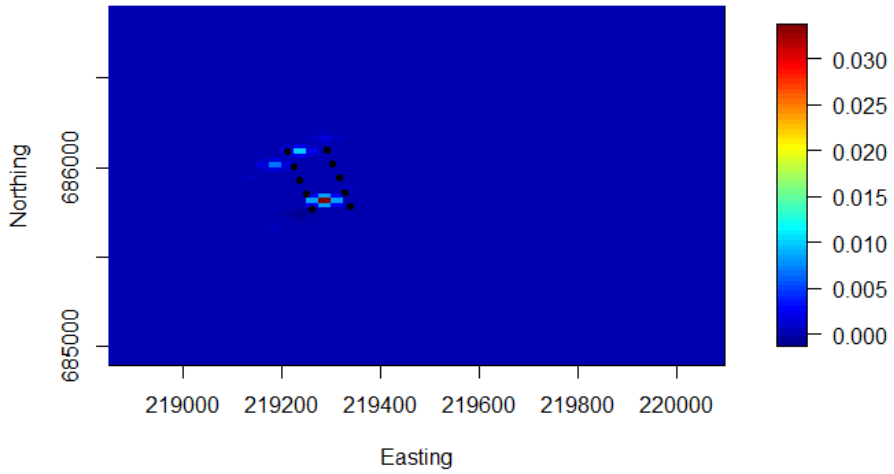


Figure 3.16: Plot of the eigenfunction for PC1 over the fish farm domain.

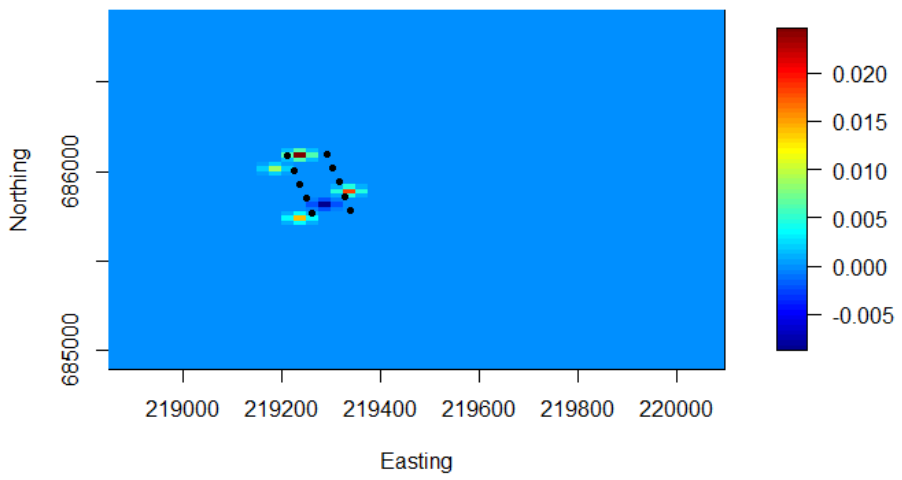


Figure 3.17: Plot of the eigenfunction for PC2 over the fish farm domain.

Due to the vast majority of the domain having zero or close to zero deposition, the variation identified by the eigenfunctions is small in comparison to the size of the domain, but tells a lot about where the main shape of the impact varies. Figure 3.16 highlights a section of the coast to the South of the cage layout as well as an area to the North of the cage layout. Figure 3.17 highlights two areas on the West of the cage layout, one at the North, and one to the South-East. As mentioned previously, the output maps can be reconstructed using the eigenfunctions and the corresponding PC scores. The PC scores can also be considered as the output for the sensitivity analysis, as they are representations of the variation of each map according the eigenfunctions.

As the PC scores can be considered as a scalar output for a sensitivity analysis, random forest modelling was used to help identify the most influential inputs for each PC. In order to draw inference from the random forest models, they need to be able to explain some of the variability. For the functional PC scores, the random forest models explain close to none of the variation in the data. Figure 3.18 shows the PC scores for the first PC across the runs.

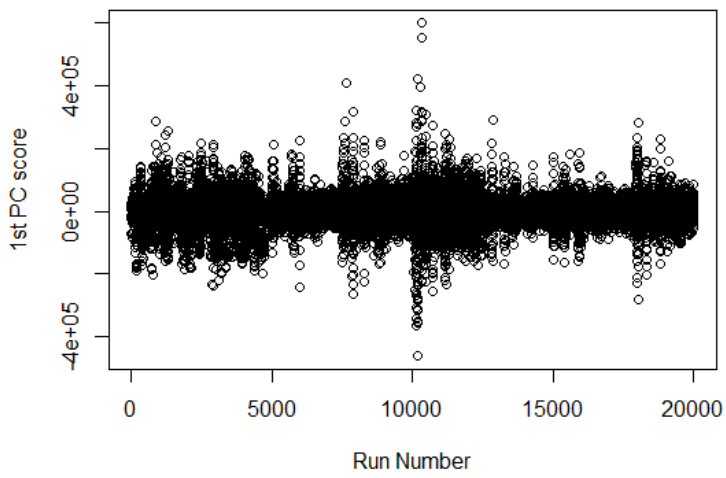


Figure 3.18: Plot of the PC scores for the first PC against the run number.

Within Figure 3.18, there appears to be different levels of variance across the runs. The run numbers are ordered based on the operational inputs, of which there are eight for this site, meaning 2500 runs for each operational setup. Figure 3.18 appears to show some sort of relationship between the variance of the PC scores and the operational setups. The random forest models were fitted again, with one term representing the operational setup as well as the physical properties inputs. However, this did not provide any improvement on the fit of the random forest models.

3.3.4 Review

The functional PCA has provided some insight as to the main areas of variation, how many PCs are required to explain a large amount of the variation, and also highlights that there could be some relationship between the variation and the operational setup. However, this approach has not produced a detailed idea of where some of the inputs are more influential over the domain, and the poor fit of the random forest models to the PC scores did not allow any conclusions to be drawn about which inputs were causing the variance described by each PC. Therefore, further approaches will be considered to provide more detail and draw better conclusions.

3.4 Individual grid cell approach for investigating NewDEPOMOD output maps

In order to overcome the issues discovered for the shape and bivariate functional analyses, the output data for individual grid cells across the domain will be considered. This will allow the whole domain to be considered, and should allow the most influential inputs to be identified.

This approach will consider multiple modelling techniques, with the aim of developing a suitable framework that could be applied to any fish farm site. The initial analysis will feature the data from Ardentinny that was used in the previous section.

3.4.1 Eta-Squared as a sensitivity measure

The eigenfunctions for each of the PCs provide some information about the variation they are describing within the domain, but not in great detail. In addition, no conclusions were able to be made from the random forest modelling of the PC scores. So, in order to gain a more accurate description of which inputs are most influential across the domain, grid cell data will be considered independently.

For a global sensitivity analysis of a univariate output and discrete input factors, it can be considered as the equivalent of an ANOVA decomposition (Saltelli et al. 2000). This method decomposes the output variance and can attribute it to the main effects of each input factor as well as the second order interactions between the inputs. The Sums of Squares are used to decompose the variance, and a unique decomposition exists when a complete factorial design is used (Lamboni et al. 2011). To test an ANOVA-style method, the continuous inputs (physical properties inputs) were converted to discrete inputs, each with 4 levels containing the same number of values. Due to the nature of how the operational setup for a farm was chosen, it does not produce a balanced design with the same number of input sets for each level of the operational inputs. As a result, the ANOVA decomposition must be altered to account for the unbalanced design.

Within R, the *multisensi* package (Bidot et al. 2018) can be utilised to complete a sensitivity analysis on a model with multivariate output. With this package, however, it is not applicable to the case where the design is unbalanced, as it follows an ANOVA style approach for decomposing the variance using type I sums of squares. As a result, the approach will have to be altered to account for the unbalanced design in this case.

Unbalanced factorial designs were studied as far back as 1934, (Yates 1934), who described the three different approaches for calculating sums of squares for unbalanced data to test hypotheses in ANOVA. These methods were considered further by Speed et al. (1978), Herr (1986), who reviewed different approaches to modelling unbalanced data. The three different methods for calculating sums of squares are known as Type I, Type II and Type III. Despite being considered as three different 'sums of squares', each approach differs in the hypothesis testing strategies for the ANOVA - which then lead to different sums of squares values when considering an unbalanced design. It should be noted that, for a balanced design, each approach will produce identical results for the sums of squares. Each of the types are described in more detail below:

- **Type I** - This corresponds to a sequential approach of adding the inputs - beginning with the main effects one at a time before adding each of the interactions for the model comparisons. This approach is dependent on the order if the inputs - and different results can be produced by altering the order.

- **Type III** - This is simpler to discuss than Type II initially. For each hypothesis being tested, the alternative model is always the full model containing all of the main effects and interactions, while the null model deletes the one term that is being tested.

- **Type II** - These are similar to Type III in that it compares a full model with a null model where a single term is removed. The difference between the two is that Type II tests are based on the 'marginality principle', which advises that a you should not omit a lower order term if there are higher order terms that are dependent on it.

Example 3. To illustrate the different hypotheses that are being considered, a simple example will be given. Suppose that two inputs, A and B, are being considered in an unbalanced ANOVA. Tables 3.9, 3.10 and 3.11 explain the different approaches used for each type. First, considering the models being

Table 3.9: Type I

Term being tested	Null Model	Alternative Model
A	1	A
B	A	A + B
A:B	A + B	A + B + A:B

test for the type I approach, in Table 3.9, when input A is tested, the null and

Table 3.10: Type II

Term being tested	Null Model	Alternative Model
A	B	A + B
B	A	A + B
A:B	A + B	A + B + A:B

Table 3.11: Type III

Term being tested	Null Model	Alternative Model
A	B + A:B	A + B + A:B
B	A + A:B	A + B + A:B
A:B	A + B	A + B + A:B

alternative models ignore input B, whereas, when input B is tested, input A is considered in the both models. Therefore, changing which input is being tested first will produce different results. In Table 3.10 for the type II approach, when a main effect is being considered, the alternative model does not include any interaction terms involving that main effect. Finally, in Table 3.10 for the type III approach, when testing main effects the null and alternative models contain the interaction term involving that main effect.

Example 3 highlights some of the issues seen with type I and type III analyses. Much controversy has surrounded the type to use when dealing with unbalanced data, which is considered in Herr (1986), but the choice essentially comes down to the hypothesis being tested. Due to the fact that Type I is dependent on the order of the inputs, it is rarely considered in the circumstance where there is an unbalanced design. It is rare to know which order the specific inputs should be considered, therefore which rules out type I sums of squares when considering unbalanced data. Langsrud (2003) came to the conclusion that type II was preferable, a suggestion that was mentioned previously by Nelder (1977) and Nelder (1994). Langsrud (2003) identified type II sums of squares as being a more powerful tool when no interaction is present between inputs. For type III, the main effects are being tested in the presence of interaction terms which are uninteresting hypotheses (Nelder 1977, 1994). Langsrud (2003) observed that type II methods were previously not considered due to the fact that the interactions are considered to be negligible or non-existent.

Eta squared (η^2) is a standardized measure of effect size for an ANOVA, meaning it can be compared across different units of measurement. It can be summarised as the ratio of variance in an output that is explained by an input.

It can be calculated as follows:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}. \tag{3.21}$$

Here, SS_{effect} refers to the sum of squares of the input, and SS_{total} is the total sum of squares. In the case of an unbalanced design, it is possible for some variance to 'go missing'. The missing variance corresponds to variance in the output that is attributable to the inputs, but where it is not clear which input is responsible. This only occurs when considering Type II & Type III tests, which are more conservative.

To apply the η^2 approach to the output maps, each grid cell within the domain was considered independently - with the Solids Flux values from each of the 20,000 runs considered as the output data. The aim of this analysis was to determine more accurately the inputs that are more influential at specific areas of the domain. This approach will allow η^2 values to be calculated for each of the inputs as well as any interactions. In addition, an η^2 value can be calculated for the residuals. As a result of how η^2 is calculated in Equation 3.21 for each of the inputs, the following condition should hold, when considering all of the η^2 values for the first order effects and any interactions.

$$\eta^2_{Res} + \sum_{i=1}^{n} \eta_i^2 \leq 1. \tag{3.22}$$

As mentioned previously, it is possible for 'missing variance' to be present, where it cannot be attributed to one specific input which is why Equation 3.22 features an inequality. Here, $i = 1, \ldots, n$ represents the element of the model, including interaction terms.

Considering the NewDEPOMOD output maps, η^2 values were calculated using type II sums of squares, for a model containing two-way interactions for all of the inputs. The first step of the process was to look at the sum of the η^2 values (Equation 3.22), to review any areas of missing variance. Figure 3.19 produces some interesting insights - the main one being that the area directly below the farm have values much less than 1, indicating that there is a lot of variance missing in these areas. In addition, it can be seen that areas of the domain have values equal to zero as no deposition occurs in these grid cells. It would be expected that the area directly below the farm would be influenced by the operational inputs. The cage setup plays a big role in the initial deposition of waste below the cages, so it is possible that the missing variance in these areas is a result of the variance not being assigned with confidence to any of

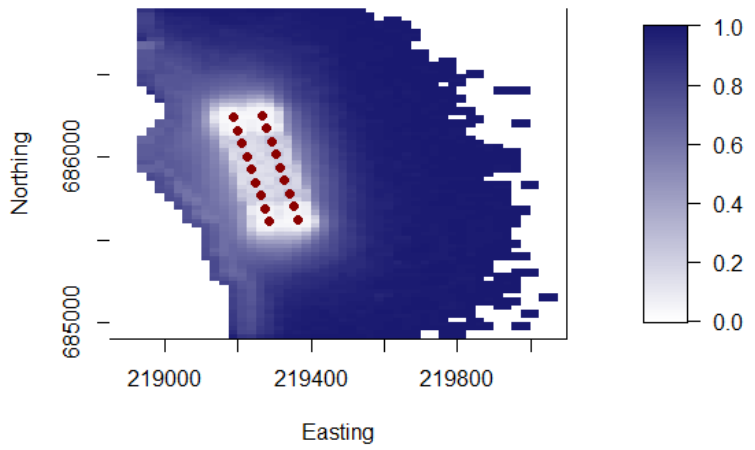


Figure 3.19: Map displaying the sum of the η^2 values for each grid cell.

the operational inputs.

To overcome this problem and reduce the missing variance, only one operational input will be considered which corresponds to the operational setup. Overall at this site, there were 8 different operational setups considered, to the input is a categorical variable with each level corresponding to a given setup. This could reduce the amount of missing variance in some grid cells as it is able to be attributed with confidence to the one operational input being considered. The sums of the η^2 values for each grid cell are given in Figure 3.21. Clear improvements are seen in the areas directly below the farm and along the coast in Figure 3.21, indicating that the previous suspicion that the variance could not be assigned confidently to any of the operational inputs in these areas, was true. However, this plot also highlights one area of concern with the η^2 calculations. Some of the grid cells within the domain violate the condition in Equation 3.22, with the sum of the η^2 values being greater than 1. This means that more variance is being explained than the variance that is available. As a result, any inference drawn from this analysis would have to be considered with caution. To avoid the case where the grid cells violate the condition in Equation 3.22, all inputs should be considered. Further methods will therefore be considered to attempt to overcome the issue of the condition being violated.

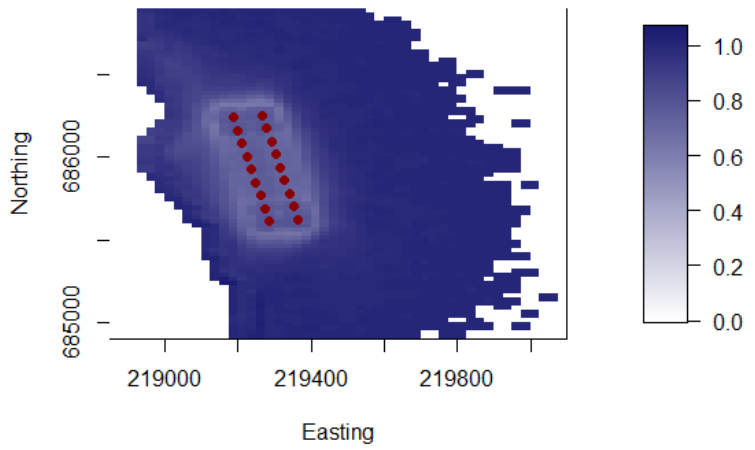


Figure 3.20: Map displaying the sum of the η^2 values for each grid cell for the reduced model with one operational input.

3.4.2 Sobol Indices approach

In order to overcome the issue of the η^2 condition being violated, an alternative approach is considered. A number of different strategies have been proposed for completing global sensitivity analyses over the years, but one of the most popular is the approach proposed by Sobol' (1993). This method computes indices that measure the variance contribution of each input to the total variance of a given output. The common indices that are calculated are the first order and total order indices, which refer to the contribution of each input individually and the total contribution that includes the interaction effects. Variance decomposition methods began with a Fourier implementation (Cukier et al. 1973), before Sobol' (1993) introduced what are now called Sobol indices. The concept of total sensitivity indices were proposed by Jansen et al. (1994), and the expansion of Sobol indices for calculation of total sensitivity indices were introduced by Homma & Saltelli (1996). The estimates of the first-order sensitivity indices for an input, X_i, for the output data, Y, are given as (Sobol' 1993):

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)}. \tag{3.23}$$

One thing that was identified in Chapter 2, was that there was that outliers can be present, shown in the skewed histograms of the scalar outputs. This is also true when considering the output maps, with some outliers being present within the grid cell data. To demonstrate this, the data for the grid cells where deposition has occurred across the runs is combined, and plotted in a histogram in Figure 3.21. The data is heavily skewed, and multiple transformations such

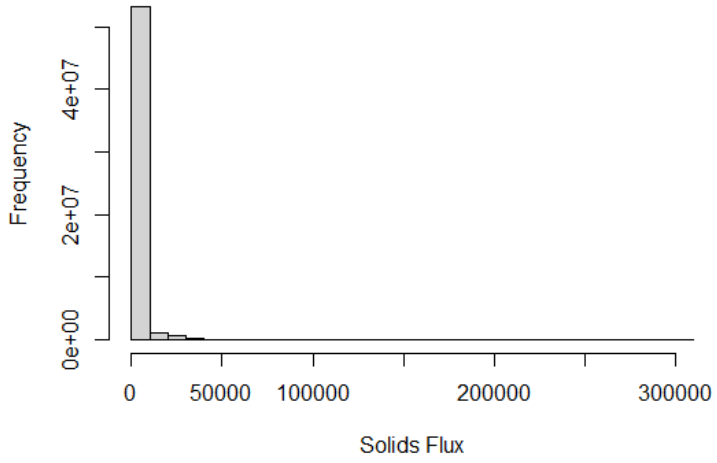


Figure 3.21: Histogram of the combined grid cell data.

as log, square root and cube root were considered but unsuccessful in reducing the skew. One drawback to the Sobol indices are that they are sensitive to outliers. The computation of the numerators in Equations 3.23 can produce values which are greater than the global variance. When calculating the Sobol indices for the data, approximately 14% of the first-order Sobol indices were outwith the required range of $[0, 1]$. As a result, a method for calculating robust Sobol indices was established. A common approach when dealing with outliers, is to trim the data, by discarding a proportion of the smallest and largest values - it is commonly used for calculating robust measures of the mean, variance and regression coefficients (Huber 1981). In order to calculate the robust Sobol indices, the proportion of data to be trimmed was considered. Removing the upper and lower 5% of the data was considered, but still resulted in approximately 8% of the first-order indices being outwith the required $[0, 1]$ range.

Although the robust method shows a reduction in the number of first order Sobol indices outwith the plausible range, it does not solve the problem. The method also works by trimming the data, and therefore removing the 'extreme' values from the analysis. As a result, it is appropriate to consider an approach that is not sensitive to outliers, and it could also be of interest to consider the 'extreme' values in more detail to establish if the are combinations of the inputs that are driving them.

3.4.3 Random forest approach for sensitivity measure

Random forests have been used previously for sensitivity analyses due to their flexibility, and the easily interpretable importance values as a measure of sensitivity ranking. Therefore, they will be considered as an alternative to the variance decomposition methods. As with the η^2 and Sobol approach, each grid cell will be considered independently, and random forest models fitted. As with the η^2 approach, each grid cell in the domain is considered independently, with the output data consisting of the Solids Flux values in each grid cell from the 20,000 runs. A total of 20,000 runs were completed for this site, containing data for 400 different input sets and will all be considered in the analysis due to the efficiency of creating the random forest models. Importance values for the inputs were able to be extracted for each grid cell as a ranking measure. For each grid cell, the highest ranking input could then be extracted to consider any patterns across the domain. Grid cells where zero deposition occurs across all of the runs are indicated by '0' in the legend in Figure 3.22. For this analysis, a number of grid cells have been removed as they feature no deposition across the runs, leaving a rectangular domain where each row or column of the map features at least one grid cell where deposition occurs. There are some clear patterns that can be seen in Figure 3.22. First of all, the dominant input that appears to be highest ranked across the most grid cells is the Settling Velocity of Faeces. The areas that it is highest ranked are below the cages and in the areas surrounding the cages, indicating that it is playing a big role in the build up of waste material near the cages. This can be explained by the fact that it determines the length of time that the faeces remains in the water column and therefore how far it is transported initially from the cages. Below the cages, the Number of Cages appears to be influential in the areas where the additional cages are added. In the areas of the domain slightly further from the cages and along the coast, the resuspension inputs appear to be most important. The one that is be highest ranked in more grid cells is the Settling Velocity of Sediment. The additional resuspension inputs such as the Release

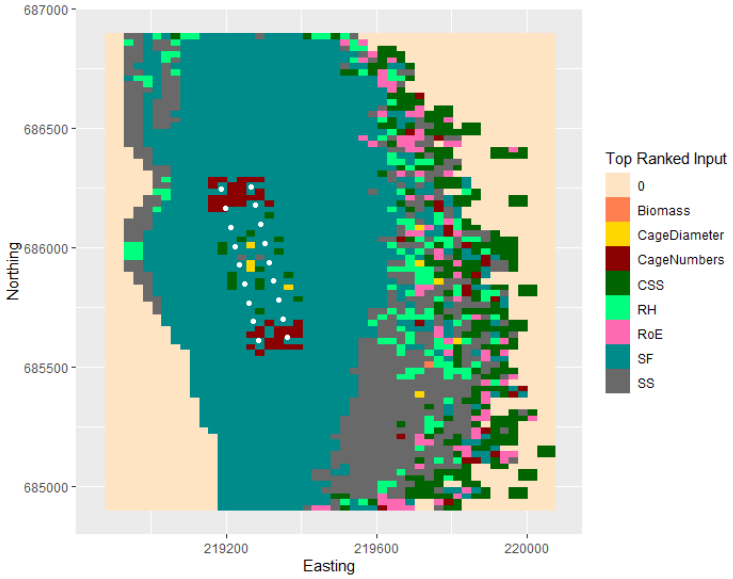


Figure 3.22: Map of the highest ranking input in each grid cell according to the random forest importance.

Height of Resuspended Material, the Critical Shear Stress for Erosion and the rate of Erosion are ranked highest in a number of the grid cells further from the cages. Only a small number of grid cells found Biomass and Cage Diameter to be the highest ranking inputs. The area to the East of the farm features a large amount of variation in the top ranking input, signalling that it is possible the amount of deposition in these areas is potentially low and that more than one of the inputs have similar influence in these areas.

This approach has been able to highlight the different areas within the domain that are dominated by the different inputs. Each of the inputs that were highlighted from Figure 3.22 can be explained logically by the characteristics of each input. The benefit of this approach is the ability to create a single map of the top ranked input for each grid cell. This will allow comparisons to be made between sites to help identify any similarities or differences that might occur.

3.4.4 Considering the extremes

Previously, the Sobol approach was identified as being sensitive to outliers, producing uninterpretable results. One feature of the data for some grid cells within the domain is that it is heavily skewed, with some extreme values present. To consider this extreme data, all of the grid cells were considered together, rather than independently, with the raw data being used rather than the smoothed maps. As previously highlighted, there are a large number of grid

cells within the domain where zero waste deposition occurs across all runs, and so these grid cells were removed from the analysis. Figure 3.23 confirms that

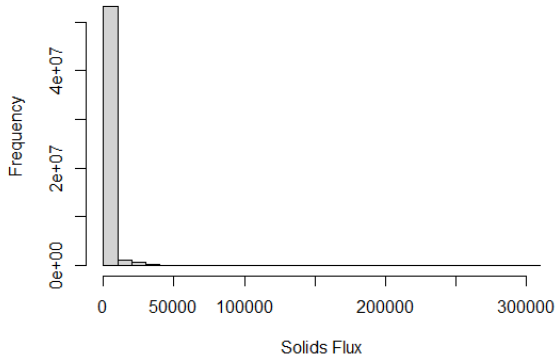


Figure 3.23: Histogram of the output data being used to investigate the extremes, after removing the grid cells with zero deposition.

even after removing the grid cells with zero deposition, the data is still skewed heavily, and the extreme values will have to be identified. Different approaches could be considered to review whether there are any links between the inputs and the extremes. As a starting point with skewed data, transformations were considered, such as log, square and cube roots. Even after transforming the data, it was still heavily skewed, and so other methods are considered.

The first approach that could be considered is a logistic regression, where the binary output refers to whether or not an observation is an extreme. In order to identify the extreme values, a robust Z-score approach will be used. The formula for a standard Z-score is given as:

$$Z = \frac{x - \mu}{\sigma},$$

where x is the observed data, μ is the mean of the data and σ is the standard deviation of the data. In order to calculate a robust version, the median can be considered instead of μ, and a robust measure of scale can be considered instead of σ. Potential robust measures of scale are Median Absolute Deviation ('MAD') or the inter-quartile range. Using these two robust alternatives, Z-scores could be calculate for the output data. A common approach is to consider a cut-off value of 3 for identifying outliers when considering Z-scores, which would refer to three standard deviations away from the mean. After identifying

the extremes, the next step of the process would look at whether there are differences between the groups in terms of the inputs. A discriminant analysis would be an ideal approach for this, however, there are categorical inputs which do not meet the assumption of normality required for the inputs, so logistic regression could be considered.

3.4.4.1 Quantile Regression

An alternative approach which does not require alterations to the data, or calculation of Z-scores, is to consider a quantile regression for the original output data, rather than the binary output data. This approach does not require any calculations of Z-scores or definitions of what is an extreme value. For a set of quantiles $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_m\}$, the model equation for the τ_jth quantile is:

$$Q_{\tau_j}(y_i) = \beta_0(\tau_j) + \beta_1(\tau_j)x_{i1} + \ldots + \beta_p(\tau_j)x_{ip}, \quad i = 1, \ldots, n.$$

Quantile regression is therefore an extension to linear regression, where the beta coefficients are changed from constants, to function with a dependency on the quantile.

For this approach, quantile regression models were fitted with the quantiles $\boldsymbol{\tau} = \{0.9, 0.95, 0.99\}$, after removing the grid cells where zero deposition occurred. Due to computational cost, samples from each set of replicates had to be considered. Out of the 50 replicates for each input set, 5 observations were considered and the quantile regression models fitted for 10 different samples. The models fitted, featured the five continuous physical properties inputs, as well as a categorical variable representing the operational setup for the runs, and additionally two-way interactions between all of the inputs. Due to the interactions and the categorical term, there were a total of 64 variables in the quantile regression models, including the intercept. After fitting the quantile regression models, Table 3.12 shows the percentage of the inputs that were considered to be significant. Table 3.12 shows that the number of significant inputs in the quantile regression models decreases from approximately 70% to 45% as the quantile value increases. Before reviewing the inputs that are significant, a measure of how well the quantile regression models fit are considered. Koenker & Machado (1999) described a process for measuring the fit of a quantile regression model, similar to R^2, which will be summarised below. First, consider a linear quantile regression model,

$$Q_{y_i}(\tau|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}(\tau),$$

Table 3.12: % of Significant inputs for each quantile model

Model Sample	0.9 Quantile	0.95 Quantile	0.99 Quantile
1	74.1%	60.3%	44.8%
2	74.1%	56.9%	43.1%
3	72.4%	60.3%	43.1%
4	74.1%	63.8%	46.6%
5	72.4%	60.3%	44.8%
6	70.7%	65.5%	43.1%
7	69.0%	62.1%	43.1%
8	69.0%	60.3%	55.1%
9	69.0%	65.5%	43.1%
10	67.2%	63.8%	44.8%

with $\widehat{\boldsymbol{\beta}}(\tau)$ being the minimizer of the following,

$$\widehat{V}(\tau) = \min_{\mathbf{b}\in\mathbb{R}^p} \sum \rho_\tau(y_i - \mathbf{x}\mathbf{b}).$$

The above equations correspond to an unrestricted problem, containing all of the inputs. Next, consider a restricted problem, where only the intercept is considered. Then $\widetilde{\boldsymbol{\beta}}$ is the minimizer of the constrained problem,

$$\widetilde{V}(\tau) = \min_{\mathbf{b}_1\in\mathbb{R}} \sum \rho_\tau(y_i - \widetilde{\mathbf{x}}\mathbf{b}_1).$$

In other words, $\widehat{\boldsymbol{\beta}}(\tau)$ and $\widetilde{\boldsymbol{\beta}}(\tau)$ refer to the quantile regression estimates for the restricted and unrestricted models. The goodness-of-fit criterion can then be estimated as (Koenker & Machado 1999),

$$R^1(\tau) = 1 - \frac{\widehat{V}(\tau)}{\widetilde{V}(\tau)}. \tag{3.24}$$

This measure of goodness-of-fit will provide an approximation of how well the quantile regression models explain the variation in the data. For the models fitted for the 0.9 and 0.95 quantiles, $R^1(\tau)$ was approximately 0, indicating that the variation in the data was not explained by these models. There was a slight improvement for the 0.99 quantile, which explained approximately 7% of the variation. Due to the poor fit of the models, the significant inputs will only be considered briefly, focusing on the inputs which were significant across all of the 10 samples.

For the 0.9 quantile, approximately 38% of the inputs were significant across all of the 10 samples. The significant inputs featured were the intercept, and the first order terms included the Rate of Erosion and the Cage Setup. Looking

at the interaction terms, the only interaction terms including only the physical properties inputs were the Critical Shear Stress for Erosion and Rate of Erosion, Rate of Erosion and Settling Velocity of Faeces, and Settling Velocity of Faeces and Settling Velocity of Sediment. The interaction between Critical Shear Stress for Erosion and Rate of Erosion will be influential in determining whether or not waste on the seabed is resuspended and transported, therefore a lack of material being resuspended and transported could produce extreme values. The remaining interaction terms were between each of the physical properties inputs and the Cage Setup - indicating that the combined influence of the Cage Setup and each physical properties input could be influential in modelling the extremes. As previously mentioned, due to the poor quality of the fit of these models, the review of the significant inputs cannot be deemed conclusive.

Moving on to consider the 0.95 quantile in a similar way, the number of significant inputs across the 10 samples was approximately 22%. Again the intercept was included as being significant, and the first order term for Rate of Erosion is no longer included, along with the interactions between Rate of Erosion and Settling Velocity of Faeces, and the interaction between Settling Velocity of Faeces and Settling Velocity of Sediment. The interaction terms between the Cage Setup and Settling Velocity of Faeces, Settling Velocity of Sediment and the Release Height of Resuspended Material remain significant across all of the samples.

Finally, considering the 0.99 quantile, the number of significant inputs across the 10 samples was reduced further to approximately 14%. The intercept was again significant across all of the 10 samples, along with the first order terms for the Cage Setup. In contrast to the previous two quantiles, none of the interaction terms between the physical properties inputs and the Cage Setup were significant across all 10 samples, except the one for Settling Velocity of Faeces. The only other interaction term that was included, is between Critical Shear Stress for Erosion and Rate of Erosion.

From reviewing the inputs that were significant across the 10 samples for each quantile, there are some patterns seen. The Cage Setup and the interaction between the Critical Shear Stress for Erosion and Rate of Erosion appear in all of the quantiles, along with the interaction between Settling Velocity of Faeces and Cage Setup. It was previously highlighted that the fits of these quantile regression models are poor, and so the conclusions have to be considered with caution.

3.4.5 Conclusions

Having tested several approaches to investigate the influence of altering the inputs on the NewDEPOMOD output maps, a framework focusing on a subset of these approaches can be developed which can be applied at additional sites. Considering the individual grid cells within the output maps independently, before combining the results to consider as a map allowed all of the data to be used, and produced plausible results when identifying the most influential inputs over the domain.

The framework that will be used for the additional sites will use the η^2, random forest and quantile regression approaches to assess the impact of the inputs across the domain, as well as investigating the extreme values seen across the domain.

3.5 Framework applied to additional sites

As previously mentioned, a subset of the methods that were considered for Ardentinny will be used to investigate the effects of altering the inputs at the additional sites. An additional low energy site will be considered first, before looking at the two high energy sites. The aim is to look across the different sites for any similarities or differences that are present in the distribution of the influential inputs over the domain. The approach will use the framework described at the start of the chapter, with the random forest, η^2 and quantile regression approaches considered.

3.5.1 Low energy sites

The additional low energy site, West Strome, will be considered first to assess if there are similar patterns seen across the domain compared to Ardentinny. Firstly, the random forest approach will be considered, before looking at the η^2 values and concluding with the modelling of the extremes using quantile regression.

As with the analysis at Ardentinny, the random forest approach will consider the raw data from each cell individually. For each grid cell, a random forest model was fitted, with the importance values then taken for each of the inputs. This allowed the input with the largest importance value in each grid cell to be identified, then used to create a map of the most important grid cells over the domain. Figure 3.25 is a map of the highest ranking inputs, which can be compared to the map produced for Ardentinny, Figure 3.24. There appear

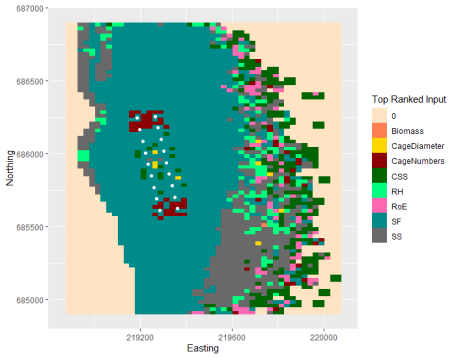


Figure 3.24: Map of the highest ranking input in each grid cell according to the random forest importance - Ardentinny.

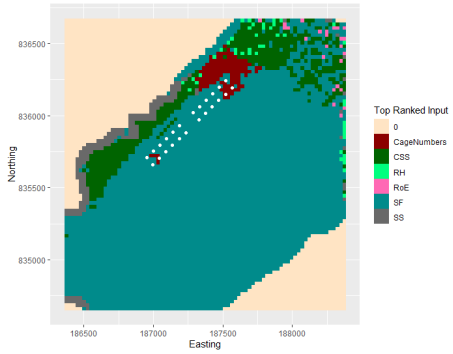


Figure 3.25: Map of the highest ranking input in each grid cell according to the random forest importance - West Strome.

to be some similarities between the two sites at the area where the cages are located. At either end of the cage setup, the Number of Cages is influential for a number of grid cells, as these are the areas where the additional cages are placed. In addition, the Settling Velocity of Faeces plays a key role in the areas below the cages and the areas surrounding the cages at both sites. However, at Ardentinny, there are a larger number of inputs that appear to be influential in the areas to the East of the cages, at the outskirts of the deposition. This area features the Cage Diameter, Number of Cages, Critical Shear Stress, Rate of Erosion, Release Height of Resuspended Material and the Settling Velocity of Faeces. At West Strome, the areas on the outskirts of the deposition appear to be dominated by the Settling Velocity of Faeces and the Critical Shear Stress, with some other additional inputs being highest ranked in some grid cells. The inputs related to the resuspension module appear to play a bigger role at Ardentinny, suggesting resuspension plays a bigger role at this site. The potential reason for the larger number of inputs being highest ranked in certain areas of the domain at Ardentinny are likely that the inputs have similar importance values across those areas, which doesn't appear to be the case at West Strome.

The next step of the framework is to consider η^2 as a variance decomposition technique. Previously, for the analysis at Ardentinny, the map illustrating the highest ranking inputs in each grid cell was not considered. Therefore the η^2 maps for both sites will be considered together. As previously mentioned, to calculate η^2, the continuous inputs have to be converted into categorical inputs with 4 levels, and all operational inputs were considered to avoid violating the condition in Equation 3.22. In the event that any grid cells do violate the condition, they were considered as missing data, and the total η^2 values for the inputs in the remaining grid cells were assessed. The total η^2 values were

calculated by summing the first and second order effects relating to each input. Figures 3.26 and 3.27 show that only the Critical Shear Stress for Erosion was

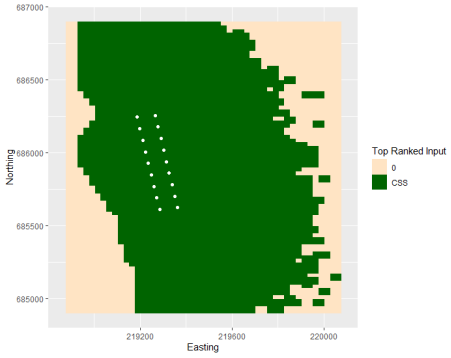


Figure 3.26: Map of the highest ranking input in each grid cell according to η^2 - Ardentinny. (*CSS - Critical Shear Stress for Erosion*)

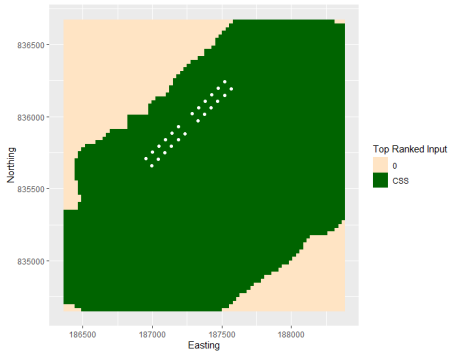


Figure 3.27: Map of the highest ranking input in each grid cell according to η^2 - West Strome. (*CSS - Critical Shear Stress for Erosion*)

identified as being the top ranked input in the grid cells across the domain where deposition occurs, with no grid cells being identified as violating the condition in Equation 3.22. In addition, the use of 4 different levels for converting the continuous inputs to categorical inputs could be a reason for only one input being identified as the top ranked.

In order to test if the number of levels used when converting the continuous inputs to categorical inputs is the reason for only one input being identified as the top ranked, an increased number of levels will be considered. To test this, a total of 8 levels will be considered when converting the continuous inputs to categorical inputs. This approach was considered for Ardentinny, with η^2 calculations run. These calculations were much more computationally expensive compared to the calculations when using 4 levels for the continuous outputs. There are two things to notice when comparing Figure 3.28 to Figure 3.26: 1) Critical Shear Stress for Erosion is again the only input that is identified as being the top ranked and 2) no grid cells were identified as violating the condition for the η^2 calculations. The adjustment in the number of levels being used to convert the continuous inputs to categorical inputs did not help pick out any other inputs as being identified as the most influential in any grid cells. The drawback of using this approach is the computational cost of using the additional levels, with the time taken to fit the models increasing by several times.

Next, considering the extreme values at West Strome, quantile regression models will be fitted in a similar way. Rather than providing the percentage of significant inputs for each of the 10 samples, the mean percentages are given

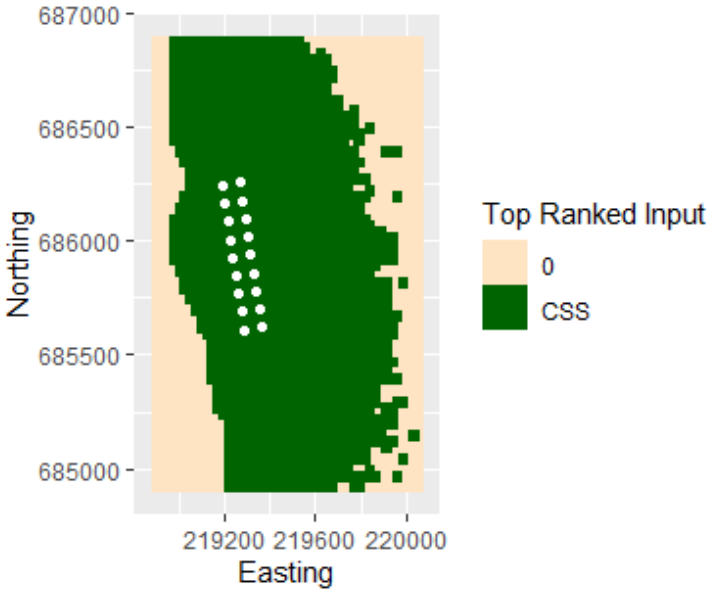


Figure 3.28: Map of the highest ranking input in each grid cell according to η^2, with additional levels used for converting continuous inputs to categorical - Ardentinny. (*CSS - Critical Shear Stress for Erosion*)

in Table 3.13 for each quantile. First, considering the measure of fit for the

Table 3.13: Mean % of Significant inputs for each quantile model across 10 samples

0.9 Quantile	0.95 Quantile	0.99 Quantile
100.0%	57.3%	26.3%

different quantiles, again, the 0.9 and 0.95 quantiles have values close to zero, and the 0.99 quantile has a slight improvement, explaining approximately 4.4% of the variation. As with the results at Ardentinny, they inputs which were significant across the 10 samples are briefly summarised. For the 0.9 quantile, all of the inputs were significant across the 10 samples. Next, considering the 0.95 quantile, approximately 52% of the inputs were significant over all the samples. The intercept was not significant across the 10 samples, and the first order terms that were significant included the Critical Shear Stress for Erosion and the Cage Setup. In addition, the interactions between: Critical Shear Stress for Erosion and Rate of Erosion; Critical Shear Stress for Erosion and Release Height of Resuspended Material; Rate of Erosion and Settling Velocity of Sediment; Rate of Erosion and Release Height of Resuspended Material; and Settling Velocity of Sediment and Release Height of Resuspended Material. In addition to those interactions, the interactions between each physical properties input and the Cage Setup was included. Similar to Ardentinny, the number of inputs significant across the 10 samples decreases again for the 0.99 quantile.

The only first order term is the Cage Setup, and the interaction terms include the interaction between Critical Shear Stress for Erosion and Rate of Erosion, Settling Velocity of Faeces and Release Height of Resuspended Material, and Settling Velocity of Faeces and Cage Setup. The significant inputs appear to be similar to those identified for the analysis at Ardentinny, but the poor fit of the models make the conclusions cautious.

When considering the two low energy sites, it was expected that there would be some similarities in the patterns seen over the domain for the highest ranking inputs. However, there are some large differences between the ranking of the inputs at both sites. This highlights the potential need to consider each site separately, instead of being able to group them by their characteristics.

3.5.2 High energy sites

After considering the low energy sites, the high energy sites will be considered together to investigate any differences or similarities that may be present within the domains. The two high energy sites being considered are Muck and Djuba Wick. As with the additional low energy site, the random forest approach for each grid cell will be considered. The two sites will be analysed together to look for any similarities between the high energy sites, and also compared to the low energy sites. Looking only at Figures 3.29 and 3.30, there appears to

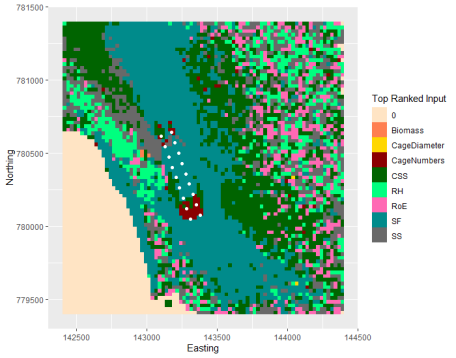


Figure 3.29: Map of the highest ranking input in each grid cell according to the random forest importance - Muck.

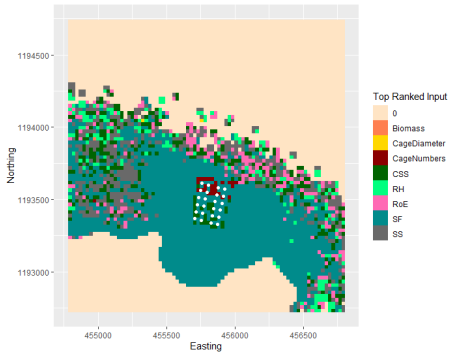


Figure 3.30: Map of the highest ranking input in each grid cell according to the random forest importance - Djuba Wick.

be some similarities between the two sites, but also some differences in certain areas. The Settling Velocity of Faeces appears to be the dominant input in the areas surrounding the cages, but at Djuba Wick, this area appears to be larger. The Critical Shear Stress for Erosion features as the highest ranking input more at Muck, along with the Release Height of Resuspended Material

and Rate of Erosion. These inputs are related to the resuspension module within NewDEPOMOD, indicating that resuspension plays a bigger role at Muck. At both sites, the resuspension inputs feature more in the areas on the outside of the deposition.

Next, Figures 3.29 and 3.30 will be compared to the low energy sites, Figures 3.22 and 3.25. There are some similarities across all of the sites in the area surrounding the cages, with the Settling Velocity of Faeces playing a big role, and the Number of Cages being influential in the areas at the ends of the cage setup. The Settling Velocity of Faeces appears to have the biggest influence at West Strome, with a similar pattern seen at Ardentinny and Djuba Wick. At Ardentinny, Djuba Wick and Muck, the resuspension inputs play a bigger role in the areas on the outside of the deposition. There appear to be differences between the sites that have the same characteristics, suggesting that the grouping of sites with similar characteristics is not suitable.

Following the fitting of the random forest models, η^2 calculations were considered as a measure of sensitivity using variance decomposition. The η^2 calculations were done in the same way as the calculations for the low energy sites. Again, total η^2 values were calculated by summing the values for the first and second order effects for each input.

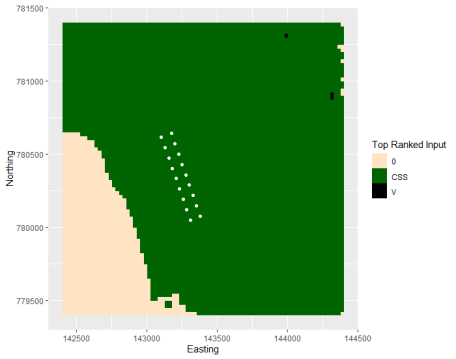


Figure 3.31: Map of the highest ranking input in each grid cell according to η^2 - Muck. (*CSS - Critical Shear Stress for Erosion, Vio - η^2 condition violated*)

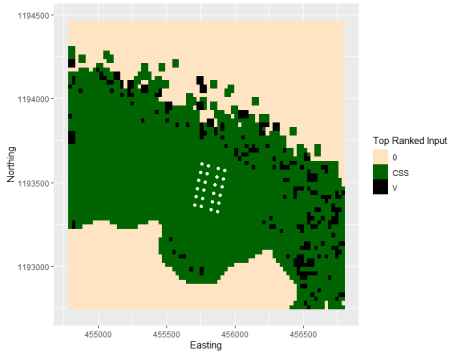


Figure 3.32: Map of the highest ranking input in each grid cell according to η^2 - Djuba Wick. (*CSS - Critical Shear Stress for Erosion, Vio - η^2 condition violated*)

As with the low energy sites, the extreme values seen across the domain will be considered using a quantile regression approach. The mean percentages of significant inputs for each quantile are given in Table 3.14. At Muck, the observations from one of the 10 samples were considered for fitting the additional models. The linear regression model identified 75% of the inputs as being significant, however the fit of the model is poor, with less than 1% of

Table 3.14: Mean % of Significant inputs for each quantile model across 10 samples

Site	0.9 Quantile	0.95 Quantile	0.99 Quantile
Djuba Wick	97.9%	98.8%	96.2%
Muck	100.0%	98.4%	79.7%

the variation explained. For the additional quantile regression models, none of the inputs were identified as being significant - indicating that the noise at the lower values is having an effect. Although there are no significant inputs identified for the 0.9, there are more inputs that were significant across the 10 samples at the 0.95 and 0.99 quantiles.

3.6 Discussion

A number of challenges were presented when considering the output maps from NewDEPOMOD within a sensitivity analysis. The idea of exploring the main shape of the impact and the variations was considered, but it meant that data was discarded from analyses, and in some cases a shape was not able to be identified.

In order to explore the full dataset, different approaches were considered. The functional representations of the maps produced surfaces that were good representations of the output maps, but the functional PCA was unable to determine the inputs responsible for the main areas of variation.

Random forest and η^2 approaches were also considered for the output maps, however, these required each grid cell in the domain to be considered independently. Considering the results from these analyses, it highlighted that each site has it's own characteristics which determine the influence of the inputs. There are some similarities seen between all of the sites, but also a number of differences, which suggest that the best approach for future work is to consider the sites separately. The Settling Velocity of Faeces was identified as the most important input across most of the domain at each site, but the influence of the resuspension inputs varies across the different sites. The fact that there are differences between sites with different characteristics suggest that the energy of a site does not determine the influence of the inputs, and that each site has to be considered individually.

Chapter 4

Emulation of Scalar Outputs

4.1 Introduction

Statistical emulation is a common technique that is used to approximate complex process-based models using statistical modelling techniques in order to reduce the computational cost (Conti & O'Hagan 2010). These complex process-based models (also referred to as simulators), are costly to run, and the building of a statistical emulator is a fundamental step when trying to gain a greater understanding of the simulator (Overstall & Woods 2016). Multiple different modelling approaches can be considered for emulation, such as linear regression, generalized linear models, regression splines and Gaussian processes (Grow & Hilton 2018).

Running NewDEPOMOD under multiple different scenarios can be computationally demanding, and take days or weeks to run depending on the number of runs required. Therefore, statistical emulation will be considered as a tool for approximating the output from NewDEPOMOD without the computational cost. This Chapter will focus on the scalar outputs considered in Chapter 2 with the aim of approximating the Total Area Impacted and the 99th Percentile of Solids Flux. These two outputs are important as the provide a measure of the size and intensity of the environmental impact of fish farms on the seabed.

The random forest models used in the sensitivity analyses in Chapter 2 were able to explain over 90% of the variation in the data. As a result, these can be used as an emulator, with predictions made using the test sets. A common method within emulation literature is to use Gaussian processes (Conti et al. 2009, Conti & O'Hagan 2010, Rajabi & Ketabchi 2017, Parker et al. 2019). However, the complexity of Gaussian processes means that, although they are much more efficient than the simulators they are approximating, the extra computational cost in comparison to more efficient, simpler regression model

approaches for fitting the models are not always beneficial in terms of the information gain, so this must be considered when emulating a simulator (Salter & Williamson 2016). Within the literature, random forests and Gaussian processes have been compared, with both showing similar predictive capabilities (Mlaker et al. 2019, Shabani et al. 2020). Gaussian processes and random forests will be considered in this Chapter for the emulation of the scalar outputs from NewDEPOMOD, with their predictive performance measured using test data.

4.2 Data being used for Emulation

To build an emulator of a given simulator, a set of costly training runs generated by the simulator are required. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ be the training input data, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})\top$ for p inputs and n input sets. For a simulator, $f(\cdot)$, the training output data is given as:

$$\mathbf{Y} = f(\mathbf{X}).$$

In other words, runs are completed for the training input data, $\mathbf{X}$, to create the training output data, $\mathbf{Y} = (y_1, \ldots, y_n)^\top$, which is then used for building the emulator model, $\hat{f}(\cdot)$. After building the emulator model, it can then be used to create approximations of the simulator for new data, $\widetilde{\mathbf{X}}$. It therefore provides a substitute for $f(\widetilde{\mathbf{X}})$, given as,

$$\hat{\mathbf{Y}} = \hat{f}(\widetilde{\mathbf{X}}).$$

To test the quality of the prediction model, the simulator is run for $\widetilde{\mathbf{X}}$, to get $\widetilde{\mathbf{Y}} = f(\widetilde{\mathbf{X}})$. This will then be compared to the emulator predictions, $\hat{\mathbf{Y}}$. The methods for comparing $\widetilde{\mathbf{Y}}$ and $\hat{\mathbf{Y}}$ will be described later in the Chapter.

Within this Chapter, the data being used as the training data, $\{\mathbf{Y}, \mathbf{X}\}$, will be the same data from the combined sensitivity analysis at the two low energy sites, Ardentinny and West Strome, and the two high energy sites, Djuba Wick and Muck, from Chapter 2. The main focus of the Chapter will be in developing a framework for the emulation using the data from Ardentinny, before applying this to the remaining sites. For each site, a smaller test set of input data, $\widetilde{\mathbf{X}}$, is created. The test sets, $\widetilde{\mathbf{X}}$, are created using the sliced LHS approach from Chapter 2, but with a total of 10 input sets in each slice. NewDEPOMOD was then run at the test sets, with a total of 5 replicate runs for each test set, before calculating the scalar outputs, $\widetilde{\mathbf{Y}}$. The performance of the statistical

emulators will be assessed by comparing their predictions, $\hat{\mathbf{Y}} = \hat{f}(\widetilde{\mathbf{X}})$, with the NewDEPOMOD output from the test sets, $\widetilde{\mathbf{Y}}$.

One of the potential investigations to be considered within this Chapter is whether or not it is possible to use a statistical emulator created for one site, and use it to predict the scalar outputs at another site. Chapter 2 showed that different inputs were more influential at different sites which suggest that every site has to be considered individually. However, it is important to be able to compare the emulator performance at each for the sites, so the data is converted to the same scale, $[0, 1]$. For each output at a given site, the training data can be expressed as, $\{\mathbf{Y}, \mathbf{X}\}$, and the test data given as $\{\widetilde{\mathbf{Y}}, \widetilde{\mathbf{X}}\}$, where $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}$, are vectors for one of the scalar outputs, and $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ are matrices, where each column features data for a given input. Each output and input parameter are considered individually for the conversion, by combining the training and test data to create a vector $(\mathbf{Y} \quad \widetilde{\mathbf{Y}})^\top$ for the output, and $(\mathbf{x}_i \quad \widetilde{\mathbf{x}}_i)$. Given a vector, $\mathbf{x}$, the formula used to transform the data to the new scale is as follow:

$$z_i = \frac{x_i - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}. \tag{4.1}$$

Equation 4.1 will be used to standardize the data for each site. It will be implemented on the combined training and test data for each output and input at every site. By standardizing the data in this way it will allow the data from all of the sites to be on the same scale to allow for better comparisons. Histograms of the training data for the two scalar outputs at Ardentinny are given in Figure 4.1 and Figure 4.2. Figure 4.1 shows that much of the data

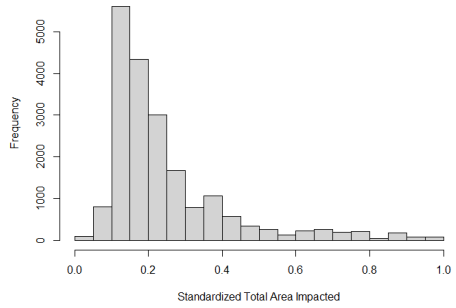


Figure 4.1: Histogram of the standardized Total Area Impacted at Ardentinny for the training data.

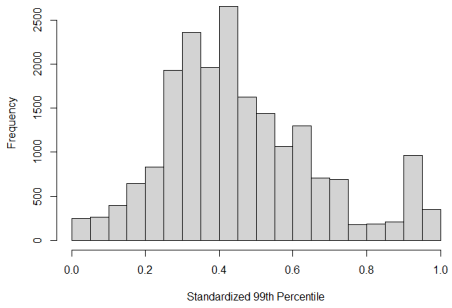


Figure 4.2: Histogram of the standardized 99th Percentile of Solids Flux at Ardentinny for the training data.

appears to be at the lower end of the scale, around 0.2. Looking at Figure 4.2, the data appears to be centred around 0.5. As the standardised data for Total

Area Impacted is slightly skewed, it could suggest that a transformation would be required, but it will initially be considered without transformation.

4.3 Methods for Emulation

The main aim of statistical emulation is to imitate a complex and computationally expensive simulator using a statistical model. It was mentioned in Chapter 1 that these emulation models can take many forms, with the main focus being to fit a model using a set of costly training runs generated by the simulator, and using the emulation model to predict the output at unknown input sets.

4.3.1 Random Forests

Random forest models, as described in Chapter 2, can be considered as a statistical emulator, and used to predict the values for the scalar outputs, Total Area Impacted and 99th Percentile of Solids Flux, using the test data. Random forests have become a popular tool for prediction methods in multiple sectors due to their flexibility and speed (Segal 2004, Zahedi et al. 2018, Iannace et al. 2019).

One aspect of producing a statistical emulator is quantifying the statistical uncertainty of the predictions. Zhang et al. (2019) explained that, within machine learning, prediction intervals for random forests are often overlooked. Meinhausen (2006) used quantile regression forests to obtain prediction intervals. This involved estimating the conditional distribution of a response variable, Y, given the predictor vector, $\mathbf{X} = \mathbf{x}$ to obtain lower and upper quantiles, $Q_L(x)\text{and}Q_U(x)$, for a prediction interval, typically a 95% prediction interval,

$$I(x) = [Q_{0.025}(x), Q_{0.975}(x)].$$

Meinhausen (2006) described the key difference between quantile regression forests and random forests as follows:

> 'for each node in each tree, random forests keeps only the mean of the observations that fall into this node and neglects all other information. In contrast, quantile regression forests keeps the values of all observations in this node, not just their mean, and assesses the conditional distribution based on this information.'

This approach to calculating prediction intervals for random forests can then be used to calculate the coverage probability for predictions.

4.3.2 Gaussian Processes

In order to create the emulators, Gaussian processes will be used due to their flexibility and their ability to capture uncertainty. They are a common method that is used for emulating complex simulators across many branches of statistics (Conti et al. 2009, Conti & O'Hagan 2010, Rajabi & Ketabchi 2017, Parker et al. 2019). Rasmussen & Williams (2006) previously described a Gaussian process as 'a collection of random variables, any finite number of which have a joint Gaussian distribution'. For a given input, $\mathbf{x}$, a Gaussian process can then be written as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')),$$

indicating that the function f is distributed as a Gaussian process with a mean function, m, and a covariance function, k (Rasmussen & Williams 2006). The mean function, m, is generally set to be zero, or a constant. The choice of covariance function is generally considered in more detail and will therefore be discussed in more detail.

4.3.2.1 Covariance Functions

Covariance functions are a crucial element when creating a Gaussian process emulator, as they define nearness or similarity of inputs to produce outputs that are also close (Rasmussen & Williams 2006). There are a range of different covariance functions that can be used, and these can be grouped into different categories (Rasmussen & Williams 2006). Generally, a function $k(\mathbf{x}, \mathbf{x}')$, mapping inputs $\mathbf{x}$, $\mathbf{x}' \in \mathbf{X}$ into $\mathbb{R}$ is referred to as a kernel, originating from the theory of integral operators (Rasmussen & Williams 2006). A function $k(\cdot)$ is said to be symmetric if $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$, which is true of a covariance function by it's definition (Rasmussen & Williams 2006).

Firstly, a stationary covariance function is a function of $\mathbf{x} - \mathbf{x}'$, therefore it is invariant to translations in the input space (Rasmussen & Williams 2006). One stationary covariance function that is often used is the squared exponential, which can have different variations to the one seen below. It is defined as follows for a pair of random variables (Rasmussen & Williams 2006):

$$\text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{|\mathrm{x} - \mathbf{x}'|^2}{2l^2}\right) + \sigma^2_{noise}\delta, \tag{4.2}$$

where $\sigma^2 > 0$ is the **signal variance**, $l > 0$ is the **lengthscale**, and $\sigma^2_{noise} \geq 0$ is the **noise variance**. The **signal variance** is a scaling factor that determines the variation of function values from their mean. Large values for σ^2 allow

for more variation in the function, however, if this value is too large it can result in the function trying to model outliers, with an illustration seen of the effect of increasing and decreasing σ^2 seen in Figure 4.3. The **lengthscale**

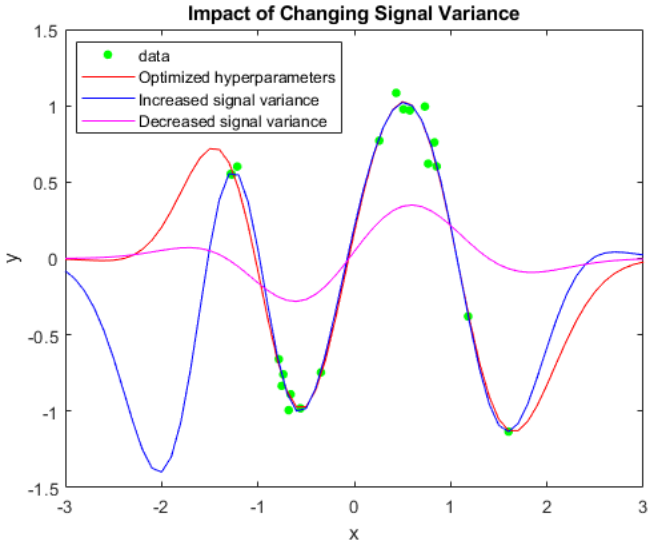


Figure 4.3: Plot illustrating the effect of increasing and decreasing the **signal variance** (σ^2) in the squared exponential covariance function, with σ^2_{noise} kept constant.

determines how smooth the function is, with small values indicating that a function values can change quickly, and large values resulting in a function that changes at a slower rate, seen in Figure 4.4. The **noise variance** allows

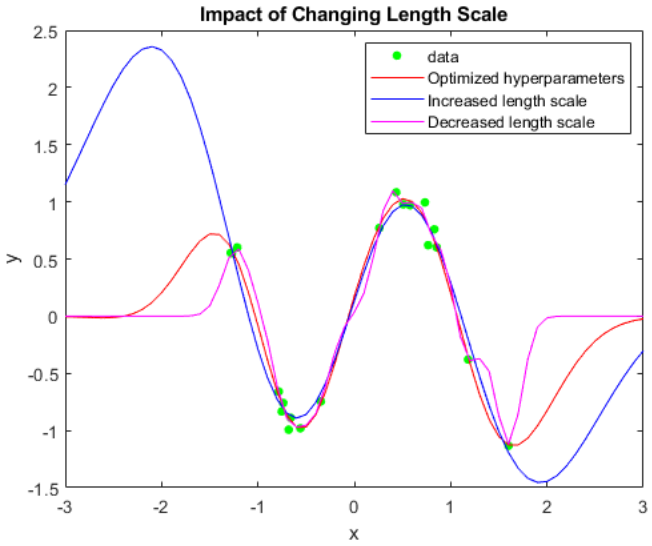


Figure 4.4: Plot to demonstrate the effect of increasing and decreasing the **length scale** (l) in the squared exponential covariance function.

the Gaussian process to account for some noise in the training data, so a value

greater than zero will mean that the uncertainty of the Gaussian process will be seen at the design points from the training data when there would have been no uncertainty if there was no noise. The squared exponential is referred to as isotropic since it is a function of $|\mathbf{x} - \mathbf{x}'|$, and is therefore invariant to all rigid motions (Rasmussen & Williams 2006). Equation 4.2 details that for inputs $\mathbf{x}$ and $\mathbf{x}'$ that are similar, the resulting covariance will be close to 1, and hence their outputs will be similar, unless the noise variance, σ^2_{noise}, is large.

Next, the dot-product covariance function depends on $\mathbf{x} \cdot \mathbf{x}'$. An example of a dot-product covariance function would be $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x} \cdot \mathbf{x}'$, which can be acquired from linear regression by placing $\mathcal{N}(0, 1)$ priors on the coefficients of x_i $(i = 1, \ldots, p)$ and a prior of $\mathcal{N}(0, \sigma_0^2)$ on the constant function, 1 (Rasmussen & Williams 2006). The dot-product covariance function is not used often, but has been successfully used in high-dimensional classification problems (Rasmussen & Williams 2006).

When using Gaussian processes, the covariance functions can include large numbers of hyperparameters (such as lengthscale), and the information known about their values is rather vague. As a result, it is essential that methods are available to assist with the selection of the form of the covariance function and it's hyperparameters (Rasmussen & Williams 2006). Within each choice of covariance function, there are a number of possibilities for the different hyperparameters. It is therefore essential to be able to compare Gaussian process models with different values for hyperparameters, different covariance function shapes, and even with models that are not Gaussian processes (Rasmussen & Williams 2006). Rasmussen & Williams (2006) referred to the selection of the covariance function and its hyperparameters as '*training* of a Gaussian process'.

Covariance functions, such as the squared exponential in Equation 4.2, can be parameterized in terms of the hyperparameters, demonstrated below.

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top V (\mathbf{x} - \mathbf{x}')\right) + \sigma^2_{noise}\delta \tag{4.3}$$

Here, $\boldsymbol{\theta} = (\{V\}, \sigma^2, \sigma^2_{noise})^\top$ is a vector that contains all the hyperparameters. The noise parameter, σ^2_{noise}, is not always considered a hyperparameter, but it plays a similar role, therefore it is treated as a hyperparameter in these circumstances. $\{V\}$ represents the symmetric matrix V that contains the parameters, which can be denoted by the following,

$$V_1 = \ell^{-2} I, \quad V_2 = \text{diag}(\boldsymbol{\ell})^{-2}, \quad V_3 = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \text{diag}(\boldsymbol{\ell})^{-2}. \tag{4.4}$$

Here, $\boldsymbol{\ell}$ denotes a vector of positive values, and $\boldsymbol{\Lambda}$ is a $D \times k$ matrix where

$k < D$ (Rasmussen & Williams 2006). In the case of the squared exponential covariance function in Equation 4.3, using $\boldsymbol{\ell} = \ell_1, \ldots, \ell_D$ and V_2 from 4.4, the hyperparameters ($\boldsymbol{\ell}$) can be described as 'characteristic lengthscales'. Rasmussen & Williams (2006) explained it in simpler terms - 'how far do you need to move (along a particular axis) in input space for the function values to become uncorrelated'. Using this approach for the squared exponential covariance function, the inverse of the lengthscale dictates how relevant an input is, which therefore applies automatic relevance determination ('ARD') (Neal 1996). In other words, the larger the lengthscale, the less of an effect the input will have on the covariance, and therefore reducing its involvement in the inference. In the case where there are multiple lengthscales for the different predictor variables (V_2 in Equation 4.4), the lengthscales can be used to determine whether or not all the predictor variables are required. If a Gaussian process model is fitted using all of the predictor variables, the Mean Squared Error ('MSE') can be calculated, and then compared to a model fitted without the predictor variable corresponding to the largest lengthscale. If the MSE remains at a similar value, then it could be said that the simpler model with less predictors could be used to emulate the output. Using Gaussian processes presents a number of major challenges such as the need to choose and define the structures of the hyperparameters, specifically, the lengthscale and signal variance.

4.3.2.2 Model selection

As previously stated, it is essential to be able to compare different models and it requires a systematic and practical approach to model selection (Rasmussen & Williams 2006). Model selection will refer to the choice of the functional form of the covariance function, as well as the values of any hyperparameters included in the Gaussian process. There are three general principles that cover many of the different methods available for model selection (Rasmussen & Williams 2006):

1. Compute the probability of the model given the data.

2. Calculate an estimate of the generalization error - this is the average error on unseen test examples.

3. Produce bounds for the generalization error.

The main approaches for model selection include a Bayesian approach, where the marginal likelihood is used in the computation of the probability of the

model given the data, and the other approach involves using cross validation. Using a Bayesian approach, the marginal likelihood includes a complexity penalty term to automatically incorporate a trade-off between model fit and model complexity (Rasmussen & Williams 2006). The general idea of cross validation is to split the training set into two disjoint sets, one of which is a 'validation set' that is used to monitor performance.

The marginal likelihood is used in a model selection scenario to calculate the probability of the data, $\mathbf{y}$, given the model, $\mathcal{M}$. For a set of model parameters $\mathbf{\Lambda}$, the marginal likelihood of $\mathcal{M}$ is given by:

$$p(\mathbf{y}|\mathcal{M}) = \int p(\mathbf{y}|\mathbf{\Lambda}, \mathcal{M}) p(\mathbf{\Lambda}|\mathcal{M}) d\mathbf{\Lambda}. \tag{4.5}$$

In the case of Gaussian Processes, Rasmussen & Williams (2006) referred to a hierarchical specification of models that was described by MacKay (1992) - (1) the lowest level included the parameters, $\mathbf{w}$, (2) the second level included the hyperparameters, $\boldsymbol{\theta}$ and (3) the top level which included a discrete set of possible model structures, $\mathcal{H}_i$. Bayesian inference takes place on a level by level basis, and using Bayes' rule, the posterior over the parameters, $\mathbf{w}$, the hyperparameters, $\boldsymbol{\theta}$, and the model structure, $\mathcal{H}_i$ can be expressed as follows (Rasmussen & Williams 2006):

$$p(\mathbf{w}|\mathbf{y}, X, \boldsymbol{\theta}, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i) p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i)}{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i)} \tag{4.6}$$

$$p(\boldsymbol{\theta}|\mathbf{y}, X, \mathcal{H}_i) = \frac{p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i) p(\boldsymbol{\theta}|\mathcal{H}_i)}{p(\mathbf{y}|X, \mathcal{H}_i)} \tag{4.7}$$

$$p(\mathcal{H}_i|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathcal{H}_i) p(\mathcal{H}_i)}{p(\mathbf{y}|X)}, \tag{4.8}$$

In each of the equations, the two elements on the numerator are known as the *likelihood* and the *prior*, and the normalizing constant on the denominator is referred to as the *marginal likelihood*, or in some instances, the evidence. In Equation 4.8, the prior $p(\mathcal{H}_i)$ over the model structures is such that it does not allow any of the models to be favoured over another, so it is often taken to be flat (Rasmussen & Williams 2006). In each case, the marginal likelihood is

calculated as follows (Rasmussen & Williams 2006):

$$p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i) = \int p(\mathbf{y}|X, \mathbf{w}, \mathcal{H}_i) p(\mathbf{w}|\boldsymbol{\theta}, \mathcal{H}_i)\, d\mathbf{w} \tag{4.9}$$

$$p(\mathbf{y}|X, \mathcal{H}_i) = \int p(\mathbf{y}|X, \boldsymbol{\theta}, \mathcal{H}_i) p(\boldsymbol{\theta}|\mathcal{H}_i)\, d\boldsymbol{\theta} \tag{4.10}$$

$$p(\mathbf{y}|X) = \sum_i p(\mathbf{y}|X, \mathcal{H}_i) p(\mathcal{H}_i). \tag{4.11}$$

The execution of Bayesian inference requires the above integrals to be evaluated and for certain models it may be required that analytical approximations are required. One approximation that is used is Type II maximum likelihood, where the marginal likelihood in equation 4.9 is maximised with respect to the hyperparameters, $\boldsymbol{\theta}$, rather than evaluating equation 4.10 (Rasmussen & Williams 2006). In the case where there are many hyperparameters, this approximation can result in overfitting. Rasmussen & Williams (2006) advised that the Laplace approximation could be used to approximate the integral in equation 4.10, and that it is a good approximation in the case that the posterior over $\boldsymbol{\theta}$ is well peaked. As previously mentioned, marginal likelihood's automatic trade-off between model fit and model complexity, make it useful in model selection problems. An example of this trade-off is described by Rasmussen & Williams (2006) and is seen in the Figure 4.5. The 3 different

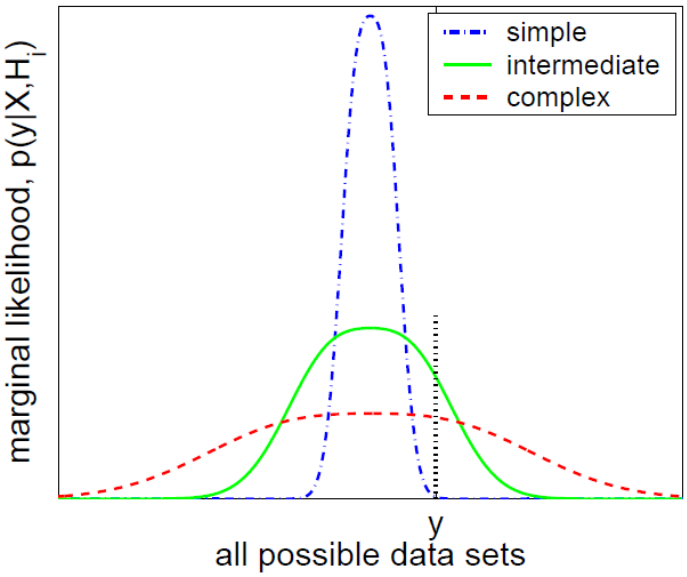


Figure 4.5: Figure from Rasmussen & Williams (2006) illustrating the trade-off between model fit and model complexity.

models are created with the same number of inputs, X, and the same number of data points, n. The y-axis of Figure 4.5 represents the marginal likelihood,

$p(\mathbf{y}|X, \mathcal{H}_i)$, and the x-axis represents the possible vectors of the target data, $\mathbf{y}$. The more complex the model, the wider the range of possible target vectors, y that could be accounted for. The marginal likelihood is a probability distribution over y, and so it must sum to 1, so more complex models have wider and lower peaks as they can account for a larger number of possible target vectors, y. In Figure 4.5, a particular dataset, y, is highlighted to show that model of intermediate complexity is preferred over the other two using a marginal likelihood approach. A marginal likelihood is therefore useful in the selection of a model that is suited to the data in terms of complexity, but care must be taken when approximations of the marginal likelihood are made.

Moving on to look at the cross validation method of model selection, the standard cross validation method of splitting the data into two disjoint sets can be improved by using k-fold cross validation to reduce the variance of the performance estimate (Rasmussen & Williams 2006). K-fold cross validation requires the training set to be split up into k disjoint, equally-sized subsets, where training is completed using the union of $k-1$ subsets. This is repeated k times with a different subset being used for the validation (Rasmussen & Williams 2006). Another method of cross validation is 'Leave-one-out' cross validation, where each observation is removed and used as test data, while the model is trained using the remaining observations. This is approach is therefore computationally expensive, as for a large number of observations, an equally large number of models are trained. For k-fold cross validation, Rasmussen & Williams (2006) advised that values for k tend to vary from 3 to 10. The hyperparameter choices are an important part of Gaussian process modelling, and will therefore be considered in more detail.

4.3.2.3 Hyperparameter optimization

Within Gaussian process models, and other machine learning models, the optimization of the hyperparameters is an essential task. These optimization processes often begin with initial hyperparameter values being specified, before optimizing a cost function via gradient-based methods (Ulapane et al. 2020).

Quasi-Newton methods (Davidon 1991) are a class of optimization algorithms which are based on Newton methods, but can be used in the case where the Jacobian or Hessian matrix are not available or computational expensive to compute at each iteration. One of the most widely used quasi-Newton methods is the BFGS optimization, which was proposed independently by Broyden, Fletcher, Goldfarb and Shanno in 1970 (Broyden 1970, Fletcher 1970, Goldfarb 1970, Shanno 1970). The BFGS algorithm is a type of second-order optimiza-

tion that approximates the Hessian matrix, using the gradient. The way that the inverse Hessian is calculated is different across the different quasi-Newton algorithms, and the BFGS is one specific way for updating this calculation, instead of recalculating it during every iteration. It is recognised as one of the most popular quasi-Newton algorithms (Nocedal & Wright 2006).

Consider a real-valued, differentiable objective function, $f(\mathbf{x})$. To find a local minimum, Newton's method uses an iterative scheme, with the following update at each iteration,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - H(\mathbf{x})^{-1}\nabla f(\mathbf{x}_k),$$

where $H(\mathbf{x}_k)^{-1}$ is the inverse Hessian matrix, and $\nabla f(\mathbf{x}_k)$ is the gradient, where, for each step, the Hessian is computed and inverted. Newton's method has two main disadvantages:

1. It is sensitive to the initial conditions - the iterative process could lead to a local maximum or a saddle point rather than a minimum.

2. It is computationally expensive - the computation of $H(\mathbf{x}_k)^{-1}$ scales as $\mathcal{O}(n^3)$.

In order to address the computational cost of Newton's method, the quasi-Newton method was developed. In order to improve the computational time, the quasi-Newton method uses an approximation of the Hessian matrix, B, which is a positive definite matrix that is updated between iterations using information from the previous steps. Any quasi-Newton method has one condition, known as the *quasi-Newton condition*, that the Hessian approximation, B must satisfy:

$$B_{k+1}[\mathbf{x}_{k+1} - \mathbf{x}_k] = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k). \tag{4.12}$$

This condition is obtained from the first order Taylor expansion of $\nabla f(\mathbf{x}_{k+1})$ about $\nabla f(\mathbf{x}_k)$. The updated Hessian approximation, B_{k+1}, from Equation 4.12, can be calculated in different ways, but with a common theme being that it only uses the previous gradient information. In addition, Equation 4.12 will be rewritten to simplify future equations, with $[\mathbf{x}_{k+1} - \mathbf{x}_k] = \Delta\mathbf{x}_k$, and $\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) = \mathbf{y}_k$.

$$B_{k+1}\Delta\mathbf{x}_k = \mathbf{y}_k \tag{4.13}$$

One drawback to the quasi-Newton condition in Equation 4.12, is that it is underdetermined for $n > 1$ dimensions. As a result, further additional constraints are required for the update method for B.

The BFGS method is a type of quasi-Newton method, and is considered to be one of the most popular (Nocedal & Wright 2006). The different quasi-Newton methods place constraints on the Hessian approximation, B, and the next steps will focus on the BFGS method. In addition to the quasi-Newton condition in Equation 4.13, the BFGS method imposes two additional constraints on the updating scheme for the Hessian approximation, B.

- B_k and B_{k+1} are characterized as being close. In other words, $\min_{B_{k+1}} \|B_{k+1} - B_k\|$.
- B_{k+1} is symmetric and positive-definite, $B_{k+1}^\top = B_{k+1}$
- Finally, the quasi-Newton condition, $B_{k+1}\Delta\mathbf{x}_k = \mathbf{y}_k$

Consider again Newton's method in Equation 4.3.2.3 and notice that it is the inverse of the Hessian matrix that is being considered, so the above constraints have to be altered to account for the inverse of the Hessian approximation, B.

- $\min_{B_{k+1}^{-1}} \|B_{k+1}^{-1} - B_k^{-1}\|$,
- $(B_{k+1}^{-1})^\top = B_{k+1}^{-1}$,
- and $\Delta\mathbf{x}_k = B_{k+1}^{-1}\mathbf{y}_k$.

In other words, the constraints mean that the change in B^{-1} at each iteration is minimized, subject to B^{-1} being symmetric, and the inverted quasi-Newton condition holding, as well as B^{-1} being positive-definite. The matrix norm that is used in the BFGS method is the Frobenius norm:

$$\|A\|_F = \sqrt{\sum_i^m \sum_j^n |a_{ij}|^2}.$$

A detailed derivation of the conditions for ${B_{k+1}}^{-1}$ can be found in Nocedal & Wright (2006), with the final setups described here. The derivation by Nocedal & Wright (2006) leads to the approximate Hessian at each iteration being updated using:

$$B_{k+1} = B_k + U_k + V_k, \tag{4.14}$$

where U and V are symmetric, rank-one matrices of the form, $U = a\mathbf{u}\mathbf{u}^\top$ and $V = b\mathbf{v}\mathbf{v}^\top$, with $\mathbf{u}$ and $\mathbf{v}$ being linearly independent, non-zero vectors and a and b are constants. The matrices U and V are both symmetric, therefore the approximate Hessian update in Equation 4.14 results in B being symmetric

following each iteration. Substituting in our values for U and V, we get,

$$B_{k+1} = B_k + a\mathbf{u}\mathbf{u}^\top + b\mathbf{v}\mathbf{v}^\top. \tag{4.15}$$

As both $a\mathbf{u}\mathbf{u}^\top$ and $b\mathbf{v}\mathbf{v}^\top$ are rank-one, their sum is rank-two, which is known as a rank-two update. This rank-two update allows the condition of closeness between B_k and B_{k+1} to be guaranteed. The next step is to consider the quasi-Newton condition, Equation 4.13.

$$\begin{aligned} B_{k+1}\Delta\mathbf{x}_k &= \mathbf{y}_k \\ B_k\Delta\mathbf{x}_k + a\mathbf{u}\mathbf{u}^\top\Delta\mathbf{x}_k + b\mathbf{v}\mathbf{v}^\top\Delta\mathbf{x}_k &= \mathbf{y}_k. \end{aligned}$$

Choosing $\mathbf{u} = \mathbf{y}_k$ and $\mathbf{v} = B_k\Delta\mathbf{x}_k$, we then have,

$$B_k\Delta\mathbf{x}_k + a\mathbf{y}_k\mathbf{y}_k^\top\Delta\mathbf{x}_k + bB_k\Delta\mathbf{x}_k\Delta\mathbf{x}_k^\top B_k^\top\Delta\mathbf{x}_k = \mathbf{y}_k \tag{4.16}$$

$$\mathbf{y}_k(1 - a\mathbf{y}_k^\top\Delta\mathbf{x}_k) = B_k\Delta\mathbf{x}_k(1 + b\Delta\mathbf{x}_k^\top B_k^\top\Delta\mathbf{x}_k). \tag{4.17}$$

Solving Equation 4.17, produces the following for a and b,

$$\begin{aligned} a &= \frac{1}{\mathbf{y}_k^\top\Delta\mathbf{x}_k}, \\ b &= -\frac{1}{\Delta\mathbf{x}_k^\top B_k^\top\Delta\mathbf{x}_k}. \end{aligned}$$

Using these values for a and b, and substituting them back in to Equation 4.15, produces the BFGS update:

$$B_{k+1} = B_k + \frac{1}{\mathbf{y}_k^\top\Delta\mathbf{x}_k} - \frac{B_k\Delta\mathbf{x}_k\Delta\mathbf{x}_k^\top B_k^\top}{\Delta\mathbf{x}_k^\top B_k^\top\Delta\mathbf{x}_k}. \tag{4.18}$$

This iterative formula for the approximate Hessian uses only the previous gradient information to update it. In practice, referring back to the Newton method in Equation 4.3.2.3, it is the inverse Hessian matrix that is required. Therefore, Equation 4.18, will have to be inverted, which can be done using the Woodbury formula (Woodbury 1950). This formula provides a way to invert the sum of an invertible matrix, A and a rank-m correction.

$$(A + SCT)^{-1} = A^{-1} - A^{-1}S(C^{-1} + TA^{-1}S)^{-1}TA^{-1}. \tag{4.19}$$

In order obtain the inverse of B from the BFGS formula, Equation 4.18 has to

be rewritten in a more suitable form:

$$B_{k+1} = B_k + \underbrace{\begin{pmatrix} B_k \Delta \mathbf{x}_k & \mathbf{y}_k \end{pmatrix}}_{S} \underbrace{\begin{pmatrix} -\frac{1}{\Delta \mathbf{x}_k^\top B_k \Delta \mathbf{x}_k} & 0 \\ 0 & \frac{1}{\mathbf{y}_k^\top \Delta \mathbf{x}_k} \end{pmatrix}}_{C} \underbrace{\begin{pmatrix} \Delta \mathbf{x}_k^\top B_k \\ \mathbf{y}_k^\top \end{pmatrix}}_{T}. \tag{4.20}$$

Using the Woodbury formula and the values for S, C and T in Equation 4.20, the matrix manipulation produces the following result for the inverse of the Hessian approximation.

$$B_{k+1} = \left(I - \frac{\Delta \mathbf{x}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \Delta \mathbf{x}_k} \right) B_k^{-1} \left(I - \frac{\mathbf{y}_k \Delta \mathbf{x}_k^\top}{\mathbf{y}_k^\top \Delta \mathbf{x}_k} \right) + \frac{\Delta \mathbf{x}_k \Delta \mathbf{x}_k^\top}{\mathbf{y}_k^\top \Delta \mathbf{x}_k}. \tag{4.21}$$

Equation 4.21 provides the detail for the computation that is required within the BFGS approach to optimization, using B^{-1} instead of calculating the Hessian matrix at each iteration for Equation 4.3.2.3. This updating of the approximate Hessian matrix removes the $\mathcal{O}(n^3)$ operations of inverting the Hessian matrix in the original Newton method. There are two common ways to initialize B_0^{-1} in practice:

1. Set B_0^{-1} to the identity matrix, I.
2. Compute and invert the true Hessian at the initial point and use the BFGS approach to update it.

The disadvantage of the second approach would be that there is an initial cost of computing the true Hessian and then inverting it. Although the BFGS method was computationally more efficient than Newton's method, its computational efficiency was further improved through the introduction of the limited memory BFGS approach ('L-BFGS') (Nocedal 1980, Liu & Nocedal 1989). The BFGS approach requires the storage of an $n \times n$ approximation of the inverse Hessian matrix, whereas the L-BFGS approach requires the storage of a small number of vectors that represent the approximation. It is able to reduce the computational storage that would be required for the BFGS method, and is therefore an effective approach with larger datasets (Liu & Nocedal 1989).

Due to the large datasets being used within this thesis, the L-BFGS approach for optimization will be considered when optimizing the hyperparameter values within a Gaussian process.

4.3.2.4 Sparse Gaussian Processes

Sparse Gaussian processes are a useful tool when fitting models for large data sets as they can reduce the computational cost significantly. There are multi-

ple methods for using sparse Gaussian processes such as Pseudo-input approximation (Snelson & Ghahramani 2006), subset of data approaches (Silverman 1985, Smola & Bartlett 2001) and variational approximations (Titsias 2009, Matthews et al. 2016). Exact Gaussian process regression requires the inversion of an $n \times n$ matrix, which becomes computationally expensive, with large storage requirements for large data sets with n observations. Approximate Gaussian process regression is a common approach used for large data sets to avoid the large computational cost and storage demands.

One approach to reducing the computational complexity is to use the 'subset of data approximation'. In order to overcome the inversion of the $n \times n$ matrix, a selection of $m \ll n$ of the total n observations are used to then apply the exact Gaussian process regression fitting methods. This set of m points is often referred to as the active set, in this case it will be called $\mathcal{A}$. Using this approach reduces the size of the matrix to be inverted, producing the reduced computational complexity as well as a smaller kernel matrix to be stored in comparison. A simple approach to selecting the active set, $\mathcal{A}$, would be to choose the points at random, but experimental studies have shown that this can result in poor results (Lawrence et al. 2003). The selection of the inducing points has been investigated in the literature, with different approaches considered, such as greedy algorithms (Smola & Schölkopf 2000, Smola & Bartlett 2001, Lawrence et al. 2003), and variational approaches (Titsias 2009). The sparse greedy matrix approximation (Smola & Schölkopf 2000, Smola & Bartlett 2001) is a greedy apporach that can be used within sparse Gaussian processes for selecting the active set, $\mathcal{A}$. The subset of data approach is often considered as the simplest form of sparse Gaussian processes (Quinonero-Candela et al. 2007), but is less computationally demanding than other approaches (Quinonero-Candela et al. 2007).

An alternative approach for the approximation is the 'subset of regressors' which was proposed by Wahba (1990) and further summarised in Rasmussen & Williams (2006). This approach involves replacing the kernel function, $k(\mathbf{x}, \mathbf{x}_i)$, with an approximation, $\hat{k}_{SR}(\mathbf{x}, \mathbf{x}_i|\mathcal{A})$, given the active set $\mathcal{A} \subset \mathcal{N} = \{1, \ldots, n\}$, where $\mathcal{N}$ is the set of indices for all observations. For an exact Gaussian process approach, the set of $\mathcal{N}$ functions $\mathcal{S}_{\mathcal{N}} = \{k(\mathbf{x}, \mathbf{x}_i), i = 1, \ldots, n\}$ is used to calculate the expected prediction. The subset of regressors approach uses the set of functions $\mathcal{S}_{\mathcal{A}} = \{k(\mathbf{x}, \mathbf{x}_j), j \in \mathcal{A}\}$, to approximate the span of the functions in $\mathcal{S}_{\mathcal{N}}$. Consider the kernel function $k(\mathbf{x}, \mathbf{x}_i)$, for $i \in \mathcal{N}$, then the

approximation can be calculated using functions from $\mathcal{S}_{\mathcal{A}}$:

$$\hat{k}(\mathbf{x}, \mathbf{x}_i) = \sum_{j \in \mathcal{A}} \alpha_{ji} k(\mathbf{x}, \mathbf{x}_j). \tag{4.22}$$

Here, $\alpha_{ji} \in \mathbb{R}$, are the corresponding coefficients for the linear combinations of the elements of $\mathcal{S}_{\mathcal{A}}$ which are used to approximate $k(\mathbf{x}, \mathbf{x}_i)$. Allow $\boldsymbol{\alpha}$ to be the $|\mathcal{A}| \times n$ matrix containing all of the coefficient values, α_{ji}. The following error function is minimized to find the best approximation of the elements of $\mathcal{S}_{\mathcal{N}}$ using the linear combinations from Equation 4.22.

$$E(\mathcal{A}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \|k(\mathbf{x}, \mathbf{x}_i) - \hat{k}(\mathbf{x}, \mathbf{x}_i)\|^2. \tag{4.23}$$

The corresponding coefficient matrix $\boldsymbol{\alpha}$ that minimizes Equation 4.23 is then given by,

$$\hat{\boldsymbol{\alpha}}_{\mathcal{A}} = K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}})^{-1} K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}). \tag{4.24}$$

Here $K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}})$ represents the covariance function matrix, with each element corresponding to $k(\mathbf{x}_a, \mathbf{x}_b)$ for $(a, b) \in \mathcal{A}$. The kernel approximation from Equation 4.22 can then be expressed in matrix form:

$$\hat{k}(\mathbf{x}, \mathbf{x}_i) = \sum_{j \in \mathcal{A}} \alpha_{ji} k(\mathbf{x}, \mathbf{x}_j) = K(\mathbf{x}^\top, \mathbf{X}_{\mathcal{A}}) \boldsymbol{\alpha}(:, i). \tag{4.25}$$

Following this, the subset of regressors approximation to the kernel function is given as:

$$\hat{k}_{SR}(\mathbf{x}, \mathbf{x}_i) = K(\mathbf{x}^\top, \mathbf{X}_{\mathcal{A}}) \hat{\boldsymbol{\alpha}}_{\mathcal{A}}(:, i) = K(\mathbf{x}^\top, \mathbf{X}_{\mathcal{A}}) K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}})^{-1} K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}), \tag{4.26}$$

using the coefficient matrix from Equation 4.24. Then considering the overall kernel function matrix, $K(\mathbf{X}, \mathbf{X})$ is defined as:

$$\hat{K}_{SR}(\mathbf{X}, \mathbf{X}) = K(\mathbf{X}, \mathbf{X}_{\mathcal{A}}) K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}})^{-1} K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}). \tag{4.27}$$

As with the subset of data approach, the subset of regressors approach requires the points to be selected for the active set, $\mathcal{A}$, which can again be done using a greedy algorithm, such as the sparse greedy matrix approximation (Smola & Schölkopf 2000, Smola & Bartlett 2001). One drawback to the subset of regressors approach is the unreasonably small predictive variances that can be produced when making predictions far away from the active set (Rasmussen & Williams 2006).

An alternative approach to the subset of regressors that overcomes the predictive variance problem is 'fully independent conditional approximation' ('FIC') (Candela 2005). This approach also approximates the kernel function, without the predictive variance problem that was mentioned for the subset of regressors approach. Given the active set, $\mathcal{A}$, the FIC approximation of $k((\mathbf{x}_p, \mathbf{x}_q)$, for $(p.q) \in \mathcal{N}$ is given by:

$$\hat{k}_{FIC}(\mathbf{x}_p, \mathbf{x}_q) = \hat{k}_{SR}(\mathbf{x}_p, \mathbf{x}_q) + d_{pq}\left(k((\mathbf{x}_p, \mathbf{x}_q) - \hat{k}_{SR}(\mathbf{x}_p, \mathbf{x}_q)\right), \tag{4.28}$$

where $d_{pq} = 1$ if $p = q$, or $d_{pq} = 0$ if $p \neq q$. In other words, it uses the exact kernel value rather than the approximation when $p = q$. To assist with the matrix version of Equation 4.28, define an $n \times n$ diagonal matrix, $\mathbf{\Delta}(X)$ as:

$$[\mathbf{\Delta}(X)]_{pq} = d_{pq}\left(k((\mathbf{x}_p, \mathbf{x}_q) - \hat{k}_{SR}(\mathbf{x}_p, \mathbf{x}_q)\right) \tag{4.29}$$

$$= \begin{cases} k(\mathbf{x}_p, \mathbf{x}_q) - \hat{k}_{SR}(\mathbf{x}_p, \mathbf{x}_q) & \text{if } p = q, \\ 0 & \text{if } p \neq q. \end{cases} \tag{4.30}$$

Using the matrix, $\mathbf{\Delta}(X)$, the matrix form of the FIC approximation of $K(X, X)$ is defined as:

$$\hat{(K)}_{FIC}(X, X) = \hat{K}_{SR}(\mathbf{X}, \mathbf{X}) + \mathbf{\Delta}(X) \tag{4.31}$$

$$= K(\mathbf{X}, \mathbf{X}_{\mathcal{A}})K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}})^{-1}K(\mathbf{X}_{\mathcal{A}}, \mathbf{X}) + \mathbf{\Delta}(X) \tag{4.32}$$

As with the previous approaches, the inducing points in the active set have to be defined, and are often done using a greedy algorithm.

Each of the sparse approximation approaches described can be implemented and their performance measured to determine the most appropriate approach to the NewDEPOMOD data.

4.3.3 Measurements of predictive performance for comparing emulators

As the statistical emulators are being created to predict the NewDEPOMOD output at unknown input sets, the measure of their performance is an important aspect. One such measure to be considered is the Root Mean Squared Error ('RMSE') which is a common measure of predictive performance. It is given

by the following formula,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}{N}},$$

where $\hat{\mathbf{y}}$ is a set of N predicted values, and $\mathbf{y}$ is a set of N observed values. One benefit of the RMSE is that, because the data has been standardized, comparisons of the RMSE for different emulators can be made. However, one potential drawback to it, is that it is sensitive to outliers, with the square of the errors resulting in a much larger effect for the outliers (Willmott & Matsuura 2005). In some cases though, it is reasonable that the effect for outliers should be punished more, as an outlier could have a much larger impact in some modelling scenarios. Therefore this should be reflected in the measure of the performance by giving the outliers a larger weighting (Chai & Draxler 2014). In the case of NewDEPOMOD, a larger under or over-prediction could result in irreparable damage to the seabed being missed in the case of under-prediction, or resources being wasted monitoring a site more closely in the case of over-prediction of the environmental impacts.

In the case where RMSE is not appropriate, an alternative that is often considered is the Mean Absolute Error ('MAE'). The MAE is given as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|.$$

The MAE is therefore not sensitive to outliers in the same way that RMSE is, as the MAE provides equal weight to the errors, where the squared element of RMSE penalised predictions that were further from the observed value (Chai & Draxler 2014).

Additionally, when assessing the predictive performance of emulators, bias is an important tool. Bias is a measure of how close the predicted values are to the true values. In order to calculate the overall bias for a set of predictions, the differences between the predicted values and the true, observed values is averaged.

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{y}_i.$$

This equation calculates an average of the differences between predictions and observed values, where an unbiased estimator would produce a value of zero. In the case where the estimator is biased, underpredictions will be identified by a positive value for Bias, and overpredictions will be identified by a negative

value.

Further to the above measurements, coverage probability is a measure of the proportion of observed values that are in a given prediction interval for a set of predictions. Coverage probability accounts for uncertainty in predictions, including it in the measure of how well the model predicts. For a set of predictions to have good coverage and indicate a good emulator model, it would be expected that the coverage probability would be greater than 0.95, indicating that the prediction intervals are a good representation of the observed values. However, if there is some bias present within the emulator, this would impact the coverage probability, producing lower values than would be expected.

Throughout this Chapter, the RMSE, MAE, Bias and coverage will be considered as measures of the predictive performance of the emulators, where appropriate.

4.4 Results from Emulation

This section will provide an overview of the results for each of the methods being considered for emulation, before comparing the results from the two approaches in a further section later in the Chapter.

4.4.1 Random forest emulation

As was mentioned previously, the data from the combined sensitivity analysis in Chapter 2 was used as the training data to fit the statistical emulators. The first site to be considered will be Ardentinny, which has a total of 20,000 runs that were included in the training set. As before, the 20,000 runs include 400 different input sets, at which NewDEPOMOD was run 50 times to create replicate runs for each input set. The test data at Ardentinny consists of 80 different input sets, which were run 5 times each to create a total of 400 runs.

4.4.1.1 Total Area Impacted

Using the standardized data, a random forest model was fitted to the training data, with Total Area Impacted as the output, and the five continuous physical properties inputs, and the three categorical operational inputs. As was seen in Chapter 2, the random forest model is able to explain approximately 94% of the variation in the training data, indicating that it is a very good fit. In addition, the random forest model was created in less than a minute, which is fast when considering the amount of data being used, and will be compared

with the Gaussian process approach. The next challenge for this model is to assess how well it performs at predicting the Total Area Impacted for the test data. Using the fitted random forest model, the input test data was used to create predictions of the Total Area Impacted. NewDEPOMOD was run using the test data to provide the output that the predictions will be compared against. Using the RMSE and the MAE, the performance of the random forest predictions can be assessed. The values for RMSE and MAE are fairly close

RMSE	**MAE**	**Bias**	**Coverage**
0.078	0.050	0.0078	0.97

Table 4.1: Table of the predictive performance of the random forest model for Total Area Impacted - Ardentinny.

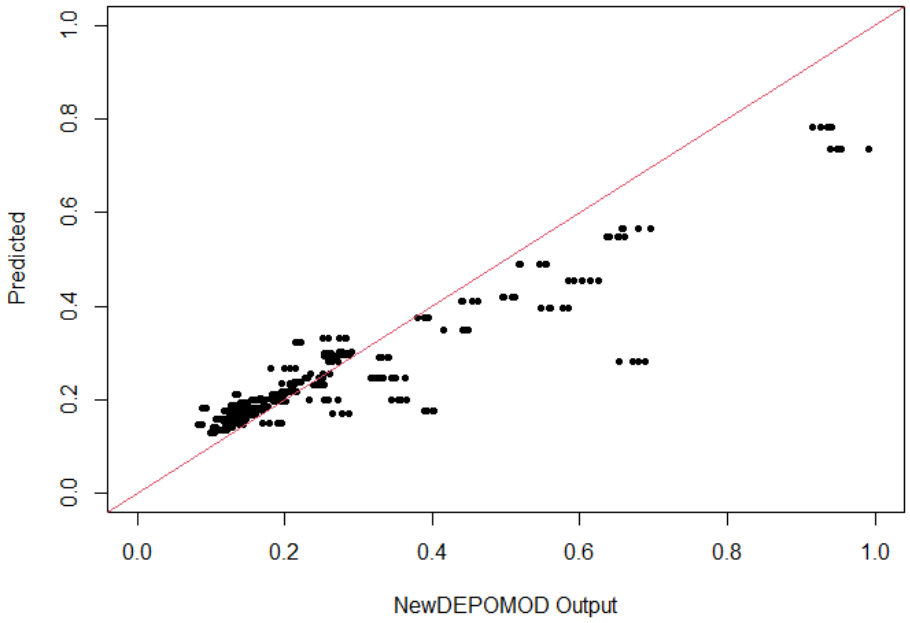


Figure 4.6: Plot of the predicted Total Area Impacted from the random forest model against the output from NewDEPOMOD - Ardentinny.

to zero, when considering the data are on the scale $[0, 1]$, indicating reasonably good performance. This is also supported by the large value for the coverage probability. This is emphasized when considering the plot of the predicted values against the output from NewDEPOMOD for the test data. The data shows a linear pattern, which is close to the line of equality, with some slight underpredictions where the NewDEPOMOD output is greater than 0.3. The underpredictions are highlighted by the positive value for bias. From Figure

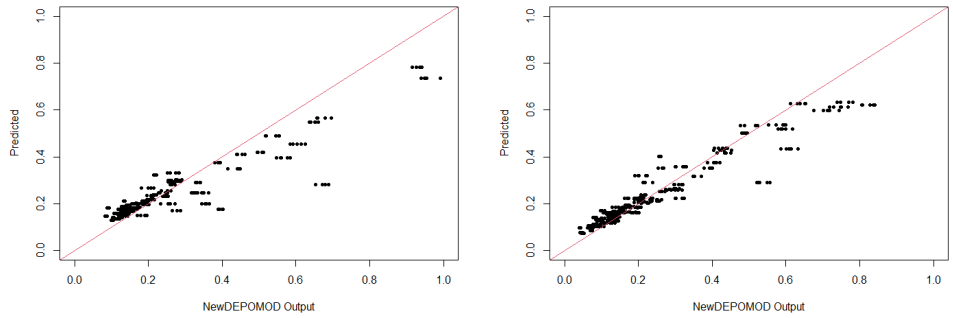
4.7a, the positive bias is related to the underpredictions seen for output values greater than 0.3. This is likely a result of the smaller amount of data available for these larger values of Total Area Impacted, seen in Figure 4.1. In general though, there seems to be reasonable agreement between the random forest predictions and the output from NewDEPOMOD at Ardentinny. In addition, the importance values of the random forest model, calculated using Equation 2.9 from Chapter 2, are similar to the values from Table 2.13, which would be expected as the only change to the data is standardizing by adding in the data from the test set, $\widetilde{\mathbf{X}}$.

Next, the other remaining sites will be considered. West Strome is the other low energy site, which had a total of 22,500 runs from 450 different input sets, together with the test set consisting of 450 runs from 90 different input sets. At Muck, one of the high energy sites, it had a total of 4,500 runs from 450 input sets in the training data and 450 runs from 90 different input sets for the test data. At the other high energy site, Djuba Wick, it had 4,000 runs from 400 input sets in the training data, and 400 runs from 80 input sets in the test data. The reason for the differences in the number of input sets between sites is described in Chapter 2, but relates to the additional operational setup that is required at West Strome and Muck based on the setup of the combined analysis. In addition, it was explained in Chapter 2 that a reduced number of replicate runs were considered for the high energy sites due to the increased computational cost of running for sites with faster current speeds. Random forest models were fitted for each of the sites, producing high values for the % of variance explained - all above 90%. The predictive performance of the random forest models was assessed using the RMSE and MAE, with the results given in Table 4.2. Table 4.2 shows lower values for the RMSE and MAE at

Site	**RMSE**	**MAE**	**Bias**	**Coverage**
West Strome	0.059	0.039	0.00012	0.99
Muck	0.143	0.102	0.01726	0.77
Djuba Wick	0.055	0.028	0.00789	0.98

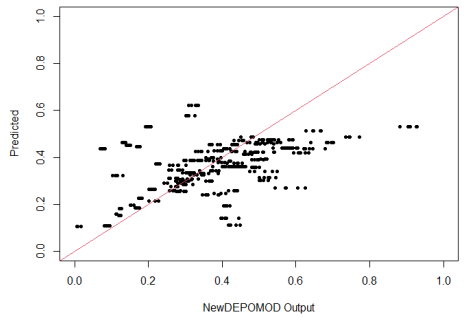
Table 4.2: Table of the predictive performance of the random forest models for Total Area Impacted - additional sites.

West Strome in comparison to the values for Ardentinny in Table 4.1, indicating better predictive performance at this site. Djuba Wick has a similar RMSE value to West Strome, but with a lower MAE value, indicating it performs slightly better. Muck has much larger values for RMSE and MAE than the other sites, indicating that the variance in the Total Area Impacted at this site was not explained well by the changes in the inputs. Considering the coverage

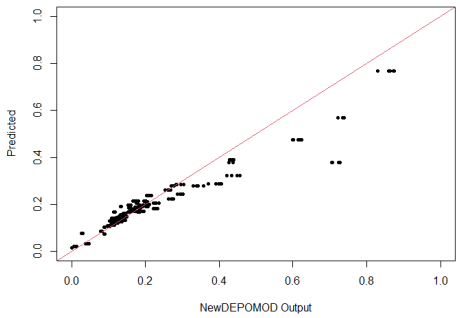


(a) Ardentinny

(b) West Strome



(c) Muck



(d) Djuba Wick

Figure 4.7: Plots of the predicted Total Area Impacted from the random forest model against the output from NewDEPOMOD for different sites.

probabilities, Muck has a much lower value than the other sites, which all have similar values. As with the random forest model for Ardentinny, the models for the additional sites all have a similar ranking structure to the models used in Chapter 2, with the Settling Velocity of Faces being the dominant input at each site when considering the Total Area Impacted as the output, and no changes to the top three ranked inputs at each site.

Considering Figures 4.7, the plot for Muck supports the results from Table 4.2 with a large number of points not close to the line of equality. In contrast, the majority of the points at West Strome and Djuba Wick lie on or close to the line of equality. At Ardentinny, there appears to be a bit more variation around the line of equality, which is expected based on the RMSE and MAE values. At all sites, the random forest emulator appears to under-predict the Total Area Impacted where the standardized NewDEPOMOD output is greater than 0.4, with the majority of the points lying below the line of equality, which is supported by the small, positive values for bias in Tables 4.2 and 4.1.

It was previously mentioned that the possibility of using one emulator to predict the scalar outputs at all sites would be investigated. It has been men-

tioned previously that the operational outputs were considered as categorical inputs, but that due to the nature of the sampling design, only a subset of the possible combinations were considered. As a result, if the test data at a new site featured combinations that were not included in the training data used to fit a model at another site, the predictions could not be calculated. One possible solution to this would be to consider the operational inputs as continuous variables. The best performing model from Table 4.7 was Djuba Wick, and so this model will be fitted using continuous variables for the operational inputs, and used to predict at the remaining sites. Using continuous variables for the operational inputs still explained over 90% of the variation in the training data, and the RMSE and MAE for the predictions at each site are given in Table 4.3. When comparing the predictive performance of the Djuba Wick random

Site	**RMSE**	**MAE**
Ardentinny	0.134	0.088
West Strome	0.133	0.090
Muck	0.259	0.220
Djuba Wick	0.054	0.028

Table 4.3: Table of the predictive performance of the Djuba Wick random forest model of the Total Area Impacted for all sites.

forest emulator for Total Area Impacted in Table 4.2, to the individual models for each site in Table 4.7, it can be seen that there is a large decrease in the predictive performance when using the Djuba Wick random forest emulator all sites. This suggests that this approach for a single emulator is not appropriate and that each site should be considered individually. An alternative approach combined the data from West Strome and Djuba Wick, as they had the best predictive performance in Table 4.7, before fitting a random forest model. Predictions for the test data at each site were made, but the RMSE and MAE values were also greater than the values for all sites using the individual model. This again suggested that each site should be considered individually when modelling the Total Area Impacted.

4.4.2 99th Percentile of Solids Flux

In addition to the Total Area Impacted, measures of the 99th Percentile of Solids Flux were calculated for each of the NewDEPOMOD runs. This allowed random forest models to be fitted to the data and used for predictive purposes in a similar way. The random forest models for each of the sites were able to explain over 90% of the variation in the training data and their predictive

performance was assessed using the RMSE and MAE, with the results given in Table 4.4. Table 4.4 shows that the random forest models for 99th Percentile

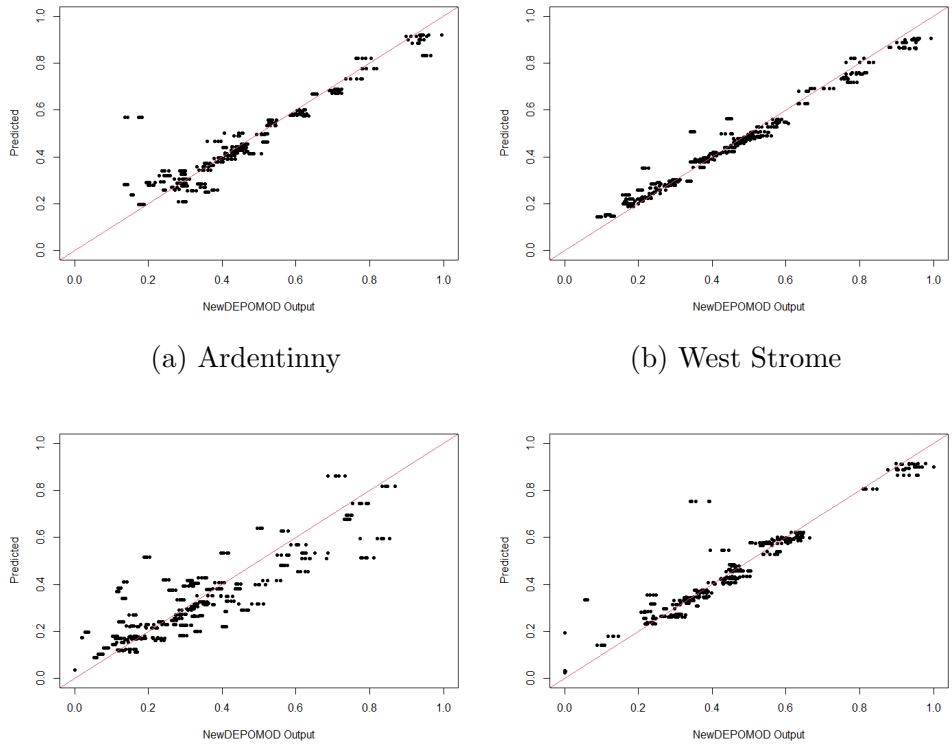
Site	**RMSE**	**MAE**	**Bias**	**Coverage**
Ardentinny	0.067	0.041	0.00078	0.96
West Strome	0.039	0.027	-0.00049	0.99
Muck	0.107	0.079	-0.00152	0.90
Djuba Wick	0.069	0.038	-0.00782	0.92

Table 4.4: Table of the predictive performance of the random forest models for the 99th Percentile of Solids Flux - all sites.

at Ardentinny, West Strome and Djuba Wick perform well when looking at the RMSE and MAE, with Muck having slightly higher values. Muck is again the worst performing of the sites, with the largest RMSE and MAE values along with the lowest coverage probability. The coverage probabilities are higher for the low energy sites, Ardentinny and West Strome. One possible reason for this is the high energy sites are subject to more variation due to the higher current speeds, meaning the variation seen in the NewDEPOMOD runs may not be explained as well by the changes in the inputs as they are for the low energy sites. Again, the ranking of the inputs based on their importance values for these models were compared to the rankings from Table 2.14 in Chapter 2, with similar ranking seen and no changes to the top three ranked inputs. In order to confirm, plots of the predicted values for 99th Percentile against the output from NewDEPOMOD are given in Figure 4.8. The plots for Ardentinny, West Strome and Djuba Wick in Figure 4.8 show most of the points are close to the line of equality, with a small number of points that appear to over-predict for each site. Again, Muck has a lot more variation around the line of equality, with the random forest model under and over predicting.

As with the Total Area Impacted, a single emulator approach was considered - using the best performing site from Table 4.4, which was West Strome, and also considering fitting a model using the training data for both West Strome and Djuba Wick. For both approaches, the RMSE and MAE values were larger than the individual approaches, with three of the sites having values almost double the size of the values from Table 4.4.

Comparing the results for modelling the 99th Percentile of Solids Flux to the results for modelling the Total Area Impacted, the models perform better for the 99th Percentile of Solids Flux at all sites except Djuba Wick, when considering RMSE and MAE. When considering the coverage probability, the emulator for the 99th Percentile has a much larger value at Muck, and for Djuba Wick, the coverage probability is larger for the Total Area Impacted. In



(a) Ardentinny (b) West Strome

(c) Muck (d) Djuba Wick

Figure 4.8: Plots of the predicted 99th Percentile of Solids Flux from the random forest model against the output from NewDEPOMOD for different sites.

addition, for the two low energy sites, the coverage probabilities are similar for both outputs. The model for West Strome was identified as performing well for both inputs, indicating that for both inputs the variation in the outputs can be explained well by the variations in the inputs. In contrast, the models for Muck did not predict well for the test data. When considering the bias values for both outputs, they are all positive for the Total Area Impacted, indicating some small levels of bias, where the underpredictions are likely to have occurred due to the lack of data for the larger values of Total Area Impacted, as seen in the histogram in Figure 4.1. In contrast, for the 99th Percentile, Ardentinny is the only site with a positive value for bias, with the others having small negative values. In comparison, the data is spread more evenly across the full range for the 99th Percentile of Solids Flux in Figure 4.2. A further investigation considered whether a single emulator could be used to predict for all sites, but for both outputs, investigations determined that each site should be considered individually as the predictive performance decreased when considering a single emulator.

4.4.3 Gaussian process emulation of NewDEPOMOD scalar outputs

The scalar output Gaussian process emulation will consider the Total Area Impacted and the 99th Percentile. The same data that was used to fit the random forest models will be used for fitting the Gaussian process models, with different sparse approaches being considered. Due to the large number of runs being considered for each site, sparse Gaussian processes are an effective tool for reducing the computational cost.

Let $\mathbf{X}$ be the input sets for the training data at a given site. Each of the sparse approximation methods above require the selection of an active set, $\mathcal{A}$, which is a subset of the training data containing $m \ll n$ of the total n observations, chosen using sparse greedy matrix approximation (Smola & Schölkopf 2000, Smola & Bartlett 2001). Using the active set, $\mathcal{A}$, the input sets, $\mathbf{X}$, can be sub-setted and defined as $\mathbf{X}_{\mathcal{A}}$. The subset of data approach fits an exact Gaussian process using only the data from the active set, whereas the subset of regressors and FIC approach reduce the computational cost by approximating the kernel matrix, $K(\mathbf{X}, \mathbf{X})$, using the active set. The approximations of the kernel matrices are given in Equations 4.27 and 4.32. These sparse approaches will be considered in detail for the Total Area Impacted at Ardentinny, before applying the specified framework to the remaining sites, and also the 99th Percentile of Solids Flux.

One main choice when fitting a Gaussian process is the kernel function. Throughout this work, the squared exponential kernel function is considered as it is considered the most commonly used kernel function due to its flexibility (Rasmussen & Williams 2006). In this work, the ARD squared exponential is used as it allows for separate lengthscales for each of the 8 inputs in order to provide added flexibility. The lengthscales, along with the signal and noise variance parameters will be optimized using the L-BFGS method described previously.

This application of sparse Gaussian processes to the NewDEPOMOD data will aim to produce efficient models that are able to approximate NewDEPOMOD for the test data without the computational cost. The predictive performance of the models will be assessed and compared to the random forest emulators.

4.4.3.1 Total Area Impacted

The first scalar output that will be considered is Total Area Impacted. The data from the combined physical properties and operational inputs analysis at Ardentinny will be considered in more detail, before summarising for the other sites.

As previously mentioned, there are multiple methods to fitting an approximate Gaussian process model. To assess the efficiency and quality of the approaches, each of the methods were considered for fitting an approximate Gaussian process model using 50 inducing points, and the L-BFGS method for optimizing the hyperparameters. The number of inducing points will be considered later after reviewing the approximation methods. After fitting the approximate Gaussian processes for each of the methods, predictions were made for the test data. RMSE values were then calculated for each of the methods, as well as the time to fit each of the models. The time required to find the predictions was negligible, and so it is not considered. Table 4.5 shows the

Fitting Method	**RMSE**	**MAE**	**Time to fit & optimize model**
SD	0.168	0.119	15s

Table 4.5: Table of the predictive performance of each Gaussian process model for Total Area Impacted using different approximation methods (*SD - Subset of Data*) and 50 inducing points - Ardentinny.

computational time for fitting and optimizing the model for the SD approach, as well as the predictive performance. However, the L-BFGS optimization was unable to converge for the Subset of Regressors and Fully Independent Conditional approaches, therefore they are not included in the results table. It was mentioned previously that the lengthscales can be used as a measure of influence for an input, with small values meaning that the output changes quickly for changes in that input. The smallest lengthscales that were identified for the SD method were Settling Velocity of Faeces, which is consistent with the random forest approach were the Settling Velocity of Faeces had a much larger importance value than the other inputs. Looking at Figure 4.9, the majority of the predictions appear to be below the line of equality, with a small number above the line for the lower values. This would suggest that the model appears to be under-predicting in most cases. Further models will be considered with more inducing points to assess their performance.

The SD approximation was the only approach that was able to converge when fitting, so this approach will be considered, with more inducing points, to assess if this will improve the predictive performance. To test this, 100,

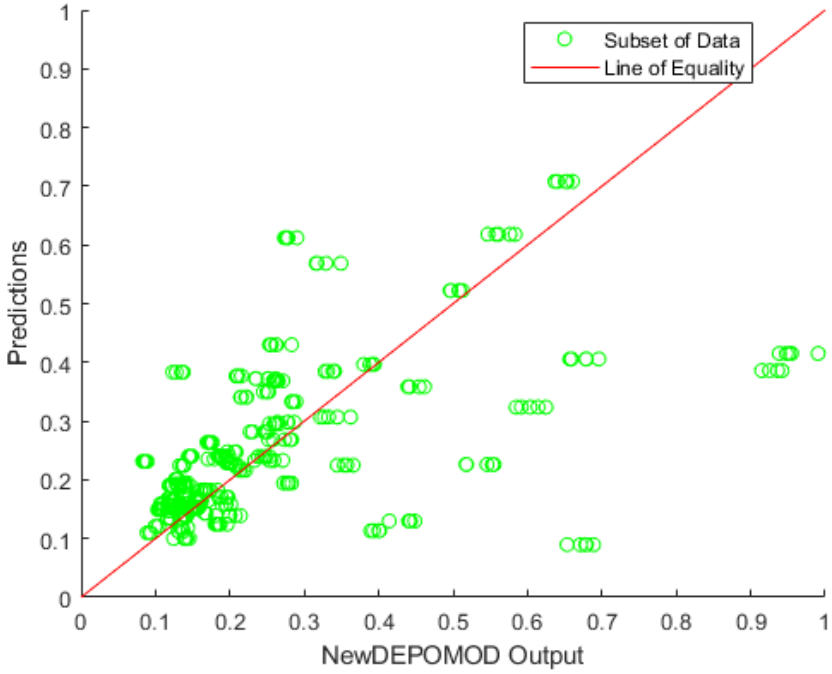


Figure 4.9: Plot of the predicted Total Area Impacted against the observed values for each of the approximation methods.

200 and 400 inducing points will be considered along with the full Gaussian process model, and again the computational time to fit the models as well as the regression loss will be reviewed. Comparing the time to fit the models and the

No. of Inducing Points	RMSE	MAE	Time to fit & optimize model
100	0.139	0.100	30s
200	0.096	0.077	63s
400	0.105	0.079	144s
Full Dataset	0.151	0.108	6224s

Table 4.6: Table of the predictive performance of each Gaussian process model for Total Area Impacted with different numbers of inducing variables using SD approximation, as well as the full Gaussian process model - Ardentinny.

RMSE and MAE from Table 4.5 and Table 4.6, there are improvements in the accuracy compared to 50 inducing points. The increased number of inducing points does result in a larger computational cost, but these are still less than the times seen for the SR and FIC methods in Table 4.5. It should also be noted that increasing the number of inducing points to 400 actually reduced the accuracy of the predictions by a small amount, likely a result of noise within the data. Figure 4.10 shows that there is still an issue of some over and under prediction in some instances. Table 4.6 indicates that increasing the number of inducing points from 200 to 400, and then fitting a full Gaussian process, did

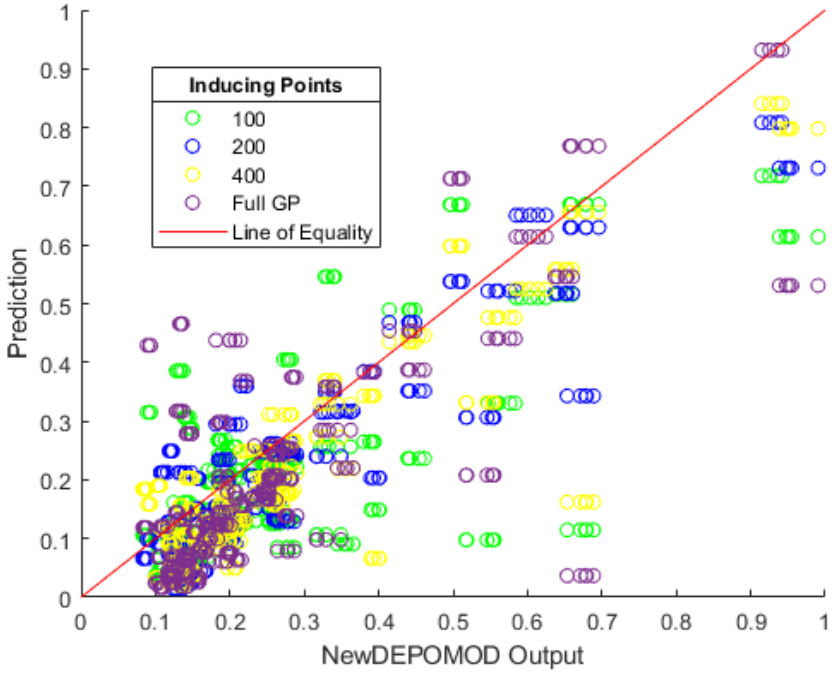


Figure 4.10: Plot of the predicted Total Area Impacted against the observed values for SD approximation models with varying numbers of inducing points, as well as the full Gaussian process model.

not produce better predictions of the Total Area Impacted. Possible reasons for this lack of improvement when using more training data, could be the variation from the random walk within in the model. Looking at the lengthscales for each of the models, the Settling Velocity of Faeces had the smallest values for the models with 200 and 400 inducing points, which is what we would expect as it was identified as the highest ranking input in the sensitivity analysis and random forest model. As a result, there is no evidence to suggest increasing the number of inducing points as there are no performance gains, and increased computational time.

As the optimization process did not converge for the SR and FIC methods, it suggests that the SD approach is the most appropriate for the remaining sites. In addition, Table 4.6, as well as additional investigations, suggested that there were no improvements in the predictive performance when the number of inducing points was increased above 200.

Future Gaussian process modelling will be done using the SD approximation with 200 inducing points for the sparse approach, as well as using the full training data set to fit full Gaussian process models. This will include the modelling of the Total Area Impacted for the additional sites, as well as the modelling of the 99th Percentile of Solids Flux. In order to compare between

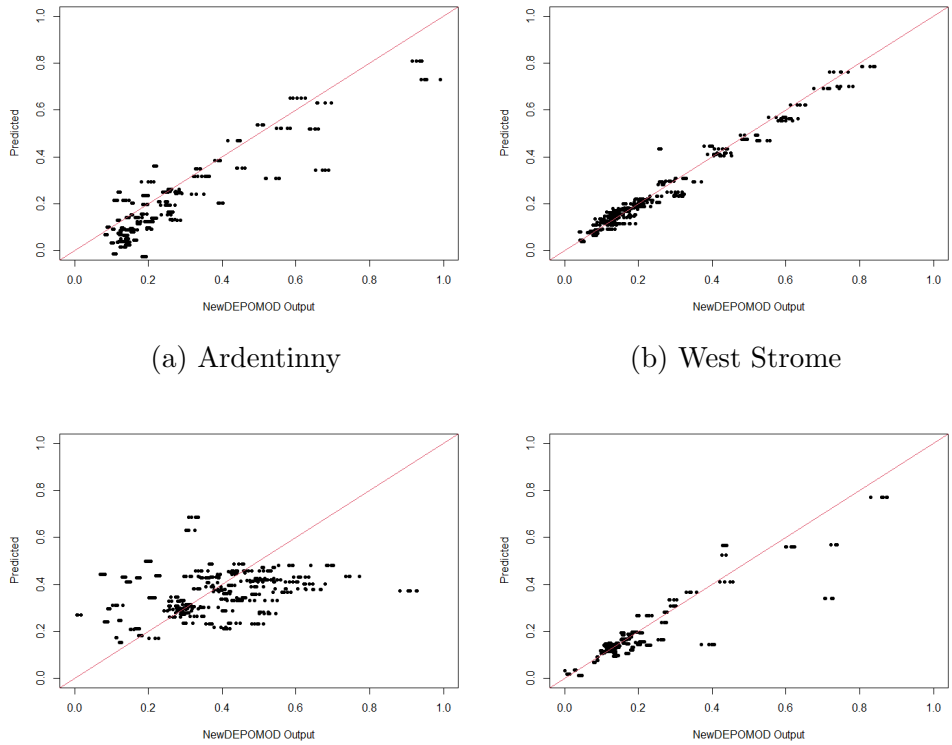
the two emulation approaches, the models have to be fitted using the same training data, so the full Gaussian process models are required for comparisons. However, as the computational time for the sparse approach is much better than the full Gaussian process, this will be considered independently for the remaining sites, as well as for the 99th Percentile of Solids Flux as the output, to assess their predictive abilities. First, the sparse approach will be considered for the additional sites to assess how well this performs, as well as considering the 99th Percentile of Solids Flux as the output, using the sparse approach again.

Next, the Total Area Impacted for the additional sites will be considered. The sparse Gaussian processes were fitted using the SD approach and 200 inducing points, and the performance of their predictions was assessed using RMSE, MAE, Bias and Coverage Probabilty, with the results displayed in Table 4.7 and Figure 4.11. Comparing the results from Table **??** to the results

Site	**RMSE**	**MAE**	**Bias**	**Coverage**
Ardentinny	0.096	0.077	0.0569	0.95
West Strome	0.034	0.024	0.0033	0.95
Muck	0.158	0.118	0.0139	0.85
Djuba Wick	0.065	0.034	0.0126	0.95

Table 4.7: Table of the predictive performance for each Gaussian process model for Total Area Impacted at the additional sites using 200 inducing points and SD approximation.

for the additional sites in Table 4.7, there are big improvements in the predictive performance at West Strome compared to Ardentinny, with the emulation at Djuba Wick also performing better than Ardentinny. The predictive performance of the Gaussian process for Muck was much poorer in comparison to the other sites, suggesting that the variation in the Total Area Impacted is not explained well by the changes in the inputs. There appears to be no pattern between the predictive performance and the site characteristics, with West Strome and Djuba Wick having the lowest RMSE and MAE values. West Strome is a low energy site and Djuba Wick is a high energy site, so this does not appear to play a role in how well the models perform. All of the sites had small, positive values for bias, indicating that the models may underpredict. The bias value for Ardentinny is larger than the other sites, which is confirmed in Figure 4.11a, where the majority of the points lie below the line of equality.



(a) Ardentinny (b) West Strome

(c) Muck (d) Djuba Wick

Figure 4.11: Plots of the predicted Total Area Impacted from the Gaussian process model against the output from NewDEPOMOD for different sites.

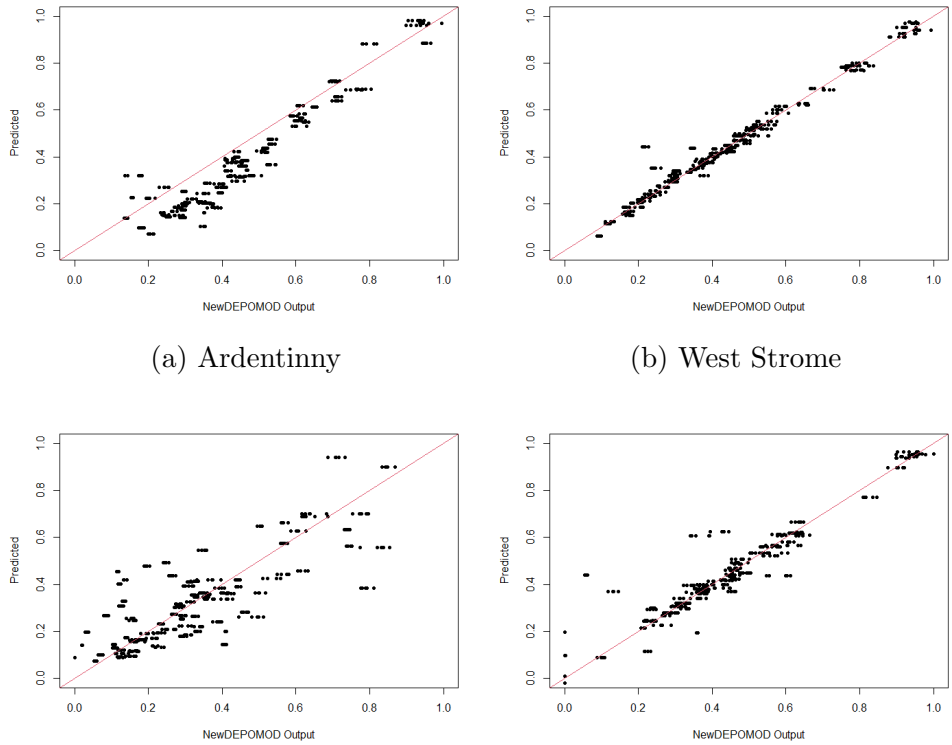
4.4.3.2 99th Percentile

Having considered the Total Area Impacted, now Gaussian process models will be fitted for the 99th Percentile at each of the sites. In line with the previous models, their predictive performance was measured using RMSE, MAE, Bias and Coverage Probabililty, with the results presented in Table 4.8 and Figure 4.12. Table 4.8 shows fairly large differences in the performance between West

Site	**RMSE**	**MAE**	**Bias**	**Coverage**
Ardentinny	0.104	0.091	0.0768	0.74
West Strome	0.035	0.021	-0.0037	0.95
Muck	0.129	0.094	0.0003	0.89
Djuba Wick	0.077	0.042	-0.0087	0.92

Table 4.8: Table of the predictive performance for each Gaussian process model for the 99th Percentile of Solids Flux at all sites using 200 inducing points and SD approximation.

Strome and the other sites, with West Strome performing better. The values for RMSE and MAE are larger at Ardentinny and Muck, and together with the lower values for the coverage probabilities indicate a poor fit. Ardentinny



(a) Ardentinny (b) West Strome

(c) Muck (d) Djuba Wick

Figure 4.12: Plots of the predicted 99th Percentile of Solids Flux from the Gaussian process model against the output from NewDEPOMOD for different sites.

has a positive value for bias, indicating the model underpredicts, which is supported by Figure 4.12a. The bias in this case could therefore be affecting the calculation of the coverage probability and causing it to be the lowest at all of the sites. The bias values for the other sites are lower, with West Strome and Djuba Wick having small, negative values, indicating that the model is slightly over-predicting.

4.5 Comparison of the random forest and Gaussian process emulation approaches

As mentioned previously, after considering the two emulation approaches, these will be compared to assess how they perform. In order to make comparisons between the two approaches, the same training data has to be used for fitting the models. As the full Gaussian process models are more computationally expensive than the sparse approaches and without any improvement on the predictive performance, the sparse Gaussian processes will be considered. To

make comparisons to the random forest approach, the data used to fit the sparse Gaussian process models will be used to fit the random forest models before calculating their predictive performance using the test data.

The first comparison considers the Total Area Impacted as the output, with the predictive performance statistics summarised in Table 4.9. The first thing

Site	Method	**RMSE**	**MAE**	**Bias**	**Coverage**
Ardentinny	RF	0.085	0.059	0.0036	0.98
Ardentinny	GP	0.096	0.077	0.0569	0.95
West Strome	RF	0.074	0.051	0.0007	0.98
West Strome	GP	0.034	0.024	0.0033	0.95
Muck	RF	0.144	0.108	0.0152	0.78
Muck	GP	0.158	0.118	0.0139	0.85
Djuba Wick	RF	0.066	0.037	0.0048	0.98
Djuba Wick	GP	0.065	0.034	0.0126	0.95

Table 4.9: Table of the predictive performance for each method at all sites for the Total Area Impacted.

to note from Table 4.9 is that the RMSE and MAE values for Muck for both the random forest and Gaussian process approaches are higher than for the other sites. This indicates that the variance in the Total Area Impacted at this site is not explained as well by the changes in the inputs. At Ardentinny, Djuba Wick and Muck, the RMSE and MAE values are fairly close, indicating that neither of the approaches appears to perform consistently better. In contrast, at West Strome, the predictive performance statistics indicate that the Gaussian process emulation approach performs best, with RMSE and MAE less than half of the values for the random forest approach. Considering bias, all of the values for both random forests and Gaussian processes are positive, indicating that they slightly underpredict, which is likely due to the lack of data available for the larger values of Total Area Impacted. When comparing between random forests and Gaussian processes, the bias values are lower at all of the sites except for Muck. Finally, looking at the coverage values, these are equal to, or about 0.95 for all of the sites except Muck, where the variation did not appear to be explained well by the changes in the inputs.

Next, the 99th Percentile of Solids Flux was considered as the output, and the predictive performance of the two methods is compared again. Comparing the results for the two approaches in Table 4.10, West Strome is the only site where the Gaussian process approach performs better than the random forest when considering RMSE and MAE. Ardentinny and Muck have lower values for RMSE and MAE, and Djuba Wick has similar values for both statistics for the two approaches. When considering the coverage probability, the random

Site	Method	**RMSE**	**MAE**	**Bias**	**Coverage**
Ardentinny	RF	0.072	0.052	0.0082	0.95
Ardentinny	GP	0.104	0.091	0.0768	0.74
West Strome	RF	0.051	0.040	0.0046	0.99
West Strome	GP	0.035	0.021	-0.0037	0.95
Muck	RF	0.106	0.083	-0.0032	0.89
Muck	GP	0.129	0.094	0.0003	0.89
Djuba Wick	RF	0.071	0.048	0.0014	0.93
Djuba Wick	GP	0.077	0.042	-0.0087	0.92

Table 4.10: Table of the predictive performance for each method at all sites for the 99th Percentile of Solids Flux.

forest approach has higher values at all of the sites except Muck, where the coverage values are the same.

One additional comment relates to the random forest models for both outputs. In order to compare between the two methods, the same datasets had to be used to train the emulator models. This meant that the random forest emulators in Tables 4.9 and 4.10 used less data than the models from earlier in the Chapter. When comparing the predictive performance though, the values for each of the statistics appear to be similar, indicating the models with less data fit just as well.

4.6 Discussion

The main aim of this Chapter was to build statistical emulators to approximate the scalar outputs produced from running NewDEPOMOD without the computational time. For this, two approaches were considered - random forest modelling and Gaussian process modelling. These two approaches have the benefit of being flexible and efficient to run when predicting at new input sets.

One benefit to using random forests for predicting the scalar outputs at the test data, is that the computational time required to fit the model using all of the training data is small - approximately one minute. In contrast, for the Gaussian process approach, it was determined that a sparse approach was required, where a subset of the training data is used to fit the model. Different sparse approaches were considered, before using the SD approach, which had a similar computational time to the random forest approach in comparison to the SR and FIC approaches.

Looking at the predictive performance of the random forest and Gaussian process models at multiple sites, one point to highlight was that there were differences seen between the sites with similar characteristics, with West Strome (a

low energy site) and Djuba Wick (a high energy site) having the lowest RMSE and MAE values for each emulation approach for both outputs. This suggested that each site should be considered independently, and was confirmed when investigating whether a single random forest emulator could be used for all of the sites. Comparing the performance of the random forest approach to the Gaussian process approach, some of the metrics indicated that one approach performed better, while other metrics suggested the other approach performed better. Overall, there did not appear to be clear evidence to suggest that one method would be preferred to the other. One thing that could be considered to suggest that random forests are a more appropriate choice would be the fact that they can be trained efficiently while using all of the data, whereas the Gaussian process models are more computationally demanding when using all of the data, which is why a sparse approach was considered. These analyses has provided a foundation for extending the random forest and Gaussian process emulation framework in the next Chapter to consider the outputs as a multivariate output with correlations, where each site will be considered independently.

Chapter 5

Emulation of Multivariate Outputs

5.1 Introduction

Having considered the emulation of univariate outputs from a process-based model, the next logical step is to consider the extension where a process-based model produces a multivariate output. Rougier (2008) highlighted that increasing emulation complexity from univariate output to multivariate output provides extra challenges such as the extra data being considered for additional inputs, and the introduction of relationships between outputs increases the computational complexity. The modelling of multivariate outputs can be considered in two different ways: 1) using multiple independent single output emulators, in which case the outputs are not related or 2) considering the outputs directly as correlated outputs. In the case of NewDEPOMOD data, the aim of this Chapter is to consider whether accounting for a relationship between the Total Area Impacted and the 99th Percentile of Solids Flux results in improved predictive performance in comparison to the single output emulators in Chapter 4.

For the case where multiple independent single output emulators are considered, dimension reduction techniques could be used to reduce/transform multiple, correlated outputs. After applying dimension reduction, the reduced/transformed data can be considered as independent outputs, in which case multiple single output emulators can be applied, an approach considered by (Higdon et al. 2008, Bayarri et al. 2007). Alternatively, the correlated outputs can be considered directly in emulation approaches, where the correlation between the outputs are incorporated into the emulator structure (Conti & O'Hagan 2010), or alternatively through convolution processes as a way to in-

corporate non-trivial correlations between outputs within Gaussian processes (Alvarez & Lawrence 2009, 2011). Both approaches have drawbacks, with dimension reduction techniques resulting in some loss of information through the new representation of the data and the requirement to fit multiple models can be computationally expensive, while considering the outputs directly as correlated outputs requires the correlation structure to be included within the model, which can be computationally expensive for many outputs.

Chapter 4 considered random forests and Gaussian processes as emulation methods, both of which can be extended to account for multivariate outputs. The multivariate output from NewDEPOMOD being considered only contains two outputs, and so these will be considered directly as correlated outputs to avoid any loss of information. Segal & Xiao (2011) extended the single-output random forest model to account for multiple outputs which are linearly related. The extension proposed by Segal & Xiao (2011) considered the combination of multivariate regression trees (De'ath 2002) and the traditional random forest approach (Breiman 2001). More detail of this approach will be considered within this Chapter.

The alternative approach to be considered within this Chapter are Gaussian processes for multivariate outputs, an extension of the univariate output Gaussian processes in Chapter 4. Different approaches are considered for implementing the multivariate output Gaussian processes, such as the Linear Model of Coregionalization (Journel & Huijbregts 1978) and convolution processes (Alvarez & Lawrence 2009, 2011). Multivariate Gaussian processes can present computational challenges related to the optimization of the hyperparameters, which will be discussed in more detail throughout the Chapter.

Both of these approaches will be applied to the output from NewDEPOMOD to consider whether modelling of the two scalar outputs, Total Area Impacted and 99th Percentile of Solids Flux, together to account for any relationship between them will improve the predictive accuracy of the models for new data. In a similar manner to Chapter 4, the predictive performance of these emulators will be assessed using the RMSE and MAE. The aim of this Chapter is to make predictions of the Total Area Impacted and 99th Percentile of Solids Flux without the computational cost of running NewDEPOMOD, while accounting for any potential relationships between the two outputs that could improve the predictive performance.

5.1.1 Data being used

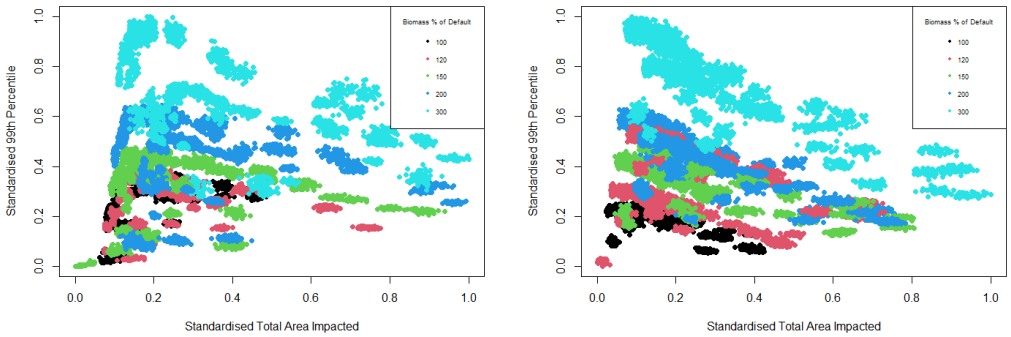
It was mentioned in Chapter 4 that in order to build a statistical emulator, the complex process-based model has to be run at a number of different input sets. These input sets and the corresponding output data are often referred to the training data and are used to build a statistical emulator model. In order to test the performance of the emulator, the process-based model will be run at some new input sets, at which predictions will be made using the statistical emulator, before comparing the output from the process-based model to the output from the statistical emulator. The new input sets and their simulator output are referred to as the test data.

Within this Chapter, the data that will be considered is the same data as was used in Chapter 4. The four sites will be considered, with training and test data which include calculations of both the Total Area Impacted and the 99th Percentile. These two scalar outputs are what will be considered within this Chapter as the multivariate output. The two outputs plotted against each other for each of the sites can be seen in Figure 5.1, where the observations are coloured based on the corresponding Biomass values. There are some patterns present in Figure 5.1, when accounting for the different Biomass values. There appears to be a lot of variation for all of the Biomass values where the Standardised Total Area Impacted is less than 0.2. As the Standardised Total Area Impacted increases above 0.2, there are potentially some weak, decreasing, linear trends which have some outliers. Considering the full data, the relationship for each Biomass value does appear to be non-linear, with a sharp increase in the standardised 99th Percentile as the Standardised Total Area Impacted increases above 0.1, but with quite a bit of variance.

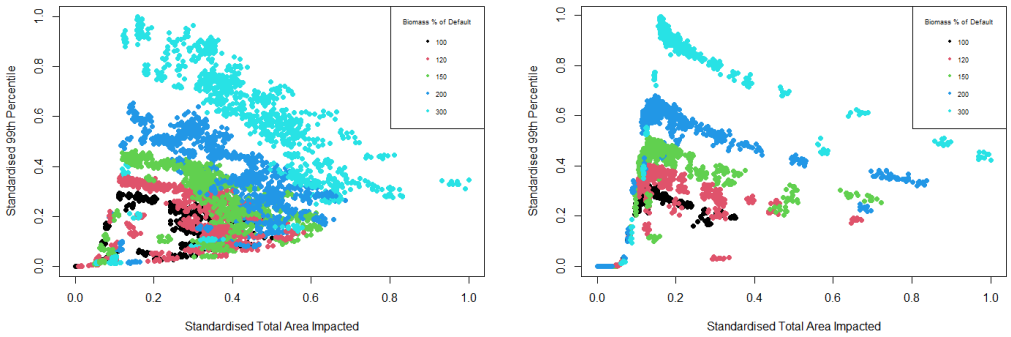
5.2 Multivariate output random forests

Random forests were identified as an effective modelling tool for both regression and classification problems due to their ability to deal with non-linear relationships, incorporate interactions, and provide easy to interpret importance values that can be used as a measure of the influence of the inputs (Pianosi et al. 2016). Segal & Xiao (2011) looked to extend the standard single-output random forest model (Breiman 2001) to the scenario where there are multiple outputs to be considered. Segal & Xiao (2011) provided an example of the multivariate random forest applied to an ecology problem, and the approach has also been considered in other settings (Miller et al. 2014, Browne et al. 2021).

Multivariate random forests were proposed as an extension of traditional



(a) Ardentinny (b) West Strome



(c) Muck (d) Djuba Wick

Figure 5.1: Plots of the predicted outputs for the test data against the NewDEPOMOD output, coloured by the Biomass values.

random forests by combining the traditional method (Breiman 2001), with multivariate regression trees (De'ath 2002). De'ath (2002) introduced multivariate regression trees as a new technique for modelling species-environment relationships. Breiman et al. (1984) introduced a regression tree framework that consisted of four components:

1. A set of binary questions or splits relating to the inputs, that aim to partition the input space. The subsamples of the data created by the splits are defined as nodes.

2. A measure of node impurity that relates to variation in the output.

3. For each split, s, of each node, t, a split function, $\phi(s,t)$, is evaluated, with the best split that optimizes, ϕ, in order that the response distribution in the resultant children nodes are most similar among the competing splits, which is assessed via the impurity measure.

4. A way of determining the appropriate tree size.

First, the univariate response for regression trees will be considered, where Y_i is the output, and x_{ij} are the inputs for $i = 1, \ldots, n; j = 1, \ldots, p$. For a node, t, containing a sub-sample of cases, the aim is to partition t into two child nodes, that can be considered as a left node, t_L and a right node, t_R. Consider one of the inputs with index, j, then binary splits that are order preserving can be considered as, $\{t_L = i \in t : x_{ij} \leq c\}$ and $\{t_R = i \in t : x_{ij} > c\}$, where the cut-point, c ranges over all possible values (Segal & Xiao 2011). The mean-squared error split statistic is then given as,

$$D(s,t) = \frac{1}{n}\sum_{i \in t_L}(Y_i - \mu(t_L))^2 + \frac{1}{n}\sum_{i \in t_R}(Y_i - \mu(t_R))^2, \tag{5.1}$$

where $\mu(t_L)$ and $\mu(t_R)$ refers to the sample means for t_L and t_R. Then, the best split for x_j is the split, S which minimizes $D(s,t)$.

The extension for the multivariate case for multivariate regression trees requires the modification of the split statistic in Equation 5.1 (Segal & Xiao 2011). To illustrate the extension, consider the q-dimensional multivariate output data, $Y_{i,j}$, for $i = 1, \ldots, n; j = 1, \ldots q$. Then the modified split statistic is given as (Ishwaran et al. 2021),

$$D_q(s,t) = \sum_{j=1}^{q}\left\{\sum_{i \in t_L}(Y_{i,j} - \mu(t_{L_j}))^2 + \sum_{i \in t_R}(Y_{i,j} - \mu(t_{R_j}))^2\right\} \tag{5.2}$$

where $\mu(t_{L_j})$ and $\mu(t_{R_j})$ represents the sample means of the j-th response for the left and right children nodes. For the multivariate output, the goal is then to minimize $D_q(s,t)$. It should be noted that all of the outputs being considered should be on the same scale, otherwise the contribution of an output with large values would dominate $D_q(s,t)$ (Ishwaran et al. 2021).

The standard multivariate regression splitting rule in equation 5.2 did not take into account any correlation between the outputs. In order to introduce correlations between the outputs, the Mahalanobis distance (Mahalanobis 1936) was incorporated. For a given element, $\mathbf{Z}$, with mean, $\mu_{\mathbf{Z}}$, and variance-covariance, $\Sigma_{\mathbf{Z}}$, the Mahalanobis distance from $\mathbf{Z}$ to the mean, $\mu_{\mathbf{Z}}$ is given as,

$$\mathcal{D}_M(\mathbf{Z}) = (\mathbf{Z} - \mu_{\mathbf{Z}})^\top \Sigma_{\mathbf{Z}}^{-1}(\mathbf{Z} - \mu_{\mathbf{Z}}). \tag{5.3}$$

One problem with the Mahalanobis distance in practice is that $\Sigma_{\mathbf{Z}}$ may be singular. To overcome this problem, the Moore-Penrose generalized inverse (Penrose 1955) is introduced. For an efficient multivariate splitting rule based on the Mahalanobis distance, consider continuous outputs, $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_q)^\top \in \mathbb{R}^q$.

For a splitting tree node, t, let the centered output matrix for t be given as,

$$\mathbf{L}_t^* = \begin{pmatrix} (\mathbf{Y}_1 - \boldsymbol{\mu}(t))^\top \\ \vdots \\ (\mathbf{Y}_n - \boldsymbol{\mu}(t))^\top \end{pmatrix}_{n \times q},$$

where the sample means for $\mathbf{Y}$ in t are given in the q-dimensional vector, $\boldsymbol{\mu}(t)$. The sample covariance matrix for the data is given as $n^{-1}\mathbf{Q}_t^*$, where $\mathbf{Q}_t^* = (\mathbf{L}_t^*)^\top \mathbf{L}_t^*$. Here, $\mathbf{Q}_t^*$ has the generalized inverse, $(\mathbf{Q}_t^*)^+$. For t, suppose that it is split into left and right children nodes, t_L and t_R, based on inputs, $\mathbf{X}$. The q-dimensional sample mean vectors for $\mathbf{Y}$ in t_L and t_R are given by $\boldsymbol{\mu}(t_L)$ and $\boldsymbol{\mu}(t_R)$. The Mahalanobis multivariate split-statistic is (Ishwaran et al. 2021),

$$\mathcal{D}_{M,t}(t_L, t_R) = \frac{n_L}{n} \sum_{i \in t_L} (\mathbf{Y}_i - \boldsymbol{\mu}(t_L))^\top (\mathbf{Q}_t^*)^+ (\mathbf{Y}_i - \boldsymbol{\mu}(t_L)) \tag{5.4}$$

$$+ \frac{n_R}{n} \sum_{i \in t_R} (\mathbf{Y}_i - \boldsymbol{\mu}(t_R))^\top (\mathbf{Q}_t^*)^+ (\mathbf{Y}_i - \boldsymbol{\mu}(t_R)). \tag{5.5}$$

By minimizing $\mathcal{D}_{M,t}(t_L, t_R)$, the best split for t can be obtained. An alternative approach to determine the best split for t, is to maximize the following (Ishwaran et al. 2021),

$$\mathcal{D}_{M,t}^*(t_L, t_R) = 1 - \frac{1}{q} \mathcal{D}_{M,t}(t_L, t_R). \tag{5.6}$$

The implementation of the Mahalanobis split-statistic allows correlations between outputs to be included within the multivariate random forest.

The extension of random forests for multivariate output will be considered within this Chapter as an emulation approach. These results can therefore be compared to the independent random forest models from Chapter 4, as well as comparing to the alternative approach of multi-output Gaussian processes that will be considered within this Chapter. (Browne et al. 2021) considered the predictive performance of independent random forest models compared to a multivariate approach, finding the multivariate approach performed better in some instances, but not all. Browne et al. (2021) noted that the simplicity of the Mahalanobis splitting method could be a promising direction for future work, but it has not yet been considered in the literature. Within this Chapter, this approach will be considered in order to determine if the joint modelling of the two outputs using multivariate random forests will provide better predictive performance.

5.2.1 Application of multivariate random forests for emulation

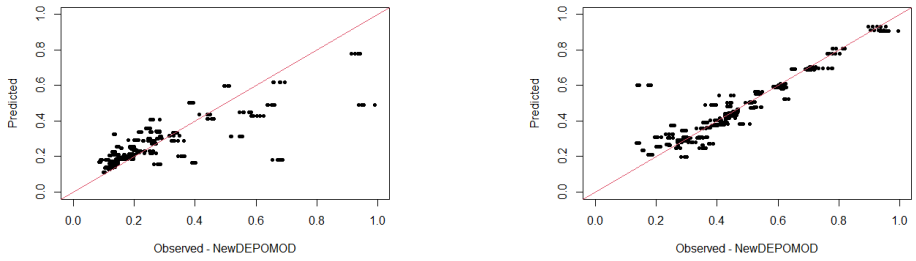
Browne et al. (2021) noted that there are a lack of more sophisticated multivariate random forest models, and that the Mahalanobis splitting method is available which is only able to account for linear relationships between inputs. Previously, Figure 5.1 appeared to show non-linear trends were present when accounting for the different Biomass values, with some large sections of weak, linear trends. As Browne et al. (2021) mentioned, there are no methods available at this time that are able to account for more complex relationships within multivariate random forests, and so, the Mahalanobis splitting method will be considered to test how the assumption of a linear trend would compare to the independent random forest approach. The non-linearity at the lower values for the Standardised Total Area Impacted will be considered when reviewing the results to assess for under or over predictions.

The data used within this modelling will be the same standardized data that was used for the emulation of the outputs independently in Chapter 4. In this case, the standardized versions of the Total Area Impacted and the 99th Percentile of Solids Flux are considered as a multivariate output for the multivariate random forest modelling. For each site, the multivariate random forest model will be fitted using all of the training data that was used in Chapter 4, before using the test data to assess the predictive performance of the models. Each of the random forest models were again able to explain over 90% of the variation in the data and could be fitted in 10 minutes, with predictions at the test data being produced in less than a second. The results from the investigation of their predictive performance is given in Table 5.1. When comparing

		Total Area Impacted		**99th Percentile**	
Site	**RF Type**	**RMSE**	**MAE**	**RMSE**	**MAE**
Ardentinny	Multi	0.108	0.068	0.071	0.043
Ardentinny	Independent	0.078	0.050	0.067	0.041
West Strome	Multi	0.050	0.032	0.041	0.026
West Strome	Independent	0.059	0.039	0.039	0.027
Muck	Multi	0.147	0.109	0.116	0.086
Muck	Independent	0.143	0.102	0.107	0.079
Djuba Wick	Multi	0.064	0.031	0.088	0.046
Djuba Wick	Independent	0.055	0.028	0.069	0.038

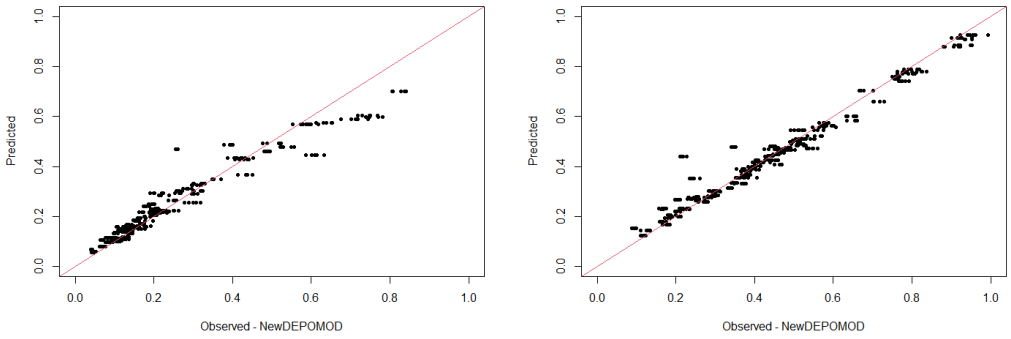
Table 5.1: Table of RMSE and MAE for each output at each site, when predictions are made using the multivariate random forest models.

the results in Table 5.1 of the multivariate random forests to the independent



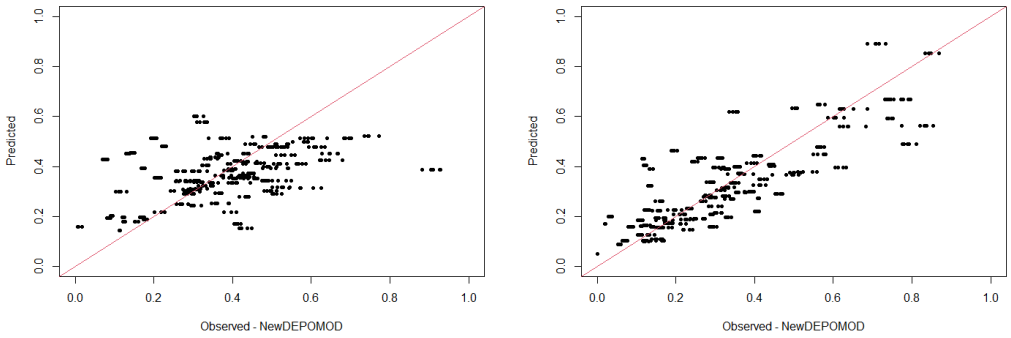
(a) Area (b) 99th Percentile of Solids Flux

Figure 5.2: Plots of the predicted outputs for the test data against the NewDEPOMOD output for multivariate random forests - Ardentinny.



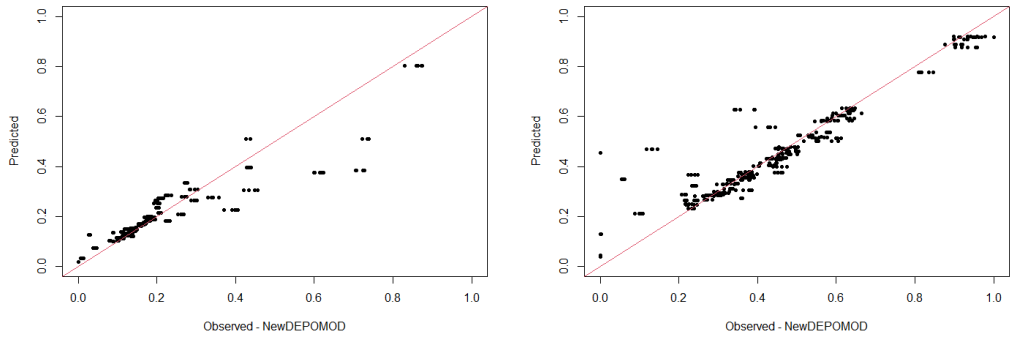
(a) Area (b) 99th Percentile of Solids Flux

Figure 5.3: Plots of the predicted outputs for the test data against the NewDEPOMOD output for multivariate random forests - West Strome.



(a) Area (b) 99th Percentile of Solids Flux

Figure 5.4: Plots of the predicted outputs for the test data against the NewDEPOMOD output for multivariate random forests - Muck.



(a) Area (b) 99th Percentile of Solids Flux

Figure 5.5: Plots of the predicted outputs for the test data against the NewDEPOMOD output for multivariate random forests - Djuba Wick.

random forests for each output, the majority of the results are slightly worse when fitting a multivariate random forest. At West Strome, the multivariate random forest improves the predictive performance slightly when considering the RMSE and MAE for Total Area Impacted, and the MAE for 99th Percentile of Solids Flux, though the differences are small. As was mentioned previously, the assumption of a linear trend between the outputs in order to use the Mahalanobis splitting rule was ambitious, and this is highlighted by the fact that considering the outputs together does not improve the predictive performance. Therefore, it suggests that further work could focus on determining a way to account for non-linear relationships between outputs. Figures 5.2 and 5.5 show reasonable agreement between the predictions for the test data and the NewDEPOMOD output for Ardentinny and Djuba Wick with some variation and areas of under and over prediction. At West Strome, Figure 5.3 highlight the better predictive performance in comparison to the other sites which was noted in Table 5.1, with the opposite seen at Muck in Figure 5.4 where the RMSE and MAE were higher than the other sites. Looking at the plots for the 99th Percentile, there appears to be some over-prediction for the lower observed values, which are potentially a result of the assumption of linearity between the two outputs.

The multivariate random forest models are able to rank the inputs based on their importance values for each input, as described in Chapter 2. The importance values produced relate to each output, with the top three ranked inputs at each site given in Table 5.2 for the Total Area Impacted and Table 5.3 for the 99th Percentile of Solids Flux. The importance values from Tables 5.2 and 5.3 can be compared to the top ranking inputs that were discussed in Chapter 4. The first thing to note is that the operational inputs (Biomass,

Ranking	**Ardentinny**	**West Strome**	**Muck**	**Djuba Wick**
1	Settling Velocity of Faeces	Settling Velocity of Faeces	Settling Velocity of Faeces	Settling Velocity of Faeces
2	Biomass	Biomass	Biomass	Critical Shear Stress for Erosion
3	Critical Shear Stress for Erosion	Number of Cages	Critical Shear Stress for Erosion	Biomass

Table 5.2: Table of top three ranked inputs at each site for the Total Area Impacted from the multivariate random forest.

Ranking	**Ardentinny**	**West Strome**	**Muck**	**Djuba Wick**
1	Biomass	Biomass	Biomass	Biomass
2	Cage Diameter	Cage Diameter	Cage Diameter	Number of Cages
3	Critical Shear Stress for Erosion	Settling Velocity of Faeces	Critical Shear Stress for Erosion	Critical Shear Stress for Erosion

Table 5.3: Table of top three ranked inputs at each site for the 99th Percentile of Solids Flux from the multivariate random forest.

Cage Diameter and Number of Cages) play a much bigger role when modelling the univariate outputs together as a multivariate output. One possible reason for this is that the operational inputs determine how much waste is released from the cages, and so, when the two outputs are considered together, they play a more dominant role. This is confirmed by the Biomass being the top ranked input for all of the sites when considering the 99th Percentile of Solids Flux. This suggests that the amount of production and therefore the amount of waste is heavily influential when considering the two outputs together.

The results from Table 5.1 highlight that, in most cases, including the linear relationship between the Total Area Impacted and the 99th Percentile of Solids Flux does not provide any information gain when using the multivariate random forests to predict the NewDEPOMOD outputs, when comparing to the results from Chapter 4. This would be expected based on the initial analysis of the relationship between the two outputs. This suggests that the linear relationship assumption for the outputs is not suitablefor providing information gain, and an extension to the multivariate random forest could be considered where more complex relationships between the outputs are incorporated.

5.3 Multi-output Gaussian processes

An alternative approach to modelling multivariate outputs that can be considered are Gaussian processes, where the single output case described in Chapter 4 can be extended to the case where multiple outputs are present. Multiple approaches have been proposed in the literature, and some of these will be introduced below. These will be introduced by introducing multi-output Gaussian processes for the case where the outputs are independent and no correlation exists between them, as an extension to the univariate Gaussian processes considered in Chapter 4. Following this, the extensions of the multi-output Gaussian processes for correlated outputs will be considered.

First consider a single output Gaussian process, $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. For given data, $\mathcal{D} = \{(\mathbf{x}_i, f(\mathbf{x}_i)) : i = 1, \ldots, N\}$, the Gaussian process can be expressed as follows, for a zero mean function:

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \right) \tag{5.7}$$

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}). \tag{5.8}$$

Consider another Gaussian Process, $g(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ with the same mean function and kernel as $f(\mathbf{x})$. This can then be expressed in a similar way to Equation 5.8, $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$. The two processes can then be represented together as:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{K} \end{bmatrix} \right) \tag{5.9}$$

$$\mathbf{h} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{f,g}\right). \tag{5.10}$$

Equation 5.10 highlights that the covariance between the two outputs, $\mathbf{f}$ and $\mathbf{g}$ is zero, and the multi-output GP therefore considers the outputs as two independent single-output Gaussian processes. This approach is simple and can be considered for the case where the outputs are independent and no correlation exists between them. Dimension reduction techniques such as principal components (Pearson 1901) are an effective approach for reducing high-dimensional, correlated data into a small number of independent variables approximating the original data. Therefore, dimension reduction techniques can be used to create a number of independent variables to approximate the original output data, before modelling them as an independent multi-output Gaussian process.

As previously mentioned, the covariance structure of a multi-output Gaussian process can be created to account for correlations between the outputs. Unless otherwise known, treating multiple outputs as independent and modelling them separately can result in significant loss of information and can be a restrictive assumption (Noè et al. 2019). Part of the approach to these multi-output Gaussian processes is to choose a prior on the correlation of the outputs, which was discussed by Alvarez et al. (2012), with some of the choices summarised below (van der Wilk et al. 2020).

5.3.1 Linear model of coregionalization

The linear model of coregionalization (Journel & Huijbregts 1978) was a consideration in geostatistics literature for expressing correlation between multiple outputs (Alvarez & Lawrence 2011). For this approach, the sum of Kronecker products between coregionalization matrices and a set of underlying covariance functions are considered, where the coregionalization matrices contain the correlations across the outputs, and the underlying covariance functions describe the correlation between the input points (Alvarez & Lawrence 2011). For a multi-output function, $\mathbf{f}(\cdot) : \mathcal{X} \to \mathbb{R}^P$, mapping data from the input space, $\mathcal{X}$, to the P-dimensional output space. The idea is then to describe the multi-output function $\mathbf{f}(\cdot)$ as follows (Journel & Huijbregts 1978):

$$\mathbf{f}(\mathbf{x}) = A\mathbf{g}(\mathbf{x}).$$

Here, $A \in \mathbb{R}^{P \times L}$, where L is the number of independent functions $g_l(\cdot) \sim \mathcal{GP}(0, k_l(\cdot, \cdot'))$. The set of independent functions $g_l(\cdot)$ are then expressed as $\mathbf{g}(\mathbf{x}) = \{g_l(\mathbf{x})\}_{l=1}^{L}$.

5.3.1.1 Intrinsic Coregionalization Model

The coregionalization approach can be broken down into a simpler format, the Intrinsic Coregionalization Model ('ICM') (Goovaerts 1997) refers to the case where the functions, $g_l(\cdot)$, have the same covariance function, $k(\cdot, \cdot')$. To illustrate the process, consider the case where there are two outputs, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, with the input $\mathbf{x} \in \mathbb{R}^2$. We will also choose two independent functions $g_1(\mathbf{x})$, $g_2(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$, which will be used to describe the two outputs

as follows.

$$f_1(\mathbf{x}) = a_1^1 g_1(\mathbf{x}) + a_1^2 g_2(\mathbf{x})$$
$$f_2(\mathbf{x}) = a_2^1 g_1(\mathbf{x}) + a_2^2 g_2(\mathbf{x}).$$

To proceed, the two outputs, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, will be grouped together as a vector, $\mathbf{f}(\mathbf{x})$, when considering a fixed value of $\mathbf{x}$.

$$\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} a_1^1 g_1(\mathbf{x}) + a_1^2 g_2(\mathbf{x}) \\ a_2^1 g_1(\mathbf{x}) + a_2^2 g_2(\mathbf{x}) \end{bmatrix}$$
$$\mathbf{f}(\mathbf{x}) = \left[\mathbf{a}^1 g_1(\mathbf{x}) + \mathbf{a}^2 g_2(\mathbf{x}) \right].$$

Here, $\mathbf{a}^1 = [a_1^1 \quad a_2^1]^\top$ and $\mathbf{a}^2 = [a_1^2 \quad a_2^2]^\top$. The computation for the covariance of $\mathbf{f}(\mathbf{x})$ is then given as,

$$\begin{aligned}
\text{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) &= \text{cov}(\mathbf{a}^1 g_1(\mathbf{x}) + \mathbf{a}^2 g_2(\mathbf{x}), \mathbf{a}^1 g_1(\mathbf{x}') + \mathbf{a}^2 g_2(\mathbf{x}')) \\
&= \text{cov}(\mathbf{a}^1 g_1(\mathbf{x}), \mathbf{a}^1 g_1(\mathbf{x}')) + \text{cov}(\mathbf{a}^2 g_2(\mathbf{x}), \mathbf{a}^2 g_2(\mathbf{x}')) + \\
&\quad \text{cov}(\mathbf{a}^1 g_1(\mathbf{x}), \mathbf{a}^2 g_2(\mathbf{x}')) + \text{cov}(\mathbf{a}^2 g_2(\mathbf{x}), \mathbf{a}^1 g_1(\mathbf{x}')) \\
&= \text{cov}(\mathbf{a}^1 g_1(\mathbf{x}), \mathbf{a}^1 g_1(\mathbf{x}')) + \text{cov}(\mathbf{a}^2 g_2(\mathbf{x}), \mathbf{a}^2 g_2(\mathbf{x}')) \\
&= \mathbf{a}^1 (\mathbf{a}^1)^\top \text{cov}(g_1(\mathbf{x}), g_1(\mathbf{x}')) + \mathbf{a}^2 (\mathbf{a}^2)^\top \text{cov}(g_2(\mathbf{x}), g_2(\mathbf{x}')) \\
&= (\mathbf{a}^1 (\mathbf{a}^1)^\top + \mathbf{a}^2 (\mathbf{a}^2)^\top) k(\mathbf{x}, \mathbf{x}') \\
&= \mathbf{W} k(\mathbf{x}, \mathbf{x}').
\end{aligned}$$

The above covariance function can then be extended for the general case. Given a set of outputs, $\{f_d(\mathbf{x})\}_{d=1}^D$, which can then be expressed in terms of L functions that are Gaussian processes sampled independently with the same covariance function $k(\mathbf{x}, \mathbf{x}')$ (Goovaerts 1997),

$$f_d(\mathbf{x}) = \sum_{l=1}^{L} a_d^l g_l(\mathbf{x}).$$

Then, for $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) \cdots f_D(\mathbf{x})]^\top$, the covariance function, $\text{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}'))$ is given as,

$$\text{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) = \left(\sum_{l=1}^{L} \mathbf{a}^l (\mathbf{a}^l)^\top \right) k(\mathbf{x}, \mathbf{x}') \tag{5.11}$$

$$= \mathbf{A}\mathbf{A}^\top k(\mathbf{x}, \mathbf{x}') \tag{5.12}$$

$$= \mathbf{W} k(\mathbf{x}, \mathbf{x}'), \tag{5.13}$$

where $\mathbf{A} = [\mathbf{a}^1 \quad \mathbf{a}^2 \cdots \mathbf{a}^L]$, and $\mathbf{W}$ is a positive-definite matrix, with the rank of $\mathbf{W} \in \mathbb{R}^{D \times D}$ equal to L.

5.3.1.2 Semiparametric Latent Factor Model

The next step in the coregionalization approach is the Semiparametric Latent Factor Model ('SLFM') (Teh et al. 2005) which is an extension of ICM, where the requirement for the Gaussian processes $g_l(\cdot)$ to have the same covariance function is relaxed. Consider two outputs, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$. Then, given two functions sampled from Gaussian processes, $g_1(\mathbf{x}) \sim \mathcal{GP}(0, k_1(\mathbf{x}, \mathbf{x}'))$ and $g_2(\mathbf{x}) \sim \mathcal{GP}(0, k_2(\mathbf{x}, \mathbf{x}'))$, scaled versions of $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ can be used to obtain the outputs,

$$f_1(\mathbf{x}) = a_{1,1} g_1(\mathbf{x}) + a_{1,2} g_2(\mathbf{x})$$
$$f_2(\mathbf{x}) = a_{2,1} g_1(\mathbf{x}) + a_{2,2} g_2(\mathbf{x}).$$

The two equations above can then be expressed as a vector-valued function, as was done with the ICM approach.

$$\mathbf{f}(\mathbf{x}) = \mathbf{a}^1 g_1(\mathbf{x}) + \mathbf{a}^2 g_2(\mathbf{x}).$$

Again, $\mathbf{a}^1 = [a_1^1 \quad a_2^1]^\top$ and $\mathbf{a}^2 = [a_1^2 \quad a_2^2]^\top$. The computation of the covariance function for $\mathbf{f}(\mathbf{x})$ is similar to ICM, but with the addition of the different covariance functions.

$$\begin{aligned} \text{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) &= \mathbf{a}^1 (\mathbf{a}^1)^\top \text{cov}(g_1(\mathbf{x}), g_1(\mathbf{x}')) + \mathbf{a}^2 (\mathbf{a}^2)^\top \text{cov}(g_2(\mathbf{x}), g_2(\mathbf{x}')) \\ &= \mathbf{a}^1 (\mathbf{a}^1)^\top k_1(\mathbf{x}, \mathbf{x}') + \mathbf{a}^2 (\mathbf{a}^2)^\top k_2(\mathbf{x}, \mathbf{x}') \\ &= \mathbf{W}_1 k_1(\mathbf{x}, \mathbf{x}') + \mathbf{W}_2 k_2(\mathbf{x}, \mathbf{x}'). \end{aligned}$$

There are now two parts to the covariance function, corresponding to the different covariance functions $k_1(\cdot, \cdot')$ and $k_2(\cdot, \cdot')$, with $\mathbf{W}_1 = \mathbf{a}^1 (\mathbf{a}^1)^\top$ and $\mathbf{W}_2 = \mathbf{a}^2 (\mathbf{a}^2)^\top$ each of rank 1. The next step is to extend this approach to

the general case for a set of D outputs, $\{f_d(\mathbf{x})\}_{d=1}^{D}$, which can be expressed as $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) \cdots f_D(\mathbf{x}]^\top$. The outputs can then be expressed as follows,

$$f_d(\mathbf{x}) = \sum_{q=1}^{Q} a_{d,q} g_q(\mathbf{x}),$$

where $g_q(\mathbf{x})$ are Gaussian processes with covariance functions $k_q(\mathbf{x}, \mathbf{x}')$. Next, the covariance function, $\text{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}'))$ is expressed as,

$$\begin{aligned}
\text{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) &= \sum_{q=1}^{Q} \mathbf{a}^l (\mathbf{a}^l)^\top k_q(\mathbf{x}, \mathbf{x}') \\
&= \sum_{q=1}^{Q} \mathbf{A}_q \mathbf{A}_q^\top k_q(\mathbf{x}, \mathbf{x}') \\
&= \sum_{q=1}^{Q} \mathbf{W}_q k_q(\mathbf{x}, \mathbf{x}').
\end{aligned}$$

For the SLFM approach, we now have a sum of multiple matrices, $\mathbf{W}_q$ of rank 1, and the covariance functions corresponding to each latent function $g_q(\mathbf{x})$.

5.3.1.3 Linear Model of Coregionalization

Finally, the Linear Model of Coregionalization ('LMC') (Journel & Huijbregts 1978) combines the ICM and SLFM approaches to allow samples from Gaussian processes with different covariance functions as well as samples from Gaussian processes with the same covariance functions. Take a set of outputs $\{f_d(\mathbf{x})\}_{d=1}^{D}$. Then, consider Q different groups of samples. Each group of samples is taken from a Gaussian process with zero mean, and covariance function $k_q(\mathbf{x}, \mathbf{x}')$. Within each group, there are R_q samples obtained independently from the given Gaussian process. Essentially, the LMC approach is the sum of Q different

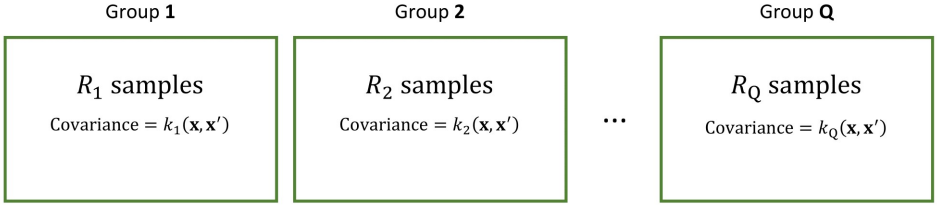


Figure 5.6: Figure to illustrate the different groups and sampling within LMC approach.

ICM's, shown in Figure 5.6. To illustrate this approach and show the calculation of the covariance, consider the case where there are two outputs ($D = 2$), two groups ($Q = 2$), and two samples within each group ($R_1, R_2 = 2$). In this case, $g_1^1(\mathbf{x}), g_1^2(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_1(\mathbf{x}, \mathbf{x}'))$, and $g_2^1(\mathbf{x}), g_2^2(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_2(\mathbf{x}, \mathbf{x}'))$. Then, the outputs are expressed as,

$$f_1(\mathbf{x}) = a_{1,1}^1 g_1^1(\mathbf{x}) + a_{1,1}^2 g_1^2(\mathbf{x}) + a_{1,2}^1 g_2^1(\mathbf{x}) + a_{1,2}^2 g_2^2(\mathbf{x})$$
$$f_2(\mathbf{x}) = a_{2,1}^1 g_1^1(\mathbf{x}) + a_{2,1}^2 g_1^2(\mathbf{x}) + a_{2,2}^1 g_2^1(\mathbf{x}) + a_{2,2}^2 g_2^2(\mathbf{x}).$$

As with the previous approaches, this will then be expressed as a vector-valued function.

$$\mathbf{f}(\mathbf{x}) = \mathbf{a}_1^1 g_1^1(\mathbf{x}) + \mathbf{a}_1^2 g_1^2(\mathbf{x}) + \mathbf{a}_2^1 g_2^1(\mathbf{x}) + \mathbf{a}_2^2 g_2^2(\mathbf{x}).$$

Here, $\mathbf{a}_q^{R_q} = [a_{1,q}^{R_q} \quad a_{1,q}^{R_q}]$. The covariance function is then expressed as follows,

$$\begin{aligned}
\operatorname{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) &= \mathbf{a}_1^1(\mathbf{a}_1^1)^\top k_1(\mathbf{x}, \mathbf{x}') + \mathbf{a}_1^2(\mathbf{a}_1^2)^\top k_1(\mathbf{x}, \mathbf{x}') + \\
&\quad \mathbf{a}_2^1(\mathbf{a}_2^1)^\top k_2(\mathbf{x}, \mathbf{x}') + \mathbf{a}_2^2(\mathbf{a}_2^2)^\top k_2(\mathbf{x}, \mathbf{x}') \\
&= (\mathbf{a}_1^1(\mathbf{a}_1^1)^\top + \mathbf{a}_1^2(\mathbf{a}_1^2)^\top) k_1(\mathbf{x}, \mathbf{x}') + (\mathbf{a}_2^1(\mathbf{a}_2^1)^\top + \mathbf{a}_2^2(\mathbf{a}_2^2)^\top) k_2(\mathbf{x}, \mathbf{x}') \\
&= \mathbf{A}_1 \mathbf{A}_1^\top k_1(\mathbf{x}, \mathbf{x}') + \mathbf{A}_2 \mathbf{A}_2^\top k_2(\mathbf{x}, \mathbf{x}') \\
&= \mathbf{W}_1 k_1(\mathbf{x}, \mathbf{x}') + \mathbf{W}_2 k_2(\mathbf{x}, \mathbf{x}').
\end{aligned}$$

Each of the matrices, $\mathbf{W}_q$ are known as the coregionalization matrices with rank R_q. Expanding on the example above, the LMC approach can be generalised. Consider the set of outputs, $\{f_d(\mathbf{x})\}_{d=1}^D$, as well as Q groups of samples, each with a given number of samples, R_q, for $q = 1, \ldots Q$. Each of the outputs can then be expressed as (Journel & Huijbregts 1978, Goovaerts 1997),

$$f_d(\mathbf{x}) = \sum_{q=1}^{Q} \sum_{i=1}^{R_q} a_{d,q}^i g_q^i(\mathbf{x}),$$

where $g_q^i(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_q(\mathbf{x}, \mathbf{x}'))$. Then for $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) \cdots f_D(\mathbf{x})]^\top$, the covariance function is specified as,

$$\begin{aligned}
\operatorname{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) &= \sum_{q=1}^{Q} \mathbf{A}_q \mathbf{A}_q^\top k_q(\mathbf{x}, \mathbf{x}') \\
&= \sum_{q=1}^{Q} \mathbf{W}_q k_q(\mathbf{x}, \mathbf{x}'),
\end{aligned}$$

where $\mathbf{A}_q = [\mathbf{a}_q^1 \quad \mathbf{a}_q^2 \cdots \mathbf{a}_q^{R_q}]$, and $\mathbf{W}_q = \mathbf{A}_q(\mathbf{A}_q)^\top$ are the coregionalization matrices with rank R_q (Alvarez et al. 2012). The covariance functions $k_q(\mathbf{x}, \mathbf{x}'))$ can be chosen from the same covariance functions that are used in the single output Gaussian processes, with one of the most popular choices being the squared exponential, as mentioned in Chapter 4.

The linear model of coregionalization is considered as a simple way of introducing correlations in the outputs, where the outputs are expressed as linear combinations of independent random functions (Alvarez & Lawrence 2009). Using this approach within a Gaussian process framework, the independent random functions are Gaussian processes, which results in the model also being a Gaussian process (Alvarez & Lawrence 2009). The linear model of coregionalization can be considered as an efficient way to incorporate correlations between outputs within a multi-output Gaussian process (van der Wilk et al. 2020).

5.3.2 Convolution processes

The linear model of coregionalization approach has it's limitations, and is considered to be a restrictive approach to constructing multi-output covariance functions. One example of the limitations are that it is not able to capture outputs which are delayed versions of each other. Convolution processes are able to overcome this problem and can account for time-lag relationships and general dependence on past observations (Alvarez & Lawrence 2009). Convolution processes were considered in different forms as an alternative to the linear model of coregionalization approach (Higdon 2002, Alvarez & Lawrence 2009, 2011), to allow the correlation structures to account for relationships such as time-lags and general linear dependence on past observations (van der Wilk et al. 2020). Alvarez et al. (2010) constructs $\mathbf{f}(\cdot)$ from a convolution of $\mathbf{g}(\cdot)$ as,

$$\mathbf{f}(\mathbf{x}) = \int G(\mathbf{x} - \mathbf{z})\mathbf{g}(\mathbf{z})d\mathbf{z}, \qquad \text{with } G(\mathbf{z}) \in \mathbb{R}^{P \times L}.$$

The resulting covariance function can be expressed as follows, when taking the same prior on $\mathbf{g}(\mathbf{x})$ as before (Alvarez et al. 2010):

$$\begin{aligned} k(\{\mathbf{x}, p\}, \{\mathbf{x}', p'\}) &= \mathbb{E}_{\mathbf{g}}\left[\int\int G(\mathbf{x} - \mathbf{z})\mathbf{g}(\mathbf{z})\mathbf{g}(\mathbf{z}')^\top G(\mathbf{x}' - \mathbf{z}')^\top d\mathbf{z}d\mathbf{z}'\right] \\ &= \sum_{q=1}^{L}\int\int G_{pq}(\mathbf{x} - \mathbf{z})G_{p'q}(\mathbf{x}' - \mathbf{z}')k_l(\mathbf{z}, \mathbf{z}')d\mathbf{z}d\mathbf{z}'. \end{aligned}$$

In order to balance flexibility against susceptibility to overfitting, $G(\cdot)$ is normally parameterised such that it makes the integral tractable and adds a number of parameters (van der Wilk et al. 2020).

One drawback to the convolution processes approach is the computational cost of considering the full covariance function of the joint Gaussian process (Alvarez & Lawrence 2009). The computational complexity can be considered as $\mathcal{O}(N^3D^3)$, and the storage expressed as $\mathcal{O}(N^2D^2)$, for a Gaussian process with D output dimensions and N data points. This lead to a sparse approximation being considered by Alvarez & Lawrence (2009, 2011) which would reduce the computational burden. However, the convolutional approach proposed by Alvarez & Lawrence (2009, 2011) requires the inversion of a $DN \times DN$ matrix, which is not feasible for large datasets (Davies et al. 2019). In addition, it was mentioned previously that the convolutional Gaussian process approach was able to overcome the time-lag relationship limitations within the linear model of coregionalization approach, which is not an issue that will occur within this research, as there is no time feature within the data from NewDEPOMOD being considered. As a result, the linear model of coregionalization will be considered as the multi-output Gaussian process approach.

5.3.3 Application of multi-output Gaussian processes to NewDEPOMOD

The fitting of the multi-output Gaussian processes will be done using the linear model of coregionalization approach. Previously, three versions of the linear model of coregionalization were considered which depended on the latent structure of the outputs. For the NewDEPOMOD data, there are two outputs being considered, Total Area Impacted and 99th Percentile of Solids Flux. In the simplest case the ICM method considers latent functions to describe the outputs, each with the same covariance structure. The outputs can be expressed in terms of L latent functions, $g_l(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot'))$ for $l = 1, \ldots, L$.

$$Y_1 = f_1(\mathbf{x}) = \sum_{l=1}^{L} a_1^l g_l(\mathbf{x})$$

$$Y_2 = f_2(\mathbf{x}) = \sum_{l=1}^{L} a_2^l g_l(\mathbf{x}).$$

As was seen in Equation 5.13, the covariance structure for the outputs, $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$, is given as,

$$\begin{aligned}\text{cov}(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}')) &= \mathbf{A}\mathbf{A}^\top k(\mathbf{x}, \mathbf{x}') \\ &= \mathbf{W}k(\mathbf{x}, \mathbf{x}'),\end{aligned}$$

where $\mathbf{A} = [\mathbf{a}^1 \quad \mathbf{a}^2 \cdots \mathbf{a}^L]$, and $\mathbf{W}$ is positive-definite with rank L. This approach can be applied to the NewDEPOMOD data, with a choice for L being required. The extensions of the ICM method, such as the SLFM and LMC, relate to different structures of the latent functions. For the modelling of the NewDEPOMOD output, the ICM approach will be considered as it reduces the complexity of the model by only including one covariance function in the covariance structure of the outputs. This will reduce the computational time required to optimize the Gaussian process model, as less hyperparameters are included. The choice that has to be made then relates to the number of latent functions, L, that are to be used when fitting the model. This will be investigated further.

This application and optimization of the Gaussian process models can be computationally demanding, and so steps are required to reduce the computational time and storage of fitting and optimizing the Gaussian process models. Sparse approaches were considered in Chapter 4 to overcome the computational demands of fitting and optimizing the scalar output Gaussian processes. The most efficient and best performing sparse approach was the Subset of Data ('SD') method, which was the most simple. This approach required a subset of the data being used to fit the Gaussian process, which reduces the size of the matrix to be inverted and therefore the computational cost. The data being used across this analysis considers either 400 or 450 different input sets, at which a number of replicate runs were completed. Therefore, a reasonable approach for the SD method would be to consider one of the runs for each input set, chosen at random. The detailed investigation at Ardentinny will consider a number of different subsets to assess whether the samples chosen have a large impact on the performance of the model.

5.3.3.1 Detailed investigation at Ardentinny

For the multi-output Gaussian process application to NewDEPOMOD, the data for Ardentinny will be considered first in more detail. The aim of this is to determine the best approach for fitting the multi-output Gaussian process model, before applying this framework to the remaining sites. One of the first

considerations that is made is the base kernel function that will be used. It has been mentioned previously that the squared exponential is a popular choice within the literature, and this will therefore be considered throughout this investigation. Within this kernel function, the choice can be made as to whether or not to include separate lengthscales for each of the inputs. In addition, the choice of the number of latent functions to be used will be investigated.

As with the single-output Gaussian processes in Chapter 4, the L-BFGS method will be used to optimize the hyperparameters in the models. The first investigation will focus on the choices for the hyperparameters such as the lengthscale and the number of latent functions. Following this, an investigation will consider multiple subsets of the data to assess whether this has any effect on the performance of the models.

For the first investigation, one of the subsets of the data containing 400 NewDEPOMOD observations (one sample for each input set), are considered. The choices that will be considered in more detail relate to the rank of the matrix $\mathbf{W}$ from Equation 5.13, which defines the number of latent functions, and the choice of a single lengthscale or separate lengthscales for each input, for the squared exponential covariance function introduced in Chapter 4. The different Gaussian process setups that will be considered are as follows:

- $L = \{1, 2, 3\}$ with a single lengthscale for each input,
- $L = \{1, 2, 3\}$ with separate lengthscales for each input.

Gaussian process models will be fitted and optimized for models considering a single lengthscale for all of the inputs, and different numbers of latent functions, L. In addition, the Gaussian process models with separate lengthscales will be fitted and optimized using different numbers of latent functions, L. After fitting the Gaussian process models with the above settings, their predictive performance was assessed by estimating the outputs using the test data and comparing them to the NewDEPOMOD output by calculating the RMSE and MAE, with the results given in Table 5.4. Table 5.4 shows that altering the number of latent functions, L, does not affect the predictive performance when considering the RMSE and MAE to 3 decimal places. When considering more decimal places, there are some differences between the RMSE and MAE values, but it suggests the impact of changing these is small. The other thing seen in Table 5.4 is that the predictive performance improves when using separate lengthscales, which would be expected as using separate lengthscales allows for a more flexible model.

The next stage of the investigation considers multiple different subsets of the

	Total Area Impacted		99th Percentile	
Hyperparameter Settings	**RMSE**	**MAE**	**RMSE**	**MAE**
$L = 1$, single lengthscale	0.172	0.118	0.189	0.151
$L = 2$, single lengthscale	0.172	0.118	0.189	0.151
$L = 3$, single lengthscale	0.172	0.118	0.189	0.151
$L = 1$, separate lengthscales	0.076	0.047	0.056	0.035
$L = 2$, separate lengthscales	0.076	0.047	0.056	0.035
$L = 3$, separate lengthscales	0.076	0.047	0.056	0.035

Table 5.4: Table of RMSE and MAE for the different hyperparameter settings for multi-output Gaussian process at Ardentinny.

data. The previous investigation determined that the number of latent functions did not alter the predictive performance, but that separate lengthscales provided better results due their increased flexibility. A total of five different subsets of the data are considered, and multi-output Gaussian processes are fitted with separate lengthscales and $L = 1$. The RMSE and MAE are provided in Table 5.5. Considering the predictive performance of the multi-output Gaus-

	Total Area Impacted		99th Percentile	
Subset number	**RMSE**	**MAE**	**RMSE**	**MAE**
Subset 1	0.076	0.047	0.056	0.035
Subset 2	0.078	0.050	0.049	0.031
Subset 3	0.080	0.050	0.050	0.032
Subset 4	0.073	0.044	0.048	0.031
Subset 5	0.080	0.051	0.051	0.034

Table 5.5: Table of RMSE and MAE for the different multi-output Gaussian process at Ardentinny using multiple subsets of the data.

sian processes for the different subsets, there is some variation in the RMSE and MAE values for each of the outputs. The variation across the RMSE and MAE values is small when considering the data lies in the interval $[0, 1]$. Next, comparing the RMSE and MAE for the two outputs, the multi-output Gaussian process has better predictive performance for the 99th Percentile of Solids Flux. As noted in Chapter 4, this suggests that the changes in the inputs perform better at explaining the variation in the 99th Percentile. This will be considered further when looking at the multi-output Gaussian processes fitted for the remaining sites. The small differences between the RMSE and MAE for the different subsets suggest that each of the fitted models perform equally well and the sub-setting approach is appropriate.

5.3.4 Multi-output Gaussian process emulation for all sites

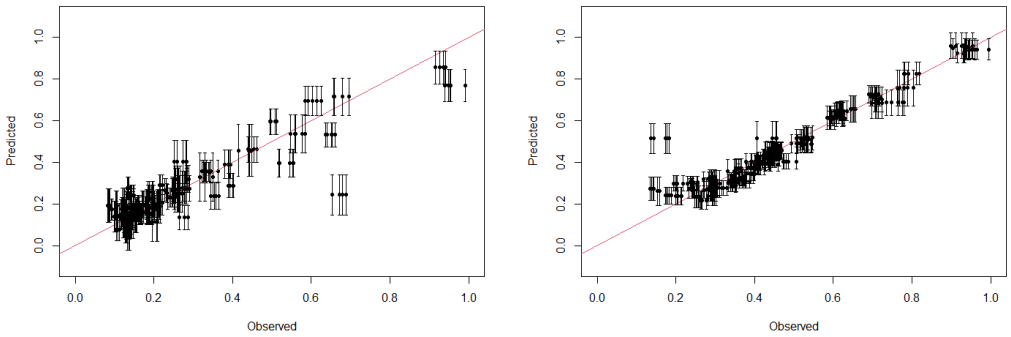
It was mentioned previously that a sub-setting approach will be considered which takes the outputs from one of the NewDEPOMOD runs for each input set at all sites. The multi-output Gaussian process models were fitted using the same hyperparameter settings as were used for Ardentinny, where $L = 1$ and separate lengthscales are used for each of the inputs. The RMSE and MAE values for each of the sites are given in Table 5.6. Firstly, looking at Table

	Total Area Impacted		**99th Percentile**	
Site	**RMSE**	**MAE**	**RMSE**	**MAE**
Ardentinny	0.076	0.047	0.056	0.035
West Strome	0.034	0.021	0.036	0.019
Muck	0.152	0.108	0.145	0.097
Djuba Wick	0.070	0.039	0.099	0.055

Table 5.6: Table of RMSE and MAE for the different multi-output Gaussian process at all of the sites using multiple subsets of the data.

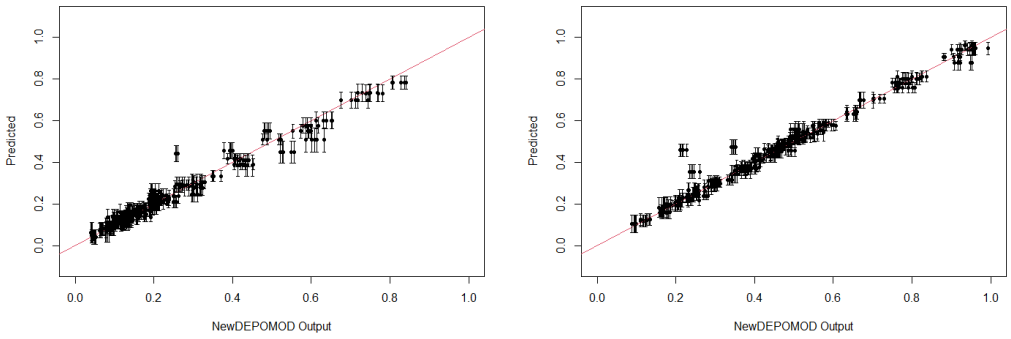
5.6, the RMSE and MAE values for the Total Area Impacted are similar for Ardentinny and Djuba Wick, with the predictions for West Strome being the most accurate and the predictions for Muck being the least accurate. The variation in the Total Area Impacted at West Strome appears to be well explained by the changes in the inputs, such as Settling Velocity of Faeces which had a much larger importance value than the other inputs in the sensitivity analysis from Chapter 2. The RMSE and MAE at Muck for both of the inputs are higher in comparison to the other sites. This could be a result of the faster flow speeds resulting in much more variation in the deposition of waste on the seabed which is unable to be explained by the changes in the inputs. The predictions for the test set are plotted against the NewDEPOMOD output to explore the predictive performance further in Figures 5.7 to 5.10, with error bands for the predictions included. At Ardentinny, in Figures 5.7a and 5.7b, the majority of the points lie close to the line of equality with a large amount of the error bands overlapping the line, particularly for the 99th Percentile of Solids Flux. This is supported by the RMSE and MAE in Table 5.6, which had lower values for the 99th Percentile of Solids Flux. The points with error bands that do not overlap the line of equality appear to be related to under-predictions in most cases for Total Area Impacted, and over-predictions for the 99th Percentile of Solids Flux.

In Figures 5.8a and 5.8b, the predictions for the test set at West Strome



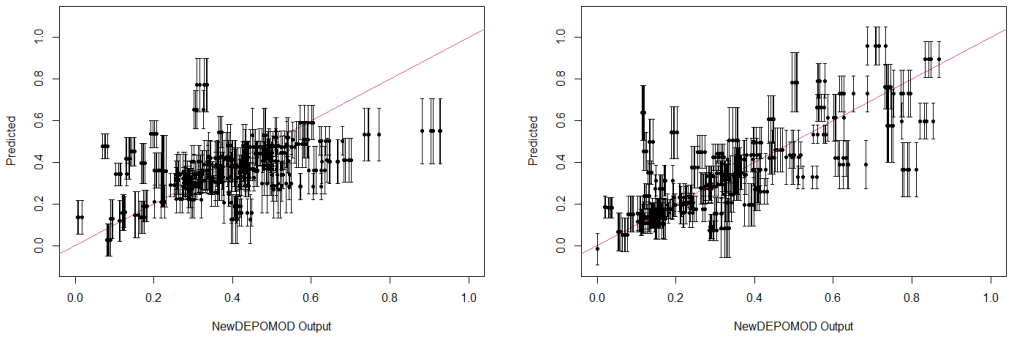
(a) Area (b) 99th Percentile of Solids Flux

Figure 5.7: Plots of the predicted outputs for the test data against the NewDEPOMOD output - Ardentinny.



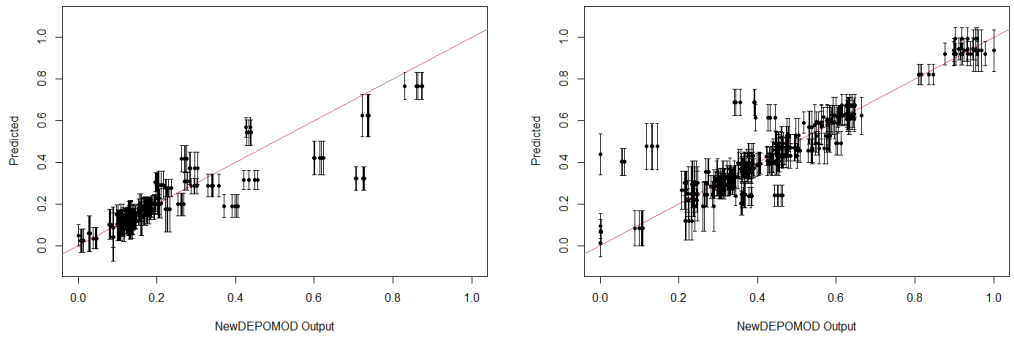
(a) Area (b) 99th Percentile of Solids Flux

Figure 5.8: Plots of the predicted outputs for the test data against the NewDEPOMOD output - West Strome.



(a) Area (b) 99th Percentile of Solids Flux

Figure 5.9: Plots of the predicted outputs for the test data against the NewDEPOMOD output - Muck.



(a) Area (b) 99th Percentile of Solids Flux

Figure 5.10: Plots of the predicted outputs for the test data against the NewDEPOMOD output - Djuba Wick.

show good agreement with the NewDEPOMOD output for both outputs. This was expected as the RMSE and MAE had the lowest values in Table 5.6. In addition, it should be highlighted that the error bands at West Strome are much smaller than the other sites.

Considering the predictions at Muck, these were expected to be poor based on the RMSE and MAE from Table 5.6. Figures 5.9a and 5.9b highlight this, with a number of over and under-predictions for both outputs. In addition, the error bands for each of the outputs are larger at Muck than they are at the other sites. This again indicates that the variation in the two outputs at Muck cannot be explained by the changes in the inputs.

Finally, looking at figures 5.10a and 5.10b, there is good agreement between the predictions and the NewDEPOMOD output for the Total Area Impacted for the lower values, but there appears to be a mix of over and under-prediction for the larger values of Total Area Impacted. For the 99th Percentile of Solids Flux, there is mostly good agreement between the predictions and the NewDEPOMOD output, but with some areas of over-prediction. In comparison to the other sites, the Total Area Impacted in Figure 5.10a has most of the data situated around low values for Total Area Impacted.

To summarise, it is clear from Table 5.6 that there are differences between the sites in terms of the performance of the multi-output Gaussian processes. West Strome performed better than the other sites based on the RMSE and it can also be highlighted when considering Figure 5.8. The multi-output Gaussian process for Muck did not perform well when predicting at the new data, suggesting the variance cannot be explained by the changes in the inputs.

5.4 Comparison of multivariate output emulation to independent scalar output emulation

The investigations within this Chapter considered the extension of the independent modelling of the scalar outputs, with the aim of assessing whether or not there is any information gain by modelling the outputs jointly to account for correlations between them. In order to compare between the different approaches for each site, a subset of the data was used fit the emulator models for each approach. A sparse subset of data approach was considered, the same approach as for the multi-output Gaussian processes, where one replicate run from each input set was considered, resulting in either 400 or 450 runs being used to fit the models depending on the site.

It was mentioned previously that there are differences between the sites, even between the sites with similar flow speeds and characteristics. As a result, each site will be considered separately for comparing the different approaches, with Ardentinny being considered first. The best performing method at Ardentinny for both outputs, for RMSE and MAE, is the multi-output Gaussian processes, but with only a small difference between independent random forests and multi-output Gaussian processes for Total Area Impacted. This suggests that this flexible approach that accounts for correlations between the inputs is successful at improving the predictive performance. Next, West Strome will be considered to determine if there are similarities in the performances of each method. Table 5.8 also shows that the Gaussian process approaches performs best for both outputs, with the independent emulators performing slightly better. The random forest methods perform similarly, with the RMSE and MAE being approximately double the values from the Gaussian process approaches for both outputs.

	Total Area Impacted		**99th Percentile**	
Emulation Approach	**RMSE**	**MAE**	**RMSE**	**MAE**
Independent RFs	0.086	0.057	0.071	0.045
Multivariate RFs	0.127	0.080	0.079	0.052
Independent GPs	0.126	0.069	0.078	0.043
Multi-output GPs	0.076	0.047	0.056	0.035

Table 5.7: Table of RMSE and MAE for each output at Ardentinny for the different emulation methods. (*RF = Random forest, GP = Gaussian process*)

Now the high energy sites will be considered, looking at Muck first in Table

	Total Area Impacted		99th Percentile	
Emulation Approach	**RMSE**	**MAE**	**RMSE**	**MAE**
Independent RFs	0.062	0.042	0.045	0.033
Multivariate RFs	0.075	0.048	0.056	0.041
Independent GPs	0.025	0.018	0.032	0.017
Multi-output GPs	0.034	0.021	0.036	0.019

Table 5.8: Table of RMSE and MAE for each output at West Strome for the different emulation methods. (*RF = Random forest, GP = Gaussian process*)

5.9. It was seen earlier, in Table 5.6 and Figure 5.9, that the performance of

	Total Area Impacted		99th Percentile	
Emulation Approach	**RMSE**	**MAE**	**RMSE**	**MAE**
Independent RFs	0.139	0.101	0.106	0.080
Multivariate RFs	0.136	0.106	0.104	0.079
Independent GPs	0.154	0.115	0.137	0.095
Multi-output GPs	0.152	0.108	0.145	0.097

Table 5.9: Table of RMSE and MAE for each output at Muck for the different emulation methods. (*RF = Random forest, GP = Gaussian process*)

the multi-output Gaussian process was poor for predicting at the test data. This was also seen for the other approaches when considering the Total Area Impacted, with similar RMSE and MAE values seen. When comparing the methods for both outputs, the two random forest approaches appear to perform best, with similar RMSE and MAE values for this dataset. As was mentioned previously, at this site there was a large amount of variation that could not able to be explained by the changes in the inputs, potentially a result of the fast flowing currents. Finally, the predictive performance of the approaches is considered for Djuba Wick in Table 5.10. For Djuba Wick, in Table 5.10, the

	Total Area Impacted		99th Percentile	
Emulation Approach	**RMSE**	**MAE**	**RMSE**	**MAE**
Independent RFs	0.064	0.032	0.074	0.043
Multivariate RFs	0.094	0.045	0.110	0.061
Independent GPs	0.054	0.028	0.068	0.037
Multi-output GPs	0.070	0.039	0.099	0.055

Table 5.10: Table of RMSE and MAE for each output at Djuba Wick for the different emulation methods. (*RF = fandom Forest, GP = Gaussian process*)

independent Gaussian processes have the best predictive performance for both outputs. There does not appear to be a big difference between the RMSE and MAE for the independent random forests and independent Gaussian processes

for the two outputs, indicating that the independent modelling of the outputs is better suited for this site. Considering both the high energy sites, Tables 5.9 and 5.10, there does not seem to be any improvement when modelling the two outputs together, with the independent random forests and independent Gaussian processes performing best at these sites.

In summary, considering the outputs jointly for the low energy sites, Ardentinny and West Strome, using multi-output Gaussian processes produces the best predictive performance when considering RMSE and MAE. However, for the two high energy sites, Muck and Djuba Wick, the independent random forest approach for each output out-performs the other methods. It is possible that the relationships between the outputs at these sites are more complex than the ones being considered for the multivariate output approaches that were considered, and also that the higher current speeds mean that the variations in the output are not explained as well by the changes in the inputs. Overall, the comparisons have indicated that the best performing method appears to be site specific.

5.5 Discussion

The main aim of this Chapter was to investigate the possibility of considering the two scalar outputs as a multivariate output in an extension to the previous scalar output random forests and Gaussian processes. In addition, the main focus of emulation is to approximate a complex model without the computational cost. The extensions to the scalar output emulation models in Chapter 4 are able to incorporate multivariate outputs with correlations. Each of the approaches that were considered were able to predict the outputs at new data with low computational cost - with predictions taking less than a second.

The first approach that was considered was a multivariate extension to random forest models which was able to account for a linear relationship between the outputs, through the use of a Mahalanobis splitting rule (Segal & Xiao 2011). Initial exploration of the Total Area Impacted and the 99th Percentile of Solids Flux showed a non-linear relationship between the outputs for the different Biomass values, but with some linearity across the range of values. The Mahalanobis splitting rule was used to assess it's suitability for modelling the outputs jointly. Comparison of the RMSE and MAE for the multivariate random forest models to the independent random forest models for each output did not indicate any improvement in the prediction when introducing a linear relationship between the outputs, indicating that a more complex relationship

is more suitable. The multivariate random forest could therefore be developed further through the introduction of more complex relationships between the outputs. Additionally, by modelling the outputs using a multivariate random forest, the ranking of the inputs changed, with the operational inputs having a greater influence. This is possibly related to the fact that the operational inputs determine how much waste is released from the cages, and so when modelling the two outputs together this influence is increased, compared to when they are modelled separately.

The second approach to modelling the outputs jointly considered a multi-output Gaussian process method. Linear model of coregionalization (Journel & Huijbregts 1978) was considered, as well as convolution processes (Alvarez & Lawrence 2009, 2011). The linear model of coregionalization approach is simpler, and was considered more appropriate for the data, as described previously. Investigations relating to the choices to be made around some of the hyperparameters were conducted, which identified the number of latent processes as having no influence on the RMSE and MAE. In addition, a single lengthscale was considered as well as separate lengthscales, with the separate lengthscales improving the fit with the added flexibility. A further investigation considered the sub-setting of the data to allow for a sparse Gaussian process approach to be considered. This identified a sub-setting method that was applied to all sites, with the predictive performance of the multi-output Gaussian processes assessed for the test data using RMSE and MAE. Analysis of the RMSE and MAE for the two outputs at the different sites identified West Strome as the best performing, and the variation at Muck not being well explained by the changes in the inputs.

Finally, the performance of the multi-output Gaussian processes and random forests were compared to the independent modelling of the scalar outputs that was considered in Chapter 4. The comparisons between the performance of the methods for each site indicate that the method with the best predictive performance is site specific.

Chapter 6

Conclusions, Discussion & Future Work

Process-based models such as NewDEPOMOD are effective tools that are used to assess environmental challenges where collecting data is not practical or effective. The challenges of process-based modelling include their computational cost, and they do not account for uncertainty.

Uncertainty in process-based models can be investigated through the use of sensitivity analyses and uncertainty quantification to quantify uncertainty in the model outputs and attribute them to variations in the model inputs (Saltelli et al. 2004). Using sensitivity analysis techniques can increase confidence in model predictions by improving the understanding of how the model output reacts to changes in the inputs (Saltelli et al. 2000). There are three main types of sensitivity analyses that can be used: 1) Ranking of the inputs by their influence on the variability of the output, 2) Screening to identify the inputs that do not contribute to the variability of the output and 3) Mapping to identify the areas of the input space that produce extreme output values. Saltelli et al. (2000) developed a framework that could be applied to sensitivity analyses to answer specific questions about models. Classical sensitivity analysis techniques focus on univariate or multivariate output, but one of the novelties within this thesis is the extension to the classic approaches to handle maps as the output.

As was mentioned above, process-based models suffer from computational challenges due to their complexity. The computational cost can be reduced by using statistical emulation techniques which approximate the output from process-based models (Conti & O'Hagan 2010). Development of statistical emulators are considered as a fundamental step when looking to gain a deeper understanding of a complex process-based model (Overstall & Woods 2016).

The aim of a statistical emulator is to create a statistical model to imitate a process-based model, using a set of costly training runs that were completed using the process-based model. Different statistical methods such as linear regression, generalized linear models, regression splines and Gaussian processes can be used to emulate the process-based models (Grow & Hilton 2018). Statistical emulators are then an effective tool for investigating uncertainty within a process-based model and used as an approximation which can be used for predicting at a range of input sets (Overstall & Woods 2016).

The main aims of this thesis were to develop novel sensitivity analysis and emulation tools through the consideration of NewDEPOMOD and the modelling of environmental waste from fish farms. The thesis investigated the impact on NewDEPOMOD output of uncertainty within the NewDEPOMOD inputs, and to develop a statistical emulator of NewDEPOMOD to overcome the computational challenge of running it several times when testing multiple model setups. The impact of uncertainty within the NewDEPOMOD inputs on the NewDEPOMOD output were analysed in this research through sensitivity analyses for both the univariate output data from NewDEPOMOD, and the output maps produced by NewDEPOMOD. The aim of the sensitivity analyses were to rank the inputs and determine the most influential inputs at different sites. Next, emulation of NewDEPOMOD was considered, with initial investigations focusing on the emulation of the univariate output data, before expanding on this to develop a statistical emulator for multiple, correlated outputs. Random forest regression and Gaussian processes are considered as emulation techniques for the univariate output data from NewDEPOMOD, before using extended approaches to account for the multiple, correlated outputs. Summaries from each of the analyses will be considered before discussing the overall success of the research and identifying areas for future work.

6.1 Sensitivity analysis for univariate outputs

The aim of this analysis in Chapter 2 was to determine which of the uncertain inputs had the biggest impact on the univariate outputs from NewDEPOMOD - Total Area Impacted, 99th Percentile of Solids Flux and Mass Balance. Two groups of inputs were considered: 1) inputs based on the physical properties (continuous inputs) and 2) operational inputs (categorical inputs). Initially, separate sensitivity analyses were conducted for each group of inputs, before completing a combined analysis to assess the impact of altering both at the same time.

The inputs based on the physical properties consisted of a set of inputs that were considered due to uncertainty surrounding their default parameter values, based on the previous modelling experience of SEPA. The inputs based physical properties inputs initially consisted of 11 different inputs, which consisted of the Settling Velocity of Faeces, four inputs related to resuspension of waste on the seabed, and six inputs related to the random walk within NewDEPOMOD. After identifying the inputs to be considered, their ranges were identified through analysis of the literature, as well as through collaboration with SEPA where ranges could not be determined through the literature alone. The inputs based on the physical properties were all continuous, and a LHS approach with a restricted pairing procedure was used to capture as much of the input space as possible while accounting for correlations between some inputs (McKay et al. 1979, Iman & Conover 1982).

An initial sensitivity analysis of the inputs based on the physical properties was conducted in order to rank the inputs by their influence on the univariate outputs. These were ranked for each of the outputs at two different fish farm sites using importance values obtained through random forest modelling (Breiman 2001, Harper et al. 2011). The two sites being considered had different characteristics, with one being considered as a low energy site and the other as a high energy site. At both sites, the Critical Shear Stress for Erosion and the Settling Velocity of Faeces were consistently ranked highly for each of the univariate outputs being considered. In contrast, the inputs related to the random walk element of NewDEPOMOD were consistently ranked lower at both sites for each of the outputs. This indicated that altering the size of the step within the random walk element of NewDEPOMOD did not have a big impact on the univariate outputs. As a result, it was concluded that the random walk inputs would be removed from future analyses, but with a number of replicate runs being completed for each input set in the analyses to capture stochasticity arising from the random walk component in NewDEPOMOD.

Next, the consideration of the operational inputs required a different approach as these inputs relate to the farm setup and could be altered by a farm operator. The three outputs that were considered were Biomass, Cage Diameter and Number of Cages, all of which were treated as categorical inputs. These inputs were considered to explore the impact on NewDEPOMOD predictions of increasing the Biomass to allow for future expansion within the industry. Two ways to increase the Biomass without reducing the levels of fish welfare are: 1) to increase the Cage Diameter or 2) increase the Number of Cages. Altering the operational outputs had to be kept realistic in relation to the possible choices

of cage numbers and size which allowed for different expansion scenarios to be considered. A flowchart was developed that allowed different scenarios to be tested.

This sensitivity analysis was again conducted using random forests and concluded that the characteristics of the low energy site, such as slow current speed, meant that the Number of Cages had the biggest influence on the Total Area Impacted. The additional cages meant a larger footprint on the area of the seabed directly below the farm. Biomass had the biggest influence on the Total Area Impacted at the high energy site, as well as the 99th Percentile of Solids Flux, due to larger amounts of waste being produced by the larger number of fish in the farm. This analysis considered different operational setups that could be implemented at farms, and how the characteristics of a site were influential in the ranking of the inputs.

Finally, the inputs based on the physical properties and the operational inputs were considered together. The number of inputs based on the physical properties were reduced following the initial analysis, with the inputs related to the random walk elements removed. This meant that there were five continuous inputs being considered in addition to the three operational inputs. This required an alternative sampling approach. The approach that was considered was a sliced LHS that was able create a space-filling design when continuous and categorical inputs are considered (Qian & Wu 2009, Ba et al. 2015).

Using the information gained from the previous sensitivity analyses, the framework developed to assess the influence of the two types of inputs was applied to two low energy and two high energy sites to allow comparisons to be made. Random forests were again used in order to rank the inputs based on their influence on the univariate outputs. Similarities were seen between the high and low energy sites, with the Settling Velocity of Faeces playing a dominant role. However, the influence of the operational inputs appeared to be different between the high and low energy sites, having a bigger effect at the high energy site.

6.2 Sensitivity analysis for output maps

Chapter 3 looked to expand on the sensitivity analysis techniques for the univariate outputs, to develop methods for considering multivariate outputs, such as NewDEPOMOD maps. The aim was to develop a framework that would be suitable for application to multiple sites. Three different approaches were considered for multivariate output data before deciding on a suitable framework

to be applied to additional sites.

The first approach considered a shape analysis for the NewDEPOMOD output maps. This involved identifying a main shape of the impact using landmarks, before performing a shape PCA to identify the main areas of variation (Dryden & Mardia 2016). Applying this approach to the data from the sensitivity analysis of the inputs based on the physical properties, it was able to identify the main areas of variation in the shapes. Modelling of the PC scores related to the main areas of variation identified the most influential inputs, with more of the resuspension inputs being identified as influential at the high energy site, where more resuspension will take place. One drawback to this approach was that it did not consider the deposition over the whole domain and only a main shape of deposition. It is not always possible to identify the main shape of the impact using the landmark approach, which led to the consideration of further methods which considered the whole domain.

The second approach aimed to use the data across the full domain, considering the output map as a surface using a bivariate functional approach. The functional representations of the output maps were produced using an adaptive smoothing approach with irregular basis functions to capture the large amounts of variation. A bivariate functional PCA was used to investigate the areas of variation across the domain (Gong et al. 2015). The main areas of variation were identified as being on the seabed below the cages. Despite being able to identify the main areas of variation across the domain, modelling of the PC scores provided limited explainability, therefore, conclusions could not be made about the inputs contributing to the variation.

The final approach considered the output from individual grid cells independently. Different modelling techniques were used to identify the inputs that had the biggest influence on the variance of the output in each grid cell. Variance decomposition approaches are common within sensitivity analyses to decompose the output variance and attribute it to the inputs (Saltelli et al. 2000). One measure of the variance explained by the inputs is η^2, which is a standardized measure of effect size for an ANOVA. It required the continuous inputs to be converted to categorical data. An alternative variance decomposition approach that is common within sensitivity analysis literature are Sobol indices (Sobol' 1993). A robust extension of Sobol indices were proposed to overcome the issue of Sobol indices being sensitive to outliers. Finally, random forest regression was considered as a way to rank the inputs in each of the grid cells across the domain. In addition, the outliers that affected the Sobol indices approach were considered in more detail to identify the inputs that may have

been responsible for the extremes.

The approach which considered the output from the individual grid cells produced the most interpretable results, and so this approach was used for the framework to be applied to multiple sites. The framework involved η^2 and random forest modelling to investigate the ranking of the inputs across the domain, as well as the consideration of the extreme values through quantile regression. The framework was applied to 4 sites in total, featuring two low and two high energy sites in order to compare the influence of the inputs across the domain for the sites with different characteristics. This analysis concluded that the effect of the inputs across the domain appears to be site dependent, with differences seen between the sites with similar characteristics. Similarities are seen between all of the sites with the Number of Cages being the highest ranking input in the areas of the seabed below where the additional cages are positioned. In addition, the Biomass and the Cage Diameter only feature as the highest ranking inputs in a small number of grid cells across the domains of all sites. When comparing the results from the low energy sites to the results from the high energy sites, the Settling Velocity of Faeces plays a bigger role across the domain at the low energy sites, with the resuspension inputs featuring more as the highest ranking inputs at the high energy sites. Quantile regression models fitted to the data to investigate the extreme values did not fit the data well, and so the analysis of the output data was considered with caution.

A framework was developed in order to assess the contribution of a set of inputs to variation seen in output maps. The framework was applied to the NewDEPOMOD data for multiple sites to allow comparisons to be made between low and high energy sites. This work identified differences between the influence of inputs across the domain for the different types of sites. The random forest approach was able to identify the most influential inputs across the domain, whereas the η^2 approach was only able to identify one input across all sites, suggesting that it did not perform well. The results for the random forest model were plausible when considering the characteristics of the sites, but did highlight that even for sites with similar characteristics, there are differences which suggests that sites should be considered independently in future work.

6.3 Emulation of univariate outputs

The objective for Chapter 4 was to develop a novel statistical emulation framework for the environmental impacts of fish farms to approximate the univariate

outputs without the computational cost of running NewDEPOMOD. The univariate outputs that were considered throughout this Chapter were the Total Area Impacted and the 99th Percentile of Solids Flux which provide a measure of the size and the intensity of the impact on the seabed. Two different modelling approaches were considered for this: 1) random forest regression and 2) Gaussian process regression. Training and test data was established for fitting the models and testing their predictive performance. Comparisons of random forests to Gaussian processes have previously been carried out (Mlaker et al. 2019, Shabani et al. 2020), which showed that the two methods had similar predictive capabilities.

The random forest approach was considered due to the high percentage of variation explained by the models used in Chapter 2. In addition, Gaussian processes were considered due to their flexibility and common use within emulation literature, as referenced in Chapter 4. The predictive performance of the two approaches was measured through the calculations of RMSE and MAE.

The predictive performance of the random forest models for each output was good for three of the four sites, with low RMSE and MAE values. However, at Muck, the RMSE and MAE values were much higher than for the other sites, which was seen when looking at the plots of the predictions for the test set against the NewDEPOMOD output in Figures 4.7c and 4.8c, where there was a lot more variation around the line of equality. Different approaches were considered to determine if a single emulator could be used for prediction at all sites, however, the RMSE and MAE for the sites was worse than for the independent random forest models. This indicated that each site should be considered independently and a single emulator was not appropriate.

The random forest and Gaussian process approaches have the benefit of being flexible and efficient to run for predicting at new input sets. One drawback to the Gaussian process approach is that it is computationally expensive to fit and optimize the model when using all of the data, and requires a sparse approach to reduce the computational time. A number of different sparse approaches were considered for one site, Ardentinny, before deciding on a method that would be applied to additional sites. However, an investigation of the predictive performance of a Gaussian process fitted using all of the data compared to the sparse approach, showed that the sparse approach actually out-performed the full Gaussian process. Multiple sparse approaches were considered such as the subset of data method which fitted an exact Gaussian process using a subset of the data. Additionally, the subset of regressors and fully independent conditional approximations were considered, which both

use a subset of the data to approximate the computationally expensive covariance function within the Gaussian process. The subset of data approach was identified as being more efficient and had better predictive performance when predicting for a test set at Ardentinny, and was therefore applied to all sites for both outputs.

For the Gaussian process regression, a number of investigations were considered to determine the sparse approach to be used as well as the number of inducing points to be included in the active set. The results indicated that a subset of data approach with 200 inducing points was best, and this was applied to both outputs at each site. The Gaussian process models had slightly higher RMSE and MAE values than for the random forest models at three of the four sites. At West Strome, the Gaussian process had better predictive performance for the Total Area Impacted, and similar RMSE and MAE values for the 99th Percentile of Solids Flux. These comparisons identified only one occasion where the Gaussian process approach performed better than the random forests - the Total Area Impacted at West Strome. This suggested that the random forest approach for emulation is better for the case where the univariate outputs are considered independently.

The two emulation approaches that were considered performed well for all of the sites, excluding Muck. Emulation is commonly used to approximate the output from complex mathematical models without the computational cost, before studying the uncertainty of the model output to variation in the model inputs. The emulators for each site that were developed could be used further to investigate the effect on the output variation of the inputs in more detail through uncertainty quantification, with a larger number of input sets considered than would be possible when using NewDEPOMOD.

6.4 Emulation of multivariate outputs

An expansion of the emulation framework from Chapter 4 was considered in Chapter 5 to create a multivariate output emulation framework that accounted for correlation between the outputs. The two univariate outputs featured in Chapter 4 are considered as a multivariate output to assess if there is any information gain from . Again, a random forest and Gaussian process approach were considered, where each method is expanded to account for multivariate outputs where correlations exist. (Segal & Xiao 2011) proposed an extension to the standard random forest that accounts for multiple outputs that are linearly related, which can be done through a Mahalanobis splitting method.

A number techniques have been considered for the multiple output extension for Gaussian processes (Conti & O'Hagan 2010, Higdon et al. 2008, Alvarez & Lawrence 2009, 2011), with a linear model of coregionalization being used for the application to NewDEPOMOD (Alvarez & Lawrence 2011).

Browne et al. (2021) noted that there are a lack of more sophisticated multivariate random forest models that can account for relationships other than linear relationships. Initial scatterplots of the Total Area Impacted against the 99th Percentile of Solids Flux showed non-linear relationships between the outputs for each Biomass value, but with some areas of linearity. The Mahalanobis splitting method was applied with the multivariate random forest models to the data to investigate if the assumption of linearity provides any information gain. After fitting the multivariate random forest models for each site, their predictive performance was assessed using test data and calculations of the RMSE and MAE. Reviewing the RMSE and MAE values for each output using the multivariate random forest, they are slightly higher for three of the four sites than the values from the independent random forests. For West Strome, similar values are seen for the independent and multivariate random forests. In addition, when considering the importance values of the inputs for each output, differences were seen when comparing to the importance values from the univariate random forests in Chapter 4. The Biomass value had a bigger influence for the multivariate random forest, having the largest importance value for the 99th Percentile across the four sites. The operational inputs determine how much waste leaves the cages and enters the water column, and the multivariate random forest suggests that, when modelling the two outputs together, the operational inputs play a bigger role.

Alternatively, an extension to the Gaussian processes for correlated, multivariate outputs was considered. The linear model of coregionalization (Journel & Huijbregts 1978) approach was applied to Gaussian processes to allow for correlated, multivariate outputs to be considered. Considerations of the hyperparameter structure were required for this approach, such as the number of latent processes within the model, and the choice of a single or separate lengthscales for each input. These choices were investigated in more detail and determined that the number of latent processes had no major impact on the predictive performance of the models, but the inclusion of the separate lengthscales significantly improved the predictions. The hyperparameters within the multi-output Gaussian process were optimized using the L-BFGS algorithm, similar to the univariate output Gaussian processes. After fitting the multi-output Gaussian processes, the same training data from this model was used to

fit emulator models using independent random forests, independent Gaussian processes and multivariate random forests. When comparing the performance of the emulators, none of them performed consistently better across the 4 sites, and the best performing approach appears to be site specific.

6.5 Discussion, limitations and future work

This thesis has investigated how uncertainty of inputs within process-based models can affect the output from the model, using the application of sensitivity analysis techniques to the modelling of the environmental impacts of aquaculture with NewDEPOMOD. Sensitivity analyses were conducted for both the univariate outputs and the multivariate output maps from NewDEPOMOD. Random forests were an effective tool for the purpose of identifying the most influential inputs in relation to the univariate outputs that were considered. In addition, the random forest approach was incorporated in the framework that was developed for analysing the multivariate output maps produced by NewDEPOMOD. This was an effective tool that was able to identify areas of variation across the maps. The other measures did not perform as well, which is potentially a result of the outliers present within the output for the grid cells. Transformations of the data were considered to overcome these problems, but did not solve the problems of the data being skewed. When considering the grid cell data, a large proportion of it features Solids Flux values close to or equal to zero, so a possible approach to combat this problem would be to only consider grid cells where the average Solids Flux values are greater than a predefined value. In addition to the problems with the outliers, the approach that was considered within the framework considered the grid cells independently and did not account for spatial correlation. Not accounting for the spatial correlation when considering the output maps in as the output for the sensitivity analysis, limits the conclusions that can be drawn from the analysis. The analysis provided a general idea of the influence of the inputs across the domains. Therefore, the analysis could be extended to account for spatial correlation within the framework for the multivariate output maps.

Following the sensitivity analysis, statistical emulation approaches were considered to approximate the univariate and multivariate output from NewDEPOMOD. The application of emulation to modelling the environmental impacts of aquaculture is a novel approach. Initially, the univariate outputs were considered, with random forests and Gaussian processes proposed as emulation techniques due to their flexibility and low computational cost when predicting

for new input sets. The performance of the univariate emulators was good for three of the four sites being considered. For each output, a single random forest emulator was considered for all of the sites, but resulted in poor predictive performance, suggesting the inputs should be considered individually to gain accurate predictions. The emulation framework was then extended to account for correlated, multivariate outputs. The first approach to be considered was multivariate random forests, which were able to account for linear relationships between the outputs. In most cases, this did not produce better predictive performance, which indicated that the assumption of a linear relationship between the outputs was not suitable. As a result, a further extension to this work could be the consideration of more complex relationships between outputs. Following the multivariate random forest approach, the univariate Gaussian process approach was extended to account for correlated multivariate outputs. The multi-output Gaussian processes performed well in comparison to the other approaches for some of the sites, but not all. Across the four sites for each of the outputs, there was no method that performed consistently better than the rest for prediction. This is potentially a result of each site having individual characteristics, which were identified in the sensitivity analysis, and also the investigation of a single emulator to be used for all sites.

Further analysis of additional sites could provide extra information for both the sensitivity analysis and emulation methods that were investigated. In addition, the emulators created for each of the sites could be used to explore the uncertainty within NewDEPOMOD without the computational cost of running NewDEPOMOD for multiple input settings. The sensitivity analysis methods described in this thesis could be applied to any setting with univariate or multivariate outputs, such as maps.

In this thesis, different statistical frameworks have been considered and contrasted, for the sensitivity analysis and emulation of process-based models. Novel approaches have been developed for dealing with maps as model outputs, as well as the novel application of statistical emulation of the environmental impacts of aquaculture.

Bibliography

Abraham, C., Cornillon, P., Matzner-Lober, E. & Molinari, N. (2003), 'Unsupervised curve clustering using b-splines', *Scandanavian Journal of Statistics* **30**, 581–595.

Alvarez, M. & Lawrence, N. (2009), Sparse convolved gaussian processes for multi-output regression, *in* 'Advances in Neural Information Processing Systems 21', pp. 57–64.

Alvarez, M. & Lawrence, N. (2011), 'Computationally efficient convolved multiple output gaussian processes', *Journal of Machine Learning Research* **12**, 1459–1500.

Alvarez, M., Luengo, D., Titsias, M. & Lawrence, N. (2010), Efficient multioutput gaussian processes through variational inducing kernels, *in* 'Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)'.

Alvarez, M., Rosasco, L. & Lawrence, N. (2012), 'Kernels for vector-valued functions', *Foundations and Trends in Machine Learning* **4**(3), 195–266.

Antoniadis, A., Lambert-Lacroix, S. & Poggi, J. (2021), 'Random forests for global sensitivity analysis: A selective review', *Reliability Engineering & System Safety* **206**, 107312.

Asche, F. & Bjorndal, T. (1996), *The Economics of Salmon Aquaculture*, second edn, John Wiley & Sons Ltd.

Ba, S. (2015), *SLHD: Maximin-Distance (Sliced) Latin Hypercube Designs*. R package version 2.1-1.
URL: *https://CRAN.R-project.org/package=SLHD*

Ba, S., Myers, W. & Brenneman, W. (2015), 'Optimal sliced latin hypercube designs', *Technometrics* **57**, 479–487.

Bannister, R., Johnsen, I., Hansen, P., Kutti, T. & Asplin, L. (2016), 'Near- and far-field dispersal modelling of organic waste from atlantic salmon aquaculture in fjord systems', *ICES Journal of Marine Science* **73**, 2408 – 2419.

Bastos, L. & O'Hagan, A. (2009), 'Diagnostics for gaussian process emulators', *Technometrics* **51**, 425–438.

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J. & Walsh, D. (2007), 'Computer model validation with functional output', *The Annals of Statistics* **35**, 1874–1906.

Bayarri, M., Berger, J., Kennedy, M., Kottas, A., Paulo, R., Sacks, J., Cafeo, J., Lin, C. & Tu, J. (2005), Bayesian validation of a computer model for vehicle crashworthiness, Technical Report 163, National Institute of Statistical Sciences.

Bidot, C., Lamboni, M. & Monod, H. (2018), *multisensi: Multivariate Sensitivity Analysis*.
URL: *https://CRAN.R-project.org/package=multisensi*

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **26**, 123–140.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and regression trees*, Chapman & Hall.

Browne, C., Matteson, D., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J. & Barrett, C. (2021), 'Multivariate random forest prediction of poverty and malnutrition prevalence', *PLOS ONE* **16**, 1–23.

Broyden, C. (1970), 'The convergence of a class of double-rank minimization algorithms', *Journal of the Institute of Mathematics and Its Applications* **6**, 76–90.

Campbell, B. & Pauly, D. (2013), 'Mariculture: A global analysis of production trends since 1950', *Marine Policy Magazine* **39**, 94 – 100.

Candela, J. (2005), 'A unifying view of sparse approximate gaussian process regression', *Journal of Machine Learning Research* **6**, 1939–1959.

Canon Jones, H., Noble, C., Damsgard, B. & Pearce, G. (2011), 'Social network analysis of behavioural interactions that influence the development of fin damage in atlantic salmon parr (salmo salar) held at different stocking densities', *Applied Animal Behaviour Science* **133**, 117–126.

Chai, T. & Draxler, R. (2014), 'Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature', *Geoscientific Model Development* **7**(3), 1247–1250.

Chen, Y., Beveridge, M., Telfer, T. & Roy, W. (2003), 'Nutrient leaching and settling rate characteristics of the faeces of atlantic salmon (salmo salar l.) and the implications for modelling of solid waste dispersion', *Journal of Applied Ichthyology* **19**, 114 – 117.

Conti, S., Gosling, J., Oakley, J. & O'Hagan, A. (2009), 'Gaussian process emulation of dynamic computer codes', *Biometrika* **96**, 663–676.

Conti, S. & O'Hagan, A. (2010), 'Bayesian emulation of complex multi-output and dynamic computer models', *Journal of Statistical Planning and Inference* **140**, 640–651.

Cromey, C., Black, K., Edwards, A. & Jack, I. (1998), 'Modelling the deposition and biological effects of organic carbon from marine sewage discharges', *Estuarine, Coastal and Shelf Science* **47**, 295 – 308.

Cromey, C., Nickell, T. & Black, K. (2002), 'Depomod - modelling the deposition and biological effects of waste solids from marine cage farms', *Aquaculture* **214**, 211 – 239.

Cukier, R., Fortuin, C., Schuler, K., Petschek, A. & Schaibly, J. (1973), 'Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. *i* theory', *The Journal of Chemical Physics* **59**, 3873–3878.

Dandekar, R., Cohen, M. & Kirkendall, N. (2001), 'Applicability of latin hypercube sampling techniques to create multivariate synthetic microdata'.

Dannenmaier, J., Kaltenbach, C., Kolle, T. & Krischak, G. (2020), 'Application of functional data analysis to explore movements: walking, running, and jumping - a systematic review', *Gait & Posture* **77**, 182–189.

Davidon, W. (1991), 'Variable metric method for minimization', *SIAM Journal on Optimization* **1**(1), 1–17.

Davies, V., Noè, U., Lazarus, A., Gao, H., Macdonald, B., Berry, C., Luo, X. & Husmeier, D. (2019), 'Fast parameter inference in a biomechanical model of the left ventricle by using statistical emulation', *Journal of The Royal Statistical Society: Series C (Applied Statistics)* **68**(5), 1555–1576.

De'ath, G. (2002), 'Multivariate regression trees: a new technique for modeling species–environment relationships', *Ecology* **83**(4), 1105–1117.

Dryden, I. & Mardia, K. (2016), *Statistical shape analysis: with applications in R*, 2 edn, John Wiley & Sons Ltd.

Fletcher, R. (1970), 'A new approach to variable metric algorithms', *Computer Journal* **13**(3), 317–322.

Fricker, T., Oakley, J. & Urban, N. (2013), 'Multivariate gaussian process emulators with nonseparable covariance structures', *Technometrics* **55**, 47–56.

Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., Ye, A., Miao, C. & Di, Z. (2014), 'A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model', *Environmental Modelling & Software* **51**, 269–285.

Gillibrand, P. & Turrell, W. (1997), 'Simulating the dispersion and settling of particulate material and associated substances from salmon farms'.

Goldfarb, D. (1970), 'A family of variable metric updates derived by variational means', *Mathematics of Computation* **24**(109), 23–26.

Gong, M., Miller, C. & Scott, M. (2015), 'Functional pca for remotely sensed lake surface water temperature data', *Procedia Environmental Sciences* **26**, 127–130.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University Press.

Gramacy, R. & Apley, D. (2015), 'Local gaussian process approximation for large computer experiments', *Journal of Computational and Graphical Statistics* **24**, 561–578.

Grow, A. (2016), *Regression Metamodels for Sensitivity Analysis in Agent-Based Computational Demography*, Vol. 41, Springer, chapter 7, pp. 185–210.

Grow, A. & Hilton, J. (2018), *Statistical Emulation*, John Wiley & Sons Ltd, pp. 1–8.

Harper, E., Stella, J. & Fremier, A. (2011), 'Global sensitivity analysis for complex ecological models: A case study of riparian cottonwood population dynamics', *Ecological applications : a publication of the Ecological Society of America* **21**, 1225–1240.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 edn, Springer.

Herr, D. (1986), 'On the history of anova in unbalanced, factorial designs: The first 30 years', *The American Statistician* **40**, 265–270.

Higdon, D. (2002), Space and space-time modeling using process convolutions, *in* 'Quantitative Methods for Current Environmental Issues'.

Higdon, D., Gattiker, J., Williams, B. & Rightley, M. (2008), 'Computer model calibration using high-dimensional output', *Journal of the American Statistical Association* **103**(482), 570–583.

Homma, T. & Saltelli, A. (1996), 'Importance measures in global sensitivity analysis of model output', *Reliability Engineering and System Safety* **52**, 1–17.

Huber, P. (1981), *Robust Statistics*, John Wiley & Sons.

Iannace, G., Ciaburro, G. & Trematerra, A. (2019), 'Wind turbine noise prediction using random forest regression', *Machines* **7**(4), 69.

Iman, R. & Conover, W. (1982), 'A distribution-free approach to inducing rank correlation among input variables', *Communications in Statistics - Simulation and Computation* **11**, 311 – 334.

Ishwaran, H., Tang, F., Lu, M. & Kogalur, U. (2021), 'randomForestSRC: multivariate splitting rule vignette'.
URL: *https://luminwin.github.io/randomForestSRC/articles/mvsplit.html*

Jalal, H., Dowd, B., Sainfort, F. & Kuntz, K. (2013), 'Linear regression metamodeling as a tool to summarize and present simulation model results', *Medical Decision Making* **33**, 880–890.

Jansen, M., Rossing, W. & Daamenm, R. (1994), *Monte Carlo Estimation of Uncertainty Contributions from Several Independent Multivariate Sources*, Springer, pp. 334–343.

Jin, R., W., C. & Sudjianto, A. (2005), 'An efficient algorithm for constructing optimal design of computer experiments', *Journal of Statistical Planning and Inference* **134**, 268–287.

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

Journel, A. & Huijbregts, C. (1978), *Mining Geostatistics*, Academic Press.

Keeley, N., Cromey, C., Goodwin, E., Gibbs, M. & Macleod, C. (2013), 'Predictive depositional modelling (depomod) of the interactive effect of current flow and resuspension on ecological impacts beneath salmon farms', *Aquaculture Environment Interactions* **3**, 275 – 291.

Kennedy, M., Anderson, C., Conti, S. & O'Hagan, A. (2006), 'Case studies in gaussian process modelling of computer codes', *Reliability Engineering & System Safety* **91**, 1301–1309.

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

Kleijnen, J. (1979), 'Regression metamodels for generalizing simulation results', *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 93–96.

Koenker, R. & Machado, J. (1999), 'Goodness of fit and related inference processes for quantile regression', *Journal of the American Statistical Association* **94**(448), 1296–1310.

Lamboni, M., Monod, H. & Makowski, D. (2011), 'Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models', *Reliability Engineering & System Safety* **96**, 450–459.

Langsrud, O. (2003), 'Anova for unbalanced data: Use type ii instead of type iii sums of squares', *Statistics and Computing* **13**(2), 163–167.

Lawrence, N., Seeger, M. & Herbrich, R. (2003), Fast sparse gaussian process methods: The informative vector machine, *in* 'Proceedings of the 16th annual conference on neural information processing systems', number CONF, pp. 609–616.

Liaw, A. & Wiener, M. (2002), 'Classification and regression by randomforest', *R News* **2**, 18–22.

Liu, D. & Nocedal, J. (1989), 'On the limited memory bfgs method for large scale optimization', *Mathematical Programming* **45**, 503–528.

Liu, F., Bayarri, M. & Berger, J. (2009), 'Modularization in bayesian analysis, with emphasis on analysis of computer models', *Bayesian Analysis* **4**, 119–150.

MacKay, D. (1992), 'Bayesian interpolation', *Neural Computation* **4**, 415–447.

Madu, C. (1990), 'Simulation in manufacturing: A regression metamodel approach', *Computers & Industrial Engineering* **18**, 381–389.

Mahalanobis, P. (1936), On the generalized distance in statistics, National Institute of Science of India.

Matthews, A., Hensman, J., Turner, R. & Ghahramani, Z. (2016), On sparse variational methods and the kullback-leibler divergence between stochastic processes, *in* 'Artificial Intelligence and Statistics', pp. 231–239.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman & Hall/CRC.

McFarland, J., Mahadevan, S., Romero, V. & Swiler, L. (2008), 'Calibration and uncertainty analysis for computer simulations with multivariate output', *AIAA Journal* **46**, 1253–1265.

McKay, M., Beckman, R. & Conover, W. (1979), 'A comparison of three methods for selecting values of input variables in the analysis of output from a computer code', *Technometrics* **21**, 239–245.

Meinhausen, N. (2006), 'Quantile regression forests', *Journal of Machine Learning Research* **7**, 983–999.

Metropolis, N. & Ulam, S. (1949), 'The monte carlo method', *Journal of the American Statistical Association* **44**(247), 335–341.

Miller, K., Huettmann, F., Norcross, B. & Lorenz, M. (2014), 'Multivariate random forest models of estuarine-associated fish and invertebrate communities', *Marine Ecology Progress Series* **500**, 159–174.

Mitchener, H. & Torfs, H. (1996), 'Erosion of mud/sand mixtures', *Coastal Engineering* **29**, 1 – 25.

Mlaker, M., Tušar, T. & Filipic, B. (2019), 'Comparing random forest and gaussian process modelling in the gp-demo algorithm', *Information Security Education Journal (ISEJ)* **6**(9), 06.

Morris, M. & Mitchell, T. (1995), 'Exploratory designs for computational experiments', *Journal of Statistical Planning and Inference* **43**, 381–402.

Neal, R. (1996), *Bayesian Learning for Neural Networks*, Vol. 118 of *Lecture Notes in Statistics*, Springer.

Nelder, J. (1977), 'A reformulation of linear models (with discussion)', *Journal of the Royal Statistical Society Series A* **140**, 48–77.

Nelder, J. (1994), 'The statistics of linear models: back to basics', *Statistics and Computing* **4**, 221–234.

Noè, U., Lazarus, A., Gao, H., Davies, V., Macdonald, B., Mangion, K., Berry, C., Luo, X. & Husmeier, D. (2019), 'Gaussian process emulation to accelerate parameter estimation in a mechanical model of the left ventricle: a critical step towards clinical end-user relevance', *Journal of The Royal Society Interface* **16**.

Nocedal, J. (1980), 'Updating quasi-newton matrices with limited storage', *Mathematics of Computation* **35**(151), 773–782.

Nocedal, J. & Wright, S. (2006), *Numerical Optimization*, Springer Science & Business Media.

O'Hagan, A. (2010), 'Bayesian analysis of computer code outputs: A tutorial', *Reliability Engineering and System Safety* **91**, 1290–1300.

Overstall, A. & Woods, D. (2016), 'Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**, 483–505.

Owen, A. (1992), 'Orthogonal arrays for computer experiments, integration and visualization', *Statistica Sinica* **2**(2), 439–452.

Parker, K., Ruggiero, P., Serafin, K. & Hill, D. (2019), 'Emulation as an approach for rapid estuarine modeling', *Coastal Engineering* **150**, 79–93.

Pearson, K. (1901), 'On lines and planes of closest fit to systems of points in space', *Philosophical Magazine* **2**, 559–572.

Pearson, K. (1926), 'On the coefficient of racial likeness', *Biometrika* **18**, 105–117.

Penrose, R. (1955), A generalized inverse for matrices, *in* 'Mathematical proceedings of the Cambridge philosophical society', Vol. 51, Cambridge University Press, pp. 406–413.

Pianosi, F., Beven, K., Freer, J., Hall, J., Rougier, J., Stephenson, D. & Wagener, T. (2016), 'Sensitivity analysis of environmental models: A systematic review with practical workflow', *Environmental Modelling and Software* **79**, 214 – 232.

Preece, D. (1983'), Latin squares, latin cubes, latin rectangles, etc., *in* 'Encyclopedia of Statistical Sciences', Vol. 4.

Qian, P. & Wu, J. (2009), 'Sliced space-filling designs', *Biometrika* **96**, 945–956.

Quinonero-Candela, J., Rasmussen, C. & Williams, C. (2007), Approximation methods for gaussian process regression, *in* 'Large-scale kernel machines', MIT Press, pp. 203–223.

Rajabi, M. & Ketabchi, H. (2017), 'Uncertainty-based simulation-optimization using gaussian process emulation: Application to coastal groundwater management', *Journal of Hydrology* **555**, 518–534.

Ramsay, J. & Dalzell, C. (1991), 'Some tools for functional data analysis', *Journal of the Royal Statistical Society: Series B (Methodological)* **53**(3), 539–561.

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis*, second edn, Springer.

Rasmussen, C. & Williams, C. (2006), *Gaussian Processes for Machine Learning*, MIT press.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. & Aigrain, S. (2013), 'Gaussian processes for time-series modelling', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**(1984), 20110550.

Rougier, J. (2008), 'Efficient emulators for multivariate deterministic functions', *Journal of Computational and Graphical Statistics* **17**, 827–843.

Sacks, J., Welch, W., Mitchell, T. & Wynn, H. (1989), 'Design and analysis of computer experiments', *Statistical Science* **4**, 409–423.

Saltelli, A., Chan, K. & Scott, M. (2000), *Sensitivity Analysis*, John Wiley & Sons Ltd.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. (2008), *Global Sensitivity Analysis. The Primer*, John Wiley & Sons Ltd.

Saltelli, A., Tarantola, S., Campolongo, F. & Ratto, M. (2004), *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models.*, John Wiley & Sons Ltd.

Salter, J. & Williamson, D. (2016), 'A comparison of statistical emulation methodologies for multi-wave calibration of environmental models', *Environmetrics* **27**(8), 507–523.

Seber, G. (2007), *Special Products and Operators*, John Wiley & Sons Ltd, chapter 11, pp. 233–255.

Segal, M. (2004), 'Machine learning benchmarks and random forest regression'.

Segal, M. & Xiao, Y. (2011), 'Multivariate random forests', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**, 80–87.

Serban, N. & Wasserman, L. (2005), 'Cats: Clustering after transformation and smoothing', *Journal of the American Statistical Association* **100**, 990–999.

Shabani, S., Samadianfard, S., Sattari, M., Mosavi, A., Shamshirband, S., Kmet, T. & Várkonyi-Kóczy, A. (2020), 'Modelling pan evaporation using gaussian process regression, k-nearest neighbours, random forest and support vector machines; comparative analysis', *Atmosphere* **11**(1), 66.

Shanno, D. (1970), 'Conditioning of quasi-newton methods for function minimization', *Mathematics of Computation* **24**(111), 647–656.

Silverman, B. (1985), 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **47**, 1–52.

Smola, A. & Bartlett, P. (2001), Sparse greedy gaussian process regression, pp. 619–625.

Smola, A. & Schölkopf, B. (2000), Sparse greedy matrix approximation for machine learning, *in* 'Proceedings of the Seventeenth International Conference on Machine Learning', Morgan Kaufmann Publishers Inc.

Snelson, E. & Ghahramani, Z. (2006), Sparse gaussian processes using pseudo-inputs, *in* 'Advances in neural information processing systems', pp. 1257–1264.

Sobol', I. (1993), 'Sensitivity analysis for non-linear mathematical models', *Mathematical Modelling and Computational Experiment I* pp. 407–414.

Speed, F., Hocking, R. & Hackney, O. (1978), 'Methods of analysis of linear models with unbalanced data', *Journal of the American Statistical Association* **73**(361), 105–112.

Sundberg, L., Ketola, T., Laanto, E., Kinnula, H., Bamford, J., Penttinen, R. & Mappes, J. (2016), 'Intensive aquaculture selects for increased virulence and interference competition in bacteria', *Proceedings of the Royal Society B: Biological Sciences* **283**(1826), 20153069.

Taranger, G., Karlsen, ., Bannister, R., Glover, K., Husa, V., Karlsbakk, E., Kvamme, B., Boxaspen, K., Bjørn, P., Finstad, B., Madhun, A., Morton, H. & Svåsand, T. (2015), 'Risk assessment of the environmental impact of norwegian atlantic salmon farming', *ICES Journal of Marine Science* **72**(3), 997–1021.

Teh, Y., Seegar, M. & Jordan, M. (2005), Semiparametric latent factor models, *in* 'International Workshop on Artificial Intelligence and Statistics', pp. 333–340.

Titsias, M. (2009), Variational learning of inducing variables in sparse gaussian processes, *in* 'Artificial Intelligence and Statistics', pp. 567–574.

Turnbull, J., Bell, A., Adams, C., Bron, J. & Huntingford, F. (2005), 'Stocking density and welfare of cage farmed atlantic salmon: application of a multivariate analysis', *Aquaculture* **243**, 121–132.

Ulapane, K., Thiyagarajan, K. & Kodagoda, S. (2020), Hyper-parameter initialization for squared-exponential kernel-based gaussian process regression, *in* '2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)', pp. 1154–1159.

van der Wilk, M., Dutordoir, V., John, S., Artemev, A., Adam, V. & Hensman, J. (2020), 'A framework for interdomain and multioutput gaussian processes', *arXiv:2003.01115* .

Wahba, G. (1990), *Spline models for observational data*, Society for industrial and applied mathematics.

Willmott, C. & Matsuura, K. (2005), 'Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance', *Clim. Res.* **30**, 79–82.

Willoughby, S. (1999), *Manual of Salmonid Farming*, Fishing News Books.

Wood, S. (2017), *Generalized Additive Models: An Introduction with R*, 2 edn, Chapman & Hall/CRC.

Woodbury, M. (1950), *Inverting Modified Matrices*, Memorandum Report / Statistical Research Group, Princeton, Statistical Research Group.

Xiao, L. (2012), Topics in Bivariate Spline Smoothing, PhD thesis, Cornell University.

Xiao, L., Li, Y. & Ruppert, D. (2013), 'Fast bivariate p-splines: the sandwich smoother', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **75**(3), 577–599.

Yates, F. (1934), 'The analysis of multiple classifications with unequal numbers in the different classes', *Journal of the American Statistical Association* **29**, 51–66.

Zahedi, P., Parvandeh, S., Asgharpour, A., McLaury, B., Shirazi, S. & McKinney, B. (2018), 'Random forest regression prediction of solid particle erosion in elbows', *Powder Technology* **338**, 983–992.

Zhang, H., Zimmerman, J., Nettleton, D. & Nordman, D. (2019), 'Random forest prediction intervals', *The American Statistician* .