NTNU
Kunnskap for en bedre verden

Department of Mathematics

TMA4500 - Industrial Mathematics - Project Thesis

Bayesian inference on land-based salmon farming data using latent Gaussian models

Author: Sivert Laukli
Supervisor: Professor Ingelin Steinsland

December, 2022

Abstract

In this study, we used Bayesian inference on latent Gaussian models to analyze data collected from a single tank at Salmon Evolution's production facility. The data, along with expert assumptions, allowed us to investigate models based on oxygen and carbon dioxide concentrations in the tank. Oxygen consumption and carbon dioxide production were assumed smooth and modelled as random effects using random walk 1 (RW1) priors. The process were evaluated by fitting five models with gradually more prior information. The models were evaluated on how they fitted appropriate to the data and the model assumptions. If several models fitted equally well to the data, the deviance information criteria was used to evaluate their model performance. We detected that strict priors on the Gaussian effect parameters were necessary for the model to fit the data accurately, and introducing strict penalized complexity (PC) priors for the variance of the random walk 1 priors for the random effects improved the accuracy of the accumulated oxygen and carbon dioxide further. using the delta method were used to compute prior variance to the parameters of the linear effects in addition to the PC priors for the variance of the RW1 priors of the random effects appeared to perform the best in terms of the the deviance information criteria. Our analysis revealed that the salmon in the tank appeared to be consuming more oxygen than it was producing in carbon dioxide, which could be explained by effects not taken into consideration in the model assumptions. To improve the model, we suggest evaluating sensor performances, obtaining more data from the facility, and fitting the model on different similar datasets. However, more sensor data could also introduce more uncertainties to the model. Overall, the results of this study provide valuable insights into the oxygen and carbon dioxide dynamics in a salmon tank and highlight the importance of considering all sources of uncertainty in modeling efforts.

Sammendrag

Vi har i denne oppgaven brukt Bayesiansk inferens på latente Gaussiske modeller for å analysere data samlet fra en enkelt tank ved produksjonsanlegget til Salmon Evolution. Data har blitt brukt sammen med kunnskap fra eksperter til å undersøke modeller basert på konsentrasjoner av oksygen og karbondioksid i tanken. Forbruk av oksygen og produksjon av karbondioksid ble antatt glatt og modellert som en *random effect* med en *random walk 1* (RW1) prior. Prosessen ble evaluert ved å tilpasse fem modeller som gradvis ble tilegnet mer prior-informasjon, og vurdert ut ifra hvilke modeller som passet best til dataene og antagelsene. Dersom noen modeller presterte tilsynelatende like bra ble de vurdert ut ifra *the deviance information criteria* (DIC). Resultatene viste at strenge prior-fordelinger på parameterene til de lineære effektene var nødvendig for at modellen skulle passe med dataene. Videre oppdaget vi at ved å velge *penalized complexity* priorer for variansen til RW1 priorene ble posteriori estimatene til forbruket av oksygen og produksjonen av karbondioksid mer presis. Delta-metoden ble brukt til å beregne variansen til priorfordelingen til de Gaussisk fordelte koeffisientene til de lineære effektene. Dette viste seg å gi det beste resultatet med hensyn på DIC. Vi oppdaget videre at laksen konsumerte mer oksygen enn hva man kunne forvente ut ifra produsert karbondioksid. Dette kan forklares av effekter som ikke ble tatt med til vurdering i antagelsene gjort på forhånd. For å forbedre modellen ble det foreslått å vurdere og kalibrere sensorene brukt til å samle data videre, å samle mer data fra anlegget og å teste modellen på flere lignende datasett. Det er imidlertid verdt å mere seg at mer sensordata sannsynligvis vil øke usikkerheten i modellen. Totalt sett gir resultatene av oppgaven verdifull innsikt i dynamikken mellom oksygen og karbondioksid i en tank med laks, samt viktigheten av å vurdere feilkilder i modelleringen.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Ingelin Steinsland for her valuable guidance and support throughout this project. I would also like to thank the field experts at Salmon Evolution, Jon Åge Stakvik-Løvås and Sondre Kvalsvik, for their valuable insights and expertise. My sincere thanks also go to Assoc. Prof. Bjarne Grimstad, my supervisor at Solution Seeker, for his support and guidance. I would also like to extend my appreciation to Professor of Chemical Engineering Sigurd Skogestad for his support with literature and initial inspiration. Finally, I would like to thank the team at Salmon Evolution for providing me with the opportunity to work on this project. I am deeply grateful to all of these individuals for their invaluable contributions to my procect and for their help in making this project a success. I also would like to thank everyone else who has supported me during the project and writing of this work.

Contents

1 Introduction **1**

2 Case study of a land-based salmon farming facility **4**
2.1 Salmon Evolutions hybrid flow-through system 5
2.2 Sensor data from a land-based salmon farming facility 6
2.3 Exploratory data analysis 7
2.4 Assumptions 9
2.5 Parameters of interest 11

3 Bayesian inference using latent Gaussian models **13**
3.1 An introduction to Bayesian inference 13
3.2 Latent Gaussian models 16
3.3 Bayesian inference on latent Gaussian models 17
3.4 Accounting for model discrepancy as a temporal random effect 20
3.5 Prior distributions 20
3.5.1 Vague prior distributions 20
3.5.2 The delta method 20
3.5.3 Penalized complexity priors 21
3.6 The deviance information criterion 21
3.7 Key aspects from the R-INLA package 22

4 Modelling substance accumulations as latent Gaussian models **23**
4.1 Expressing the physical case as latent Gaussian models 23
4.1.1 Oxygen concentration as a latent Gaussian model 23
4.1.2 Carbon dioxide concentration as a latent Gaussian model 24
4.2 Model fitting by assigning prior distributions 25
4.2.1 Initially: Linear regression model 25
4.2.2 Model 0: Flat priors 26
4.2.3 Model 1: Gaussian priors for intercepts and coefficients of linear effects . . 26
4.2.4 Model 2: Adding PC priors to the hyperparameters 26
4.2.5 Model 3: Gaussian priors using the delta method 26
4.3 Bayesian inference on latent Gaussian model using integrated nested Laplace approsimation 27

5 Results **27**
5.1 Posterior Gaussian parameters 28
5.1.1 Posterior parameters for the linear models with vague priors 28
5.1.2 Posterior parameters for the random effect models with vague priors 29
5.1.3 Posterior parameters with assigned Gaussian and PC priors 31
5.2 Posterior discrepancy functions 34
5.2.1 Posterior discrepancy function for random effect models with vague priors . 35
5.2.2 Posterior discrepancy functions for models assigned priors 37
5.3 Posterior relevant hyperparameters 40
5.4 Evaluation criteria 42

6 Model assessments and physical interpretations **43**
6.1 Choice of models 43
6.2 Model errors 44
6.3 Model approach improvement 46

7 Conclusion **47**

References **48**

1 Introduction

As the global demand for food continues to rise, the United Nations Food and Agriculture Organization (FAO) has in FAO (2020) identified aquaculture as a potential sustainable source of protein. However, the practice of aquaculture in the sea has scaling limitations as increased production can have negative impacts on local ecosystems, including the treatment of parasites and diseases, feed spill, and the escape of salmon from production sites, as noted by Taranger *et al.* (2015). Additionally, data mining in the aquaculture industry has been challenging due to the often difficult or remote locations of production sites. One alternative to traditional aquaculture in the sea is the practice of salmon farming on land. This approach eliminates many of the negative impacts associated with sea-based aquaculture and offers new opportunities for data-driven production. With several traditional methods and prior information in hand, a Bayesian approach can be used to effectively exploit data in the land-based salmon farming industry. Bayesian methods allow for the incorporation of prior information and expert opinions into the modeling process, enabling a more informed and accurate model.

Land-based aquaculture and fish-farming has traditionally been used primarily for the hatchery and smolt production stages of fish cultivation. Once the fish reach a certain size, they are typically transferred to cages near the shore to continue growing until they are ready for harvest. However, this approach has several challenges, including the potential for the spread of diseases and parasites, the escape of farmed fish and the following negative impact on wild salmon populations, as discussed by Sætre and Østli (2021). These issues highlight the need for more sustainable and responsible approaches fish-farming.

In recent years, the concept of using land-based facilities for the grow-out production of fish, keeping the fish on land until they reach the desired harvest weight, has gained increasing attention and support. The return in terms of sustainability, minimized environmental impacts and economic growth potential has made governments and investors support pioneers in the industry in the development of such production facilities. Further, building a new production site is a greenfield process, of which a brief definition is given at ProductionPlan (2022). Greenfield processes are characterized by the lack of established industry standards for the construction of these facilities, allowing producers to customize the installation of sensors to gather data that would not be possible in sea-based production. The ability to install sensors in this way enables the continuous monitoring of the entire production process to a greater extent than is possible in sea-based facilities, as well as the collection of data that can be used for analysis and decision-making in future production. Land-based production facilities offer a promising alternative to traditional sea-based aquaculture, with the potential to mitigate many of the negative impacts and challenges associated with this approach.

There are several additional advantages to land-based salmon farming. One key benefit is the increased biosecurity, as these systems can be isolated from the sea and divided into distinct biological departments within the production facility. This reduces the risk of disease transmission. Control over environmental conditions as pH, oxygen saturations and water temperature can improve health and growth. Land-based systems also allow for more efficient resource use, including the reuse of water and the monitoring of feed spill to minimize waste. The isolation of the production facilities from the sea also eliminates the risk of escape and reduces the impact on local ecosystems. Finally, the ability to monitor and control the entire production process offers the opportunity for technological development and innovation, which may be more limited in sea-based production.

Historically, the use of data analysis in the salmon farming industry has been limited, and many decisions in the industry have been based on personal experience and interactions between experienced individuals rather than on robust data analysis. Additionally, much of the research and technology development in the industry has been focused on addressing challenges in sea-based production facilities, such as problems with parasites described by Sætre and Østli (2021). In contrast, land-based production facilities do not face many of these challenges, and the main focus is on optimizing water quality parameters and feeding regimes for optimized growth. However, this also introduces new technical challenges and opportunities related to data analysis and decision-

making. Land-based production presents an opportunity to develop and apply new approaches to data-driven decision-making in the salmon farming industry.

There are several technologies that have been developed to support the safe and sustainable production of marine species on land, including recirculating aquaculture systems (RAS), entire flow-through systems (FTS), and hybrid flow-through systems (HFS). RAS, as described by Davison (2018), recycle up to 95% of their circulating water using a biological filter. FTS, as implemented by Andfjord Salmon (2022) and others, use fresh seawater that flows continuously through the production tanks and back into the ocean. HFS, used by Salmon Evolution (SE) and described by Salmon Evolution (2022), combines elements of both RAS and FTS. These three technologies dominate the market for land-based salmon production in Norway today and offer different approaches to meeting the challenges and opportunities of this industry.

Salmon Evolution (SE) is in the early stages of establishing a steady-state, industrial production facility at their site in Hustadvika, Norway. By December 2022, a significant portion of their facility is in active production, and they successfully completed their first harvest in November 2022, as described at Salmon Evolution (2022). SE has also installed a number of sensors in their facility to collect data on feeding and water quality parameters, explained by Stakvik-Løvås (2022). This data will be critical in optimizing the production process and supporting data-driven decision-making in their land-based salmon farming operation.

The facility is planned to consist of 48 tanks divided into four departments, each with 12 tanks, which are currently viewed as two production lines, explained by Lyngstad (2022). The purpose of the production lines is to split the biomass of the salmon into smaller groups as they grow, in order to maintain suitable water quality throughout the production process. In this thesis, we focus on a single operational tank containing salmon, referred to as "the tank," "fish tank," or "tank."

For the purposes of this thesis, SE has provided data on various water quality parameters. The data consists of real-time time series data measuring oxygen added from two injectors, oxygen saturation from two sensors, carbon dioxide concentration, water flow of fresh seawater and recycled water, and temperature from two sensors. This data will be used to understand the dynamics of the water quality parameters in the tank and to inform decision-making in the land-based salmon farming operation.

The HFS technology applied by SE, explained by Lyngstad (2022) and at Salmon Evolution (2022), allows the producer to adjust certain parameters in order to optimize growth while maintaining appropriate threshold values for water quality quantities. The adjustable parameters include water flow, temperature, oxygen concentration, and feeding regimes. These parameters are controlled using water pumps, heat pumps, and oxygen injectors, and there are some natural limitations to the effects of these adjustments due to the capabilities of the machinery. In addition to these technical constraints, there are also biological thresholds that must be considered in land-based salmon production. These thresholds are determined in consultation with marine veterinarians, who monitor the health of the fish throughout the production process by taking random salmon samples. Technical operators ensures that the carbon dioxide concentration, ammonium concentration, turbidity, and biomass density remain below certain levels (15 mg/l, 4 mg/l, 27 mg/l, and 85 kg/m3, respectively, given by Lyngstad (2022)). These biological thresholds take into account the inherent biological variability of the fish, including genetic factors that can affect their growth and health.

Further interactions with Løvås gives insights in the consequences of adjusting parameters. Increased water flow decreases substance concentrations, and has an upper bound decided by the water pumps. Increased temperature increases the growth of the salmon, which in turn increases the concentration of critical substances. Oxygen injection is adjusted to keep the saturation approximately at *100%*, and feeding intensity is adjusted after measured feed spill through the system.

Stakvik-Løvås (2022) explains further the impacts of adjusting various parameters on the production process. Increasing water flow can decrease substance concentrations, but is limited by the capabilities of the water pumps. Increasing temperature can improve the growth of the salmon, but also increases the concentration of critical substances. Oxygen injection is adjusted to maintain an approximate saturation level of 100%, and feeding intensity is adjusted based on measured feed

spill through the system.

Using the data provided, it is possible to explore the impact of adjusting various quantities on the production process and to evaluate existing assumptions. For example, dos Santos *et al.* (2022) propose the following equation:

$$1\text{kg feed} + 0.45\text{kg } O_2 \longrightarrow 0.9\text{kg biomass} + 0.48\text{kg } CO_2 + 0.047\text{kg } NH_3.$$

By analyzing the provided data, it may be possible to examine the effect of increasing oxygen (a growth-stimulating parameter) on the concentration of carbon dioxide (a limiting parameter). This can provide valuable insights into the processes involved in land-based salmon production and inform future production planning and decision-making, as well as help to better understand surveillance data from the production facility.

The following introduction of Bayesian models is based on the work of Blangiardio and Cameletti (2015), Moraga (2019) and Gómez-Rubio (2020). Bayesian modelling is a framework for statistical models that allow for updates of the probability of a conditional events as more evidence or information becomes available. In Bayesian modeling, we begin with a prior belief about the likelihood of different events and update this belief using data and statistical inference. One of the key advantages of Bayesian modeling is that it allows for the incorporation of prior knowledge and uncertainty into the model. For example, if we have some prior information about the range of values that a parameter is likely to take, we can encode this information in the form of a prior distribution, which the model can use to guide the inference process and produce more accurate estimates. Bayesian models also have a natural way of expressing uncertainty through the use of probability distributions, which can be useful for decision making as we can use the model to compute the probability of different outcomes and choose the course of action that has the highest expected value. Overall, Bayesian models are a powerful tool for data analysis and decision making, and are widely used in fields such as machine learning by Theodoris (2015), health economics by Baio (2012), and psychology by Bivand *et al.* (2014).

This introduction to latent Gaussian models is based on the work of Johnson and Wichern (2007) and Gelman *et al.* (2013). Latent Gaussian models are a class of statistical models used to analyze data that is generated by an underlying process that cannot be directly observed. Instead, we can only observe a set of noisy or incomplete measurements that are related to the underlying process through a set of known or unknown relationships. In a latent Gaussian model, the underlying process is assumed to be a Gaussian process, which means it is characterized by a mean and a covariance matrix. These mean and covariance matrix are typically treated as latent or hidden variables that are not directly observed, but can be estimated from the data. One of the key benefits of latent Gaussian models is their ability to explicitly model the uncertainty and noise in the observed data, which can be useful for tasks such as prediction and outlier detection, where we want to estimate the most likely values of the latent variables given the observed data as well as the uncertainty in these estimates. In a Bayesian framework, latent Gaussian models can be used to model data and make inferences about the latent variables and model parameters. In this context, LGMs can be considered a subclass of Bayesian models because they are based on Bayesian principles and involve the use of Bayesian inference to estimate the model parameters. However, it is important to note that LGMs can also be used in a non-Bayesian framework, in which case they would not be considered a subclass of Bayesian models in general.

There are several approaches to performing Bayesian inference on latent Gaussian models, including Markov chain Monte Carlo (MCMC) methods and variational methods. MCMC methods, explained by Casella and George (1992) and Gelfand and Smith (1990) and Betancourt (2018), involve sampling from the posterior distribution of the latent variables using a Markov chain, which is a sequence of random samples drawn from the distribution. By repeatedly sampling from the distribution, we can estimate the posterior distribution of the latent variables and compute various quantities of interest, such as the mean and variance of the latent variables. However, MCMC can quickly become computationally intensive for complex models or large datasets. Variational methods, described by Blei *et al.* (2017), involve approximating the posterior distribution with a simpler distribution that can be efficiently computed and optimized. These methods typically involve minimizing a divergence measure between the approximate posterior and the true posterior using optimization algorithms such as gradient descent. Overall, both MCMC and variational

methods can be useful approaches for performing Bayesian inference on latent Gaussian models, depending on the specific requirements and constraints of the problem at hand.

To address the computational demands of Bayesian inference, Rue *et al.* (2009) introduced a method for approximating the posterior distribution of latent variables in latent Gaussian models using the integrated nested Laplace approximation (INLA). This approach, described by Blangiardi and Cameletti (2015), is based on the Laplace approximation and is specified for LGMs in which the likelihood function is known up to a constant of proportionality. INLA allows for the estimation of the posterior distribution of latent variables as well as summary statistics such as means and variances, and has proven effective in handling large data sets and complex models. One key advantage of INLA is its computational efficiency, making it a useful tool for Bayesian inference in these contexts.

The inclusion of expert knowledge, prior assumptions, and observed data allows for the modeling of various aspects of production processes. In the present case, we aim to estimate the oxygen consumed and carbon dioxide produced by salmon, as well as the relationship between these two variables. We utilize two separate latent Gaussian models, one for each aspect of interest. By applying Bayesian methods to these latent Gaussian models, we are able to incorporate prior information obtained from expert knowledge and assumptions to fit the model to the data. This approach enables us to obtain posterior information about each of the targeted variables and their relationship, providing a more comprehensive understanding of the production process.

In addition to estimating the relevant variables and their relationships, we are also interested in evaluating the extent to which incorporating prior information can improve the accuracy of our models. By fitting these prior distributions to the data, we aim to assess the extent to which they can increase the model accuracy and fit to initial assumptions and the data. To investigate the impact of assigning prior distributions, we compare five models that vary in the amount of prior information incorporated. These models are similar for oxygen and carbon dioxide, and include a linear model with no specified priors and a model with a random intercept and a chosen random walk 1 (RW1) prior, which assumes a smooth accumulation of substances with a vague hyperprior of variance. As we gradually increase the amount of prior information assigned to the models, we first assign Gaussian priors to the linear effects and then add penalized complexity (PC) hyperpriors for the non-Gaussian hyperparameters of model variance and variance of the RW1 process. In the final model, we apply the delta method to calculate the assigned prior variance of the Gaussian priors for the parameters of the linear effects.

To evaluate the performance of the obtained posterior distributions, we compare them to previous assumptions, expert opinions, and observed data. According to Stakvik-Løvås (2022) and Kvalsvik (2022), we expect the salmon to produce a roughly equal amount of carbon dioxide as consumed oxygen. As such, a reasonable evaluation quantity would be the difference between the posterior carbon dioxide production and the posterior oxygen consumption. For comparing posterior models that fit the data and assumptions similarly well, we use the deviance information criterion (DIC) described in Spiegelhalter *et al.* (2002). However, the fit of the model with assumptions is prioritized over the DIC, which will only be used for comparison between models that perform similarly in this regard.

Our aim is to provide SE with a deeper understanding of the biological processes in their specific production facilities, enabling them to improve the health and growth conditions of the salmon in a safe manner. Additionally, we aim to evaluate the performance of the latent Gaussian model approach in this case, as a baseline for the potential future use of Bayesian methods on latent Gaussian models in the land-based salmon farming industry. By doing so, we hope to contribute to the development of more sustainable and efficient production practices in this field.

2 Case study of a land-based salmon farming facility

The transition from traditional salmon farming at sea to producing salmon to harvest weight on land has introduced new challenges, primarily related to optimizing the growth of the salmon

through the use of machines able to adjust water quality quantities, rather than combating lice and diseases and minimizing environmental impact. Salmon Evolution (SE) has provided time series data from nine sensors of adjustible and biological quantities, which we will utilize in order to approach our aims.

2.1 Salmon Evolutions hybrid flow-through system

The focus of this thesis is on a single operational tank containing salmon. A schematic representation of the water circuit of the tank is shown in Figure 1.

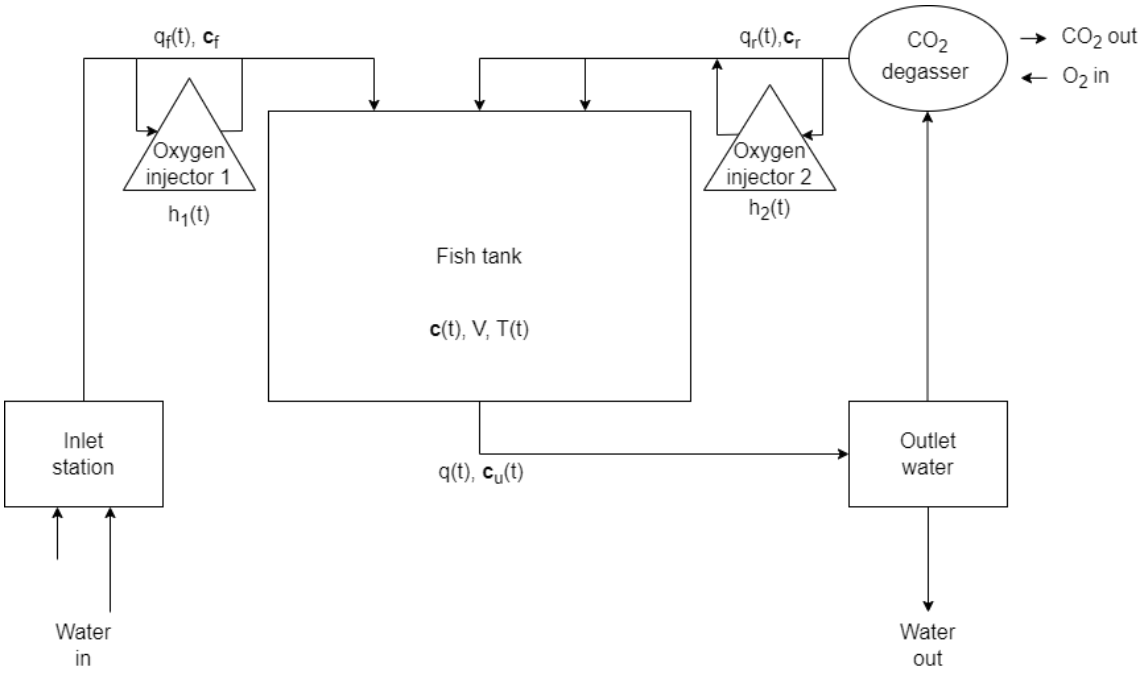

Figure 1: A sketch of the water flow and its compounds in a tank at Salmon Evolutions production facility.

Figure 1 illustrates a simplified view of the water flow through the tank at SE's production facility. Seawater is pumped into the intake station from two different depths to exploit seasonal variation and ensure redundancy. It is then treated with a particle filter and UV filter before being distributed to all tanks at the production site, explained by Lyngstad (2022). The seawater is also heated on its way from the intake station to the fish tank using customized heat pumps. The flow rate of the fresh seawater at time t is denoted as $q_f(t)$, with constant substance concentrations $\boldsymbol{c}_f = [c_{f,O_2}, c_{f,CO_2}]^T$ of oxygen and carbon dioxide. The magnitude of $q_f(t)$ is measured using electromagnetic flow sensors.

Additional oxygen is added to some of the water from $q_f(t)$ to achieve satisfactory oxygen levels in the tank with a saturation of approximately 100%, compensating for the oxygen consumed by the salmon, explained by Stakvik-Løvås (2022).The added oxygen at time t is denoted as $h_1(t)$ and is assumed to be independent of $q_f(t)$ and $\boldsymbol{c}_f$ for simplicity. The water is then introduced to the tank, which has a volume of $V = 4900m^3$ and substance concentrations $\boldsymbol{c}(t) = [c_{O_2}(t), c_{CO_2}(t)]^T$ of oxygen and carbon dioxide, and temperature $T(t)$ at time t. The tank is cylindrical and has an adjustable flow rate with the water circulating around the center of the tank.

The measured substance concentrations in the tank at time t are denoted as $\boldsymbol{y}(t) = [y_{O_2}(t), y_{CO_2}(t)]^T$ for oxygen and carbon dioxide. The oxygen concentration is measured near the center and near the edge of the tank using optical sensors, while the carbon dioxide concentration is measured at one location in the tank using another optical sensor. Water is drawn from the tank near the center

at a flow rate $q(t) = q_f(t) + q_r(t)$ at time t, where $q_r(t)$ will be described later, and has output substance concentrations $\boldsymbol{c}_u(t) = [c_{u,O_2}(t), c_{u,CO_2}(t)]^T$.

From $q(t)$, a portion of the water is recycled and denoted as $q_r(t)$, while the remainder is treated before being released into the ocean. The recycled water is then sent through a carbon dioxide degasser, where it is passed through several layers of filters that allow carbon dioxide to dissipate from the water and oxygen to dissolve into it, approaching equilibrium concentrations. It is assumed by the provider of the carbon dioxide degasser, Artec Aqua (n.d.) and by Kvalsvik (2022) and Stakvik-Løvås (2022) that the degasser has a constant efficiency of $\xi = 67\%$. This means that the water leaving the degasser has a carbon dioxide concentration reduction of $\xi = 67\%$ of what it was upon entering the degasser, i.e. $(1 - \xi)c_{u,CO_2}(t)$. It is also assumed that the oxygen saturation of the water leaving the degasser is 97%. Additionally, $q_r(t)$ is measured in a similar manner to $q_f(t)$ with electromagnetic sensors, and additional oxygen $h_2(t)$ is added to this part of the circuit, similar to $h_1(t)$.

2.2 Sensor data from a land-based salmon farming facility

The data provided by SE consists of time series measurements of various water quality parameters collected from nine sensors positioned inside and around a tank, as depicted in Figure 1. These sensors measure oxygen added, oxygen concentrations, water flows, carbon dioxide concentrations, and temperatures in the tank. The data was collected over a period of two and a half hours, from 03:00 to 05:30 one night in the fall of 2022, with a sampling resolution of one minute (denoted as timestep t). The data was obtained using a data system provided by Clarify (Clarify (2022)). This data was collected during a break in the current feeding regime, during which we would expect both oxygen consumption and carbon dioxide production to decrease over time. This is due to the fact that the feed amount in the tank gradually decreases after feeding has stopped, and is consumed in proportion to the oxygen input and resulting carbon dioxide production. The measured data is summarized in Table 1. We denote the 'actual' concentrations in the tank as $\boldsymbol{c}(t)$, and the 'observed' concentrations as $\boldsymbol{y}(t)$. We have also assigned notation for each measured quantity, as shown in Table 1. It is worth noting that the units of oxygen concentration are given as a percentage of saturation, while the preferred unit for estimating oxygen consumption is mg/l. This is to facilitate comparison between added oxygen $h_1(t)$ and $h_2(t)$ and oxygen concentration $y_{O_2}(t)$, as well as between carbon dioxide concentration $y_{CO_2}(t)$ and oxygen concentration. The other given and preferred units are also outlined in Table 1.

Table 1: Measured water quality parameters in the considered tank at the production facility of SE.

Notation	Measure	Given units	Preferred units
$h_1(t)$	Instant oxygen added from oxygen cone 1	kg/h	kg/min
$h_2(t)$	Instant oxygen added from oxygen cone 2	kg/h	kg/min
$y_{O_2}^C(t)$	Oxygen concentration near center	%	mg/l
$y_{O_2}^E(t)$	Oxygen concentration near edge	%	mg/l
$q_f(t)$	Fresh sea water flow	l/min	l/min
$q_r(t)$	Recycled water flow	l/min	l/min
$y_{CO_2}(t)$	Carbon dioxide concentration	mg/l	mg/l
$T^C(t)$	Temperature near center	$^\circ C$	$^\circ C$
$T^E(t)$	Temperature near edge	$^\circ C$	$^\circ C$

Preferred units for comparison between different quantities are listed in Table 1. The transformation of $h_1(t)$ and $h_2(t)$ from kg/h to kg/min was performed by multiplying the quantities by $1/60$ h/min. Oxygen concentration is given as a percentage of saturation, while the preferred unit for estimating oxygen consumption is mg/l, to facilitate comparison between added oxygen $h_1(t)$ and $h_2(t)$ and oxygen concentration $y_{O_2}(t)$, as well as between carbon dioxide concentration $y_{CO_2}(t)$ and oxygen concentration. Since oxygen saturation is dependent on factors such as salinity, pressure, and temperature, the transformation to the preferred unit is not straightforward. After consulting with Kvalsvik (2022), we utilized the calculator provided by Logio Systems (2022) to

perform this transformation. All other quantities were already provided in reasonable units for analysis.

2.3 Exploratory data analysis

We will now conduct an initial exploration of the data by plotting the data from each sensor over the specified time period and calculate the empirical mean and variance of the time series. Before doing so, we will transform the units to the preferred units. Specifically, we will convert the oxygen concentration from a percentage of saturation to mg/l using the calculator provided by Logio Systems (2022), assuming a salinity of 35‰, an atmospheric pressure of 1 atm, and a water temperature of $11.9°C$, as obtained from Figure 6 as described above.

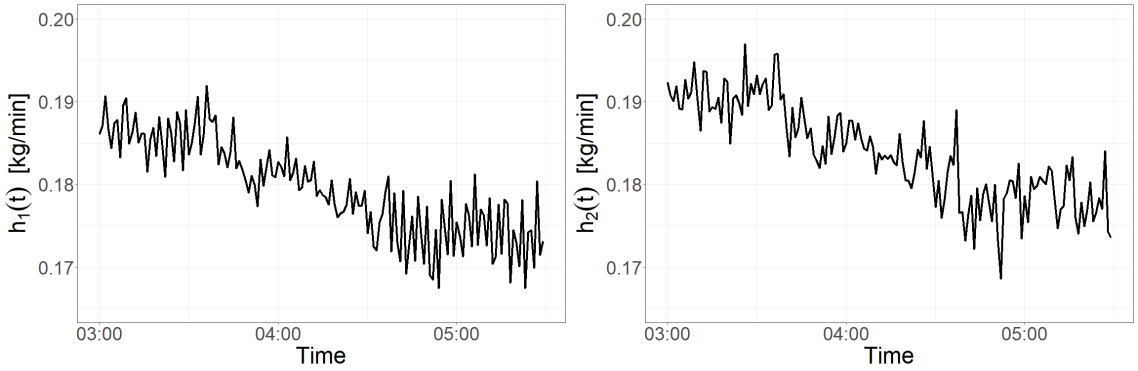

Figure 2: Plotted values of oxygen added to the tank form injector 1 and 2, $h_1(t)$ and $h_2(t)$, respectively.

As seen in Figure 2, the amount of added oxygen from the two oxygen injectors is roughly equal, with a slight decrease over time. According to Kvalsvik (2022) and Stakvik-Løvås (2022), this can be attributed to the fact that the injectors are programmed to inject oxygen in order to maintain a saturation level in the water close to 100%. During the selected time period, the feeding regime involved continuous feeding throughout the day with periodic breaks at night. The time period in question falls within one of these breaks, which explains the decrease in oxygen consumption as more of the feed is converted into energy, biomass, and carbon dioxide.

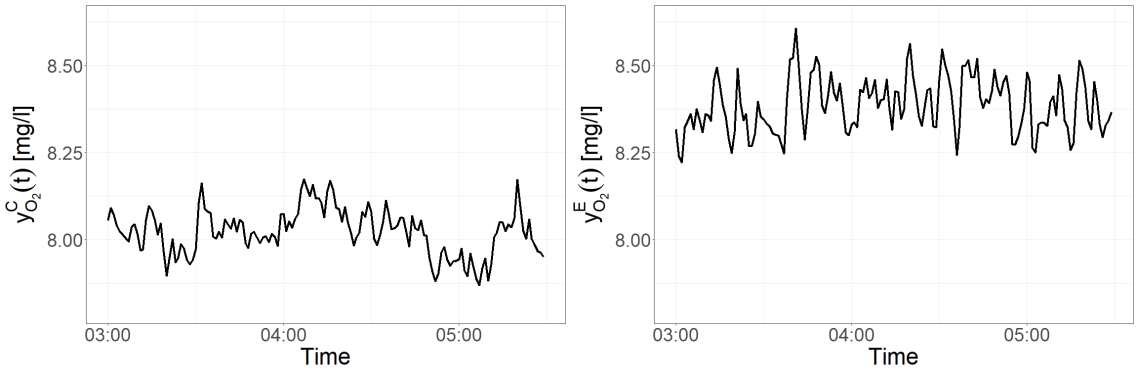

Figure 3: Plotted values of measured oxygen concentrations near the center and edge of the tank, $y^C_{O_2}(t)$ and $y^E_{O_2}(t)$, respectively.

Figure 3 illustrates approximately constant values for the oxygen concentrations at the center and edge of the tank, as it is desirable to maintain these concentrations within certain bounds of saturation. It can be observed that the oxygen concentration is higher at the edge than at the center. This can be explained by a combination of two factors. Firstly, the water injected into the

tank, which has higher oxygen concentrations, is injected near the edge, leading to higher oxygen saturation in this region. Secondly, the current in the water creates a slight water profile that "pushes" the dissolved oxygen towards the edge, as described by Forsberg (1994).

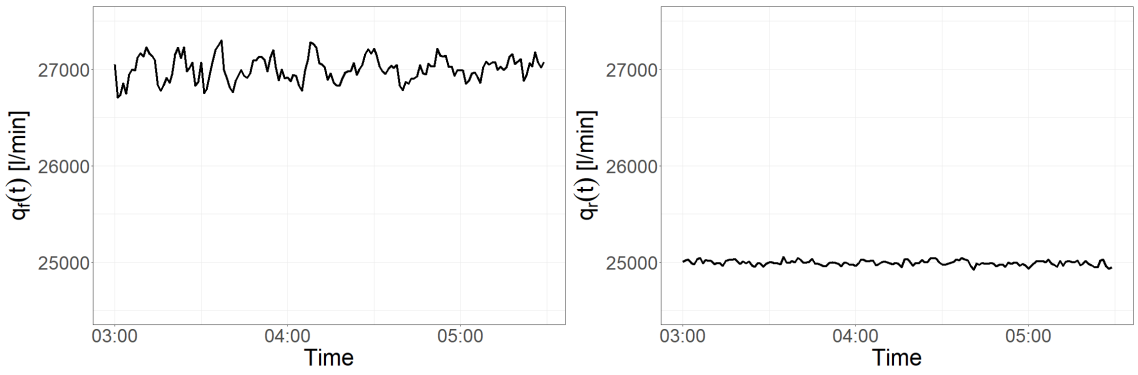

Figure 4: Plotted values of flow rates into the tank of fresh seawater and recycled water, $q_f(t)$ and $q_r(t)$, respectively.

Figure 4 demonstrates that the flows of fresh sea water and recycled water are approximately constant. The amount of fresh sea water is slightly higher than that of recycled water, but both appear to remain constant over the given time period.

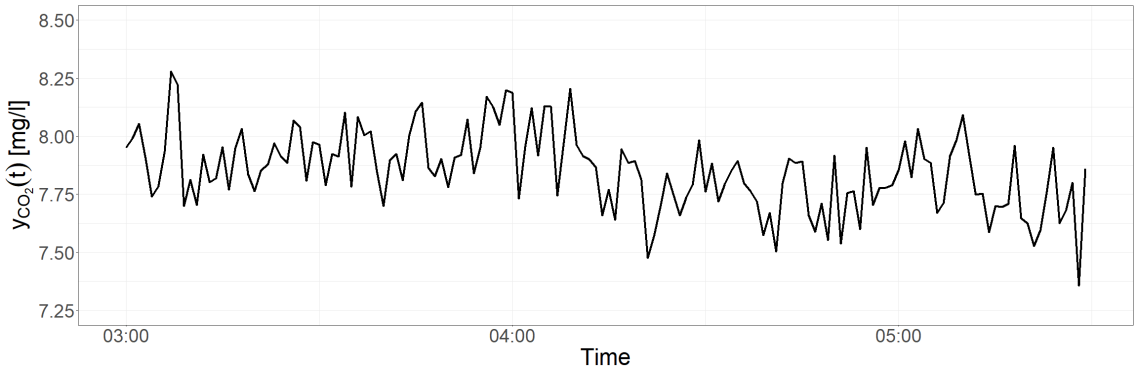

Figure 5: Plotted values of the measured carbon dioxide contentration in the tank, $y_{CO_2}(t)$.

Figure 5 reveals that the oxygen concentration in the tank has changed little over the time period. A slight decline may be observed, which could potentially be attributed to the feeding break mentioned earlier as a cause for the decrease in oxygen injection.

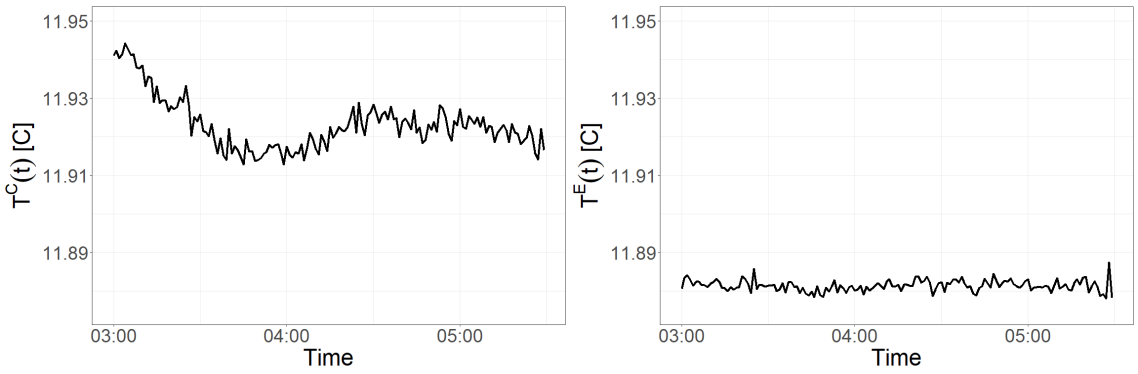

Figure 6: Plotted values for the temperature in the tank measured near the center and near the edge, $T^C(t)$ and $T^E(t)$, respectively.

Figure 6 indicates that the temperature measured by both sensors is essentially equal and constant throughout the period, with a numerical value of $11.9°C$.

I addition to plots, the empirical mean and variance of the quantities of the time series data are systematized in Table 2.

Table 2: Empirical mean and variance of the quantities of the time series data.

Notation	Mean	Unit	Variance
$h_1(t)$	0.180	kg/min	$3.39 \cdot 10^{-5}$
$h_2(t)$	0.184	kg/min	$3.50 \cdot 10^{-5}$
$y^C_{O_2}(t)$	8.02	mg/l	0.00431
$y^E_{O_2}(t)$	8.39	mg/l	0.00621
$q_f(t)$	$2.70 \cdot 10^4$	l/min	$1.61 \cdot 10^5$
$q_r(t)$	$2.50 \cdot 10^4$	l/min	614
$y_{CO_2}(t)$	7.85	mg/l	0.0278
$T^C(t)$	11.9	$°C$	$4.70 \cdot 10^{-5}$
$T^E(t)$	11.9	$°C$	$2.14 \cdot 10^{-6}$

2.4 Assumptions

In order to model the consumed oxygen and produced carbon dioxide by the salmon, we need to make some assumptions about the process before developing the model.

Our first assumption is that the total amount of added oxygen is equal to the sum of the added oxygen from the two oxygen injectors. We express this as

$$h(t) = h_1(t) + h_2(t).$$

The total amount of added oxygen, which is used in the subsequent analysis of this study, is plotted in Figure 7.

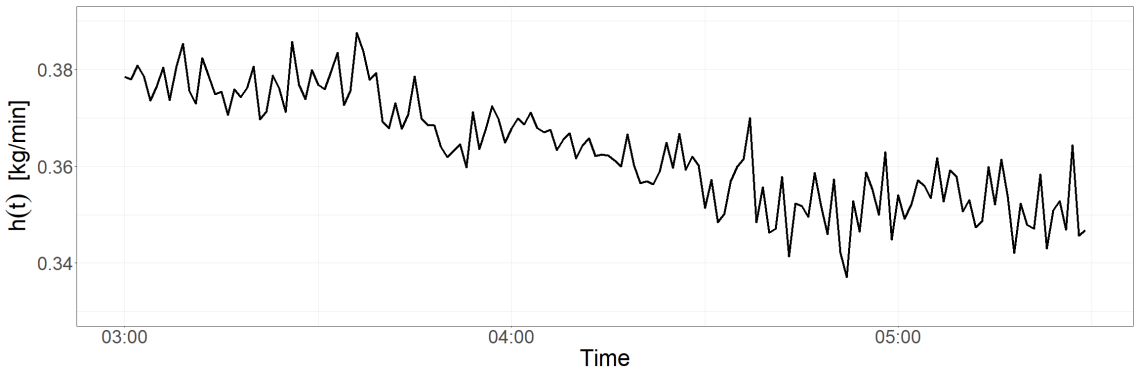

Figure 7: Plotted values of total amount of oxygen added to the tank form injector 1 and 2 combined, $h(t)$.

Figure 7 displays a similar trend for the total oxygen added to the tank, $h(t)$, as the two injectors, $h_1(t)$ and $h_2(t)$, plotted in Figure 2, with a decrease in added oxygen over the time period.

We have two sensors measuring the oxygen concentrations in the tank. In order to compare the changes in oxygen and carbon dioxide amounts in the tank, and given that we only have one measurement for carbon dioxide concentration, it is necessary to model the changes in these two quantities in a similar manner. To do so, we need an estimate of the substance concentrations in the outtake water. A reasonable assumption would be to estimate the outtake concentrations based on the concentrations near the center. However, as this is only possible for oxygen and not for carbon dioxide, we assume that the oxygen concentration in the tank at time t is equal to the average of the two provided measurements, i.e.

$$y_{O_2}(t) = \frac{1}{2}\left[y^C_{O_2}(t) + y^E_{O_2}(t)\right].$$

The data for the average observed oxygen concentration, calculated as described above, is plotted together with the observed carbon dioxide concentration from Figure 5 in Figure 8.

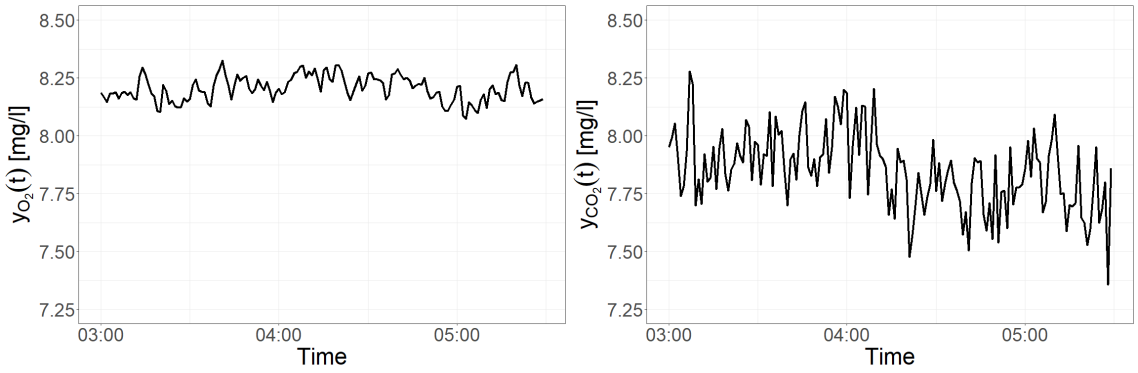

Figure 8: Plotted avgerage value of oxygen concentration in the tank over the two measuring points, $y_{O_2}(t)$, and plotted value of the observed carbon dioxide concentration $y_{CO_2}(t)$.

We observe from Figure 8 that the observed carbon dioxide concentration, $y_{CO_2}(t)$, exhibits more noise than the average observed oxygen concentration, $y_{O_2}(t)$.

Given that we have similar measurements for the concentrations of oxygen and carbon dioxide in the tank, we make another assumption that the observed concentrations in the outlet water, $\boldsymbol{y}_u(t)$, are equal to the concentrations in the tank in the previous timestep, i.e.

$$\boldsymbol{y}_u(t) = \boldsymbol{y}(t-1), \tag{1}$$

where $\boldsymbol{y}(t) = [y_{O_2}(t), y_{CO_2}(t)]^T$. This assumption is not necessarily as accurate as considering the concentration deviation between the center and the edge, but it allows us to compare the substances using the available data.

We assume that the flow rates $q_f(t)$ and $q_r(t)$ are constant quantities, q_f and q_r, with value equal to the empirical mean of the time series data, given in Table 2. We denote the total flow rate, $q(t)$, which is the flow out of the tank, as

$$q(t) = q_r(t) + q_f(t).$$

The total flow, $q(t)$, is plotted in Figure 9 and is assumed to have the same properties as q_f and q_r, as a constant quantity q with value equal to the empirical mean over the time series, given in Table 2.

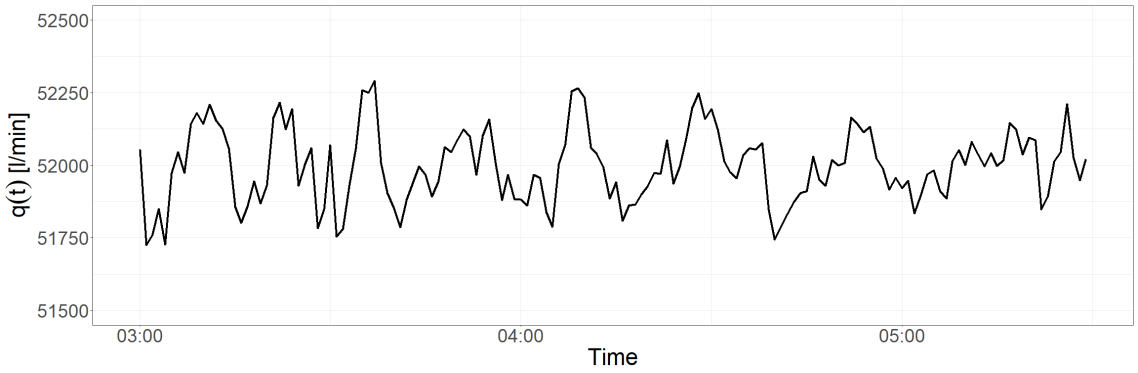

Figure 9: Plotted value of total water flow through the tank, $q(t)$.

In addition, we make some assumptions about constant values for the physical phenomena related to the tank. Specifically, we assume that the oxygen saturation in fresh sea water is 100%, which is converted into the unit of mg/l by assuming that the temperature is equal to that inside the tank and applying the calculator at Logio Systems (2022). We also assume that the fresh sea water has a carbon dioxide concentration of 1 mg/l according to Kvalsvik (2022). We further assume that the water from the recycled water has an oxygen saturation of 97%, and that the carbon dioxide degasser has an effectiveness of 67%, denoted as ξ, as described in Chapter 2.1. These assumptions are summarized in Table 3.

Table 3: Assumed parameter values by Salmon Evolution.

Notation	Measure	Assumed value	Denomination
c_{f,O_2}	O_2 concentration fresh sea water	100	%
c_{f,CO_2}	CO_2 concentration fresh sea water	1.0	mg/l
c_{r,O_2}	O_2 concentration recycled water	97	%
ξ	CO_2 degasser effect	67	%

To summarize, we are left with the temporal variables $\boldsymbol{y}(t) = [y_{O_2}(t), y_{CO_2}(t)]$, $\boldsymbol{y_u}(t) = [y_{u,O_2}(t), y_{u,CO_2}(t)]$ and $h(t)$, and the assumed constants q, q_r, q_f, c_{f,O_2}, c_{f,CO_2}, c_{r,O_2} and ξ. As for the measured data in Chapter 2.3, the empirical mean and variance of the temporal quantities are summarized in Table 4. With our definition of $\boldsymbol{y}_u(t)$, the empirical mean and variance of $\boldsymbol{y}_u(t)$ is essentially the same as for $\boldsymbol{y}(t)$.

2.5 Parameters of interest

The quantities that we want to explore are the amount of consumed oxygen and the amount of produced carbon dioxide by the salmon. Based on our current assumptions and data, we can

Table 4: Empirical mean and variance of the temporal quantities after model assumptions.

Notation	Mean	Unit	Variance
$y_{O_2}(t)$	8.20	mg/l	0.00294
$y_{CO_2}(t)$	7.85	mg/l	0.0278
$h(t)$	0.363	kg/min	$1.31 \cdot 10^{-4}$

calculate the oxygen consumption and carbon dioxide production using the observed data. For now, we denote the calculated consumed oxygen and produced amounts of substances as $\boldsymbol{g}(t) = [g_{O_2}(t), g_{CO_2}(t)]$. To facilitate comparison, we consider accumulated amounts in the tank, with a positive direction indicating accumulation of the substance. Therefore, the consumed oxygen, $g_{O_2}(t)$, is negative and the produced carbon dioxide, $g_{CO_2}(t)$, is positive. We assume that the difference in oxygen amount in the tank from time $t-1$ to t is $(Vy_{O_2}(t) - Vy_{O_2}(t-1))$. This is equal to the difference between added oxygen, from the injectors and the inlet water, and the removed oxygen from the outlet water and consumed by the salmon. The added oxygen comes from the fresh sea water in the amount of $c_{f,O_2}q_f$, from the recycled water in the amount of $c_{r,O_2}q_r$, and from the injectors in the amount of $h(t)$. The removed amounts consist of the oxygen concentration in the tank multiplied by the flow out, in addition to the amount consumed by the salmon. Therefore, the oxygen consumed by the salmon can be estimated using the observed values as

$$g_{O_2}(t) = Vy_{O_2}(t) - Vy_{O_2}(t-1) - h(t) - q_f c_{f,O_2} - q_r c_{r,O_2} + qy_{u,O_2}(t).$$

Similarly, this approach can be used to calculate the carbon dioxide produced by the salmon. In this case, the concentration of carbon dioxide in the recycled water is $(1-\xi)y_{u,CO_2}(t)$, such that

$$g_{O_2}(t) = Vy_{CO_2}(t) - Vy_{CO_2}(t-1) - q_f c_{f,CO_2} - q_r(1-\xi)y_{u,CO_2}(t) + qy_{u,O_2}(t).$$

It is important to note that $h(t)$ is measured in kg/min and other parameters are measured in mg/l and l/min. To get the magnitudes in an appropriate order, we measure the consumed and accumulated substances in kg/min. We plot the estimated values based on the observed data in Figure 10. All temporal terms in the expressions are based on observed values.

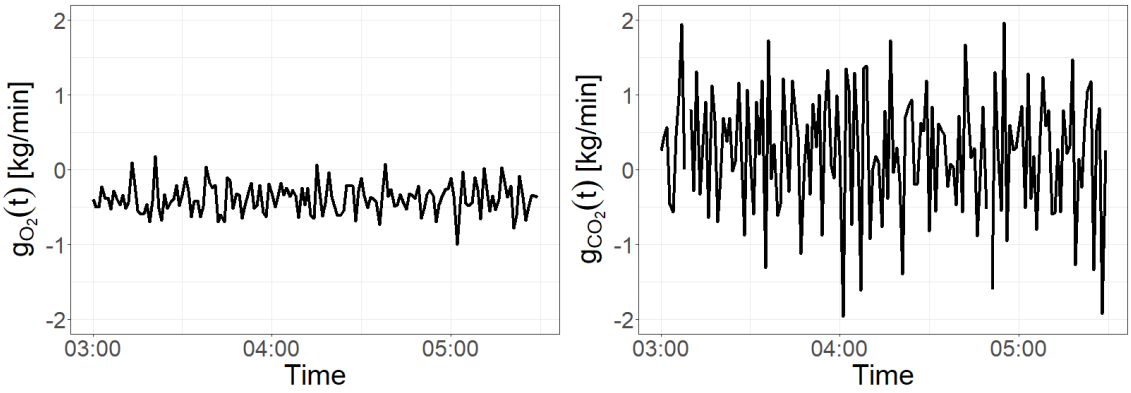

Figure 10: Plotted value of estimated accumulated oxygen and carbon dioxide, $g_{O_2}(t)$ and $g_{CO_2}(t)$, respectively.

From Figure 10, we can see that the calculated accumulated amount of oxygen varies slightly around a value right below zero. The calculated accumulated amount of carbon dioxide, on the other hand, is considerably more noisy due to the fact that the observed carbon dioxide concentration in the tank is more noisy than the observed oxygen concentration, as we can see in Figure 8.

To evaluate the difference between produced carbon dioxide and consumed oxygen, we can examine the sum of the calculated accumulated substance time series along with a histogram of the differences, as shown in Figure 11.

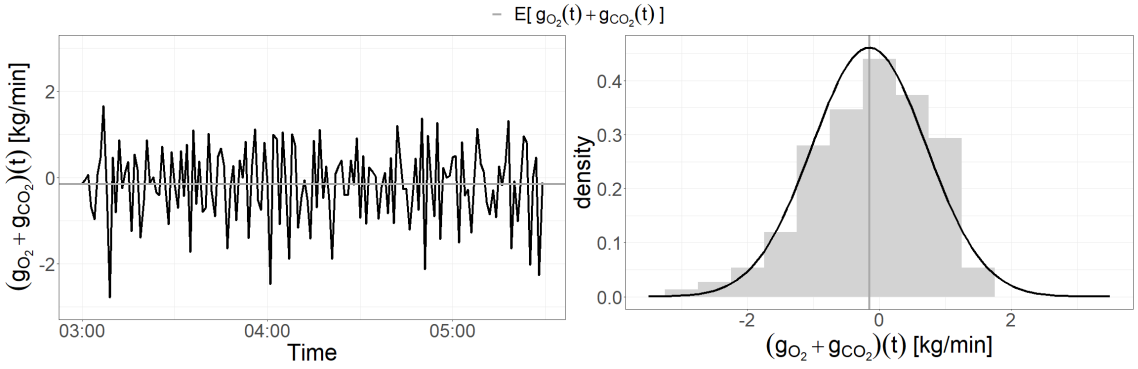

Figure 11: Plot of the value of the difference between the calculated consumed oxygen and produced carbon dioxide, $g_{O_2}(t) + g_{CO_2}(t)$, as well as a histogram and density plot using the empirical mean and variance of the data of the observed differences. The mean of the data is also displayed as $\mathrm{E}[g_{O_2}(t) + g_{CO_2}(t)]$.

We observe that the value of the sum of accumulated carbon dioxide and consumed oxygen fluctuates around zero with noise from the two estimated observed values. The histogram show that the mode of observed differences are close to zero, which we could expect, but that there is a heavier tail for negative observations, indicating that the salmon consumes more oxygen than it is expected to compared to the accumulated carbon dioxide. The mean is slightly less than zero, with a numerical value of $\mathrm{E}[g_{O_2}(t) + g_{CO_2}(t)] = -0.15$. As well, the distribution does not look perfectly Gaussian as we can see from the comparizon of the histogram to the Gaussian distributed density plot, which can be useful to have in mind when evaluating the results.

3 Bayesian inference using latent Gaussian models

In this chapter, we will begin by providing a brief overview of Bayesian inference. We will then delve into the subject of latent Gaussian models and discuss various techniques for conducting Bayesian inference using integrated nested Laplace approximation (INLA). We will also cover some basic theory on random effects and methods for fitting prior distributions. Finally, we will introduce the R-INLA package and the deviance information criterion (DIC).

3.1 An introduction to Bayesian inference

Bayesian inference, as described by Gómez-Rubio 2020, is a statistical approach that allows us to make predictions or draw conclusions about unknown quantities based on data and prior knowledge or assumptions. It involves estimating the probability distribution of the unknown quantities, referred to as the posterior distribution, given the data and prior information. This is done through the use of Bayes' theorem, which states that the posterior distribution is proportional to the product of the likelihood of the data given the unknown quantities and the prior distribution of the unknown quantities. By updating the prior distribution with the data, we can improve our understanding of the unknown quantities and make more accurate predictions or conclusions about them. Bayesian inference is particularly useful when the data are limited or the relationships between the unknown quantities are complex, as it allows us to incorporate subjective information and consider different possible scenarios.

When considering Bayesian statistics, one first introduce conditional probabilities and from there derive Bayes theorem, which for two events A and B states that

$$Pr(B|A) = \frac{Pr(A|B) \cdot Pr(B)}{Pr(A)}, \tag{2}$$

i. e. that the probability of an event occurring, given certain evidence or conditions A, is equal to the probability of the evidence or conditions occurring, given the event B, multiplied by the probability of the event occurring, divided by the probability of the evidence or conditions occurring.

In this study, we define the likelihood function as

$$L(\boldsymbol{\theta}) = \pi(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta}),$$

where $\pi(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta})$ represents the probability density of the observed variable $\boldsymbol{Y}$, given a set of parameters $\boldsymbol{\theta}$.,from now on refered to as the likelihood $\pi(\boldsymbol{y}|\boldsymbol{\theta})$. We also introduce the prior density of the parameters, $\pi(\boldsymbol{\theta})$. Using Bayes theorem in Equation 2, we can then derive the posterior distribution of the parameters, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, as follows:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})}. \tag{3}$$

This allows us to incorporate the prior knowledge about the parameters into our analysis and make inferences about their values given the observed data. The marginal likelihood, denoted as $\pi(\boldsymbol{y})$ is computed by

$$\pi(\boldsymbol{y}) = \int_{\Theta} \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

This can be challenging to compute in many cases, leading to the posterior distribution only being available in closed form for a few models as it is often highly multivariate. However, since it is independent of $\boldsymbol{\theta}$, we can express the posterior distribution as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}),$$

where the symbol $\propto$ represents proportionality. Through Bayesian calculations, we are able to extract parameters of interest on the posterior distribution, such as the posterior mean, posterior quantiles, and posterior credible sets, using the joint posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. We can obtain the same properties of each marginal distribution of each element of $\boldsymbol{\theta}$ by integrating out all other parameters in $\boldsymbol{\theta}$ than θ_i, denoted $\boldsymbol{\theta}_{-i}$, i. e.

$$\pi(\theta_i|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-i},$$

where $i \in \{1, ..., \dim(\boldsymbol{\theta})\}$.

To estimate parameters in models where the closed form posterior distribution is not available, we can use sampling methods such as Markov chain Monte Carlo (MCMC). In the case of a Gaussian likelihood with a Gaussian prior, the conjugate prior for the mean is also Gaussian, resulting in a Gaussian posterior distribution for the mean. When this is not the case, we must use other methods to estimate the posterior distribution, such as MCMC. These methods allow us to obtain estimates for the posterior mean, quantiles, and credible sets for each parameter in the model, by sampling from the joint posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

The likelihood function represents the probability of collecting a certain set of data given a set of parameters. By introducing prior knowledge or expert opinions through the use of a prior distribution, we can incorporate this information into our model. When the prior distribution contains a significant amount of prior knowledge, it is known as an informative prior, and can be closer examined in O'Hagan *et al.* (2006). On the other hand, when our prior knowledge is limited, the prior distribution is less strict and the posterior distribution is more heavily influenced by the observed data. These are known as noninformative priors, and are described closer in Lesaffre and Lawson (2012).

We may also separate between proper and imroper priors on to which extent they sum to 1 or not, respectively, which is described in Gelman *et al.* (2013). Using an improper prior can have a number of consequences, as that the posterior distribution may not exist or may not be well-defined. This can lead to difficulties in estimating the posterior distribution and in making predictions based on the posterior. Another consequence of using an improper prior is that the posterior distribution may not have desirable properties, such as being unbiased or having low variance. This can lead

to poor performance of the model, particularly if the data are limited or noisy. In general, it is generally recommended to use proper prior distributions in Bayesian analysis to ensure that the posterior distribution is well-defined and has desirable properties. However, in some cases it may be appropriate to use an improper prior, particularly if the data are very informative or if the model is being used for exploratory purposes. In these cases, it is important to carefully consider the potential consequences of using an improper prior and to check the robustness of the results to different prior choices.

Computational methods for Bayesian inference generally aim to obtain estimates appearing in the inference process, as described in among others Blangiardio and Cameletti (2015) and Moraga (2019). The posterior mean of a parameter θ_i, $\mathrm{E}[\theta_i|\boldsymbol{y}]$, is for instance defined as

$$\mathrm{E}[\theta_i|\boldsymbol{y}] = \int_{\Theta_i} \theta_i \pi(\theta_i|\boldsymbol{y}) d\theta_i,$$

where $\pi(\theta_i|\boldsymbol{y})$ is the marginal posterior distribution of the univariate parameter θ_i. Similar integrals also takes place for among others the variance and quantiles. These can be solved using Laplace approximation or numerical integration, further described in Tierney and Kadane (1986). Bayesian inference can be extracted using other methods, dependent on the framework of the model. For instance, on could use Monte-Carlo methods, which simulates samples from marginal and conditional posteriors, to extract posterior parameters of interest. Such methods does however have restrictions when approaching spaces of high dimensions. By maximizing the product of the prior and the likelihood, one can obtain point estimates by Maximum a posteriori estimation. It can be atimated in using several different methods, such as mor instance the Expectation Maximization (EM) algorithm described in Gelman *et al.* (2013).

Another class of methods for Bayesian inference is Markov chain Monte Carlo (MCMC) methods, described in among others Casella and George (1992) and Brooks *et al.* (2011). The concept of these is to draw samples from the joint posterior distribution, based on constructiong a Markov chain whose stationary distribution is the posterior distribution, such that the Markov chain approaches the joint posterior distribution after a certain amount of iterations. The Metropolis-Hastings algorithm and Gibbs sampler constitutes the two mainly used algorithms for Bayesian inference by MCMC methods, both described in Brooks *et al.* (2011).

The Metropolis-Hastings algorithm, as described in Gómez-Rubio (2020), uses a proposal distribution $q(\cdot|\boldsymbol{\theta})$ conditional on the latest accepted value of the parameter $\boldsymbol{\theta}$ to draw a new proposal value $\boldsymbol{\theta'}$ from the proposal density. The proposal distribution can be chosen in several ways, such as for instance a Gaussian distribution with mean as the latest accepted value and a predetermined variance. When a new proposal $\boldsymbol{\theta'}$ is drawn it is accepted wit the probability α given as

$$\alpha = \min\left\{1, \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta'})\pi(\boldsymbol{\theta'}|\boldsymbol{y})}{q(\boldsymbol{\theta'}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})}\right\}.$$

The term $q(\boldsymbol{\theta}|\boldsymbol{\theta'})$ is the proposal distributions, $q(\cdot|\boldsymbol{\theta'})$, density evaluated at $\boldsymbol{\theta}$. The acceptance probability α also contains the posterior distribution, which is only known after the computation. This is however taken care of using Bayes theorem such that

$$\alpha = \min\left\{1, \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta'})\pi(\boldsymbol{y}|\boldsymbol{\theta'})\pi(\boldsymbol{\theta'})}{q(\boldsymbol{\theta'}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right\},$$

independent of $\pi(\boldsymbol{y})$ such that we avoid computing the marginal distribution. This probability causes the new proposed values to be accepted within the target posterior distribution at an appropriate rate, and it favours the prior probability to move when there is a chance of coming back to the current distribution to avoid many unneccesary sampes in low posterior probability regions.

The Gibbs sampler explained by Casella and George (1992) works in somewhat the same manner, and can be consider as a charactereistc case of the Metropolis-Hastings algorithm. The Gibbs sampler sets the proposal distribution equal to the full conditional distribution, conditional on all model parameters. Mathematically, this yields the proposal distribution $q(\theta'|\boldsymbol{\theta}) = \pi(\theta_i|\boldsymbol{y}, \boldsymbol{\theta}_{-i})$. As a consequence, the acceptance probability α is always equal to 1 such that the Gibbs sampler allways accepts all proposals automatically.

The convergence of the posterior distribution generally takes a given number of iterations, and different techniques of concergence diagnostics are given in Gelman and Rubin (1992). Advantages of MCMC methods are that they are applicaple for a wide rarnge of problems, and that they are easy to implement. However, they might get cumputationally intensive making them fragile for very large or very complex models.

Integrated nested Laplace Approximation is a computational method to perform approximate Bayesian inference on latent Gaussian models on a stricter computational budget. It was invented by Rue *et al.* (2009) as a proposed method of performing Bayesian inference fatser than often computer intensive MCMC methods.

3.2 Latent Gaussian models

Latent Gaussian models, described in among others Rue *et al.* (2009), Blangiardio and Cameletti (2015) and Gómez-Rubio (2020), is a class of models with certain characteristics such that inference tasks can be generalized for the whole class. Such models are built of a likelihood, a latent Gaussian field, and a vector of hyperparameters. We assume for these models that the observations $\boldsymbol{y}$ are independent, conditioning on the latent Gaussian random field $\boldsymbol{x}$ and the subset of hyperparameters $\boldsymbol{\theta_1}$, where the whole set of hyperparameters are $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$. Mathematically this yields

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta_1} \sim \prod_{i \in I} \pi(y_i|x_i, \boldsymbol{\theta_1}).$$

The generality of inference schemes for latent Gaussian models indicates that the Gaussian latent field is conditionally independent and normally distributed given the subset of hyperparameters $\boldsymbol{\theta_2}$, i. e.

$$\boldsymbol{x}|\boldsymbol{\theta_2} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta_2}), \boldsymbol{Q}^{-1}(\boldsymbol{\theta_2})),$$

where $\boldsymbol{Q}(\boldsymbol{\theta_2})$ is the precision matrix of $\boldsymbol{x}$ given $\boldsymbol{\theta_2}$. This property includes all random terms of a statistical model and describes dependency structures in data. The hyperparameters control the likelihood of the data and the Gaussian latent field and has positerior on the form

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i \in I} \pi(y_i|x_i, \boldsymbol{\theta}).$$

Latent Gaussian model are often mentioned together with generalized linear models. If we interpretate the likelihood of the latent Gaussian model $\pi(y_i|x_i, \boldsymbol{\theta})$ to the extent that any oservation y_i only depend on its linear predictor η_i through a link function, $\eta_i = g(y_i)$, the generalized linear model setup defined by Fahrmeir and Tutz (2001) in Equation 4 is valid. The link function can take several forms, such as among others log links where $g(y_i) = \log y_i$, logit links where $g(y_i) = \log(y_i/(1 - y_i))$ or simple idrentity links where $g(y_i) = \mathrm{E}[y_i]$, all described by Fahrmeir and Tutz (2001). Setting the element of the Gaussian latent field $\{x_i, i \in I\}$ equal to the linear predictor η_i, we can write

$$\eta_i = \beta_0 + \sum_j \beta_j z_{ij} + \sum_k f_{k,j_k(i)} + \epsilon_i, \tag{4}$$

where β_0 is the model intercept, and $\boldsymbol{z} = [z_{i1}, ..., z_{ip}]^T$ are linear effects with coefficients $\boldsymbol{\beta} = [\beta_1, ..., \beta_p]$, the terms $\{f_k\}$ random or mixed effects, and $\epsilon_i \sim N(0, \sigma_\eta^2)$ a noise term.

The random effects f_k are explained in detail in Fahrmeir and Tutz (2001) and represents specific Gaussian processes and each element f_k are labelled as model components where element j affects linear predictor η_i. These variables are typically included to account for unobserved heterogeneity or variation that is not explained by the fixed effects in the model. Random effects can be included in a latent Gaussian model using a variety of different specifications, depending on the nature of the data and the research question. For example, random effects can be included as intercept terms to allow for individual-level differences in the mean response, or they can be included as slope terms to allow for individual-level differences in the relationship between the response and one or more predictor variables. The inclusion of random effects in a latent Gaussian model allows the model to account for the correlated structure of the data, which can improve the accuracy

and precision of the estimates. The random effects can be estimated using maximum likelihood or Bayesian methods, and their estimates can be used to make inferences about the population from which the sample was drawn. The components f_k may among others represent random-walk processes, autoregressive processes or identical independelt distributed (iid) elements.

We now assign assumed gaussian priors on the intercept β_0 and fixed effects $\boldsymbol{\beta}$ in Equation 4 such that that the model formulation in this equation and latent Gaussian models refer to the same class of models. The joint distribution of the Gaussian latent field $\boldsymbol{x} = (\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{f})$, where $\boldsymbol{f} = [\boldsymbol{f_1}, ..., \boldsymbol{f_n}]$ and $\boldsymbol{f_i} = [f_1, ..., f_k]$, is Gaussian, and by considering the noise term $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ in Equation 4, $\boldsymbol{x}$ is also nonsingular, such that the column vectors of the model matrix are independent. This is essential for the model to be identifiable, which means that the model parameters can be uniquely estimated from the data.

We now have the latent field $\boldsymbol{x}$ in terms of the hierarchial latent Gaussian model formulation with $\dim(\boldsymbol{x}) = n$ where n is the sum of the number of observations, intercepts, fixed effects and dimension of all model components. The hyperparameters $\boldsymbol{\theta}$ contains the parameters of the model components and of the likelihood, which each have zero or more of for instance a scale, variance, shift or correlation parameters. This makes the number of hyperparameters small and independent of the number of observations nor the dimension of the latent field n. This is necessary for the computational efficiency for latent Gaussian models as the number of hyperparameters remains the same for any number of observations.

3.3 Bayesian inference on latent Gaussian models

Integrated Nested Laplace Approximation (INLA) is a method for performing Bayesian inference on latent Gaussian models. It is a computationally efficient alternative to traditional Markov chain Monte Carlo (MCMC) methods and can be used to estimate posterior distributions and make predictions based on the model. INLA works by approximating the posterior distribution using a Laplace approximation, which involves approximating the posterior distribution with a Gaussian distribution centered at the mode of the posterior. This approximation is then nested within an integration scheme, which allows for the estimation of the posterior distribution for the latent variables and the hyperparameters of the model.

Laplace approximation is a technique for approximating integrals. This brief introduction is inspired by Blangiardio and Cameletti (2015). For now, we assume that we are interested in computing the value of the integral of a density function $f(x)$ of a random variable X. We can write

$$\int f(x)dx = \int \exp \log f(x)dx.$$

Taylor expanding $\log f(x)$ of second order around some value $x = x_0$, we can write

$$\log f(x) \approx \log f(x_0) + (x - x_0)\frac{\partial}{\partial x} \log f(x)\bigg|_{x=x_0} + \frac{1}{2}(x - x_0)^2 \frac{\partial^2}{\partial x^2} \log f(x)\bigg|_{x=x_0}.$$

Defining x_0 as $x^* = \text{argmax}_x \log f(x)$ such that $(\partial/\partial x) \log f(x)|_{x=x^*} = 0$, the approximation can be written as

$$\log f(x) \approx \log f(x^*) + \frac{1}{2}(x - x^*)^2 \frac{\partial^2}{\partial x^2} \log f(x)\bigg|_{x=x^*},$$

such that the integral we wanted to approximate can be written as

$$\int f(x)dx \approx f(x^*) \int \exp\left\{\frac{1}{2}(x - x^*)^2 \frac{\partial^2}{\partial x^2} \log f(x)\bigg|_{x=x^*}\right\} dx,$$

which seemingly has the same form as a Gaussian distribution. Setting $(-1)/((\partial^2/\partial x^2) \log f(x)|_{x=x^*}) = \sigma^{2^*}$, we can write our integral approximation as

$$\int f(x)dx \approx f(x^*) \int \exp\left\{-\frac{(x - x^*)^2}{2\sigma^{2^*}}\right\} dx,$$

with the integrand as the exponential factor of a Gaussian distribution with mean x^* and variance σ^{2^*}. Finally, we can write the integral approximation as

$$\int_a^b f(x)dx \approx f(x^*)\sqrt{2\pi\sigma^{2^*}}(\Phi(b) - \Phi(a)),$$

where $\Phi(\cdot)$ represents the density function of a Gaussian distribution with mean x^* and variance σ^{2^*}. If now $f(x)$ ia the density function of any other distribution, this can thus be approximated as a normal distribution, which can be used to perform inference of the distribution.

We conisder some assumptions to ensure accuracy of approximations and computational necessities. The assumptions yields that the number of hyperparameters $\boldsymbol{\theta}$ should be small, the distribution of the latent field $\boldsymbol{x}|\boldsymbol{\theta}$ should be Gaussian and close to a Gaussian Markov random field (GMRF) when the dimension of the data n is large, and the observations $\boldsymbol{y}$ are mutually conditionally independent of the Gaussian latent field $\boldsymbol{x}$ and the hyperparameters $\boldsymbol{\theta}$ such that each observation y_i only depend on one latent field component x_i, as given in Rue *et al.* (2016). In this case, we consider small as $\dim(\boldsymbol{\theta}) \leq 5$ and large as $\dim(\boldsymbol{\theta}) \geq 10^4$, aligned with the assumptions of Gómez-Rubio (2020).

As mentioned, the latent field should satisfy, or be close to the requirements of a sparse Gaussian Markov random field (GMRF). A latent field $\boldsymbol{x}$ is GMRF if it is Gaussian and that the elements x_i and x_j of $\boldsymbol{x}$ are conditionally independent given all remaining elements $\boldsymbol{x}_{-ij}$. An introduction to the concept is given in Held and Rue (2010). The advantage of a GMRF is that the pairs of conditionally independent entries in $\boldsymbol{x}$ contributes with zeros to the precision matrix making it sparse, simplifying computations significantly. A sparse matrix is a matrox of which most entries are zero simplifying computational operations. Combining GMRFs and additive/generalized linear models gives, as mentioned in Rue *et al.* (2016), that when $\boldsymbol{x}$ is a GMRF, the joint distribution for $\boldsymbol{x}$ is also a GMRF with precision matrix equal to the sum of precision matrixes of fixed effects and other model components. The derivation of the joint distribution of the latent field is a process useful to repeat frequently as it is depnendent on the hyperparameters $\boldsymbol{\theta}$, such that the necessity of simple precision matrix approximations is present. For a GMRF with a sparse precision matrix, this is fulfilled.

We have now defined what we need for designing a model on the form of a latent Gaussian model as in Equation 4 with the conditional model for the observations also known as the likelihood

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \prod_i \pi(y_i|\eta_i,\boldsymbol{\theta}),$$

the latent field $\boldsymbol{x}$ as a GMRF conditional on $\boldsymbol{\theta}$,

$$\boldsymbol{x}|\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})),$$

and the prior distribution of $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}),$$

aligned with the notation of Martino and Riebler (2019). We also have the distribution of the latent field $\boldsymbol{x}$,

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x},\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \propto \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x},\boldsymbol{\theta}), \tag{5}$$

where the expression of the marginal likelihood $\pi(\boldsymbol{y})$ is avoided for similar reasons as in Chapter **??**. We can define the joint distribution of the latent effects $\boldsymbol{x}$ and the parameters $\boldsymbol{\theta}$ as the product of the distribution of latent field conditioning on the hyperparameters $\boldsymbol{x}|\boldsymbol{\theta}$ and the distribution of the hyperparameters $\boldsymbol{\theta}$, i. e.

$$\pi(\boldsymbol{x},\boldsymbol{\theta}) = \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

As $\boldsymbol{x}$ is assumed to be or be close to a GMRF, we can write the posterior distribution of the latent field as

$$\pi(\boldsymbol{x}|\boldsymbol{\theta}) \propto \sqrt{\det(\boldsymbol{Q}(\boldsymbol{\theta}))}\exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})x\right\}.$$

Inserting this into Equation 5 together with the assumption that $\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_i \pi(y_i|x_i,\boldsymbol{\theta})$ for conditionally independent observations $\boldsymbol{y}$ conditional on the latent field $\boldsymbol{x}$ and the hyperparameters

$\boldsymbol{\theta}$, we get the posterior distribution of the latent effects to be

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\sqrt{\det(\boldsymbol{Q}(\boldsymbol{\theta}))}\exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})x + \sum_i \log \pi(y_i|x_i, \boldsymbol{\theta})\right\},$$

where $\prod_i \pi(y_i|x_i, \boldsymbol{\theta}) = \exp\{\sum_i \log \pi(y_i|x_i, \boldsymbol{\theta})\}$. INLA estimates the posterior marginals of the latent effects and hyperparameters and not the posterior distribution directly. We extract these marginals by estimating

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \tag{6}$$

for any latent effect x_i, and

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j} \tag{7}$$

for any hyperparameter θ_j. Defining $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ as an approximation for the joint posterior of $\boldsymbol{\theta}$, Rue *et al.* (2009) suggests this approximation to be

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})},$$

where $\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})$ is the joint distribution of the latent field, hyperparameters and observations, $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is a Gaussian approximation of the distribution of the latent effects conditional on the hyperparameters $\boldsymbol{\theta}$ and the observations $\boldsymbol{y}$, and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode of a full conditional distribution of $\boldsymbol{x}$ for a given $\boldsymbol{\theta}$. The approximated distribution of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ can then be directly used to approximate marginal posteriors for θ_j inserting it as the integrand in Equation 7. A similar approach accounts for posterior marginals for x_i, replacing $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ by an approximation $\tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ in Equation 6. The approximation becomes sufficiently accurate using Laplace approximation according to Rue *et al.* (2009).

When approximating marginal posterior distributions of latent effects as in Equation 7, INLA applies the approximation

$$\pi(x_i|\boldsymbol{y}) \approx \sum_{k=1}^{K} \tilde{\pi}(x_i|\boldsymbol{\theta}^{(k)}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}^{(k)}|\boldsymbol{y})\Delta_k$$

of the integral in Equation (6). Here the weights Δ_k are associated with the subset of hyperparameters $\boldsymbol{\theta}^{(k)}$ used for the numerical integration. INLA approximates the integration points by establishing a grid around the posterior mode or a central composite design centered at the posterior mode, described in Box and Darper (1987).

In Rue *et al.* (2009), approximation for the distribution $pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ are suggested using Guassian approximation, Laplace approximation and simplified Laplace approximation. Gaussian approximation approximate the mean and variance of x_i as functions of theta, and is computationally cheap as the Gaussian approximation of $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is computed when computing $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ and $\tilde{\pi}_G(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ is obtained thorugh simple marginalizing. Approximating $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ using Laplace approximation is more computationally expensive than the Guassian approximation, and gives the approximation

$$\pi_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \left.\frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^*{}_{-i}(x_i, \boldsymbol{\theta})},$$

where $\tilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation of $\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y}$ with corresponding mode $\boldsymbol{x}^*{}_{-i}(x_i, \boldsymbol{\theta})$. As this must be evaluated at each x_i it is significantly more computer intensive than the Gaussian approximation, and to deal with this, Rue *et al.* (2009) sugests a simplified approximation, where

$$\pi_{LA} \propto e^{S(x_i)}N(\mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})),$$

where $S(x_i)$ represents a cubic spline around x_i such that the Laplace approximation is approximated to the Gaussian approximation corrected by the cubic spline around each x_i. The simplified Laplace approximation, denoted $\tilde{\pi}_{SLA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ and uses the estimated mean of x_i, $\mu_i(\boldsymbol{\theta})$, as a center for series expanding $\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, such that the parameters of the Gaussian approximation effectively gets corrected for each x_i.

3.4 Accounting for model discrepancy as a temporal random effect

We explain deviations in the model with no attached explainable data by a functional discrepancy function $\delta(t)$. Using a latent Gaussian model approach, the temporal discrepancy functioncan be accounted for as a random effect denoted $f_{k,j(i)}$ in the latent Gaussian model, written as a generalized linear mixed model in Equation 4, as described by Fahrmeir and Tutz (2001). In Chapter 3.2, it was suggested that the random effect could follow a random-walk process, an autoregressive process, an independent identical distribution, or others. Assuming $f_{k,j(i)} = \delta(t)$ a temporal discrepancy following a random walk 1 (RW1) process, we have the relation

$$\delta(t) - \delta(t-1) \sim N(0, \sigma_\delta^2), \tag{8}$$

where σ_δ^2 is the variance of the random walk process and is considered as a non-Gaussian hyperparameter of the model. If other processes are assigned to the temporal discrepancy, the parameters of the processes are also assigned to the set of non-Gaussian hyperparameters, similarly as σ_δ^2 for the discrepancy $\delta(t)$ with a RW1 prior.

3.5 Prior distributions

For the latent Gaussian model, we want to specify priors for the intercept denoted β_0 in the expression for the latent Gaussian model in Equation 4, the coefficients for the linear effects denoted β_j, the hyperparameters of the random effect and the variance of the noise in the observed likelihood data σ_ϵ^2. We assume that the intercept β_0 and the coefficients of the linear effects β_j are Gaussian distributed, such that the eventual priors to assign to them are their prior mean and their prior precision.

3.5.1 Vague prior distributions

We note that we can in practice choose between fitting an informative prior, or applying an uninformative prior when fitting the latent Gaussian model. When using INLA by Rue *et al.* (2009) in R, as described in Chapter 3.7, we can correspondingly choose between fitting informative priors or applying the default priors, which we refer to as vague priors.

The vague priors for the intercept and the coefficients of the linear effects are Gaussian distribution with zero mean and high variance, typically around 10^3.

Considering one random effect, assuming it can be modelled as a random walk 1 process as in Chapter 3.4, we have one non-Gaussian hyperparameter σ_δ^2. Additionaly we have one non-Gaussian hyperparameter form the variance of the noise term in the observed likelihood data in σ_ϵ^2. The vague prior for these hyperparameters is the inverse Gamma distribution over sigma, given as

$$\frac{1}{\sigma^2} \sim Gamma(\alpha, \beta) \tag{9}$$

with shape α and rate β, as described in Moraga (2019). If one, on the other hand, has prior information of the range of these variances, there are several ways to implement this prior information. One way could be to define the parameters in the default hyperprior, or one could implement the prior knowledge through other prior distrubution.

3.5.2 The delta method

When assigning informative prior distributions to the Gaussian model parameters as the model intercept and the coefficients of the linear effects, the priors fitted are the means and the variances of the model parameters. The prior means can be decided as the expected value of each parameter based on observed data and field experts opinions. The variances can also be decided from these, but when the parameters are functions of observed and unobserved physical parameters, we need a method to extract the variances of the parameters.

When we have an expression of several random variables, say $Y = f(X_1, ..., X_n)$, where every X_i has a variance $\text{Var}[X_i]$. Assume we want to find the variance of Y, $\text{Var}[Y]$. Using the formulation of Klein 1953, we write

$$\text{Var}[Y] = \sum_{i=1}^{n} \left(\frac{\partial Y}{\partial X_i} \right)^2 \text{Var}[X_i] + \sum_{i=1}^{n} \sum_{j \neq i} \left(\frac{\partial Y}{\partial X_i} \right) \left(\frac{\partial Y}{\partial X_j} \right) \text{Cov}[X_i, X_j]. \tag{10}$$

This is useful when we want to apply the variance of a parameter whose eqpression is dependent on several random variables.

3.5.3 Penalized complexity priors

When assigning prior distributions to the non-Gaussian hyperparameters, several priors can be chosen, which of some are preented in Gómez-Rubio (2020). Penalized complexity (PC) priors, described by Simpson *et al.* (2017), are a type of prior distributions used in Bayesian statistical modeling that are designed to encourage parsimony in models by penalizing more complex models. They are constructed based on principles of inference that favor the base model unless evidence is provided against it. It measures the distance from the base model using the Kullback-Leibler distance, and penalizes departure from the base model at a constant rate on this distance. PC priors are defined using probability statements on the model parameters and are particularly well suited for use in additive models defined by different components, such as latent effects.

The PC prior takes in two arguments u and α, and is given for the standart deviation σ as

$$Pr(\sigma_\delta > u) = \alpha,$$

for $u > 1$, $\alpha \in (0, 1)$. This means that one is able to decide the approximate magnitude of the cariance of the variances. To approximate the model discrepancy as precise as possible, one want to make sure that most of the noise is captured in the noise term of the observed likelihood. In such a case, one want to assign a loose prior for the hyperparameter σ_ϵ^2 and a strict prior for σ_δ^2.

3.6 The deviance information criterion

When having models which seems to fit the data equally well, we want a formal metric to evaluate the performance between models. The deviance information criterion (DIC) was introduced by Spiegelhalter *et al.* (2002), and is a measure of the fit of a Bayesian statistical model. It is used to compare the fit of different models and to choose the model that provides the best fit to the data. The DIC is defined as the mean deviance of the fitted model plus a complexity penalty. The deviance $D(\theta)$ of a fitted model is a measure of how well the model fits the data. Mathematically, using the notation of Spiegelhalter *et al.* (2002), the deviance or "Bayesian deviance" is given as

$$D(\theta) = -2 \log\{p(y|\theta)\} + 2 \log\{f(y)\},$$

where $p(y|\theta)$ is the model likelihood and $f(y)$ is some standardizing term which is fully specified as a function of the data alone. The DIC is derived through the number of effective parameters in the model p_D, which in Spiegelhalter *et al.* (2002) is given as

$$p_D = \overline{D(\theta)} - D(\overline{\theta}),$$

where $\overline{D(\theta)}$ denotes the expected value of the deviance and $\overline{\theta}$ denotes the expected value of θ. The DIC is thus given as

$$DIC = p_D + \overline{D(\theta)} = D(\overline{\theta}) + 2p_D.$$

The complexity penalty is a term that penalizes more complex models, similar to the way that the Akaike information criterion (AIC) by Akaike (1973) or the Bayesian information criterion (BIC) by Schwartz (1978) do. The DIC can be used to compare the fit of different models, with lower values of DIC indicating a better fit. It is a widely used criterion for model selection in Bayesian statistics, and is particularly useful for comparing models with different numbers of parameters. One disadvantage of the DIC is that it is not always straightforward to calculate, particularly for models with complex likelihood functions or models with latent variables. In these cases, approximations to the DIC may be used.

3.7 Key aspects from the R-INLA package

The r-inla package provides a suite of functions for fitting Bayesian models using INLA in R. It includes functions for model specification, model fitting, and model evaluation, as well as a variety of diagnostic plots and other tools for exploring the fitted models. A maunal of the package is given in Martino and Rue (2009). This chapter contains some take-aways from the R-INLA package applied in this thesis. A INLA object in R is defined as a function with several abilities of function specifications. An example of a defined inla object is written as follows.

```
inla.model = inla(formula = formula, data = data,
                  control.fixed = prior.fixed, family = "gaussian",
                  control.family = list(hyper = prior.likelihood),
                  control.predictor(compute = TRUE),
                  control.compute = list(dic = TRUE)))
```

The `formula` object specifies the latent Gaussian model in terms of the linear predictor given in Equation 4. This object may be specified as follws.

```
formula = y ~ 1 + x1 + x2 + f(ID, model = "rw1", hyper = prior.random,
                               constr = FALSE)
```

In the given terms, `y` is the observed data fo the linear predictor, `1` is represented as the interccept, `x1` and `x2` the data of the observed linear effects. The function `f()` represents a random effect. The argument `ID` represents the index of the observations the `model` represents the name of the model, and can for instanke be `"iid"`, `"ar1"`, or as in this case `"rw1"`. The `"hyper"` argument specifies the hyperprior of the model, and in case of the `"rw1"` process, this is the assumed prior distribution of the non-gaussian parameter σ_δ^2 from the expression of a random walk 1 process given in Equation 8. Unless another prior is specified, the default prior distribution is given as an inverse gamma prior as described in Chapter 3.5 with $\alpha = 1$ and $\beta = 5 \cdot 10^{-5}$. One might fit other priors to the random effect by defining the argument `hyper` to be as follows.

```
prior.random = list(prec = list(prior = "pc.prec", param = c(1, 0.05)))
```

In this case, we have defined a penalized complexity prior (PC prior) setting constraints on the variance of the random walk hyperparameter. In this case, the PC prior has been defined with $u = 1$ and $\alpha = 0.05$. The argument `constr` is a boolean variable regarding the sum-to-zero constraint of the random effect. When the arguments is set to `TRUE` the posterior random effect sums to zero while when `FALSE` it approaches the true value of the random effect.

The `data` argument in the `inla.model` object is the data the model is fit to as a data frame consisting of the response `y` and the covariates of the linear effects `x1` and `x2` in the considered case. The `control.fixed` arguments lets us fit priors to the Gaussian distributed intercept and to the Gaussian distributed coefficients of the linear effects through defining prior mean and precisions for the coefficients. When this is not specified, the default option is fitted. This assigns a gaussian prior to the intercept with mean equal to zero and a precision equal to zero, indicating a variance approaching infinity. For the coefficients of the linear effects, a gaussian prior with mean zero and a precision of 0.001 is fitted. When assigning a prior to the intercept and the coefficients of the linear effects, the `prior.fixed` object may look like follows.

```
prior.fixed = list(mean.intercept = mu0, prec.intercept = tau0,
                   mean = list(mu1, mu2), prec = list(tau1, tau2))
```

The `family` argument specifies the likelihood family, and may take inputs like `"poisson"`, `"binomial"`, or as in this case `"gaussian"` which also is the default alternative. The `control.family` arguments lets us specify the prior distribution of the variance of the likelihood through what we have written as `prior.likelihood`, where the default is similar as for `prior.random`, and other priors can be specified in a similar way as for `prior.random`. The entry `control.predictor(cumpute = TRUE)` states that the marginal densities for the linear predictor should be computed, and the function also enables us to define an eventual link function, if present. To decide which diagnostic statistics to compute the `control.compute` takes input on which statistics to compute, as the deviation information criterion (DIC).

4 Modelling substance accumulations as latent Gaussian models

We have so far a physical case considering the relationship between consumed oxygen and accumulated carbon dioxide by salmon in a land-based production facility. As well, we have introduced a framework on how to perform Bayesian calibration of latent Gaussian models using physics-informed and expert-informed prior distributions for Gaussian and non-Gaussian model parameters. This chapter combines the data and assumptions with the theory of Bayesian inference and latent Gaussian models to explore relations between oxygen and carbon dioxide, and to which extent prior information can improve the models.

4.1 Expressing the physical case as latent Gaussian models

We set up the model for what is general for what is general for oxygen consumption and carbon dioxide accumulation. We then derive each of the processes as latent Gaussian models. We use the notations from Chapter 2, but we rewrite the observed substance accumulation

$$\boldsymbol{g}(t) = [g_{O_2}(t), g_{CO_2}(t)]^T$$

to a modelled discrepancy function

$$\boldsymbol{\delta}(t) = [\delta_{O_2}(t), \delta_{CO_2}(t)]^T.$$

We assume the change in amount of substance form time $t-1$ to time t is the difference between the substance added to the tank and the substance extracted from the tank. This corresponds to

$$V\boldsymbol{y}(t) - V\boldsymbol{y}(t-1) = \boldsymbol{m}_{in}(t) - \boldsymbol{m}_{out}(t) + \boldsymbol{\delta}(t) + \boldsymbol{\epsilon}, \tag{11}$$

where

$$\boldsymbol{y}(t) = [y_{O_2}(t), y_{CO_2}(t)]^T$$

is the measured oxygen concentration at time t, and

$$\boldsymbol{m}_{in}(t) = [m_{in,O_2}(t), m_{in,CO_2}(t)]^T,$$

and

$$\boldsymbol{m}_{out}(t) = [m_{out,O_2}(t), m_{out,CO_2}(t)]^T$$

are the amounts of substance transported in and out of the tank by the water at time t, respectively, and $\boldsymbol{\delta}(t)$ the accumulated substance amounts. The term $\boldsymbol{\epsilon} = [\epsilon_{O_2}, \epsilon_{CO_2}]$ represents the noise in each of the models, where $\epsilon_{O_2} \sim N(0, \sigma^2_{\epsilon,O_2})$ and $\epsilon_{CO_2} \sim N(0, \sigma^2_{y,CO_2})$

4.1.1 Oxygen concentration as a latent Gaussian model

From Section 2.1, we obtain that the oxygen entering the tank is the oxygen transported in with the water and the amount injected from the oxygen injectors. Thus we can define

$$m_{in,O_2}(t) = q_f c_{f,O_2} + q_r c_{r,O_2} + h(t).$$

The oxygen let out of the tank by the outlet water at time t can be eqpressed as

$$m_{out,O_2}(t) = q y_{u,O_2}(t).$$

Inserting the obtained expressions for $m_{in,O_2}(t)$ and $m_{out,O_2}(t)$ into Equation 11, we obtain for oxygen

$$Vy_{O_2}(t) - Vy_{O_2}(t-1) = q_f c_{f,O_2} + q_r c_{r,O_2} + h(t) - q y_u(t) + \delta(t) + \epsilon_{O_2},$$

where $\epsilon_{O_2} \sim N(0, \sigma_y^2)$. Solving for $y_{O_2}(t)$, setting $y_{O_2}(t-1) = y_{u,O_2}(t)$ as assumed in Section 2.4, and collecting with respect to temporal variables, we get

$$y(t) = \frac{c_{r,O_2} q_r + c_{f,O_2} q_f}{V} + \left[1 - \frac{q}{V}\right] y_{u,O_2}(t) + \frac{1}{V} h(t) + \frac{1}{V} \delta_{O_2}(t) + \epsilon_{O_2}. \tag{12}$$

The quantity V is the volume of the tank.

We now formalize this model as a latent Gaussian models as in Equation 4. We use the identity link function, such that the linear predictor $\eta_{O_2}(t) = E[y_{O_2}(t)]$. We can thus rewrite Equation 12 as

$$\eta_{O_2}(t) = \beta_{0,O_2} + \beta_{1,O_2} y_{u,O_2}(t) + \beta_{2,O_2} h(t) + \delta_{O_2}(t), \tag{13}$$

where

$$\beta_{0,O_2} = \frac{c_{r,O_2} q_r + c_{f,O_2} q_f}{V}$$

is the model intercept, and

$$\beta_{1,O_2} = 1 - \frac{q}{V}$$

and

$$\beta_{2,O_2} = \frac{1}{V}$$

are the coefficients of the linear effects. These are assigned independent Gaussian priors

$$\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}, \Sigma),$$

where

$$\pi(\boldsymbol{\beta}) = [\pi(\beta_0^{O_2}), \pi(\beta_1^{O_2}), \pi(\beta_2^{O_2})]^T$$

are the prior distributions for the linear coefficients, and

$$\boldsymbol{\mu} = [\mathrm{E}[\beta_0^{O_2}], \mathrm{E}[\beta_1^{O_2}], \mathrm{E}[\beta_2^{O_2}]]^T$$

and

$$\Sigma = diag(\mathrm{Var}[\beta_0^{O_2}], \mathrm{Var}[\beta_1^{O_2}], \mathrm{Var}[\beta_2^{O_2}])$$

are the corresponding prior means and prior variances. The discrepancy function $\delta_{O_2}(t)$ is here considered as accumulated amount per liter. This function is assumed smooth, as $\boldsymbol{g}(t)$ seems to fluctuate around an approximately constant value in Figure 10. This is modelled as a random effect following with a RW1 prior such that

$$\delta_{O_2}(t) - \delta_{O_2}(t-1) = \epsilon_{\delta,O_2},$$

where $\epsilon_\delta \sim N(0, \sigma_\delta^2)$. The non-Gaussian hyperparameter σ_δ^2 may then be assigned a hyperprior, or the vague inverse Gamma prior described in Section 3.5.1. The variance of the model noise σ_{y,O_2}^2 is also a non-Gaussian hyperparameter and will be assigned a prior similarly as σ_{δ,O_2}^2.

4.1.2 Carbon dioxide concentration as a latent Gaussian model

We now follow the same analysis as for the oxygen concentration, basing the model derivation on Equation 11. Using the data presented in Chapter 2.1 and the assumptions in Chapter 2.4, we obtain that the carbon dioxide entering the tank from the intake water can be written as

$$m_{in,CO_2}(t) = c_{f,CO_2} q_f + (1 - \xi) q_r y_{u,CO_2}(t).$$

The carbon dioxide removed from the tank through the outtake water can be written as

$$m_{out,CO_2}(t) = q y_u(t).$$

The carbon dioxide produced by the salmon is modelled as $\delta_{CO_2}(t)$. We can thus write the expression for carbon dioxide concentration in the tank as

$$V y_{CO_2}(t) - V y_{CO_2}(t-1) = (1 - \xi) q_r y_{u,CO_2}(t) + q_f c_{f,CO_2} - q y_{u,CO_2}(t) + \delta_{CO_2}(t) + \epsilon_{CO_2},$$

where $\epsilon_{CO_2} \sim N(0, \sigma^2_{y,CO_2})$. Solving for $y_{CO_2}(t)$, setting $y_{CO_2}(t-1) = y_{u,CO_2}(t)$ as assumed in Section 2.4, and collecting each temporal variable for itself, as for the oxygen concentration, we get

$$y_{CO_2}(t) = \frac{q_f c_{f,CO_2}}{V} + \left[1 - \frac{q_f - \xi q_r}{V}\right] y_{u,CO_2}(t) + \frac{1}{V}\delta_{CO_2}(t) + \epsilon_{CO_2}. \tag{14}$$

Expressing this as a latent Gaussian model follows the same assumptions as for the oxygen concentration, with the identity link function. Thus we have

$$\eta_{CO_2}(t) = \beta_{0,CO_2} + \beta_{1,CO_2} y_{u,CO_2}(t) + \delta_{CO_2}(t), \tag{15}$$

where

$$\beta_{0,CO_2} = \frac{q_f c_{f,CO_2}}{V}$$

is the model intercept and

$$\beta_{1,CO_2} = 1 - \frac{q_f - \xi q_r}{V}$$

is the coefficient of one linear effect in this case. These are both assigned Gaussian priors similarly as for oxygen. The discrepancy function also considers accumulated substance per volume in this case, as for the oxygen concentration. Similarly as for the oxygen concentration, $\delta_{CO_2}(t)$ assumed smooth and is considered a random effect with an assigned RW1 prior with a hyperparameter σ^2_{δ,CO_2} which will be assigned a prior. The variance of the noise term in Equation 14, σ^2_{y,CO_2} is also a non-Gaussian hyperparameter is be assigned a non-Gaussian hyperprior similarly as for σ^2_{y,O_2}.

4.2 Model fitting by assigning prior distributions

Having expressed our physical models as latent Gaussian models, we now want to consider what prior information we should include in our model for calibration. We have a latent field consisting of the Gaussian parameters β_{0,O_2}, β_{1,O_2}, β_{2,O_2}, β_{0,CO_2} and β_{1,CO_2}, and the non-Gaussian hyperparameters σ^2_{ϵ,O_2}, σ^2_{δ,O_2}, σ^2_{ϵ,CO_2} and σ^2_{δ,CO_2}.

We thus have the prior distribution

$$\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}, \Sigma)$$

for the fixed effects, where $\pi(\boldsymbol{\beta}) = [\pi(\beta_0^{O_2}), \pi(\beta_1^{O_2}), \pi(\beta_2^{O_2}), \pi(\beta_0^{CO_2}), \pi(\beta_1^{CO_2})]^T$ are the prior distributions for the linear coefficients, and $\boldsymbol{\mu} = [\mathrm{E}[\beta_0^{O_2}], \mathrm{E}[\beta_1^{O_2}], \mathrm{E}[\beta_2^{O_2}], \mathrm{E}[\beta_0^{CO_2}], \mathrm{E}[\beta_1^{CO_2}]]^T$ and $\Sigma = diag(\mathrm{Var}[\beta_0^{O_2}], \mathrm{Var}[\beta_1^{O_2}], \mathrm{Var}[\beta_2^{O_2}], \mathrm{Var}[\beta_0^{CO_2}], \mathrm{Var}[\beta_1^{CO_2}])$ are the corresponding prior means and variances.

The method of calibrating the model will be to initially fit a model with default options containing no prior information about the data nor the model. Furtermore, models with more prior information will be fitted, and the results will be compared to the expected outcome and to each other if they seem til fit the case sufficiently, using the deviance information criterion (DIC).

4.2.1 Initially: Linear regression model

We initially fit a linear regression model to identify the necessity of accounting for random effect in our model. The linear regression model leaves the discrepancy function $\boldsymbol{\delta}(t)$ out in the modelling of the oxygen and carbon dioxide concentrations in Equation 13 and Equation 15. The models for oxygen and carbon dioxide can thus respectively be expressed as

$$\eta_{O_2} = \beta_{0,O_2} + \beta_{1,O_2} y_{u,O_2}(t) + \beta_{2,O_2} h(t)$$

and

$$\eta_{CO_2} = \beta_{0,CO_2} + \beta_{1,CO_2} y_{u,O_2}(t).$$

4.2.2 Model 0: Flat priors

Model 0 for both oxygen concentration and carbon dioxide concentration is fitted not assigning any prior information to the data, keeping the vague priors. This assigns default Gaussian priors to the intercepts and the coefficients of the fixed effects, where the intercepts are assigned mean and precision zero and the coefficients of the linear effects are assigned mean zero and precision 0.001, systematized in Table 5. The hyperparameters σ^2_{y,O_2}, σ^2_{δ,O_2}, σ^2_{y,CO_2} and σ^2_{δ,CO_2} are all assigned the same inverse gamma prior, as in equation 9, whith shape parameter $\alpha = 1$ and rate parameter $\beta = 5 \cdot 10^{-5}$.

Table 5: Prior mean and variance on linear effects in Model 0.

Coefficient	Estimating	Mean	Variance	Precision
β_{0,O_2}	$(c_{r,O_2}q_r + c_{f,O_2}q_f)/V$	0	$\approx \infty$	0
β_{1,O_2}	$1 - q/V$	0	1000	0.001
β_{2,O_2}	$1/V$	0	1000	0.001
β_{0,CO_2}	$q_f c_{f,CO_2}/V$	0	$\approx \infty$	0
β_{1,CO_2}	$1 - (q_f - \xi q_r)/V$	0	1000	0.001

4.2.3 Model 1: Gaussian priors for intercepts and coefficients of linear effects

Model 1 introduces informative priors chosen for the intercept and the coefficients of the linear effects. We used the derived expressions of these estimates from Section 4.1.1 and Section 4.1.2. Model 1 assumes that all other parameters than the water flows q, q_f and q_r are constant, such that the mean and variances are estimated in terms of the empirical means and variances for the data obtained from the water flows. As we obtain from Figure 4 and Figure 9 that the flow rates appear constant with relative little noise, the prior variances in Model 1 are relatively strict. The prior Gaussian means and variances for the intercepts and the coefficients of the linear effects are systematized with two current digits in Table 6. The hyperpriors for the non-Gaussian parameters are kept as the default as in Model 0.

Table 6: Prior mean and variance on linear effects in Model 1.

Coefficient	Estimating	Mean	Variance	Precision
β_{0,O_2}	$(c_{r,O_2}q_r + c_{f,O_2}q_f)/V$	0.091	$5.1 \cdot 10^{-8}$	$2.0 \cdot 10^{7}$
β_{1,O_2}	$1 - q/V$	0.99	$6.7 \cdot 10^{-10}$	$1.5 \cdot 10^{9}$
β_{2,O_2}	$1/V$	$2.0 \cdot 10^{-7}$	0	$\approx \infty$
β_{0,CO_2}	$q_f c_{f,CO_2}/V$	0.0055	$6.7 \cdot 10^{-10}$	$1.5 \cdot 10^{9}$
β_{1,CO_2}	$1 - (q_f - \xi q_r)/V$	0.99	$6.6 \cdot 10^{-10}$	$1.5 \cdot 10^{9}$

4.2.4 Model 2: Adding PC priors to the hyperparameters

Model 2 introduces chosen PC priors described in Chapter 3.5.3 for the non-Gaussian hyperparameters σ^2_{y,O_2}, σ^2_{δ,O_2}, σ^2_{y,CO_2} and σ^2_{δ,CO_2}. We want to reduce the cariance of the RW1 priors for the discrepancy functions assigning these priors. We thus choose a strict prior for $\boldsymbol{\sigma}^2_\delta$ and a less strict prior for the parameter $\boldsymbol{\sigma}^2_y$, where $\boldsymbol{\sigma}^2_y = [\sigma^2_{y,O_2}, \sigma^2_{y,CO_2}]$ and $\boldsymbol{\sigma}^2_\delta = [\sigma^2_{\delta,O_2}, \sigma^2_{\delta,CO_2}]$. We choose PC priors for all the hyperparameters, with $u = 1$ and $\alpha = 0.01$ for the parameters in $\boldsymbol{\sigma}^2_\delta$ and with $u = 1$ and $\alpha = 0.99$ for the parameters in $\boldsymbol{\sigma}^2_\epsilon$.

4.2.5 Model 3: Gaussian priors using the delta method

We now assess the variances of the intercepts and the priors of the linear effects, to take uncertainties in assumed quantities into consideration. Using expert information, we can assume 2%

measuring uncertainty in the sensors measuring the flow rates, 2% uncertainty in the constants c_{r,O_2}, c_{f,O_2}, c_{f,CO_2} and 10% uncertainty in the effect of the carbon dioxide degasser ξ. When the uncertaity is refered to in per cent, standard deviation is the percentage of the mean value. Assuming all the additional parameters are independent, we use the Delta method given in Chapter 3.5.2 to derive the expressions for the variances the intercepts and the coefficients of the linear effects. The expression variance of β_{0,O_2} is through the delta method calculated to be

$$\text{Var}[\beta_{0,O_2}] = \frac{1}{V^2}\left(q_r^2\text{Var}[c_{r,O_2}] + c_{r,O_2}^2\text{Var}[q_r] + q_f^2\text{Var}[c_{f,O_2}] + c_{f,O_2}^2\text{Var}[q_f]\right).$$

The expression for the variance of β_{1,O_2} and β_{2,O_2} remains the same as before. The variance of β_{0,CO_2} becomes

$$\text{Var}[\beta_{0,O_2}] = \frac{1}{V^2}\left(q_f^2\text{Var}[c_{f,CO_2}] + c_{f,CO_2}^2\text{Var}[q_f]\right).$$

Finally, the expression of the variance of β_{1,CO_2} can be written

$$\text{Var}[\beta_{0,O_2}] = \frac{1}{V^2}\left(\text{Var}[q_f] + \xi^2\text{Var}[q_r] + q_r^2\text{Var}[\xi]\right).$$

The numerical values of the prior Gaussian parameters of Model 3 are systematized in Table 7. As we see, introduced uncertainties increases the variance in the prior distributions, but the variances are still small.

Table 7: Prior mean and variance on linear effects in Model 3.

Coefficient	Estimating	Mean	Variance	Precision
β_{0,O_2}	$(c_{r,O_2}q_r + c_{f,O_2}q_f)/V$	0.091	$3.34 \cdot 10^{-6}$	$3.0 \cdot 10^5$
β_{1,O_2}	$1 - q/V$	0.99	$4.5 \cdot 10^{-8}$	$2.2 \cdot 10^7$
β_{2,O_2}	$1/V$	$2.0 \cdot 10^{-7}$	0	$\approx \infty$
β_{0,CO_2}	$q_f c_{f,CO_2}/V$	0.0055	$2.4 \cdot 10^{-8}$	$4.1 \cdot 10^7$
β_{1,CO_2}	$1 - (q_f - \xi q_r)/V$	0.99	$1.3 \cdot 10^{-7}$	$7.5 \cdot 10^6$

The priors for the non-Gaussian hyperparameters may be tuned according to DIC and the model fit to the initial data. after having fitted Model 3.

4.3 Bayesian inference on latent Gaussian model using integrated nested Laplace approsimation

Having defined different models for oxygen and carbon dioxide concentrations, accounting for different amounts of prior information, we fit the models to the data presented in Chapter 2.3. The models are fitted using the R-INLA package presented in Chapter 3.7, performing INLA methodology described in Chapter 3.3. We evaluate the posterior distributions of the fixed effects β_j, and their posterior discrepancy functions δ_j. Additionally, we evaluate the sum of the discrepancy functions, considering them as two independent random variables when calculating the means and the confidence bounds. Lastly, we evaluate the posterior distribution of the model uncertainties $\boldsymbol{\sigma}_y$. The models are primarily evaluated on their fit to the data and assumptions in Chapter 2, but if several models are performing similarly well, they are compared based on DIC, described in Chapter 3.6.

5 Results

This chapter presents the posterior distributions of the linear model, Model 0, Model 1, Model 2 and Model 3 defined in Chapter 4.2. Relevant comparisons between the models, and between the models and the data and assumptions presented in Chapter 2.4 will also be displayed and discussed. The results are reported parameterwise through visualizations and comparisons.

5.1 Posterior Gaussian parameters

We will now assess the posterior intercept and fixed effect coefficients for oxygen and carbon dioxide of the linear model, Model 0, Model 1, Model 2 and Model 3. We will compare the models to each other and evaluate how they fit with initial assumptions.

5.1.1 Posterior parameters for the linear models with vague priors

The posterior intercept $\beta_0^{O_2}$ and coefficients for the linear effects $\beta_1^{O_2}$ and $\beta_1^{O_2}$ for oxygen for the linear model witout a random effects, and with a vague assigned prior are plotted and compared to the calculated values of these coefficient with respect to the parameter assumptions in Figure 12. The posterior intercept and linear effect coefficient for carbon dioxide, $\beta_0^{CO_2}$ and $\beta_0^{CO_2}$, are plotted and compared to the computed values from the parameter assumptions in Figure 13.

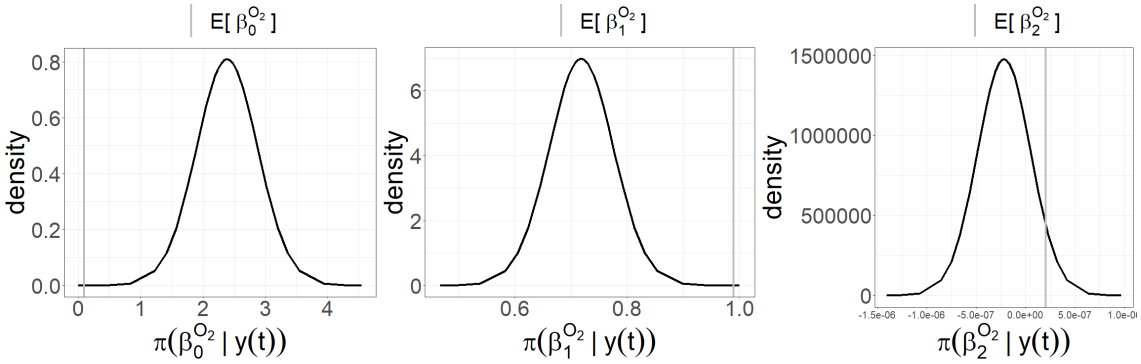

Figure 12: Plotted posterior densities of the fixed effect coefficients for the linear model for oxygen, together with the expected prior values of each parameter.

We see from Figure 12 that the posterior estimate of $\beta_0^{O_2}$, $\pi(\beta_0^{O_2}|y(t))$, is considerably larger than the calculated value from the parameter assumptions. For the linear model for oxygen, we stated in Chapter 4.2 that the oxygen concentration in the tank at time t was equal to the intercept $\beta_0^{O_2}$ plus the oxygen concentration at $t-1$ weighted by $\beta_1^{O_2}$. Additionally the added oxygen $h(t)$ is weighted by $\beta_2^{O_2}(t)$. Taking the model fitted into consideration, the linear model balances the equation

$$y_{O_2}(t) = \beta_0^{O_2} + \beta_1^{O_2} y_{u,O_2}(t) + \beta_2^{O_2} h(t),$$

where the expected parameter values are calculated to account for a negative effect on the right hand side for negatively accumulated oxygen, the linear model has compensated for this effect by weighting the parameters in another way. While the intercept $\beta_0^{O_2}$ is calculated to have an expected value close to zero, the linear model weights more of the oxygen concentration to this value, which is mainly ranging between 1.5 and 3.5. The calculated model assigns the expected value of $\beta_1^{O_2}$ to be close to 1, while the linear model estimates this posterior parameter to mainly range between 0.6 and 0.8. As $y_{O_2}(t)$ and $y_{u,O_2}(t)$ are assumed to be equal as described in Equation 1 and evident from Figure 8 in Chapter 2.4, it is a reasonable estimate from the calculated model to state that $y_{O_2}(t) \approx y_{u,O_2}(t)$. From the linear model, y_{u,O_2} (t) is assigned less weight, but as y_{O_2} seems to fluctuate around an approximately constant value, it is reasonable to assign more of the value to the intercept. The parameter for the added oxygen $h(t)$, $\beta_2^{O_2}$, is of approximately the same absouute value as its calculated expected value, but with a negative sign. Note that this is considered as a parameter even if its value is a known constant $1/V$, to let $h(t)$ be treated as the other model parameters. Its negative value might be explained by that the model does not account for the negatively accumulated oxygen which in reality takes place in the tank, making $\beta_2^{O_2}$ explain this effect. The posterior variances of the model are most probably fitted by the noise in the data when vague priors for all model parameters and hyperparameters are fitted.

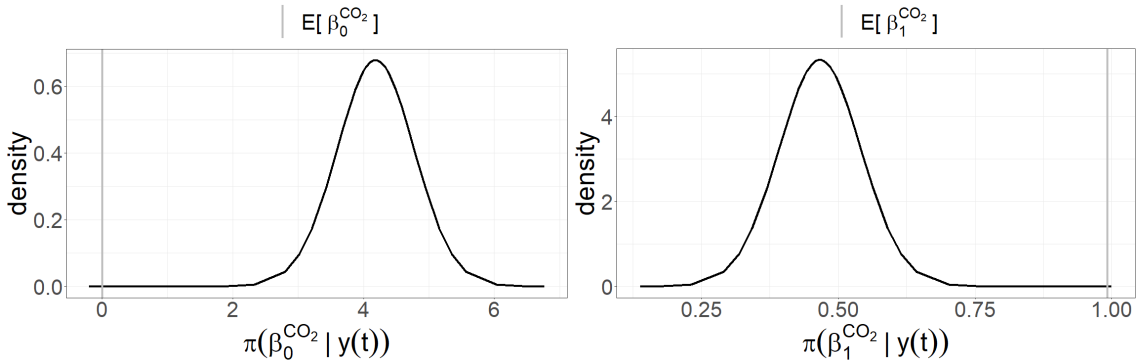

Figure 13: Plotted posterior densities of the fixed effect coefficients for the linear model for carbon dioxide, together with the expected prior values of each parameter.

Figure 13 shows that the calculated values from the parameter assumptions are close to zero for $\beta_0^{CO_2}$ and close to one for $\beta_1^{CO_2}$. Considering the linear model fitted for the carbon dioxide concenration, whose posterior distribution of the model parameters are plotted in Figure 13, the model is assigned to balance the equation

$$y_{CO_2}(t) = \beta_0^{CO_2} + \beta_1^{CO_2} y_{u,CO_2}(t).$$

It is evident from Figure 13 that the expected values are calculated to have the intercept $\beta_0^{CO_2}$ close to zero and $\beta_1^{CO_2}$ close to one, assigning the posterior discrepancy with a relatively small value compared to the value of $y_{CO_2}(t)$. This is however reasonable as the posterior discrepancy is multiplied by a factor $1/V$. The posterior estimates of de linear model does however deviate considerably from the calculated values, with the intercept $\beta_0^{CO_2}$ ranging mainly between 3 and 5 and the coefficient of $y_{u,CO_2}(t)$, $\beta_1^{CO_2}$, ranging mainly between 0.3 and 0.6. Figure 5 shows that the carbon dioxide concentration seem to fluctuate around a relatively constant level, such that more of the data may be captured by the intercept in the linear model. The model does not, as for the oxygen model, account for accumulated carbon dioxide, which also causes the deviantions from the expected values in the model. The posterior variances can also as for the oxygen model be assumed to accunt for noise in the data.

The results of the posterior parameter estimates, plotted in Figure 12 and Figure 13, indicate that they significantly differ from their expected values based on the data and the assumed parameters. This is expected, as the model assumes that the presence of salmon in the tank leads to a negative accumulation of oxygen and a positive accumulation of carbon dioxide. None of the posterior distributions from the linear model seem to fit to the data. It will therefore be reasonable to introduce a random effect accounting for negatively accumulated oxygen and accumulated carbon dioxide in the model as for Model 0 presented in Chapter 4.2.

5.1.2 Posterior parameters for the random effect models with vague priors

Figure 14 shows plots of the posterior distributions of the intercept and the linear effect parameters $\beta_0^{O_2}$, $\beta_1^{O_2}$ and $\beta_2^{O_2}$ for the oxygen consumption modelled by Model 0, together with calculated values of the parameters based on parameter assumptions. Figure 15 shows plots of the posterior distributions for the intercept and the linear effect parameter $\beta_0^{CO_2}$ and $\beta_1^{CO_2}$ for the carbon dioxide concentration modelled by Model 0, together with calculated parameter values based on the parameter assumptions. Recall that Model 0 is fitted with a random effect for oxygen and carbon dioxide accumulation, and all parameters and hyperparameters has vague priors.

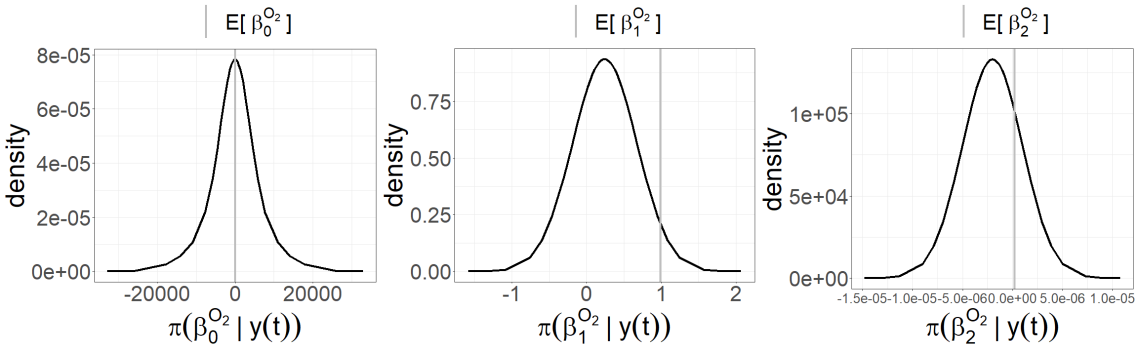

Figure 14: Plotted posterior densities of the fixed effect coefficients for the Model 0 for oxygen, together with the expected prior values of each parameter.

Assessing the posterior results of the Gaussian parameters for Model 0, containing a random effect assigned a RW1 prior with a vague variance hyperparameter, and vague priors for the linear coefficients and the variance hyperparameter of the model noise, the results are evident in Figure 14. The figure shows that the intercept takes its posterior mean value relatively close to zero and the calculated expected value, compared to the posterior variance of the parameter. Compared to the linear model, the intercept for the oxygen model, $\beta_0^{O_2}$ has a significantly higher posterior variance, and seems to have absorbed a significant amount of the total model noise. The posterior mean and variance of the coefficient for the oxygen concentration outside the tank, $\beta_1^{O_2}$, are closer to zero in Model 0 compared to the linear model. However, the posterior variance is larger, which means that the expected value of the parameter falls within reasonable confidence bounds of the posterior distribution, even though it is farther from the calculated expected value. As vague priors are fitted with zero mean and a high variance, the probability of the posterior values to deviate from the calculated expected values is present, which is evident for $\beta_1^{O_2}$ for Model 0. The posterior distribution for the coefficient of $h(t)$, $\beta_2^{CO_2}$ is negative and has a higher absolute value for Model 0 than for the linear model. The uncertainty is as for the other parameters of Model 0 large enough to make the calculated expected value to be within reasonable confidence bounds of the model. The negative value of this coefficinent implies that the variable for added oxygen is negative removed oxygen from the tank.

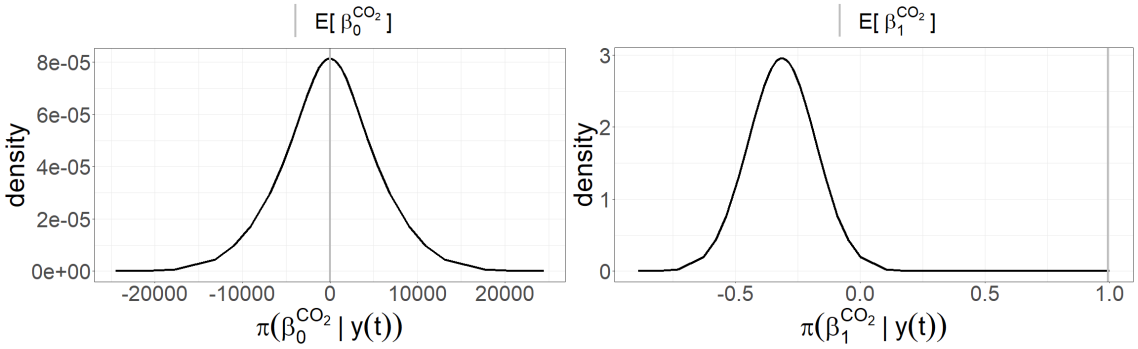

Figure 15: Plotted posterior densities of the fixed effect coefficients for Model 0 for carbon dioxide, together with the expected prior values of each parameter.

We see form Figure 15 that as for the oxygen concentration, that the posterior distribution for $\beta_0^{CO_2}$, $\pi(\beta_0^{CO_2}|y(t))$ seems to have a mean coinciding with the calculated value of $\beta_0^{CO_2}$ based on the parameter assumptions. However, the variance is also large for $\beta_0^{CO_2}$, as for the oxygen concentration. The posterior distribution of $\beta_1^{CO_2}$, $\pi(\beta_1^{CO_2}|y(t))$, has a negatve mean at approximately $\beta_1^{CO_2} \approx -0.3$, and the calculated value for $\beta_1^{CO_2}$close to 1 does not seem to lie within any reasonable credible set of the posterior distribution.

Overall, introducing a random effect for the model does not seem to make the posterior distribution of the Gaussian parameters to fit perfectly with their calculated values based on the parameter assumption. Thus Model 0 described in Chapter 4.2 does not seem to fit with the data and assumptions. More of the posterior distributions seems to have reasonable credible sets containing the calculated parameter values, but the posterior means seems to deviate from the calculated estimates. It is therefore reasonable to fit prior means for the Gaussian parameters and the non-Gaussian hyperparameters of the model, as in Model 1, Model 2 and Model 3 presented in Chapter 4.2.

5.1.3 Posterior parameters with assigned Gaussian and PC priors

Recall that Model 1 has assigned Gaussian priors to the Gaussian parameters $\beta_0^{O_2}$, $\beta_1^{O_2}$ and $\beta_2^{O_2}$ for the oxygen concentration model, and $\beta_0^{CO_2}$ and $\beta_1^{CO_2}$ for the carbon dioxide concentration model. The priors are fitted with a mean equal to the calculated value of the parameters based on the data and parameter assumptions and a variance decided by the noise in the measured sub-parameters in each parameters. Model 2 has the same Gaussian priors as Model 1, but has assigned penalized complexity (PC) hyperpriors to the non-Gausian hyperparameters σ_δ^2 and σ_y^2. Recall again that σ_δ^2 is the variance of the RW1 prior for the random effect $\delta(t)$ and σ_y^2 is the variance of the model noise ϵ_y. Model 3 has the same PC prior as Model 2, but the Gaussian prior variances are updated to assuming that the assumed sub-parameter values having uncertainties. The posterior distributions of Model 1, Model 2 and Model3 for the oxygen concentration are plotted parameterwise for $\beta_0^{O_2}$, $\beta_1^{O_2}$ and $\beta_2^{O_2}$ in Figure 16, Figure 17 and Figure 18, respectively. The posterior distributions of Model 1, Model 2 and Model 3 for the carbon dioxide concentration are plotted parameterwise for $\beta_0^{CO_2}$ and $\beta_1^{CO_2}$ in Figure 19 and Figure 20, respectively.

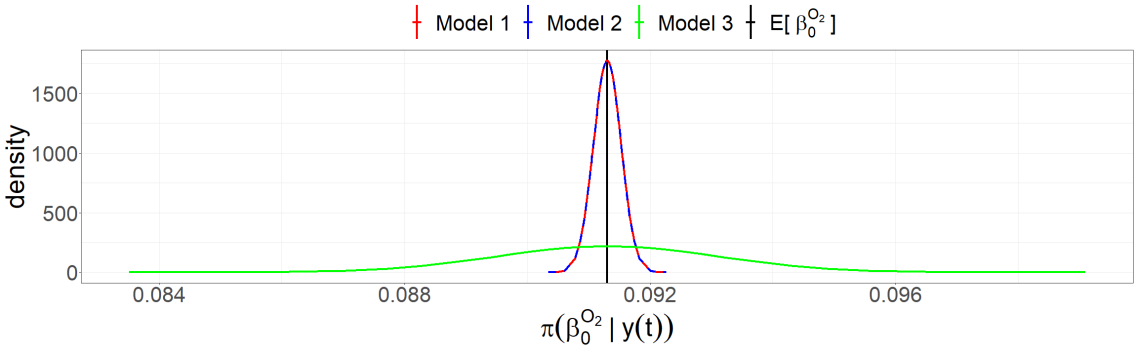

Figure 16: Plotted posterior distribution of the intercept for oxygen for Model 1, Model 2 and Model 3, together with prior mean of the intercept $\beta_0^{O_2}$.

Figure 16 shows the posterior distributions of the parameter $\beta_0^{O_2}$ fitted by Model 1, Model 2 and Model 3. Recall from Chapter 4.1 that

$$\beta_0^{O_2} = \frac{c_{r,O_2} q_r + c_{f,O_2} q_f}{V},$$

where c are concentrations q are waterflows and the subscrips r and f represents recycled and fresh sea water, respectively. The posterior distrbutions have, as visualized in Figure 16, their mean essentially equal to the calculated expected mean of $\beta_0^{O_2}$ according to the data and the parameter assumptions. The figure shows that the posterior distribution of $\beta_0^{O_2}$, $\pi(\beta_0^{O_2}|y(t))$ appears identical for Model 1 and Model 2. We note that these two models are assigned the same Gaussian priors for the Gaussian priors, such that the presence or absence of the PC prior for the variance of the RW1 prior for the model discrepancy and the variance of the model noise parameter does not affect the posterior distribution of the intercept for the model of oxygen concentration. Model 3, which is assigned a larger prior variance than Model 1 and Model 2, has a larger posterior variance than the two other models. All of the posterior variances does still appear to be relatively small for

the model. There is difficult to determine per preferred model based on the posterior distribution for $\beta_0^{O_2}$ fitted by Model 1, Model 2 and Model 3 alone. Traditionally, one perfer unbiased models with the least variance. However, assessing the parameter components, the parameters c_{r,O_2} and c_{f,O_2} do contain some significant uncertainty as the ability to absorbe gasses vary with salinity, preassure and temperature, which varies through the water circuit arount the tank. Therefore, perferring a model considering some uncertainty regarding this parameter would arguably be the better choice. Thus, by assessing the intercept parameter for the oxygen model, $\beta_0^{O_2}$, Model 3, giving the highest posterior variance of the considered models would be the perferred choice.

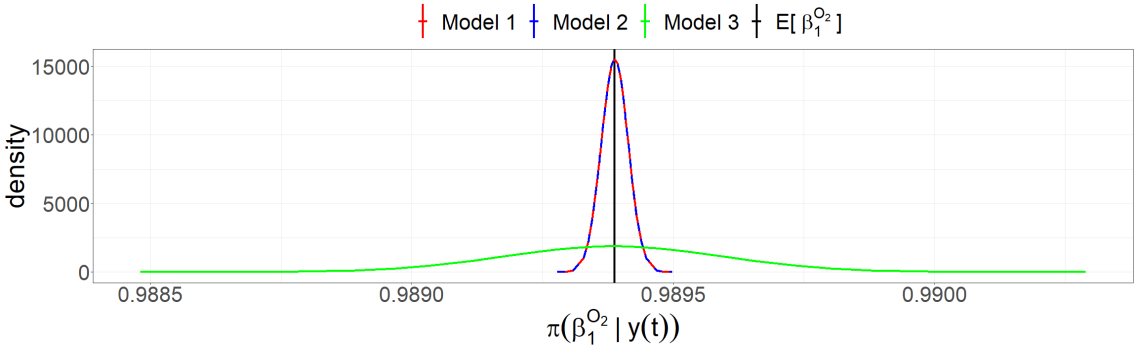

Figure 17: Plotted posterior distribution of the coefficient for $y_{u,O_2}(t)$, $\beta_1^{O_2}$, for Model 1, Model 2 and Model 3, together with its prior mean.

The parameter of $y_{u,O_2}(t)$, $\beta_1^{O_2}$ is given analytically in Chapter 4.1 as

$$\beta_1^{O_2} = 1 - \frac{q}{V},$$

and has its posterior distributions for Model 1, Model 2 and Model 3 plotted in Figure 17. Similar trends as for the oxygen model intercept is evident also for this model. Model 1 and Model 2 seems to have identical posterior distributions for the parameter and Model 3 has higher posterior variance, as the prior variance is larger. All models have their mean close to the calculated expected parameter value which also is their prior mean. As for the oxygen model intercept $\beta_0^{O_2}$, the assigned PC prior to the hyperparameters $\sigma^2_{y,O_2}$and σ^2_{δ,O_2} does not affect the posterior distribution of the parameter $\beta_1^{O_2}$. The perferrable model based on $\beta_1^{O_2}$ is as for $\beta_0^{O_2}$ hard to determine isolated. However, the variance of the parameter q, which is the only factor of the parameter containing variance, is according Kvalsvik (2022), larger than the variance in the observed data itself. The prior distribution of Model 3 has taken this into consideration, and could thus arguably be the best model choice to the parameter assumptions.

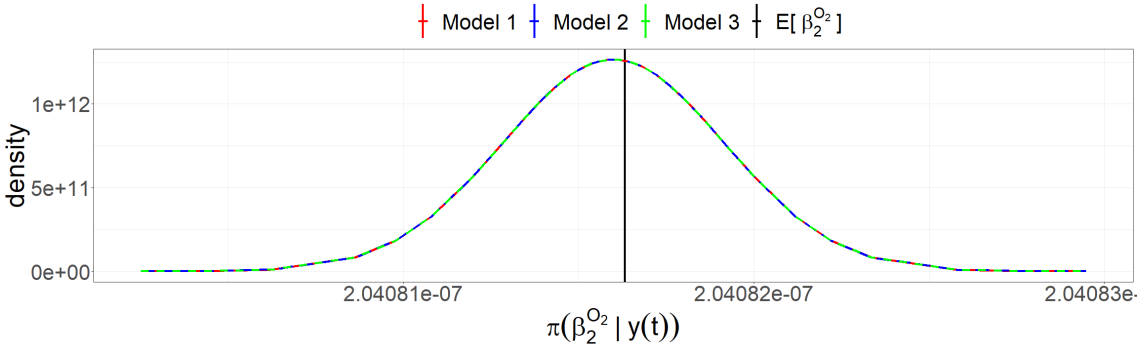

Figure 18: Plotted posterior distribution of the second coefficient of the fixed effects, i. e. $1/V$ for oxygen for Model 1, Model 2 and Model 3, together with prior mean of the coefficient $\beta_2^{O_2}$, which is equal to $1/V$.

The parameter of the linear effect $h(t)$, which is added oxygen to the oxygen model, $\beta_2^{O_2}$, is

analytically equal to

$$\beta_2^{O_2} = \frac{1}{V}$$

from Chapter 4.1, and has its posterior distribution plotted in Figure 18. This is a known constant parameter given the properties of a random variable parameter to be able to interpretate $h(t)$ as a linear effect similarly as $y_{u,O_2}(t)$. For the parameter to be as close to the real known value as possible, we want its posterior distribution to have its mean equal to the known value, with as small variance as possible, which was assigned as the prior distribution for all considered models. Figure 18 shows that the posterior distribution appears to be equal for all of Model 1, Model 2 and Model 3, such that the change in Gaussian priors for the other linear effect parameters nor the change in hyperpriors for the hyperparameters have any effect on the posterior distribution of the parameter. It does also apear from Figure 18 that the posterior mean seems to deviate some from the calculated expected value. However, by looking at the magnitudes on the first axis, this difference seems to be negigible. As well, taking the magnitudes of the first and second axis into account, the uncertainty is also sufficiently small to evaluate the posteror distribution of the parameter to be as we wanted.

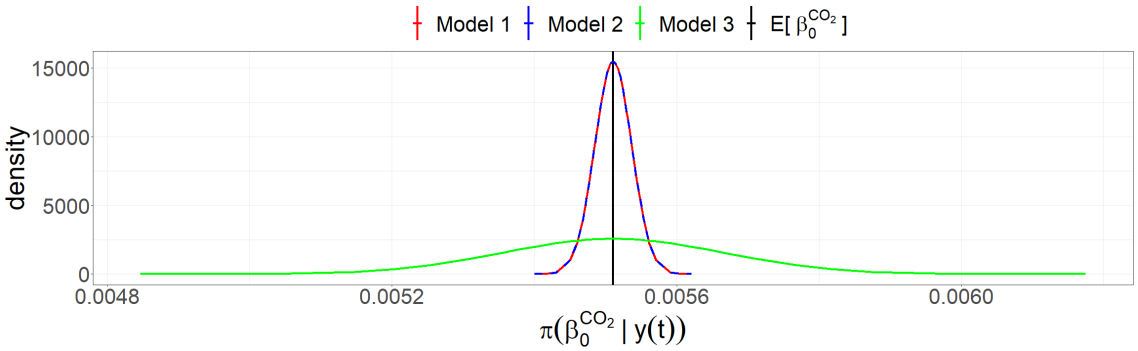

Figure 19: Plotted posterior distribution of the intercept for carbon dioxide for Model 1, Model 2 and Model 3, together with prior mean of the coefficient $\beta_0^{CO_2}$.

The carbon dioxide model intercept $\beta_0^{CO_2}$ has its posterior distributions $\pi(\beta_0^{CO_2}|y(t))$ plotted in Figure 19, and has the analytical expression from Chapter 4.1 given as

$$\beta_0^{CO_2} = \frac{q_f c_{f,CO_2}}{V}.$$

From Figure 19, the same trends as for the posterior distribution of $\beta_0^{CO_2}$ as for the posterior distributions for $\beta_0^{O_2}$ and $\beta_1^{O_2}$. All of Model 1, Model 2 and Model 3 has their posterior mean as good as equal to the calculated expected value, which is also their prior mean. The posterior distributions of $\beta_0^{CO_2}$ by Model 1 and Model 2 appears to be equal, indicating that the assigned hyperpriors do not affect the posterior distribution of the parameter. The posterior distribution of Model 3 has a higher posterior variance than Model 1 and Model 2, probably caued by a larger prior variance. The parameter expected and posterior mean is still small, with relative small variances for all of the three models. As for the parameters of the oxygen model, one can assume that q_f has a higher measurment uncertainty than captured by the data, and c_{f,CO_2} has a present parameter uncertainty by Kvalsvik (2022), indicating that Model 3 is the most appropriate model assessing $\beta_0^{CO_2}$.

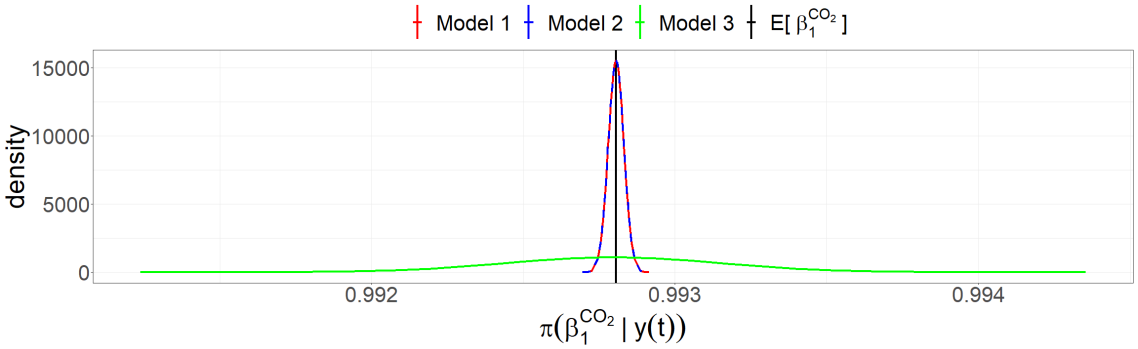

Figure 20: Plotted posterior distribution of the coefficient for the fixed effect for carbon dioxide for Model 1, Model 2 and Model 3, together with prior mean of the coefficient $\beta_1^{CO_2}$.

The coefficient for $y_{u,CO_2}(t)$, $\beta_1^{CO_2}$ is analytically given as

$$\beta_1^{CO_2} = 1 - \frac{q_f - \xi q_r}{V}$$

from Chapter 4.1 and has its posterior distribution plotted in Figure 20. From the figure, similar trends as for the other linear effect parameters are present, where the posterior mean of Model 1, Model 2 and Model 3 are essentially equal to the calculated expected value, which lso is assigned as the prior mean. Model 1 and Model 2 seems to have equal posterior distributions for the parameters, indicating that the assigned PC prior does not have any effect on the parameter posterior distribution. Unlike the other linear effect parameters, the priorvariance of Model 3 is larger than the variance of Model 1 and Model 2, most likely caused by an assigned prior variance which is based on a large variance in the parameter component ξ. Choosding a perferred model by the posterior distribution of this parameter yields similar argumets as for the other linear effect parameters. The variance of the parameter components q_f, q_r and ξ is most likely larger than the variance obtained by the noise in the data, such that Model 3, having the most posterior variance, appears to be the most appropriate model for the model assumptions.

5.2 Posterior discrepancy functions

We denote the posterior discrepancy $\boldsymbol{\delta}(t)|\boldsymbol{y}(t) = [\delta_{O_2}(t)|y_{O_2}(t), \delta_{CO_2}(t)|y_{CO_2}(t)]^T$. We plot the posterior discrepancy functions for Model 0, Model 1, Model 2 and Model 3 with confidence bounds to see how the discrepancies are affected from model to model. This is as well compared to the calculated values, denoted $\boldsymbol{g}(t) = [g_{O_2}(t), g_{CO_2}(t)]$, visualized in Figure 10. We compare the discrepancy functions of carbon dioxide for every model, and visualize their sum and compare it to the calculated results for $\boldsymbol{g}(t)$. Note that the considered linear model does not have any discrepancy function.

5.2.1 Posterior discrepancy function for random effect models with vague priors

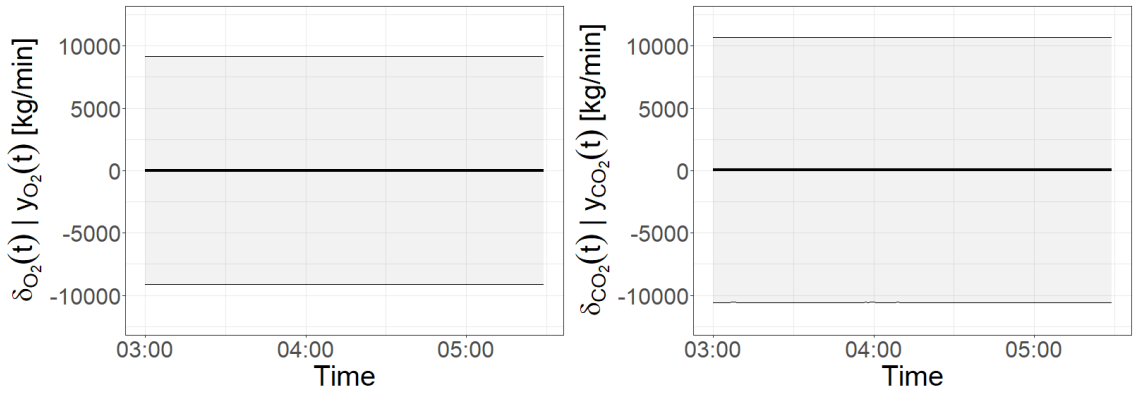

Figure 21: Plotted posterior discrepancy function means with a 95% credible interval for oxygen and carbon dioxide using Model 0.

Figure 21 shows a plot of the mean and 95% confidence bounds of the posterior discrepancy functions $\delta_{O_2}(t)|y_{O_2}(y)$ and $\delta_{CO_2}(t)|y_{CO_2}(y)$ for Model 0. We see that the uncertainty in this plot is extraordinary large compared to the estimated value. This makes sense, as no prior infromation about the parameters were fitted to the model, and that the fitted random effect, represented as the discrepancy function accounts for most of the noise in the model. We also note that the posterior variance of disrcepancy for produced carbon dioxide is higher than for the consumed oxygen. This is aligned with the higher noise of the plot of observed carbon dioxide consumption in Figure 10 coming from higher noise in the observed carbon dioxide concentration than for the oxygen in Figure 8. It is worth to take into consideration that the observed data lies within the confidence bounds.

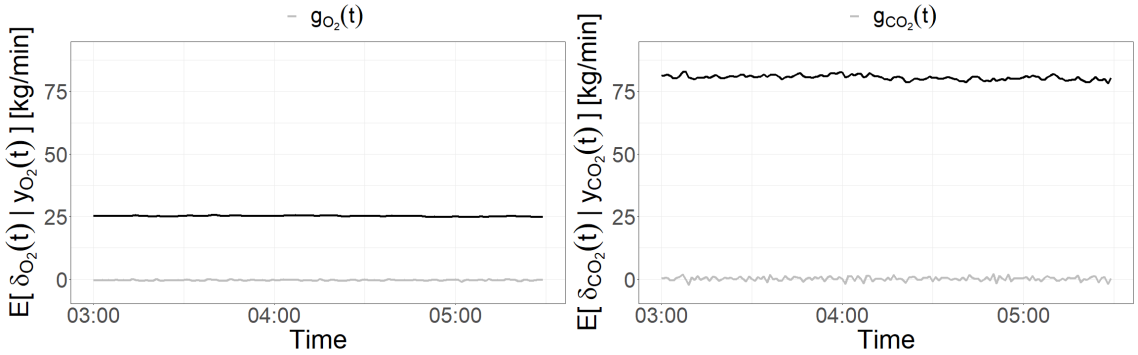

Figure 22: Plotted mean of posterior discrepancy functions $\mathrm{E}[\boldsymbol{\delta}(t)|\boldsymbol{y}(t)]$ compared to the calculated functions $\boldsymbol{g}(t)$.

To be able to compare the posterior discrepancy functions to what we obtained from the observed data, the posterior mean of the discrepancies are plotted in Figure 22, with the observed consumed and accumulated substances in the background. We see that the posterior mean of both the discrepancy functions are significantly higher than what we could expect. The posterior mean of the discrepancy function for oxygen accumulation is positive in contrast to the observed $g_{O_2}(t)$ and a value of approximately 25 kg/min higher than what we observed. The posterior mean for the discrepancy function ofthe produced carbon dioxide is over 75 kg/min higher than observed. This can be explained that the intercept and the coefficients of the linear effects are not assigned any priors in Model 0, and takes the "missing" amount of oxygen between what comes in and goes out.

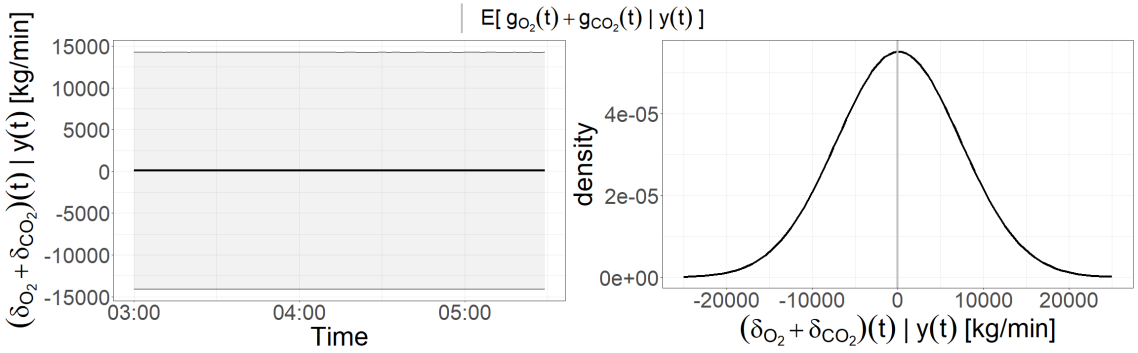

Figure 23: Plotted difference in posterior discrepancy functions for oxygen and carbon dioxide using Model 0 by mean with a 95% credible interval. Density plot of the posterior differences using the mean of the means for the whole distribution and means for the standard deviations.

Figure 23 shows that the difference in consumed oxygen and produced carbon dioxide is affected by large uncertainties, reinforcing our assumption of that no specified prior distribution causes most of the noise in the data to be captured by the discrepancy function. The large uncertainties makes the distribution incomparable to the same quantity for the observed data $\boldsymbol{g}(t)$ visualized in Figure 11 in Chapter 2, but we may compare the mean posterior difference to the observed data, as done in Figure 24.

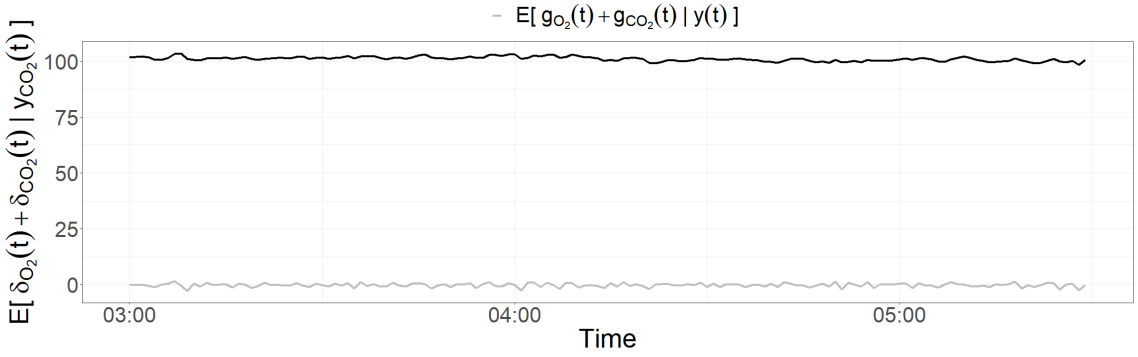

Figure 24: Plotted difference in the meam posterior discrepancy functions for oxygen and carbon dioxide using Model 0. The observed difference is plotted in grey labelled $g(t)$.

Figure 24 shows that the posterior sum of the mean of accumulated carbon dioxide and oxygen is over 100 kg/min higher than for the observed data $\boldsymbol{g}(t)$ and our prior expectation. This is, as mentioned, probably because the discrepancy functions accounts for some of the effects that should be captured by the intercepts and the coefficients for the fixed effects. Physically, this means that the salmon produces more carbon dioxide than is expected to based on its oxygen consumption.

5.2.2 Posterior discrepancy functions for models assigned priors

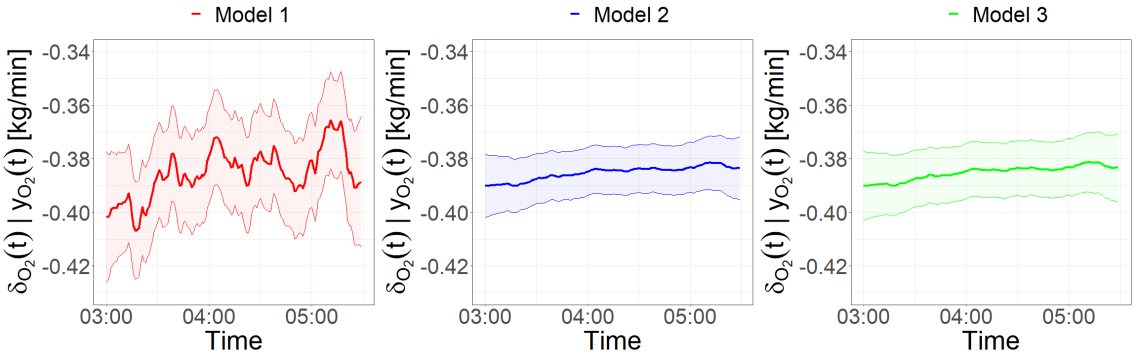

Figure 25: Plotted posterior discrepancy function for accumulated oxygen in the tank, $\delta_{O_2}(t)$, as the mean with a 95% confidence interval for Model 1, Model 2 and Model 3.

The posterior discrepancy for the oxygen accumulation $\delta_{O_2}(t)$ for Model 1, Model 2 and Model 3 is plotted as the mean value with 95% confidence bounds in Figure 25. All three models appears to yield increasing posterior discrepancies, which is decreasing in absolute value. This is aligned with the assumptions of the salmon consuming a decreasing amount of oxygen in the feeding breaks as discussed in Chapter 2. Model 1, with assigned Gaussian priors for the linear effect parameters and vague priors for the variance of the model noise, σ^2_{y,O_2}, and the variance of the RW1 prior for the discrepancy, σ^2_{δ,O_2}, appears to have a posterior discrepancy with sime large fluctuation and present noise. When a relatively strict hyperprior is fitted to σ^2_{δ,O_2} in Model 2, the posterior discrepancy becomes more smooth with an appearing smaller cnfidence interval. This is reasonable, as more of the noise in the model is captured by the model noise parameter and less of the model noise are left to the random effect and the fixed effect parameters. From Model 2 to Model 3, the confidence interval of the discrepancy appears to increase, when the prior variance of $\beta_0^{O_2}$ and $\beta_1^{O_2}$ increases and the other priors are unchanged relative to Model 2. This can be explained by when the variance of the two other elements in the model increases, the variance of the third element also increases to keep the balance in the model. By comparing the posterior discrepancy functions for oxygen of Model 1, Model 2 and Model 3, Model 2 seems to give the most accurate estimate, and would thus be the perferred model based on this parameter.

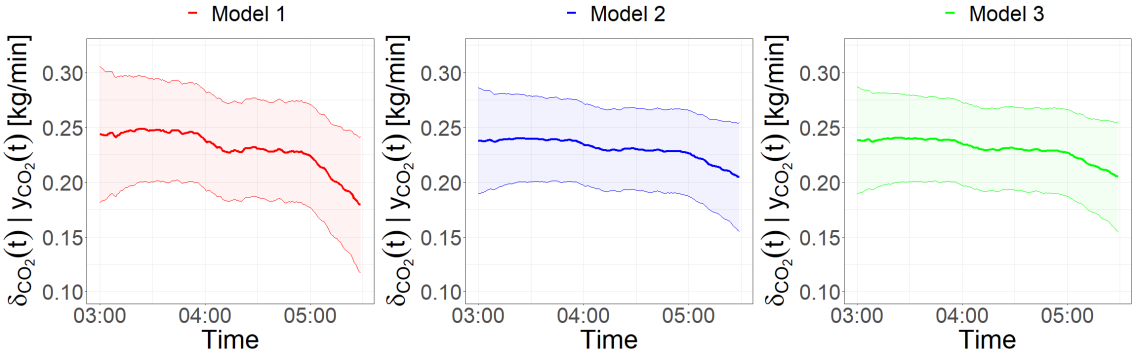

Figure 26: Plotted posterior discrepancy function for accumulated carbon dioxide, $\delta_{CO_2}(t)$, as the mean-value with a 95% confidence interval for Model 1, Model 2 and Model 3.

Figure 26 shows the posterior discrepancy function of carbon dioxide accumulation $\pi(\delta_{CO_2}(t)|y(t))$ fitted by Model 1, Model 2, and Model 3. The carbon dioxide accumulation appears to be decreasing for all three models, aligned with the assumption of reduced carbon dioxide production by the salmon through feeding breaks, as discussed in Chapter 2. Model 1 appears in Figure 26 to yield a

slightly more noisy posterior discrepancy function with a larger confidence interval than Model 2 and Model 3. The two latter models seems further to give essentially equal posterior discrepancy functions for the carbon dioxide model, indicating that the increased prior variances of the Gaussian parameters from Model 2 to Model 3 does not affect the posterior discrepancy function for the carbon dioxide models. Choosing the best of the three model based on the posterior discrepancy function for carbon dioxide is a hcoice between the two approximately equally accurate models Model 2 and Model 3. As they perform similarly on this model component, it is difficult to choose either of them as their complexities are essentially the same.

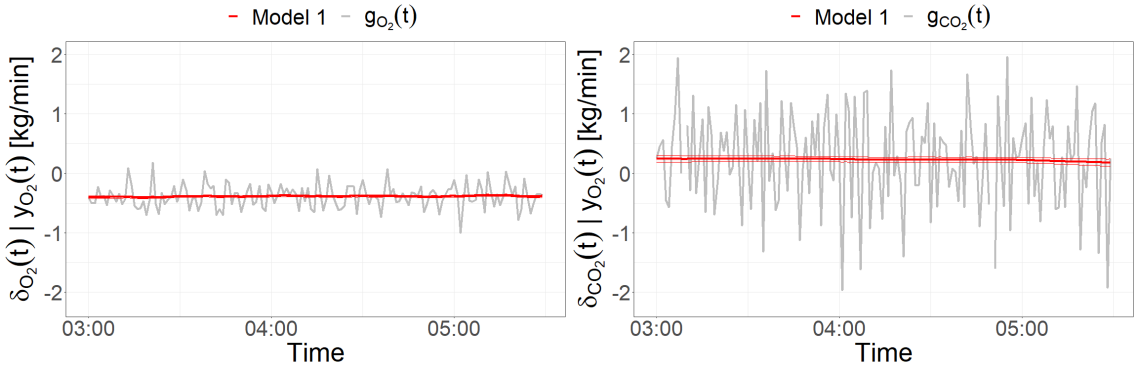

Figure 27: Plotted posterior discrepancy function means with a 95% credible interval for oxygen and carbon dioxide using Model 1 compared to the calculated values $g_{O_2}(t)$ and $g_{CO_2}(t)$.

The posteror discrepancy functions in Model 1 are plotted in Figure 27, and seem more aligned with the data than the mean of the discrepancy of Model 0, plotted in 22. The priors chosen for the variance in the linear predictor and the variance in the discrepancy functions are both vague and relatively open. The prior distributions assigned to the model intercept and coefficients of the linear effects seem to reduce the variance of the discrepancy functions and model them such that they are in the neighborhood of the observed data.

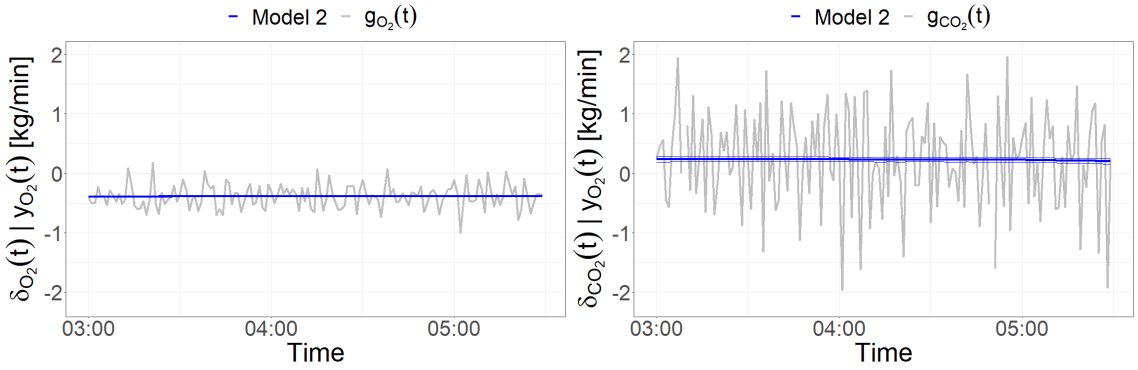

Figure 28: Plotted posterior discrepancy function means with a 95% credible interval for oxygen and carbon dioxide using Model 2 compared to the calculated values $g_{O_2}(t)$ and $g_{CO_2}(t)$.

The discrepancy functions for oxygen and carbon dioxide with 95% confidence bounds fitted in Model 2 are plotted in Figure 28. The priors in this model includes PC-priors for the linear predictor and for the discrepancy function, and is otherwise identical to Model 1. The prior variance assigned to the linear predictor is relatively large ($Pr(\sigma_y > 1) = 0.99$), and the prior assigned to the random walk parameter in the discrepancy function is relatively strict ($Pr(\sigma_\delta > 1) = 0.01$). We could expect this to cause the variance of the discrepancy function in Model 2 to be smaller than for Model 1, which is evident in the plots of the posterior discrepancies for Model 1, Model 2 and Model 3 in Figure 25 and Figure 26.

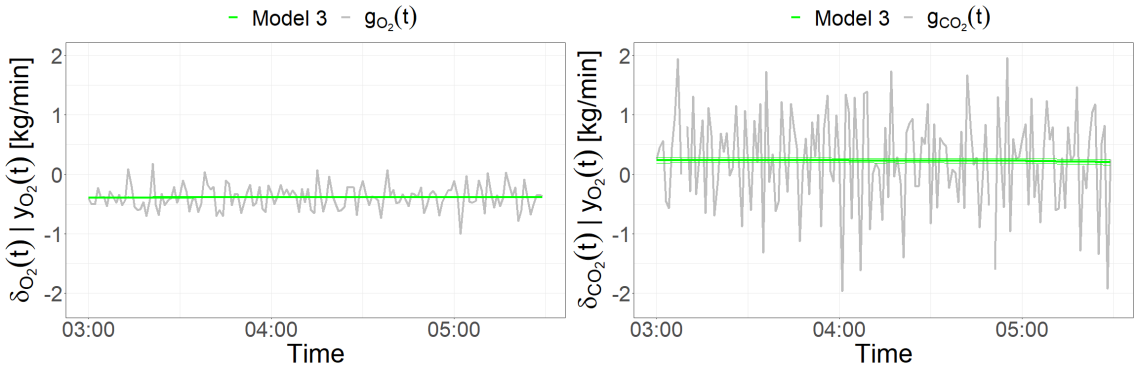

Figure 29: Plotted posterior discrepancy function means with a 95% credible interval for oxygen and carbon dioxide using Model 3 compared to the calculated values $g_{O_2}(t)$ and $g_{CO_2}(t)$.

Figure 29 shows the plotted posterior discrepancy functions for the accumulated oxygen and carbon dioxide estimated using Model 3. Model 3 has looser priors for the fixed effects and similar hyperpriors for the non-Gaussian hyperpriors of the variances of the linear predictor and of the random walk process of the discrepancy. I. e. the prior variance of the Gaussian distributed intercepts and coefficients for the fixed effects for the models is higher than for Model 2. We could thus expect that these intercepts and coefficients captures slightly more of the variance in the model such that the variance of the discrepancy function is shrinked further.

Figure 27, Figure 28 and Figure 29 compares the posterior discrepanciy functions $\delta_{O_2}(t)$ and $\delta_{CO_2}(t)$ to each other and to the calculated oxygen and carbon dioxide accumulations $g_{O_2}(t)$ and $g_{CO_2}(t)$ for Model 1, Model 2 and Model 3, respectively. All three figures have plotted the mean values with the 95% confidence intervals as in Figure 25 and Figure 26. All three figures shows similar properties of the three considered model, where the variances of the posterior discrepancy functions are small compared to the noise of the calculated model. Moreover, the posterior discrepancy functions seems to be around the same vlues, where approximately 0.39 kg oxygen vanishes from the tank per minute on average, and approximately 0.23 kg carbon dioxide accumulates on average every minute. The posterior discrepancy functions appear constant compared to the calculated model for all fitted models, which is explained by relatively large noise in the observed data compared to what is obtained by the fitted models.

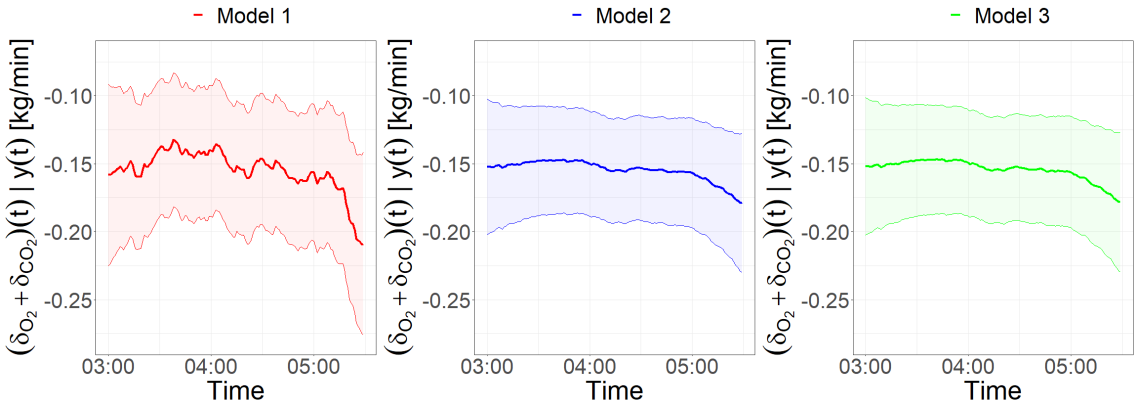

Figure 30: Plotted differnce between posterior discrepancy functions for oxgygen and carbon dioxide as the mean value with a 95% credible interval using Model 1, Model 2 and Model 3.

Figure 30 shows the sum of the posterior discrepancy functions for Model 1, Model 2 and Model 3, $(\delta_{O_2}(t) + \delta_{CO_2}(t))$. This has essentially the same properties as the posterior discrepancy functions of oxtgen and carbon dioxide with more noise in Model 1 than in the two other models. However,

the figure shows a negative trend in the sum of the posterior discrepancy functions for all of the three fitted models. This indicates that less carbon dioxide is accumulated than oxygen vanished as the time goes. We would initially expect this curve to be approximately constant, as most of the oxygen consumed gets transformed into carbon dioxide at approximately the same rate of mass. As a feeding break takes place in the period considered, and we expect a decay in both oxygen consumption and carbon dioxide production, this might indicate that there is a delay in the carbon dioxide production compared to the oxygen consumption. This decay might have a natural explaination in the reaction time of the process of transforming oxygen into carbon dioxide, which is an aspect to take into consideration.

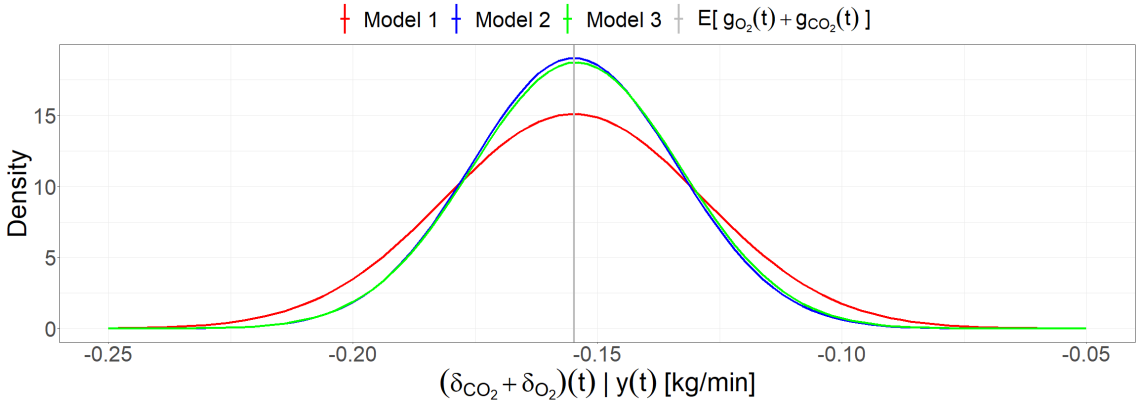

Figure 31: Plotted density of the differnce between posterior discrepancy functions for oxgygen and carbon dioxide using Model 1, Model 2 and Model 3.

Considering Figure 31, which shows a distribution approximation of the sum of the discrepancies over the whole period, it is evident that Model 1 has the most variance in the sum of the discrepancy functions, and that Model 2 has slightly less variance than Model 3. This can be explained by what is shown in Figure 25, where the posterior discrepancy function for oxygen has slightly more variance in Model 3 than in Model 2, where larger prior variances for the Gaussian parameters are assigned. Moreover, all of the models considered have their mean of the difference of the posterior discrepancy functions close to the mean of the differnce of the calculated model, showing that the posterior discrepancy functions fits well with the data and the model assumptions for Model 1, Model 2 and Model 3. Based on the differences in the posterior discrepancy functions, all thre models performs relatively well on the data, but Model 2 might be perferrable as it has slightly less uncertainty than Model 3.

5.3 Posterior relevant hyperparameters

This chapter shows the results of the posterior distributions of the hyperparameter $\boldsymbol{\sigma}_y^2 = [\sigma_{y,O_2}^2, \sigma_{y,CO_2}^2]^T$ of the different fitted models, which is the variance of the Gaussian observations. Recall that the linear model, Model 0 and Model 1 has assigned the vague inverse Gamma prior to this hyperparameter, while Model 2 and Model 3 has assigned the same PC prior to the parameter, as described in Chapter 4.2.

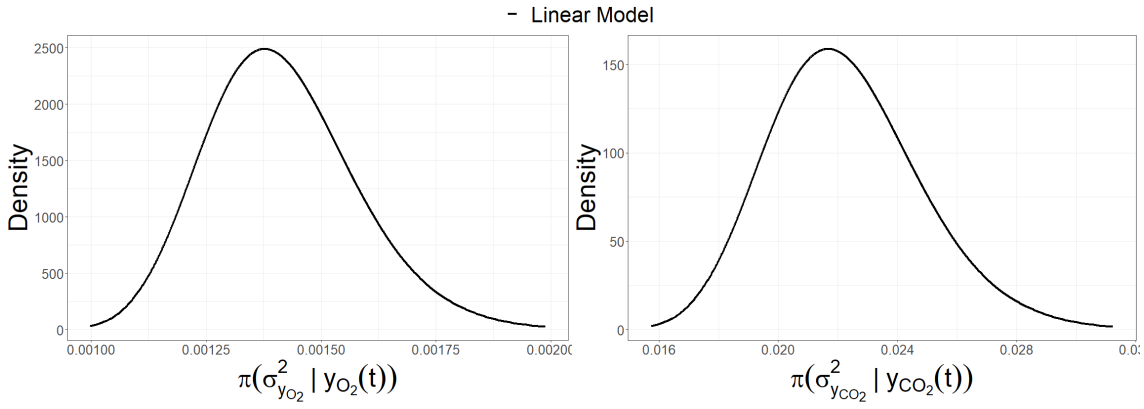

Figure 32: Plotted posterior marginal density of the hyperparameter of the variance of the observations for oxygen and carbon dioxide, σ^2_{y,O_2} and σ^2_{y,CO_2} for the linear model.

Figure 32 shows the posterior distribution of the parameter $\boldsymbol{\sigma}^2_y$ for the oxygen and carbon dioxide model fitted by the linear model. The posterior distributions have their modes close to 0.0014 and 0.021 for oxygen and carbon dioxide, respectively. This is smaller than the empitical variance of the oxygen and carbon dioxide concentrations from the observed data, where the oxygen has variance 0.00294 and carbon dioxide has variance 0.0278, given in Table 4 in Chapter 2.4.

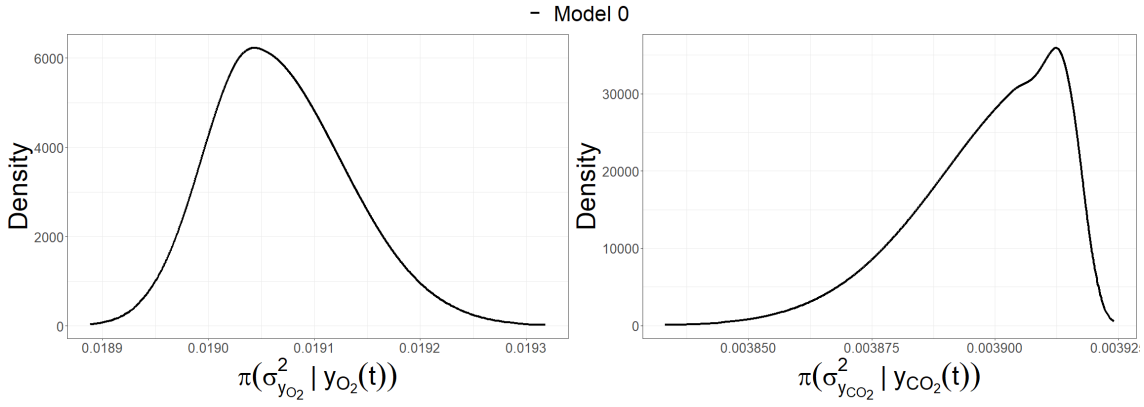

Figure 33: Plotted posterior marginal density of the hyperparameter of the variance of the observations, σ^2_{y,O_2}.

Figure 33 shows the posterior distribution of the hyperparameters σ^2_{y,O_2} and σ^2_{y,CO_2} fitted by Model 2, which has vague priors for all parameters and hyperparameters. The posterior distribution for σ^2_{y,O_2} has its mode at approximately 0.0195, and the posterior distribution of σ^2_{y,CO_2} has its mode at approximately 0.00391. The oxygen model as thus larger variance in the Gaussian observation for Model 0 than for the linear model, while the opposite is the case for the carbon dioxide model. Comparing the plots of the posterior parameter estimates in Figure 14 and Figure 15 and the posterior discrepancy functions in Figure 21 of oxygen and carbon dioxide, we see that the posterior fixed effect parameters for the two substances has essentially equally much uncertainty for oxygen and carbon dioxide, while the carbon dioxide posterior discrepancy has more variance than the posterior discrepancy of oxygen. Thus, more of the posterior variance of carbon dioxide can be captured by the discrepancy, while the posterior variance is captured more by σ^2_{y,O_2} for oxygen.

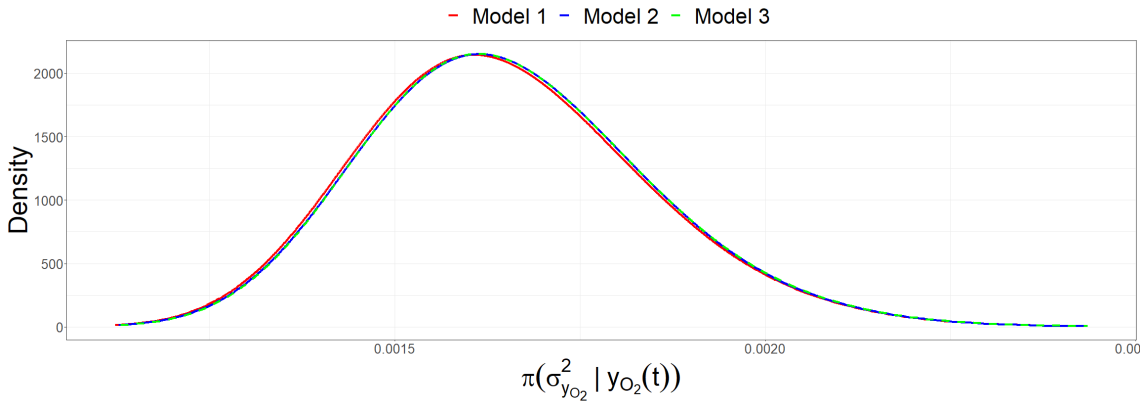

Figure 34: Plotted posterior marginal density of the hyperparameter of the variance of the observations, σ^2_{y,O_2}.

Figure 34 shows the posterior distribution of the variance of the Gaussian observation, σ^2_{y,O_2}. All three models appear to have approximately the same posterior distribution for the parameter, with the posterior mode at approximately 0.0016, in the neighborhood of the posterior mode for the parameter by the linear model, and smaller than the empirical variance of the observed data of 0.00294. The posterior mode of Model 1 is slightly smaller than for Model 2 and Model 3, but overall all three posterior distributions appear relatively similar. The assined PC prior to the parameter does thus not have any major effect on the parameters posterior distribution.

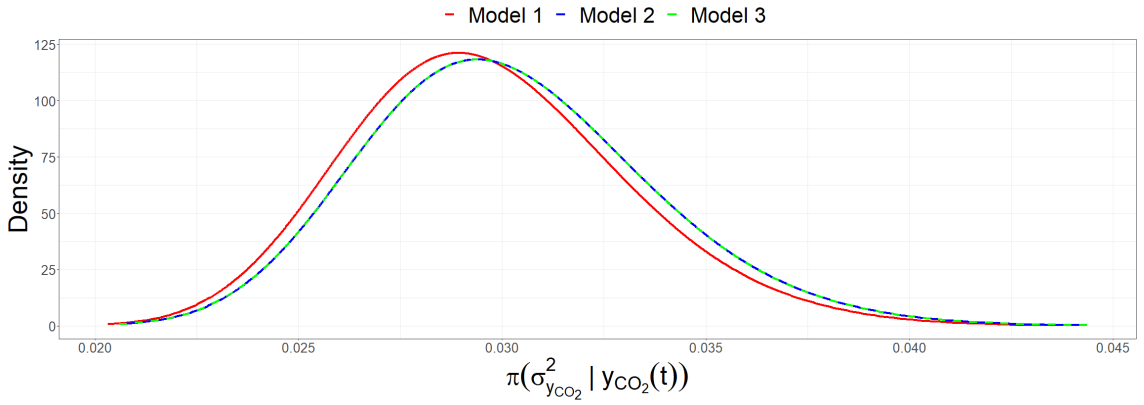

Figure 35: Plotted posterior marginal density of the hyperparameter of the variance of the observations, σ^2_{y,CO_2}.

Plots of the posterior distribution of the hyperparameter σ^2_{y,CO_2} fitted by Model 1, Model 2 and Model 3 are shown in Figure 35. The posterior distribution of the parameter fitted by Model 2 and Model 3 appears similar with the posterior mode close to 0.029, while the posterior distribution of the parameter fitted by Model 1 has a slightly lower posterior mode of approximately 0.027. With an empirical variance of the observed data at 0.0278, the posterior variance is in the same neighborhood as the observed. The PC prior assigned to Model 2 and Model 3 thus appears to cause a slight increase in the posterior variance of the Gaussian observations.

5.4 Evaluation criteria

The evaluation criteria DIC, described in Chapter 3.6 for the linear model, Model 0, Model 1, Model 2 and Model 3 for oxygen and carbon dioxide are systematized in Table 8.

Table 8: Evaluation criteria of each model.

Model	DIC
Linear Model O_2	-553.32
Linear Model CO_2	-140.96
Model 0 O_2	-187.71
Model 0 CO_2	-278.70
Model 1 O_2	-525.39
Model 1 CO_2	-96.65
Model 2 O_2	-533.67
Model 2 CO_2	-98.42
Model 3 O_2	-533.72
Model 3 CO_2	-98.43

From the deviance information criteria (DIC) given in Table 8, which has the lowest magnitude for the model best fitting the data, it is evident that the linear model performs best on modelling oxygen in the tank and Model 0 performs best on modelling carbon dioxide in the tank with DIC of -553.32 and -278.70. Even these models are observed to not fit well to the data, they obtain such a results on the deviance information criteria as they under vague priors is fitted "freely" which is good in terms of the DIC. Having identified the linear model and Model 0 not to fit the assumptions and data, we are left with comparing Model 1, Model 2 and Model 3. Here, Model 3 performs slightly better than Model 2 for both oxygen and carbon dioxide in terms of DIC, and they do both perform better than Model 1. Thus considering DIC, Model 3 appears to be the best fit to the data, which is aligned with what is discussed regrading the posterior distributions of the coefficients of the fixed effects.

6 Model assessments and physical interpretations

This chapter performs an assessment of the results of the fitted models defined in Chapter 4.2, which are found in Chapter 5. The discussion will try to identify the best model, with respect to the data and assumptions. The models are further compared to the data and to the prior assumptions, where eventual deviations from experts expectations are discussed. The chapter ends with suggestions on how to improve the model approaches, and how the model approach needs to be improved to give value for the industry.

6.1 Choice of models

In this chapter, we will evaluate the performance of five different models: the linear model, Model 0, Model 1, Model 2, and Model 3. Our goal is to determine which model best fits the data, and how much prior information should be assigned to the model parameters and hyperparameters in order to make meaningful inferences about the data and the assumptions of the model. The assessment will be based on the results presented in Chapter 5.

Having assessed the linear model without any random effects fitted with vague priors, and Model 0, with a random effect assigned a RW1 prior with vague Gaussian priors and non-Gaussian hyperpriors in Chapter 5, it is evident that neither of the models seems to fit satisfactory to the data nor the physical assuptions. There is thus a neecessity of assigning regulations for which values the parameters may take for the Model to fit appropriately to the data. For improving the models, Model 1, Model 2 and Model 3 described in Chapter 4.2 are fitted, to see how different assigned prior infromaion affects the accuracy of the model with respect to its parameters and the expected discrepancies.

By evaluating the models fitted, it is evident that the linear model with no discrepancy function explaining oxygen consumed and carbon dioxide produced by the salmon does not fit well with prior assumptions and observed data. Introducing a discrepancy function accounting for this effect does neither perform aligned with prior expectations and observed data. Assigning gradually more prior information to the models improves their fit to their assumptions and modelled data. Thus, expert informed prior distributions are necessary to assign to the model for it to make sense.

To decide between the best model fit to the data, the decision stands between Model 1, Model 2 and Model 3. They all appear to perform similarly in terms of posterior distributions of all parameters. It is evident that the posterior distributions corresponds to prior information assigned to the parameters in all of the models. As there is of interest for the model to capture a reasonable share of the model uncertainty in the parameters, Model 3 seems to capture the most of it by assessing the posterior distribution of its linear effect in figures 16, 17, 18, 19 and 20. This corresponds with The results from considering the DIC, showing that Model 3 fits best to the data of the models giving reasonable posterior estimates to the assumptions and the data.

6.2 Model errors

Assessing the fitted models, comparing them to the data and evaluating their performance against each other, showed that the three models where priors were assigned some or all of the model parameters and Hyperparameters coincided relatively well with the data and the model assumptions. There were however some deviations for some of the model components worth discussing, with some possible physical aspects provided Kvalsvik (2022) and Stakvik-Løvås (2022).

The fitted linear model and Model 0 deviates from the assumptions and observed data as a consequence of lack of initial information fitted through the discrepancy function and through vague assigned priors to the models, as discussed in Chapter 6.1. To evaluate model deviations from field experts expectations, we thus only consider the three models having a reasonable fit to the data, calculated models and parameter assumptions.

As written in the introduction, dos Santos *et al.* (2022) expect 0.45 kg oxygen to produce 0.48 kg carbon dioxide, corresponding to a sum of accumulated substances of 0.03 kg. From Figure 31, it is evident that the posterior mean of the sum of accumulated substances is slightly lower than -0.16 kg/min. From Figure 25, it is evident that the posterior mean of the accumulated oxygen is in the neighborhood of -0.39 kg/min for all t, and Figure 26 shows a posterior mean difference in the neighborhood of 0.23 kg/min for all t, which is slightly lower in absolute value than what is assumed by dos Santos *et al.* (2022). Scaling to similar values of the posterior discrepancy functions gives a posterior sum of discrepancy functions even lower than -0.16 kg/min. Thus, taking these results into consideration, the fitted models indicates that it vanishes more oxygen than it is produced corresponding carbon dioxide in the tank through the time period, with an enforced effect as time goes, evident through the downward sloping curves in Figure 30. These results may have several physical explainations, among others a significant reaction time of transforming oxygen to carbon dioxide, degassing inside the tank not captured by the assumptions nor the data, bacteria and possible algae in the tank, pH of the water in the tank and sensor uncertainties, which are given by Kvalsvik (2022) and Stakvik-Løvås (2022).

When oxygen and carbon, considered as feed, reacts to carbon dioxide, some time are required for the reaction. This is caused by the biological process of digestion and combustion, which requires time for all spieces in the nature. A following consequence is that the oxygen consumed by salmonids at time t is emitted into the water as carbon dioxide at time $t + \delta t$ for some lag δt, which is not taken into consideration in the model. As a feeding break is considered further, and a decay in oxygen consumption and carbon dioxide production is expected over the time period, the lag causes the decay in carbon dioxide emmision happens a time δt after the oxygen consumption, causing the sum of accumulated oxygen and accumulated carbon dioxide to be lower than it would be with a model adjusted for this lag, as the results in Figure 31 and Figure 30 shows.

Degassing in the tank causes the observed concentration of gasses dissolved is lower than the expected concentration by the amount of gasses injected into the tank. The effect may be present

for both oxygen and carbon dioxide, but is expected by Kvalsvik (2022) and Stakvik-Løvås (2022) to be stronger for oxygen than for carbon dioxide as oxygen is injected throug high preassure meachines, and the preassure decreases rapidly to approximately 1 atm., which may have the consequence of lost gasses into the surrounding air. Assuming the degassing effect is significantly stronger for oxygen than for carbon dioxide, it can be interpreted through the data as a negative bias in the observed added oxygen $h(t)$. The amount of added oxygen $h(t)$ is negatively correlated to the accumulated oxygen in the tank, i. e. $(\partial/\partial h(t))\delta_{O_2}(t) < 0$ and $(\partial/\partial h(t))g_{O_2}(t) < 0$. Thus, a negative bias in $h(t)$ causes the calculated accumulated oxygen and posterior discrepancy funtion of the accumulated oxygen shows a higher oxygen consumption equivalent to a lower oxygen accumulation than what it would be accunting for this effect. A model accounting for this would therefore show a sum of accumulated substances higher than a model not accunting for this, aligned with the result that the fitted model shows a lower posterior sum of discrepancy functions than what would be expecited by dos Santos *et al.* (2022).

Bacteria and algea affects the concentration of substances in the tank by combustion and photosyntesis reactions, respectively. Combustion of oxygen and carbon (feed) into carbon dioxide by bacteria in the tank causes more oxygen to be consumed and more carbon dioxide to be produced in the tank than what is consumed and produced by the salmon alone. Thus, the accumulated oxygen would be lower and the accumulated carbon dioxide would be higher than what would be expected by the salmon. For the sum of the accumulated substances, this effect would cancel out, but taking scaling into consideration would yield a lower relative sum of the accumulated substances. Algae performs photosyntesis reducing the carbon dioxide concentration and increasing the oxygen concentration in the tank. Taking this into account would dampen the consumption of oxygen and accumulation of carbon dioxide, and as for the bacteria the effects would cancel out considering the sum of the accumulated substances. It would however increase the relative sum, as the bacteria would decrease the relative sum. According to Kvalsvik (2022) and Stakvik-Løvås (2022), the presence of bacteria can be assumed to take place with the salmon, but the algea is filtered away by particle filters and UV lights. As the effect of the bacteria is expected to cancel out in the sum of the accumulated substances, it is not a directly explainable factor of the deviation between the assumptions of dos Santos *et al.* (2022) and our data and assumptions, but the effect is probably still a part of the discrepancy function.

Another effect affecting the concentrations in the tank is the pH and alkalinity in the water. For explaining how this affects the processes in the tank, the chemical equilibrium

$$\mathrm{CO_2 + H_2 = \rightleftharpoons H_2\,CO_3},$$

where carbon dioxide reacts with water into carbonic acid in basic environments and carbonic acid reacts into carbon dioxide and water in acicid environments. The carbon dioxide concentration thus becomes reduced if the encironment in the tank is basic and increased if it is acidic. Ammonia, NH_3 is also emitted by the salmon affecting the alkalinity of the water by reducing the pH giving a more acidic environment in the tank. According to Kvalsvik (2022), the pH in the tank is approximately 6.7, and there is thus a slight acidic environment. Carbon dioxide emitted by the salmon does thus not react into carbonic acid directly as it would in a basic environment, and the alkalinity effects of the water does not have a large impact on the obtained results.

There is in Chapter 2.4 assumed that the substance concentration in the outlet water of the tank is equal to the substance concentration in the tank at the previous timestamp. In reality, we could expect lower substance concentrations near the center of the tank where the water is taken out than near the edge where the water is injected, with a stronger effect for oxygen than for carbon dioxide according to Kvalsvik (2022) and Stakvik-Løvås (2022). This causes the model to show that less oxygen has vanished from the tank than there actually has. This does however not explain the deviation of the sum of posterior discrepancies in the model compared to what one would expect, but it is an effect to keep into cosideration.

The last effect mentioned by Kvalsvik (2022) and Stakvik-Løvås (2022), are sensor uncertainties. As introduced in Chapter 2, the salmon production facility is a greenfield production site. Thus, SE are early in the process of introducing the sensors for the obtained data, and there are no common practices in the business on how to install and calubrate sensors. This has an advantage of competition as SE is one of a fiew prosecutor in the market able to keep track of their collected

data, but there are challenges in calibrating the sensors and the data flow to obtain results based on the data with a large level of confidence. As the facility has been in operation for a short period of time, the sensors have been tested to give values making sense, but they are according to Stakvik-Løvås (2022), not yet fully calibrated to yield perfect measurements at any time. As a note, the data considered in this thesis is thoroughly considered throughout the whole period of collected data and updated several times as appearingly better data becomes tracked.

There are as well made assumptions on physical parameters that vary with natural parameters. The oxygen and carbon dioxide concentrations in fresh sea water vary with temperature, preassure and salinity which in turn may depend on sea currents, weather and nearby activities of the production facility. The effect of the carbon dioxide degasser also introduces a potential error, as it has been adjusted several times as when this project was performed.

To sum up, the time required by the process of combusting oxygen into carbon dioxide into causing a lag of the carbon dioxide emission compared to the oxygen consumption, and degassing of oxygen inside the tank may explain the deviation of posterior mean sum of accumulated oxygen and carbon dioxide. Bacteria, algea, alkalinity and pH does seem to have a that large impact on this result. The obtained results seem, taking the potential sensor and assumption uncertainties into account, to fit prior knowledge about the physical process relatively well.

6.3 Model approach improvement

It is now evident that fitting observed data and parameter assumptions regarding oxygen and carbon dioxide concentrations in a tank of salmon as latent Gaussian models may give reasonable results for the relationship between consumed oxygen and produced carbon dioxide. There are however some challenges and uncertainties evident from the results that could be solved or eliminated improving or developing the approach of the model further. In this thesis, much time has been spent on finding and treating appropriate data for the purpose. Trying to fit more prior distributions to the model could possibly improve the model results.

A natural step in improving the model apprach, making in valuable for the industry will be to fit the model and data over different similar datasets. The salmon in the tank had a certain average weight and age, and fitting the mode over several other datasets could improve the unterstanding of the modelled phenomenon.

There are several aspects that are left out in the model, limited by the availability of data of parameters. For instance, the feeding has a natural impact of the oxygen consumption and carbon dioxide production i the tank. The data in this project was considered during a feeding break simplifying the feeding aspect, but taking feed into consideration would improve the understanding further, enabling the model to account for the relation between feed consumption and oxygen consumption. Nitrogen combinations in the tank, denoted total ammonium nitrate (TAN) are also a limiting factor of production, which could have been consudered in such a model. As well, the average growth rate of the salmon in the tank is affected by the average weight, and theese parameters both in turn affects the feed and oxygen consumed and carbon dioxide emitted. It is difficult to keep track of this continuously, even in a land-based facility, but systematic measurments could include this understanding into the model as well.

Some simplified assumptions were made when assessin the provided data, as describred in Chapter 2.4. A more accurate model could be obtained if some assumed measurements or properties where accurately measured. The oxygen and carbon dioxide of fresh sea water could for instance be measured taking a water sample of fresh sea water after preliminary treatment into the tank. Further, in Chapter 2.4 it is assumed that the substance concentrations out of the tank is equal to the substance concentration of the precious timestamps, as discussed above. In reality, according to Kvalsvik (2022) and Stakvik-Løvås (2022), the water is taken out of the tank near the center, and injected near the edge. Taking measurements of the oxygen and carbon dioxide concenrations near the center and near the edge of the tank, could create a possibility of obtaining concenration profiles as a function of distance to the center and asen on this be able to assume the concentrations in the outtake water based on what is measured in the tank. Another way could be to place sensors to

measure time series in the outtake and intake water pipes to get data for all assumed parameters.

7 Conclusion

In land-based aquaculture, the focus has shifted from primarily addressing issues related to diseases and parasites at sea, to optimizing growth through manipulation of water temperatures and feeding schedules, while also considering biological limits such as biomass density and carbon dioxide concentrations. This has allowed the salmon farming industry to increase production without negatively impacting the local environment. The use of land-based systems also allows for the collection of data that would not be feasible in marine environments, enabling improved monitoring of production and the use of data for analysis and decision-making.

Through the use of Bayesian inference, we were able to analyze data from a single tank at Salmon Evolution's production facility to investigate the relationship between models based on oxygen and carbon dioxide concentrations. We used latent Gaussian models with assumed smooth random effects with chosen random walk 1 priors for the accumulated carbon dioxide and negatively accumulated oxygen in the tank. Our analysis revealed that strict priors on the Gaussian effect parameters were necessary for the model to fit the data accurately. Additionally, implementing strict penalized complexity (PC) priors for the variance of the random walk 1 priors improved the accuracy of the accumulated oxygen and carbon dioxide estimates. Overall, the model that included PC priors for the variance of the random walk 1 priors and accounted for the variance of all quantities within the fixed effect parameters using the delta method performed the best according to the deviance information criteria. However, the differences in model performance between the different models with prior distributions were relatively small.

Our analysis using Bayesian inference on latent Gaussian models of the oxygen and carbon dioxide concentration in the tank revealed that the salmon appeared to be consuming more oxygen than it was producing carbon dioxide corresponding to. This discrepancy may be explained by factors not accounted for in the model assumptions. One possibility is that the salmon undergoes a process of converting oxygen to carbon dioxide that requires time, which is not captured in the model. Additionally, the tank exhibits a degassing effect, leading to the dissipation of oxygen from the water, which is not accounted for in the model assumptions. Other factors such as bacteria, the alkalinity of the water, and sensor uncertainties may also be contributing to the observed results, but are not considered to have a significant impact. Additionally, the properties of individual salmon and the differences between batches of salmon can impact the results, and overall, the obtained results were reasonably accurate given the uncertainty inherent in this case.

In order to further improve our analysis of oxygen and carbon dioxide concentrations in a tank for salmon farming, there are several potential areas for improvement. One possibility is to carefully evaluate the performance of the sensors used to measure these concentrations, as sensor errors could potentially impact the accuracy of our results. Additionally, obtaining more data from the facility, particularly from different locations within the tank, could help to better understand the processes at play. It may also be beneficial to fit the model to different datasets, as individual and batch-to-batch variations in properties could affect the results. However, it is important to consider that adding more sensor data could also introduce additional uncertainties into the model. Ultimately, by carefully considering these factors and collecting more data, it may be possible to further refine our understanding of the relationships between oxygen and carbon dioxide concentrations in a tank for salmon farming.

References

Akaike, H. 1973. 'Information Theory and an Extension of the Maximum Likelihood Principle'. *Second International Symposium on Information Theory,* 267–281.

Andfjord Salmon. 2022. 'Technology'. Accessed 1 November 2022. https://www.andfjordsalmon.com/technology.

Artec Aqua. n.d. 'Technologies'. Accessed 1 November 2022. http://www.artec-aqua.com/technologies/.

Baio, Gianluca. 2012. *Bayesian Methods in Health Economics.* London, UK: Chapman Hall/CRC Biostatistics Series.

Betancourt, M. 2018. 'The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo'. *Special Issue: Physics of Information* 531 (3).

Bivand, Roger S., Virgilio Gómez-Rubio and Håvard Rue. 2014. 'Approximate Bayesian inference for spatial econometrics models'. *Spatial Statistics* 9:146–165.

Blangiardio, Marta, and Michela Cameletti. 2015. *Spatial and Spatio-tempotal Bayesian Models with R-INLA.* John Wiley Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom: John Wiley Sons Ltd.

Blei, D. M., A. Kucukelbir and J. D. McAuliffe. 2017. 'Variational inference: A review for statisticians'. *Journal of the American Statistical Association* 112 (518): 859–877.

Box, George E. P., and Norman Darper. 1987. *Empirical Model-Building and Response Surfaces.* John Wiley Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom: John Wiley Sons Ltd.

Brooks, Steeve, Andrew Gelman, Galin L. Jones and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo.* Boca Raton, FL: Chapman Hall/CRC Press.

Casella, G., and E. I. George. 1992. 'Explaining the Gibbs Sampler'. *The American Statistican* 46 (3): 167–174.

Clarify. 2022. 'Clarify - Time series intelligence for industrial teams'. Accessed 18 November 2022. https://www.clarify.io/.

Davison, Andy. 2018. *Recirculating Aquaculture Systems : A Guide to Farm Design and Operations.* Seattle, Washington, United States: Farmfish LLC.

Fahrmeir, Ludwig, and Gerhard Tutz. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer Publishing 11 West 42nd Street, Manhattan, New York, United States: Springer.

FAO. 2020. *The State of World Fisheries and Aquaculture 2020 - Sustainability in Action.* Technical report. Rome, Italy: FAO.

Forsberg, Odd I. 1994. 'Modelling Oxygen Consumption Rates of Post-Smolt Atlantic Salmon in Commercial-Scale, Land-Based Farms'. *Aquaculture International* 2:180–196.

Gelfand, A. E., and A. F. Smith. 1990. 'Sampling-Based Approaches to Calculating Marginal Densities'. *Journal of the American Statistical Association* 85 (410): 398–409.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin. 2013. *Bayesian Data Analysis.* Boca Raton, FL: Chapman Hall/CRC Press.

Gelman, Andrew, and Donald B. Rubin. 1992. 'Inference from Iterative Simulation Using Multiple Sequences'. *Statistical Science* 7:457–511.

Gómez-Rubio, Virgilio. 2020. *Bayesian Inference with INLA.* Boca Raton, FL: Chapman Hall/CRC Press.

Held, Leonhard, and Håvard Rue. 2010. 'Conditional and Intristic Autoregressions'. *Handbook of Spatial Statistics,* 201–216.

Johnson, Richard A., and Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis.* 6. ed. Upper Saddle River, NJ 07458: Pearson Prentice Hall.

Klein, Lawrence Robert. 1953. *A Textbook of Econometrics.*

Kvalsvik, S. 2022. *Water Quality Specialist Salmon Evolution.* private communication, 16 December 2022.

Lesaffre, Emmanuel, and Andrew B. Lawson. 2012. *Bayesian Biostatistics.* John Wiley Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom: John Wiley Sons Ltd.

Logio Systems. 2022. 'LOLIGO ONLINE OXYGEN CONVERTER'. Accessed 18 December 2022. https://www.loligosystems.com/convert-oxygen-units?menu=77.

Lyngstad, O. J. 2022. *Technical Manager Biology and Aquaculture.* private communication, 24 February 2022.

Martino, Sara, and Andrea Riebler. 2019. 'Integrated Nested Laplace Approximations (INLA)'. August.

Martino, Sara, and Håvard Rue. 2009. *Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: a manual for the inla program.* Department of Mathematical Sciences NTNU.

Moraga, Paula. 2019. *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny.* Chapman Hall/CRC Biostatistics Series.

O'Hagan, Anthony, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley and Tim Rakow. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities.* John Wiley Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom: John Wiley Sons Ltd.

ProductionPlan. 2022. 'Greenfield Project'. Accessed 17 December 2022. https://www.productplan.com/glossary/greenfield-project/.

Rue, Håvard, Sara Martino and Nicholas Chopin. 2009. 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations'. *Journal of the Royal Statistical Society* 71 (2): 319–392.

Rue, Håvard, Andrea Riebler, Sigrunn H. Sørbye, Hanine B. Illian, Daniel P. Simpson and Finn K. Lindgren. 2016. 'Bayesian Computing with INLA: A Review'. *Annual Review of Statistics and Its Application* 4:395–421.

Sætre, Simen, and Kjetil Østli. 2021. *Den nye fisken - Om temmingen av laksen og alt det forunderlige som fulgte.* Pb 6673 St. Olavs Plass, 0129 Oslo, Norway: Spartacus Forlag AS.

Salmon Evolution. 2022. 'Technology - Hybrid flow-through system'. Accessed 1 November 2022. https://salmonevolution.no/technology/.

Schwartz, G. E. 1978. 'Estimating the Dimension of a Model'. *Annals of Statistics* 6 (2): 461–464.

Simpson, D., H. Rue, A. Riebler, T. G. Martins and S. H. Sørbye. 2017. 'Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors'. *Statistical Science* 32 (1): 1–28.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin and A. van der Linde. 2002. 'Bayesian measures of model complexity and fit'. *Statistical Methodology Series B* 64 (4): 583–639.

Stakvik-Løvås, J. Å. 2022. *Automation Engineer Salmon Evolution.* private communication, 16 December 2022.

Taranger, G. L., Ø. Karlsen, R. J. Bannister, K. A. Glover, Husa V., E. Karlsbakk, B. O. Kvamme *et al.* 2015. 'Risk assessment of the environmental impact of Norwegian Atlantic salmon farming'. *ICES Journal of Marine Science* 72 (3): 997–1021.

Theodoris, Sergios. 2015. *Machine Learning - A Bayesian and Optimization Perspective.* London Wall, London, UK: Elsevier Academic Press.

Tierney, Luke, and Joseph B. Kadane. 1986. 'Accurate Approximation of Posterior Moments and Marginal Densities'. *Journal of the American Statistical Association* 81 (393): 82–86.