Master's thesis
Master's Programme in Data Science

Model selection and sensitivity analysis in a class of Bayesian spatial distribution models

Neli Noykova

May 19, 2022

Supervisor(s): Dr. Jarno Vanhatalo

Examiner(s): Dr. Indrè Žliobaitè

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Data Science

Tekijä — Författare — Author

Neli Noykova

Työn nimi — Arbetets titel — Title

Model selection and sensitivity analysis in a class of Bayesian spatial distribution models

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	May 19, 2022	55

Tiivistelmä — Referat — Abstract

This work is focused on Bayesian hierarchical modeling of geographical distribution of marine species *Coregonus lavaretus L. s.l.* along the Gulf of Bothnia. Spatial dependences are modeled by Gaussian processes. The main modeling objective is to predict whitefish larvae distribution for previously unobserved spatial locations along the Gulf of Bothnia. In order to achieve this objective, we have to solve two main tasks: to investigate the sensitivity of posterior parameters estimates with respect to different parameter priors, and to solve model selection task. In model selection, among all candidate models, we have to choose the model with best predictive performance. The candidate models were divided into two main groups: models that describe spatial effects, and models without such description. The candidates in each group involved different number (6 or 8) and expressions of environmental variables. In the group describing spatial effects, we analyzed four different models of Gaussian mean, and for every mean model we used four different prior parameters combinations. The same four models of latent function were used in the candidates where spatial dependences were not described. For every such model we assigned four different priors of overdispersion parameter. Thus, all at all, 32 candidate models were analyzed.

All candidate models were estimated with Hamiltonian Monte Carlo MCMC algorithm. Model checks were conducted using the posterior predictive distributions. The predictive distributions were evaluated using the logarithmic score with 10 fold cross validation.

The analysis of posterior estimates in models describing spatial effects revealed, that these estimates were very sensitive to prior parameters choices. The provided sensitivity analysis helped us to choose the most suitable priors combination. The results from model selection showed that the model, which showed best predictive performance, does not need to be very complicated and to involve description of spatial effects when the data are not informative enough to detect well the spatial effects. Although the selected model was simpler, the corresponding predictive maps of log larvae intensity correctly predicted the larvae distribution along the Gulf of Bothnia.

ACM Computing Classification System (CCS):
Applied computing → Physical sciences → Earth and atmospheric sciences → Environmental sciences

Avainsanat — Nyckelord — Keywords

SDMs, Bayesian hierarchical modeling, GP, Model validation, Model selection, Sensitivity analysis

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

Species distribution models (SDMs) play central role for studying dynamics of different species and their geographical distribution [27, 29, 30]. Since different SDMs are created for different purposes and describe different phenomena, it is very important to know well the main features of both ecological processes and existing modeling approaches. It is essential to understand the geographical and environmental dimensions of species, specifics of modeling marine species, and the problems that could appear because of data quality. Based on this knowledge, the modeling goals, and information about strengths and weakness of different SDMs, it is possible to make the most appropriate model choice.

The era of big data and development of the geographical information system (GIS) have brought new information and changed a lot the trends in developing SDMs. The new available data are on different scale and quality, and possess different structure [5, 15, 29]. In order to transform this information to valuable knowledge, novel ways to produce, store and analyze the data are needed. Thus the field of data science was born to bring together computational, algorithmic, statistical and mathematical methods and techniques towards extrapolating knowledge from big data. The new models have to be able to answer to new, more detailed questions.

Bayesian approach to data science is model-based, capable to develop new understanding, can appropriately quantify and propagate uncertainty, and through hierarchical models is able to use population-level information to make inferences and predictions about species distribution on unobserved locations. Bayesian SDMs are hierarchical statistical models that explain spatial pattern using environmental covariates [32, 33, 36, 38]. These models are complex, cannot be solved analytically and require integrating over uncertainty. Such high dimensional integration is computationally expensive comparing to other optimization-based data science methods [27, 29].

Big data availability have inspired also development of different numerical methods, among them Markov Chain Monte Carlo (MCMC) method [4, 32, 33, 35]. MCMC is very efficient for solving such complicated models that use big data, and made feasible and attractive the implementation of Bayesian approach to SDMs.

The recent achievements in data science approach to SDMs led to intensive research on Bayesian hierarchical spatial regression models describing dynamics of some marine

species [22, 33, 34, 35, 36, 38]. These models are not only capable to make predictions, but also provide new insights and understanding on species dynamics.

This work in this thesis is focused on Bayesian hierarchical modeling of geographical distribution of marine species *Coregonus lavaretus L. s.l.* along the Gulf of Bothnia. Descrbed whitefish has different environmental preferences during its lifespan (larvae or adult). Larvae occupy smaller areas and are very sensitive to the environmental conditions. These smaller reproduction areas can limit the fish production [36]. The main modeling objective is to predict whitefish larvae distribution for previously unobserved spatial locations along the Gulf of Bothnia.

In order to achieve this objective, we have to select the model with best predictive performance among the set of models with different level of complexity and different prior combinations of involved parameters. Thus, we have to solve two main tasks: to investigate the sensitivity of posterior parameters estimates with respect to different parameter priors, and to solve model selection task. The choice of right combination of parameter priors in this class of complicated hierarchical models determines the quality of posterior results. In model selection task, among all candidates we have to choose the model with best predictive performance. By selecting the best model we also clarify which environmental variables are more influential and have to be included in the model, and how important is to model spatial effects for particular data that are available.

The work starts with description of theoretical concepts used in hierarchical species distribution modeling and representation of species as points in space. Next two chapters introduce the foundations of Bayesian inference and prediction for modeling spatial data, and describe the main principles and algorithms for Markov chain Monte Carlo sampling. The prior distributions of all parameters, which are used in this work, and reasoning to choose them, are explained in Chapter 5. Theoretical foundations of information criterion, used for model selection, are presented in Chapter 6. Next the used data and modeling workflow are described. The modeling experiment, where all model candidates are stated, is explained in Chapter 9. In Chapter 10, the results of model performance for all candidates are summarized. The model selection results are presented in Chapter 11. Next the predictive performance of the best chosen model is shown. At the end all results are discussed, and conclusions are drawn.

2. Theoretical background

2.1 Hierarchical species distribution modeling

2.1.1 Data types

The data type and quality of measurements affect strongly the choice of appropriate modeling method. Environmental variables describe abiotic environment. They could be related to climate (temperature, precipitation), topography or seabed type in marine ecosystems [28]. In the experiment described here 22 environmental variables were measured as Geographical Information System (GIS) map layers. Raster layers were converted to prediction grids.

The most common data type of observed species are occurrence or abundance data. Occurrence data are usually in binary form, and describe presence (1) or absence (0) of species at given locations. In many cases only records about the locations, where species live, are available (so called presence-only data) [26]. This introduces uncertainty about the unobserved sites - if they are occupied or not [18].

Abundance data could be count or continuous. Continuous data express species biomass. Count data correspond to the number of individuals at given location.

Count data could be collected by systematic surveys, such as point counts or quadrat counts [20]. In both cases the data do not cover the whole study region. The data are collected from a number of sampling sites of finite area (or volume for marine species) during a fixed period of time [33]. In this work we use count data.

2.1.2 Species as points in space, modeling of spatial point patterns

In this work, we aim to analyze and predict the underlying spatial pattern of whitefish larvae abundance in the Gulf of Bothnia. For this purpose we need to model the process that constructs patterns of points in the study region, denoted by D. Since an individual of a species exists as a point in space, we treat the whitefish larvae counts as points in space [20, 25, 33]. The total number of points in D is $n' = N(D)$. We assume that the available

count (or occurrence presence/absence data) arise from point process distribution. The set of locations, where the whitefish larvae are observed, $\mathbf{S} = \{\mathbf{s_1}, \ldots, \mathbf{s'_n}\}$, is a random variable and we need to assign a probabilistic model to it. For any $n' \in 0, 1, 2, \ldots$ we denote a multivariate probability density over $D^{n'}$ by $p(\mathbf{s_1}, \ldots, \mathbf{s'_n})$. If we denote by $|\partial s_i|$ the area of arbitrary small circular neighborhood around s_i, the joint probability for point pattern $\mathbf{S}$ is expressed as [2]

$$P(\mathbf{S}) \approx P(N(D) = n')p(\mathbf{s_1}, \ldots, \mathbf{s'_n}|P(N(D) = n') \prod_{i=1}^{n'} |\partial s_i| \tag{2.1}$$

In Eq. (2.1) we have to define the probability distribution $P(N(D) = n')$ and density $p(\mathbf{s_1}, \ldots, \mathbf{s'_n}|P(N(D) = n')$.

In the Poisson process, the number of species $N(B)$ in the subset $B \subset D$, where D is a bounded region, follows the Poisson distribution, $N(B) \sim \text{Poisson}(\lambda(B))$, where $\lambda(B) = \int_B \lambda(\mathbf{s})d\mathbf{s}$ and $\lambda(\mathbf{s})$ is the intensity function of the process [2, 33]. The intensity function $\lambda(s)$ is given. In Poisson process if B_1 and B_2 are disjoint, then $N(B_1)$ and $N(B_2)$ are independent.

The probability distribution of species counts in region D is

$$P(N(D) = n') \sim \text{Poisson}(\lambda(D)) \tag{2.2}$$

where $\lambda(D)$ is the total intensity over D, $\lambda(D) = \int_D \lambda(\mathbf{s})d\mathbf{s}$.

Then the probability $p(\mathbf{s_1}, \ldots, \mathbf{s'_n}|N(D) = n')$ in a Poisson process over region D is expressed as

$$p(\mathbf{s_1}, \ldots, \mathbf{s'_n}|N(D) = n') = \prod_{i=1}^{n'} p(s_i) = \prod_{i=1}^{n'} \lambda(s_i)/\lambda(D) \tag{2.3}$$

The Poisson process is a conditional process $p(\mathbf{s_1}, \ldots, \mathbf{s'_n}|\lambda(s))$, where the intensity $\lambda(s)$ is known. When the intensity surface λ is assumed as random, the resulting process is called Cox process. In the Cox point process we need to marginalize over $\lambda(s)$ and obtain $p(\mathbf{s_1}, \ldots, \mathbf{s'_n}) = E[p[\mathbf{s_1}, \ldots, \mathbf{s'_n}|\lambda(s)]]$. In this work we model the spatial whitefish larvae distribution by so called log Gaussian Cox process, where $\log(\lambda(s))$ follows a Gaussian process [33].

The likelihood after observing the point pattern $\mathbf{S} = \{\mathbf{s_1}, \ldots, \mathbf{s'_n}\}$ in fully observed region D is

$$L(\mathbf{s_1}, \ldots, \mathbf{s'_n}, N(D) = n'|\lambda(\mathbf{s})) \propto e^{-\lambda(D)} \prod_{i=1}^{n'} \lambda(s_i) \tag{2.4}$$

Since the integral in $\lambda(D) = \int_D \lambda(\mathbf{s})d\mathbf{s}$ cannot be solved in closed form, we need to find an approximate solution of Eq. (2.4). For this purpose D can be partitioned into

dense grid, and the counts $N(B_i)$ of every grid cell $B_i, i = 1, \dots n$ are assumed to be available. If the grid is dense enough, it is possible to approximate $\lambda(B_i) \approx \lambda(s_i)|B_i|$, where s_i is the center location of B_i.

The likelihood of counts over grid cells are approximated as:

$$p(N(B_1), \dots, N(B_n)|\lambda(s)) = \prod_{i=1}^{n} \text{Poisson}(N(B_i)|\lambda(s_i)|B_i|) \tag{2.5}$$

In the model of whitefish larvae counts the intensity function $\lambda(s)$ is the probability that one larvae present at location $\mathbf{s}$, $B_i = V_i$ is the sampled volume of water, $N(B_i) = y_i$ are observations of whitefish larvae. In the experimental data, used in this work, many sampling sites were observed, but not the whole region D. Hence, we can model the data as obtained from Poisson process by so called thinning method [33].

Then the species counts are modeled as:

$$p(y_1, \dots, y_n|\lambda(s)) = \prod_{i=1}^{n} \text{Poisson}(y_i|\lambda(s_i)V_i) \tag{2.6}$$

In the cases when counts from sampling sites have larger variance than predicted by Poisson distribution, it is plausible to assume a Negative-Binomial distribution for the count observations. The Negative-Binomial model can be derived from Eq. (2.6) by adding random effect to the rate parameter of Poisson distribution ($\lambda(s_i)V_i\varepsilon_i$ instead of $\lambda(s_i)V_i$). If we give a Gamma prior for the random effect and marginalize over it, the observation model becomes:

$$\begin{aligned} p(y_1, \dots, y_n|\lambda(s)) &= \prod_{i=1}^{n} \text{Negative Binomial}(y_i|\lambda_i V_i, r) \\ &= \frac{\Gamma(y_i + r)}{y_i!\Gamma(r)} \left(\frac{r}{r + V_i\lambda_i}\right)^r \left(\frac{V_i\lambda_i}{r + V_i\lambda_i}\right)^{y_i} \end{aligned} \tag{2.7}$$

where r is an overdispersion parameter. The expectation of the Negative Binomial distribution in Eq. (2.7) is $E[y_i] = V_i\lambda_i$, and the variance is $Var[y_i] = V_i\lambda_i(1 + V_i\lambda_i/r)$. The expectation and variance of Poisson observation model, given in Eq. (2.6), are $E[y_i] = Var[y_i] = V_i\lambda_i$. In the limit case when $r \to \infty$ the variance of Negative Binomial approaches the variance of Poisson model.

2.1.3 Hierachical modeling framework

As it was described in the previous section, the point pattern process is chosen as log Gaussian Cox process, described as:

$$\log \lambda(\mathbf{s}, \mathbf{x}(\mathbf{s})) = f(\mathbf{s}, \mathbf{x}(\mathbf{s})) \tag{2.8}$$

$$f(\mathbf{s}, \mathbf{x}(\mathbf{s})) = \beta_0 + \mathbf{x}(\mathbf{s})^{\mathbf{T}}\beta + \phi(\mathbf{s}) \tag{2.9}$$

where $f(\mathbf{s}, \mathbf{x}(\mathbf{s})$ is a Gaussian latent function, and $\mathbf{x}(\mathbf{s}) = [\mathbf{x_1}(\mathbf{s}), \ldots, \mathbf{x_n}(\mathbf{s})]^T$ is a vector of environmental variables, measured at locations $\mathbf{s}$. The linear weights $\beta \sim \mathcal{N}(0, \Sigma_\beta)$ are mutually independent, β_0 is the intercept, and $\phi(\mathbf{s})$ is a spatial Gaussian process.

Eq. (2.9) allows to place the Negative Binomial point process observational model, given by Eq. (2.7), within a Bayesian hierarchical structure, described as [33]:

$$\text{[Data | process, parameters]} \qquad \mathbf{y} \sim p(\mathbf{y}|f(\mathbf{s}, \mathbf{x}(\mathbf{s}), \gamma)) \tag{2.10}$$

$$\text{[process | parameters]} \qquad f(\mathbf{s}, \mathbf{x}(\mathbf{s}))|\theta) \sim GP(m(\cdot|\theta), k(\cdot, \cdot|\theta)) \tag{2.11}$$

$$\text{[(hyper)parameters]} \qquad \theta, \gamma \sim p(\theta, \gamma) \tag{2.12}$$

Eq. (2.10) - (2.12) specify a flexible model where the process f and observations $\mathbf{y}$ are separated. The first layer corresponds to the observational model given by Eq. (2.7) and Eq. (2.8). γ denotes the observation model parameters. In our case $\gamma = r$, the oversdispersion parameter in Eq. (2.7). The second layer specifies the prior for latent process f conditionally to the parameters of covariance and mean functions θ. In our model $\theta = [\alpha, \beta, \sigma^2_{exp}, 1/l]$. On the third level priors of all parameters are defined.

A Gaussian process GP is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a joint multivariate normal distribution [33]. The spatial Gaussian process in the model used here is defined as

$$\phi(\mathbf{s}) = GP(m(\mathbf{x}(\mathbf{s})), k(\mathbf{s}, \mathbf{s}')) \tag{2.13}$$

The prior mean m describes the linear part of latent variable f

$$m = \beta_0 + \mathbf{x}(\mathbf{s})^T \beta \tag{2.14}$$

The spatial covariance function $k(\mathbf{s}, \mathbf{s}')$ in Eq. 2.13 is modeled here by exponential covariance function

$$k_{exp}(\mathbf{s_i}, \mathbf{s_j}) = \sigma^2_{exp} e^{-\|\mathbf{s_i} - \mathbf{s_j}\|/l} \tag{2.15}$$

where $\|\mathbf{s_i} - \mathbf{s_j}\|$ is the Euclidean distance between $\mathbf{s_i}$ and $\mathbf{s_j}$, σ^2_{exp} is the magnitude, and l is the length scale, which determines how fast correlation function decreases when distance increases. Exponential covariance function is stationary and isotropic because it is invariant to transitions in index domain, and depends only on Euclidean distance $\|\mathbf{s_i} - \mathbf{s_j}\|$ between two different geographical locations $\mathbf{s_i}$ and $\mathbf{s_j}$ [33]. This is in accordance with "the first law of geography", saying that "near things are similar because they influence each other or are influenced by the same pattern generating processes" [28]. The spatial effect, describing point pattern process in Eq. 2.7, does not depend directly on the distance between different locations $\mathbf{s}$ and $\mathbf{s}'$. The observations $\mathbf{y}(\mathbf{s})$ and $\mathbf{y}(\mathbf{s}')$ can be spatially dependent, but not need to be close to each other [2, 11].

2.2 Bayesian inference and prediction

2.2.1 Basic ideas of Bayesian inference

Bayesian inference answers the question "What we can learn about parameters θ given that we have observed y?"

The answer is provided in three steps. First we define all parameter priors in Eq. (2.12). Next we model the sampling distributions of f and y (Eq. (2.10) and (2.11)) given θ. We update our knowledge about the unknown parameters by computing posterior distribution $p(\theta|y)$ according the Bayes' theorem [6, 13]

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2.16}$$

The marginal distribution $p(y) = \int p(y|\theta)p(\theta)d\theta$ does not depend on θ, but can be intractable to solve. In such cases for determining posterior $p(\theta|y)$ we need to use methods that do not explicitly compute $p(y)$.

We need to analyze the posterior estimates of β since one of our main goals is to understand the effect of environmental variables on species distribution. Posterior samples of β reveal weather a particular environmental covariate has a significant impact on species intensity. Posterior analysis of covariance parameters σ^2_{exp} and $1/l$ describes the strength of spatial association between neighboring locations after adjusting for covariate effects.

The second, even more important goal of Bayesian analysis provided here, is to solve prediction task. We aim to construct maps of whitefish larvae intensities $\lambda(s)$ over entire study region, Gulf of Bothnia. To do this, we need to predict the distribution of latent variable $\tilde{f}$ and larvae counts $\tilde{y}$ on unobserved locations $\tilde{s}$. The corresponding posterior predictive distributions $p(\tilde{f}|f)$ and $p(\tilde{y}|y)$ can be computed using Eq. (2.17) and Eq. (2.18)

$$p(\tilde{f}|f) = \int p(\tilde{f}|\theta)p(\theta|f)d\theta \tag{2.17}$$

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta \tag{2.18}$$

2.2.2 Posterior inference and spatial prediction of latent function and observational larvae counts

The latent function defined in Eq.(2.9) is normally distributed since both additive components $\beta_0 + x(s)^T\beta$ and ϕ are Gaussian. Then the marginal distribution of $\mathbf{f} = [f(s_1), \ldots, f(s_n)]$ is also Gaussian [33]:

$$\mathbf{f}|\mathbf{S}, \mathbf{X}(\mathbf{S}), \theta \sim GP(\mathbf{0}, \mathbf{X}(\mathbf{S})\mathbf{\Sigma}_\beta\mathbf{X}(\mathbf{S})^\mathbf{T} + \mathbf{K}_{\phi,\phi}) \tag{2.19}$$

where $\mathbf{X}(\mathbf{S}) = [x(s_1), \ldots, x(s_n)]$, $\mathbf{K}_{\phi,\phi} = Cov(\phi, \phi)$, $\phi = [\phi(s_1), \ldots, \phi(s_n)]^T$. The elements of the spatial covariance matrix $\mathbf{K}_{\phi,\phi}$ are computed according Eq.(2.15). The covariance of linear part is $\mathbf{k}(\mathbf{x}(\mathbf{s}), \mathbf{x}'(\mathbf{s}')) = \mathbf{x}(\mathbf{s})\mathbf{\Sigma}_\beta\mathbf{x}(\mathbf{s})^\mathbf{T}$

If we denote the value of latent variable at unobserved locations $\tilde{S}$ by $\tilde{f}$, and environmental covariates at $\tilde{S}$ by $\tilde{X}$, the joint prior for latent variables at locations $\mathbf{S}$ and $\tilde{\mathbf{S}}$ is [33]

$$\begin{bmatrix} \mathbf{f} \\ \tilde{\mathbf{f}} \end{bmatrix} |\mathbf{S}, \mathbf{X}, \tilde{\mathbf{S}}, \tilde{\mathbf{X}}, \theta \sim N\left(0, \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}} \\ \mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}} & \mathbf{K}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}} \end{bmatrix}\right) \tag{2.20}$$

where $\mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}} = \tilde{\mathbf{X}}(\tilde{\mathbf{S}})\mathbf{\Sigma}_\beta\mathbf{X}(\mathbf{S}) + \mathbf{K}_{\tilde{\phi},\phi}, \mathbf{K}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}} = \tilde{\mathbf{X}}(\tilde{\mathbf{S}})\mathbf{\Sigma}_\beta\tilde{\mathbf{X}}(\tilde{\mathbf{S}}) + \mathbf{K}_{\tilde{\phi},\tilde{\phi}}, \mathbf{K}_{\mathbf{f},\mathbf{f}} = \mathbf{X}(\mathbf{S})\mathbf{\Sigma}_\beta\mathbf{X}(\mathbf{S}) + \mathbf{K}_{\phi,\phi}$

The computation of posterior distribution in hierarchical Bayesian models has to be performed in several steps [33, 35]. To improve readability, we further use the brief notations $\mathbf{f} = \mathbf{f}(\mathbf{X}(\mathbf{S}), \mathbf{S})$, $\mathbf{y} = \mathbf{y}(\mathbf{X}(\mathbf{S}), \mathbf{S})$, $\tilde{\mathbf{f}} = \tilde{\mathbf{f}}(\mathbf{X}(\mathbf{S}), \tilde{\mathbf{X}}(\tilde{\mathbf{S}}), \tilde{\mathbf{S}}, \mathbf{S})$, $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}(\mathbf{X}(\mathbf{S}), \tilde{\mathbf{X}}(\tilde{\mathbf{S}}), \tilde{\mathbf{S}}, \mathbf{S})$

1. The full posterior of hyperparameters θ and γ and latent variable $\mathbf{f}$ can be expressed by applying the Bayes theorem to hierarchical model Eq. (2.10) - Eq. (2.12):

$$p(\mathbf{f}, \theta, \gamma|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f}, \gamma)p(\mathbf{f}|\theta)p(\theta, \gamma)}{p(\mathbf{y})} \tag{2.21}$$

The evaluation of the full posterior (Eq. (2.21)) can be provided in two steps:

1.1 First compute the marginal posterior for the hyperparameters $p(\mathbf{f}, \theta, \gamma|\mathbf{y})$

$$p(\theta, \gamma|\mathbf{y}) = \frac{1}{Z}p(\mathbf{y}|\theta, \gamma)p(\theta, \gamma) \tag{2.22}$$

where the normalizing constant $Z = \int p(\mathbf{y}|\theta, \gamma)p(\theta, \gamma)d\theta d\gamma$

1.2 Next compute the marginal posterior of latent variables $p(\mathbf{f}|\mathbf{y})$ by marginalizing over the posterior of hyperparameters, obtained in step 1.1.

$$p(\mathbf{f}|\mathbf{y}) = \int p(\mathbf{f}|\mathbf{y}, \theta, \gamma)p(\theta, \gamma|\mathbf{y})d\theta d\gamma \tag{2.23}$$

2. Determine posterior predictive distribution of $\tilde{\mathbf{f}}$ given latent variables $\mathbf{f}$ at observed locations. The conditional distribution of set of latent variables given other set of latent variables is also Gaussian [33]:

$$p(\tilde{\mathbf{f}}|\mathbf{f}, \theta, \gamma) \sim N(\mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}|\theta}\mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta}\mathbf{f}, \mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}|\theta} - \mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta}\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}|\theta}) \tag{2.24}$$

3. Determine the marginal posterior predictive distribution of $\tilde{\mathbf{f}}$ given the observations $\mathbf{y}$

$$p(\tilde{\mathbf{f}}|\mathbf{y}) = \int p(\tilde{\mathbf{f}}|\mathbf{f}, \theta, \gamma)p(\theta, \gamma|\mathbf{y})p(\mathbf{f}|\mathbf{y}, \theta, \gamma)df d\theta d\gamma \tag{2.25}$$

4. Compute posterior predictive distribution of new observation $\tilde{\mathbf{y}}$ at location $\tilde{\mathbf{s}}$:

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int p(\tilde{\mathbf{y}}|\tilde{\mathbf{f}}, \gamma)p(\tilde{\mathbf{f}}|\mathbf{y}, \theta, \gamma)p(\gamma|\mathbf{y})d\gamma d\tilde{\mathbf{f}} \tag{2.26}$$

2.3 Conditional posterior mean and covariance of the latent function

We aim to construct maps of log larvae predictive densities $\log\lambda(s)$ in the water along Gulf of Bothnia, which is equivalent to compute the conditional mean and variance of the latent function $\tilde{\mathbf{f}}$ (see Eq. (2.8)). If we have already computed $p(\tilde{\mathbf{f}}|\mathbf{f},\theta,\gamma)$ and $p(\tilde{\mathbf{f}}|\mathbf{y},\theta,\gamma)$ by Eq.(2.24) and Eq.(2.25), according Eq. (2.24) the mean

$$m_p(\tilde{\mathbf{f}}|\theta,\gamma) = \int E_{\tilde{\mathbf{f}}|\mathbf{f},\theta,\gamma}[\tilde{\mathbf{f}}]p(\mathbf{f}|\mathbf{y},\theta,\gamma)d\mathbf{f} = \mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}|\theta}\mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta}E_{\mathbf{f}|\mathbf{y},\theta,\gamma)}[\mathbf{f}] \tag{2.27}$$

The posterior predictive covariance between any set of latent variables $\tilde{\mathbf{f}}$ can be computed by applying the rule of total variance [13]

$$Cov_{\tilde{\mathbf{f}}|\mathbf{y},\theta,\gamma}[\tilde{\mathbf{f}}] = E_{\mathbf{f}|\mathbf{y},\theta,\gamma}[Cov_{\tilde{\mathbf{f}}|\mathbf{f}}[\tilde{\mathbf{f}}]] + Cov_{\mathbf{f}|\mathbf{y},\theta,\gamma}[E_{\tilde{\mathbf{f}}|\mathbf{f}}[\tilde{\mathbf{f}}]] \tag{2.28}$$

The first term in Eq. (2.28) simplifies to the conditional variance in Eq.(2.24), and the second term is expressed as [33, 35]

$$Cov_{\mathbf{f}|\mathbf{y},\theta,\gamma}[E_{\tilde{\mathbf{f}}|\mathbf{f}}[\tilde{\mathbf{f}}]] = \mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}|\theta}\mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta}Cov_{\mathbf{f}|\mathbf{y},\theta,\gamma}[\mathbf{f}]\mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta}\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}|\theta} \tag{2.29}$$

After grouping posterior predictive covariance of $\tilde{\mathbf{f}}$ Eq. (2.28) becomes:

$$K_p(\tilde{\mathbf{f}},\tilde{\mathbf{f}}|\theta) = \mathbf{K}_{\tilde{\mathbf{f}},\tilde{\mathbf{f}}|\theta} - \mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}|\theta}\Big(\mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta} - \mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta}Cov_{\mathbf{f}|\mathbf{y},\theta,\gamma}[\mathbf{f}]\mathbf{K}^{-\mathbf{1}}_{\mathbf{f},\mathbf{f}|\theta}\Big)\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}|\theta} \tag{2.30}$$

2.4 Bayesian computation, Markov chain Monte Carlo sampling

2.4.1 Monte Carlo estimates

The practical problem in Bayesian inference is how to compute the integrals in normalizing constant Z in Eq. (2.23) or predictive distributions in Eq. (2.25) and Eq. (2.26). Since these integrals do not have analytical solution, we have to compute them numerically. The expectation of posterior distribution $p(\theta|y)$ can be obtained by integration over posterior density:

$$E(\theta|y) = \int \theta p(\theta|y)d\theta \tag{2.31}$$

On the other hand, by the law of large numbers, the same expectation can be computed as:

$$\lim_{M \to +\infty} \frac{1}{M} \sum_{h=1}^{M} \theta_h = E[\theta|y] \tag{2.32}$$

where $\theta_h \sim p(\theta|y)$.

The quantity

$$\frac{1}{M} \sum_{h=1}^{M} \theta_h \tag{2.33}$$

is called Monte Carlo estimate for expectation of θ [33]. Once the samples θ_h are available, we can apply different transformations $g(\theta)$ to them $g_h = g(\theta_h))$ for all $h = 1, \ldots, M$. For example, the posterior variance could be computed using Monte Carlo estimate as:

$$Var[\theta|y] \approx \frac{1}{M} \sum_{h=1}^{M} (\theta_h - \hat{\theta})^2 \tag{2.34}$$

where $\hat{\theta} = \frac{1}{M} \sum_{h=1}^{M} \theta_h$. Thus, using Monte Carlo estimates, we are able to compute different measures related to summary statistics (quantiles, posterior density estimates) or to sample from joint distribution using conditional distributions.

2.4.2 Markov chain Monte Carlo sampling

The main problem in obtaining Monte Carlo estimates is how to sample θ_h, $h = 1, \ldots, M$ from arbitrary distributions. For this purpose different Markov chain Monte Carlo methods (MCMC) are created [13, 33]. The idea is to construct a Markov chain with stationary distribution corresponding to the desired distribution, from which Monte Carlo estimates are sampled.

Metropolis-Hastings and Hamiltonian Monte Carlo algorithms

Metropolis-Hastings (MH) algorithm performs a random walk according to a Markov chain whose stationary distribution is the desired target distribution. At each step in the chain, a new state is proposed according to some proposal distribution, and either accepted or rejected in agreement with dynamically calculated probability, called an acceptance criteria. The key property of MH algorithm is that for computing acceptance probability only the likelihood function $p(y|\theta)$ and the prior probability $p(\theta)$ are needed, but not the intractable marginal likelihood $p(y)$ (Eq. (2.16)) [33]. If the MH algorithm is run for long enough (until the Markov chain converges), then the probability of being on a given state on the chain is equal to the probability of the state [19]. Thus, walking on the Markov chain and recording its states is like sampling from target distribution.

The Hamiltonian Monte Carlo (HMC) algorithm is a modification of MH algorithm, in which the random walk behavior is improved to move more rapidly and avoid a long time zigging and zagging through the target distribution [13]. Transition proposal in HMC is enhanced by adding a "momentum" variable ψ_h to each component θ_h in target space. Although we are interested only in the simulation of θ, the sampling is provided for the joint distribution $p(\theta, \psi|y)$. The vector ψ is an auxiliary variable, used to move faster through the parameter space. A multivariate normal distribution with mean zero and diagonal covariance matrix K is assigned to the momentum ψ. The matrix K is called "mass matrix". Other parameters required by HMC algorithm, are the number of leapfrog steps L and a scaling factor parameter ϵ. The word "leapfrog" is used because the momentum updates are split into half steps. HMC also uses the information from the gradient of the log-posterior density $\nabla \log p(\theta|y)$

HMC algorithm proceeds by $h = 1, \ldots, M$ iterations with each iteration executing the following steps:

1. Update ψ with a draw $\psi \sim$ Multivariate-Normal$(0, K)$
2. Repeat the following leapfrog steps L times:
   (a) Use the gradient of log-posterior density of θ to make a half-step of ψ:
   $$\psi \leftarrow \psi + \frac{1}{2}\epsilon \, \nabla \log p(\theta|y)$$
   (b) Use the momentum vector ψ to update the parameter vector θ:
   $$\theta \leftarrow \theta + \epsilon K^{-1}\psi$$
   (c) Again use the gradient of the log-posterior density of θ to make a half-step of ψ:
   $$\psi \leftarrow \psi + \frac{1}{2}\epsilon \, \nabla \log p(\theta|y)$$
3. Denote the values of parameter and momentum vectors at the start of leapfrog process as θ_{h-1}, ψ_{h-1}, and the corresponding parameter values after L leapfrog steps as θ_*, ψ_*. Calculate acceptance probability as
   $$r = \frac{p(\theta_*|y)p(\psi_*)}{p(\theta_{h-1}|y)p(\psi_{h-1})}$$
4. Set
   $$\theta_h = \begin{cases} \theta_* & \text{with probability min(r,1)} \\ \theta_{h-1} & \text{otherwise} \end{cases} \tag{2.35}$$

The tuning of sampling parameters in HMC algorithm may be problematic, because of which latent variables may be heavily dependent in their posterior distribution [33]. To avoid this, we use the Cholesky decomposition of prior covariance matrix $\mathbf{K_{f,f}} = \mathbf{LL^T}$, and sample from $\mathbf{z} = \mathbf{L^{-1}f}$

Assessment of convergence and autocorrelation in Markov chain simulation

The material in this section is based on theoretical foundations reported in [13] and [33]. The building of Markov chain starts from randomly chosen initial state θ_0, which may be very far from the desired stationary distribution. Independently on the chosen θ_0, after sufficiently large sample, drawn in agreement with the chosen MCMC algorithm, the chain has to converge to the desired proposal distribution. According to the Markov property, every next state θ_h depends only on the current state θ_{h-1}. These dependences introduce some autocorrelation that may lead to less precise inference of MCMC sample comparing to the inference of independent sample of the same size. Thus, we have to assess two main properties of the obtained MCMC samples - convergence to the desired stationary distribution, and autocorrelation between the samples from the posterior distribution.

Assessment of convergence Since the early draws are influenced by the initial state, they are not representative for the stationary distribution and have to be discarded from further analysis (so called burn-in). We further analyze only the rest of the samples. To check convergence, we have to simulate several sequences with starting points dispersed through parameter space, and compare the resulting draws. Visual impression about convergency can be obtained by constructing trace plots. The trace plot is a time series plot of the Markov chains. It shows the evolution of parameter vector over the iterations of one or many Markov chains [4]. In convergent models, trace plots look like random noise jumping around a relatively constant number. The chains should look like they are drawn from the same distribution. Since the visual inspection can be misleading, we need to use some quantitative measures. The $\hat{R}$-statistics measures the potential scale reduction of the current MC estimate for the distribution of θ if the number of iterations of the Markov chain was increased to infinity. We denote by L the number of independent Markov chain and by M the number of samples in each chain. The mean of the samples from l-th chain is $\overline{\theta}_l = \frac{1}{M}\sum_{i=1}^{M}\theta_{lh}$, and the total mean of all samples in all chains is $\overline{\theta} = \frac{1}{L}\sum_{i=1}^{L}\overline{\theta}_l$. Then the between-chain variance B and within-chain sample variance W are defined as

$$B = \frac{M}{L-1}\sum_{l=1}^{L}(\overline{\theta}_l - \overline{\theta})^2 \tag{2.36}$$

$$W = \frac{l}{L}\sum_{l=1}^{L}\frac{1}{M-1}\sum_{h=1}^{M}(\theta_{lh} - \overline{\theta}_l)^2 \tag{2.37}$$

The marginal posterior variance of θ is estimated as weighted average of the between and within chain variances:

$$\widehat{var}^{+}(\theta|y) = \frac{M-1}{L}W + \frac{1}{M}B \tag{2.38}$$

The estimators of posterior variance Eq. (2.38) and within chain variance Eq. (2.37) are used to define $\hat{R}$ statistics as

$$\hat{R} = \sqrt{\frac{\widehat{var}^{+}(\theta|y)}{W}} \tag{2.39}$$

Gelman et al. [13] recommend a general maximum acceptable threshold of at most 1.1 for the $\hat{R}$ statistics.

Autocorrelation within the chain, effective sample size Since we take random samples from the posterior distribution, the Monte Carlo estimate of posterior expectation contains some randomness through the actual realization of θ_h, $h = 1, \ldots, M$. One quantitative measure about the error caused by this randomness, is called a Monte Carlo Standard Error (MCSE) and it is defined as a standard deviation of the MC average of a random variable [33]

$$MCSE(\theta) = \sqrt{\frac{Var[\theta_h|y]}{M}} \tag{2.40}$$

where θ_h, $h = 1, \ldots, M$ are independent samples from the posterior distribution. MCSE is different for different random variables and therefore should be computed for all of them. In order to be able do compare the MCSE values for different variables, we can use the folowing relative measure:

$$CV(\overline{\theta}) = \frac{MCSE(\theta)}{E[\theta|y]} \tag{2.41}$$

where $\overline{\theta} = \frac{1}{M}\sum_{i=1}^{M}\overline{\theta}_l$. $CV(\overline{\theta})$ is called a coefficient of variation. By exploring Eq. (2.41) it is possible to show that with Monte Carlo methods it is harder to estimate small posterior probabilities and therefore we need to use much more samples to estimate them [33].

When we estimate the variance of samples obtained via MCMC, we have to take into account the autocorrelation within the chain. The variance of posterior mean is:

$$Var(\overline{\theta}) = Var\left[\frac{1}{M}\sum_{i=1}^{M}\theta_h\right] = \frac{Var[\theta_h|y]}{M} + \frac{2}{M}\sum_{h=1}^{M}\sum_{h'=h+1}^{M}Cov(\theta_h, \theta_{h'}) \tag{2.42}$$

When the covariance $Cov(\theta_h, \theta_{h'}) \approx 0$ the MC error with Markov chain is approximately the same as the MC error in the case of independent samples.

Another scalar complementary measure about efficiency of Markov chain is the effective sample size n_{eff}. It is defined as the sample size M multiplied by the proportion of variance estimate under the assumption of independent samples to the variance estimate for an MCMC sample. In the limit case when $M \to \infty$ theoretical $n_{eff}$is defined as [33]

$$n_{eff}(\theta) = M\frac{Var[\theta_h|y]}{Var[\theta_h|y] + 2\sum_{h=1}^{\infty} Cov(\theta_0, \theta_{h'})} = \frac{M}{1 + 2\sum_{h=1}^{\infty} Corr(\theta_0, \theta_{h'})} \tag{2.43}$$

$Cov(\theta_0, \theta_{h'})$ and $Corr(\theta_0, \theta_{h'})$ are theoretical covariance and correlation functions of Markov chains, and similarly as the variance in Eq. (2.42) cannot be computed directly. Similarly as MCSE, n_{eff} is different for different variables.

2.5 Prior distributions of hyperparameters

Prior distributions are unconditional probabilities, assigned before to observe any data. They present the initial beliefs about the parameters. In the Bayes formula (Eq. (2.16)), via the likelihood function posterior distribution adds information that comes from the data, to the prior beliefs. When the information. provided by the data, is not strong enough, posterior inference becomes very sensitive to the prior choice. If some priors are not properly chosen, the posterior may become improper and thus it will be impossible to provide any meaningful inference. The problem arises with increasing the model complexity. Therefore the choice of priors is very important task and takes special attention in scientific literature [12, 13, 14, 31, 33].

2.5.1 Prior classes

Depending on amount of available information involved, priors are uninformative, weakly informative or informative [12, 13]. Informative priors include all available knowledge from theory or previous experiments. Uninformative priors (called also noninformative) are assumed in the case of missing preliminary information. Then one strategy is to assume equal probabilities for all possibilities, like uniform priors do. Weakly informative priors incorporate only partial insight so that the information, provided by them, is weaker than the knowledge incorporated in any actual prior.

According to their exact definition, the priors are proper and improper [12, 13]. Proper priors $p(\theta)$ have a valid probability density function defined as

$$\int_{\Theta} p(\theta) = 1$$

Improper priors are extension of proper priors such that they do not integrate to a finite number:

$$\int_{\Theta} p(\theta)d\theta = \infty$$

Hence the improper priors do not have direct statistical interpretation. If the prior is improper, the resulting posterior may also be improper distribution. However, proper priors always lead to proper posteriors.

2.5.2 Uninformative and weakly informative priors

The uninformative and weakly informative priors are the most commonly used in ecological models [24]. Uninformative priors express vague or general information about the parameters and aim minimal impact on the posterior inference. Common choice of such priors are uniform $\mathcal{U}(0,1)$, uniform on wide range, e.g. $\mathcal{U}(0,100)$, or a normal $\mathcal{N}(0,\sigma^2)$ centered at 0 and standard deviation set at high values, e.g. $\sigma^2 = 100^2$ [12, 23, 24]. These distributions are flat over the interesting parameter range, but it is not guaranteed that they will remain invariant under transformation of parameters. Other examples of uninformative priors are Jeffrey's and reference priors.These priors also require very informative data, which is not always a case [21, 40, 15].

Weakly informative priors are more informative comparing to uninormative ones, but still the knowledge involved is weak and does not involve any actual preliminary expertise. These priors are proper and designed to apply to more general model classes, without describing all specific features. It is recommended to use weakly informative instead fully informative priors because weak assumptions allow to extend the relevant part of parameter space and thus to improve the model robustness. More specific prior information may increase the model accuracy, but on the other hand it may also exclude some plausible parameter values [12, 14, 15].

2.5.3 Weakly informative prior distributions

Standard prior choice for α and β is uninformative or weakly informative normal $\mathcal{N}(0,\sigma_\beta^2)$ [14, 33]. This prior is very weakly informative because σ_β^2 is larger than the posterior variances of all the β parameters. Here we assume the priors $\beta \sim \mathcal{N}(0,10)$.

The observational model here is assumed as Negative binomial distribution, described by Eq. (2.7). For the overdispersion parameter r in this model it is possible to use weakly informative prior, such as half Normal $\mathcal{HN}(0,1)$, half Cauchy $\mathcal{HC}(0,5)$, half Student $\mathcal{HT}(4,0,1)$, or Inverse Gamma $\mathcal{IG}(0.001,0.001)$ [1, 12]. If the data possess small amount of overdispersion, it is better to assume weakly informative prior for $1/r$ or $1/\sqrt{r}$. More informative prior Gamma $\mathcal{G}(9,1)$ for r is also recommended [41] and used in this work.

We also need to specify the hyperparameter priors σ_{exp} and l, involved in exponential covariance function $k(\mathbf{s},\mathbf{s}')$ (Eq. (2.15)). This task is challenging because it is known that only the ratio σ_{exp}^2/l is identifiable, but not both parameters alone [8, 15, 33]. The

exponential covariance $k(\mathbf{s}, \mathbf{s}')$ forms a ridge in the likelihood function for σ_{exp} and $1/l$, so that if we assign different values of both parameters, but their ratio is a constant (for example $\sigma^2_{exp} = 1/l$), the pattern in $k(\mathbf{s}, \mathbf{s}') = f(\mathbf{S}, \sigma_{exp}, 1/l)$ remains the same and only the values of covariance function $k(\mathbf{s}, \mathbf{s}')$ differ [8]. The data can restrict the range of possible covariance values, but even in the case of strong data, the prior will affect the shape of posterior along the ridge. This means that the prior of $1/l$ will influence the posterior of σ^2_{exp}. If the prior of $1/l$ has light right tail that restricts short l values, this means that in the posterior only bigger values of σ^2_{exp} are feasible. The idea of penalized complexity priors is to use coordinates parallel and orthogonal to the ridge to specify the priors of both parameters (σ_{exp}, l) [8, 15]. Another possibility is to fix one of the parameters. This solution is not very good because in the case when both parameters vary, the model better fits the data, but most of all - because we cannot fully understand the prior by fixing some parameters and assessing the effect of the others. We need do understand the joint effect of prior as a multivariate distribution [10, 15]. Therefore the careful choice of prior distributions for length scale $1/l$ and magnitude σ^2_{exp} is very important, especially in our case when we use data, for which it is known that they do not express very strong spatial dependence [38]. The priors of variance and length scale should be chosen so that they do not restrict some plausible values of observed fish counts. The length scale prior has to be restricted between some minimal and maximal values, $l_{min} < l < l_{max}$. The limit l_{max} determines shorter right tails of l and thus avoids an overlap between the spatial effects modeled in Gaussian process, and linear regression term affected in parameters space by larger variance of fixed effects β. The variance σ^2_{exp} should be penalized towards zero. We prefer to use prior for standard deviation σ_{exp} instead of variance σ^2_{exp} to avoid the problems that may occur with the very small values towards zero. Here we assume that the spatial random effect describes the variability that is not explained by the covariates. Here we use prior for $1/l$ instead for the length scale l. This ensures that the prior is less flexible and therefore it will not explain variability captured by linear term $\mathbf{X}(\mathbf{S})^{\mathbf{T}}\beta$ [33]. For both σ_{exp} and $1/l$ parameters weakly informative priors are typically used [4, 14, 15, 33]. Priors for standard deviation σ_{exp} have to have a peak near zero and very heavy right tail. The long right tail ensure that variance σ^2_{exp} may increase if the data become more informative at these higher variance values. Since the fixed terms have higher values of their variance priors, the combination of the priors of σ_{exp} and σ_β ensures that the variability explained by spatial effects (lower σ_{exp} values) and fixed terms (higher σ_β values) are separated. Suitable prior distributions of σ_{exp} are those from half-t family - Cauchy or Student-t, or half Gaussian. Gamma or χ^2 distributions are not recommended because both distributions are not defined at zero [4, 33]. For the prior of length scale $1/l$ the weakly informative half Student t $1/l \sim \mathcal{HT}(\nu, \mu = 0, s)$ is recommended [33]. The location parameter ν controls the mass along the distribution

tail. The scale parameter s controls the width of the length prior. If we have some preliminary information where could be the most probable parameter value of $1/l$, we can use wide unimodal distribution with peak at this guess. Other possibility is to use the generalized inverse Gaussian distribution, which has an inverse gamma left tail and an inverse Gaussian right tail [4].

2.5.4 Sensitivity to prior distributions: choice of priors to be analysed

Ideally, we should choose the prior before to obtain any data. This is possible only when we have very strong preliminary knowledge and therefore we are completely sure about prior model assumptions. When uncertainty arises, we try to obtain additional information via likelihood function and detect some possible problems in posterior distribution. If there are some problems in posterior distribution, one possible reason may be because some important features or parts in parameter space were not involved in the prior. Then we have to go back and change the prior. It might be that posterior cannot distinguish between the priors since under a common likelihood several priors will lead to the same posterior [10, 15]. In this iterative modeling process we actually investigate the sensitivity of posterior to different prior distributions. Here we provide sensitivity analysis by setting different prior combinations of hyperparameters σ_{exp} and $1/l$, and overdispersion r, and investigate the joint effect of the chosen priors on posterior distribution. After that we analyze the model performance and provide model selection to find the model with best predictive performance.

The probability density functions of prior distributions, used for overdispersion parameter r are shown in Fig. 2.1. For standard deviation σ_{exp} and inversed of length scale $1/l$ parameters we here explore Half Student-t prior distributions with different values of location and scale parameters. The probability density function of these distributions is shown in Fig. 2.2.

2.6 Model selection

One of the main modeling goals is to choose the best model structure and parameter values that describe the investigated phenomena. For this purpose different models have to be compared before choosing the best among them. In Bayesian models this comparison is based on predictive performance. Model comparison and model selection are different tasks. Since we investigate different properties of the compared models, model comparison is related to the inference and analysis. Model selection refers to decision theory because we have to decide which model to choose. In both tasks the key problem is how to choose

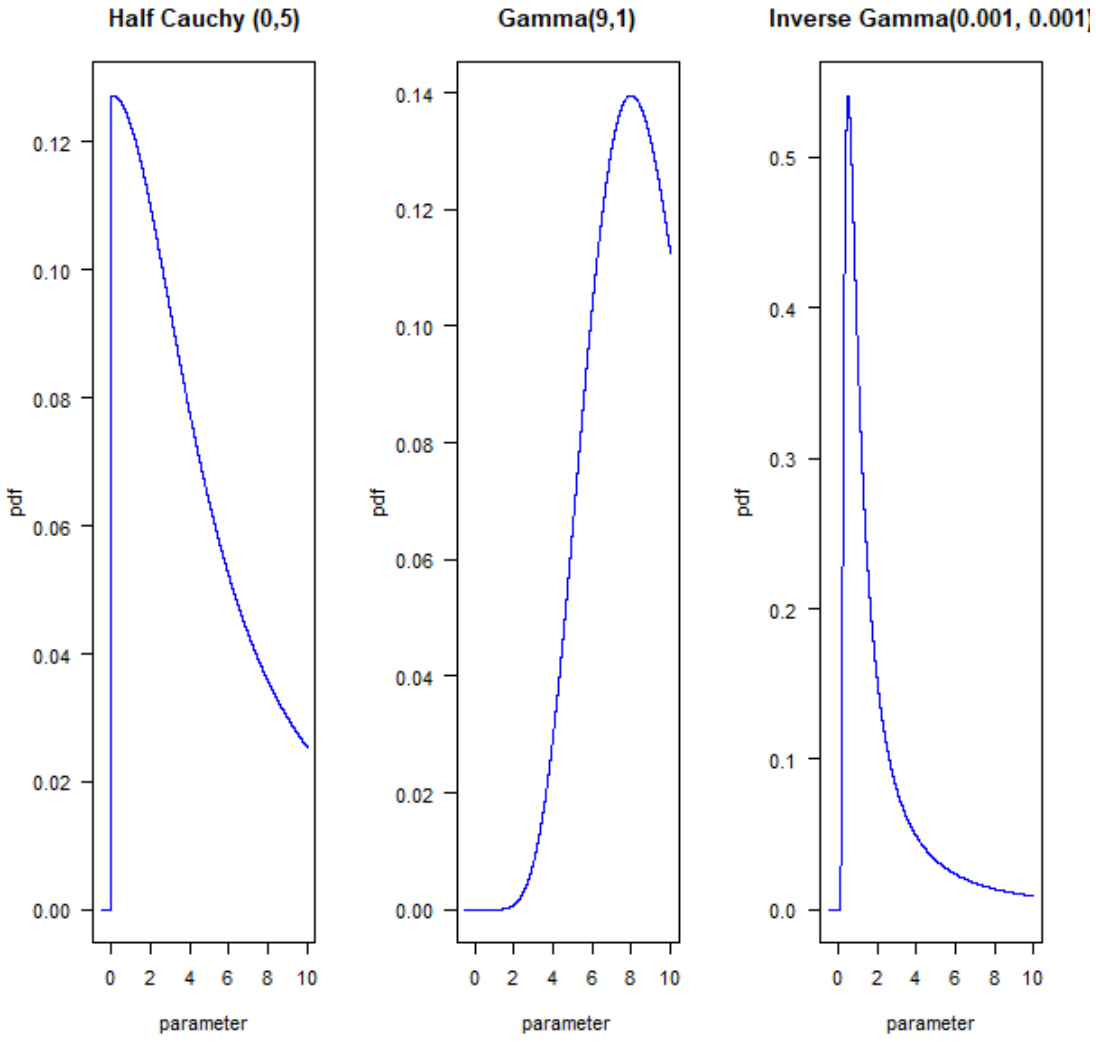


Figure 2.1: Prior distributions used for overdispersion parameter r. We use weakly informative Half Cauchy and Inverse Gamma distributions, with small effect of overdispersion. The Gamma(9,1) prior assumes significant overdispersion nearly the range of fixed effects.

a criterion about goodness of the model [32].

2.6.1 Scoring rules as a measure of predictive accuracy

One way to evaluate the model is through the properties of its predictions. For this purpose scoring rules are introduced as a summary measure that assigns a numerical score based on the realized values and reported predictive distributions. Scoring rules are defined as positively oriented rewards that we want to maximize [7, 17]. If we denote by P the realized predictive distribution for a fixed random variable $Y = y$, a scoring rule S is defined as a function $S = S(P, y)$, which describes the reward for prediction P given the realized value y. If we denote by Q the best predicted, nearest to true predictive

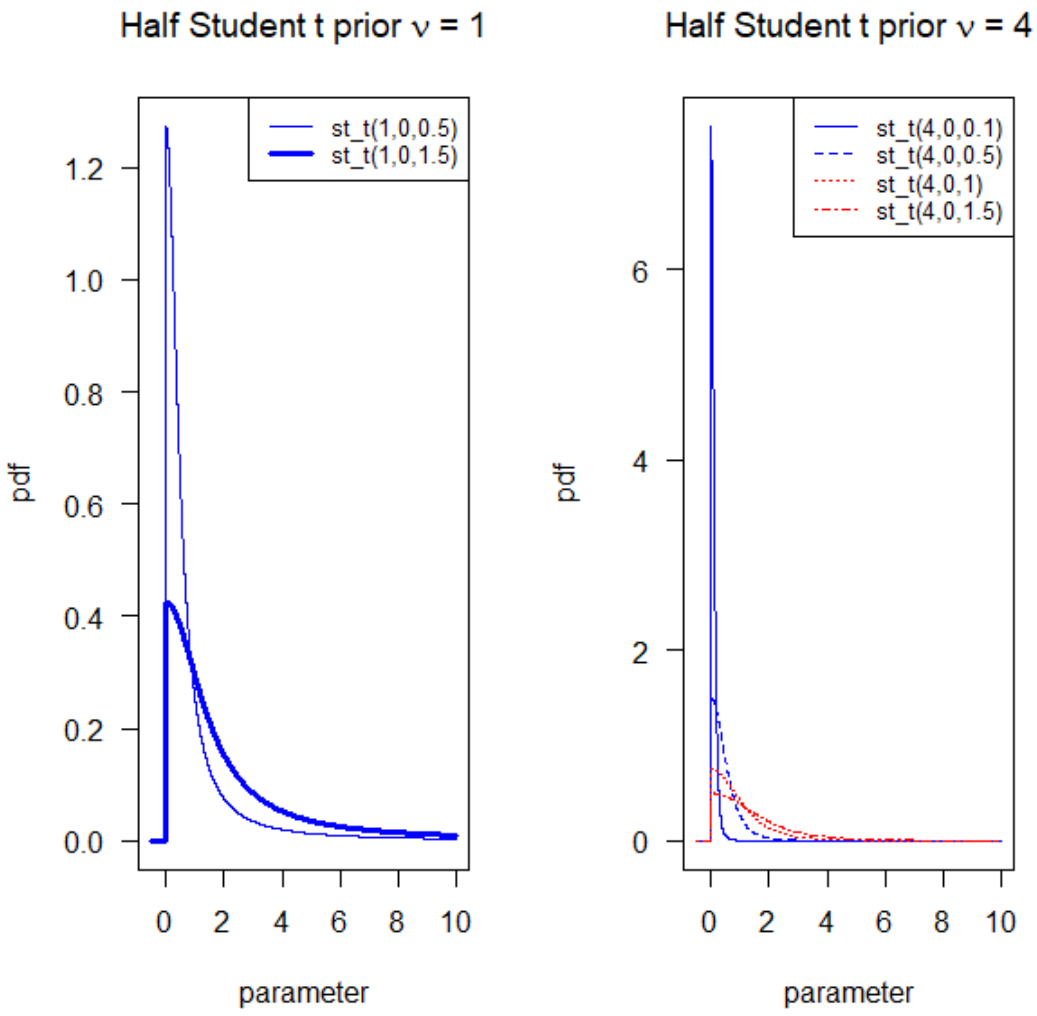


Figure 2.2: Half Student-t distribution used for prior of hyperparameters σ_{exp} and $1/l$. Priors with location $\nu = 4$ have stronger influence towards zero. Increasing the scale value s decreases the peak towards zero and increases the right tail.

distribution, the score $S(P, Q)$ is then the expected value of reward $S(P, y)$ when y is drawn from the distribution Q

$$S(P, Q) = E_Q[S(P, y)] = \int S(P, y) dQ(y) \tag{2.44}$$

A positively oriented scoring rule is called *proper* if for all probabilistic distributions P and Q the following inequality holds:

$$S(Q, Q) \geq S(P, Q) \tag{2.45}$$

Thus, for a proper score, the forecaster maximizes the expected score if he forecasts the true distribution. A *strictly proper* score is a score such that equality in Eq. (2.45) is achieved *uniquely* at $P = Q$ [7, 17].

The only strictly proper scoring rule is the logarithmic scoring rule

$$S(P, y) = \log p(y)$$

where $p(y)$ is the density for result y given by distribution P [32]. Concerning the model selection task, logarithmic scoring rule has attractive properties related to information theory. We assume that the sample sizes of all investigated models are sufficiently large. Then, among all considered models, the model with the highest logarithmic score has the lowest Kullback-Leibler information and thus the highest posterior probability [13, 32]. Therefore in this work we use logarithmic score to assess the model predictions.

2.6.2 Decision theory for model selection

In the framework of decision theory, model comparison task is defined as decision analysis, while model selection is formulated as a decision problem [32]. To solve the decision problem, one has either to minimize the expected loss function, or to maximize the expected utility function. Next we formalize the model selection task. We denote by $d = \{M_d, a(d)\}$ a decision to choose the model $M_d \in D = \{M_1, \ldots, M_k\}$. Here we assume that the true model M is among the investigated models D, $M \in D$. The prediction $a(d)$ of M_d may be probability distribution or point estimate. A utility function $U(d, \tilde{D}, D)$ measures the goodness of chosen model $d \in D$. If we denote by $\tilde{D}$ the future data that will be observed, the utility is denoted by $U(d, \tilde{D}, D)$, and the expected utility for a decision d is defined as:

$$\bar{U}(d, D) = E[U(d, \tilde{D}, D)] = \int U(d, \tilde{D}, D)\, p(\tilde{D}|M_{true})\, d\tilde{D} \tag{2.46}$$

where M_{true} is the true data generating process. Naturally, we do not know this and we will present ways to approximate it below. We search for a decision about optimal model $\hat{d}$ that maximizes the expected utility

$$\hat{d} = \arg\max_{d \in D} \bar{U}(d, D) \tag{2.47}$$

It is known that the optimal prediction for model M_d under logarithmic scoring rule is the posterior distribution of M_d [32].

If we denote by $\tilde{y}$ the predicted value of the observation y, and by $\tilde{x}$ the predicted value of covariate x, the utility function expressed by logarithmic score is defined as:

$$U(d, \tilde{y}, \tilde{x}, D) = \log p(\tilde{y}|M_d, \tilde{x}, D) \tag{2.48}$$

The expected logarithmic density utility (Eq. (2.46)) for decision d becomes

$$\bar{U}(d, D) = \int\int \log\left(p(\tilde{y}|M_d, \tilde{x}, D)\right) p(\tilde{y}, \tilde{x}|M_{true})\, d\tilde{y}\, d\tilde{x} \tag{2.49}$$

Thus, by selecting $\hat{d}$ according Eq. (2.47), where expected utility $\bar{U}(d, D)$ is defined as in Eq. (2.49), we choose the model M_d that minimizes the Kullback-Leibler information [32]. In the case of N future observations the observational vector is $\tilde{\mathbf{y}} = [\tilde{y}_1, \dots \tilde{y}_N]^T$, and vector of covariates is $\tilde{\mathbf{x}} = [\tilde{x}_1, \dots \tilde{x}_N]^T$. The future data consist of $\tilde{D} = \{\tilde{y}_i, \tilde{x}_i\}_{i=1}^N$. The expected utility $\bar{U}(d, D)$ in Eq. (2.49) depends on the full posterior predictive density $p(\tilde{y}, \tilde{x}|D)$. If we assume that the true model M_{true} is known, and regression data are of the type $D = \{y_i, x_i\}_{i=1}^n$, the expected utility of M_{true}, given by Eq. (2.46), can be computed as:

$$\bar{U}(d, D, M_{true}) = \int U(d, \tilde{y}, \tilde{x}, D)\, p(\tilde{y}, \tilde{x}|M_{true})\, d\tilde{y}\, d\tilde{x} \tag{2.50}$$

Although we do not know the true model M_{true}, we might be able to approximate the true data generating process $p(\tilde{y}, \tilde{x}|M_{true})$ by applying Monte Carlo approximation. Then approximated Eq. (2.50) becomes

$$\bar{U}(d, D, M_{true}) \approx \frac{1}{N} \sum_{i=1}^{N} U(d, \tilde{y}_i, \tilde{x}_i, D) \tag{2.51}$$

where $\tilde{y}_i, \tilde{x}_i \sim p(\tilde{y}, \tilde{x}|D)$. Additional benefit of the approximation given in Eq.(2.51), is that we do not need to construct a model for $\tilde{x}$ [32].

2.6.3 K-fold cross-validation

The goals of cross-validation are to test the model's ability to predict new, previously unseen data, to clarify some problems related to overfitting or selection bias, and to estimate how accurately the model will perform in practice. It is a resampling method that uses different portions of the data to train and test a model on different iterations [39]. In k-fold cross-validation the data is divided into k disjoint groups $D = \bigcup_{j=1}^{k} D_j$. At every iteration each group D_j in turn is used as test data. One round j of cross-validation uses one dataset partition into training subset, on which the analysis is performed, and testing set, on which validation of analysis is performed. After k rounds, in which different portion of the data is used as testing subset, the validation results are averaged over the rounds to provide an estimate of the model's predictive performance. The expected utility (Eq. (2.51)) in k-fold cross validation is computed as

$$\bar{U}(d, D, M_{true}) \approx \frac{1}{n} \sum_{i=1}^{n} U(d, y_i, x_i, D_{\setminus k(i)}) \tag{2.52}$$

where the test data are $D_k = \{x_i, y_i\}_{i=1}^l$ and training data $D_{\setminus k(i)} = \bigcup_{j \neq k(i)} D_j$ involve a collection of all $k-1$ subsets that do not include data points y_i, x_i.

When $k = 10$ the model is estimated 10 times, which is feasible from computational point of view. When we use the logarithmic scoring rule, replace Eq. (2.48) in Eq.(2.52)

and obtain:

$$\bar{U}(d, D, M_{true}) \approx \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, M_d, D_{\setminus k(i)}) \tag{2.53}$$

Since the posterior estimates in every round are based on training data that are smaller than the whole available dataset, we need to use large n to be sure that the obtained posterior quantities are reliable and we can approximate $p(y_i|x_i, M_d, D_{\setminus k(i)}) \approx p(y_i|x_i, M_d, D)$. When we use smaller data set for training, this can cause underestimation of the predictive fit for smaller data sets. The second important assumption in cross validation is that the future data $\tilde{D}$ comes from the same generating process as the training data D.

Eq.(2.53) still does not explicitly show how parameters are expressed in expected utility function $\bar{U}(d, D, M_{true})$. If we replace the true parameters θ_{true} by their posterior distribution $p_{post} = p(\theta|y)$ and add them in Eq. (2.53), we obtain expression about logarithmic pointwise predictive density (lppd) [13]:

$$lppd = \frac{1}{n}\sum_{i=1}^{n} \log \int p(y_i|x_i, \theta) p_{post}(\theta|y_{\setminus k(i)}) d\theta \tag{2.54}$$

We again use Monte Carlo approximation and take S simulated draws $\theta^s, s = 1, \ldots, S$ from posterior $p_{post}(\theta|y_{\setminus k(i)})$, where posterior estimates of θ are obtained using only the trainig data sets. Then computed lppd is

$$\text{computed lppd} = \frac{1}{n}\sum_{i=1}^{n} \log\Big(\frac{1}{S}\sum_{s=1}^{S} p(y_i|x_i, \theta^s)\Big) \tag{2.55}$$

Thus the Bayesian k-fold cross-validation (CV) estimate of out-of-sample predictive fit is [13, 32, 37]

$$lppd_{k-CV} = \frac{1}{n}\sum_{i=1}^{n} \log p_{post(-k(i))}(y_i|x_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} log\Big(\frac{1}{S}\sum_{s=1}^{S} p(y_i|x_i, \theta^s\Big) \tag{2.56}$$

Eq. (2.56) provides a measure of predictive accuracy in the case of k-fold CV. By historical reasons similar measures are called information criteria [13]. The information criteria, based on Eq. (2.56) is defined as:

$$CV = -2\, n\, \text{lppd}_{k-CV} \tag{2.57}$$

where n is the number of data points. Lower values of CV correspond to higher predictive accuracy. The CV information criterion, given by Eq. (2.57), is used in this work for the purpose of model selection.

3. Modeling experiment

3.1 Data description

The data used here are previously published and described in [36, 38]. The target species is a sea-spawning whitefish (*Coregonus lavaretus* L. s.l.). This species spawns in October to November and the eggs remain in the spawning grounds until hatching at ice breakup in April to May. During that period the embryos are very sensitive to unfavorable changes in environmental factors. The sea-spawning whitefish larvae appear in littoral areas soon after hatching. The goal in the data collection is to sample whitefish larvae along the Gulf of Bothnia. Gulf of Bothnia is the northernmost basin of Baltic sea and one of the largest brackish water basins in the world. Its size is approximately 120 km x 600 km and is characterizied by strong environmental gradients that influence the larvae distribution and survival: sloping bottom with wide shallow areas in the east coast, deep hollows forming rifts and faults close to the west coast, strong influence of wind and waves around archipelago. Sampling of whitefish larvae started about one week after the ice break from sought to north, when early larvae started feeding. The sampling was provided using a beach seine in nearshore sites and tow net sampler in open water. Larvae were sampled from 642 sampling sites in 21 sub-areas during 2009 - 2011 [36]. To avoid spatial autocorrelation in the data, the sampling sites were randomized. Sampling subareas along Gulf of Bothnia are shown in Fig. 3.1.

The data used here involve at most 8 environmental variables that are summarized in Table 3.1. Since one of the main modeling goals is to construct predictive maps, the environmental variables are available as Geographical Information System (GIS) map layers. The raster layers were converted to a prediction grid of resolution of 300 m. For all necessary GIS analyses authors have used ESRI ArcGIS or ERDAS software packages [36].

Since the values of environmental variables X are different at different spatial locations s, they are also called environmental covariates. From all listed variables $X_i, i = 1, \ldots, 8$, only X_1 (visible bottom in shallow areas) is categorical, all others are continuous. Our training data involve 216 data points measured in year 2010. The data contain measurements on whitefish larvae counts, sampled volume of water, and the en-

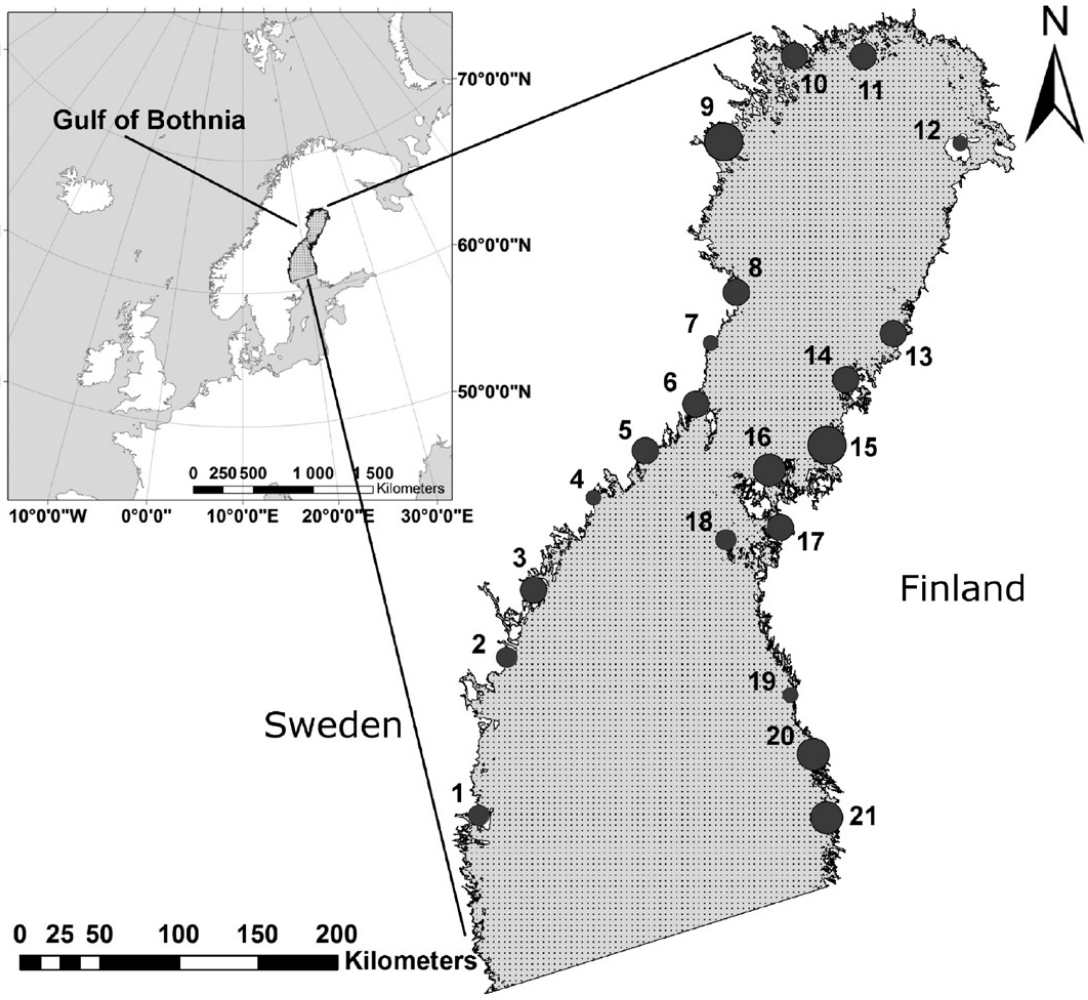


Figure 3.1: Map of Gulf of Bothnia. Sampling sub-areas are marked with circles and numbered (1-21). The size of the circles indicates the count of sampling sites in each sub-area. From [36]

vironmental covariates at observed sites. The prediction locations involve 199 266 map grid cells, corresponding to the locations, at which there are available measurement of environmental variables. The goal is to predict larvae counts at these unobserved locations.

3.2 Experiments

In this chapter we describe all different models to be compared. We investigate the impact of three main factors on posterior inference and predictions. They are:

1. Effect of modeling spatial dependences.
2. Influence of different environmental covariates and expressions with different level of complexity that describe linear part of latent variable
3. Sensitivity of posterior results with respect to different prior distributions of model parameters

Variable	Description	In	Type	Resol.(m)	Unut	Source
BOTCLS	Bottom classification (soft, silt, sand, stones, rocks, cliff)	X_1	class	200	NA	FGFRI
DSAND	Distance to sand, weighted by shallow area	X_2	cont.	90	I	FGFRI
FE300ME	The average fetch over all directions	X_3	cont.	300	m	FGFRI
ICELAST	The end of ice cover in 2009	X_4	cont.	1852 E	wk	FMI
RIVERS	Distance to rivers	X_5	cont.	150	m	FGFRI
SALSPRS	Spring salinity	X_6	cont.	10,000	psu	FMI
D20M	Distance to 20 m depth curve	X_7	cont.	200	m	FGFRI
CHLA	Chlorophyll - a summer phytoplankton concentration	X_8	cont.	2000	I	HELCOM

Table 3.1: Description of the environmental variables. The variables are available as thematic maps layers. Last column shows the source of the data. The abbreviations are: Finnish Game and Fisheries Research Institute (FGFRI), Finnish Environment Institute (SYKE) and Swedish Environmental Protection Agency (SEPA). *I* denotes index value. From [36, 38]

The notation of environmental covariates is given in the second column.

Different model versions are obtained in accordance with these criteria. The goal of the experiment is to select the best model among all candidates based on CV information criteria, described in Chapter 2.6. After choosing the best model, we conduct full posterior evaluation and analyses as it was described in Chapter 2.4.

The common structure of general hierarchical model is presented as:

$$\mathbf{y}|\mathbf{f},\mathbf{N} \sim \prod_{i=1}^{n} \text{Negative Binomial}\left(y_i|V_i\lambda(f(\mathbf{s_i},\mathbf{x_i})),r\right) \tag{3.1}$$

$$\mathbf{f}(\mathbf{s},\mathbf{x})|l,\sigma^2_{exp} \sim GP(m(\mathbf{x}),k(\mathbf{s},\mathbf{s}'|l,\sigma^2_{exp}) \tag{3.2}$$

$$p(\beta,1/l,\sigma^2_{exp},r) \sim p(\beta)p(1/l)p(\sigma^2_{exp})p(r) \tag{3.3}$$

The larval density in the water $\lambda(\mathbf{s_i},\mathbf{x_i})$ in Eq. (3.1) is modeled as $\lambda(\mathbf{f_i}) = e^{\mathbf{f_i}}$.

3.2.1 Models of the linear part of latent variable

Preliminary data transformations

Before to use experimental data in the model, we first provided some data transformations. Except for the first categorical variable X_1^{exp}, we standardized all other covariates $X^{exp} = [X_2^{exp},\ldots,X_8^{exp}]^T$. The standardized environmental data are $X_j =$

$[(X_{j,1}^{exp} - \bar{X}_j)/X_j^{SD}, \ldots, (X_{j,n}^{exp} - \bar{X}_j)/X_j^{SD})]^T$. $\bar{X}_j$ are sample means, and X_j^{SD} are standard deviations of j-th covariate $X_j, j = 2, \ldots, 8$. The purpose of standardization is to put all environmental variables on the same scale and thus to be able to compare their influence on posterior predictions. The other reason to provide it here was to facilitate MCMC model convergence.

Models of Gaussian Process mean

As it was described in Chapter 3.1, the bottom type X_1 is a categorical variable with 6 different classes. We model each class of X_1 by its own effect as:

$$m_0(X_1) = \alpha_0 + \alpha_1\delta_1(X_1) + \alpha_2\delta_2(X_1) + \cdots + \alpha_6\delta_6(X_1) \tag{3.4}$$

where $\delta_i(X_1) = 1, i = 1, \ldots, 6$ if $X_1 = i$, and $\delta_i(X_1) = 0$ otherwise. The intercept $\alpha_0 = 1$ [33].

All other environmental variables are continuous. The linear regression part in Eq.(3.2) of the model that involves 6 covariates, is:

$$m_1(X) = \mathbf{x}^\mathbf{T}\beta = m_0(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \tag{3.5}$$

When we extend this model by adding quadratic terms of all 5 continues variables, the resulting regression becomes:

$$\begin{aligned} m_2(X) = \mathbf{x}^\mathbf{T}\beta = m_0(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \\ \beta_7 X_2^2 + \beta_8 X_3^2 + \beta_9 X_4^2 + \beta_{10} X_5^2 + \beta_{11} X_6^2 \end{aligned} \tag{3.6}$$

Adding X_7 and X_8 to the linear terms of covariates results in:

$$m_3(X) = \mathbf{x}^\mathbf{T}\beta = m_0(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7^2 + \beta_8 X_8 \tag{3.7}$$

After adding the quadratic terms, the regression model becomes:

$$\begin{aligned} m_4(X) = \mathbf{x}^\mathbf{T}\beta = m_0(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \\ \beta_9 X_2^2 + \beta_{10} X_3^2 + \beta_{12} X_4^2 + \beta_{13} X_5^2 + \beta_{14} X_6^2 + \beta_{15} X_7^2 + \beta_{16} X_8^2 \end{aligned} \tag{3.8}$$

Thus, we investigate 4 different models of the mean $m(\mathbf{x})$ in Eq. (3.2).

3.2.2 Modeling spatial effects

The latent variable in Eq. (3.2) is modeled by Gaussian process, where the spatial random effect is modeled by covariance function $k_{exp}(\mathbf{s_i}, \mathbf{s_j})$ and describes the variability that is not explained by the covariates. We investigate two cases:

1. No spatial effects, $k_{exp}(\mathbf{s_i}, \mathbf{s_j}) = 0$. Here all variability is explained by involved covariates in $m(\mathbf{x})$.

2. The spatial effects are modeled by exponential covariance function (Eq. 2.15):

$$k_{exp}(\mathbf{s_i}, \mathbf{s_j}) = \sigma_{exp}^2 e^{-\|\mathbf{s_i}-\mathbf{s_j}\|/l} \tag{3.9}$$

3.2.3 Different priors of hyperparameters and overdispersion parameter

The third task of the experiment is to investigate the influence of different prior parameter distributions on posterior results. We are interesting to learn how sensitive is the posterior prediction with respect to different prior assumptions about hyperparameters.

Linear weights α and β, involved in Eq. (3.4) - Eq. (3.8) are called also fixed effects, and are mutually independent. Usually normal distribution $\mathcal{N}(0, \sigma_\beta^2)$ is assigned to them. When $\sigma_\beta^2 = 10$, this prior is very weakly informative because σ_β^2 is larger than the posterior variances of all the α and β parameters [14, 33]. Thus we assume

$$\alpha, \beta \sim \mathcal{N}(0, 10) \tag{3.10}$$

The possible prior distributions of overdispersion parameter r in Eq. (3.1), as well as hyperparameter priors of σ_{exp}^2 and $1/l$, involved in covariance function (3.9), are discussed in Chapter 2.5. All prior combinations of spatial effects and overdispersion parameters that we analyze further in this work, are listed in Table 3.2 and Table 3.3.

Param.	1	2	3	4	5
σ_{exp}	half-t(4,0,0.1)	half-t(4,0,1)	half-t(4,0,0.5)	half-t(1,0,0.5)	half-t(1,0,0.5)
l/l	half-t(4,0,1)	half-t(4,0,1)	half-t(4,0,1.5)	half-t(1,0,1.5)	half-t(1,0,0.5)
r	half Cauchy(0,5)	half Cauchy(0,5)	half Cauchy(0,5)	half Cauchy(0,5)	Gamma(9,1)

Table 3.2: Weakly informative priors of standard deviation σ_{exp}, inverse length $1/l$ and overdispersion r

Thus we analyze 36 versions of models describing spatial effects, and 16 versions of models without spatial effects. The goal is to investigate all in all 52 model versions and to choose the one with smallest CV information criterion.

3.3 Modeling workflow

In this work we use the Stan package for MCMC sampling, and R environment for further statistical computing and graphics. RStan interface allows us to fit Stan models, called

Parameters	6	7	8	9
σ_{exp}	half-t(4,0,0.5)	half-t(4,0,0.5)	half-t(4,0,0.5)	half-t(4,0,0.5)
l/l	half-t(4,0,1.5)	half-t(4,0,1.5)	half-t(4,0,1.5)	half-t(4,0,1.5)
r	Gamma(9,1)	Inv-Gamma(0.001, 0.001)	Inv-Gamma(0.001, 0.001)	half-t(4,0,0.5)

Table 3.3: Weakly informative priors of standard deviation σ_{exp}, inverse length $1/l$ and overdispersion r

from R, and access the Stan output in R. Stan provides a free, ready-made refined HMC algorithm for MCMC sampling. The tuning of the sampling parameters L and ϵ, as well as determining the gradient of the log-posterior density, is done in automated manner [4].

In Stan, the user first provides a program to import the data and specify the Bayesian model assumptions. Standard distributions (normal, gamma, binomial, Poisson) are pre-programmed. Arbitrary distributions can be entered by directly programming the log density [13]. Next the user provided program is translated to C++ and resulting C++ program is compiled and run. The resulting posterior sample of the model parameters is returned to the user. Stan also provides ready made implementation of methods for computing $\hat{R}$ statistics and n_{eff} measure.

We first choose which environmental variables will be involved in the model. After that we load in R the data about the chosen environmental covariates $x(s)$, and whitefish larvae counts $y(s)$ at observed locations s. Next we standardize the chosen covariates, and load the corresponding environmental data $\tilde{x}(\tilde{s})$ at unobserved locations $\tilde{s}$. We standardize also the data $\tilde{x}(\tilde{s})$. For the task of cross-validation only the data at observed locations $x(s)$ and $y(s)$ are used. These data are split to train and test data sets. The test data are treated as previously unseen data and used to check the predictive performance. For each of the models we use the same splitting of data in $K = 10$ groups.

The steps in the modeling workflow are:

1. Choose the model version to be analyzed and set priors of hyperparameters.

2. Import the data in R and provide the necessary data preprocessing.

3. Write a Stan program for the hierarchical model as a separate file with a *.stan* extension. Stan model is used to obtain the joint posterior for $h = 1, \ldots, M$ $f_h, \theta_h, \gamma_h \sim p(\mathbf{f}, \theta, \gamma|\mathbf{y})$ (Eq. (2.21))

4. Call Stan output via Rstan in R and analyze convergency and autocorrelation of posterior samples. Check posterior distributions of estimands in interest. Examine

scatter plots to analyze correlation between hyperparameters σ_{exp} and $1/l$. Compute in R the posterior estimates of linear weights α_h and β_h.

5. Decide whether chains are sampling from posterior distribution.

6. Check whether priors are unduly influencing posterior.

7. Complete model selection and validation tasks by computing CV information criterion. Repeat the previous steps as needed to find the final model.

8. Use the posterior estimates f_h, θ_h, γ_h of the chosen model to compute in R the posterior predictive mean $m_p(\tilde{\mathbf{f}}|\theta,\gamma)$ (Eq.(2.27)) and variance $K_p(\tilde{\mathbf{f}},\tilde{\mathbf{f}}|\theta)$ (Eq.(2.30)) of the log larvae density $\log\lambda$. Plot the corresponding predictive maps and the map of whitefish larvae density λ over the whole study region.

9. Analyze the obtained results.

In this scheme we first sample in Stan the posterior estimates f_h, θ_h, γ_h, and next compute posterior linear weights and all predictive quantities $(\tilde{f}, \tilde{y}, m_p(\tilde{\mathbf{f}}|\theta,\gamma), K_p(\tilde{\mathbf{f}},\tilde{\mathbf{f}}|\theta))$ in R. This is more efficient, feasible and fast way to proceed [33].

4. Results

4.1 Summary of the models performance.

Following the modeling workflow, described in Chapter 3.3, we have performed posterior analysis of all 52 model versions. It turned out that for the prior combinations 2, 3, 4, 7 and 8 (listed in Table 3.2 and Table 3.3), all models (m_1, m_2, m_3 and m_4) failed to produce any results due to numerical problems. Some observation values y were too huge and exceeded the largest representable number in R $1.797693e + 308$. Similar problems concerning weakly informative Inverse Gamma (0.001, 0.001) prior are reported in [12]. The weakly informative half-Student(ν=1, 0, scale) may also be problematic as a prior of either σ_{exp} or $1/l$ parameters. Therefore we next report only the results obtained by the other prior combinations (1,5,6 and 9), listed in Table 3.2 and Table 3.3.

4.1.1 MCMC diagnostics: convergence and autocorrelation

Convergence

We use potential scale reduction statistic $\hat{R}$ as a main tool for assessing convergence of MCMC. We observed that $\hat{R}$ - statistics of all investigated estimands in all model versions was lower than the 1.01 threshold. The acceptable threshold, reported in [13], is higher - 1.1. Thus, we conclude that all parameters and estimated quantities in all investigated models are convergent. In addition, we performed visual inspection of convergence using trace plots. It is recommended to apply trace plot for diagnosing convergence problems only after $\hat{R}$-statistics indicated some problems [9]. The trace plots of all investigated quantities (latent functions f, hyperparameters and overdispersion parametrs) for all investigated model versions look acceptable. Here we illustrate the trace plots of two model versions from every group - including spatial effects, and models without spatial terms. Both illustrated models are the most complicated ones in the group. All analogous trace plots of other models from the same group look very similar. The trace plots of some latent functions f, hyperparameters σ_{exp} and $1/l$ and overdispersion r of a model m_4, prior 1, describing spatial effects, are shown in Fig.4.1. The trace plots of some latent

functions f and overdispersion r of a model m_4, prior 1, without spatial effects, are shown in Fig.4.2.

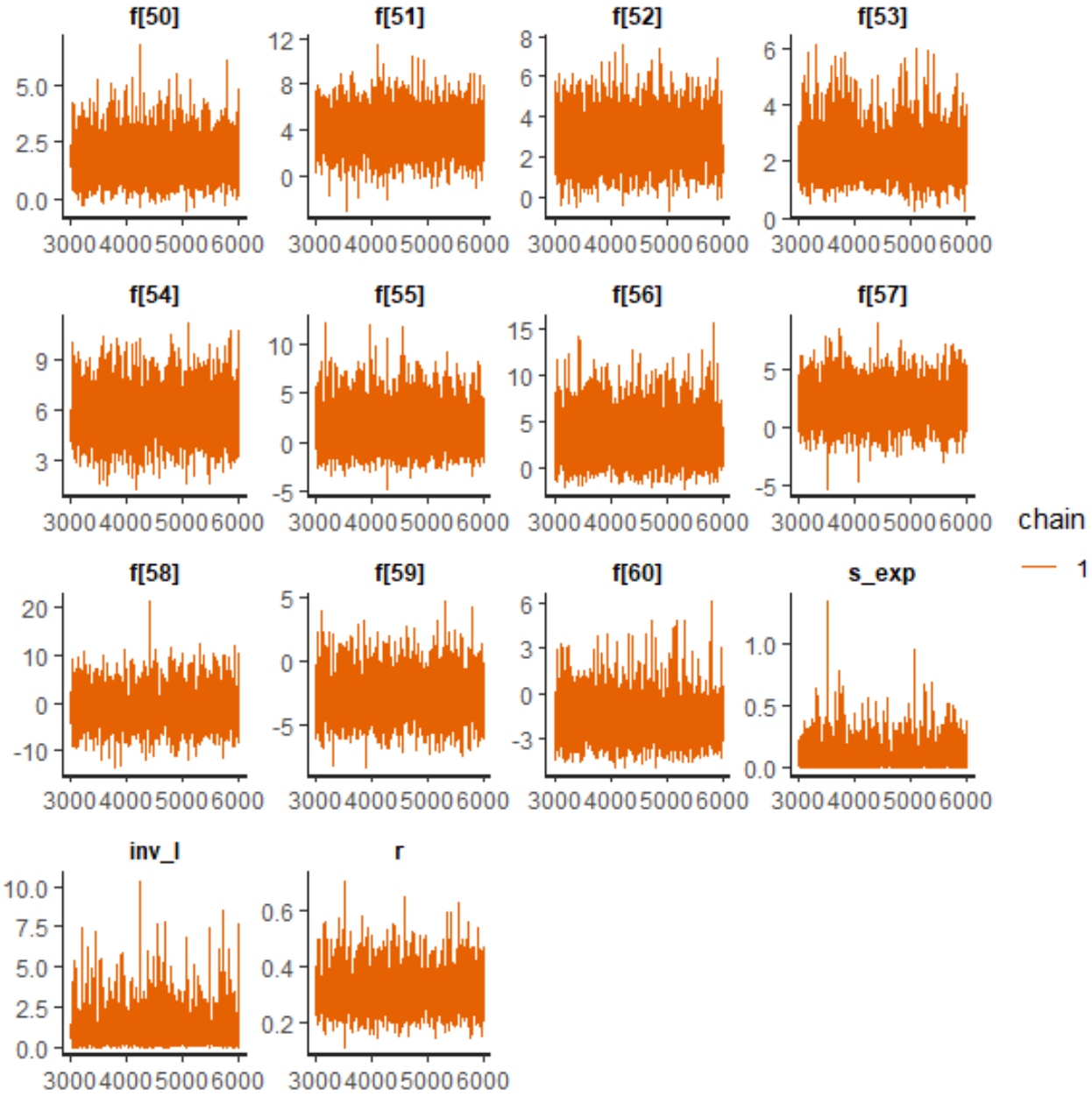


Figure 4.1: Posterior estimation of model m_4 involving eight covariates and their squared terms, with spatial effects. Trace plots of some sampled latent functions $f_i, i = 50, \ldots, 60$, hyperparameters σ_{exp}, $1/l$, and overdispersion parameter r for priors 1 in table 3.2

Although all trace plots in Fig. 4.1 look acceptable, the trace plots of hyperparameters σ_{exp} and $1/l$ look different comparing to the others. These trace plots explore the same region of parameter values, but as σ_{exp} or $1/l$ approaches zero, the chain spends some time in the same region of the parameter space.

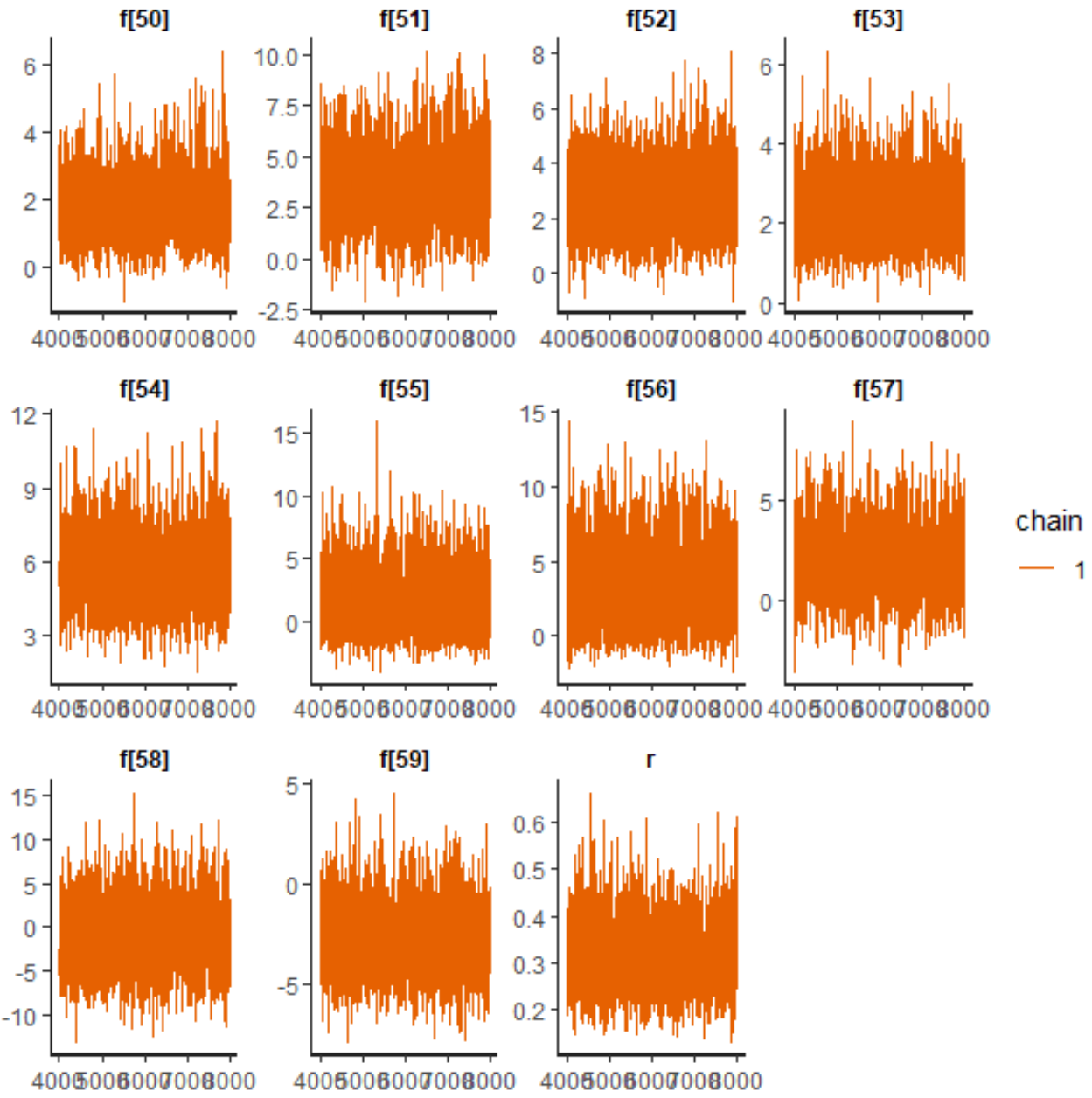


Figure 4.2: Posterior estimation of model m_4 involving eight covariates and their squared terms, without spatial effects. Trace plots of some sampled latent functions $f_i, i = 50, \ldots, 60$, and overdispersion parameter r for priors $r \sim$ Cauchy(0,5)

Autocorrelation

The effective sample size n_{eff} is an estimate of the number of independent draws from the posterior distribution of the estimand of interest. Since the draws within a Markov chain are not independent if there is autocorrelation, the effective sample size n_{eff} is smaller than the total sample size N. The exact definition of n_{eff} is given by Eq. 2.43 in Chapter

2.4. We observe that for all models describing spatial effects n_{eff} is biggest for parameter $1/l$ comparing to n_{eff} of σ_{exp} and r.

The assessment of autocorrelation is provided via visual inspection of autocorrelation plots. These plots show the values of autocorrelation function versus lag. The autocorrelation function should quickly decrease with increasing lag. If autocorrelation function does not decrease quickly with lag, this indicates that the sampler is not exploring the posterior distribution efficiently and results in increased $\hat{R}$ values and decreased n_{eff} values. All investigated models show quickly decreasing autocorrelation functions with respect to all estimands in interest. Here we show two autocorrelation plots from the group of models describing spatial effects (Fig. 4.3) and without involving spatial description (Fig. 4.4). Autocorrelation plots of all other simpler models from the corresponding group are very similar.

Thus we conclude that all 32 models show good convergence.

Correlation among the parameters

As it was reported in Chapter 2.5.4, the estimation of hyperparameters σ_{exp} and $1/l$ might be problematic. We expect some correlation to appear in posterior estimates of σ_{exp} and $1/l$. The trace plots of both parameters also suggest that there might be such correlation. We examined correlation visually by scatter plots. Here we illustrate the scatter plots of the most complicated model, describing spatial effects and Gaussian mean m_4, where 8 covariates and their squared terms and involved. The four different scatter plots in Fig. 4.5 correspond to the four different prior combinations 1, 5, 6 and 9 given in Table 3.2 and Table 3.3.

We observed that the strongest pattern produce prior combination 5, and this is a case for all models m_1, m_2, m_3 and m_4. This suggest that the parameters estimates of separate parameters σ_{exp} and $1/l$ under this prior are not very reliable. Less correlated look the hyperparameter estimates produced by prior combinations 1 and 9. The pattern, observed when prior 6 is used, is not so strong comparing to the scatter plot using prior combination 5, but the shape is similar.

4.1.2 Analysis of parameter estimates

We first examined posterior histograms of some latent functions $f[i], i = 1, \dots 10$ in all investigated models. We observe that all f's are normally distributed, as it is expected in Gaussian processes. Some of these histograms for the model involving spatial effects and described by Gaussian mean m_4 under prior combination 1, are shown in Fig. 4.6.

Posterior histograms of σ_{exp}, $1/l$ and overdispersion parameter r are shown in Fig. 4.7. We observe that the posterior distributions of both hyperparameters σ_{exp} and $1/l$

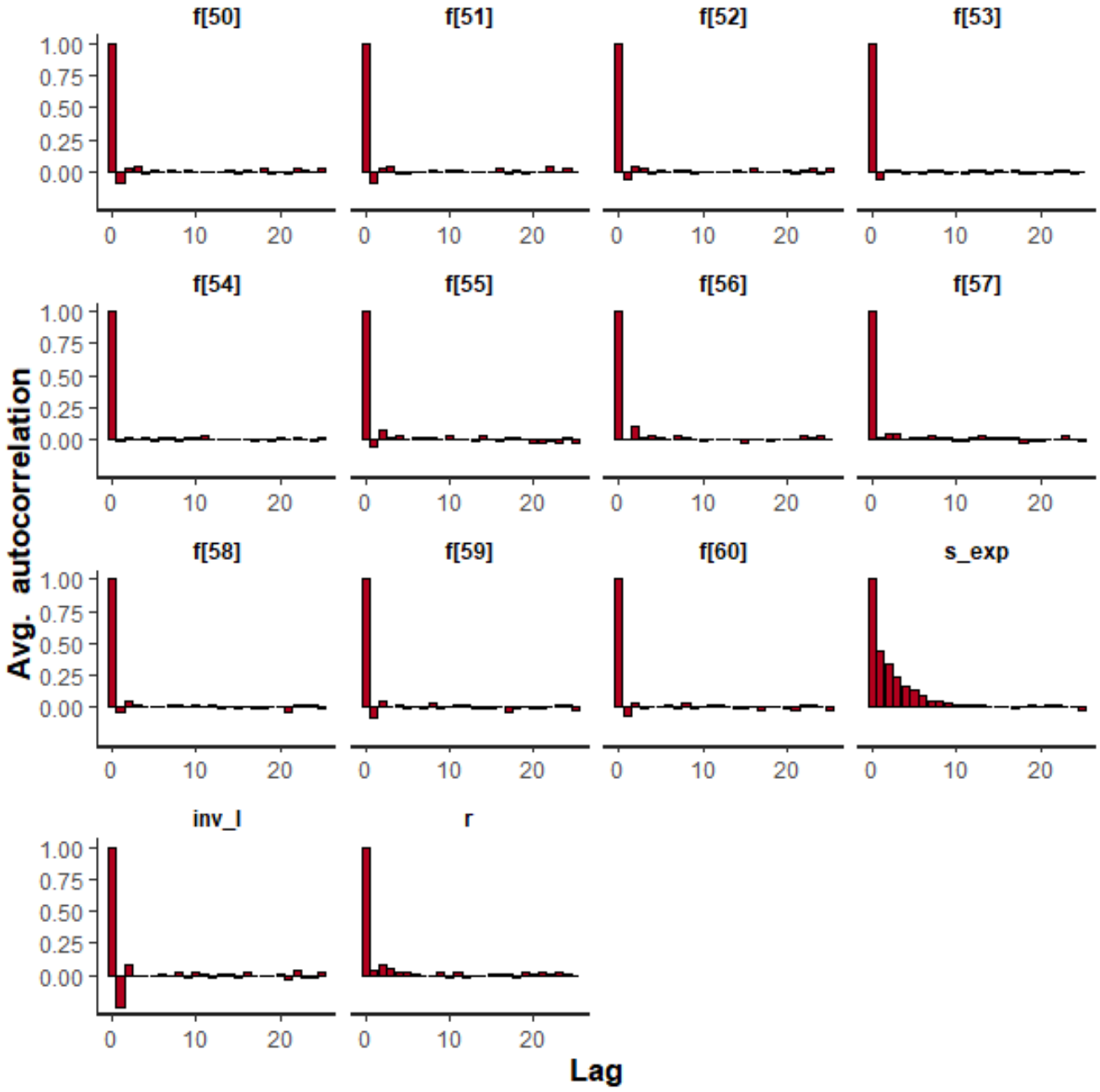


Figure 4.3: Posterior estimation of model m_4 involving eight covariates and their squared terms, with spatial effects. Autocorrelation plots of some sampled latent functions $f_i, i = 50, \ldots, 60$ and hyperparameters σ_{exp} and $1/l$, and overdispersion parameter r for priors 1 in table 3.2

are very similar to theit prior distributions $\sigma_{exp} \sim$ half-t(4,0,0.1) and $1/l \sim$ half-t(4,0,1). The posterior of overdispersion r is normally distributed.

Posterior histograms of linear weights α and β of all investigated models are nearly normally distributed, as it is expected.

The means of parameter estimates of most significant parameters β , σ_{exp}, $1/l$ and r in all models describing spatial effects, are given in Table 4.1. We assume all parameter

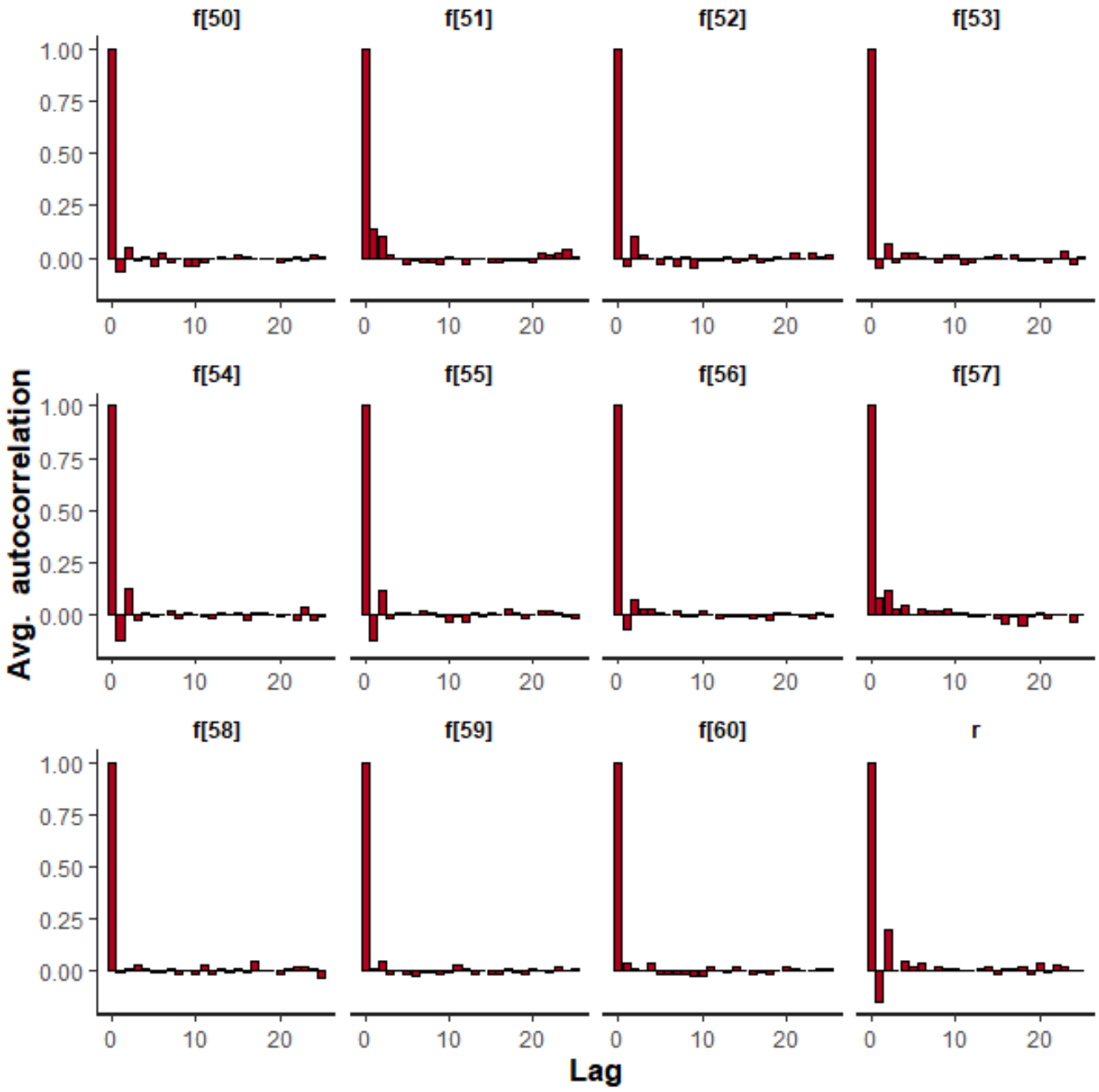
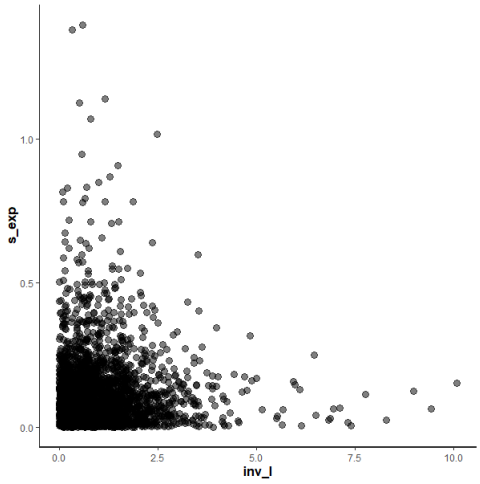
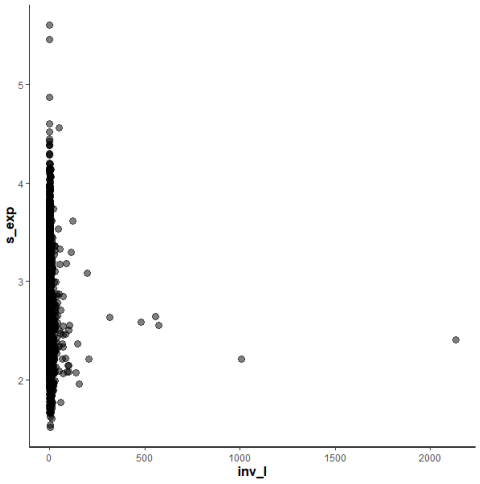


Figure 4.4: Posterior estimation of model m_4 involving eight covariates and their squared terms, without spatial effects. Autocorrelation plots of some sampled latent functions $f_i, i = 50, \ldots, 60$, and overdispersion parameter r for prior $r \sim$ Cauchy(0,5)

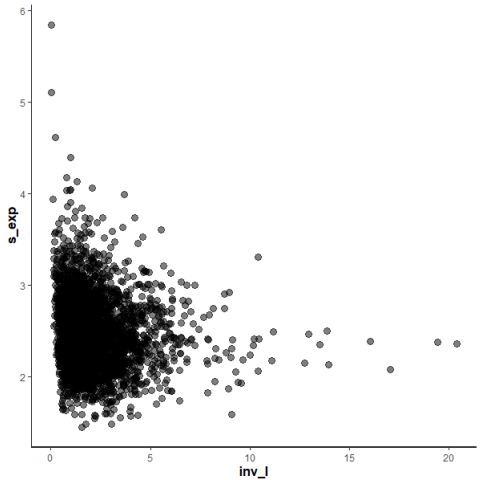
values $|\beta| > 0.5$ as significant. The means of parameter estimates of most significant parameters β and r in all models that do not include description of spatial effects, are given in Table 4.2. The mean estimates, where 2.5% and 97.5% quantiles of the estimation error are huge, are colored in gray. Since there is a big difference among the estimates of linear parameters between the models m_1, m_2, and m_3, m_4 on the other hand, for the models m_1 and m_2 we assumed that all parameter values $|\beta| > 100$ are significant. For



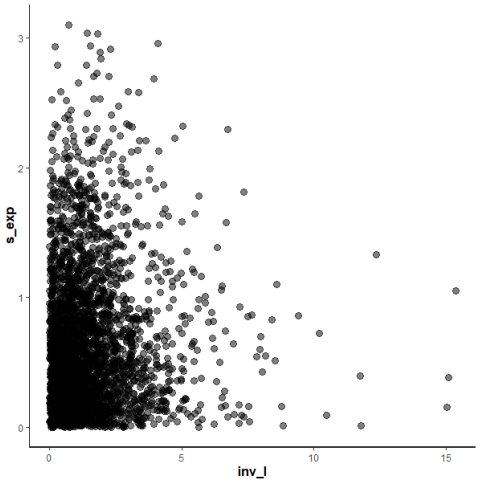
(a) Priors 1



(b) Priors 5



(c) Priors 6



(d) Priors 9

Figure 4.5: Scatter plots of the hyperparameters σ_{exp} and $1/l$ model describing spatial effects by exponential covariance function. The Gaussian mean m_4 involves 8 covariates and their squared term. The four different prior combinations 1, 5, 6 and 9 correspond to priors listed in Table 3.2 and Table 3.3

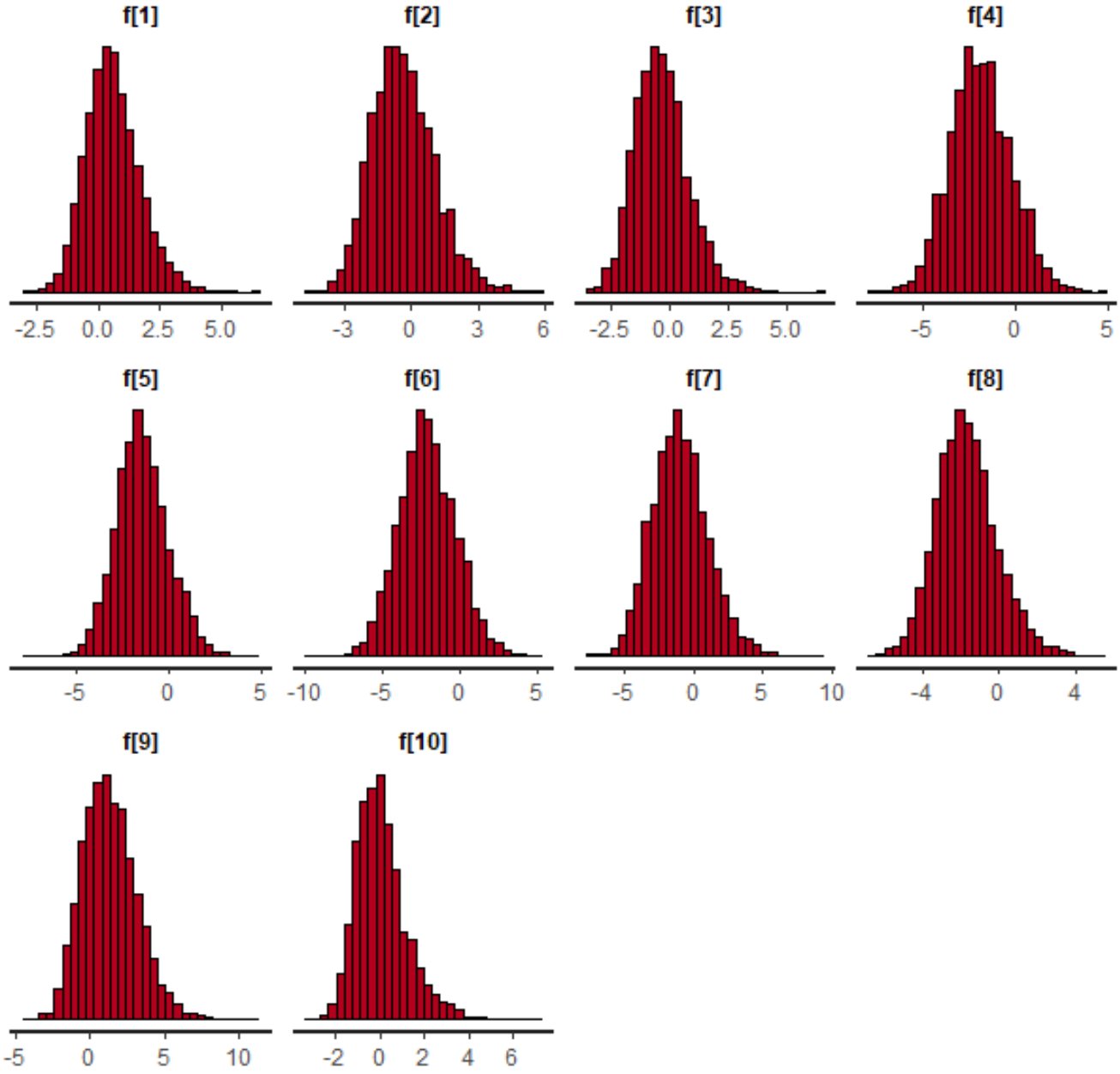


Figure 4.6: Posterior estimation of model describing spatial effects and Gaussian mean m_4 under prior combination 1. Posterior histograms of first ten samples of the latent function $f[i], i = 1, \ldots, 10$

the models m_3 and m_4 we assumed β as significant if $|\beta| > 0.5$. Since in both tables 95% confidence intervals of almost all α estimates are too large, these parameters are not listed in the tables.

More careful inspection of Table 4.1 reveals that except for the model m_1, in all other models the mean estimates of hyperparameter $1/l$ and overdispersion r for priors 5 and 6 are not reliable. The results concerning $1/l$ may be explained by the strong correlation pattern between hyperparameters shown in Fig. 4.5. We also observe that

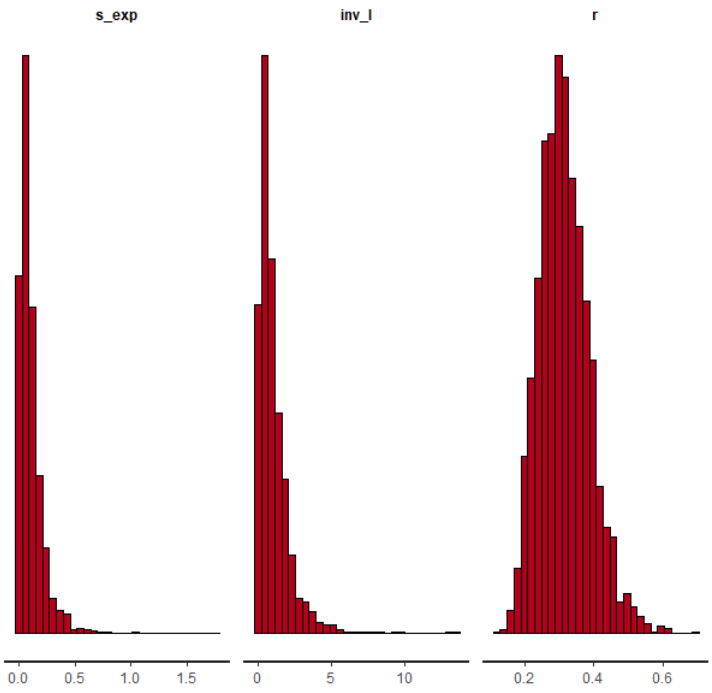


Figure 4.7: Posterior estimation of model m_4 and exponential spatial covariance. Posterior histograms of parameters σ_{exp}, l using priors 1 in table 3.2

independently on used prior combination, the most significant β parameters are almost the same for every model m_1, m_2, m_3 or m_4. The same observation is valid for the model without description of spatial effects (Table 4.2) - in every model m_1, m_2, m_3 or m_4 the most significant β parameters are the same under different priors of r. We observe also that β mean estimates for m_1 and m_2 are much bigger comparing to estimates in m_3 and m_4. The mean estimation values of r in m_1 and m_2 are very small,which indicates that the overdispersion does not influence the prediction. Next based on its predictive performance, we have to choose the best model among these 32 models.

4.2 Results from model selection

In model selection task, we aim to choose the best model by comparing CV information criteria for every candidate, as it was described in Chapter 2.6. Next given the model with lowest CV criterion, we have to analyze its posterior distribution and examine its predictive power. The cross validation information criterion was computed for all models where posterior inference did not resulted in numerical problems or NaN values due to large values that exceeded the largest possible value in R. The corresponding CV information criteria are shown in Table 4.3 and Table 4.4.

The lowest CV values in Table 4.3 is $CV = 197,0604$ for the model including

Models	Priors	Posterior mean estimates of most significant β's				σ_{exp}	$1/l$	r
m_1	1	$\beta_1 = -2.08$	$\beta_2 = -0,75$	$\beta_5 = 0,83$		2,45	1,89	0.33
	5	$\beta_1 = -2.15$	$\beta_2 = -0.71$	$\beta_5 = 1,0$		0,42	1,44	0.33
	6	$\beta_1 = -2.13$	$\beta_2 = -0.79$	$\beta_5 = 0.97$		0,70	1,67	8,9
	9	$\beta_1 = -1,79$	$\beta_2 = -0.76$			0,45	1,45	0.33
m_2	1	$\beta_1 = -1,06$	$\beta_2 = -0.67$	$\beta_7 = 0.75$		0,11	1,01	0.32
	5	$\beta_1 = -0.94$	$\beta_5 = 1,04$	$\beta_7 = 1,67$		2,49	5,42	8,91
	6	$\beta_1 = -1,12$	$\beta_5 = 1,20$	$\beta_7 = 1,56$		2,32	2,03	8,86
	9	$\beta_1 = -1,10$	$\beta_2 = -0.66$	$\beta_7 = 0.95$		0,66	1,45	0.37
m_3	1	$\beta_1 = -0,84$ $\beta_2 = -0,92$	$\beta_5 = -0,82$	$\beta_9 = -0,84$	$\beta_{10} = -1,02$	0,10	0,98	0.3
	5	$\beta_2 = -0,93$	$\beta_4 = 1,07$ $\beta_5 = -0,53$	$\beta_9 = -1,12$	$\beta_{10} = -2,11$	2,91	4,26	8,95
	6	$\beta_2 = -0.81$	$\beta_4 = 1,03$	$\beta_9 = -1,06$	$\beta_{10} = -2,05$	2,65	1,71	8,83
	9	$\beta_1 = -0,73$ $\beta_2 = -0,93$	$\beta_4 = -0.66$ $\beta_5 = -0,8$	$\beta_9 = -0.9$	$\beta_{10} = -1,05$	0,47	1,46	0.31
m_4	1	$\beta_2 = -1,57$	$\beta_6 = -1,14$	$\beta_{11} = -0.85$	$\beta_{12} = -1,40$	0,10	1,00	0.31
	5	$\beta_2 = -1,65$	$\beta_6 = -1,41$	$\beta_{11} = -1,13$	$\beta_{12} = -1,90$	2,70	4,71	9,01
	6	$\beta_2 = -1,62$	$\beta_6 = -1,49$	$\beta_{11} = -1,006$	$\beta_{12} = -1,99$	2,46	2,26	8,76
	9	$\beta_2 = -1,48$	$\beta_6 = -1,05$	$\beta_{11} = -0,87$	$\beta_{12} = -1,31$	0,47	1,46	0,31

Table 4.1: The means of parameter estimates of the most sigificant parameters β ($|\beta| > 0.5$), hyperparameters σ_{exp}, $1/l$ and overdispersion r in the models desribing spatial effects. The prior combinations in the second columns are described in Table 3.2 and Table 3.3.The mean estimates, where 95% confidence intervals are too large, are colored in gray.

spatial effects, where the Gaussian mean m_3 is described by Eq.3.7. The combination of priors $\sigma_{exp} \sim$ Student_t$(4, 0, 0.1)$, $r \sim$ Cauchy$(0, 5)$ and $1/l \sim$ Student_t$(4, 0, 1)$ is most feasible.

The values of CV information criterion for the models that do not describe spatial effects, are shown in Table 4.4.

When we inspect carefully Table 4.4, we observe that the smallest value of CV criterion is in the column for the prior $r \sim$ Cauchy$(0, 5)$ and Gaussian mean m_4, described by Eq.(3.8). This model shows the smallest value of CV criterion in both tables Table 4.4 and Table 4.4. Therefore we continue our analysis using it.

Thus, we continue and analyze posterior predictive performance of the following simpler model, described as:

Models	Priors	Posterior mean estimates of most significant β's					r
m_1	1	$\beta_3 = 238,86$	$\beta_4 = -117,30$	$\beta_5 = 174,34$			0.32
	2	$\beta_3 = 238,62$	$\beta_4 = -117,99$	$\beta_5 = 174,1$			0.43
	3	$\beta_3 = 238,91$	$\beta_4 = -117,32$	$\beta_5 = 174,32$			0,30
	4	$\beta_3 = 238,81$	$\beta_4 = -117.29$	$\beta_5 = 174,25$			0.31
m_2	1	$\beta_3 = 397,5$	$\beta_6 = -470,14$				0.30
	2	$\beta_1 = -397,5$	$\beta_6 = 469,8$	$\beta_7 = -250$			0,44
	3	$\beta_1 = -397,0$	$\beta_6 = 470,0$	$\beta_7 = -251,0$			0,30
	4	$\beta_1 = -397,52$	$\beta_6 = 470,15$	$\beta_7 = -251,13$			0.31
m_3	1	$\beta_2 = -0,84$	$\beta_5 = -2,08$	$\beta_6 = 1,15$	$\beta_{10} = -1,86$		0.30
	2	$\beta_2 = -0,84$	$\beta_5 = -2,09$	$\beta_6 = 1,24$	$\beta_{10} = -1,99$		0,42
	3	$\beta_2 = -0.85$	$\beta_5 = -2,18$	$\beta_6 = 1,14$	$\beta_{10} = -1,90$		0,28
	4	$\beta_2 = -0,91$	$\beta_5 = -2,19$	$\beta_6 = 1,17$	$\beta_{10} = -1,93$		0.30
m_4	1	$\beta_1 = -1,2$	$\beta_2 = -0,83$	$\beta_5 = -1,14$ $\beta_6 = -0,63$	$\beta_7 = 0,91$		0.31
	2	$\beta_1 = -1,36$	$\beta_2 = -0,68$	$\beta_5 = -0,91$	$\beta_7 = 0,94$	$\beta_8 = 0,80$	0,45
	3	$\beta_2 = -1,18$	$\beta_2 = -0,87$	$\beta_5 = -1,17$	$\beta_7 = 0,88$	$\beta_8 = 0,75$	0,29
	4	$\beta_1 = -1,23$	$\beta_2 = -0,83$ $\beta_3 = -0,61$	$\beta_5 = -1,13$	$\beta_7 = 0,98$ $\beta_8 = 0,75$	$\beta_{14} = 0,72$	0,31

Table 4.2: The means of parameter estimates of the most sigificant parameters β, hyperparameters σ_{exp}, $1/l$ and overdispersion r in the models without spatial effects. In the second column the priors are denoted as: $1 \sim half\ Cauchy(0,5)$, $2 \sim Gamma(9,1)$, $3 \sim Inv-Gamma(0.001, 0.001)$, $4 \sim half_t(4,0,0.5)$. For the models m_1 and m_2 parameter values $|\beta| > 100$ are assumed as significant. For the models m_3 and m_4 we assumed β as significant if $|\beta| > 0.5$

Model	Priors			
	1	5	6	9
m_1	228,3749	278,9655	280,5521	228,591
m_2	234,0359	240,46	245,9719	233,926
m_3	197,1243	200,983	208,997	198,271
m_4	236,1184	225,5029	234,0437	230,9378

Table 4.3: Values of CV information criteria for the models describing spatial effects

$$\mathbf{y}|\mathbf{f},\mathbf{N} \sim \prod_{i=1}^{n} \text{Negative Binomial}\,(y_i|V_i\lambda(f(\mathbf{x_i})), r)$$

$$\mathbf{f}(\mathbf{x}) \sim \mathrm{N}(m(\mathbf{x}), \sigma^2)$$

$$m_4(X) = \mathbf{x^T}\beta = m_0(X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 +$$
$$\beta_7 X_7 + \beta_8 X_8 + \beta_9 X_2^2 + \beta_{10} X_3^2 + \beta_{12} X_4^2 + \beta_{13} X_5^2 +$$
$$\beta_{14} X_6^2 + \beta_{15} X_7^2 + \beta_{16} X_8^2$$

$$\alpha, \beta \sim \mathcal{N}(0, 10)$$

Model	Priors of r			
	Cauchy(0,5)	Gamma(9,1)	Inv-Gamma(0.001, 0.001)	Student-t(4,0,1)
m_1	206,699	224,1808	198,63	213,69
m_2	201,3634	200,6109	208,87	210,31
m_3	196,9595	197,936	214,74	215,73
m_4	189,4997	203,0221	207	208

Table 4.4: Values of CV information criteria for the models without description of spatial dependences

4.3 Analysis and predictive performance of the selected model

The MCMC diagnostics of all investigated models was discussed in Chapter 4.1. The chosen model is convergent and does not show problems in autocorrelation plots.

Posterior results

Posterior histograms of some samples of latent functions $f[i], i = 50, \ldots, 59$ and overdispersion parameter r are shown in Fig. 4.9. All latent functions $f[i]$ are nearly normally distributed, as it is expected. Posterior histogram of r is also nearly normally distributed.

Posterior histograms of α parameters are shown in Fig. 4.9, and posterior histograms of β's are shown in Fig. 4.10. All α and β estimates are nearly normally distributed.

A posterior summary of the estimated overdispersion parameter r is shown in Table 4.5). Its 95% confidence interval is not very large. $\hat{R} = 1$ shows good convergence of MCMC.

	mean	sd	2,5%	97.5%	n_eff	Rhat
r	0.31	0.07	0.26	0.46	2662.33	1.00

Table 4.5: Summary of the posterior estimate of overdispersion parameter r in the model with mean m_4 involving eight covariates and their squared terms, without spatial effects, under prior $r \sim$ Cauchy(0,5).

Posterior summary of α and β estimates is shown in Table 4.6. The means and 95% confidence intervals of posterior α and β are presented graphically in Fig. 4.11. The confidence intervals of α_5 and α_6 are smaller comparing to the confidence intervals of other α's. The most significant β parameters are β_1, β_2, β_5, β_7 and β_8.

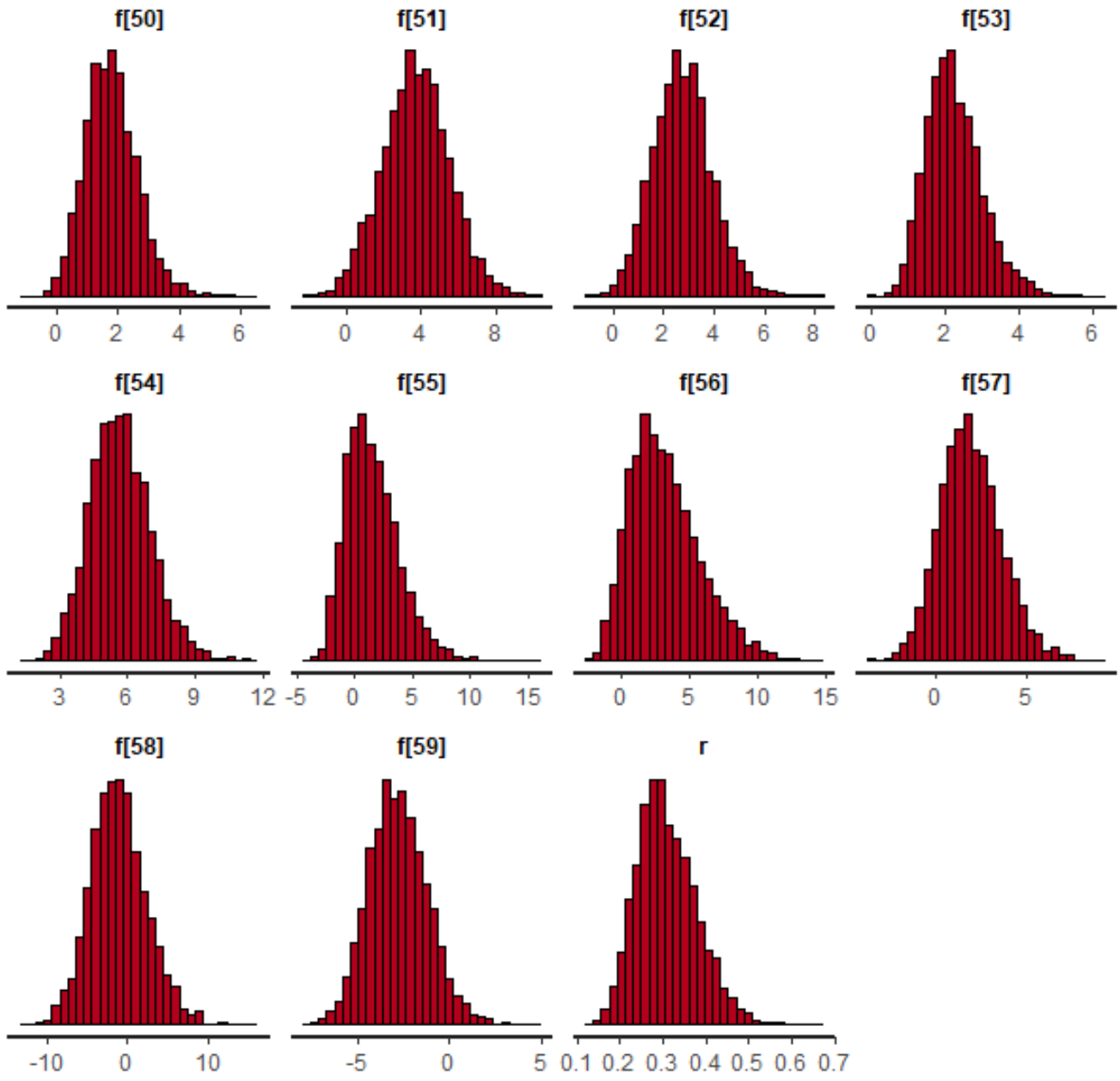


Figure 4.8: Posterior estimation of the model with Gaussian mean m_4 involving eight covariates and their squared terms, without spatial effects. Posterior histograms of some latent functions $f[i], i = 50, \ldots, 59$ and overdispersion parameter r under prior $r \sim$ Cauchy(0,5)

Predictive performance

Predictive maps, obtained from the model with Gaussian mean m_4, involving eight covariates and their squared terms, without spatial effects, under prior $r \sim$ Cauchy(0,5), are shown in Fig. 4.12. Three predictive maps are shown on this figure: log density mean $E(\mathbf{f})$, log density variance $Var(\mathbf{f})$, and intensity $\lambda = e^{E(\mathbf{f})}$. Since the counts λ vary a lot

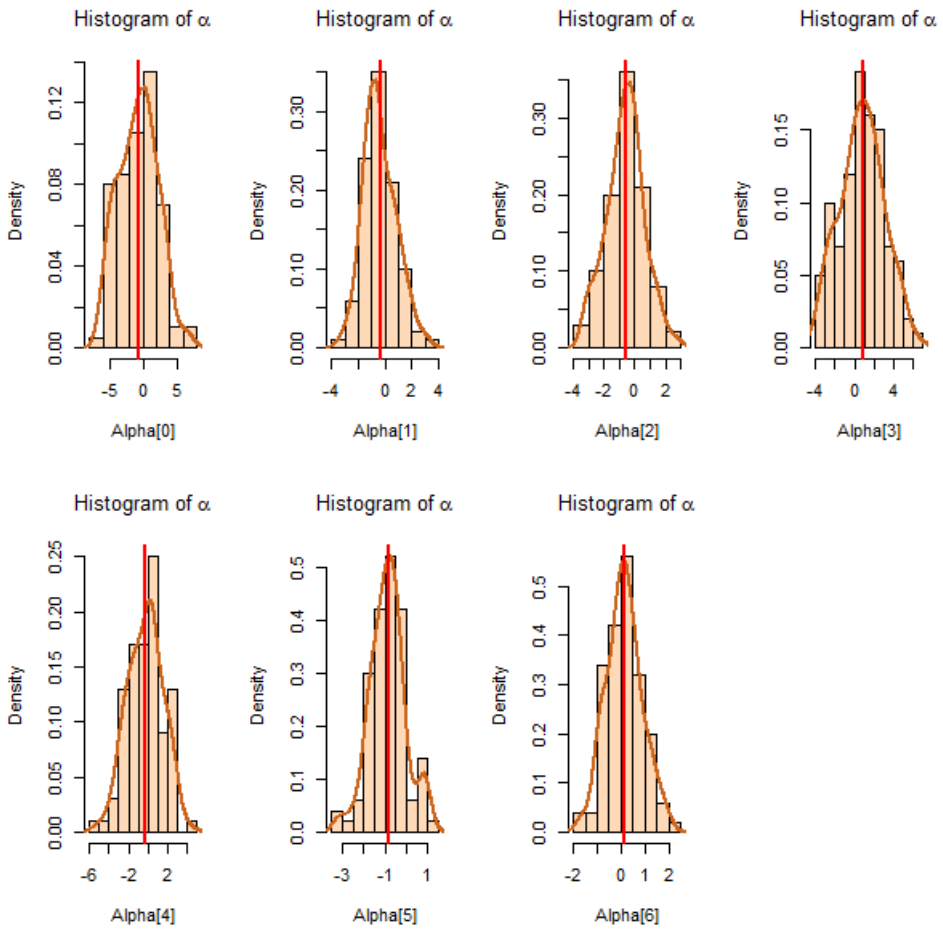


Figure 4.9: Posterior estimation of the model with mean m_4 involving eight covariates and their squared terms, without spatial effects, under prior $r \sim$ Cauchy(0,5). Posterior histograms, means and density functions of α parameters

from very small to very huge numbers, we cannot see all of them on one scale. This is illustrated on the the right map in Fig. 4.12. When we use log scale instead counts, this problem disappears, and we see that the map on the left showing log density mean $E(\mathbf{f})$ is much more informative.

We observe that whitefish larvae is distributed almost everywhere on the map, with less counts near to the south east part. This correspond to the predictive results reported in [38]. Thus we can conclude that for these particular data the simpler model without describing spatial effects is suitable choice to describe whitefish larvae distribution along the Gulf of Bothnia. Furthermore, it was previously reported that for these data the most predictive power comes from environmental variables, and the spatial component has only a slight influence on the prediction [38]. The CV information criterion also shows the best value for the chosen model.

	mean	sd	2,5%	97,5%
α_0	-0.66	2.85	-5.51	4.22
α_1	-0.36	1.22	-2.41	2.01
α_2	-0.59	1.21	-2.99	1.75
α_3	0.83	2.25	-3.31	4.96
α_4	-0.23	1.80	-3.61	2.83
α_5	-0.86	0.84	-2.55	0.90
α_6	0.14	0.75	-1.12	1.71
β_1	-1.20	1.07	-3.04	0.78
β_2	-0.83	0.74	-2.15	0.33
β_3	-0.54	0.65	-1.82	0.94
β_4	0.49	0.44	-0.43	1.25
β_5	-1.14	0.83	-2.67	0.63
β_6	-0.63	0.55	-1.65	0.55
β_7	0.91	0.44	0.13	1.88
β_8	0.71	0.59	-0.45	1.78
β_9	0.12	0.32	-0.38	0.76
β_{10}	0.12	0.26	-0.34	0.60
β_{11}	-0.28	0.19	-0.66	0.08
β_{12}	-0.36	0.85	-1.95	1.12
β_{13}	0.28	0.15	0.00	0.56
β_{14}	0.67	0.41	0.09	1.57

Table 4.6: Summary of the posterior estimates of fixed effects α and β in the model with mean m_4 involving eight covariates and their squared terms, without spatial effects, under prior $r \sim \text{Cauchy}(0,5)$.

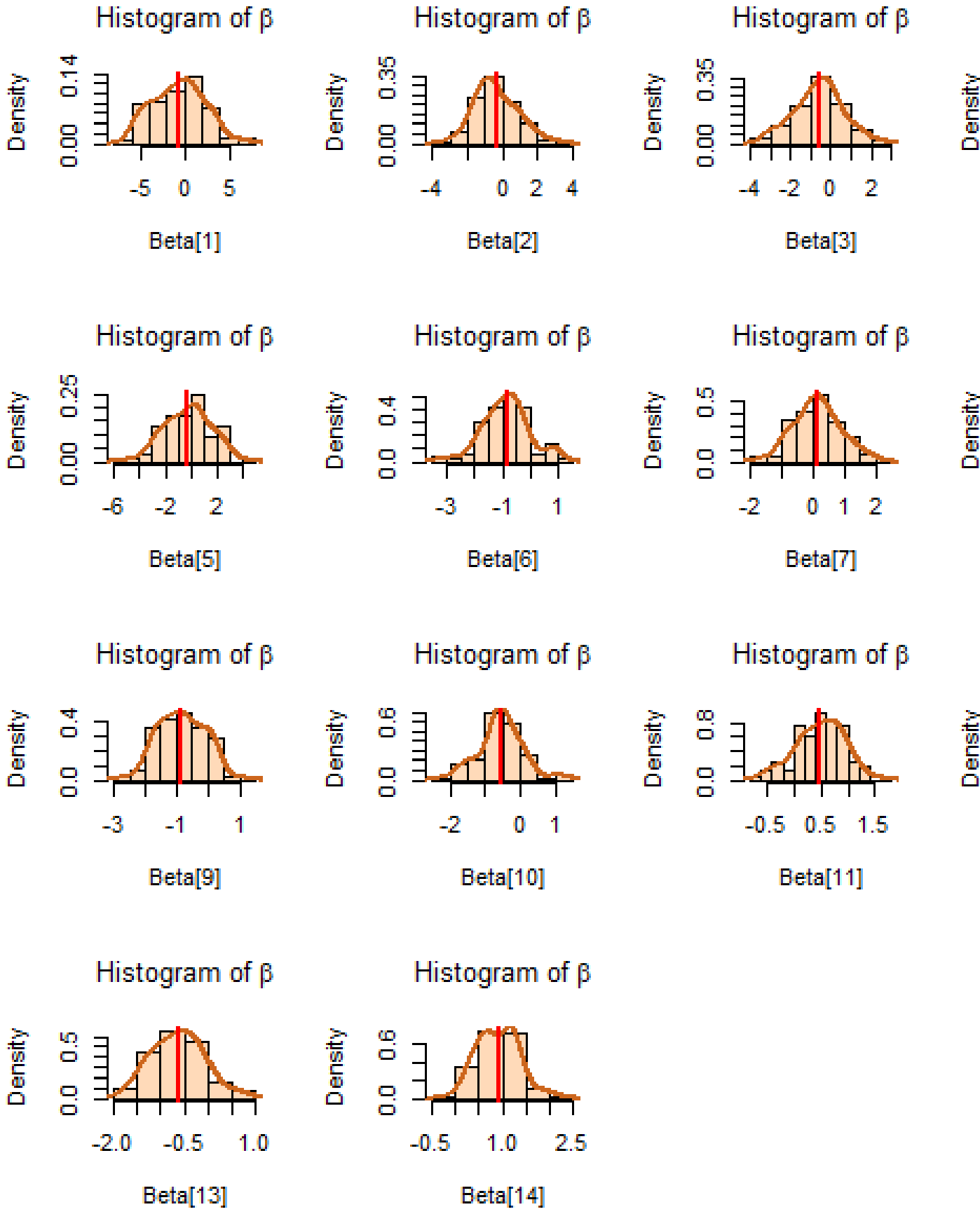


Figure 4.10: Posterior estimation of the model with mean m_4 involving eight covariates and their squared terms, without spatial effects, under prior $r \sim$ Cauchy(0,5) Posterior histograms, means and density functions of β parameters

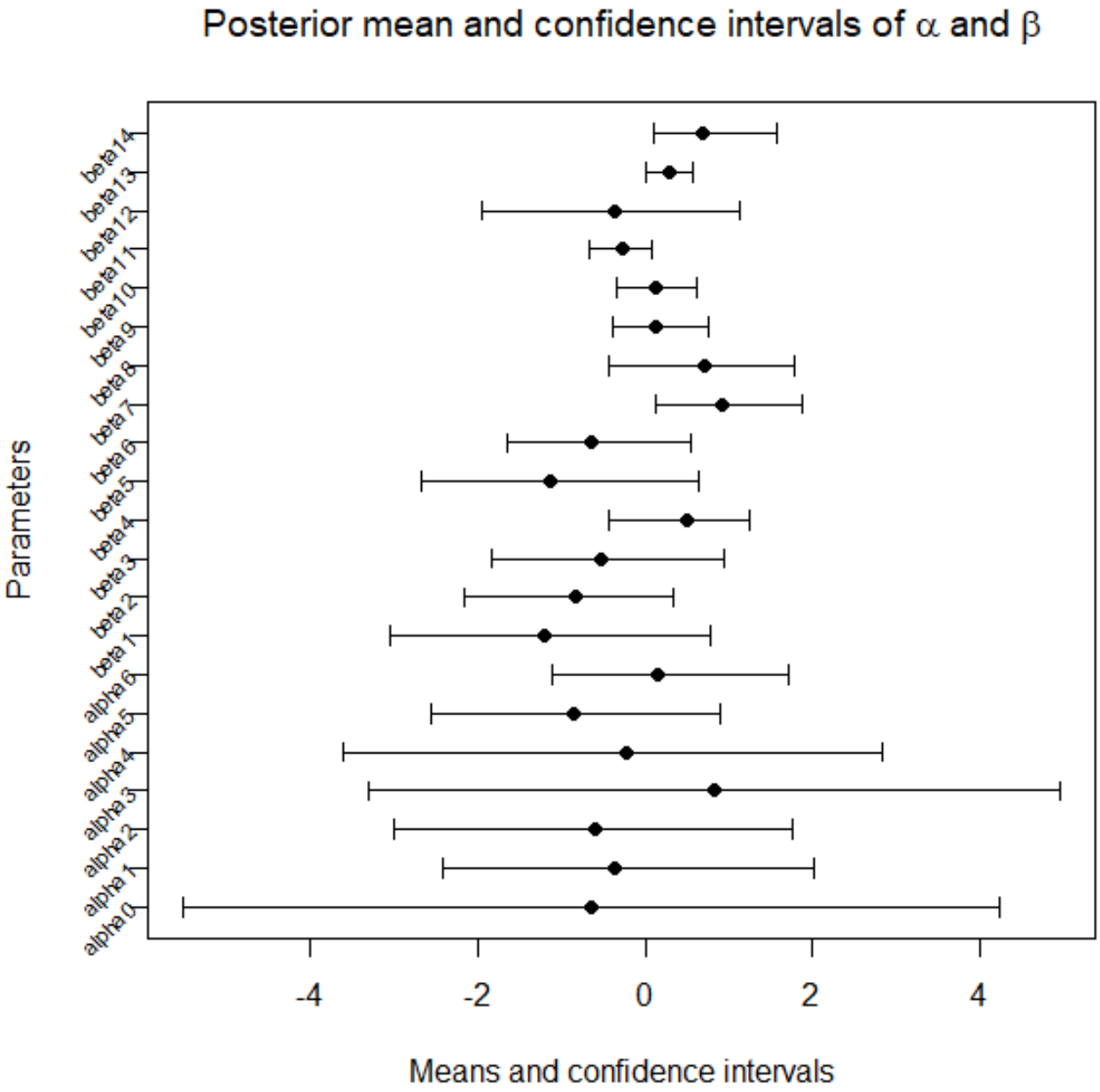


Figure 4.11: Posterior estimation of the model with mean m_4 involving eight covariates and their squared terms, without spatial effects, under prior $r \sim$ Cauchy(0,5). Posterior histograms, means and density functions of β parameters

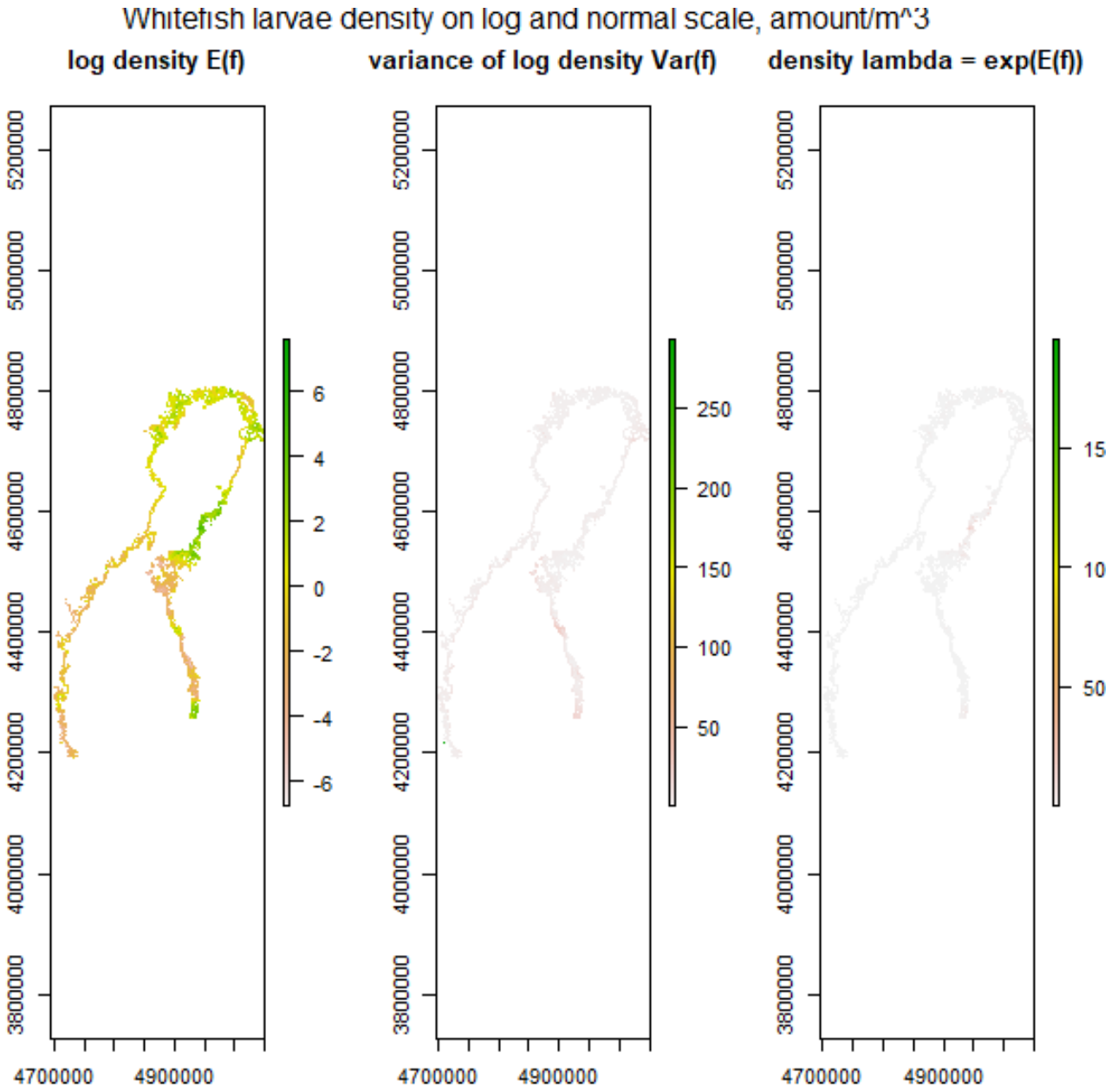


Figure 4.12: Posterior predictions of the model with Gaussian mean m_4, involving eight covariates and their squared terms, without spatial effects, under prior $r \sim$ Cauchy(0,5). Plot of log density mean $E(\mathbf{f})$ (left), log density variance $Var(\mathbf{f})$ (center), and intensity $\lambda = e^{E(\mathbf{f})}$ (right).

5. Discussion

5.1 Posterior parameter estimates of fixed effects

In this thesis we investigated two main groups of models - describing spatial dependences, and models without description of spatial effects. In each group we analysed the performance of four models of Gaussian process means: involving 6 or 8 environmental covariates, including either only linear terms, or adding also their squared terms. In the group describing spatial effects, we applied four different prior parameters combinations to every mean model. In the group where spatial effects were not described, we assigned four different priors of overdispersion parameter to every separate model of latent function. Thus, all at all, 32 candidate models were analyzed.

The common result for both groups of models was that independently on the chosen prior combinations, for the same model of the mean $m_i, i = 1, \dots 4$, the most significant linear weights β are the same. For almost all models the posterior estimates of the weights α are not very reliable because the large 95% confidence intervals. Since the first environmental covariate is categorical, where we want to model separately the effect of every class by different $\alpha_j, j = 0, \dots, 6$, these results show that it is problematic to estimate every class separately and may be better to use different coding system to record x_1 and involve it in the regression model [3].

We observe that the prior distributions of overdispersion r or hyperparameters σ_{exp} and $1/l$ (in the case spatial effects are involved) do not influence a lot the estimates of fixed effect parameters α and β. The explanation might be because the spatial effect model explains only the part of variability that is not explained by linear terms, so there is not an overlap between both model parts.

Another interesting result are the big mean posterior estimates of all α's and β's for all models including only linear terms and 6 or 8 covariates in the models without describing spatial effects. In these models the mean posterior estimate of overdispersion r is very small. Since the overdispersion is negligible small, in these models it might be better to use Poisson observational model instead Negative binomial.

5.2 Sensitivity of posterior to prior choices

For the models describing spatial effects, we examined 9 different combinations of parameters σ_{exp}, $1/l$ and r. It turned out that for 5 of these prior combinations all models of Gaussian mean (m_1, m_2, m_3 and m_4) failed to produce any posterior results results due to numerical problems. Some observation values y were too huge and exceeded the largest representable number in R $1.797693e+308$. These results are not surprising because similar problems when using weakly informative Inverse Gamma (0.001, 0.001) or half-t(ν=1, 0, scale) priors are reported in [12]. Furthermore, it is known that hyperparameters σ_{exp} and $1/l$ might be correlated and therefore difficult to estimate separately. Also, the intensity λ may take either very small or very big values, which also causes problems to show it on the predictive maps. From all other prior combinations, the most problematic was combination $\sigma_{exp} \sim$ half-t$(1, 0, 0.5)$, $1/l \sim$ half-t$(1, 0, 0.5)$ and $r \sim$ Gamma$(9, 1)$. Under this prior, the posterior correlation between hyperparameters for all models m_1, m_2, m_3 and m_4 was very strong, and therefore the posterior estimates of both hyperparameters were not reliable. The best posterior results were obtained using the prior combination $\sigma_{exp} \sim$ half-t$(4, 0, 0.1)$, $1/l \sim$ half-t$(4, 0, 1)$ and $r \sim$ half Cauchy$(0, 5)$. Under this prior, the posterior estimates of both hyperparameters were very weakly correlated, and the obtained estimates were more reliable and close to the mean estimates, obtained by using the rest of the priors. Thus, it was shown that posterior results are very sensitive to prior choices of hyperparameters. The provided sensitivity analysis helped us to choose the most suitable prior for the models that describe spatial effects. Concerning the models without description of spatial dependences, different priors of overdispersion r did not influence a lot the posterior estimates of the corresponding model m_1, m_2, m_3 or m_4.

5.3 Model selection

After computing the CV information criterion, it turned out that the best model from the group describing spatial effects, is the model that uses the best prior combination described above, and involved eight covariates and only the linear term. The best model among the models without spatial effects, and also the best for all investigated models, was the one that involves 8 covariates and their linear and squared terms. These result show that all eight covariates are important and have to be involved in the model. The better predictive performance of the model without describing spatial effects may be explained with the quality of the data, that are not informative enough to be able to model better spatial dependences.

5.4 Interpretation of the results in ecological aspect

The practical significance of this work is determined by predictive performance of the chosen model for the locations, where observations of larvae counts are not available. The predictive maps of log whitefish larvae density for unobserved locations, produced by the model, show in which regions the larvae distribution may be problematic, and where might be abundance of whitefish larvae. It was more feasible to present larvae density on log scale instead to draw directly larvae counts because the huge variability of the counts from very small to huge numbers. The log larvae intensity maps may help ecologists to take decisions about conservation in these regions, or fishing when it is feasible. Although the selected model is quite simple, the predictive maps of log intensity, obtained by it, look similar to the results, reported in [36, 38]. If more informative data are available, it might be feasible to use a model describing also spatial dependences.

5.5 Further improvements

If some additional preliminary knowledge about the maximal value of whitefish larvae counts becomes available, before real model analysis we may perform prior predictive check. Its goal is to simulate some observational data based on priors and likelihood only, and to analyze these data. If the generated data are not consistent with our expectations and understanding, this means that there is something wrong in the chosen prior we have to change it. At this step we can reject the prior if produced observational values are larger than the maximal reported values, or even exceeded the largest representable number in R. Thus prior predictive check can facilitate the sensitivity analysis task.

It became quite clear that for better modeling of spatial dependences, more informative data are needed. Once such data are available, the estimation of hyperparameters may become more reliable. In such case also more complicated covariance functions may be examined [2].

Additionally to the logarithmic score, used in CV information criterion,the predictive distributions can be analyzed by so called probability integral transform (PIT) [16]. If the model matches the true generating process, the distribution of PIT's should be uniform. The empirical PIT's can be plotted as a histogram. Any clear deviations from uniformity in the histogram shows mismatch between the predictions and true generating process. This test can be added to other analyses to improve model assessment.

6. Conclusions

The provided analyses in this thesis showed that hierarchical Bayesian approach with Gaussian processes for modeling spatial dependences is a very powerful tool to explain and predict spatial data of marine species distribution. With these types of models, it was possible to utilize geographical information and produce predictive maps where to illustrate the distribution of larvae density for unobserved locations.

In this work we also illustrated that sensitivity analysis of posterior estimates to prior distributions is a powerful method to select the most suitable prior in the case when preliminary knowledge about the prior is quite weak. Even for the examined models with known identifiability problems of hyperparameters, it was possible to choose the most suitable prior combination that leads to almost uncorrelated hyperparameters, and to produce meaningful posterior parameter estimates for the models that describe spatial effects.

During model selection we compared models with different level of complexity and different prior distributions and chose the one with the best predictive performance. The results showed that the model, which demonstrated best predictive performance, does not need to be very complicated and to involve description of spatial effects if the data are not informative enough to detect spatial effects.

References

[1] J. Arnold. Simon Jackman's Bayesian model examples in Stan, 2018. https://jrnold.github.io/bugs-examples-in-stan/index.html, [03.05.2022].

[2] S. Banerjee, B. Carlin, and A. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall - CRC, Taylor and Francis Group 6000 Broken Sound Parkway NW Suite 300 Boca Raton FL 33487 - 2742, Second edition, 2015. doi:10.1111/biom.12290.

[3] J. Bruin. Coding system for categorical variables in regression analysis, 2022. https://stats.oarc.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis-2/#:~:text=Categorical%20variables%20require%20special%20attention,entered%20into%20the%20regression%20model, [17.05.2022].

[4] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. STAN: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017. doi:10.18637/jss.v076.i01.

[5] M. R. D'Alcala. Similarities, differences and mechanisms of climate impact on terrestrial vs. marine ecosystems. *Nature Conservation*, 34:505–523, 2019. doi:10.3897/natureconservation.34.30923.

[6] E. Duncan, S. Cramb, P. Baade, K. Mengersen, T. Saunders, and J. Aitken. Developing a cancer atlas using bayesian methods: A practical guide for application and interpretation, 2019. URL: https://atlas.cancer.org.au/developing-a-cancer-atlas/index.html#suggested-citation.

[7] W. Ehm and T. Gneiting. Local proper scoring rules. In *Technical Report no. 551*. Department of Statistics, University of Washington, 2009.

[8] G. Fuglstad, D. Simpson, F. Lindgren, and R. Håvard. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 114(1):1–50, 2018. doi:10.1080/01621459.2017.1415907.

[9] J. Gabry. Visual MCMC diagnostics using the bayesplot package, 2017. https://mran.microsoft.com/snapshot/2017-12-15/web/packages/bayesplot/vignettes/visual-mcmc-diagnostics.html#general-mcmc-diagnostics, [13.05.2022].

[10] J. Gabry, D. Simpson, A. Vehtari, and M. Betancourt. Visualization in bayesian workflow. *Statistics in Society*, 182(2):389–402, 2019. doi:10.1111/rssa.12378.

[11] A. Gelfand. Hierarchical modeling for spatial data problems. *Spatial Statistics*, 1:30–39, 2012. doi:10.1016/j.spasta.2012.02.005.

[12] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006. doi:10.1214/06-BA117A.

[13] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall - CRC, Taylor and Francis Group 6000 Broken Sound Parkway NW Suite 300 Boca Raton FL 33487 - 2742, 3rd edition, 2013. doi:10.1111/j.1467-985X.2014.12096_1.x.

[14] A. Gelman, B. Carpenter, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Prior choice recommendations, 2020. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations, [10.12.2021].

[15] A. Gelman, D. Simpson, and M. Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):1–13, 2017. doi:10.3390/e19100555.

[16] J. Geweke and G. Amisano. Comparing and evaluating Bayesian predictive distributions of asset returns, 2008. https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp969.pdf, [17.05.2022].

[17] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi:10.1198/016214506000001437.

[18] G. Guillera-Arroita. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, 40:281–295, 2017.

[19] G. Gundersen. Practical algorithms for latent variable models, 2021. URL: http://arks.princeton.edu/ark:/88435/dsp01cc08hj757.

[20] T. Hefley and M. Hooten. Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1(2):87–97, 2016. doi:10.1007/s40823-016-0008-7.

[21] B. José and A. Smith. *Bayesian Theory*. John Wiley and Sons, Bafiins Lane, Chichestcr, West Sussex PO19 IUD, Engtand, 2009. doi:10.1002/9780470316870.

[22] M. Kallasvuo, J. Vanhatalo, and L. Veneranta. Modeling the spatial distribution of larval fish abundance provides essential information for management. *Canadian Journal of Fisheries and Aquatic Sciences*, 74:636–649, 2017. doi:10.1139/cjfas-2016-0008.

[23] K. Kamary and C. Robert. Reflecting about selecting noninformative priors. *Journal of Applied & Computational Mathematics*, 3(5):1–7, 2014. doi:10.4172/2168-9679.1000175.

[24] N. Lemoin. Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. *OIKOS Advancing Ecology*, 128(7):912–928, 2019. doi:10.1111/oik.05985.

[25] J. Liu and J. Vanhatalo. Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process. *Spatial Statistics*, 35:100392, 2020. doi:10.1016/j.spasta.2019.100392.

[26] J. M. Lobo, A. Jimenez-Valverde, and J. Hortal. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114, 2010. doi:10.1111/j.1600-0587.2009.06039.x.

[27] C. Marshall, G. Glegg, and K. Howell. Species distribution modelling to support concervation planning: The next steps. *Marine Policy*, 45:330–332, 2014. doi:10.1016/j.marpol.2013.09.003.

[28] J. Martinez-Minaya, M. Cameletti, D. Conesa, and M. G. Pennino. Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stochastic Environmental Research and Risk Assesment*, 32:3227–3244, 2018.

[29] S. Melo-Merino, H. Reyes-Bonilla, and A. Lira-Noriega. Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, 415:1–35, 2020. doi:10.1016/j.ecolmodel.2019.108837.

[30] A. Norberg, N. Abrego, G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araujo, T. Dallas, D. Dunson, J. Elith, S. D. Foster, R. Fox, J. Franklin, W. Godsoe, A. Guisan, B. O?Hara, N. A. Hill, R. D. Holt, F. K. C. Hui, M. Husby, J. A. Kalas,

A. Lehikoinen, M. Luoto, H. K. Mod, G. Newell, I. Renner, T. Roslin, J. Soininen, W. Thuiller, J. Vanhatalo, D. Warton, M. White, N. E. Zimmermann, D. Gravel, and O. Ovaskainen. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3):1–24, 2019. doi:10.1002/ecm.1370.

[31] T. O'Hagan. Dicing with unknown. *Significance*, 1(3):132–133, 2004. doi:10.1111/j.1740-9713.2004.00050.x.

[32] J. Vanhatalo. Advanced Bayesian inference, 2020. Lecture notes.

[33] J. Vanhatalo. Spatial modeling and bayesian inference, 2021. Lecture notes.

[34] J. Vanhatalo, G. Hosack, and H. Sweatman. Spatio-temporal modelling of crown-ofthorns starfish outbreaks on the Great Barrier Reef to inform control strategies. *Journal of Applied Ecology*, 54(1):188–197, 2017. doi:10.1111/1365-2664.12710.

[35] J. Vanhatalo, J. Riihinmäki, J. Hartikanen, P. Jylänki, P. Tolvanen, and A. Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013. http://jmlr.csail.mit.edu/papers/v14/vanhatalo13a.html, [27.9.2020].

[36] J. Vanhatalo, L. Veneranta, and R. Hudd. Species distribution modeling with Gaussian processes: a case study with the youngest stages of sea spawning whitefish (Coregonuslavaretus L. s.l.) larvae. *Ecological Modelling*, 228:49–58, 2012.

[37] A. Vehtari, A. Gelman, and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):711–735, 2018. doi:10.1007/s11222-016-9696-4.

[38] L. Veneranta, R. Hudd, and J. Vanhatalo. Reproduction areas of sea-spawning coregonids reflect the environment in shallow coastal waters. *Marine Ecology Progress*, 477:231–250, 2013.

[39] Wikipedia. Cross-validation (statistics), 2019. https://en.wikipedia.org/wiki/Cross-validation_(statistics)#k-fold_cross-validation_with_validation_and_test_set, [17.04.2022].

[40] R. Yang and J. Berger. A catalog of noninformative priors, 1998. Technical Report.

[41] K. Yoshida. Count outcome models with Stan, 2019. https://rstudio-pubs-static.s3.amazonaws.com/455021_9628fb7a86fc4516b51baf676265e016.html, [03.05.2022].