UNIVERSITY OF TORINO

M.Sc. in Stochastics and Data Science

Final dissertation

Effects of global change to Finnish coastal ecosystems: An application of species distribution models

Supervisor: Silvia Montagna
Co-supervisor: Jarno Vanhatalo

Candidate: Ilaria Pia

ACADEMIC YEAR 2019/2020

Summary

The Baltic Sea is an inland sea with brackish water, located in the Northern part of the Atlantic Ocean, between Sweden and Finland. In this sea, and in the surrounding coastal areas, effects of climate change and euthrophication are particularly pronounced. This is because Baltic Sea's extreme location and conditions have caused many changes to be concentrated and to have happened more quickly than in many other regions of the Earth. Furthermore, its unique biodiversity makes this sea a crucial area of interest for the scientific community, and past changes have been object of accurate studies. Nevertheless, the lack of understanding on the expected future changes in these areas is still high.

In this master's thesis we apply state-of-the-art statistical species distribution models to predict the likely effects of climate change to biota in the Finnish coastal region. The core part of the thesis is devoted to modelling the distribution and the extent of sea-spawning whitefish (*Coregonus lavaretus*) and vendace (*Coregonus albula*) larval in the Gulf of Bothnia, the northern part of the Baltic Sea. Whitefishes have a central role in Gulf of Bothnia ecosystem, other than having a relevant economical and social value in Finnish society. Compared to previous studies from 1990s, the extent of whitefish larval areas has decreased. This suggests that the distribution of the reproduction areas could be a good indicator of the health of the Baltic Sea shores. In this work we discuss the possibility that long-term changes in the environment, such as increasing temperature, decreasing salinity, more frequent iceless winters and increasing eutrophication, have reduced the reproductive success of sea-spawning coregonids. The outputs of this study include larval distribution maps, useful monitoring tools, that may assist integrated coastal zone management and environmental protection, e.g. by focusing conservation measures in the most appropriate place.

To conduct the analysis, extensive sampling data on larval occurrences were combined with GIS (geographic information system) raster layers on a

selection of relevant environmental variables. The resulting data were used to implement regression models known, in ecology, as species distribution models. These are hierarchical models with additional Gaussian random effect, which predict the spatial probability of species occurrence or abundance. In our study, we represent larvae abundances distributions through counting processes (Poisson or Negative-Binomial), so that we could further model the average species log-density as a linear function of a set of environmental covariates and a Gaussian spatial random effect. After fitting several models we predict larvae abundances along the whole Gulf of Bothnia coastline, at both current and future climate conditions. Finally, results and potential improvements are discussed.

Acknowledgements

Even though the current situation makes everything a bit confused, obtaining this degree has a really strong meaning for me.

This thesis marks the conclusion of a period that I can safely define as the hardest in my young life. Not everything has been perfect, as I started this master. Difficult circumstances made me hesitate more than ever. During the year I spent in Turin, finding the strength to keep going and the direction to follow was not always easy. My Erasmus in Finland was a great chance of growth and here I found new motivation and a path to pursue. This thesis was, indeed, realized as part of my working experience at the Department of Mathematics and Statistics at the university of Helsinki, under the supervision of Jarno Vanhatalo. The analysis conducted in the thesis outline the first steps for further studies related to whitefish and vendace spawning and will be useful in future publications by Helsinki EnvStats group. Hence, a special acknowledgement goes to Jarno and his research group.

I would also like to thank you all the teachers and the friends encountered through these two years, both in Italy and in Finland. A big thank you goes to my father, my sister and all my family, that have always supported me, despite the physical distances. Even if they can't be present today I am sure we will be soon able to hug again.

My last words are for my mother, who I know would be happy and proud of my achievement. To her, I dedicate this work and all the milestones I reached so far.

Contents

List of Tables 7

List of Figures 8

1 Introduction 10

2 Case study 13
2.1 Whitefishes, vendaces and Gulf of Bothnia 13
2.2 Whitefish dataset 14
2.3 SmartSea dataset 16
2.4 Data preparation 17
2.5 Target and covariates 20

3 Methods 31
3.1 Spatial data 31
3.2 Coordinate reference system 32
3.3 Spatial process models 35
3.4 Spatial predictions 36
3.5 Hierarchical spatial models 37
3.6 Gaussian latent variable models 38
3.7 Species distribution models in ecology 39
3.8 Joint species distribution models 42
3.9 Climate change and SDM 42

4 Proposed methodology and priors setting for HSDM 45
4.1 HSDM for whitefish data 45
4.2 SDM with additional random effect 47
4.3 Covariance function and priors setting 48
4.4 Posterior distributions and MCMC 50
4.5 Spatial predictions 54

5 Posterior analyses for HSDM 57
5.1 Conduct of the analysis 57
5.2 Convergence checking and sensitivity analysis 58
5.3 Models comparison in Bayesian statistics 60
5.4 Case study: model selection 62
5.5 Whitefish HSDM results 64
5.6 Vendace HSDM results 69
5.7 Software 74

6 Conclusions and further extensions 85
6.1 Final results 85
6.2 Criticisms and extensions 87
6.3 Further extension: JSDM 89

List of Tables

2.1 Environmental variables 22
5.1 logpredictive densities 63
5.2 Posterior distribution of random effects parameters for Whitefish SDM 65
5.3 Posterior distribution of linear effects parameters for Whitefish SDM 66
5.4 Posterior distribution of random effects parameters for vendace SDM 72
5.5 Posterior distribution of random effects parameters for vendace SDM 72

List of Figures

2.1 Locations of the 653 sampling sites along the coastal area of the GoB, used to recover whitefish data measurements. . . . 15
2.2 Raster layers of temperature, salinity and ice coverage 21
2.3 Raster layers of the continuous covariates. 24
2.4 Training covariates distribution. 25
2.5 whitefish and vendace larvae distributions are the target of our models. 26
2.6 Correlation plot of the continuous covariates 27
2.7 Scatterplot of the target variable, whitefish logdensity, and the covariates. On the top-right part of the plots for continuous covariates, the correlation coefficient is reported. When there were no larvae, we set target variables to 0.01, in order to get finite values of the logdensity. 28
2.8 Scatterplot of the target variable, vendace logdensity, and the covariates. 29
3.1 Main 3 developable surfaces used for projecting maps from an ellipsoid to a plane. 34
4.1 Stan pseudocode of our SDM. 53
5.1 Whitefish trace plots for the four models considered. 59
5.2 Vendace trace plots for the four models considered. 61
5.3 Predicted vs observed logdensities for training and test data 65
5.4 Residuals distribution, for whitefish 67
5.5 Residual plots, for whitefish 68
5.6 Posterior distribution of random effect parameters: σ^2, l and r, for whitefish model 69
5.7 Posterior distribution of linear parameters, for whitefish model 70
5.8 Posterior distribution of linear parameters (continuous variable), for whitefish model 71
5.9 Whitefish response curve 73

5.10 Whitefish larvae logdensity, future scenario 74
5.11 Whitefish larvae logdensity, current scenario 75
5.12 Predicted vs observed logdensities for training and test data 76
5.13 Residuals distribution 77
5.14 Residuals plot 78
5.15 Posterior distribution of random effect parameters: σ^2, l and r, for vendace model 79
5.16 Posterior distribution of linear parameters, for vendace model 80
5.17 Posterior distribution of linear parameters (continuous variable), for vendace model 81
5.18 Vendace response curve 82
5.19 Whitefish larvae logdensity, future scenario 83
5.20 Vendace larvae logdensity, current scenario 84
6.1 Current and future whitefish larvae logdensity 86
6.2 Current and future vendace larvae logdensity 86

Chapter 1

Introduction

Climate change is real.

Despite the ones denying it are still many, thanks to activists, mass media, and various organizations, climate change is slowly gaining attention among western citizens and policy makers[1], so that it is now generally accepted and recognized as one of the most urgent issue of today's society. Indeed, the consequences of global warming are already evident in several parts of the world.

In this thesis we will focus on the effects of climate change on the Baltic Sea, a shallow and close waters basin located between Finland and Sweden. According to Finnish meteorological institute's climate change research group, temperature in the Scandinavian region will rise, precipitation will increase, snow cover season will become shorter, and the amount of soil frost will decrease. Also, the sea level in the Baltic Sea will rise and the winter ice cover will reduce. The Gulf of Bothnia (GoB), in the northern part of the Baltic Sea is the most sensitive to climate change. Its unique ecosystem, combined with its location, make it a crucial element where to conduct climate change studies in order to predict the most likely scenarios induced by global warming and support decision making in coastal management and environmental protection. The aim of the thesis is to study how climate change will affect the main fishes population distributions in the GoB. We consider whitefish and vendace larvae, as these species are among the most relevant components of GoB fauna other than being economically and socially valuable in Finnish culture. To this scope, we will deal with spatial data, collected from different Finnish institutes, working in climate

[1] https://www.theguardian.com/environment/climate-change-scepticism

and environmental field. We will study how a set of, previously selected, environmental variables affects larvae spawning in the shallow waters of GoB, by using hierarchical Bayesian models, known as Species distribution models. We will then compute spatial predictions for current and future larvae density and report results by mean of raster maps, representing how the species distributions vary among the GoB.

All the work has been conducted using R and Stan software. The code to reproduce the analysis and the results presented in this thesis, is available at the following github repository [2]. The thesis is structured as follow.

In Chapter 2 we introduce the reader to the case study treated in the thesis. After a brief introduction about the area and the fishes species involved in the analysis, we discuss how we collect and then prepare the data. Care was taken to project all spatial data on the same reference system and in report them to the same scale.

In Chapter 3 we start by presenting the different types of spatial data commonly used in spatial statistics. Hierarchical spatial models are first presented in a general framework together with the main statistical tools we will use in the analysis. We then introduce the species distribution model for both single and joint species (SDM and JSDM respectively). A short summary of the state of the art in SDMs role in ecology is followed by an insight into how such models have been used to study climate change effects in previous researches.

In Chapter 4 we present and implement the SDMs for both whitefish and vendace larvae. In Chapter 5 we analyse and compare the different models fitting and select the best one for each species, in terms of prediction efficiency. We, hence, report the main outcomes of the models: random effect parameter estimates, as a measure of spatial associations and non-spatial variability in the data, and linear coefficients estimates, as a measure of association between the environmental covariates and the larvae density distribution in the GoB. We then report spatial predictive maps of larvae density, for both current and future scenarios.

In Chapter 6 we briefly report and discuss the final results of the thesis. Finally, we outline the main criticisms encountered in the analysis, proposing further possible extensions of the currently implemented models.

[2] https://github.com/ilapia/thesis-pia

Chapter 2

Case study

2.1 Whitefishes, vendaces and Gulf of Bothnia

The Baltic Sea can serve as an excellent case study area to research the consequences of climate change, as its signals are visible already now and predicted to be particularly strong in the future. Moreover, due to a faster trajectory of anthropogenic perturbations and to the presence of a diverse range of interacting pressures, that most coastal areas will experience only in the future, evidence from the Baltic Sea may deliver important insight for the future coastal ocean. Last but not least, the Baltic Sea region is also one of the most intensely studied coastal areas with high data density and many long-term data series.

The Gulf of Bothnia (GoB) is a brackish water basin between Sweden and Finland, in the northern part of the Baltic Sea, covering an area of approximately 600 × 120 km. Its coastal areas play a central role in the ecosystem and many Baltic fish stocks are dependent on the coastal regions for their reproduction. The GoB is mainly inhabited by 2 sea-spawning species of coregonids: whitefish (*Coregonus lavaretus*) and vendace (*Coregonus albula*). In the GoB, there exist two reproductive ecotypes of whitefish, an anadromous [1] river-spawning ecotype and a more stationary sea-spawning ecotype reproducing in the coastal areas (Lehtonen (1981), Sõrmus, I. & Turovski, A. (2003)). Sea-spawning whitefish and vendace, as native Baltic species, are a key element of coastal fish fauna. In addition, whitefishes are

[1] Anadromous fishes are the ones born in freshwater that spend most of their lives in saltwater and return to freshwater to spawn, such as salmons.

caught and sold for human consumption, hence, they are economically important for local fishermen, other than culturally valuable. Previous studies on the topic (Veneranta et al. (2013)) raveled that the catch of both whitefish ecotypes has decreased since the 1980s, while vendace's harvest has remained relatively stable.

The earliest larval stages of sea-spawning whitefish can be found in various habitats close to the shoreline, but the highest densities of larvae are observed along gently sloping, shallow sandy shores. Vendace reproduction occurs in the northernmost and less saline areas of the Bothnian Bay and larval stages use the shallow areas. Whitefish spawns in Autumn and the larvae have their born in spring. Vendace spawns later in the year but larvae births happen in the same season as that of whitefish.

The applied goals of the thesis are twofold:

- to compile the SmartSea data (section 2.3) with the existing whitefish data (section 2.2),
- to study how whitefish and vendace spawning distribution is affected by environmental changes.

2.2 Whitefish dataset

The whitefish dataset contains 63 different variables in total, which can be divided in two groups: environmental variables and larvae abundance data. The latter were measured at 653 different sampling sites in 26 sub-areas along the Finnish and Swedish coastal region of the GoB, during 2009-2011. Sampling sites locations are visible in Figure 2.1. Sampling sites were randomized following different approaches among the years (Veneranta et al. (2013)). The sites were visited once, approximately one week after ice break-up, in order to reach the early larval stages that had already started feeding. The sampling was made with a beach seine (509 sites) in all near shore sites (Hudd et. al. (1988)), with a Gulf- Olympia sampler (Hudd et. al. (1988), Aneer et al. (1992)) in open water in sub-areas 10, 19 and 20 (100 sites) and with a tow net sampler in open water (44 sites) in sub-areas 9-12 and 25. Using different gear in shallow and deep water was necessary in order to cover all coastal types in our sampling. For modeling purposes, we assume that, given their presence in the sampling site, the probability to catch one or more larval fish, is equal with all gear.

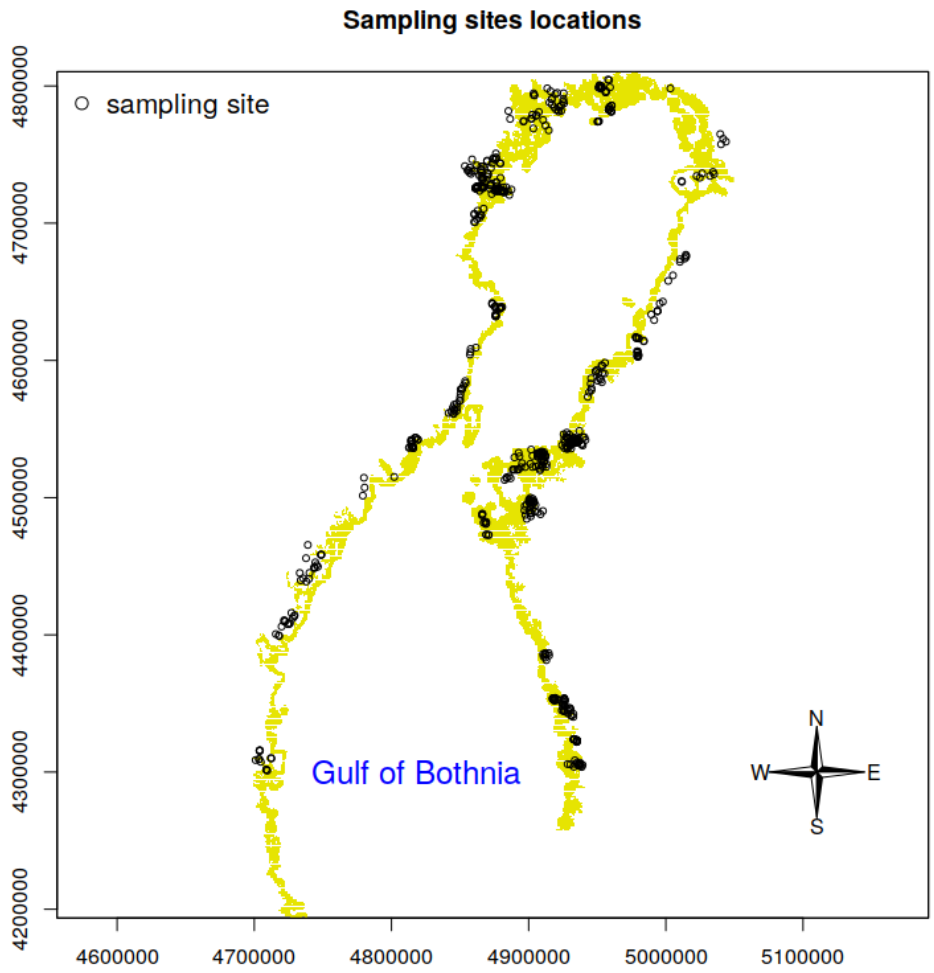


Figure 2.1: Locations of the 653 sampling sites along the coastal area of the GoB, used to recover whitefish data measurements.

Environmental variables were obtained from the Finnish Meteorological Institute (FMI), Finnish Environment Institute (FEI), Swedish Environmental Protection Agency (SEPA) and Helcom Map and Data Service (HELCOM), or constructed in the Finnish Game and Fisheries Research Institute (FGFRI). These variables were obtained by interpolations from point measurements or satellite photos and were already available as thematic Geographical Information System (GIS) map layers, converted to a grid with resolution of 300 m (see Vanhatalo et al. (2012)). All variables locations are stored according to the coordinate system ETRS89 which is a

3D Cartesian coordinate system constructed specifically to European areas. Whitefish dataset only stores the model-based values for the different variables at sampling locations. We use part of this dataset to fit our model. We then use two bigger datasets, containing only raster maps of environmental variables, related to current and future conditions, to produce predictive maps of larvae abundance on the whole GoB, which is the objective of our work.

2.3 SmartSea dataset

Information on temperature at the measurement sites was not available with the whitefish dataset, while the only present information on salinity was referred to winter measurements. In addition, the whitefish dataset contains environmental variables estimates for the current GoB condition only. For our predictive purpose, we seek to integrate the data with future values estimates for salinity, temperature and last day of ice coverage. The other variables considered, are expected to remain constant in the next years. This is why the variables temperature, salinity and future ice last coverage were extracted from a different dataset, elaborated by a team of leading experts in a variety of fields, as part of the SmartSea project [2]. SmartSea – "Gulf of Bothnia as Resource for Sustainable Growth", is part of the "Climate-Neutral and Resource-Scarce Finland"–program, funded by the Strategic Research Council of Academy of Finland.

The SmartSea project focuses on the GoB as its active maritime sector has a potential key role for "Blue Growth", the EU's long term strategy to support sustainable growth in the marine and maritime sectors in Europe. In addition, GoB is still relatively untouched compared to the other Baltic Sea areas and, due to its upper North location, climate change will most probably have the most drastic effects in this area. As the conditions of the marine areas will change in the coming decades, one of the central SmartSea project's goal is to build valuable data predictions concerning environmental conditions in the next 10-20 years, as well as projections up to 50 years in the future.

The SmartSea project aims to create new potential for development and growth of the GoB region by providing high quality data and efficient tools

[2] http://smartsea.fmi.fi

for marine spatial planning, which can supply science-based guidance and new innovations for the sustainable use of the Finland's marine resources.

SmartSea's objectives include[3]:

- Estimate the impacts of human activity changes on GoB's state.
- Identify key locations of natural resources and ecosystem functioning in the GoB.
- New innovations studied for efficient Blue Growth, including sustainable fish farming for enabling recycling of the nutrients and reduce the impact of climate change significantly.
- Estimate of the prospects and possible harmful effects of the seabed sand/gravel and mineral extraction to underwater nature of the GoB, in order to assess whether seabed extraction could provide a more sustainable way of mineral resource extraction than terrestrial methods.
- Accounting for climate change through the modeling approach to provide long-lasting guidelines for design requirements for offshore structures, economic optimization of food resources and aquaculture.
- Development of the multiform concept, which enables the combination of different offshore infrastructures in feasible and sustainable way.
- Provide open source Marine Spatial Plan toolbox with open access data for commercial and non-commercial applications
- Identify obstacles of sustainable transition
- Create a strategy for the GoB as resource for sustainable growth

Many scientific papers that base their studies on SmartSea data have already been published [4].

2.4 Data preparation

When dealing with spatial indexed data, the preliminary steps of the analysis consist of solving potential spatial misalignment present in data coming from different sources.

[3] http://smartsea.fmi.fi/what-is-smartsea-project/

[4] http://smartsea.fmi.fi/scientific-articles-published-in-smartsea/

To implement our species distribution model, we need to merge the SmartSea data containing temperature, salinity and ice coverage measurements with the whitefish dataset, so that all covariates that will be used in the model have the same coordinate system and resolution. The coordinate system (crs) in the whitefish data is ETRS89-extended / LAEA Europe, that is a UTM (easting-northing), with datum European Terrestrial Reference System 1989. The SmartSea data has latitude and longitude as coordinates, so it has a geographic coordinate system (WGS84), based on the World Geodetic System 1984 datum. The resolution for whitefish data is $300m \times 300m$ while for SmartSea data is $1M \times 1M$ (nautical miles). In SmartSea data all of the results are model based, namely the data are the output of simulations made with a climate change model which uses different data input depending on the considered year: historical simulation set is used for years 1975-2005, while, two different Representative Concentration Pathways[5] are used for years 2006-2059.

Since the last historical data in those models are run for year 2005, we use years 1995-2005 to represent the current values for temperature and salinity. We build the training temperature and salinity data as follow:

- temperature: the average of the temperature values in April-June over all years in 1995-2005 at water depths 0-9 meters,
- salinity: the average of the monthly salinity values over all months in years 1995-2005 at water depths 0-9 meters.

The original data contains monthly measurements of temperature or salinity at depth 0-9 meters from year 1975 to year 2059. In order to conform SmartSea data to whitefish data we implemented a function in R using raster and rgdal libraries.

The following procedure was applied to all the variables of interest in the SmartSea dataset:

1. Select the raster layers of interest (April-June 1995-2005) and average them.
2. Remove the land cells, encoded as 0 valued raster cells.

[5]Scenarios that include time series of emissions and concentrations of the full suite of greenhouse gases (GHGs) and aerosols and chemically active gases, as well as land use/land cover (Moss et al., (2008)) (RCP 4.5 and RCP 8.5) are used to assess the changing physical and biogeochemical properties of GoB, for years 2006-2059(Hordoir et al. (2019))

3. Extract latitude, longitude and variable values and create a raster layer with such coordinates.
4. Project the layer to the whitefish data crs system (ETRS89).
5. Upscale the resolution to the whitefish data resolution ($300m \times 300m$).
6. Crop the layer to the same extent of whitefish data.

We remark that the period taken into account consists of the months of April, June and July, since this is the season in which both whitefish and vendace larvae have birth. We then obtained these two environmental covariates by averaging the values in years 1995-2005. This is a common practice, when dealing with variables are obtained as output of a climate model, as in our case. Indeed, a typical property of climate models is that they do not necessarily produce accurate values for some particular year. By construction, their goal is to model the climate; that is the mean and variability of environmental variables over several years. They also produce more accurate predictions for differences in climate than for the actual climate itself. The upscaled raster layers for average salinity and temperature are visible in Figure 2.2. After obtaining them, the final datasets are created. The final withefish raster file, for current GoB variables, is obtained by merging the variables of interests from whitefish raster datafile and the values of salinity and temperature extracted from the raster layer created as explained above, for the corresponding raster cells. To implement the analysis, the environmental covariates are selected from the final whitefish data frame, which is created by matching each sampling location t in the whitefish dataset with the closest rastercell, with non zero value, for the variable of interest (temperature and salinity).

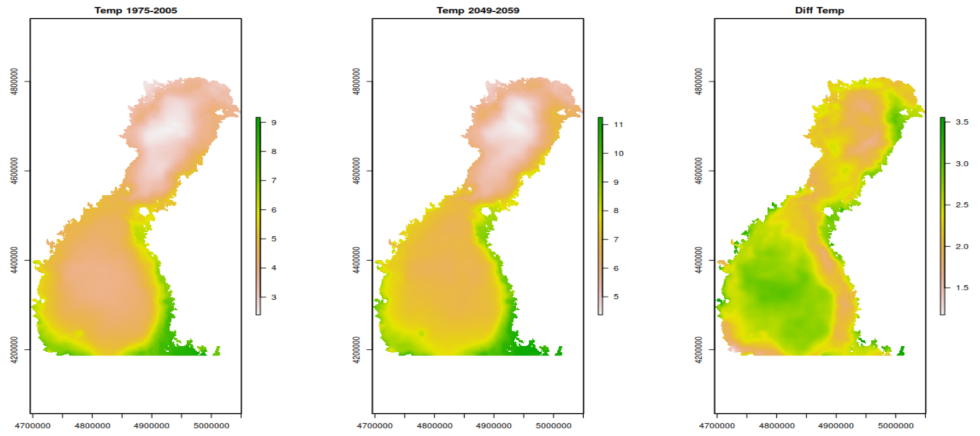
For what concerns the future values of the covariates, we applied a similar procedure, to recover average salinity, temperature and last week of ice coverage, in years 2049-2059. Raster layers for these variables were again provided by the SmartSea project, and obtained by applying the model RCP 8.5. We remark that, model based values, have been used for both salinity and temperature past and future values. For the last day of ice coverage, instead, we only use model based values for the future, as measured and hence, more accurate values were already presents in the whitefish dataset. The future covariate for this variable, was obtained by adding to the measured values the difference between ICELAST values predicted with the RCP model for years 2049-2059 and the one estimated for years 1995-2005(2.2).

2.5 Target and covariates

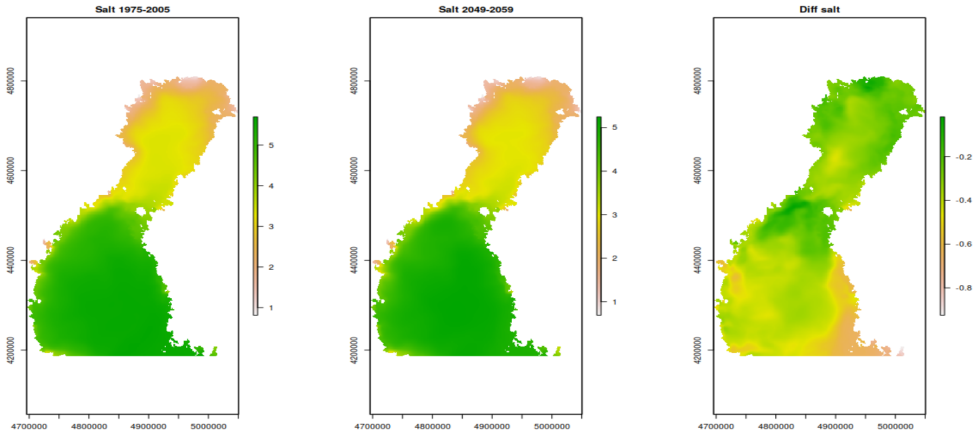
The first goal of our analysis is to study how whitefish and vendace spawning distributions are affected by environmental changes. To this purpose, a species distribution model (SDM) that models the abundance of whitefish and vendace larvae with respect to a set of nine environmental covariates has been implemented. Indeed, as well known in ecology, species distribution modeling is directly related to inferring species' responses to environmental factors.

When assessing the effects of climate change on species distribution, ecological understanding is a prerequisite to select the environmental covariates to include in SDM. When the selection is inadequate, a model may pick up irrelevant variables and its predictive power may decreases significantly (Mac Nally, R. (2000)). Therefore, care was taken to include the most relevant ecological variables in order to reach the best projections about the role of climate-driven effects on biotic patterns. GoB hosts rich variety of environmental conditions. Coastal areas are affected by inflows from land as well as shallow and complex topography. In the scale of GoB, there is a gradient in river influence, salinity, temperature and length of ice cover period from North to South. Earlier studies have shown that water salinity and temperature conditions have a primary role in shaping the large scale patterns of benthic macrophyte and invertebrate species in the Baltic Sea area, and these same variables are also expected to change the most because of future climate change (Meier et al. (2012), Meier, H. E. M. et al. (2012)).

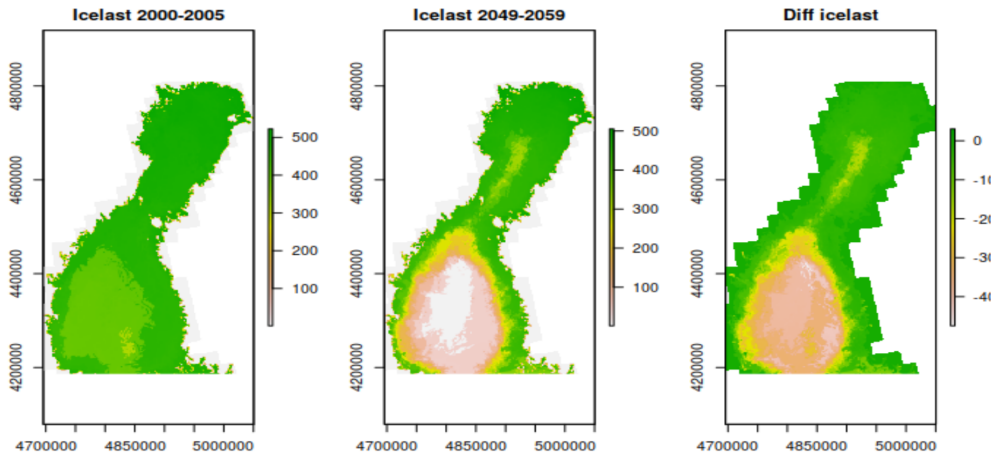
In the light of these considerations, we selected eight real-valued and one categorical abiotic environmental covariates, which have been found to be associated with whitefish and vendace distribution in earlier studies on related topics (Veneranta et al. (2013), Vanhatalo et al. (2020)).



(a) Temperature



(b) Salinity



(c) Last day of ice coverage

Figure 2.2: The figures show the final raster layers of average temperature (a) and salinity (b) and last day of ice coverage (c), measured in years 1995-2005 and predicted in years 2049-2059 and the difference between future and current values. These data were obtained from SmartSea Data project. Values for temperature and salinity are averaged on weekly outputs, obtained in the spring seasons, April-June. Last days of ice coverage are counted starting from the 1st of January of the previous year. When building the future raster layer used to make predictions, we transform the values from days to weeks and set them to 0 if there was no ice in a location for a certain year.

Covariate	Description	Source
BOTTOMCLS	Bottom type classification, a categorical variable with classes 0 = not shallow, 1 = open water, 2 = other, 3 = sand, 4 = sand/mud, 5 = sand/stone	FGFRI
DIS_SAND	distance to sandy shore, weighted by shallow area	FGFRI
FE300ME	average fetch (openness/exposure) over all directions	FGFRI
ICELAST09	last ice cover date in winter 2009-10, expressed in weeks, starting from the first week of 2009	FMI
RIVERS	Influence of rivers (weighted average distance to river mouths)	FGRI
DIST20M	distance to 20 meters deep water	FGFRI
CHL_A	Chlorophyll- a status index (as a proxy of phytoplankton biomass)	HLCOM
TEMP09M	mean temperature in April-June 1995-2005 at 0-9 meters depth	SSD
SALT09M	mean salinity in April-June 1995-2005 at 0-9 meters depth	SSD

Table 2.1: Environmental variables used in the study, available as thematic maps layers. The abbreviations are: Finnish Game and Fisheries Research Institute (FGFRI), SmartSea Data (SSD), Helcom Map and Data Service (HELCOM) and Finnish Meteorological Institute (FMI).

The quantities included in the model are reported in Table 2.1.

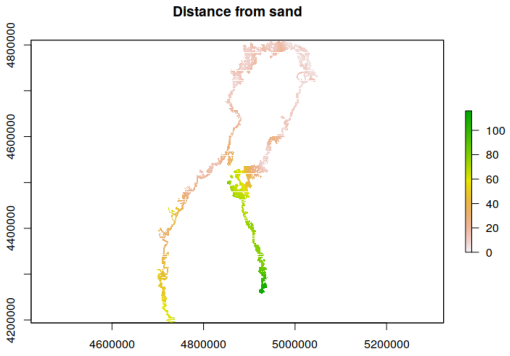
Continuous variables raster layers covering the whole GoB are visible in Figure 2.3, such values will be used to predict current larvae abundance out of the sampling sites. The distribution of the training covariates is instead shown in Figure 2.4, one can observe how, many of them, have a strongly asymmetric distribution, with the presence of potential outliers.

The dependent variables are WHISUM: the number of whitefishes caught in sampling occasion, or VENSUM: the number of vendaces caught in sampling occasion, depending on which species abundance is modeled. The two species distributions, visible in Figure 2.5, present an evident positive skew, with the majority of the measurements concentrated around quite small values, while a few potential outliers determine long tails on the right. Indeed, there are no larvae detected in almost 50% of the sites for whitefishes and 70% for vendace, while, in very few isolated cases larvae amounts greater than 500 were measured.

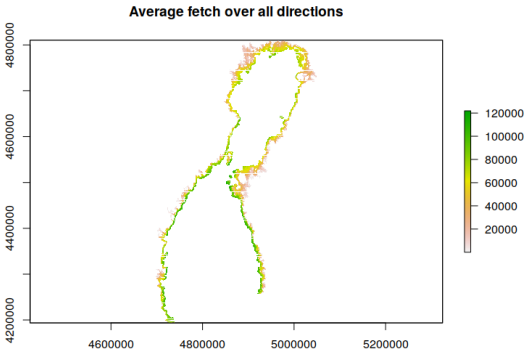
In addition we take as an "off-set" covariate for likelihood the variable VOLUME that indicates the volume of water sampled and as such, can be taken as a measure of the unit's exposure to larvae spawning. This meaning we model the larvae density per unit volume. In whitefish model, for instance, our response is given by $\log \frac{WHISUM}{VOLUME}$, or equivalently we add the quantity $\log(VOLUME)$ to the model. In the data there are 7 observations where the variable VOLUME has value 0, this is probably due to some error in collecting the data. We discard such data-points from the analysis to avoid troubles in computing the density. We further exclude from the data other 12 observations which present slightly different measurements in the same locations, as the presence of such datapoints results in affecting quite strongly the variability of the predictional outputs of our models.

Looking at the correlation among the continuous covariates the presence of high linear dependence can be identified, as shown in Figure 2.6.

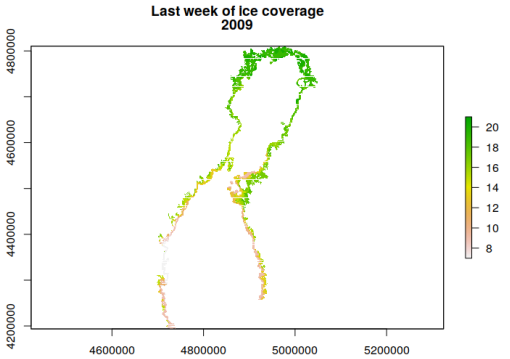
We notice how salinity and temperature are strongly correlated, nevertheless we are keeping both of them since it has been shown in laboratory experiments that both of these variables have physiological effects on larvae (Albert et al. (2004), Jäger (1981)). The target variable, instead, is not much correlated with the training covariates, when considering either whitefish either vendace larvae, as visible in Figures 2.7 and 2.8. All continuous covariates have been standardized so that they assume values on similar range and the use of zero-mean priors is justified.



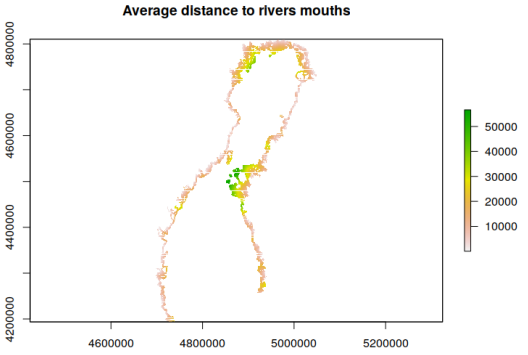
(a) Distance to sandy shore



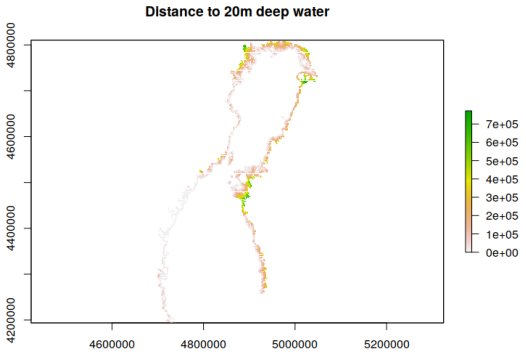
(b) Average fetch over all directions



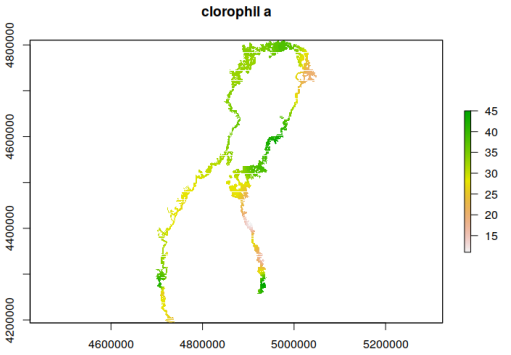
(c) last ice cover week in year 2009



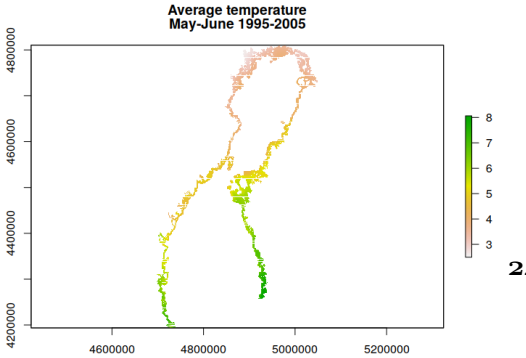
(d) Influence of rivers on average



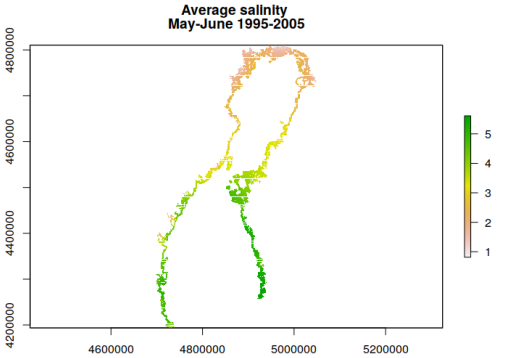
(e) Distance to 20 meters deep water



(f) Chlorophyll- a



(g) Average temperature in years 1995-2005



(h) Average salinity in years 1995-2005

Figure 2.3: Raster layers of the continuous covariates.

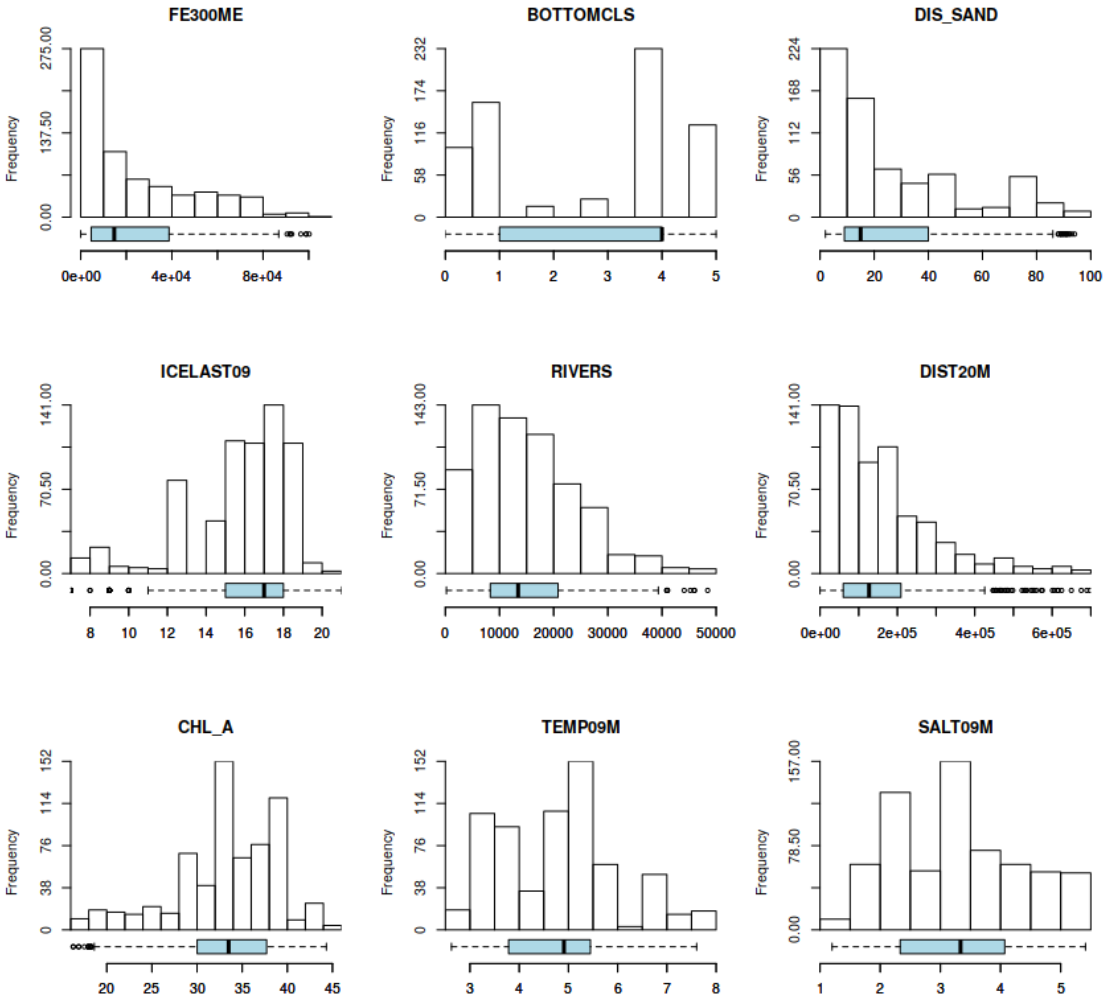


Figure 2.4: Training covariates distribution.

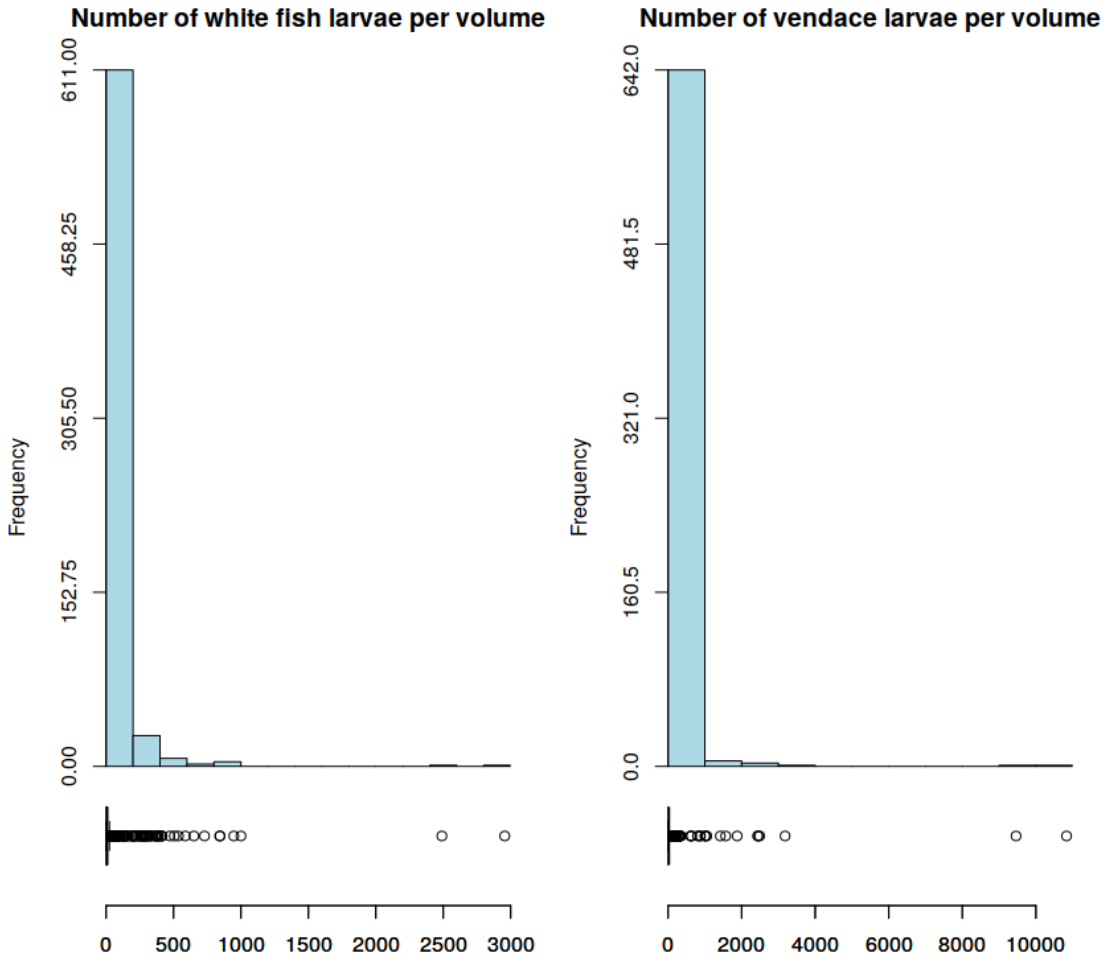


Figure 2.5: whitefish and vendace larvae distributions are the target of our models.

Figure 2.6: Correlation plot of the continuous covariates

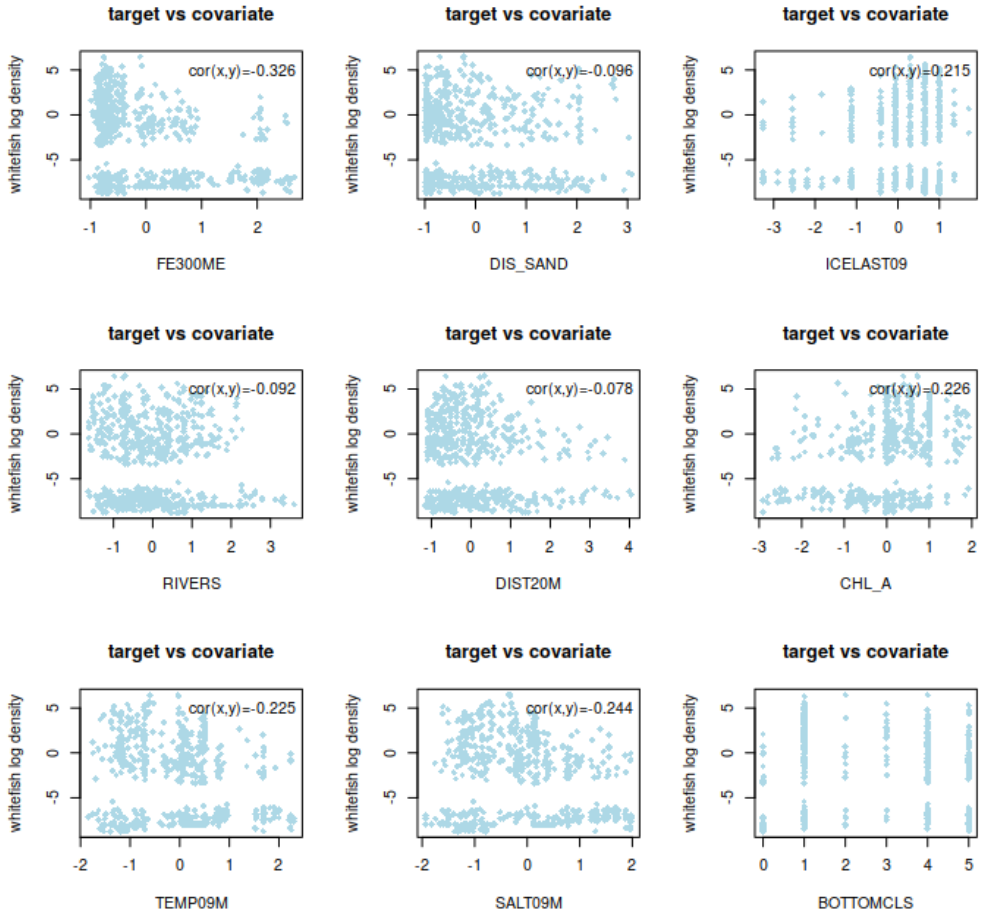


Figure 2.7: Scatterplot of the target variable, whitefish logdensity, and the covariates. On the top-right part of the plots for continuous covariates, the correlation coefficient is reported. When there were no larvae, we set target variables to 0.01, in order to get finite values of the logdensity.

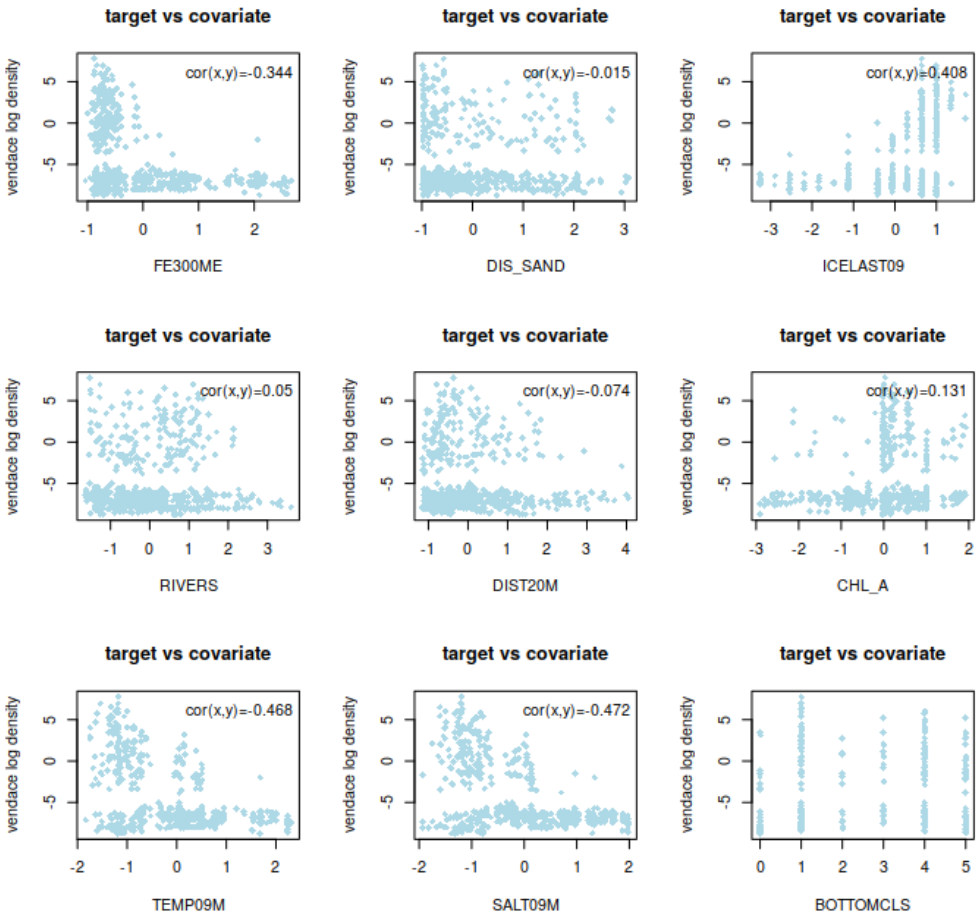


Figure 2.8: Scatterplot of the target variable, vendace logdensity, and the covariates.

Chapter 3

Methods

3.1 Spatial data

As presented in the previous chapter, in our studies, we use spatially indexed covariates and target variables. Accounting for this additional spatial information and dealing with analysis of this kind of data, in order to "reduce spatial patterns to a few clear and useful summaries" (Ripley et al. (1981)), are the objectives of a distinct area of statistical research that goes by the name of spatial statistics.

Traditional statistical theory bases its models on assumed independent observations. Even though independence is still a suitable benchmark from which to identify statistically significant non-independent phenomena, it is worth mentioning how, such a strong assumption, can not always be done when aiming at a more accurate representation of several real-world situations. The field of spatial statistics is based on the assumption that nearby units are in some way associated (Tobler (1979)), that yield to dependency among of the observations.

Typical objectives for spatial data analysis are related to inference and prediction of spatially indexed phenomena. For instance, we can be interested in what is the temperature at a spatial location $\mathbf{s} = (\mathbf{s_1}, \ldots, \mathbf{s_d})^\top$ and how can temperature measurements at n different locations $\mathbf{s_1}, \ldots, \mathbf{s_n}$ be used to predict the temperature at another location $\tilde{\mathbf{s}}$. Similarly we might be interested in analyzing geographic differences in spatial patterns between different regions, or forecasting temporal trends in spatial phenomena, such as the temporal change of annual average temperature in a certain region. Another typical objective is to explain spatial patterns with covariates which leads to spatial regression models.

Spatially indexed data are traditionally classified into three types:

- *Point referenced data* are measured at point referenced locations in space, namely, each datum has a value, $y(s)$, at location $\mathbf{s} \in \mathbf{D}$, where D is a fixed subset of $\mathbb{R}^d$, with d usually being equal to 2 or 3, that identifies the spatial area of interest. An example is the temperature measured at a specific location.

- *Areal data* describe phenomena over spatial regions. That is, a datum y_i is a value that is related to, for example, the average temperature over a region $\mathbf{A_i} \in \mathbf{D}$, fixed subset of $\mathbb{R}^d$, but does not tell what is the temperature at a specific location $\mathbf{s} \in \mathbf{A_i}$. A typical modern day example of areal data are remote sensed data, such as raster maps from satellite images. Another, traditional, example are administrative data, such as employment rates or disease prevalences within administrative regions (cities, counties etc.).

- *Point pattern data* describes the spatial presence of a phenomenon. Now D is itself random; its index set gives the locations of random events that are the spatial point pattern. $Y(\mathbf{s})$ itself can simply equal one for all $\mathbf{s} \in \mathbf{D}$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a marked point pattern process). A classical example is the spatial pattern of trees in a forest. Here, each datum is a location of a tree, $\mathbf{s_i}$, and the aim is to analyze the process that leads to a specific presence pattern.

In this thesis we will deal with point referenced data.

One inherent property of spatial data that needs to be taken into account is spatial (residual) correlation. We will discuss it in more details in the next sections.

3.2 Coordinate reference system

When working with spatially indexed data, it is fundamental to know how our data locations have been encoded.

A coordinate reference system (CRS) is a coordinate-based local, regional or global system used to locate geographical entities. In order to analyze spatial data, we need a homogeneous CRS for the area of interest.

A coordinate reference system is made up of four key components:

- Coordinate system: the X, Y grid upon which the data is overlayed and the method used to define where a point is located in space.
- Horizontal and vertical units: the units used to define the grid along the x, y axis.
- Datum: a modeled version of the shape of the Earth which defines the origin used to place the coordinate system in space.
- Projection Information: The mathematical equation used to flatten objects that are on a round surface (e.g. the Earth) so you can view them on a flat surface.

We are interested in areas on the surface of the Earth. As the Earth is not exactly spherical but has an irregular ellipsoidal shape, called geoid, one first needs to select a good approximation for the Earth surface. A geodetic datum or geodetic system is a coordinate system, and a set of reference points, used for locating places on the Earth (or similar objects). A datum is characterized by a set of numbers that define the shape, size, and position of an ellipsoid which best approximates the true surface of the Earth, either locally or globally. There are four main parameters that uniquely identify the ellipsoid (that is the datum) in use: semi-major axis (the equatorial radius), semi-minor axis (the polar radius), the degree of flattening, and the ellipsoid's position with respect to the center of the Earth. A standard example of datum used in cartography, geodesy and satellite navigation is the World Geodetic System: WGS84. Other examples are GRS80, NAD83 (the North American datum) and ETRS89 (the European Datum).

As part of a CRS, there are several coordinate systems that can be used to describe the location on the Earth. The simplest one is the Geographic coordinate system, in decimal degrees, where the location is described by latitude and longitude.

However, often the purpose is to analyze only a subset of the Earth's surface. If this subset is small enough, it is practical to use a map projection which projects the subregion from sphere to two dimensional plane. There are two main reasons for this. The map projections allow easy visualization on two dimensional plane and they allow the use of Euclidean metric to measure distances between locations. A map projection is a systematic representation of all or part of Earth's surface on a plane. It is well known from topology that it is impossible to construct a distortion-free representation of a globe on a two dimensional plane. Hence, when building maps,

one has to decide which aspects of the reality we want to reconstruct well and which parts of earth's surface the map should represent well, and select the most proper method accordingly. The general strategy to build maps is to use an intermediate surface that can be flattened. The globe (or part of it) is projected onto this intermediate surface, developable surface, after which it is flattened to a plane to produce a map. The most commonly used developable surfaces are the cylinder, the cone, the plane and the sinusoidal (as shown in Figure 3.1, realized by Battersby, (2017)).

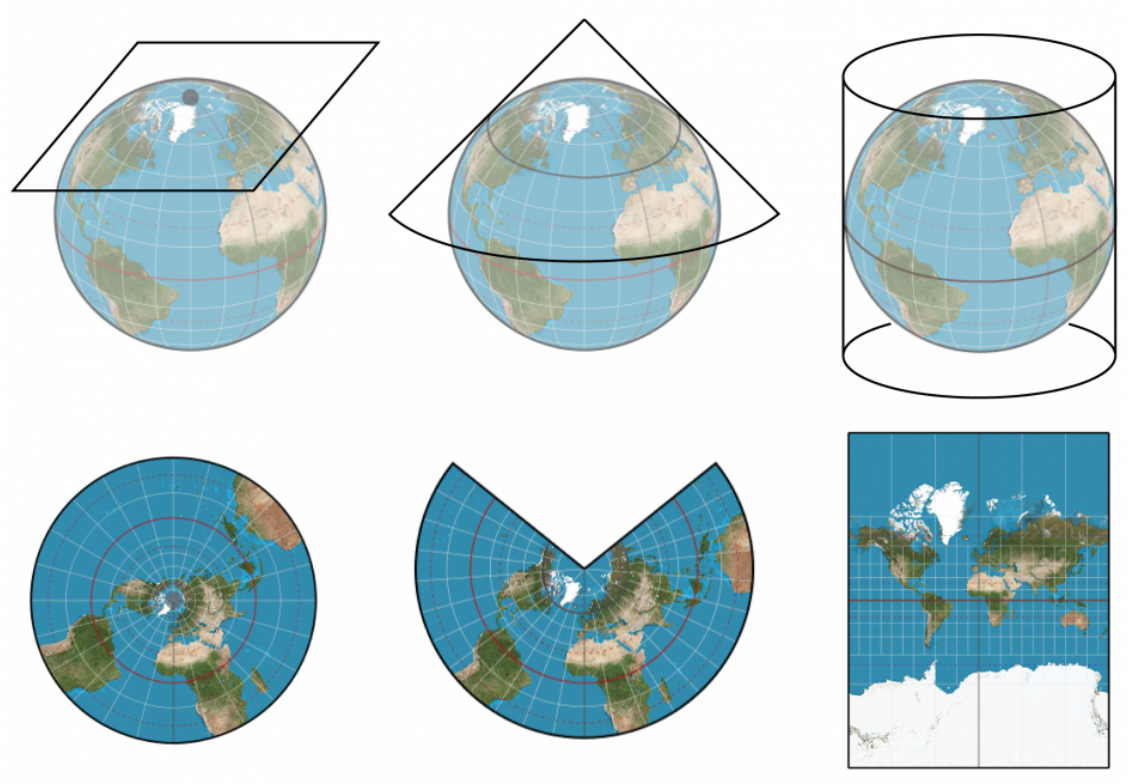


Figure 3.1: Main 3 developable surfaces used for projecting maps from an ellipsoid to a plane.

Map projections can be further classified based on which geometrical measure is preserved: area or angle. Equal-area maps are used for statistical displays of areal-referenced data, an example is the Lambert cylindrical projection. Angle (and hence, direction) preserving projections are also known as conformal, a classical example is Mercator projection. We cannot produce a map projection that preserves exact distances, hence, there are numerous projections available and one needs to find the one that best suites its case depending on applications.

Even the simplest map projections lead to complex transcendental equations relating latitude and longitude to positions of points on a given map. Therefore, rectangular grids have been developed for use by surveyors. In this way, each point may be designated merely by its distance from two perpendicular axes on a flat map. The y -axis usually coincides with a chosen central meridian, y increasing North, and the x -axis is perpendicular to the y -axis at a latitude of origin on the central meridian, with x increasing east. Frequently, the x and y coordinates are called eastings and northings, respectively, and to avoid negative coordinates, may have false eastings and false northings added to them. The grid lines usually do not coincide with any meridians and parallels except for the central meridian and the equator. The most common example is the Universal Transverse Mercator (UTM) grid, first adopted by the National Imagery and Mapping Agency, that divides the Earth into 60 zones, of 6 degrees of longitude in width, which are further segmented into 20 latitude bands of 8 degrees of latitude in height. Another example is the Universal Polar Stereographic (UPS).

3.3 Spatial process models

When we denote a collection of random variables $\{\mathbf{F}(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$ for some region of interest $D \in \mathbb{R}^2$, we are considering a stochastic process indexed by $\mathbf{s}$ to capture spatial association. These variables will be pairwise dependent with strength of dependence specified by their locations.

We start by considering as an example, the classical regression model

$$\mathbf{f} = \mathbf{X}\beta + \epsilon$$

with $\mathbf{f}, \epsilon \in \mathbb{R}^{\mathbf{n}}$ and $\mathbf{X} \in \mathbb{R}^{\mathbf{n}} \times \mathbb{R}^{\mathbf{p}}$, where one assumes that the residual terms ϵ_i are mutually independent for all $i \in 1, \ldots, n$. In a spatial framework such model has to be replaced by

$$\mathbf{f}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\beta + \epsilon(\mathbf{s})$$

where it is made explicitly visible that the covariates, $\mathbf{x_j}(\mathbf{s})$, with $j \in 1, \ldots, p$ are also spatially indexed. The random component $\epsilon(\mathbf{s})$ exhibits spatial correlation, by mean of a covariance matrix $\Sigma_\epsilon(\mathbf{s})$. When such matrix is not diagonal, the dependent variables, $\mathbf{f}(\mathbf{s})$, are not mutually independent given the covariates.

We can observe how spatial models are analogous to time series models in the sense that we need to explicitly account for spatial (residual) correlation in the former and temporal (residual) correlation in the latter.

Since our focus is on hierarchical modeling (see section 3.5), often the spatial process is introduced through random effects at the second stage of the model specification.

3.4 Spatial predictions

Assuming we have observed a realization of a process, $\mathbf{f} = [\mathbf{f}(\mathbf{s_1}), \ldots, \mathbf{f}(\mathbf{s_p})]^\top$, at a set of locations $S = \{\mathbf{s_1}, \ldots, \mathbf{s_p}\}$, spatial (non-/semi-parametric) prediction consists of using this information to update our knowledge concerning the values of the function at some other locations $\tilde{S} = \{\tilde{\mathbf{s}}_1, \ldots, \tilde{\mathbf{s}}_p\}$, $\tilde{\mathbf{s}}_i \in D$. This is a classical problem, which is called *Kriging* in traditional geostatistics[1]. When spatial covariate values $\mathbf{x} = (\mathbf{x}(\mathbf{s_1}), \ldots, \mathbf{x}(\mathbf{s_p}))$ and $\tilde{\mathbf{x}} = (x(\tilde{\mathbf{s}}_1), \ldots, x(\tilde{\mathbf{s}}_p))$ are available for incorporation into the analysis, the procedure is often referred to as universal kriging.

Spatial prediction in the presence of covariates can be divided into two cases

- Prediction when all environmental covariates are available throughout the study region. That is, $\mathbf{x}(\mathbf{s})$ is known for all $\mathbf{s} \in \mathbf{D}$. A typical example is an environmental covariate that is available as a raster map, such as elevation.

- Prediction when environmental covariates are available only for observation sites. That is, $\mathbf{x}(\mathbf{s})$ is known only for $\mathbf{s} \in \mathbf{S}$. An example would be an in-situ observations such as soil/water nutrient level.

The latter case is definitely the most common, since only seldom the study domain is surveyed extensively so that we would accurately know the environmental covariates in every location.

The classical approach to spatial prediction in the point-referenced data setting consists in applying minimum mean-squared error estimations. In a Bayesian framework, spatial prediction for a new location $\tilde{\mathbf{s}}$ amounts to finding the predictive distribution $p(\tilde{\mathbf{f}}|\mathbf{f}, \mathbf{s}, \mathbf{x}, \tilde{\mathbf{x}})$, where we denoted $\tilde{\mathbf{f}} = \mathbf{f}(\tilde{\mathbf{s}})$ and $\tilde{\mathbf{x}} = \mathbf{x}(\tilde{\mathbf{s}})$.

[1] https://en.wikipedia.org/wiki/Kriging

3.5 Hierarchical spatial models

We take the Bayesian perspective to spatial data analysis. We start with a general hierarchical spatial model (HSM) definition:

$$[Data|process, parameters] \quad p(\mathbf{y}|\mathbf{f}(\cdot), \gamma) \tag{3.1}$$
$$[process|parameters] \quad p(\mathbf{f}(\mathbf{s})|\theta) \tag{3.2}$$
$$[parameters] \quad p(\theta, \gamma) \tag{3.3}$$

where we have three hierarchical layers. The first layer is the observation process which describes the conditional distribution of observations $\mathbf{y} = [\mathbf{y_1}, \ldots, \mathbf{y_n}]^\top$ given the latent process $\mathbf{f}(\mathbf{s})$ and observation model parameters, γ. To simplify the notation, from now on, we will denote $\mathbf{y} = \mathbf{y}(\mathbf{s})$, as we always assume $\mathbf{y}$ being dependent on the locations $\mathbf{s}$. The second layer specifies the model for the latent process conditionally to the process model parameters θ, and the third layer specifies the prior for the parameters, also to be called hyperparameters. At this level of generality we have not yet defined what the process model will be in practice. However, it is considered to be a function of spatial coordinates, such as $f(\mathbf{s}) : \mathbf{D} \to \mathbb{R}$ where $\mathbf{s} \in \mathbf{D} \subset \mathbb{R}^\mathbf{2}$. The latent process can be a function of other variables as well and typical examples include time and covariates. The mapping can be also to vector spaces, so that $f(\mathbf{s}) : \mathbf{D} \to \mathbb{R}^\mathbf{d}$ in multivariate spatial models. Here d corresponds to the number of different types of observations, e.g., temperature, precipitation, wind etc. It should also be noticed that the above definition does not make any assumption of a priori conditional independence among the observations y, given the latent process. We have written $f(\cdot)$ in the conditioning side to make it explicit that this formulation allows observations to depend on the latent process very generally, for example, at specific locations or over a subset of the input domain. In the former case the observation model would simplify to $p(\mathbf{y}|\mathbf{f}(\cdot), \gamma) = \prod \mathbf{p}(\mathbf{y}(\mathbf{s_i})|\mathbf{f}(\mathbf{s_i}), \gamma)$, where $\mathbf{y} = [y(\mathbf{s_1}), \ldots, \mathbf{y}(\mathbf{s_n})]$ is a vector of point-wise observations.

An example of the latter case would be an observation model for areal totals $p(\mathbf{y}|\mathbf{f}(\cdot), \gamma) = \prod \mathbf{p}(\mathbf{y_i}|\mathbf{f_i}, \gamma)$ where $f_i = \int_{\tilde{\mathbf{s}} \in A_i} f(\tilde{\mathbf{s}}) d\tilde{\mathbf{s}}$ and $A_i \subset D$ are finite subsets of the input domain and the observation model describes how each observed datum y_i is related to the areal integral f_i. Note that A_i, $i = 1, \ldots, n$ do not need to be mutually disjoint.

The hierarchical construction (3.1) allows us to separate the process from

the observations, so that we can first formulate our prior knowledge and assumptions concerning the latent process and how the observations are linked to the underlying process. There are two main approaches to construct process models: the first approach consists of considering the process as a mechanistic (deterministic or stochastic) model for the underlying phenomena, as done, for instance, in a weather forecast model where the spatiotemporal process model is built from physical principles (Di Narzo et al. (2008)). The second approach, which is more common in spatial statistics, is to build a stochastic process that describes well the spatial (and temporal) dependencies between process function values at different spatial (temporal) locations. A typical example, is a spatial Gaussian process model that is used for spatial interpolation or to describe spatially correlated noise (Sang et al. (2010)).

3.6 Gaussian latent variable models

For a given probability space (Ω, F, P) and an index set D a real valued stochastic process is a collection of random variables which can be written as $\{\mathbf{f}(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$. A Gaussian process is a stochastic process such that every finite collection of those random variables has a joint multivariate Gaussian distribution. Therefore a Gaussian process, $\mathbf{f}(\mathbf{s})$, can be specified by its mean function $\mu(\mathbf{s})$ and its covariance function $k(\mathbf{s}, \mathbf{s}')$, so that

$$f(\mathbf{s}) \sim \mathbf{GP}(\mu(\mathbf{s}), \mathbf{k}(\mathbf{s}, \mathbf{s}'))$$

A Gaussian latent variable model (GLVM) is such that the latent process model in the HSM, conditional on hyperparameters, is explicitly assumed to be Gaussian random variable or Gaussian stochastic process. A classical example of a spatial GLVM is

$$f(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \beta + \phi(\mathbf{s})$$

where all additive components are mutually independent, $\beta \sim N(0, \Sigma_\beta)$ are linear weights, $\mathbf{x}(\mathbf{s}) = [\mathbf{x_1}(\mathbf{s}), \ldots, \mathbf{x_n}(\mathbf{s})]^\top$ is a vector of covariates and $\phi(\mathbf{s})$ is a spatial Gaussian process, typically called spatial random effect. This is a GLVM since each additive component is Gaussian, which implies that the marginal distribution for any $\mathbf{f} = [\mathbf{f}(\mathbf{s_1}), \ldots, \mathbf{f}(\mathbf{s_n})]$ is again Gaussian

$$f|S, \mathbf{X}(\mathbf{S}) \sim \mathbf{N}(\mathbf{0}, \mathbf{X}(\mathbf{S})\mathbf{\Sigma}_\beta \mathbf{X}(\mathbf{S})^\top + \mathbf{K}_{\phi,\phi}),$$

where $\mathbf{X}(\mathbf{S})^\top = [\mathbf{x}(\mathbf{s_1}), \ldots, \mathbf{x}(\mathbf{s_n})]$, $\mathbf{K}_{\phi,\phi} = \mathbf{Cov}(\phi, \phi)$ and $\phi = [\phi(\mathbf{s_1}), \ldots, \phi(\mathbf{s_n})]^\top$.

The above model is the generalization of linear mixed effects models to the spatial setting. It can be seen as an extension of simple linear models to allow both fixed and random effects, and therefore, it is also known as linear random effects model. These kinds of models are particularly used when there is non independence in the data, such as arises from a hierarchical structure.

3.7 Species distribution models in ecology

Species distribution models (SDMs) are key tools in ecology. They have been widely used to address theoretical and practical issues in several applied problems. Examples are studying species habitat preferences (Midgley et al. (2002), Latimer et al., (2006)), improving identification and management of conservation areas and natural resources (Austin and Meyers (1996), Kallasvuo et al., (2017)), finding additional populations of known species or closely related sibling species (e.g., Raxworthy et al. (2003)), seeking evidence of competition among species (e.g., Leathwick (2002)), and to evaluating species responses to environmental filtering under climate change scenarios (Clark et al., (2014); Kotta et al., (2019)).

In all of these applications, the main goals of statistical inference are to use species observations and information on the associated environment to infer the relationship between these two attributes and predict over regions of unsampled locations to build thematic species distribution maps (Gelfand et al. (2006); Elith, Jane & Leathwick, J.R.. (2009)).

The literature on species distribution models covers many modeling approaches and applications. In order to infer species' responses to environmental factors, generalized linear (or additive) models as well as machine learning approaches (such as regression trees or maximum entropy modeling (Elith, Jane & Leathwick, J.R.. (2009)) have been traditionally used in the past years.

Species distribution models attempt to provide detailed predictions of distributions by relating presence/absence or abundance of species to environmental predictors. Spatial prediction of species distributions is thus directly related to the concept of *environmental niche*, that refers to the position of a species within an ecosystem, describing both the range of conditions necessary for persistence of the species, and its ecological role in the ecosystem.

A first issue encountered in the implementation of SDMs is related to data sampling, including observer error, variable sampling intensity, and gaps in sampling. There may also be a spatial mismatch between different data sources (e.g. Agarwal et al. (2002)). Most research on development of distribution modelling techniques has focused on creating models using presence/absence or abundance data, where regions of interest have been sampled systematically (Austin and Cunningham (1981), Hirzel and Guisan (2002), Cawsey et al. (2002)). However, occurrence data for most species have been recorded without planned sampling schemes, and the great majority of these data consists of presence-only records, e.g. from museum or herbarium collections, that are increasingly accessible electronically (Graham et al. (2004), Huettmann (2005), Soberon and Peterson (2005)).The main problem with such occurrence data is that the intent and methods of collecting are rarely known, so that absences cannot be inferred with certainty. These data can also have errors and biases associated.

Despite SDM are still among the most popular models in ecology, further studies have shown that contemporary environmental factors alone are not sufficient to account for species distributions. Other ecological processes including dispersal, reproduction, competition, and the dynamics of large and small populations may also affect the spatial arrangement of species distributions (Gaston (2003)). Most species distribution models ignore spatial patterns and thus are based implicitly on two assumptions:

- the environmental factors are the primary determinants of species distributions
- the species have reached or nearly reached equilibrium with these factors.

These assumptions may or may not be adequate approximations, depending on the relative influence of environmental change, including both climate change and direct human transformation of landscapes, microevolution, and dispersal-related lags in species movement across landscapes. To the extent that model assumptions are violated, or that species distribution data are inadequate, simple species distribution models may fail to provide adequate predictive power, or may underestimate the degree of uncertainty in their predictions, resulting in poor characterization of the environmental response of the species.

Another key issue that remains problematic is that the methodology associated with SDMs is difficult to test and validate with real data. Model

performance has been largely studied on the basis of how well can the statistical models predict the patterns of presence-absence. Among the factors that have been widely cited as influencing model performance, both species and sample prevalence, defined as the proportion of sites in which a species is present, have been shown to be important (Pearce and Ferrier (2000), Kadmon et al. (2003), Hernandez et al. (2006), Santika (2011)), but other factors such as the properties of the specific indices used (Fielding and Bell (1997)) or the ratio between extent and area of occupancy have also been shown to be important (Lobo et al. (2008)). Species prevalence, also referred to as true prevalence, is the frequency of occurrence of the species overall in the landscape or biogeographic region of interest. It has been shown for example, that common species (i.e. species with higher prevalence) are more difficult to model than rare species (Guisan et al. (2006), Hernandez et al. (2006)), and it has also been argued that generalist species would be more difficult to model than specialists [2] (Seoane et al. (2006), Hernandez et al. (2006)). Sample prevalence (i.e. the proportion of sites within the sample used for model fitting and testing) can be different from species prevalence because researchers often target collection sites in which they know the species of interest is likely to be found, or introduce sampling bias due to road accessibility or other logistic reasons (Phillips et al. (2009), Ward et al. (2009), Lobo et al. (2010), Lobo and Tognelli (2011)).

A further complication comes from the fact that, in order to validate model outputs, the data is often divided into a training and a testing set, the training set being used during model calibration and the testing set being used in order to calculate performance measures (Fielding and Bell (1997)). Depending on how these training and testing sets were designed, the training sample and the testing sample could also show differences in prevalence which will affect SDM performance measures differently (Lobo et al. (2008)). All these issues are difficult to explore with real-world data, since it would require extensive knowledge regarding species and sample properties, including their relationship to environmental gradients. In order to provide key insights on how and when do different statistical models differ, artificial species models have been used (Meynard et al. (2012)).

[2] A generalist species is able to thrive in a wide variety of environmental conditions and can make use of a variety of different resources. A specialist species can thrive only in a narrow range of environmental conditions or has a limited diet.

3.8 Joint species distribution models

The approaches presented so far model each species separately and cannot account for species interactions nor shared responses to the environment. However, species interaction with other species is potentially as important factor as its response to environment. Moreover, in many practical situations, data from species can be patchy or scarce, in which case, sharing information between species can significantly improve models' predictive performance. In order to model species to species interaction, many different approaches have been used through the past years, such as canonical correspondence analysis (Ter Brake (1986)), nonmetric multidimensional scaling (Kruscal (1964)), and the use of multivariate regression trees (De'ath (2002)). These approaches make it possible to utilize data on all species, allowing for a more complete representation of the data, but they fail to show in detail how the relationship between environmental covariates and the species community builds up from species-specific responses (Gelfand et al. (2005)). Moreover, these approaches can be difficult to use for predicting, e.g., how the community would respond to changing environmental conditions. A hierarchical approach instead combines a set of species-specific models with a community-level model (Ovaskainen and Soininen, (2011); Hui et al., (2013); Clark et al. (2017); Thorson et al., (2015); Hartmann et al., (2017)). Species-specific models are linked to each other by a higher-level structure, and they are thus fitted simultaneously to the data. The hierarchical structure makes it possible to include also species with very limited data, and it thus facilitates the analysis of sparse data sets with large numbers of very rare species. On top of the species-specific inference, the approach provides a compact summary of the entire community, which can be used to assess, e.g., how community similarity depends on variation in environmental covariates and other factors. For these reasons, joint species distribution models (JSDM) have gained increasing attention in recent years (Warton et al., (2015)).

3.9 Climate change and SDM

Quantifying networks of ecological interactions and assessing their changes due to global environmental change has been the focus of much recent studies (Fortuna and Bascompte (2006); Tylianakis et al. (2008); Araujo et al. (2011)). Many of the terrestrial and marine ecosystems involve latent

networks that cannot be measured directly. Hierarchical Bayesian models and Gaussian processes provide tools to model biologically reasonable joint effects of environmental covariates in the model. Moreover, they allow to combine a priori biological knowledge with all the information in heterogeneous distribution and experimental data, as well as produce posterior distributions from which uncertainty summaries can easily be extracted. We can also explicitly model the spatial autocorrelation in the data not explainable by the covariates (Vanhatalo et al. (2019)). These are the reasons why we chose hierarchical Bayesian models and Gaussian processes to study the spatial patterns of whitefish and vendace under current and future climate conditions.

As highlighted in recent studies (Kotta et al., (2019)), SDMs that account to study climate changes effects are based on hypothesis that may not always apply optimally to the case. Mainly, these models do not include mechanistic, causal, knowledge on a species' dependence to its environment or other species. This is why they are often called correlative, even though they do not rely only on correlations. To overcome this problem, the use of informative Gaussian process (GP) prior for the model latent function of the environmental covariates has been proposed (Mäkinen et al, (2018)).

Moreover, an important assumption behind most of the SDMs is that relationships between the observed patterns of environment and species distribution will remain unchanged over the study region and time. Due to non-stationarity of ecosystem processes such an assumption seems unrealistic and will likely be violated under future climate conditions, when statistical patterns between current species distributions and the environment are expected to become uncoupled. Moreover, under future climate scenarios, statistical SDMs are often applied to environmental conditions which differ from the ones they have been initially trained, this may yields to unreliable results. One potentially important aspect that may affect range shifts driven by climate change is the presence of locally adapted populations with varying potential to respond to shifts in environmental conditions. Local differentiation in tolerance is currently poorly understood but an inclusion of such genetic within-species variation likely improves the model performance.

Chapter 4

Proposed methodology and priors setting for HSDM

4.1 HSDM for whitefish data

Hierarchical specie distribution model follows the generic hierarchical structure as presented by Wikle (2003), Cressie and Wikle (2011) and Banerjee et al. (2015).

$$[data|process, parameters] : p_Y(\mathbf{y}(\mathbf{x},\mathbf{s})|\mathbf{f}(\mathbf{x},\mathbf{s}),\eta),$$

$$[process|parameters] : p_f(f(\mathbf{x},\mathbf{s})|\theta),$$

$$[parameters] : p(\eta,\theta),$$

where we have three hierarchical layers. The first layer is the observation process which describes the conditional distribution of observations $\mathbf{y} = [\mathbf{y_1}, \ldots, \mathbf{y_n}]^\top$ given the latent process $f(\mathbf{x},\mathbf{s})$ and observation model parameters, η. The data, $\mathbf{y}(\mathbf{x},\mathbf{s})$, are evaluated at spatial location $\mathbf{s} \in \mathbf{D} \in \mathbb{R}^\mathbf{2}$ with associated covariates $x \in \mathbb{R}^d$. The second layer specifies the model for the latent process conditionally to the process model parameters θ, and the third layer specifies the prior for all the unknown parameters, also called hyperparameters. Such models can be seen as an extension of Generalized linear models to (spatial) random effect. Furthermore they can be extended to Hierarchical multivariate species distribution model, by considering $j \in 1, \ldots, J$ different species.

In our application the variable of interest $\mathbf{y}$ represents the number of whitefishes caught at the various sampling locations. Let $i \in 1, \ldots, N$ represent the sampling sites and j the number of different species considered (in this initial case we have $J = 1$).

At first, we assume the response variable has distribution

$$y_{ij}|f_{ij} \sim \text{Poisson}(V_i e^{f_j(\mathbf{x_i},\mathbf{s_i})})$$

where $e^{f_j(\mathbf{x_i},\mathbf{s_i})}$ models the larval density in water, whereas V_i is the sampled volume of water, and serves as an offset. Parameters $\beta_\mathbf{j}$ are the linear weights in the latent variable's linear model

$$f_j(\mathbf{x_i}, \mathbf{s_i}) = \mathbf{x_i}(\mathbf{s_i})^\top \beta_\mathbf{j} + \phi_\mathbf{j}(\mathbf{s_i})$$

where $\mathbf{x}(\mathbf{s_i}) \in \mathbb{R}^\mathbf{15}$ contains the environmental covariates at location $\mathbf{s_i} \in \mathbb{R}^\mathbf{2}$, and specify the i-th row of the design matrix $X \in \mathbb{R}^N \times \mathbb{R}^{15}$.

Given $\mathbb{E}(y_i) = \mu_i \ \forall i \in 1, \dots, N$ observations, the model can be written as:

$$g(\frac{\mu_i}{V_i}) = \mathbf{x}_i(\mathbf{s})^\top \beta_\mathbf{j} + \phi_\mathbf{j}(\mathbf{s_i})$$

where $g(\cdot) = \log(\cdot)$ is the link function.

For the linear coefficient we assume uninformative priors $\beta_{dj} \sim N(0{,}10) \ \forall d \in 1, \dots, 15$. The variable $\phi(\mathbf{s})$ models the spatial random effects that describes temporally constant associations, unexplained by the available covariates.

We construct and analyse different GLVMs, models such that the process $f_j(\mathbf{x}, \mathbf{s})$ conditional on hyperparameters, is assumed to be Gaussian.

We first implement a model where $\phi \sim N(0, \sigma_\phi^2)$, i.e. a model where there is no spatial correlation induced by the random effects: such model fails to account for spatial correlation among data, hence predictions may result unprecise.

To account for spatial random dependences, we implement a linear mixed model (or linear random effects model). We model ϕ by using a zero mean Gaussian Process

$$\phi_j(\mathbf{s})|\sigma_\phi^2, l \sim N(0, \Sigma_\phi)$$

where l and σ_ϕ^2 are the range and intensity parameters of the covariance function. This is a GLVM since each additive component is Gaussian, which implies that the marginal distribution for any $\mathbf{f} = [\mathbf{f}(\mathbf{s_1}), \dots, \mathbf{f}(\mathbf{s_n})]$ is again Gaussian:

$$\mathbf{f}|\mathbf{S}, \mathbf{X}(\mathbf{S}) \sim \mathbf{N}(\mathbf{0}, \mathbf{X}(\mathbf{S})\mathbf{\Sigma}_\beta\mathbf{X}(\mathbf{S})^\top + \mathbf{\Sigma}_\phi)$$

where $\mathbf{X}(\mathbf{S})^\top = [\mathbf{x}(\mathbf{s_1}), \dots, \mathbf{x}(\mathbf{s_n})]$, $\Sigma_\phi = Cov(\phi, \phi)$ and $\phi = [\phi(\mathbf{s_1}), \dots, \phi(\mathbf{s_n})]^\top$. In our case $\Sigma_\beta = \sigma_\beta^2\mathbf{I}$. Denoted by $\Sigma_\beta(\mathbf{s}, \mathbf{s}') = \mathbf{x}(\mathbf{s})\mathbf{\Sigma}_\beta\mathbf{x}(\mathbf{s}')^\top$, we have that

$$f \sim GP(0, \Sigma_\beta(\mathbf{s}, \mathbf{s}') + \mathbf{\Sigma}_\phi(\mathbf{s}, \mathbf{s}')) \tag{4.1}$$

or equivalently

$$f \sim GP(\mathbf{x}(\mathbf{s})^\top \beta, \mathbf{\Sigma}_\phi(\mathbf{s}, \mathbf{s}')) \tag{4.2}$$

We first implement the model using two different covariance matrix for the random effect ϕ, from the family of Matérn covariance function (see section 4.3).

4.2 SDM with additional random effect

The estimates obtained in the first stages of our studies reveal some criticism in the models. When looking at the parameter posterior samples in the model with Matérn covariance function, and target distribution Poisson, we notice that the scale parameter, l, is smaller than the sampling resolution (0.300 km), on average. This suggests that the spatial random effect is converging to capture overdispersion in the data.

Indeed, when using a Poisson distribution to model species abundance, the variance equals the mean. However, the amount of variation for each sampling unit is typically higher than expected by a pure Poisson process(Lindén and Mäntyniemi (2011)). This extra variation, termed as overdispersion, is caused by spatiotemporal heterogeneity in the process that produces data, typically due to observation errors (sampling effects and observation inaccuracy) or process errors (variation in the Poisson intensity). One way to fix this is to add independent and identically distributed (i.i.d.) random effects in addition to the spatially structured random effect, or equivalently, one could use Negative-Binomial observation model (Jia Liu et al. (2020)).

This means we will assume now

$$y_i|\beta, \phi_i, \epsilon_i \sim \text{Poisson}(\epsilon_i V_i e^{f(\mathbf{x_i}, \mathbf{s_i})})$$

where ϵ_i is an independent random effect, that can describes, for example, non-structured stochasticity due to environmental conditions during the data collection. Since volumes V_i are approximately equal we gave a joint prior for the random effects with $\epsilon_i \sim \text{Gamma}(r, 1/r)$, so that $E[\epsilon_i] = 1$ and $Var[\epsilon_i] = 1/r$. This yields to

$$y_i|\beta, \phi_i, \epsilon_i \sim \text{Negative} - \text{Binomial}(V_i e^{f(\mathbf{x_i}, \mathbf{s_i})}, r).$$

Where $E(y_i) = V_i e^{f(x_i, s_i)}$ is the mean of the distribution, while the variance is given by $Var(y_i) = E(y_i) + \frac{E(y_i)^2}{r}$, hence, r is an overdispersion parameter,

such that when $r \longrightarrow \infty$, Y approaches a Poisson distribution. We set as prior for the overdispersion parameter

$$r \sim \text{Gamma}(2,0.1).$$

4.3 Covariance function and priors setting

The covariance matrix incorporates the spatial association between locations and it can be represented as a function that describes the decay in correlation between pairs of points with distance. There is a large literature on alternative covariance functions and the properties they impose to Gaussian process (see Rasmussen and Williams (2006)). In this thesis we introduce the widely used Matérn covariance function, a class of stationary[1] and isotropic [2] covariance functions:

$$k_{matérn}(\mathbf{s}_i, \mathbf{s}_j) = \sigma^2 \frac{2^{1-\nu}}{\Gamma\nu} \Big(2\nu \frac{\sum_{d=1}^{D}(s_{i,d} - s_{j,d})^2}{l_d^2} \Big)^{\nu/2} K_\nu \Big(\sqrt{2\nu \frac{\sum_{d=1}^{D}(s_{i,d} - s_{j,d})^2}{l_d^2}} \Big)$$

where Γ is the Gamma function and K_ν is a modified Bessel function. It is function of three hyperparameters: ν controls the smoothness of the process realizations, l controls the correlation range and σ^2 controls the magnitude of variation in process realizations.

The correlation range is a common concept in spatial statistics, usually defined as the distance c, at which the covariance function has dropped to 5 % of its maximum. If we assume l does not vary with spatial dimensions, we can write the Matérn covariance, as function of the distance, so that, c is the quantity such that:

$$Cov(c) = 0.05Cov(0), \tag{4.3}$$

When considering the exponential covariance function, c can, then, be expressed as a function of l, from equation 4.3, we easily obtain $d = l\log(20)$.

When setting the prior distributions for the hyperparameters of the covariance function, it is good practice to give them at least weakly informative priors, since information about these hyperparameters are typically

[1] A covariance function is called stationary if it is a function of $h = s - s'$ only. Note that this means weak stationarity for the corresponding stochastic process whereas strong stationarity would mean that all of its finite dimensional distributions are invariant to translations.

[2] A covariance function is called isotropic if it is a function of the distance only $||h||_2 = ||s - s'||_2$.

scarce. ν is often assumed to be fixed, since with the choice of $\nu = p + 1/2$ where $p \in \{0, 1, 2, \dots\}$ the Matérn covariance function can be solved analytically (Rasmussen and Williams (2006), Section 4).

We, hence, decide to keep ν fixed. At a preliminary phase of the study we consider two alternative settings for the parameter ν. First, we set $\nu = \frac{1}{2}$ leading to the exponential covariance function

$$\Sigma_\phi(s_i, s_h) = \sigma^2_\phi \exp\Big(- \sqrt{\sum_{d=1}^{2} \frac{|s_{i,d} - s_{h,d}|^2}{l_d^2}} \Big)$$

We then set $\nu = \frac{3}{2}$, this gives

$$\Sigma_\phi(\mathbf{s}_i, \mathbf{s}_h) = \sigma^2_\phi(1 + \sqrt{3}r(\mathbf{s}_i, \mathbf{s}_h))e^{-\sqrt{3}r(\mathbf{s}_i, \mathbf{s}_h)}$$

where

$$r(\mathbf{s}_i, \mathbf{s}_h) = \sqrt{\sum_{d=1}^{2} \frac{(s_{d,i} - s_{d,h})^2}{l_d^2}}$$

In the first stages of our work, when running our initial models, we observe that outputs are not sensitive to the choice of the spatial covariance function, based on the two different ν. This is also confirmed by other studies conducted on the same data (Vanhatalo et al. (2012)) so, once reached more advanced stages of our analysis, we select the Mátern covariance function with 1/2 degrees of freedom, which is computationally less heavy.

For the covariance hyperparameters, l and σ^2, we assume independent priors. We remark that, as shown in previous studies, the length-scale and magnitude are under identifiable and the proportion $\frac{\sigma^2}{l}$ is more important to the predictive performance than their individual values (Diggle et al. (1998); Zhang (2004); Diggle and Ribeiro, (2007)). Since we cannot expect the data to be informative on both the variance and the length-scale parameters and, hence, prior information is needed to conduct sensible inference on both parameters. In this thesis we use two different sets of quite informative priors. In both cases we assume the hyperparameter l does not vary over d, this allows one to express Matérn covariance as function of the euclidean distance between locations, and, for the exponential covariance matrix, as function of the correlation range.

At first, we set as priors:

$$\sigma^2_\phi \sim \text{Student} - t^+_{\nu=4}(\mu = 0, \sigma = 1)$$

$$1/l \sim \text{Student} - t^{+}_{\nu=4}(\mu = 0, \sigma = 1)$$

that favor smooth functions with small variability, as done in Vanhatalo et al. (2012). These priors can be considered as weakly informative priors that prefer smooth and small in magnitude spatial random effects (Gelman (2006),Vanhatalo et al. (2020)). For the variance parameter σ^2_ϕ, we chose a prior distribution concentrated around zero as we want to restrict the magnitude of the spatial random effect towards zero. While, for the length-scale, l, which governs how fast the correlation decreases as a function of distance, by setting a prior for the inverse of the length scale, $1/l$, we penalize short length-scale in favor of longer length-scale and hence long correlation range.

As second, we use the priors:

$$\sigma^2_\phi \sim \text{Student} - t^{+}_{\nu=4}(\mu = 0, \sigma = 1)$$

$$l \sim \text{Gamma}(\alpha = 10, \beta = 1)$$

While the magnitude prior is left unchanged, we set a new prior for l, so that it gives about 0 probability for the correlation range bigger than 50 km or smaller than 5 km. Given the correlation range definition (4.3), as we are using an exponential covariance function, we can recover directly the relation between length scale parameter and correlation range c, so that we have:

$$P(5 < c < 50) = P(\frac{5}{\log(20)} < l < \frac{50}{\log(20)}) \approx 0.97.$$

With the first prior setting, we can notice how the posterior of l was approaching too smaller values (posterior mean around 0.6, hence closer to the data resolution), so that the spatial random effect was accounting for overdispersion, without modeling the actual correlation between sampling sites. The overdispersion parameter, instead, got quite bigger values, on average, so that the Negative Binomial was actually converging to a Poisson distribution. With this second setting we prefer even longer length scale, bigger than the resolution (0.3km), so that overdispersion is modeled by the Negative-Binomial parameter, r, and the random effect captures correlation between farther locations.

4.4 Posterior distributions and MCMC

Markov chain Monte Carlo (MCMC) is a family of techniques for sampling probability distributions. MCMC methods are extremely popular in field

such as Bayesian statistics, as the Monte Carlo estimate is proved to converge to the correct distribution, when the chain length increases towards infinity. Even though they are theoretically correct, these algorithms are often hard to implement in practice. Reaching convergence, and hence reliable sampling estimates, can be really hard and time consuming. In addition, the sample chains might mix poorly which results in high autocorrelation and low number of efficient samples.

Given the hierarchical model described by equation (3.1), our inferential objectives are the posterior distributions of the hyperparameters and the latent function, as well as the predictive distribution for new observations. These posterior probability distributions cannot be solved analytically, in general. This is why MCMC methods are typically applied in Bayesian modelling, with target distributions being posterior distributions of unknown parameters, or predictive distributions for unobserved phenomena.

In our case-study we implement models in Stan (Stan Development Team. (2018)) and make use of MCMC algorithms from the Stan software, to recover the posteriors of our models' parameters.

When running the model to estimate the posterior density of the parameters, Stan uses a tailored Hamiltonian Monte Carlo (Duane et al., (1987); Nael (2012)) where the tuning of the sampling parameters is done in automated manner (Hoffman and Gelman, (2014)). Hamiltonian Monte Carlo utilizes the gradient information of the log posterior distribution to direct the sampling to interesting regions and, hence, to speed up the convergence and improve mixing. When using this method, understanding how to tune the sampling parameters can be challenging. In addition, when assuming a GP prior for the target distribution, the computationally most time consuming operation is the inversion of the covariance matrix, needed to recover its logdensity and its derivative. To overcome this computational limitation a number of sparse approximations for GP have been suggested in the literature (see, for instance, the fully independent training conditional (FITC) sparse approximation (Vanhatalo et al. (2007), Snelson and Ghahramani, (2006); Quiñonero-Candela and Rasmussen, (2005))).

We implement our model as follow.

We define the latent variable $\mathbf{f}$, which represents the larvae logdensity, as in equation (4.1), rather than as in equation (4.2). Namely, we assume $\mathbf{f}$ being a Gaussian process with 0 mean and covariance matrix given by the sum of two components: the covariance matrix induced by the linear effects, that is $\Sigma_\beta = \sigma^2_{lin}\mathbf{X}^\top(\mathbf{s})\mathbf{X}(\mathbf{s})$ and the covariance matrix induced

by the spatial random effect, that we assume following the exponential covariance function. A small amount of random noise (jiitter = 1e-6) was also added to the diagonal elements, to ensure the resulting matrix to be positive definite and avoid computational troubles.

Defining the model as stated above, allows one to sample from the posterior of $\mathbf{z} = \mathbf{L}^{-1}\mathbf{f}$, where we set a Standard Gaussian prior for $\mathbf{z}$, and denote by $\mathbf{L}$, the Cholesky decomposition of $\mathbf{f}$ prior covariance, chosen as a matrix that approximates the square root of the posterior covariance of the latent variable $\mathbf{f}$. This method improve significantly the computational efficiency of the model, since the elements $z_1, \ldots, z_n$ are approximately independent whereas the elements $f_1, \ldots, f_n$ have very high correlation between them. The high correlation in f slows down the convergence and mixing of the MCMC chain. Such implementation, allows us to recover posterior estimates for the latent variable, without sampling from the linear coefficients, which we recover while doing predictions with a separate function, out of the Stan model, as shown in the next section. In Figure 4.1 we report a pseudo-code, showing the core section of the model, built with the above method, which was then applied to several different combinations of priors, datasets and target distributions, as discussed in the previous sections.

In order to approximate the exact solution of the Hamiltonian dynamics we, then, need to choose a step size governing how far we move each time we evolve the system forward. That is, the step size controls the resolution of the sampler. The adapt_delta parameter in Stan, defines the target average proposal acceptance probability during Stan's adaptation period, and increasing it will force Stan to take smaller steps. To help models reaching convergence, we set adapt_delta = 0.99.

To assess the convergence of our MC we use the R-hat (Gelman & Rubin (1992)) convergence diagnostic, returned by Stan fitting function. R-hat compares the between- and within-chain estimates for model parameters and other univariate quantities of interest, considering the threshold R-hat larger than 1.1, as an indication that chains have not mixed well. We also check the effective sample size (ESS) returned by Stan. This is an estimate of how many independent draws contain the same amount of information as the dependent sample obtained by the MCMC algorithm. In particular, Stan outputs Bulk-ESS and Tail-ESS. The first refers to the effective sample size based on the rank normalized draws, and hence gives an estimate of the sampling efficiency for the location of the distribution (e.g. mean and median). Tail-ESS computes the minimum of the effective sample sizes

```
transformed data {
  vector[N] mu;
  matrix[N, N] Dist_spatial;
  matrix[N, N] Sigma_lin;
   ...
  // linear covariance
   Sigma_lin[i, j] = 10 * dot_product(x[i],x[j]);   // off-diagonal elements
   Sigma_lin[k, k] = 10 * dot_product(x[k],x[k]) + 1e-6;  // diagonal elements (add also some jitter)
  ...
}
parameters {
  real<lower=0> l;
  real<lower=0> s2_matern;
  real<lower=0> r;
  vector[N] z;
}
transformed parameters {
  matrix[N, N] Sigma;
  matrix[N, N] L;

  // Exponential cov. from random effect + cov. from linear effect
  Sigma = s2_matern*exp(-inv_l*Dist_spatial ) + Sigma_lin;

  // Cholesky decomposition of Sigma
  L = cholesky_decompose(Sigma);
}
model {
  vector[N] ff;

  // weakly informative prior for hyperparameters
  s2_matern ~ p();
  l ~ p();
  r ~ p();

  z ~ normal(0, 1);
  ff = L*z;

  for (n in 1:N)
    y[n] ~ neg_binomial(V[n]*exp(ff[n]), r);
  }
generated quantities {
    vector[N] f;
    // derived quantity (transform)
    f = L*z;
}
```

Figure 4.1: Stan pseudocode of our SDM.

(ESS) of the 5% and 95% quantiles, giving an estimate of the sampling efficiency for the scale parameters (e.g. variance and tail quantiles).

In the preliminary phase of our studies we run the models for a subset of the data, setting MCMC with 1000 iterations, 400 burn-in and 3 chains. When considering the whole dataset, we run MC for 2000 iterations, 800 burn-in, and 3 chains.

4.5 Spatial predictions

In order to estimate how the different environmental covariates affect the whitefish larvae distribution we calculate the posterior distribution for the linear predictor parameters, β_j, and report the posterior mean and 95% credible interval for each of them. We recover such quantities by sampling from the latent function $f(\cdot)$ [3].

We first sample f and the other model parameters from their posterior, using Stan. We then recover Monte Carlo approximations for β by sampling from the conditional for $\beta|f$ in equation (4.4), with all the posterior samples of f and parameters. We obtain the poster for $\tilde{f}$ in the same way.

It is worth noticing that, when we make inference concerning the covariate effects, dependent error assumption and inclusion of random effects term may have significant effect to the inferential results. Assuming Gaussian priors for the linear weights $\beta \sim N(0, \Sigma_\beta)$, and for the spatially correlated random error $\phi(\mathbf{s})$, the posterior distribution for linear weights, conditional on the known error covariance is then

$$\beta|\mathbf{f}, \mathbf{X}, \sigma^2, \Sigma_\phi \sim N(\mu_p, \Sigma_p) \tag{4.4}$$

where

$$\mu_p = \Sigma_\beta \mathbf{X}^\top (\mathbf{X}\boldsymbol{\Sigma}_\beta \mathbf{X}^\top + \sigma^{\mathbf{2}}\mathbf{I})^{-\mathbf{1}}\mathbf{f}$$
$$\Sigma_p = \Sigma_\beta - \Sigma_\beta \mathbf{X}^\top (\mathbf{X}\boldsymbol{\Sigma}_\beta \mathbf{X}^\top + \sigma^{\mathbf{2}}\mathbf{I})^{-\mathbf{1}}\mathbf{X}\boldsymbol{\Sigma}_\beta$$

We finally report the spatial predictions for the whitefish larvae density on the whole GoB area.

The posterior predictive density of latent variables, conditional on hyperparameters, is

$$\tilde{\mathbf{f}}|\mathbf{S}, \mathbf{X}(\mathbf{S}), \mathbf{f}, \tilde{\mathbf{S}}, \tilde{\mathbf{X}}(\tilde{\mathbf{S}}), \theta \sim \mathbf{N}(\mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}}(\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}}+\sigma^{\mathbf{2}}\mathbf{I})^{-\mathbf{1}}\mathbf{f}, \mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}}-\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}}(\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}}\mathbf{f}+\sigma^{\mathbf{2}}\mathbf{I})^{-\mathbf{1}}\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}})$$

where

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

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

[3] We can simulate from a Gaussian process with mean function $\mu(s)$ and covariance function $k(s, s_0)$ at locations $S = [s_1, \ldots, s_n]^\top$ as follows. Construct a vector $\mu = [\mu(\mathbf{s_1}), \ldots, \mu(\mathbf{s_n})]^\top$ and a covariance matrix $[K_{f,f}]_{i,j} = k(\mathbf{s_i}, \mathbf{s_j})$. Form a Cholesky decomposition of the covariance matrix $LL^\top$. Form an $n \times 1$ vector of i.i.d. zero mean and unit variance Gaussian random variables, $z \sim N(0, I)$. After this form a vector $f = \mu + Lz$. The vector f is then a sample from the Gaussian process at locations S.

and

$$\mathbf{K_{f,f}} = \mathbf{X(S)\Sigma_\beta X(S)} + \mathbf{K}_{\phi,\phi}$$

One of the most common methods to estimate the posterior predictive density for the environmental covariates $p(\tilde{\mathbf{X}}(\tilde{\mathbf{S}})|\mathbf{S}, \mathbf{X}(\mathbf{S}), \tilde{\mathbf{S}})$ consists in building another spatial model for the environmental covariates and use this model to first predict $\tilde{\mathbf{X}}(\tilde{\mathbf{S}})$ and after use some point estimate for the environmental covariates, such as posterior predictive mean $\hat{\tilde{\mathbf{X}}} = \mathbf{E}[\tilde{\mathbf{X}}(\tilde{\mathbf{S}})|\mathbf{S}, \mathbf{X}(\mathbf{S}), \tilde{\mathbf{S}}]$ and approximate

$$p(\tilde{\mathbf{f}}|\mathbf{S}, \mathbf{X}(\mathbf{S}), \mathbf{f}, \tilde{\mathbf{S}}, \theta) \approx \mathbf{p}(\tilde{\mathbf{f}}|\mathbf{S}, \mathbf{X}(\mathbf{S}), \mathbf{f}, \tilde{\mathbf{S}}, \tilde{\mathbf{X}})$$

In fact, almost all predictions with environmental covariates correspond to approximation of this type, even those where the environmental covariates are estimated for all locations in the study domain.

We calculate the posterior predictive mean and variance of the log density: $\tilde{f}(\mathbf{x}, \mathbf{s})$ as well as the posterior median of the density of larvae throughout the study region, using two bigger prediction raster datasets, to get the covariates values at new locations $\tilde{\mathbf{S}}$, for current and future scenarios.

Chapter 5

Posterior analyses for HSDM

5.1 Conduct of the analysis

After the preliminary phase of data preparation, our dataset contains 634 datapoints, 9 environmental (model based) covariates and the two different target variables (measured). Due to time reasons, in the first stages of the analysis, we work on a subset of the data. We select our training data by taking one every three observations, for a total of 211 datapoints. We remark, how the resulting training dataset is quite well distributed in the space, covering the whole GoB. Also, the distributions of the training variables reflect quite closely the ones from the full dataset. Other possible subsets of the data were also considered: selecting observations at random, excluding/including specific clusters of observations, corresponding to sample sites concentrated in the same area. We also try to apply agglomerative clustering techniques (k-mean and hierarchical clustering), to cluster together observations very close to each others and then construct the training data so that it contains one observation for each cluster. It turns out that the way data are distributed in the space affects quite heavily the results of the model.

We start by considering as target whitefishes abundances. We then repeat the same computations for vendace. For each of the target we consider four models: two linear models and two models with additional quadratic effect.

In ecology, it is, indeed, common practice to consider a second order polynomial for continuous environmental covariates. This is related to the niche concept, which defines the role that a species plays in its ecosystem, and is determined by biotic and abiotic factors. In SDMs we represent the niche through the environmental covariates. It is generally assumed that

each species is characterized by a specific niche, defined by some optimal values of the environmental covariates, which are further supposed to vary on a limited range, since the species niche is restricted to some specific environment. These are the main reasons why linear models might result unrealistic. One of the simplest way to acknowledge for these assumptions is, indeed, the use of quadratic effects for all the continuous covariates. We further remark that the ecological niche is influenced, not only by environmental factors, but also by species interactions, that are not captured by the simple addiction of quadratic effects.

For both, linear and quadratic models, we use exponential covariance function to model the random effect covariance and the two different prior settings for the length scale parameter, l, presented in the previous chapter: $l \sim$ Student $- t_4(0,1)$ and $l \sim$ Gamma(1,10) . We run the models for 100 iterations, 400 burn-in, and three chains.

We recover the linear coefficients estimates and spatial predictions for the remaining two thirds of the data (423 datapoints), as explained in the previous chapter. After fitting the four models, we select, for each of the two species, the one that gives better results in terms of predictive power, evaluated on the test set.

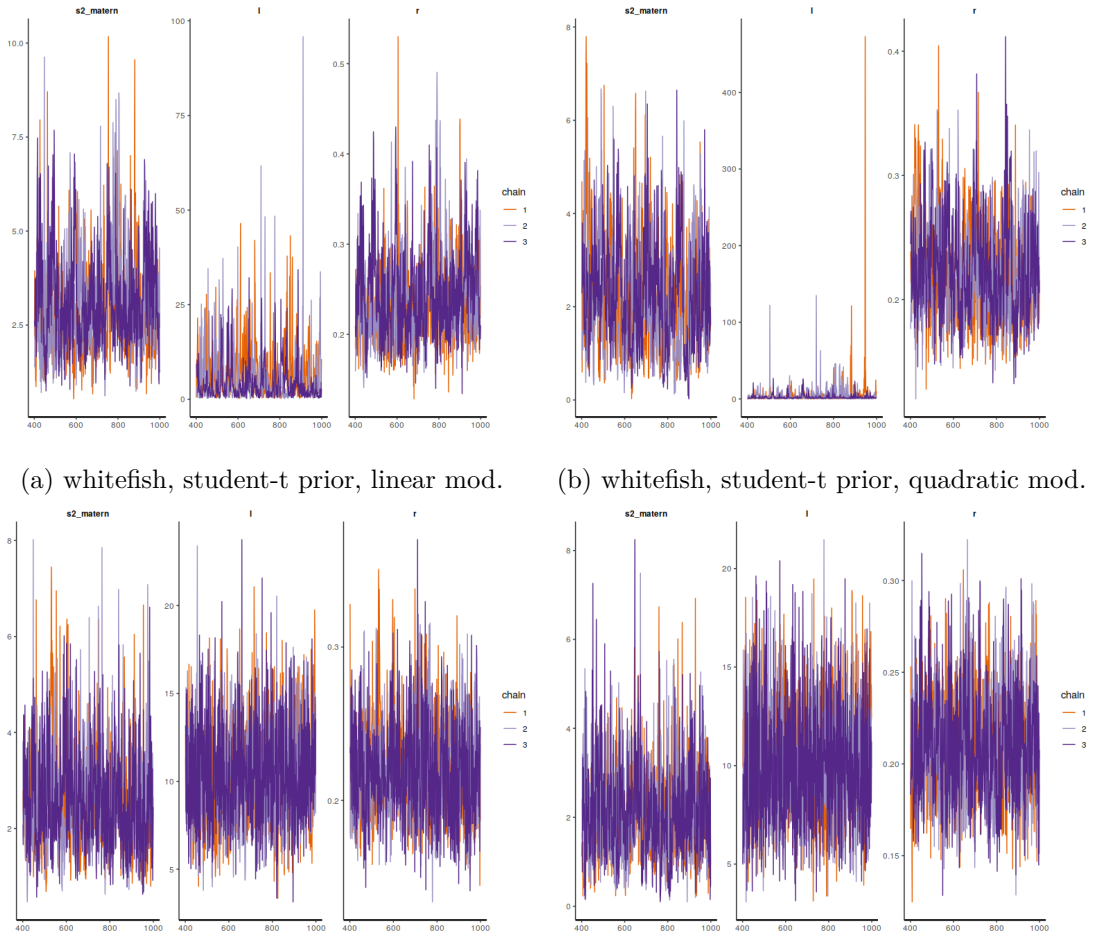
Finally, we fit the selected model with the full dataset (634 observations) and do predictions on the whole GoB region (mapped by about 180000 new locations), for both current and future values of the environmental covariates.

5.2 Convergence checking and sensitivity analysis

In assessing the convergence of the models, we use as threshold, $Rhat < 1.1$.

As visible from the trace-plots, in Figure 5.1, the length-scale l, the magnitude σ^2 and overdispersion r parameters tend to mix quite well, and do not show convergence problems for whitefish species. When assuming a Gamma(10,1) prior for l, which favors larger length scale, the autocorrelation of the posterior parameters decreases quite fast while, when assuming a Student-t prior for the inverse of l, the sample autocorellation decreases more slowly. In general we can state that the model is quite robust with respect to the prior selection of l: in all cases, its posterior prefers values around 10km on average. While setting a gamma prior leads to posterior range in $(2km, 25km)$, the use of the Student-t leads to a more skewed distribution, with some extremely high values, in the upperquantiles of the

distribution. The magnitude of the random effects and the overdispersion are both contained, and similar in all four models. The posterior of σ^2 has mean around 2 and range (0,8), r takes values in (0,0.5) with mean around 0.2. Such smaller values for the overdispersion parameter suggest the presence of high variability in the data and support our choice of using the Negative-Binomial distribution, to model the target, rather than a Poisson.



(a) whitefish, student-t prior, linear mod.

(b) whitefish, student-t prior, quadratic mod.

(c) whitefish, gamma prior, linear mod.

(d) whitefish, gamma prior, quadratic mod.

Figure 5.1: Whitefish trace plots for the four models considered.

In the case of vendace species, models reveal to be quite more sensible to the choice of the prior for the length scale parameter.

When assuming a Gamma prior for l, and hence an, a priori, preference for quite big correlation ranges, chains mix quite well and reach convergence.

Even though autocorrelation is a bit higher than the one observed for whitefish case, posterior distributions are quite concentrated around their mode, and in the same ranges obtained from the whitefish models. When, instead we set as l prior a distribution which favors much shorter length-scale, we can observe clear problems in chain mixing. As visible in the trace-plot (5.2) from the linear model, such prior force l to very small values, even smaller than our sampling resolution, so that the spatial random effect fails to account for the correlation among different sampling sites and focuses on modeling overdispersion, as an iid random effect, with high magnitude. As a result, the overdispersion parameter increases as well, to mean values around 20, so that the target distribution gets closer to the Poisson, meaning there is no need of any additional random effect, as such randomness is modeled by our spatial random effect. This confusion between l and r, and the different roles of the two random effects present in the model, lead to not uniquely identifiable results. To overcome the issue, one can set a lower bound constrain for l sampling values.

5.3 Models comparison in Bayesian statistics

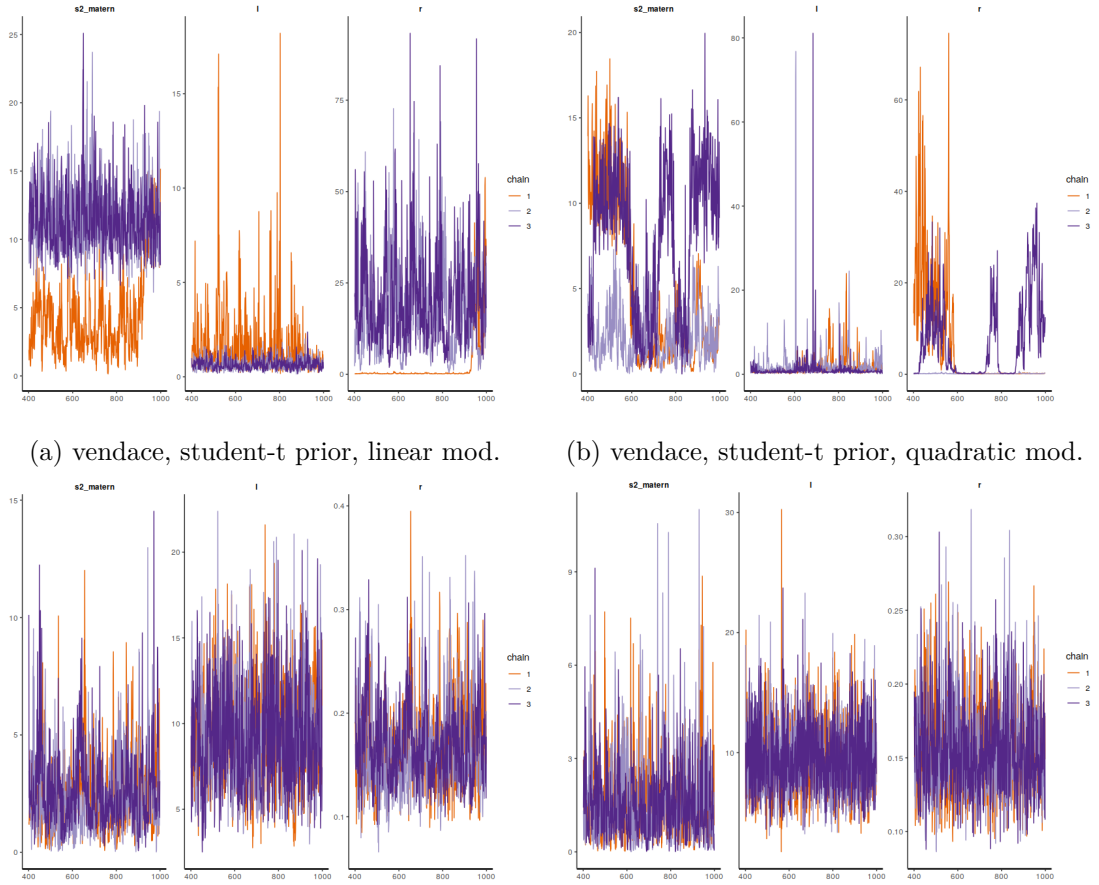
Model is the key element to any statistical inference, prediction and decision making, therefore models comparison and selection are crucial topics in statistics and machine learning. The underlying motivation for comparing models and/or selecting a model from alternatives are highly context dependent, so, when selecting the utility function to maximize in order to find the model that best fits certain objectives, the optimal strategy may vary case by case. Since there is no generally accepted "universal" method for model comparison and selection, many different methods have been developed through the years.

In Bayesian statistics, a typical approach consists in comparing models posterior probabilities $p(M|D)$. Pairwise comparison of two models, M_1 and M_2, is often done through the *posterior odds ratio*

$$\frac{\Pr(M_2|D)}{\Pr(M_1|D)} = \frac{\Pr(D|M_2)\Pr(M_2)}{\Pr(D|M_1)\Pr(M_1)}. \tag{5.1}$$

This is the prior odds ratio $\frac{\Pr(M_2)}{\Pr(M_1)}$ multiplied by the likelihood ratio, which is also called *Bayes factor*

$$B_{21} = \frac{\Pr(D|M_2)}{\Pr(D|M_1)}. \tag{5.2}$$



(a) vendace, student-t prior, linear mod.

(b) vendace, student-t prior, quadratic mod.

(c) vendace, gamma prior, linear mod.

(d) vendace, gamma prior, quadratic mod.

Figure 5.2: Vendace trace plots for the four models considered.

If the prior for models is assumed to be uniform, models comparison reduces to computing Bayes Factors. Such method works well when we condition the model comparison and selection strictly to the set of candidate models and we assume that one of the candidate models is the true model. This is also know as M-closed view (Bernardo and Smith (2000) and Vehtari (2012)).

When, instead, we do not consider that any of the candidate models would be the true model and do not aim to construct reference model we use a M-open view. In this perspective the most common approaches for model comparisons are cross validation and information criteria. They were originally developed from non-Bayesian perspective but one can also give them a Bayesian interpretation.

The leading idea in cross-validation is that a useful model or scientific

theory should produce useful predictions when applied to situations not seen yet. Hence, the goodness of our models should be assessed with external data that has not been used for training the model parameters. The simplest way to simulate external testing is to divide the data in two: the training set will be used to calculate the posterior distribution of model parameters and the test data set will be used to assess models predictive performance. In k-fold cross-validation the data is divided into k disjoint groups and each group in turn is used for testing (Vehtari (2012),Gelman et al. (2013)). By averaging the log predictive density (lpd) computed for each observation, one gets the k-fold CV predictive accuracy, that is

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, D_{\backslash k(i)}) \approx \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{i=1}^{S} p(y_i|\mathbf{x}_i, \theta^{\mathbf{s}})\right)$$

where S is the sampling size, $\theta^s \sim p(\theta|D_{\backslash k(i)})$ and $D_{\backslash k(i)}$ denotes the data from where the block including the data point i is excluded.

5.4 Case study: model selection

We now compare the four SDMs obtained by fitting one third of the full data. Out of the three methods presented in the above section, we prefer the training-test set division. To assess model predictive power we compute log predictive density (lpd) of larvae abundance, at the test dataset:

$$\text{V} = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}} \log p(y_i|\mathbf{x}_i, D_{train})$$

For each species, we select as final model, the one which returns the maximum lpd mean value for the test set $D_{test} = \{y_1, \ldots, y_{n_{test}}, \mathbf{x_1}, \ldots, \mathbf{x_{n_{test}}}\}$. Results are shown in Table 5.1: the optimal model for whitefish is the linear model with Student-t prior for the inverse of length-scale parameter, while, for vendace, the linear model with length-scale's Gamma prior, is preferred. We can observe how the lpd does not vary much for whitefish, while for vendace, there is a clear preference for the Gamma prior, meaning that the prior for length-scale is not as important for whitefish as it is for vendace. In all cases, including quadratic effects in the model does not produce any improvement.

An advisable way to conduct model selection in our future studies, would be applying k-fold CV to the full data.

model	lpd.wf	lpd.ve
fit1.gamma	-3.2372	-1.93210
fit2.gamma	-3.2167	-1.94361
fit1.stud	-3.1958	-2.27660
fit2.stud	-3.1998	-2.18870

Table 5.1: Mean lpd for whitefish (lpd.wf) and vendace (lpd.ve) abundance in the four different models considered. In order: linear with l Gamma prior, quadratic with l Gamma prior, linear with $1/l$ student prior, quadratic with $1/l$ student prior.

5.5 Whitefish HSDM results

To assess the goodness of our final model, we first do predictions on the training set, as well as on the remaining test data. By plotting the predicted values of larvae densities against the corresponding measured value, we can observe how the model is generally able to predict the observed logdensity, with an error margin of about (-10,10). The model have a clear tendency of overestimating larvae abundance, in particular, when dealing with sampling sites where no larvae were found (for these locations we set $\mathbf{y}$ equal to 0.1, to avoid having infinite values for the observed logdensity), as shown in Figures 5.3 and 5.4. As confirmed by the small r posterior estimates, the observed vs predicted plots reveal a lot of overdispersion in the data. Our results, reflect the typical predicted vs. observed plots for Poisson target distribution: the variance of Poisson distribution increases with mean, producing a lot variation on the top-right corner of Figure 5.3. When dealing with log-transformed target, the likelihood function does not penalize the mean parameter of Poisson distribution for values $y = 0$ as strongly as for $y > 0$. As a result, the lower-left corner of log-transformed observed vs. predicted plot, usually shows high variability.

We then compute the realized-residuals $r_i = y_i - \tilde{y}_i$ for $i = 1, \ldots, N$ (Gelman et al. (2000)), where $\tilde{y}_i = \exp(\tilde{f}_i(\mathbf{x_i}, \mathbf{s_i}))$. If we visualize the residuals and their locations on the Gob (Figure 5.4) we can spot some spatial patterns in the map,: the model tends to underestimate the abundances in the southern areas (where we have smaller observed values, in general), while, in certain areas where we have many sampling sites, the model overestimates larvae amount.

In Figure (5.5), we represent the standardized residuals with respect to each of the continuous covariates. Looking at these residual-plots we can observe how residuals are generally spread quite randomly, suggesting a good fitting for the model. The variable ICLAST09 may present some heteroschedasticity. The general lack of clear trends in the plots confirms the unnecessity of additional quadratic effects.

We then make predictions on the bigger raster set, and examine the posterior distributions of the model parameters. Table 5.2 and Figure 5.6 show posterior distribution for scale parameters: the model capture quite high overdispersion in the data, and returns a relatively high estimate of the length scale parameter, indicating the presence of spatial correlation among sites farther apart (measured by l) and of high variability in the

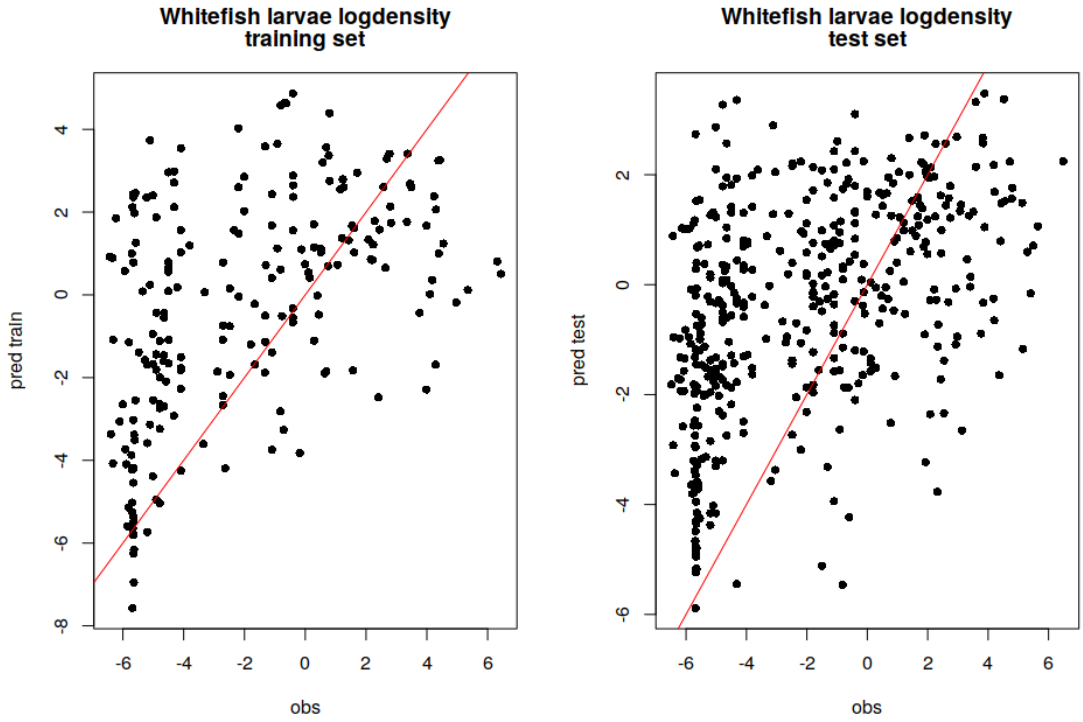


Figure 5.3: Predicted vs observed logdensities for training and test data

data (measured by r).

From Table 5.3 and Figures 5.7 and 5.8 we can see the linear coefficients posterior distribution. The absence of bottom coverage (BOTTOMCL0) and the average fetch (FE300M) are negatively associated with larvae logdensity. There are no other clear associations visible directly from the posteriors.

parameter	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
s2_matern	2.982	0.084	1.278	1.136	2.117	2.749	3.597	6.115	230.175	1.022
l	5.647	0.414	6.787	0.399	1.431	3.336	7.423	23.965	268.543	1.020
r	0.239	0.003	0.047	0.167	0.207	0.233	0.263	0.356	183.577	1.027

Table 5.2: Scale parameters posterior estimate in whitefish larvae logdensity model, with Student-t prior for $1/l$.

To asses how larvae logdensity is related to each of the fixed effects we report in Figure 5.9 the response function with respect to the environmental covariates within the range of the covariate values in the full prediction raster. The "x" symbols in the figures represent the covariate values in

Covariate	Mean	Stand.Dev	quant2.5%	quant97.5%
intercept	-0.282	3.222	-6.148	6.117
BOTTOMCLS_0	-2.600	0.677	-4.004	-1.117
BOTTOMCLS_1	-0.680	0.549	-1.536	0.426
BOTTOMCLS_2	0.854	0.725	-0.613	2.390
BOTTOMCLS_3	1.079	1.030	-0.642	3.109
BOTTOMCLS_4	0.708	0.413	-0.140	1.482
BOTTOMCLS_5	0.458	0.557	-0.850	1.366
FE300ME	-2.335	2.091	-8.420	0.009
DIS_SAND	-0.359	0.355	-0.998	0.291
ICELAST09	0.304	0.375	-0.421	1.038
RIVERS	0.166	0.348	-0.524	0.822
DIST20M	-0.155	0.394	-0.896	0.528
CHL_A	-0.438	0.594	-1.802	0.421
TEMP09M	0.644	1.153	-2.054	2.746
SALT09M	0.579	1.954	-1.700	6.279

Table 5.3: β coefficients estimates in whitefish larvae density model, Student-t prior for $1/l$.

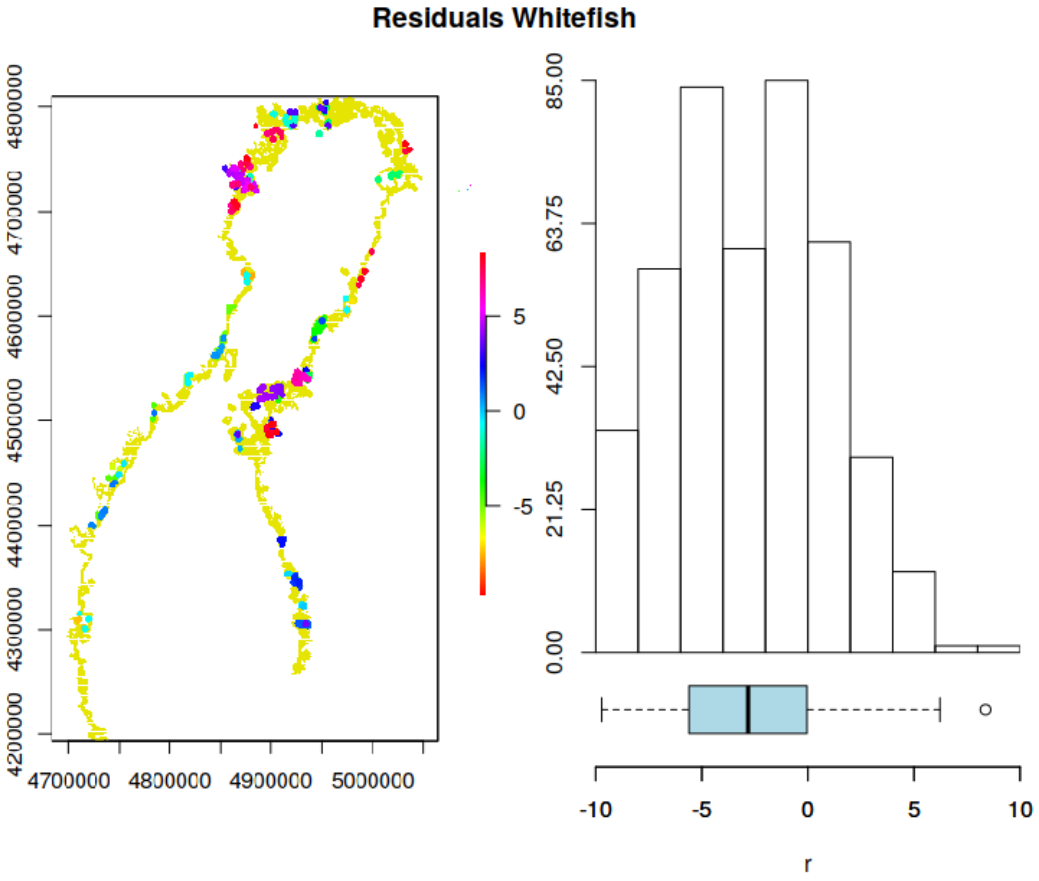


Figure 5.4: Residuals distribution, for whitefish

the training data, the continuous line indicates the mean logdensity of the continuous covariates, while the two dashed-lines its 0.05 and 0.95 order quantiles. From these figures we can conclude that whitefish larvae logdensity in the GoB, is positively associated to temperature, salinity, the number of weeks of ice coverage and the presence of rivers, while it is negatively associated with the others. The sandy/stone bottom type is the one that favors more whitefish larvae spawning.

Finally, we report the maps of larvae logdensity under current and future climate conditions in Figures 5.10 and 5.11. The larvae logdensity is evaluated for about 180000 new locations along the shallow waters of the GoB. The corresponding covariates values at each location are stored in the bigger rasterdatafile, built as described in Chapter 2. To recover our predictive maps, we first ran the model using as training data the full data

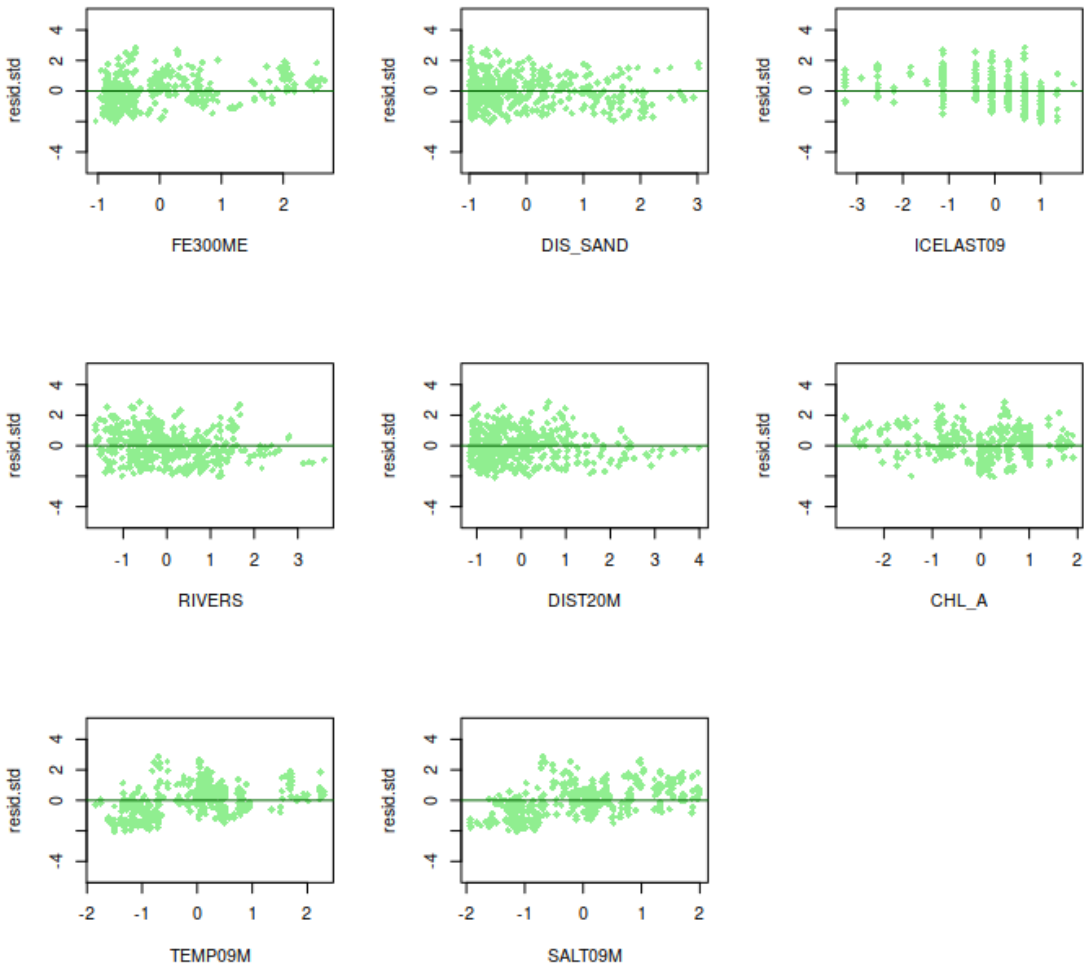


Figure 5.5: Residual plots, for whitefish

available (634 observations in total). We ran 1 chain, for 500 iterations and 150 burn in. The MCMC algorithm reaches convergence and the posteriors of the model parameters reflect the ones obtained in the previous steps.

From the final raster maps we can see how whitefish larvae logdensity is expected to decrease in the future.

We remark that, making predictions on the bigger raster dataset, using as training data only one third of the total data available, heavily affect the results: about 0.1% of the predicted log densities get unrealistically large values. If, instead we fit the model, using all the data, we do not encounter any issue of that kind. This is probably due to the fact that,

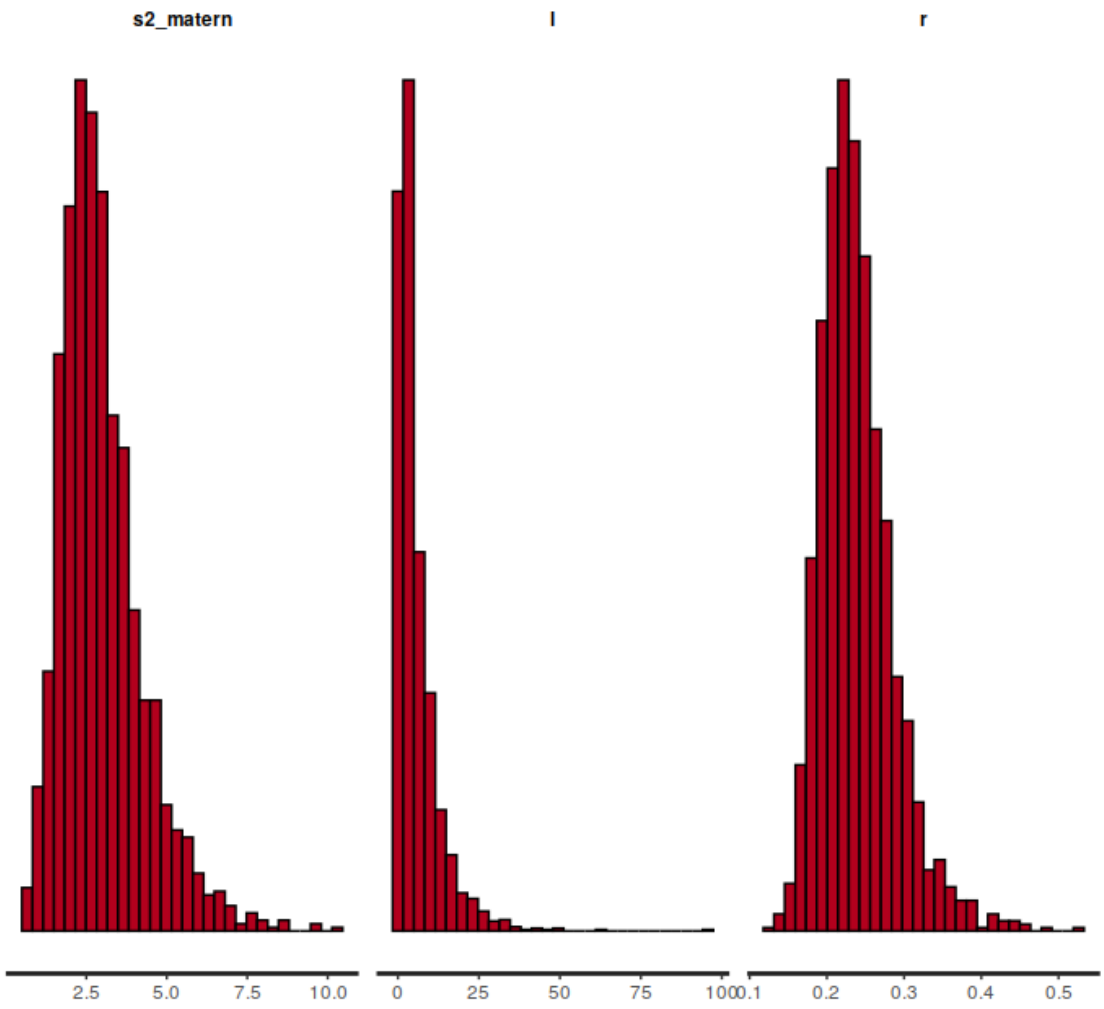


Figure 5.6: Posterior distribution of random effect parameters: σ^2, l and r, for whitefish model

when using such a small training dataset, the covariate range values at those 180000 prediction sites are so far out of the range of the covariate values in the training data. If, instead, we increase the training set, and hence its covariates variability, the model predictions work fine.

5.6 Vendace HSDM results

We analysis results from vendace species in a similar fashion, as done in the previous section for whitefish.

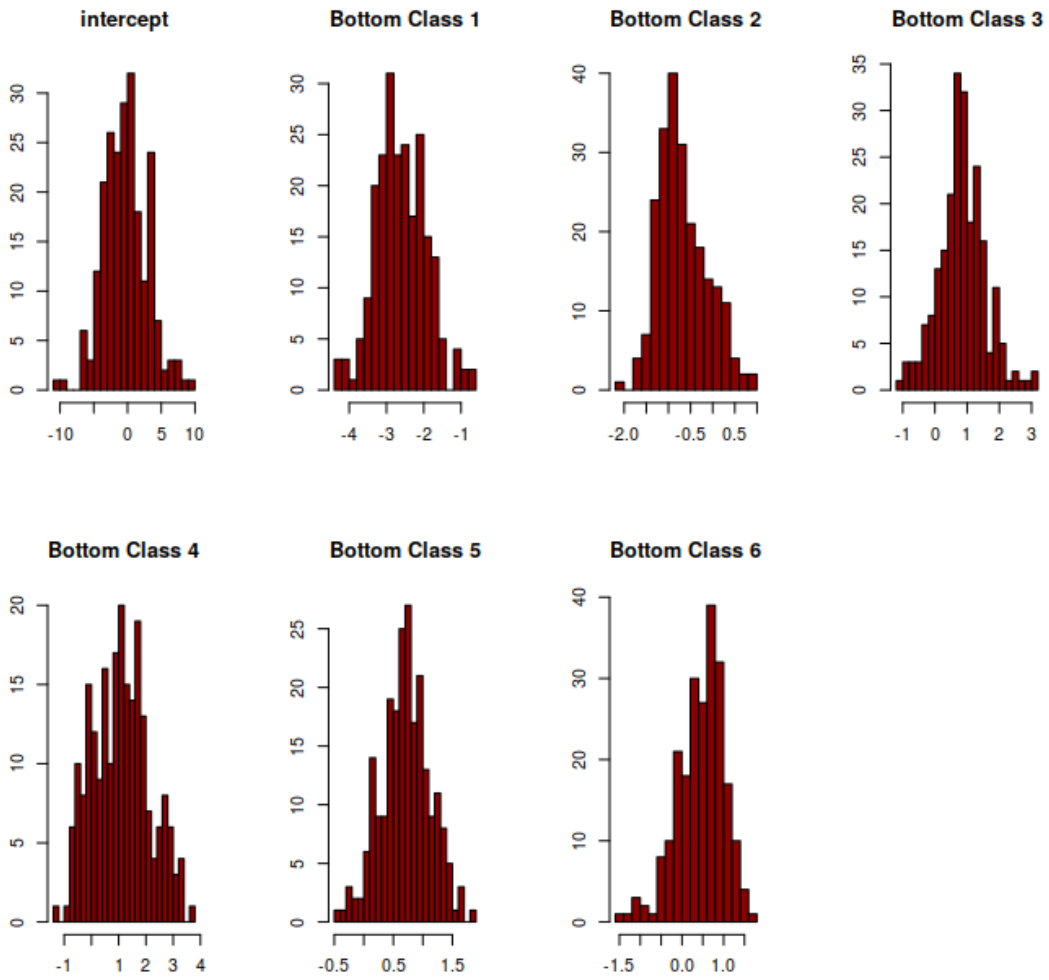


Figure 5.7: Posterior distribution of linear parameters, for whitefish model

By plotting the predicted values of larvae densities against the corresponding measured value, we can observe how the model is generally able to predict the observed logdensity, a bit better than in the previous case. Indeed, even thought the residual range increases to about (-20,5), the high variability in the data, is mainly produced by the negative values of the logdensity, so that the model seems to underestimates vendace abundances, on average, as shown in Figures 5.12 and 5.13.

Again, predictions on the test set do not seem much worse than the ones on the test set, if we look at the residual plots, we can observe how residuals are generally spread quite randomly, suggesting a good fitting for the model.

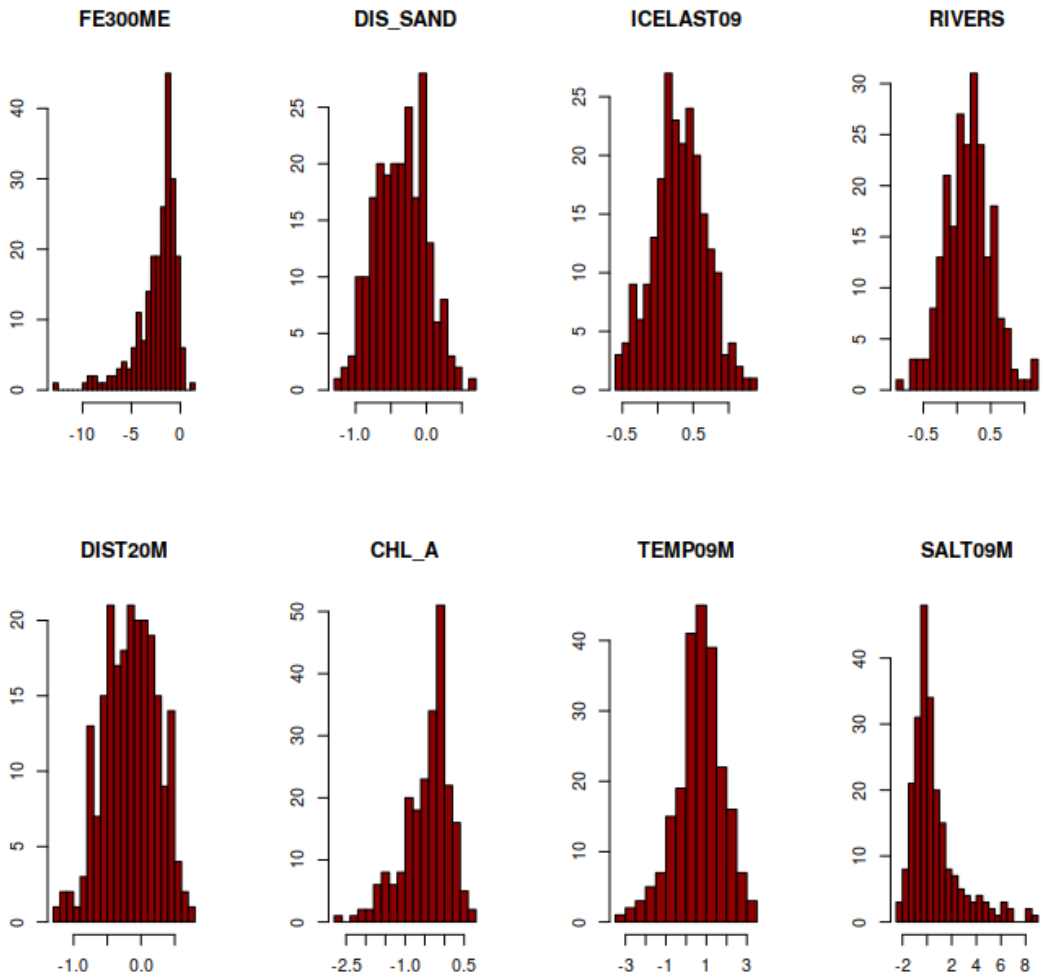


Figure 5.8: Posterior distribution of linear parameters (continuous variable), for whitefish model

The lack of clear trends in the plots confirms the unnecessity of additional quadratic effects.

Examining the posterior distributions of the model parameters, we assess the presence of high variability in the data and an even wider estimate for the correlation range (Figure 5.15 and Table 5.5). From Table 5.4 and Figures 5.16 and 5.17 we can see the linear coefficients distributions.

Looking at the response curves, in Figure 5.18, we can conclude that vendace larvae logdensity in the GoB, is positively associated to temperature, the number of weeks of ice coverage, the distance from sand and the influence of rivers, while it is negative associated with the others. The first

Covariate	Mean	Stand.Dev	quant2.5%	quant97.5%
intercept	-4.142	2.908	-9.470	1.333
BOTTOMCLS_0	-3.659	0.933	-5.784	-1.944
BOTTOMCLS_1	-2.397	0.823	-3.984	-1.004
BOTTOMCLS_2	0.077	0.686	-1.249	1.369
BOTTOMCLS_3	1.129	1.780	-2.105	4.385
BOTTOMCLS_4	-0.022	0.496	-0.961	1.112
BOTTOMCLS_5	0.744	0.569	-0.328	1.862
FE300ME	-4.015	1.772	-7.862	-1.277
DIS_SAND	0.081	0.604	-1.094	1.210
ICELAST09	0.932	0.521	0.004	1.960
RIVERS	1.522	0.485	0.648	2.514
DIST20M	-0.677	0.533	-1.673	0.273
CHL_A	-0.885	0.585	-2.281	0.139
TEMP09M	2.266	1.421	-0.228	5.385
SALT09M	-2.447	1.896	-5.791	1.461

Table 5.4: β coefficients estimate in vendace larvae logdensity model, with gamma prior for l.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
s2_matern	2.446	0.087	1.654	0.282	1.317	2.098	3.153	6.693	364.769	1.011
l	9.145	0.101	3.103	4.197	6.835	8.768	11.040	15.834	949.511	0.999
r	0.167	0.002	0.042	0.103	0.138	0.161	0.188	0.274	378.380	1.010

Table 5.5: scale parameters estimates in vendace larvae density model, gamma prior for l.

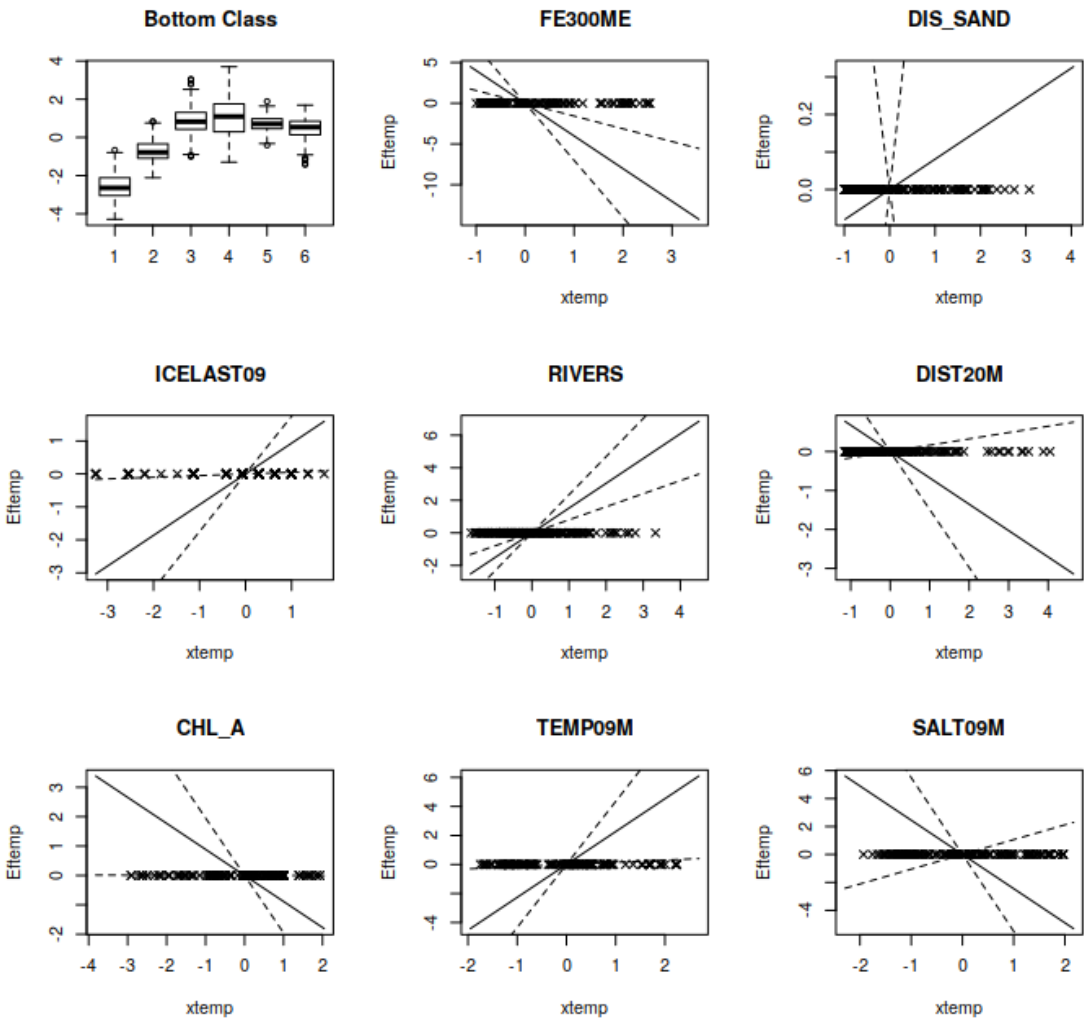


Figure 5.9: Whitefish response curve

two classes of bottom coverage (absence of coverage and open waters) do not favor vendace spawning, which prefer, instead, a sandy/stone bottom type.

As before, to recover our final raster maps, we first ran the model with all the available data, setting 1 chain, for 500 iterations and 150 burn-in. In Figures 5.19 and 5.20 we report the maps of larvae log density under current and future climate conditions. The MC algorithm reaches convergence and the posterior of the model parameters reflect the ones obtained in the previous steps.

From the final raster maps we can see how vendace larvae logdensity is

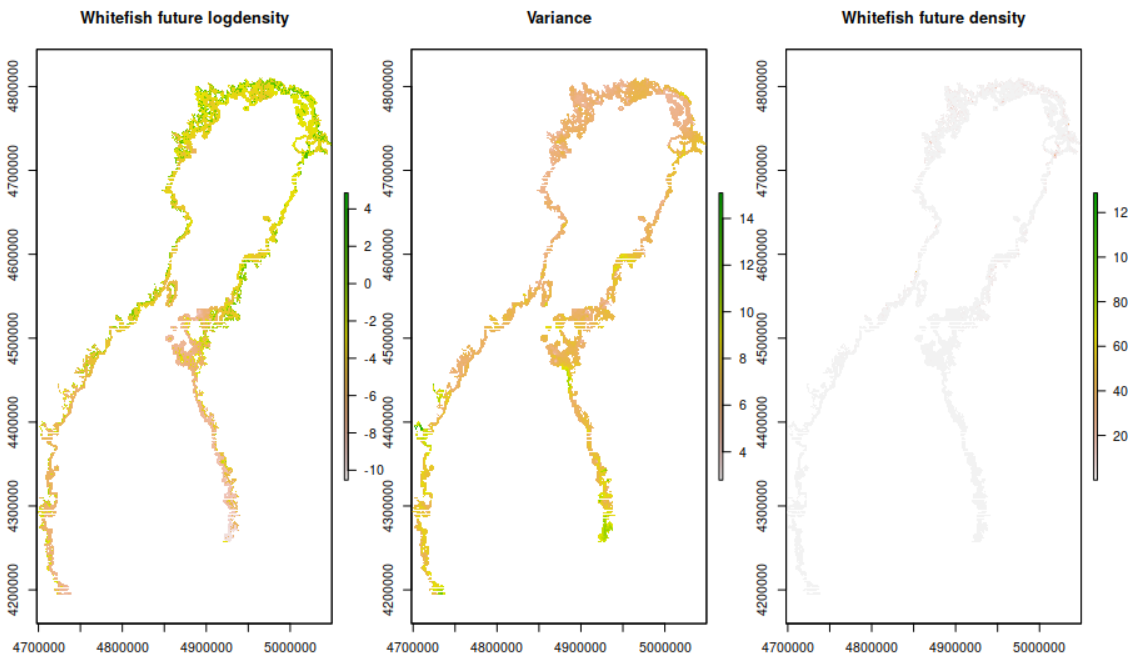


Figure 5.10: Whitefish larvae logdensity, future scenario

expected to decrease in the future.

5.7 Software

In conducting the analysis we used the software R (R Core Team (2019)) and Stan (Stan Development Team. (2018)). When preparing the data, dealing with spatially indexed data, we used the library ncdf4 (Pierce (2019)), to open and explore environmental variables obtained from the SmartSea project, and the libraries raster (Hijmans (2020)), GIStools (Brunsdon and Chen (2014)) and rgdal (Bivand et al. (2019)), to rasterize the data, rescale, project and visualize them, by mean of raster maps, as visible e.g. in Figure 2.1.

The hierarchical Bayesian models are all implemented in Stan, and ran in R, through the library rstan. Stan model codes are partially based on previous scripts provided by Jarno Vanhatalo.

When computing predictions we use a covariance function, implemented

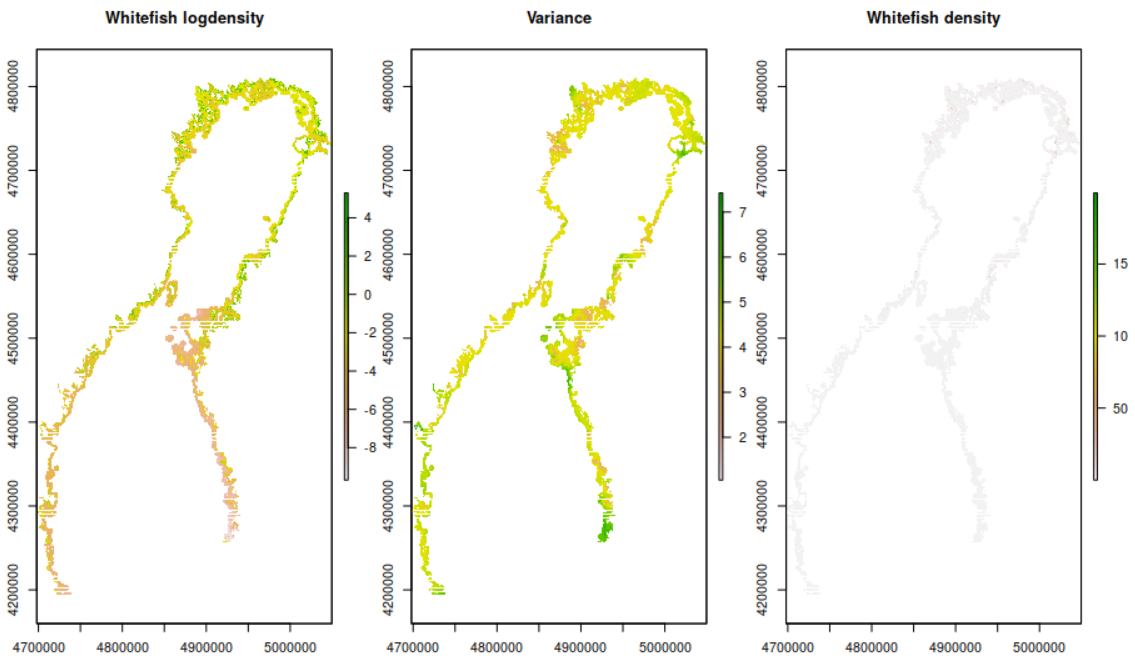


Figure 5.11: Whitefish larvae logdensity, current scenario

in C++, to speed up the loops iterations. The function was integrated in the R-code by using the library Rcpp (Eddelbuettel and Balamuta (2017)).

All the codes are available on my github repository[1]. Other than the main data-files recovered from the initial sources, one can find various Rmarkdown files, containing code, outputs and comments related to different stages of the work. The two main files are the following: 'nc-file.Rmd' (or nc-file.pdf) contains the codes and comments related to SmartSea data preparation and the merging of the two data sources while 'white-fishes.Rmd' (or white-fishes.pdf) contains the model implementations and the related analysis.

The last stages of the studies, and the results reported in the thesis can be found in the remaining markdown files. The SmartSeadata preparation file contains the function implemented to recover data related to temperature, salinity and ice coverage. The outputs of the data preparation file, are

[1]https://github.com/ilapia/thesis-pia

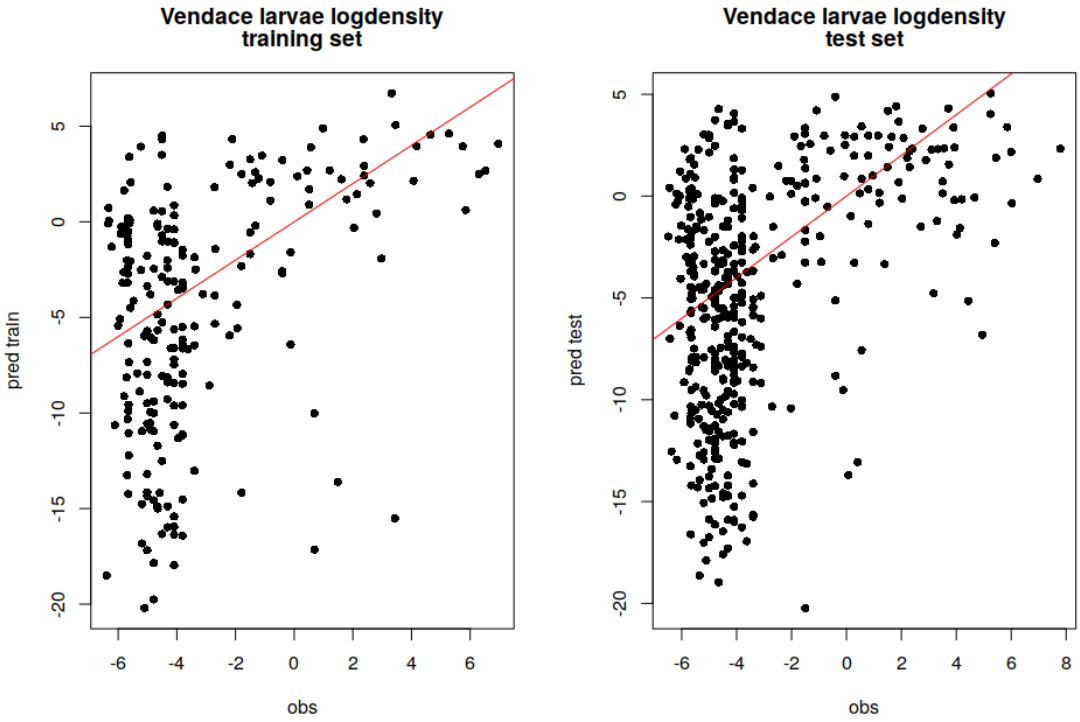


Figure 5.12: Predicted vs observed logdensities for training and test data

accessible in the 'data' folder, while the outputs of the models script are in the 'fits' folder. The file 'functions' contains all the function used in the other files. The output of the final models can be obtained from the 'whitefish-final.Rmd' file.

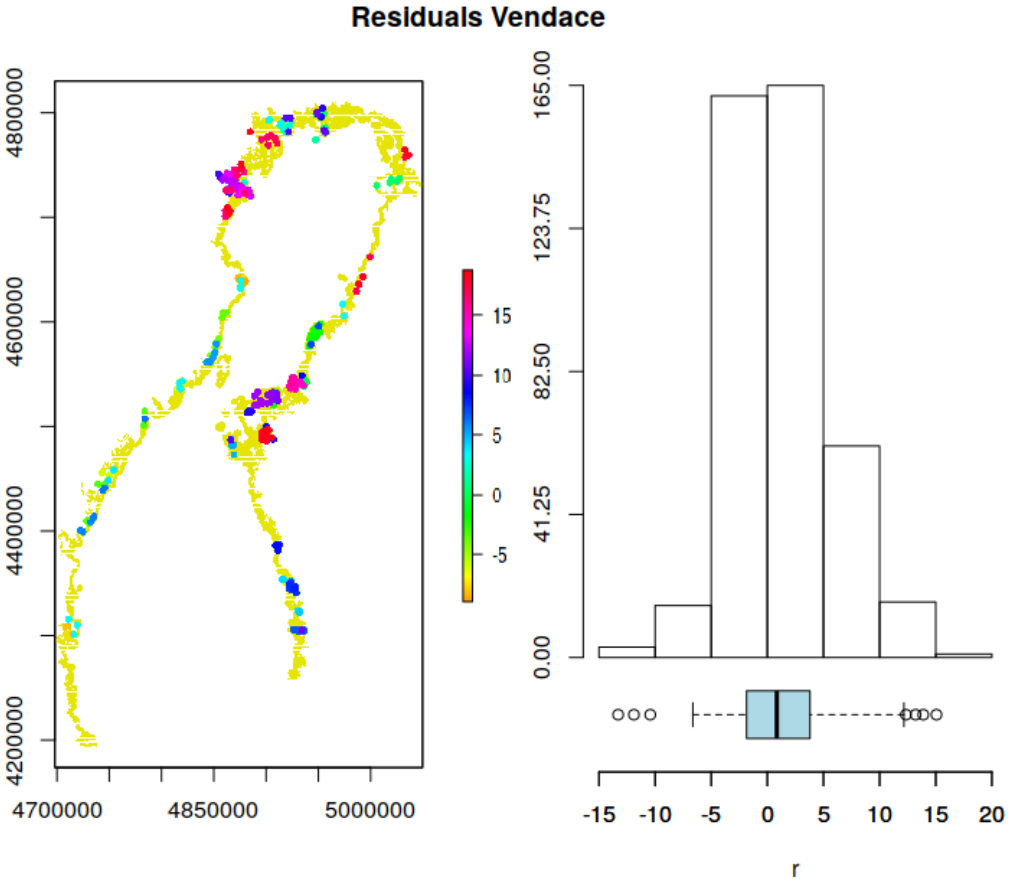


Figure 5.13: Residuals distribution

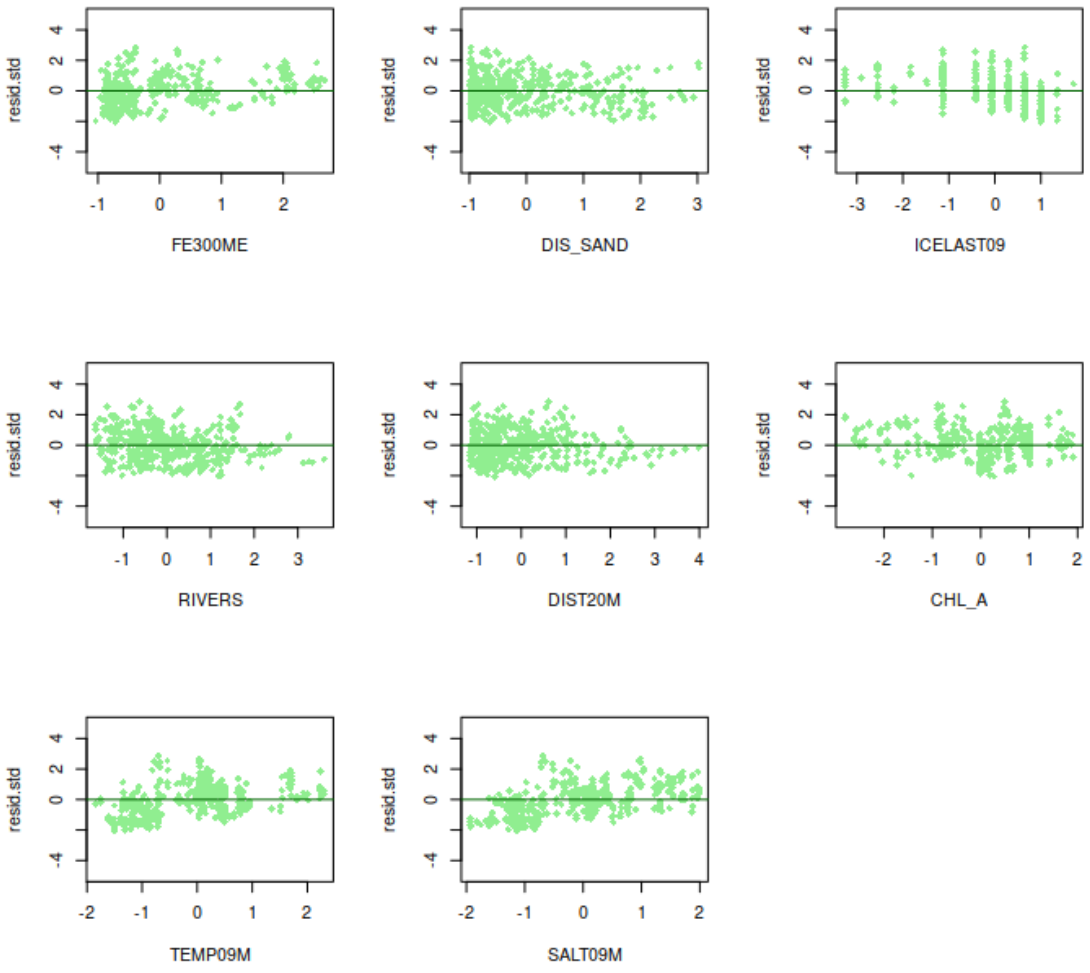


Figure 5.14: Residuals plot

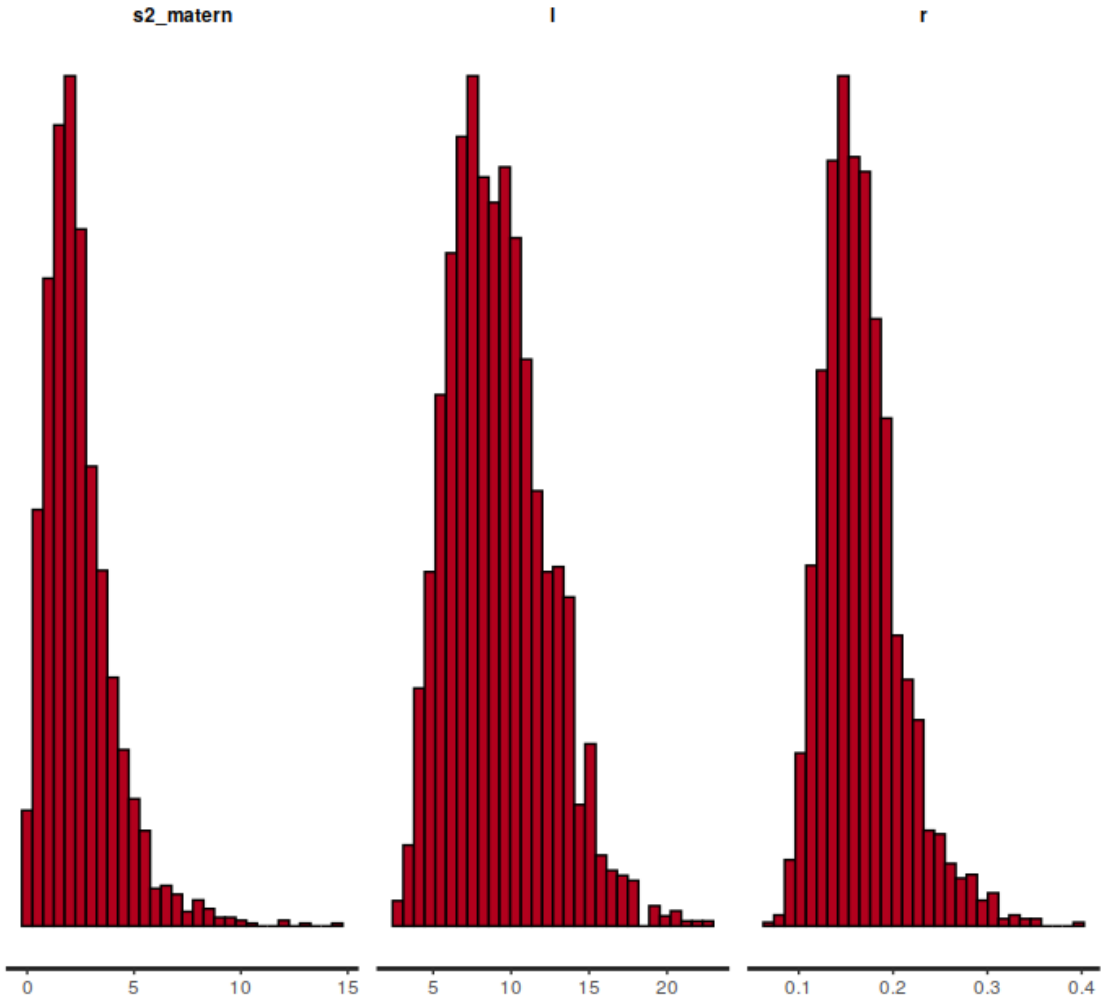


Figure 5.15: Posterior distribution of random effect parameters: σ^2, l and r, for vendace model

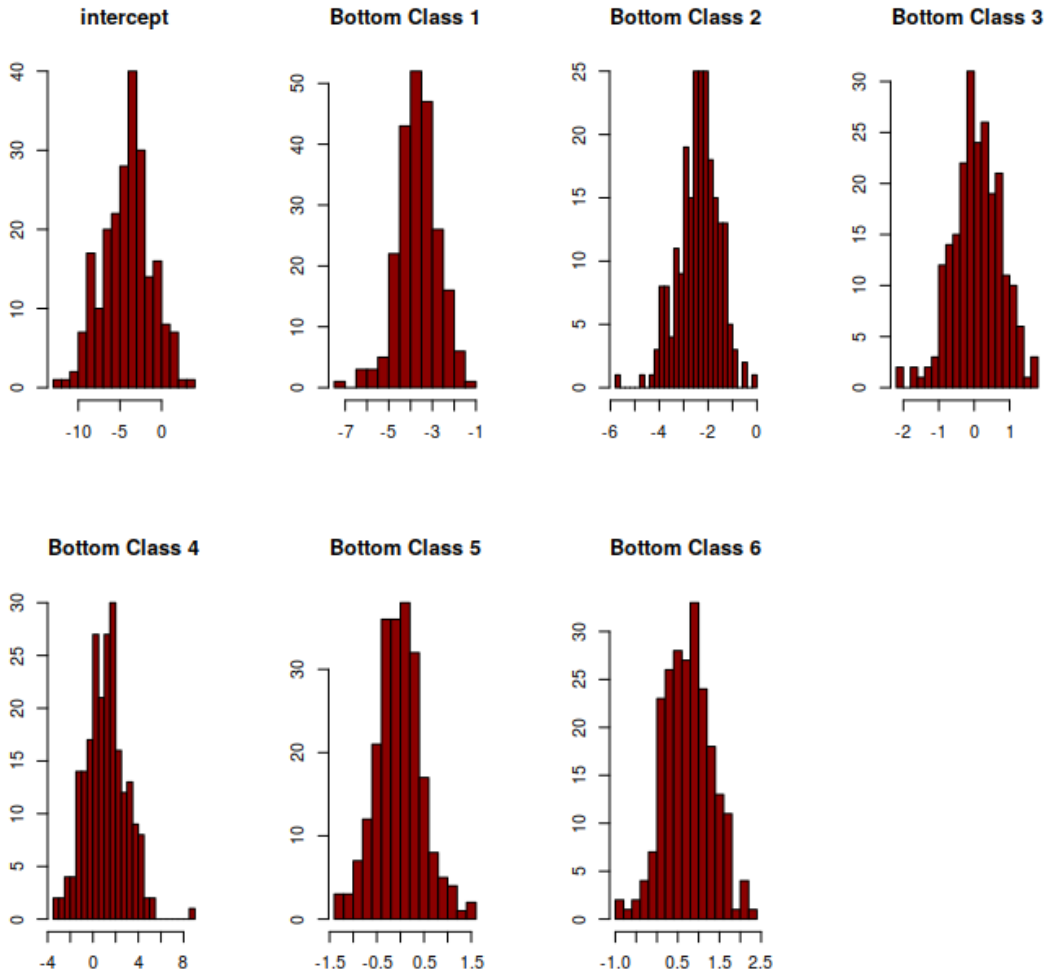


Figure 5.16: Posterior distribution of linear parameters, for vendace model

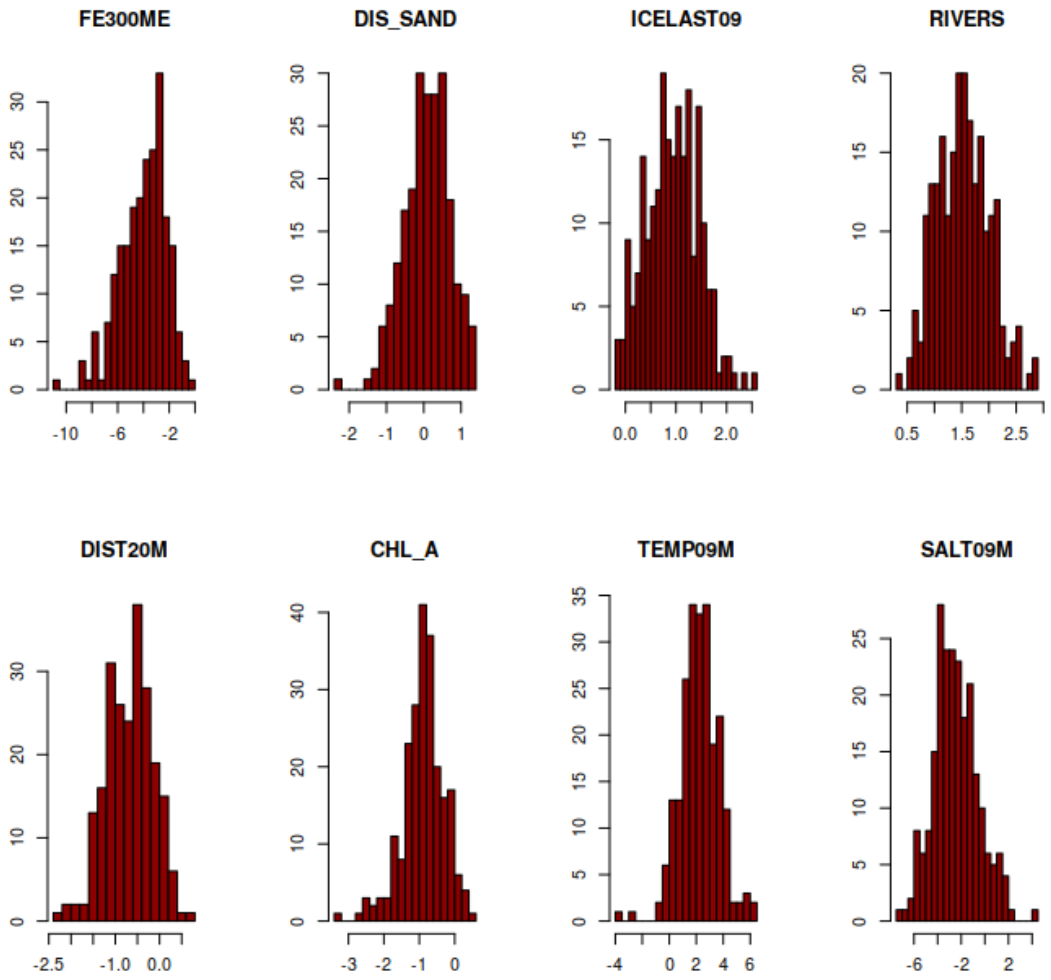


Figure 5.17: Posterior distribution of linear parameters (continuous variable), for vendace model

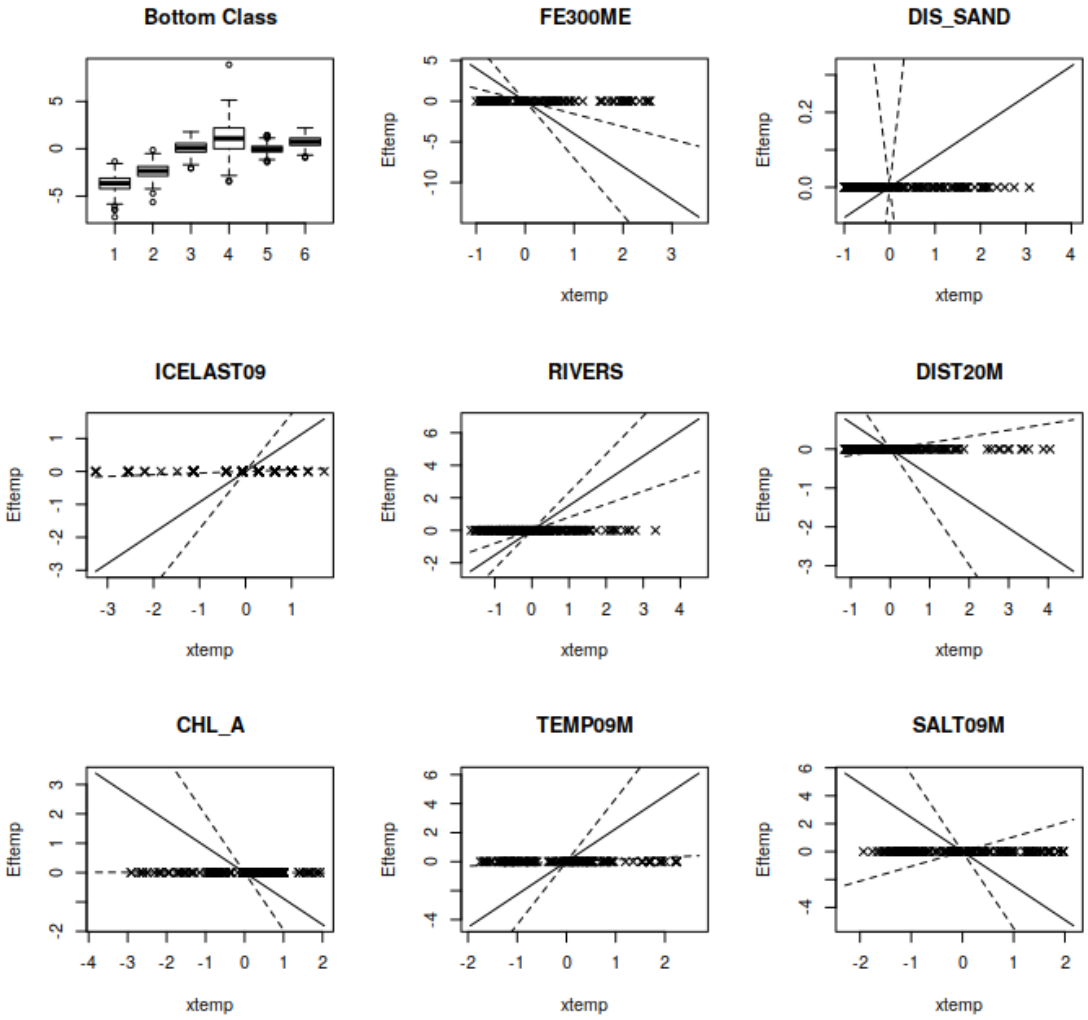


Figure 5.18: Vendace response curve

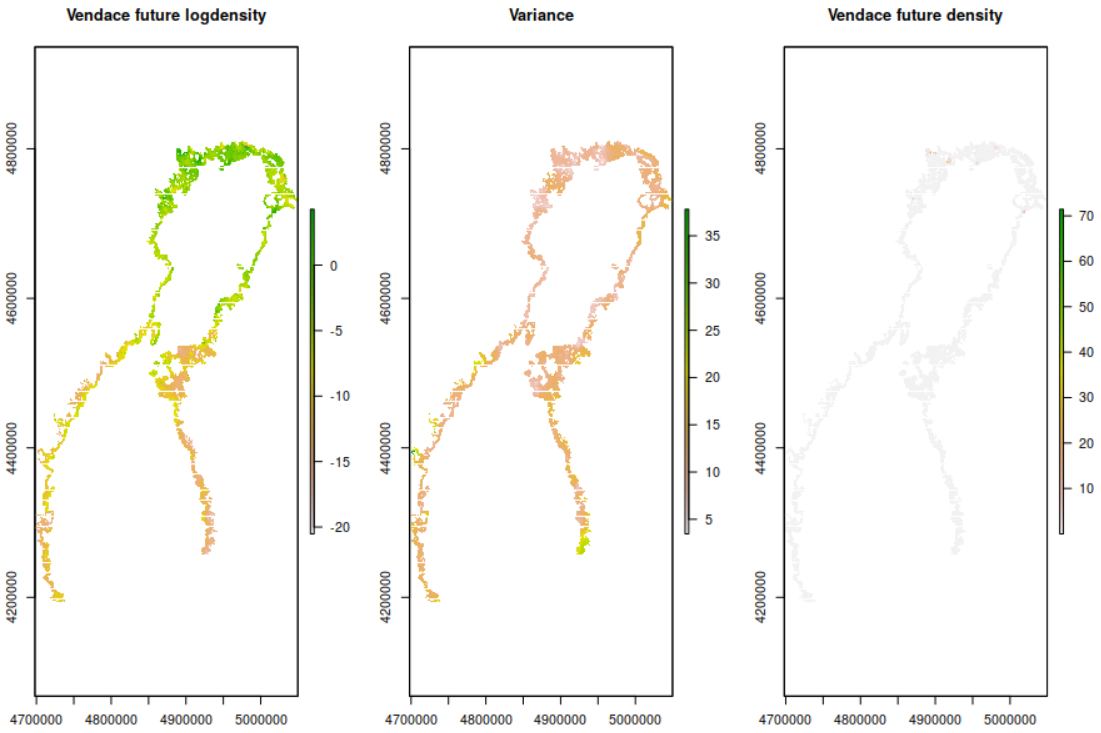


Figure 5.19: Whitefish larvae logdensity, future scenario

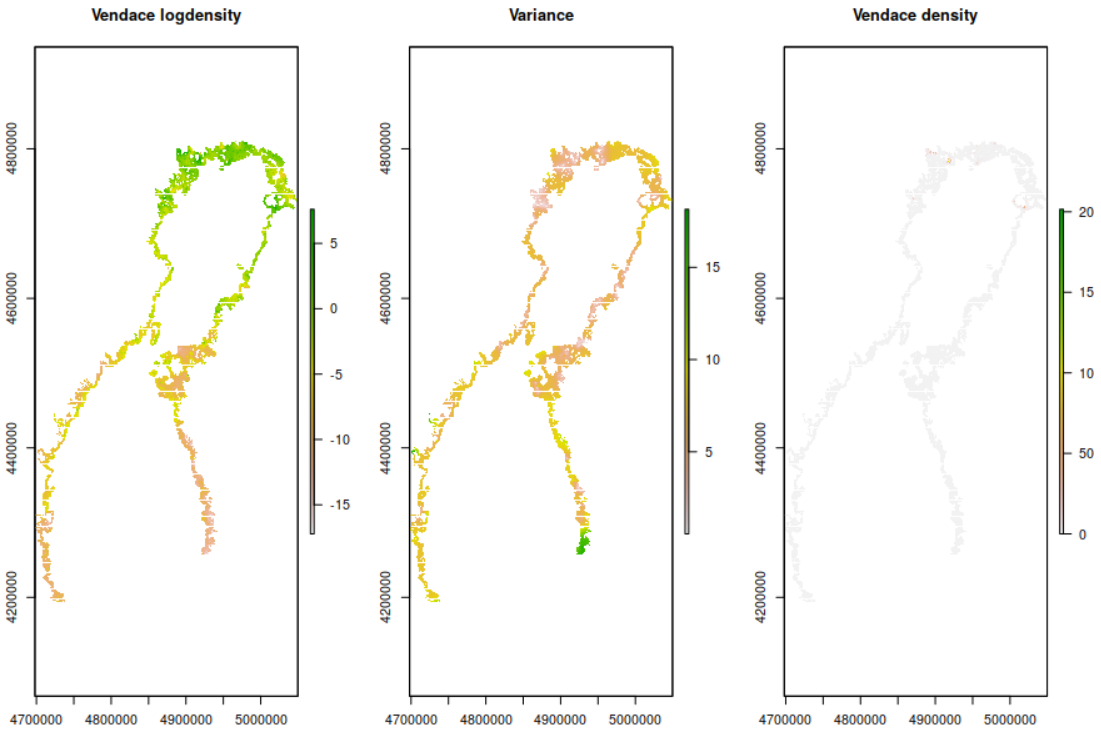


Figure 5.20: Vendace larvae logdensity, current scenario

Chapter 6

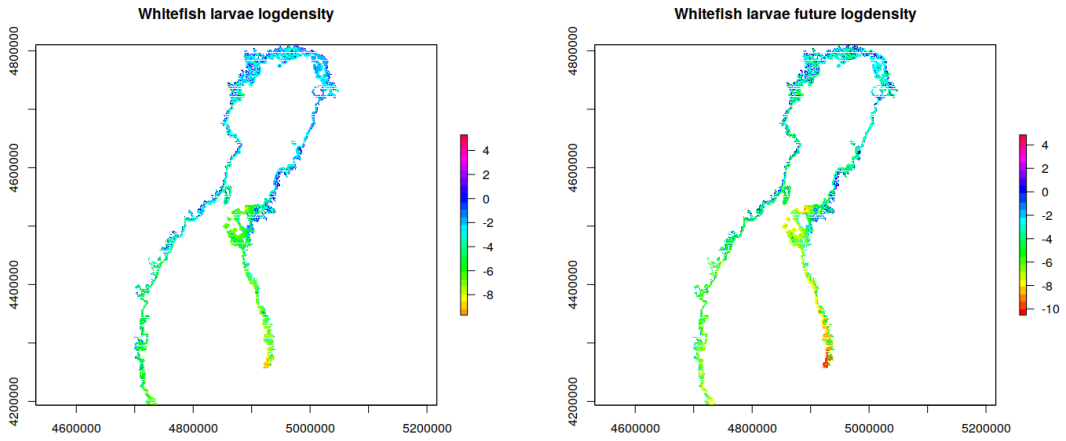
Conclusions and further extensions

6.1 Final results

One of the main objective of our study was to compile the SmartSea data with the existing whitefish data and check whether these two information sources can be used for analyzing the effects of climate change on whitefish and vendace. By mean of simple SDMs, we then try to assess how climate change will affect whitefish and vendace larvae distribution among the coastal area of the GoB.

The data has large overdispersion and, hence, the model predictions for test data contain large residuals. Nevertheless, the obtained results are generally in line with the one obtained by previous studies (Vanhatalo et al. (2012), Veneranta et al. (2013)) on the same species. Past and future scenarios can be compared through the raster maps in Figures 6.1 and 6.2. Both larvae abundances are expected to decrease in the next decades. Whitefish larvae reductions will be particularly evident in the southern and eastern areas of the GoB where their presence is already lower than in other regions of the gulf. Vendace larvae will be affected even more strongly by the future changes in climate conditions, with remarkable decrease of the logdensity in both its maximum and minimum values.

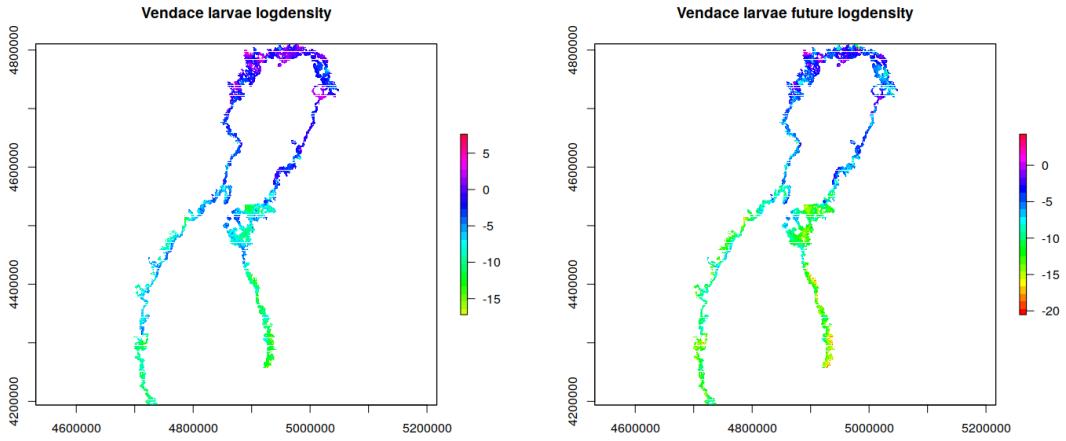
We can further observe how whitefish spawning is favored by high temperature and low salinity conditions. The near presence of rivers, and longer ice coverage of the waters also benefits their reproduction, while high rate of eutrophication (measured by the Chlorophyll -a index) and the exposure to wind (FETCH covariate) prevent it. Whitefishes spawns grow better in



(a) Measured whitefish larvae logdensity, year 2009

(b) Predicted whitefish larvae logdensity for years 2049-2059

Figure 6.1: Final raster layers obtained as outputs of our analysis, for current (a) and future (b) whitefish larvae logdensity in the GoB.



(a) Measured vendace larvae logdensity, year 2009

(b) Predicted vendace larvae logdensity for years 2049-2059

Figure 6.2: The figures show the final raster layers obtained as outputs of our analysis, for current (a) and future (b) vendace larvae logdensity, in the GoB.

shallow waters with sandy and muddy bottoms, rather than in not covered sites, located in more open waters. The distance from sand and from deep waters have indeed negative, even if less evident, impact.

Vendace spawning are affected by the environmental covariates in a similar fashion: sandy and muddy bottoms, in shallow waters, favor their reproduction together with high temperature and lower salinity conditions.

We remark that, the analyses presented in this work, are preliminary and aimed to be an initial investigation on the subject. However, the model seems to already capture the most essential aspects from the data, and hence, we believe that more careful analyses of the SmartSea and whitefish data can answer to the ultimate objectives: how will the spawning areas and larvae abundance in the Gulf of Bothnia change in the future?

6.2 Criticisms and extensions

The model presented in the study is very simplistic and, hence, we do not expect it to have very good predictive capability as such. Even if we can already get quite trustful conclusions from our predictions, the results obtained from our SDMs implementations and analysis reveal several criticisms.

When assessing the goodness of the model, on the test data, we observed how the SDMs produces quite high residuals, with the tendency of miss estimate larvae abundances, in particular for these sampling sites where there were no larvae sampled. The target distributions objective of our analysis are, indeed, very strongly skewed to the right, with mass concentrated around 0 (we have roughly 0 observed larvae in 50% of the sites for whitefishes and in 70% of the sites for vendace), while, in a few sites, a big amount of fishes was detected, determining quite long right tails in the target distributions. Nevertheless, results are in line with earlier analyses of this data. Some deviations from these previous studies could be related to the fact that, the earlier analyses have used non-parametric response curves for both species. One possible future direction could be to test the climate change predictions with such models as well.

If predicting the abundance seems to be a hard task, when reducing our objective to modeling presence-absence of the larvae, we obtain slightly better results. We did implement a HSDM, following a similar structure as in 3.1, but assuming the target distribution being a Bernoulli:

$$y_i|\theta \sim Bernoulli(\theta(\mathbf{s}_i, \mathbf{x}_i))$$

where θ is the probability that the larvae occur, and latent variable $f(\mathbf{x_i}, \mathbf{s_i}) = \mathbf{logit}(\theta(\mathbf{s_i}, \mathbf{x_i}))$ can be modeled with a linear model and a Gaussian random

effect as before

$$f(\mathbf{x_i}, \mathbf{s_i}) = \mathbf{x_i}(\mathbf{s})^\top \beta + \phi(\mathbf{s_i})$$

We fit the model on our training data, and then recover an accuracy of the predictions on the test data, of 0.71 for whitefish and 0.74 for vendace. In light of these observations, a possible improvement in the fitting could be achieved by implementing a model so that we first check for the occurrence of larvae with a presence-absence SDM, and then, given the presence, we run an abundance SDM. To this aim, we can use zero-inflated or hurdle models (Lambert (1992)). Both provide mixtures of a Poisson and Bernoulli probability mass function to allow more flexibility in modeling the probability of a zero outcome.

Zero-inflated models add additional probability mass to the outcome of zero. For zero-inflated Negative-Binomial distributions, given a probability θ of drawing a zero, and a probability 1-θ of drawing from NegBin(λ, r), the probability function is

$$p(y_n|\theta, \lambda, r) = \begin{cases} \theta + (1-\theta)\text{NegBin}(0|\lambda, r) & if\ y_n = 0, \\ (1-\theta)\text{NegBin}(y_n|\lambda) & if\ y_n > 0. \end{cases}$$

The hurdle model is similar to the zero-inflated model, but more flexible in that the zero outcomes can be either deflated or inflated. Hurdle models, are formulated as pure mixtures of zero and non-zero outcomes, so that

$$p(y_n|\theta, \lambda, r) = \begin{cases} \theta & if\ y_n = 0, \\ (1-\theta)\frac{\text{NegBin}(y|\lambda,r)}{1-\text{NegBin}(\theta|\lambda,r)} & if\ y_n > 0. \end{cases}$$

In order to account for the high variability present in the data, we modeled the target variable with a Negative Binomial distribution. Nevertheless, the posterior distribution of the overdispersion parameter r is concentrated around quite small values, suggesting the presence of overdispersion not yet accounted by the model. In order to approximate overdispersed Poisson processes with a wide range of mean–variance relationships, we suggest the implementation of a more flexible parameterization of the Negative Binomial distribution, by writing the variance, σ^2, as a quadratic function of the mean μ, and two overdispersion parameters, ω and θ, so that:

$$\sigma^2 = \omega\mu + \theta\mu^2$$

as proposed by Lindén and Mäntyniemi (2011).

Also the unbalanced distribution of the sampling sites along the GoB seems to affect quite heavily the goodness of the models. In particular, when using only a small set of training data, predictions produce very unstable and unrealistic results, when having predictive covariates ranges way wider than the one from the training sets. Considering non stationary covariance functions for the spatial random effect, may help in capture more asymmetric correlation structures in the data.

An ultimate source of error could be found in the high spatial autocorrelation present among the covariates, such as salinity and temperature or temperature and ice coverage.

We will take a causal statistics approach to analyse spatiotemporal relationships in the data. One of the main objective of our future research, will be, then, studying how to account for strong dependency among environmental variables and targets, so that future predictive results could still be reliable, when considering scenarios in which a substantial change occurs in the environment, such as climate change. Among the various questions, we are willing to investigate if (even when we assume the original relationships between the covariates and the species abundance remaining the same), it is possible to obtain different correlation patterns in species distributions, when trying to predict them with SDMs fitted in the original environment.

6.3 Further extension: JSDM

We finally point out a possible improvement of the model, motivated by more ecological-related reasons. As stated in chapter 3, biotic interactions play a key role in shaping the assembly and dynamics of species communities at different spatiotemporal scales and can be evaluated by taking into account species-to-species associations (Ovaskainen and Soininen, (2011); Morales-Castilla et al. (2015)). It has been shown that interactions among species are usually really relevant in estimating species distributions models.

A good strategy to extend single SDMs to JSDMs could be to use a latent multivariate spatial process defined over locations in a region. The general model is still formalized by equation (3.1). As we are considering two different species, now we have that the species index set $j \in \{1,2\}$ and $J = 2$ denotes the total number of species.

When we consider independent models for each species, that is implementing J different SSDM, the spatial random effect $\phi(\cdot)$ and the covariates effects are assumed to be mutually independent among all species.

To implement a JSDM, we assume $\mathbf{Y}(\mathbf{x},\mathbf{s})^\top = [Y_1,\dots,Y_J]$ to be a J-variate random vector with components $Y_j = Y_j(\mathbf{x_j},\mathbf{s})$ related to the j-th species at spatial location $\mathbf{s}^\top = [s_1, s_2]$ and with p environmental covariates $\mathbf{x}_j^\top = [x_{j,1},\dots,x_{j,p}]$. To model both vendace and whitefish larvae we can express $Y(\mathbf{x},\mathbf{s})$ with a bivariate Negative Binomial distribution and $f(\mathbf{x},\mathbf{s})$ as a bivariate Gaussian Process, given the hyperparameters. As such, its distribution is specified by its mean and its covariance function.

The crucial element in computing the latter is the so called cross-covariance matrix, a $J \times J$ matrix that accounts for the correlation among different species. The most straightforward specification of a valid cross-covariance function for a J-dimensional random field is through a Separable model (Mardia and Goodall (1993), Banerjee and Gelfand (2002)). Other possible cross-covariance matrix construction could be achieved through kernel specification or convolution (Gelfand et al. (2004)).

If implementing a Separable model for the cross-covariance in our case study, the resulting JSDM, given the species index j = 1,2, at a specific location $\mathbf{s_i}$, with $i \in 1,\dots,N$, will be

$$f(\mathbf{x}_{ij},\mathbf{s}_i) := \log(\frac{\mu_{ij}}{V_{ij}}) = \mathbf{x_{ij}^\top}\beta_\mathbf{j} + \phi_\mathbf{ij}(\mathbf{s_i})$$

where the 15-dimensional vector of linear weights β_j describes how species j responds to environmental covariates, and hence characterizes its environmental niche: $\mu_j = \mathbf{X}\beta_\mathbf{j}$ denotes the expected environmental niche of species j , and the variance-covariance matrix Σ_β the variation around this expectation (Ovaskainen and Soininen, (2011)). ϕ_{ij} models variation in species occurrence and co-occurrences. Considering all the locations $\mathbf{s} \in \mathbf{1},\dots,\mathbf{N}$ we can express our latent function as a $2N$ dimensional GP:

$$\mathbf{f}(\mathbf{x},\mathbf{s}) = \begin{pmatrix}\mathbf{f_{wf}}\\ \mathbf{f_{ven}}\end{pmatrix} \sim \mathbf{MGP}\Big(\mathbf{0},\mathbf{\Sigma_f}\Big)$$

where $\mathbf{f_{wf}}$ and $\mathbf{f_{ven}}$ are two N dimensional subvectors modeling, respectively, whitefish and vendace' larvae density, and

$$\Sigma_f = \Sigma_\beta + \Sigma_\phi$$

Σ_β accounts for covariance from linear model, and has elements $\sigma_\beta(\mathbf{s},\mathbf{s}')_\mathbf{ij} = \mathbf{x}(\mathbf{s})^\top\mathbf{x}(\mathbf{s}')\sigma_\mathbf{ij}$ as in previous chapters, while the covariance from random effects is a $2N \times 2N$ covariance block matrix, such that its $N \times N$ blocks

(each of dimension 4×4) account for species specific covariance and species interactions at each locations $\mathbf{s}, \mathbf{s}'$. The matrix has the following form:

$$\mathbf{R} \otimes \mathbf{T} = \begin{pmatrix} \rho_{11}\mathbf{T} & \dots & \rho_{1N}\mathbf{T} \\ \vdots & \ddots & \\ \rho_{N1}\mathbf{T} & \dots & \rho_{NN}\mathbf{T} \end{pmatrix}$$

where the symbol $\otimes$ stands for the Kronecker product, that in our case, gives:

$$(R \otimes T)_{ij} = r_{\lfloor (i-1)/2 \rfloor + 1, \lfloor (j-1)/2 \rfloor + 1} t_{(i-1)\%2+1,(j-1)\%2+1}$$

where $i\%2$ denotes the remainder of $i/2$. $\mathbf{R}$ is a $N \times N$ matrix with elements $r_{ih} = \rho(\mathbf{s}_i, \mathbf{s}_h)$ that measures spatial correlation among sites $\mathbf{s}_i$ and $\mathbf{s}_h$, with $i, h \in 1, \dots, N$, through e.g. a Matérn covariance function. $\mathbf{T}$ is a 2×2 matrix that accounts for correlations among the two species.

Another possible extension, to relate community-level responses to environmental variation to response traits, can be obtained by including data on species-specific traits. These data may range from morphological traits such as body size, or physiological traits such as tolerance to salinity, to functional traits such as feeding type, or to the actual position of the species within the surrounding food web. Furthermore, we may combine trait data with phylogenetic data, that concern evolutionary relationships among biological entities (Ovaskainen et al. (2017)).

Nature is complex, and the unexplored or little studied fields in environmental and ecological statistics are many. SDMs is among the most popular topics in this science, but, as we demonstrate in this thesis, the opportunities to improve these models are various, and open in many directions.

Bibliography

Agarwal, Chetan & Green, G.M. & Grove, Morgan & Evans, T.P. & Schweik, C.M.. (2002). *A Review and Assessment of Land-Use Change Models.Dynamics of Space, Time, and Human Choice.*

Albert A., Vetemaa M., Saat T. (2004). *Effects of salinity on the development of Peipsi whitefish Coregonus lavaretus maraenoides Poljakow embryos.* Ann Zool Fenn 41: 85-88

Aneer G, Blomqvist EM, Hallbäck H, Mattila J, Nellbring S,Skóra K, Urho L (1992) *Methods for sampling of shallowwater fish. Baltic Marine Biologists Publication 13*

Araújo, M.B., Alagador, D., Cabeza, M., Nogués-Bravo, D. and Thuiller, W. (2011), *Climate change threatens European conservation areas.* Ecology Letters, 14: 484-492.

Austin, M.P. & Cunningham, R.B. (1981) *Observational analysis of environ mental gradients.* Proceedings of the Ecological Society of Australia, 11, 109–119.

Austin, M. P. and Meyers, J. A. (1996). *Current Approaches to Modelling the Environmental Niche of Eucalypts: implication for management of forest biodiversity.* Forest Ecology and Management, 85:95-106.

Banerjee S.. and Gelfand A. E. (2002). *Prediction, interpolation and regression for spatial misaligned data points.* Sankhya, 64:227-245.

Battersby, S. (2017). *Map Projections. The Geographic Information Science & Technology Body of Knowledge* (2nd Quarter 2017 Edition), John P. Wilson (ed.).

Banerjee S., Carlin B. P. and Gelfand A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman-Hall CRC Press, Boca Raton.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modelling and Analysis for Spatial Data.* Chapman Hall/CRC, second edition.

Bernardo, Jose & Smith, Adrian. (2000). *Bayesian Theory.* 10.2307/2983298.

Roger Bivand, Tim Keitt and Barry Rowlingson (2019). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.4-8. https://CRAN.R-project.org/package=rgdal

Chris Brunsdon and Hongyan Chen (2014). GISTools: Some further GIS capabilities for R. R package version 0.7-4. https://CRAN.R-project.org/package=GISTools

Cawsey, E.M., Austin, M.P. & Baker, B.L. (2002) *Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling.* Biodiversity and Conservation, 11, 2239–2274.

Clark, J. S. et al. *More than the sum of the parts: forest climate response from joint species distribution models.* Ecological Applications 24, 990-999 (2014).

Clark, J. S., D. Nemergut, B. Seyednasrollah, P. Turner, and S.Zhang. (2017). *Generalized joint attribute modeling for biodi-versity analysis: median-zero, multivariate, multifarious data.* Ecological Monographs 87:34–56.

Cressie, N. A. C., & Wikle, C. K. (2011). *Statistics for spatio-temporal data.* Hoboken, N.J: Wiley.

De'ath, G. (2002), *MULTIVARIATE REGRESSION TREES: A NEW TECHNIQUE FOR MODELING SPECIES–ENVIRONMENT RELATIONSHIPS.* Ecology, 83: 1105-1117.

Di Narzo, A. F., and D. Cocchi. *A Bayesian Hierarchical Approach to Ensemble Weather Forecasting.* (2008) Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 59, no. 3, 2010, pp. 405-422.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics.* Springer Science+Business Media, LLC.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). *Model-based geostatistics.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 47(3):299–350.

Simon Duane, A.D. Kennedy, Brian J. Pendleton, Duncan Roweth, *Hybrid Monte Carlo*, Physics Letters B, Volume 195, Issue 2, (1987), Pages 216-222, ISSN 0370-2693,

Dirk Eddelbuettel and James Joseph Balamuta (2017). Extending R with C++: A Brief Introduction to Rcpp. PeerJ Preprints 5:e3188v1. URL https://doi.org/10.7287/peerj.preprints.3188v1.

Elith, Jane & Leathwick, J.R.. (2009). *Species Distribution Models: Ecological Explanation and Prediction Across Space and Time.* Annual Review of Ecology, Evolution and Systematics. 40. 677-697. 10.1146/annurev.ecolsys.110308.120159.

Fielding, A.H. and J.F. Bell. (1997) *A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models.* Environmental Conservation, 24, 38-49.

Fortuna MA, Bascompte J. *Habitat loss and the structure of plant-animal mutualistic networks.* Ecol Lett. (2006);9(3):281-286.

Gaston, Kevin. (2003). *The Structure and Dynamics of Geographic Ranges.*

Gelfand, A.E., Schmidt, A.M., Banerjee, S. et al. *Nonstationary multivariate process modeling through spatially varying coregionalization.* Test 13, 263–312 (2004).

Gelfand JM, Weinstein R, Porter SB, Neimann AL, Berlin JA, Margolis DJ. *Prevalence and treatment of psoriasis in the United Kingdom: a population-based study.* Arch Dermatol. 2005;141(12):1537-1541.

Gelfand, Alan E.; Silander, John A.; Wu, Shanshan; Latimer, Andrew; Lewis, Paul O.; Rebelo, Anthony G.; Holder, Mark. *Explaining species distribution patterns through hierarchical modeling.* Bayesian Anal. 1 (2006), no. 1, 41–92. doi:10.1214/06-BA102.

Gelman, A. (2006). *Prior distributions for variance parameters in hierarchical models.* Bayesian Analysis, 1(3):515–533.

Gelman, A., Goegebeur, Y., Tuerlinckx, F. and Van Mechelen, I. (2000), *Diagnostic checks for discrete data regression models using posterior predictive simulations.* Journal of the Royal Statistical Society: Series C (Applied Statistics), 49: 247-268.

Gelman, A. & Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, New York.

Gelman, A. and Rubin, D.B. (1992) *Inference from iterative simulation using multiple sequences*, Statistical Science, 7, 457-511.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D. (2013). *Bayesian Data Analysis.* New York: Chapman and Hall/CRC, https://doi.org/10.1201/b16018

Goulard, M., Voltz, M. *Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix.* Math Geol 24, 269–286 (1992)

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) *New developments in museum-based informatics and applications in biodiversity analysis.* Trends in Ecology and Evolution, 19, 497–503

Guisan A, Broennimann O, Engler R, et al. *Using niche-based models to improve the sampling of rare species.* Conserv Biol. 2006;20(2):501-511

Hartmann, A., T. Gleeson, Y. Wada, and T. Wagener, (2017): *Enhanced groundwater recharge rates and altered recharge sensitivity to climate variability through subsurface heterogeneity.* Proc. Natl. Acad. Sci., 114, no. 11, 2842-2847.

Hernandez, P.A., Graham, C.H., Master, L.L. and Albert, D.L. (2006), *The effect of sample size and species characteristics on performance of different species distribution modeling methods.* Ecography, 29: 773-785. doi:10.1111/j.0906-7590.2006.04700.x

Hirzel, Alexandre & Guisan, Antoine. (2002). *Which is the optimal sampling strategy for habitat suitability modelling.* Ecological Modelling. 157. 331-341. 10.1016/S0304-3800(02)00203-X.

Robert J. Hijmans (2020). raster: Geographic Data Analysis and Modeling. R package version 3.0-12. https://CRAN.R-project.org/package=raster

Hoffman MD, Gelman A (2014). *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.* Journal of Machine Learning Research, 15, 1593–1623.

Hordoir R, Axell L, Höglund A, Dieterich C, Fransner F, Gröger M, Liu Y, Pemberton P, Schimanke S, Andersson H, et al. 2019. *Nemo-Nordic 1.0:*

a NEMO-based ocean model for the Baltic and North seas – research and operational applications. Geosci Model Dev. 12:363–386. doi:10.5194/gmd-12-363-2019

Hudd R, Hilden M, Urho L, Axell MB, Jåfs LA (1984) *Fiskeriundersökning av Kyrö älvs mynningsoch influensområde 1980-82 [Fishery investigation in 1980-1982 of theKyrönjoki Estuary and its influence area in the NorthernQuarter of the Baltic Sea].* Report 242B, National Boardof Waters, Helsinki

Hudd R, Lehtonen H, Kurttila I (1988) *Growth and abun-dance of fry; factors which influence the year-classstrength of whitefish (Coregonus lavaretus widegreni) in the southern Bothnian Bay (Baltic).* Finn Fish Res 9: 213-220

Huettmann, F. (2005). *Databases and science-based managementin the context of wildlife and habitat: towards a certified ISOstandard for objective decision-making for the globalcommunity by using the internet.*-J. Wildl. Manage. 69:466-472.

Hui, F. K. C., D. I. Warton, S. D. Foster, and P. K. Dunstan.(2013). *To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models.* Ecology 94:1913–1919.

Jia Liu, Jarno Vanhatalo, *Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process*, Spatial Statistics, Volume 35, 2020, 100392, ISSN 2211-6753

Jäger T., Nellen W., Schöfer W., Shodjai F. (1981). *Influence of salinity and temperature on early life stages of Coregonus albula, C. lavaretus, R. rutilus, and L. lota.* Rapp P-V Reun Cons Int Explor Mer 178: 345-348

Kadmon, R., Farber, O. & Danin, A. (2003) *A systematic analysis of factors affecting the performance of climatic envelope models.* Ecological Applications, 13, 853–867.

Kallasvuo, M., Vanhatalo, J., & Veneranta, L. (2017). *Modeling the spatial distribution of larval fish abundance provides essential information for management.* Canadian Journal of Fisheries and Aquatic Sciences 74: 5, p. 636-649.

Kotta, J., Vanhatalo, J., Jänes, H. et al. *Integrating experimental and distribution data to predict future species patterns.* Sci Rep 9, 1821 (2019)

Kruskal, J.B. *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.* Psychometrika 29, 1–27 (1964)

Lambert, Diane. (1992). *Zero-Inflated Poisson Regression, With An Application to Defects in Manufacturing.* Technometrics. 34. 1-14. 10.1080/00401706.1992.10485228.

Latimer AM, Wu S, Gelfand AE, Silander JA (2006) *Building statistical models to analyze species distributions.* Ecol Appl 16:33-50

Leathwick, John & Morgan, Fraser & Wilson, Gareth & Rutledge, Daniel & Mcleod, Malcolm & Johnston, Kirsty. (2003). *Land Environments of New Zealand: A Technical Guide.*

Lehtonen H. 1981. *Biology and stock assessments of Core-gonids at the Baltic coast of Finland. Finn. Fish.* Res. 3: 31-83.

Lindén, A. and Mäntyniemi, S. (2011), *Using the negative binomial distribution to model overdispersion in ecological count data.* Ecology, 92: 1414-1421. doi:10.1890/10-1831.1

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) *AUC: amisleading measure of the performance of predictive distributionmodels.* Global Ecology and Biogeography, 17, 145 –151.

J.M. Lobo, A. Jiménez-Valverde, J. Hortal *The uncertain nature of absences and their importance in species distribution modelling* Ecography, 33 (2010), pp. 103-114

J.M. Lobo, M.F. Tognelli *Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data* Journal of Nature Conservation, 19 (2011), pp. 1-7

Mac Nally, R.(2000) *Regression and model-building in conservation biology, biogeography and ecology: the distinction between 'predictive' and 'explanatory' models.* Biodiversity and Conservation, 655-671

Mäkinen, J, Vanhatalo, J. (2018). *Hierarchical Bayesian model reveals the distributional shifts of Arctic marine mammals.* Divers Distrib. 24: 1381–1394.

Mardia, K.V., Goodall, C., Redfern, E.J. et al. *The Kriged Kalman filter.* Test 7, 217–282 (1998). https://doi.org/10.1007/BF02565111

Meier, H. E. M. et al.(2012) *Modeling the combined impact of changing climate and changing nutrient loads on the Baltic Sea environment in an ensemble of transient simulations for 1961–2099.* Climate Dynamics39, 2421-2441

Meier, H. E. M. et al. (2012) *Impact of climate change on ecological quality indicators and biogeochemical fluxes in the Baltic sea: A multi-model ensemble study.* Ambio41, 558-573

Meynard, C.N. and Kaplan, D.M. (2012), *The effect of a gradual response to the environment on species distribution modeling performance.* Ecography, 35: 499-509.

Midgley GF, Hannah L, Millar D, Rutherford MC, Powrie LW (2002) *Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot.* Glob Ecol Biogeogr 11:445-451

Morales-Castilla I, Matias MG, Gravel D, Araújo MB. Inferring biotic interactions from proxies. Trends Ecol Evol. 2015;30(6):347-356. Morales-Castilla I, Matias MG, Gravel D, Araújo MB. *Inferring biotic interactions from proxies.* Trends Ecol Evol. 2015;30(6):347-356.

Moss R, Babiker M, Brinkman S, Calvo E, Carter T, Edmonds J, Elgizouli I, Emori S, Erda L, Hibbard KA et al (2008) *Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. IPCC Expert Meeting Report on New Scenarios.* Intergovernmental Panel on Climate Change, Noordwijkerhout

Neal, Radford. (2012). *MCMC using Hamiltonian dynamics.* Handbook of Markov Chain Monte Carlo. 10.1201/b10905-6.

Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). *Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions.* Ecology, 91(9), 2514-2521. https://doi.org/10.1890/10-0173.1

Ovaskainen, O., Abrego, N., Halme, P., & Dunson, D. (2016). *Using latent variable models to identify large networks of species-to-species associations at different spatial scales.* Methods in Ecology and Evolution, 7, 549–555.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. and Abrego, N. (2017). *How to make more out*

of community data? A conceptual framework and its implementation as models and software. Ecol Lett, 20: 561-576.

Ovaskainen, O., and J. Soininen. 2011. *Making more out of sparse data: hierarchical modeling of species communities.* Ecology 92:289–295.

Pearce, Jennie & Ferrier, Simon. (2000). *Evaluating the predictive performance of habitat models developed using logistic regression.* Ecological Modelling. 133. 225-245. 10.1016/S0304-3800(00)00322-7.

S.J. Phillips, M. Dudik, J. Elith, C.H. Graham, A. Lehmann, J. Leathwick, S. Ferrier *Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data* Ecological Applications, 19 (2009), pp. 181-197

David Pierce (2019). ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files. R package version 1.17. https://CRAN.R-project.org/package=ncdf4

Quiñonero-Candela and C. E. Rasmussen. *A unifying view of sparseapproximate Gaussian process regression.* Journal of Machine Learning Research, 6:1939–1959, 2005.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

C. E. Rasmussen & C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT Press, 2006

Raxworthy, C. J., Forstner, M. R. J. & Nussbaum, R. A. *Chameleon radiation by oceanic dispersal.* Nature 415, 784-787 (2003).

Ripley, B. D. (1981). Lengyel, I. & Derish, P.. (2002). *Spatial statistics.* John Wiley Sons, New York..

Stan Development Team. 2018. RStan: the R interface to Stan. R package version 2.17.3. http://mc-stan.org

Sang, H., Gelfand, A.E.(2010) *Continuous Spatial Process Models for Spatial Extreme Values.* JABES 15, 49-65 .

Santika, T. (2011), *Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data.* Global Ecology and Biogeography, 20: 181-192.

Seoane, Sergio & Laza-Martinez, Aitor & Urrutxurtu, Iñaki & Orive, Emma. (2005). *Phytoplankton Assemblages and Their Dominant Pigments in the Nervion River Estuary.* Hydrobiologia. 549. 1-13. 10.1007/s10750-005-1736-6.

Snelson E. and Ghahramani Z..*Sparse Gaussian processes using pseudo-inputs.* InAdvances inNeural Information Processing Systems 18, pages 1259 U-1266. MIT Press, 2006.

Soberón, J. & Peterson, A.T. (2005) *Interpretation of models offundamental ecological niches and species' distributional areas.* Biodiversity Informatics, 2, 1–10.

Sõrmus, I. & Turovski, A. 2003. *European whitefish, Coregonus lavaretus (L.) s. l., Baltic sea forms.* In E.Ojaveer, E.Pihu & T.Saat (eds.), Fishes of Estonia. Estonian Academy Publishers, Tallinn

Stan Development Team. 2018. *Stan Modeling Language Users Guide and Reference Manual*, Version 2.18.0. http://mc-stan.org

Ter Braak, C. J. F., and C. W. N. Looman. In press 1986. *Weighted averaging, logistic regression and the Gaussian response model.* Vegetatio 65:3-11.

Thorson, J. T., M. D. Scheuerell, A. O. Shelton, K. E. See, H. J.Skaug, and K. Kristensen. (2015). *Spatial factor analysis: anew tool for estimating joint species distributions and correlations in species range.* Methods in Ecology and Evolution 6:627–637.

Tikhonov G, Opedal ØH, Abrego N, et al. (2020) *Joint species distribution modelling with the r -package H msc* . Methods Ecol Evol. 2020;11:442-447

Tobler, W.R. (1979), *Lattice Tuning.* Geographical Analysis, 11: 36-44.

Tylianakis, J.M., Didham, R.K., Bascompte, J. and Wardle, D.A. (2008), *Global change and species interactions in terrestrial ecosystems.* Ecology Letters, 11: 1351-1363.

Warton, David & Foster, Scott & De'ath, Glenn & Stoklosa, Jakub & Dunstan, Piers. (2015). *Model-based thinking for community ecology.* Plant Ecology. 216. 669-682. 10.1007/s11258-014-0366-3.

Kotta, Jonne & Vanhatalo, Jarno & Jänes, Holger & Orav-Kotta, Helen & Rugiu, Luca & Jormalainen, Veijo & Bobsien, Ivo & Viitasalo, Markku & Virtanen, Elina & Sandman, Antonia & Isaeus, Martin & Leidenberger, Sonja & Jonsson, Per & Johannesson, Kerstin. (2018). *Integrating experimental and distribution data to predict future species patterns.* Scientific Reports. 9. 10.1038/s41598-018-38416-3.

Jarno Vanhatalo, Marcelo Hartmann and Lari Veneranta (2020). *Additive multivariate Gaussian processes for joint species distribution modeling with heterogeneous data.* Bayesian Analysis, 15(2):415-447.

Vanhatalo J, Veneranta L, Hudd R (2012) *Species distribution modeling with Gaussian processes: a case study with the youngest stages of sea spawning whitefish (coregonus lavaretus L. sl) larvae.* Ecol Model 228:49–58

Vanhatalo, J., & Vehtari, A. (2007). *Sparse Log Gaussian Processes via MCMC for Spatial Epidemiology.* Journal of Machine Learning Research, 2007(1), 73-89.

Vehtari, Aki; Ojanen, Janne. *A survey of Bayesian predictive methods for model assessment, selection and comparison.* Statist. Surv. 6 (2012), 142–228. doi:10.1214/12-SS102

Veneranta, Lari & Hudd, Richard. (2013). *Reproduction areas of sea-spawning coregonids reflect the environment in shallow coastal waters. Marine Ecology Progress Series.* 477. 231-250. 10.3354/meps10169.

Ward, E.J., Holmes, E.E. and Balcomb, K.C. (2009), *Quantifying the effects of prey abundance on killer whale reproduction.* Journal of Applied Ecology, 46: 632-640.

Wikle, C.K. (2003), *Hierarchical Models in Environmental Science.* International Statistical Review, 71: 181-199.

Zhang, H. (2004). *Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics.* Journal of the American Statistical Association, 99(465):250–261.