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

Ilaria Pia

University of Torino

ilaria.pia@helsinki.fi

21/07/2020

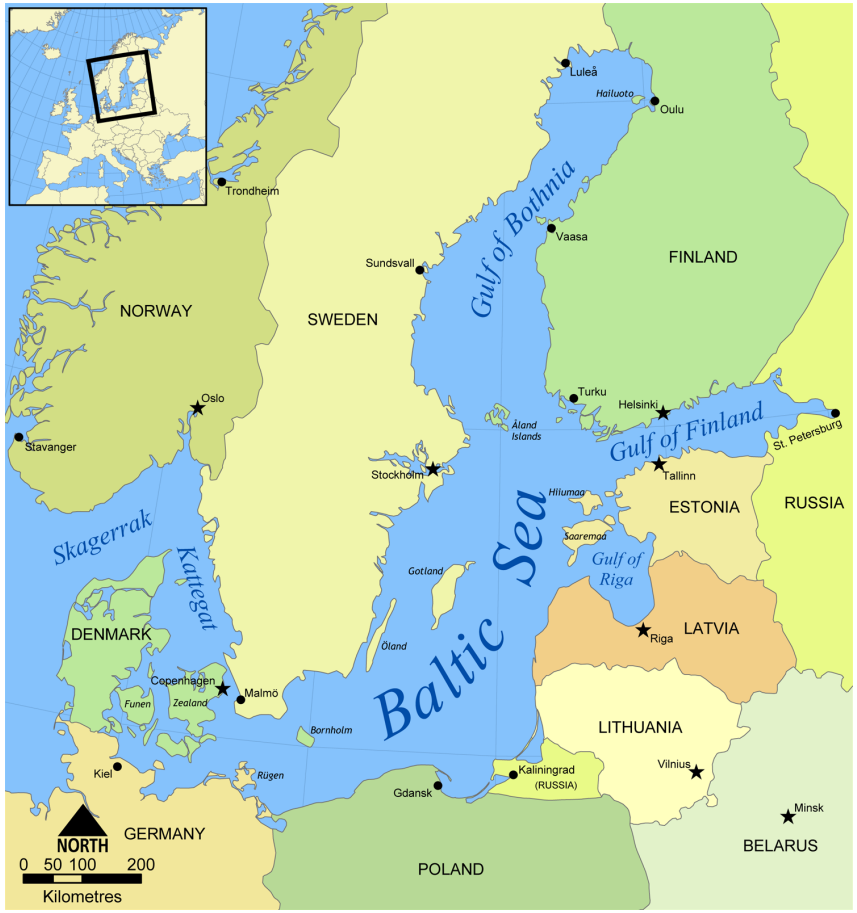
Overview

1. Case study
   - Study scenario
   - Data preparation
   - Target and covariates
2. Model
   - HSDM
   - Random effects
   - Covariance function and prior setting
   - Posterior distributions
3. Results and conclusions
   - Model selection
   - Covariates and target associations
   - Predictive logdensity

Study area

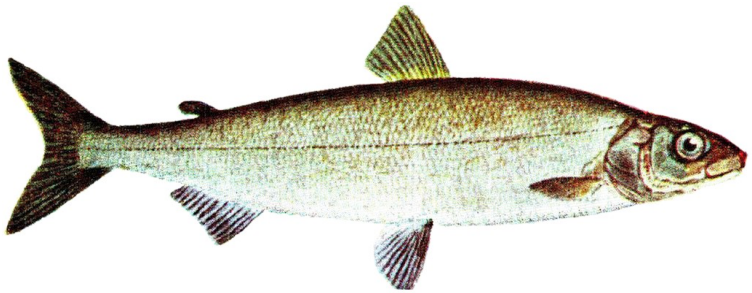
The Gulf of Bothnia (GoB) is a brackish water basin in the northern part of the Baltic Sea.

It is an excellent case study area to research the consequences of climate change.

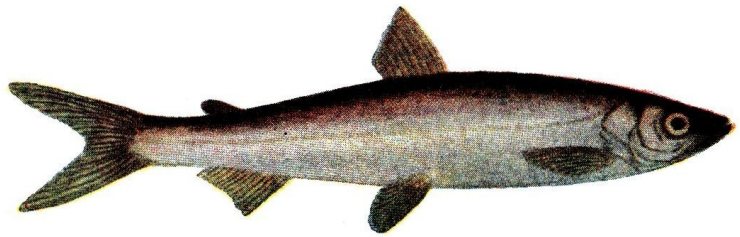


Whitefish and vendace

The GoB is mainly inhabited by 2 sea-spawning species of coregonids: whitefish (*Coregonus lavaretus*) and vendace (*Coregonus albula*).



(a) Whitefish



(b) Vendace

Thesis objectives

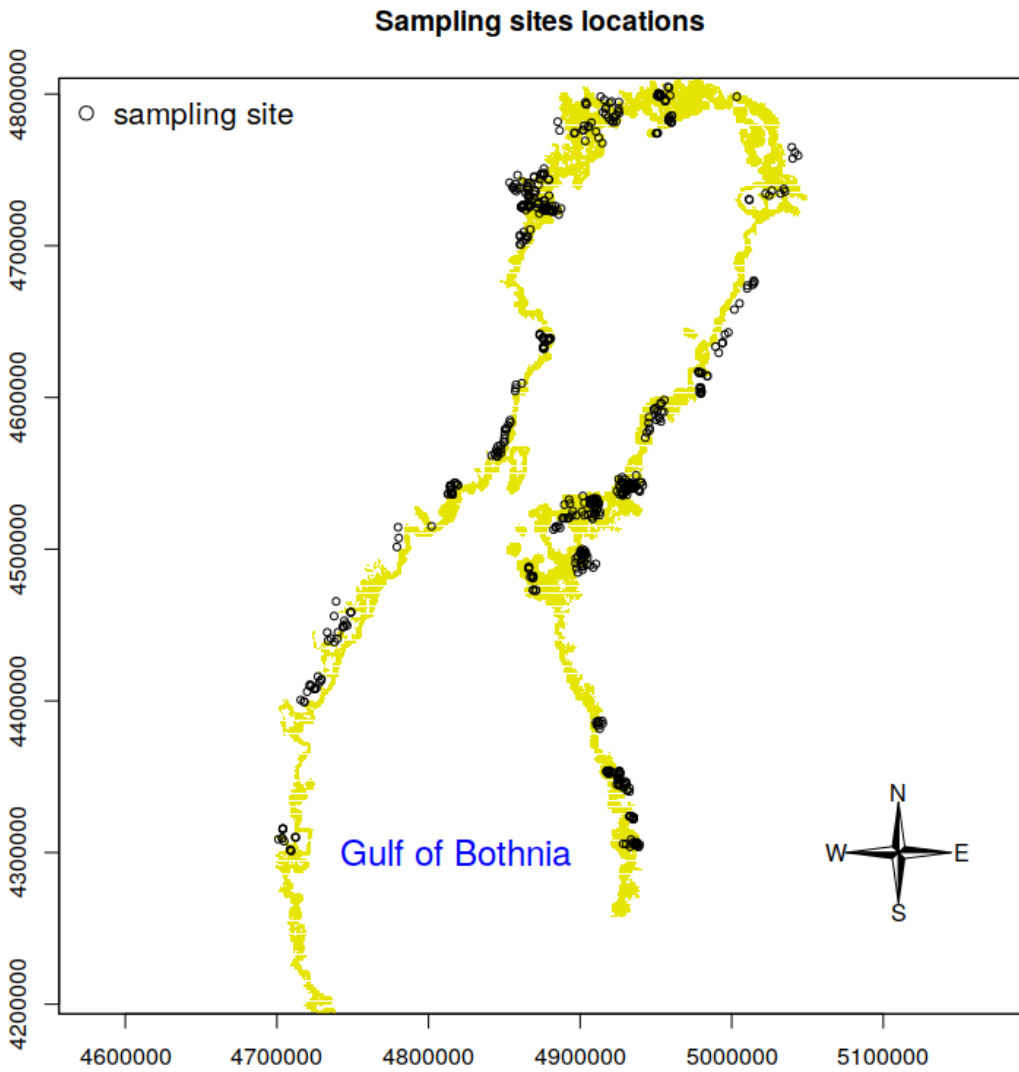
The applied goals of the thesis are twofold:

- to compile the existing whitefish data with new environmental data related to current and future Gulf of Bothnia conditions,
- to study how whitefish and vendace spawning distributions are affected by environmental changes.

Whitefish dataset

The whitefish dataset contains 653 observations and 63 different variables, which can be divided in two groups:

- Larvae abundance data
- Environmental variables



SmartSea data

The variables temperature, salinity and future last day of ice 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 [1]: SmartSea Gulf of Bothnia as Resource for Sustainable Growth.

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

Data preparation

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

1. Select the raster layers of interest and average them.
2. Remove the land cells, encoded as 0 valued raster cells.
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.

Raster layers of temperature

We report maps with current, future values of the variables and their difference.

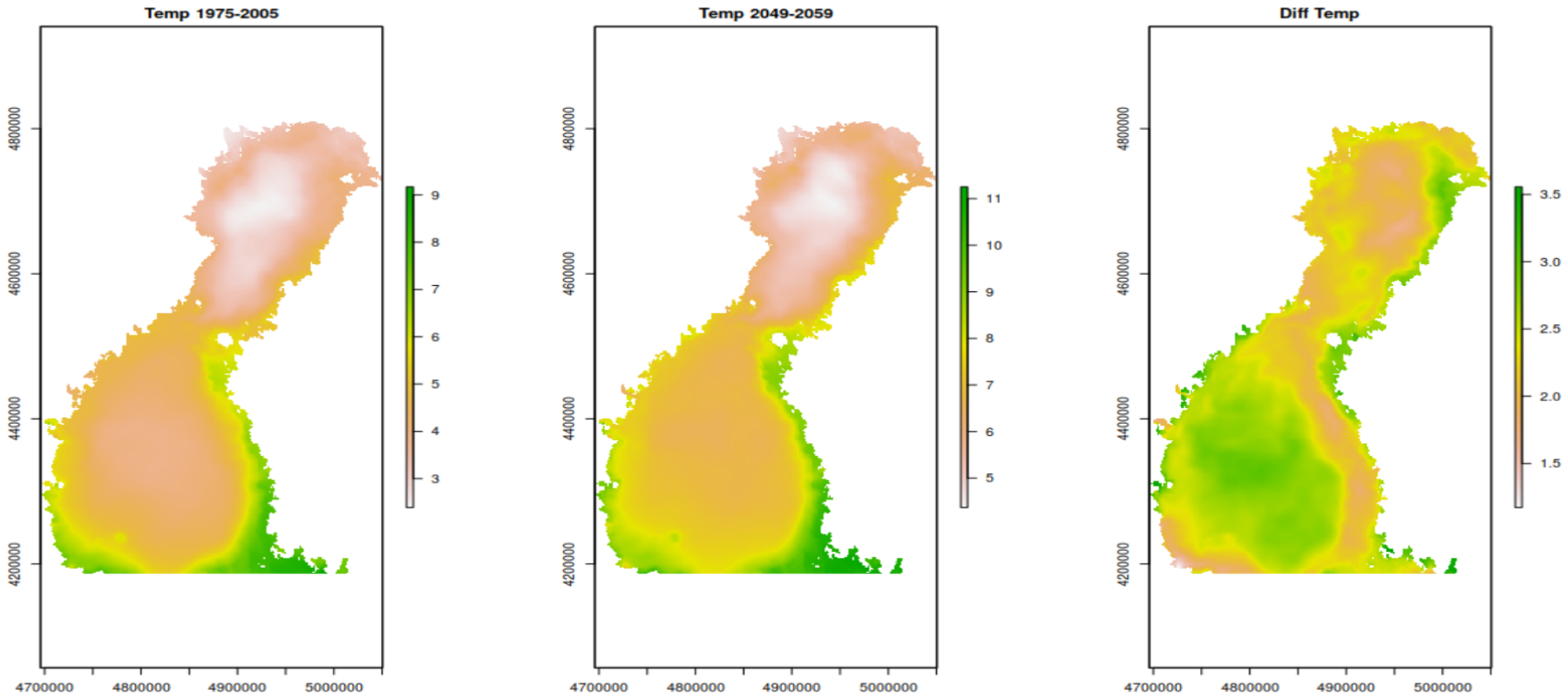


Figure: Temperature

Raster layers of salinity

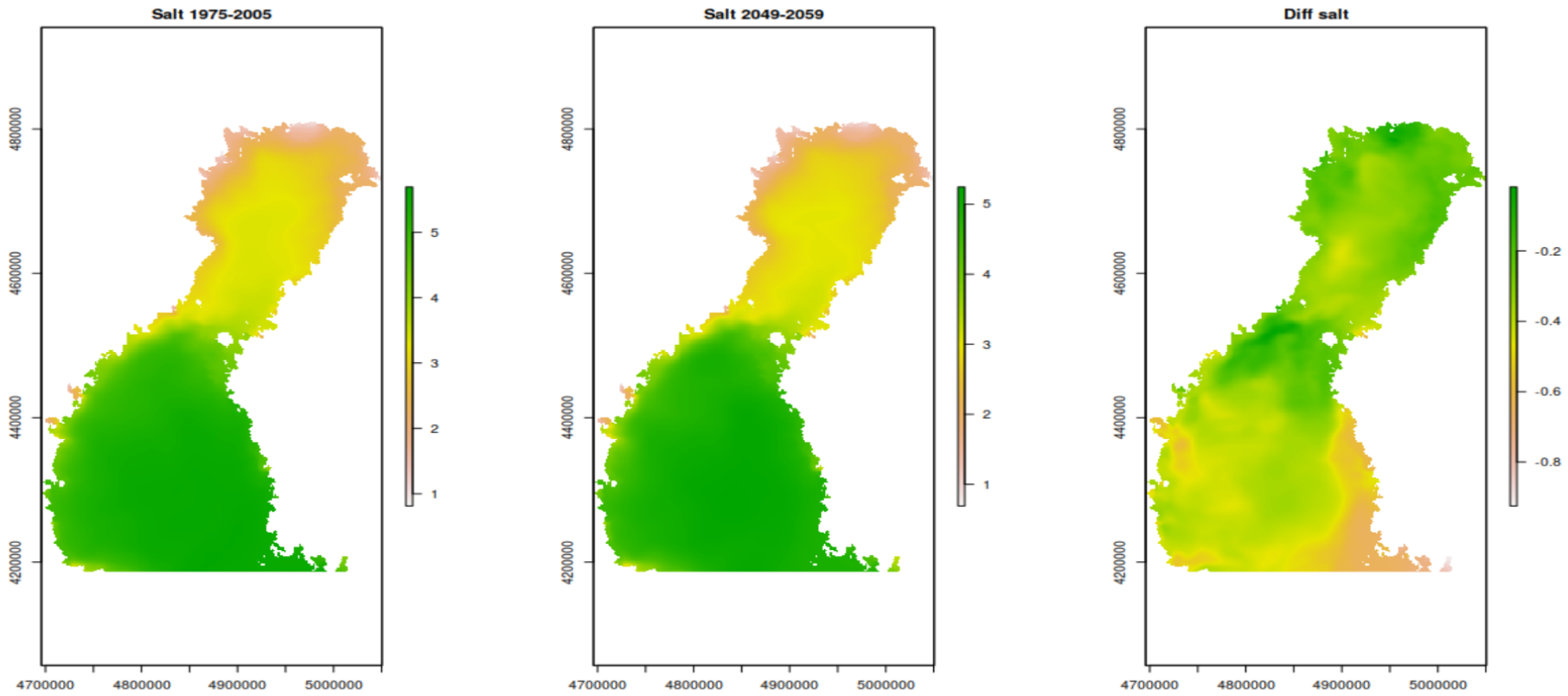


Figure: Salinity

Values for temperature and salinity are averaged on weekly outputs, obtained in the spring seasons, April-June.

Raster layers of last day of ice coverage

Last days of ice coverage are counted starting from the 1st of January of the previous year.

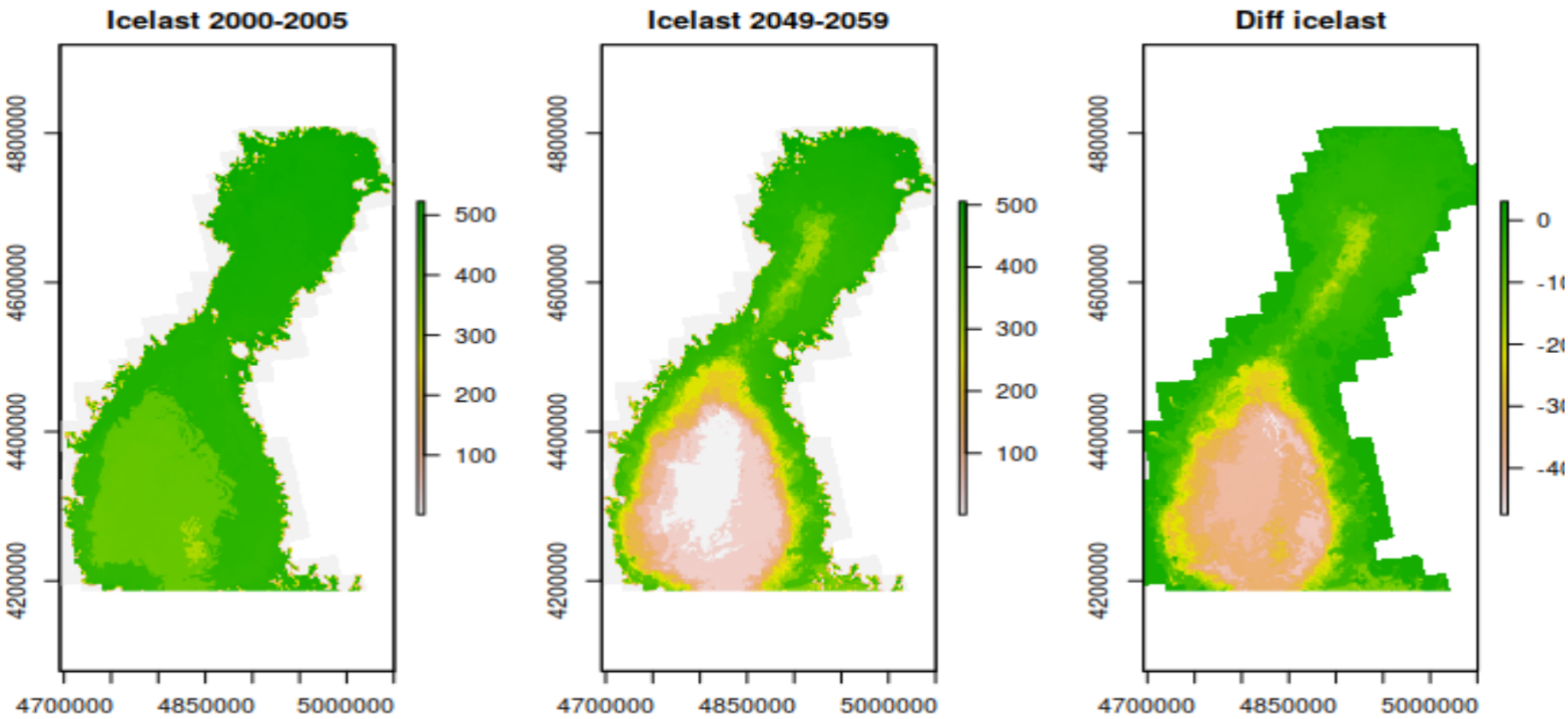


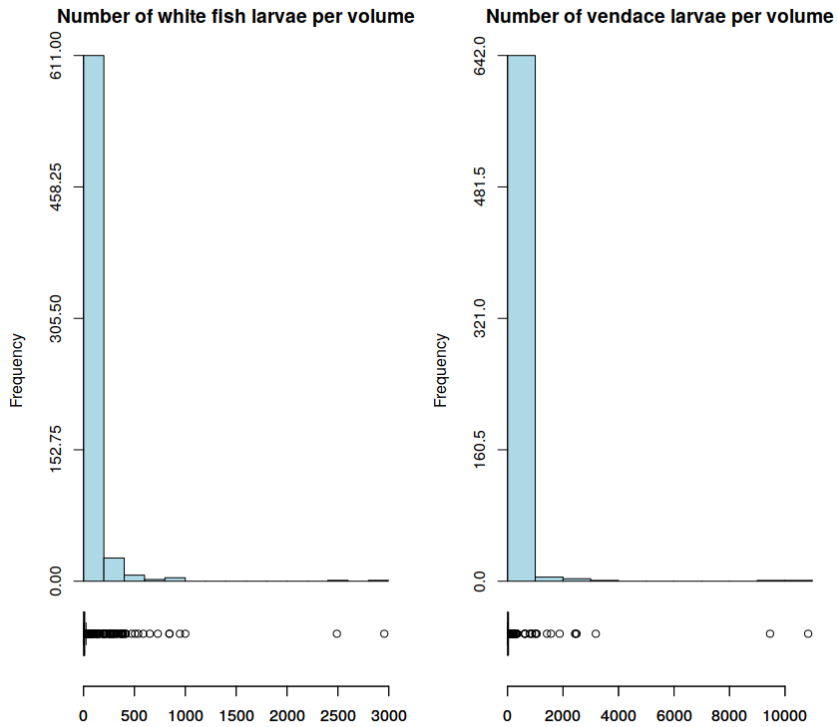
Figure: Last day of ice coverage

Target variables

The dependent variable is the larvae density, i.e. the number of larvae caught in sampling occasion, per unit volume. Depending on which species abundance is modeled, we set as target

- WHISUM/VOLUE for whitefishes
- VENSUM/VOLUME for vendaces

where the variable VOLUME, that indicates the volume of water sampled, serves as an "off-set" covariate for likelihood.



Covariates

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: Environmental covariates

Covariates

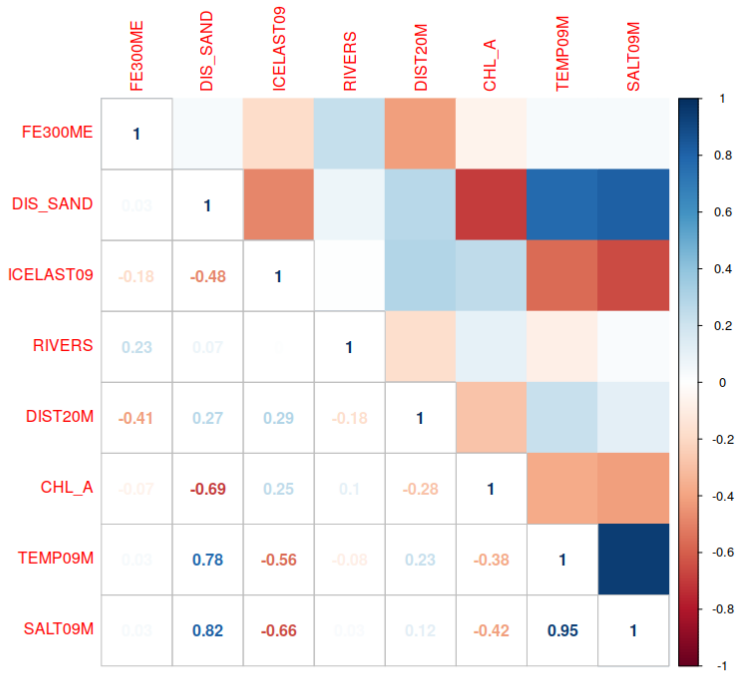


Figure: Correlation plot of the continuous covariates

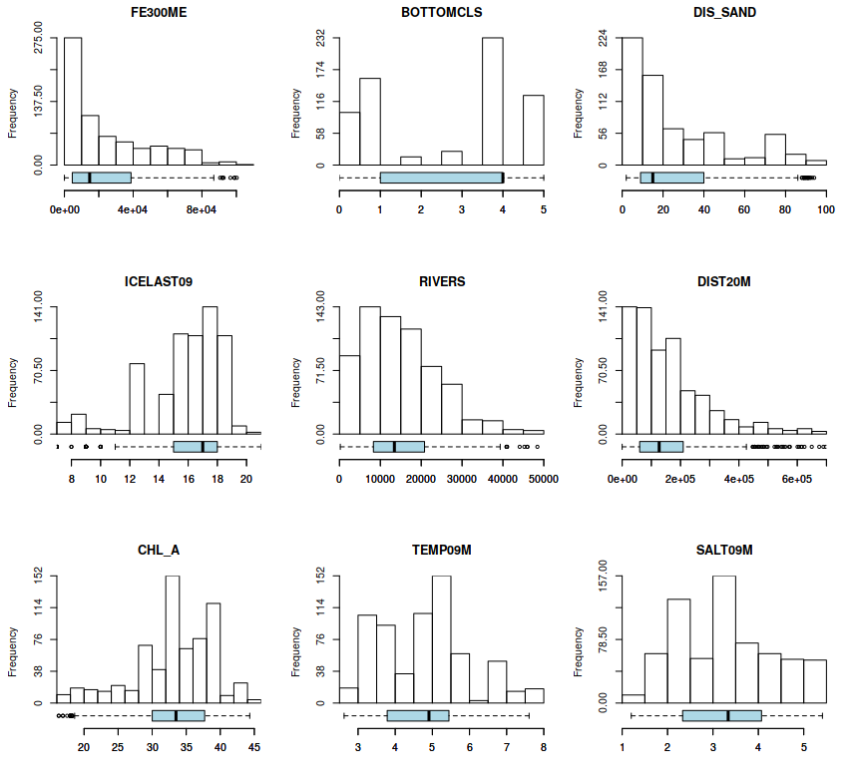


Figure: Training covariates distribution.

Species distribution models

Species distribution models (SDMs) are key tools in ecology that aim to predict distributions of species by relating presence/absence or abundance to environmental predictors.

They 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.

Spatial prediction of species distributions is thus directly related to the concept of *environmental niche*.

Hierarchical spatial models

We take the Bayesian perspective to spatial data analysis.
A general hierarchical spatial model (HSM) can be defined as:

$$[Data|process, parameters] \quad p(\mathbf{y}|\mathbf{f}(\cdot), \gamma) \tag{1}$$

$$[process|parameters] \quad p(\mathbf{f}(\mathbf{s})|\theta) \tag{2}$$

$$[parameters] \quad p(\theta, \gamma) \tag{3}$$

Larvae density SDM

To implement our SDM we let $\mathbf{y}$ represents the number of whitefishes or vendaces caught at the various sampling locations. Given $i \in 1, \ldots, N$

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

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

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

Additional iid random effect

We add some independent random effect ϵ_i , so that:

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

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

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

Spatial random effect

The variable $\phi(\mathbf{s})$ models the spatial random effects that describes temporally constant associations, unexplained by the available covariates.

$$\phi(\mathbf{s})|\Sigma_\phi \sim N(\ 0, \Sigma_\phi\ )$$

If we assume linear weights independent priors $\beta_d \sim N(0, \sigma^2_\beta)$, the marginal distribution for any $\mathbf{f} = [\mathbf{f}(\mathbf{s_1}), \ldots, \mathbf{f}(\mathbf{s_n})]$ is again Gaussian:

$$\mathbf{f}|\mathbf{S}, \mathbf{X}(\mathbf{S}) \sim \mathbf{N}(\ \mathbf{0}, \sigma^2_\beta \mathbf{X}(\mathbf{S})\mathbf{X}(\mathbf{S})^\top + \mathbf{\Sigma}_\phi\ )$$

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

Covariance function

We model the spatial covariance by mean of a stationary and isotropic function, the exponential covariance function

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

where l and σ^2_ϕ are the range and intensity parameters.

Σ_ϕ can be expressed as function of the euclidean distance between locations, and, hence, as function of the correlation range, defined as the distance c, such that:

$$\Sigma_\phi(c) = 0.05\Sigma_\phi(0) \tag{4}$$

Prior setting

We set the following priors for the model parameters

$$\beta_d \sim N(0, 10) \; \forall d \in 1, \ldots, 15$$

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

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

while for the length scale we consider two cases

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

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

Stan model and MCMC

We implement models in Stan and make use of MCMC algorithms from the Stan software, to recover the posteriors of our models' parameters.

We select a subset of the data, as training set, using the rest as test set. For each of the target we consider four models.

`https://github.com/ilapia/thesis-pia`

Whitefish convergence checking

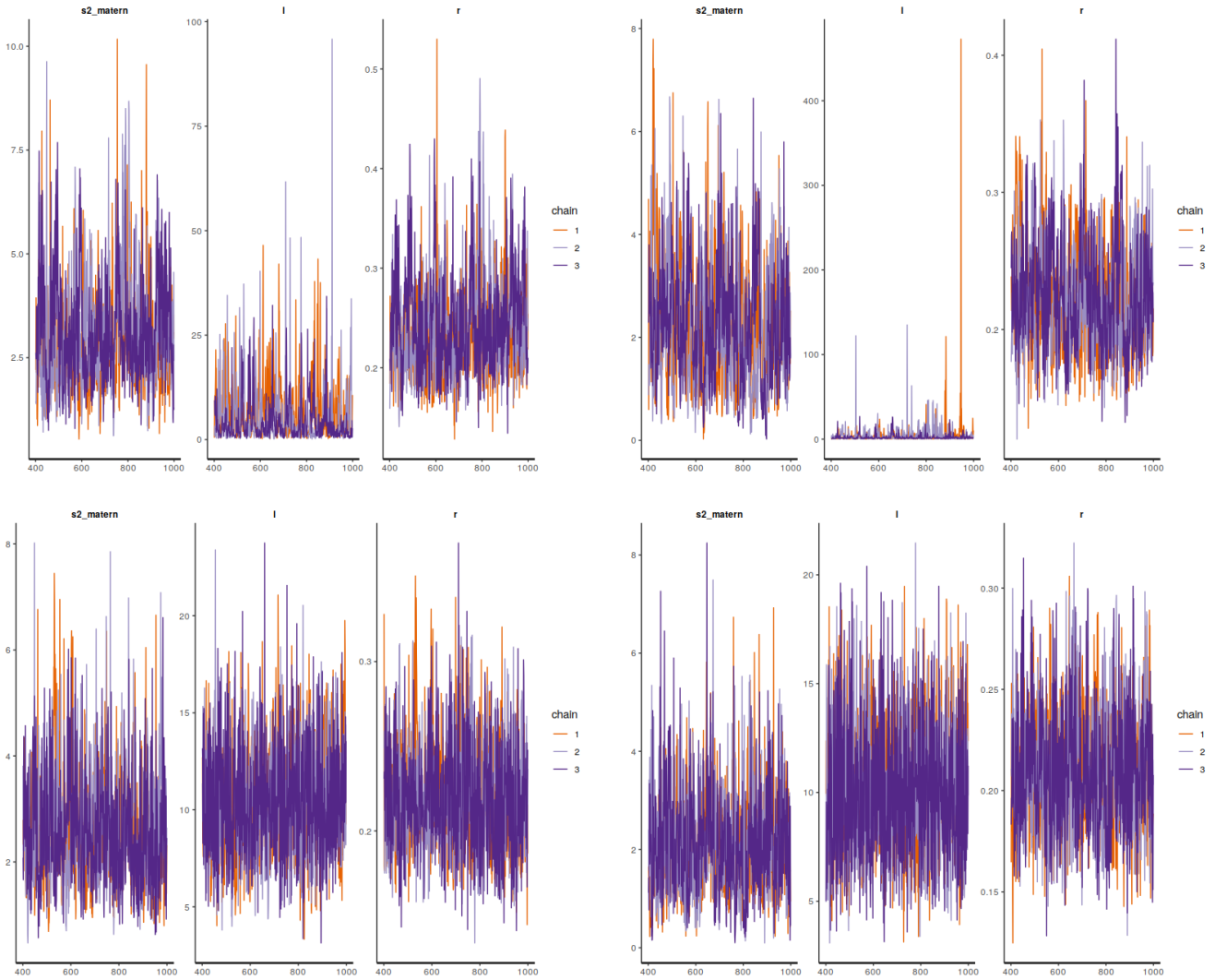


Figure: trace plots wf1.stud, wf2stud, wf1.gamma, wf2.gamma.

Vendace convergence checking

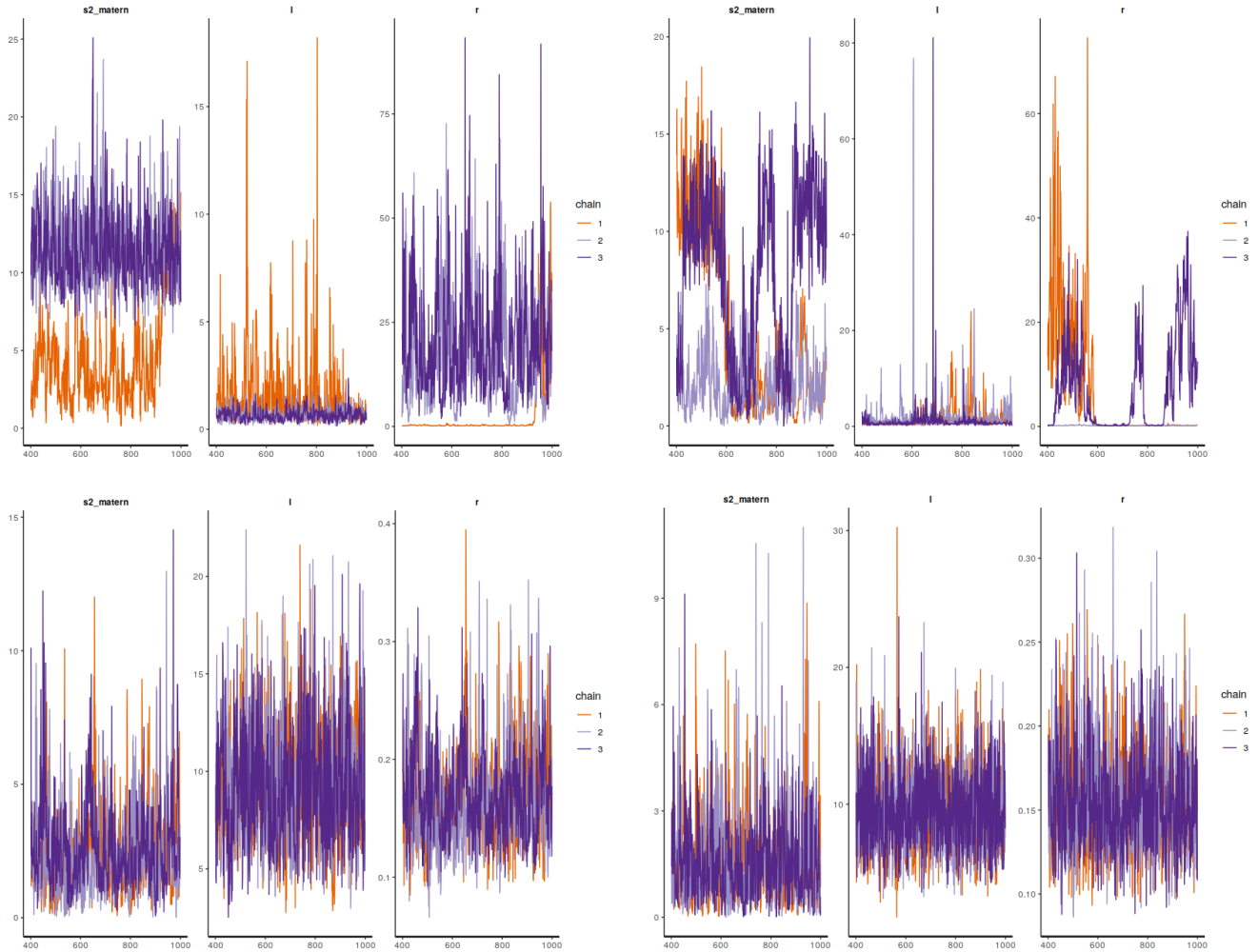


Figure: trace plots ve1.stud, ve2stud, ve1.gamma, ve2.gamma.

Model comparison

To assess the model predictive power we compute the log predictive density (lpd) of larvae abundance, at the test dataset:

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

model	lpd.wf	lpd.ve
fit1.gamma	-3.237	-1.932
fit2.gamma	-3.217	-1.944
fit1.stud	-3.196	-2.277
fit2.stud	-3.120	-2.189

Table: 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.

Models' parameters posteriors

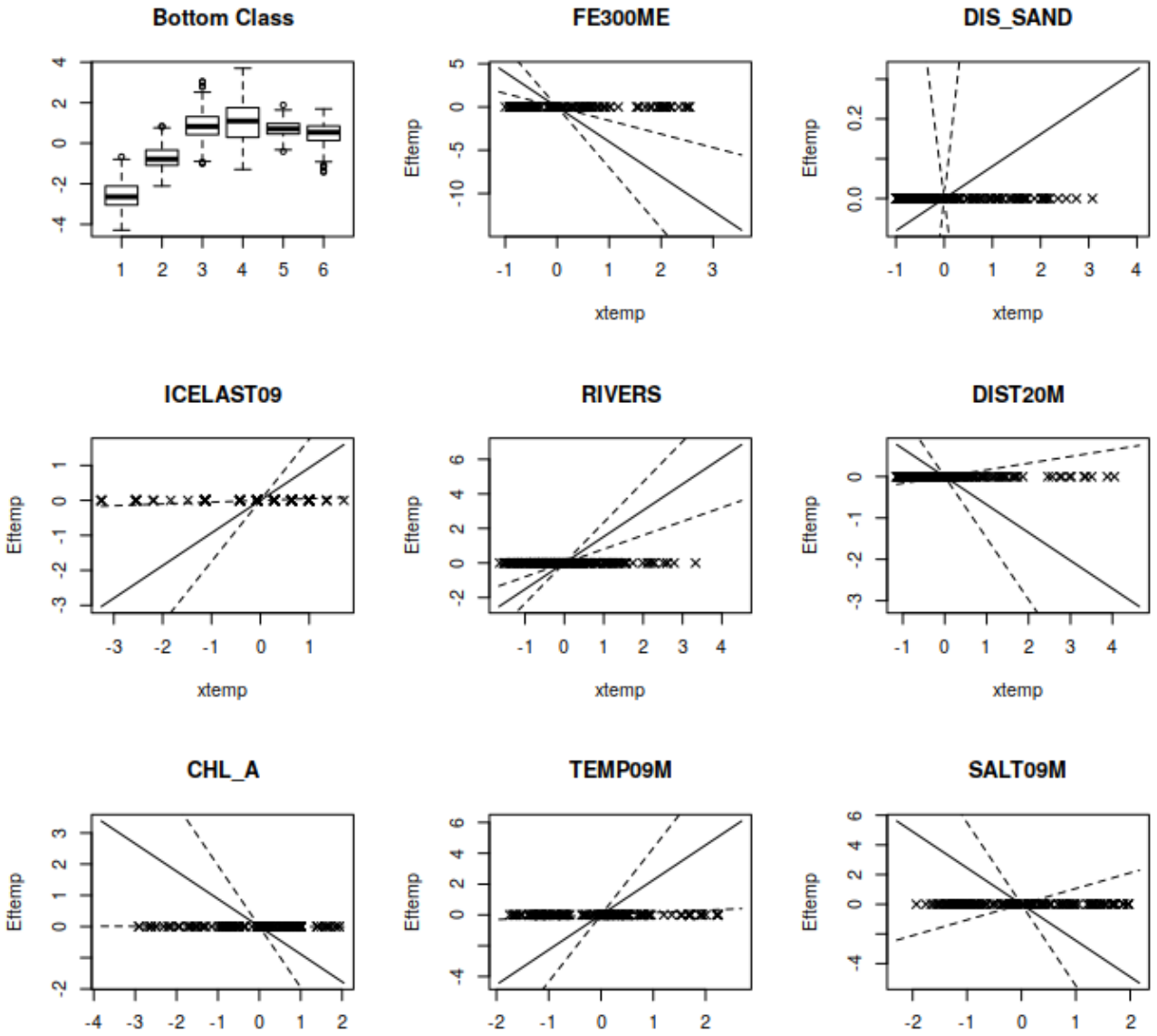
Whitefish

Covariate	Mean	Stand.Dev	quant2.5%	quant97.5%
s2_matern	2.982	0.084	1.278	6.115
l	5.647	0.414	6.787	23.965
r	0.239	0.003	0.047	0.356
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

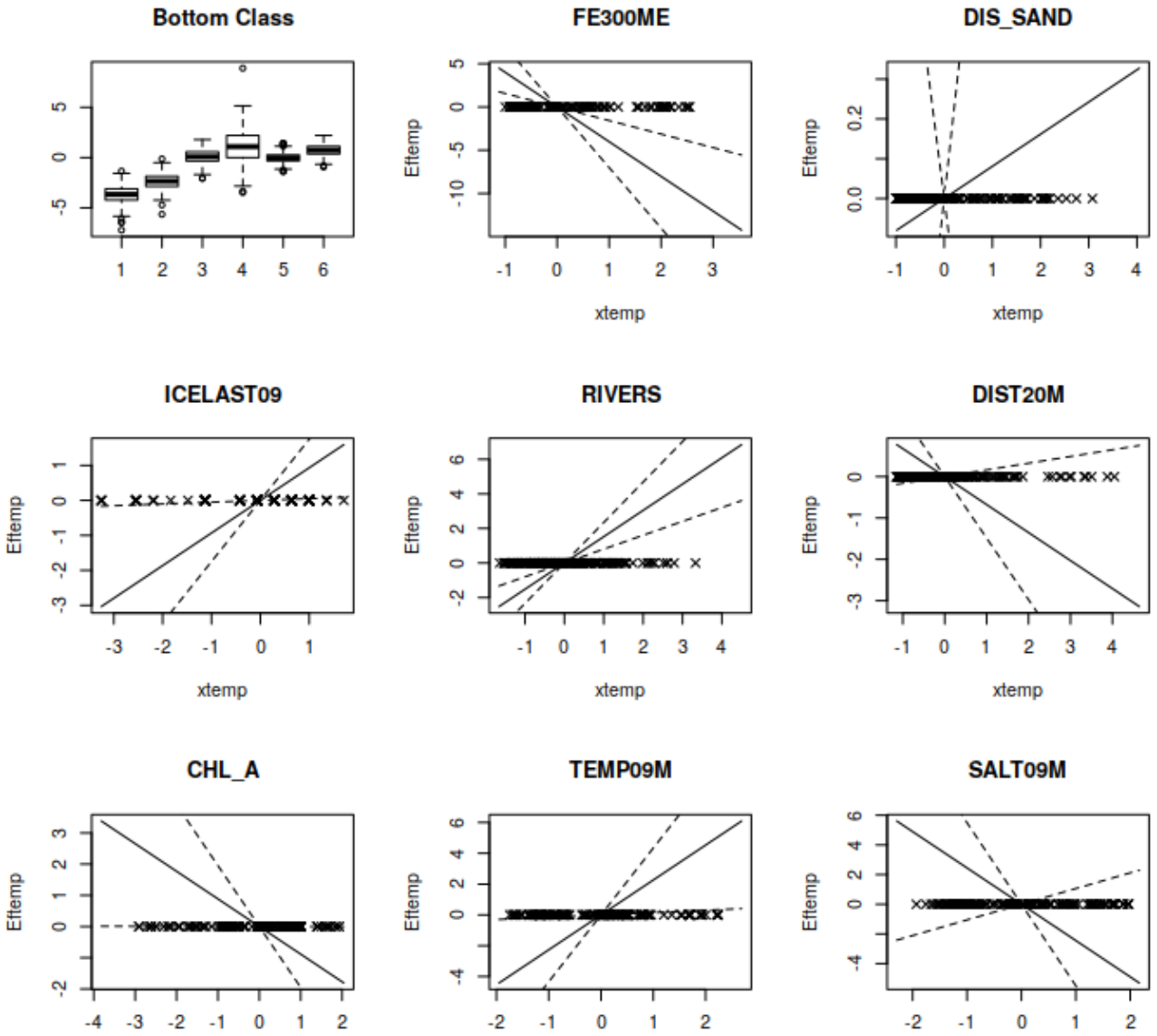
Vendace

Covariate	Mean	Stand.Dev	quant2.5%	quant97.5%
s2_matern	2.446	0.087	1.654	6.693
l	9.145	0.101	3.103	15.834
r	0.167	0.002	0.042	0.274
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

Whitefish HSDM response curve



Vendace HSDM response curve



Logdensity predictions

We finally report the spatial predictions for the whitefish and vendace larvae logdensity on the whole GoB area, for both current and future scenarios.
The posterior predictive density of latent variables, conditional on hyperparameters, is

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

$$\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}} - \mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}}(\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}}\mathbf{f} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{\mathbf{f},\tilde{\mathbf{f}}})$$

where

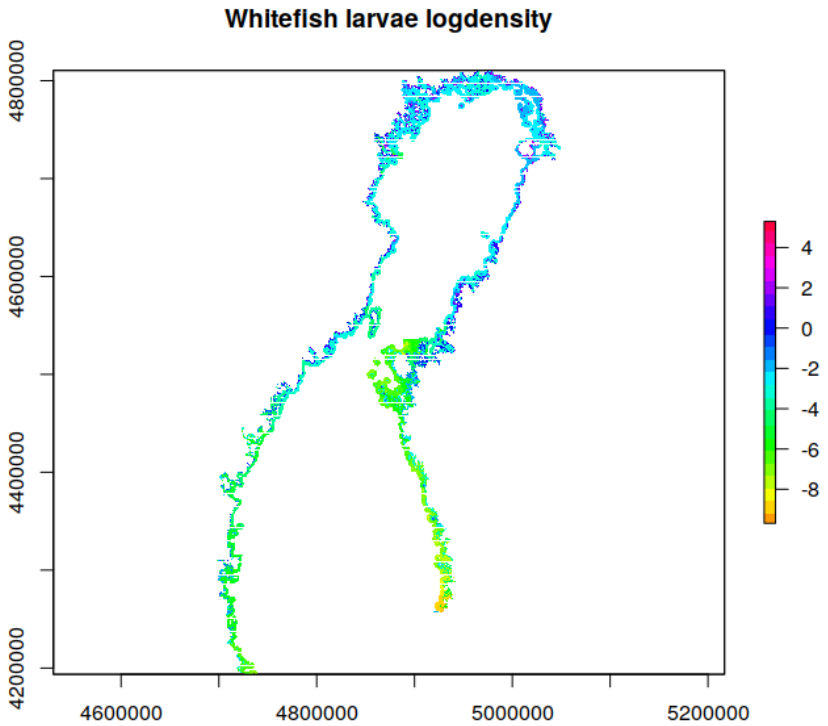
$$\mathbf{K}_{\tilde{\mathbf{f}},\mathbf{f}} = \tilde{\mathbf{X}}(\tilde{\mathbf{S}})\mathbf{\Sigma}_\beta\mathbf{X}(\mathbf{S}) + \mathbf{K}_{\phi,\tilde{\phi}}$$

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

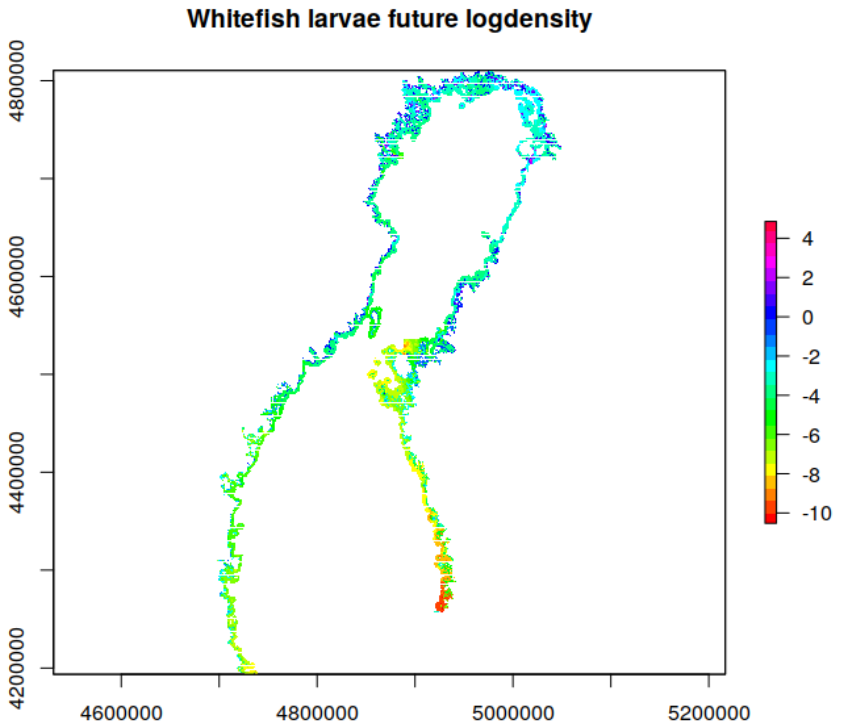
and

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

Whitefish predictive logdensity

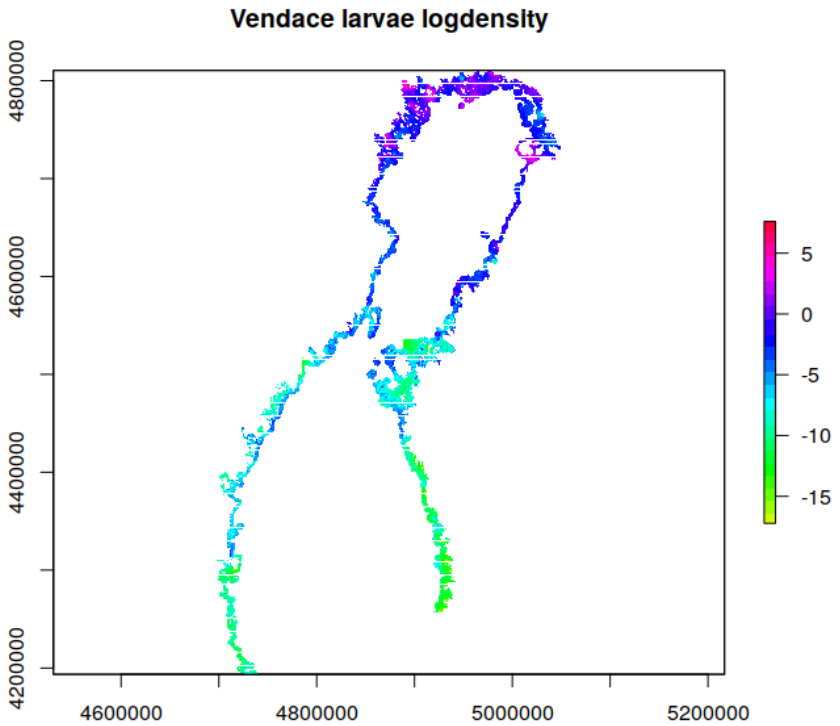


(a) Predicted whitefish larvae logdensity, current situation

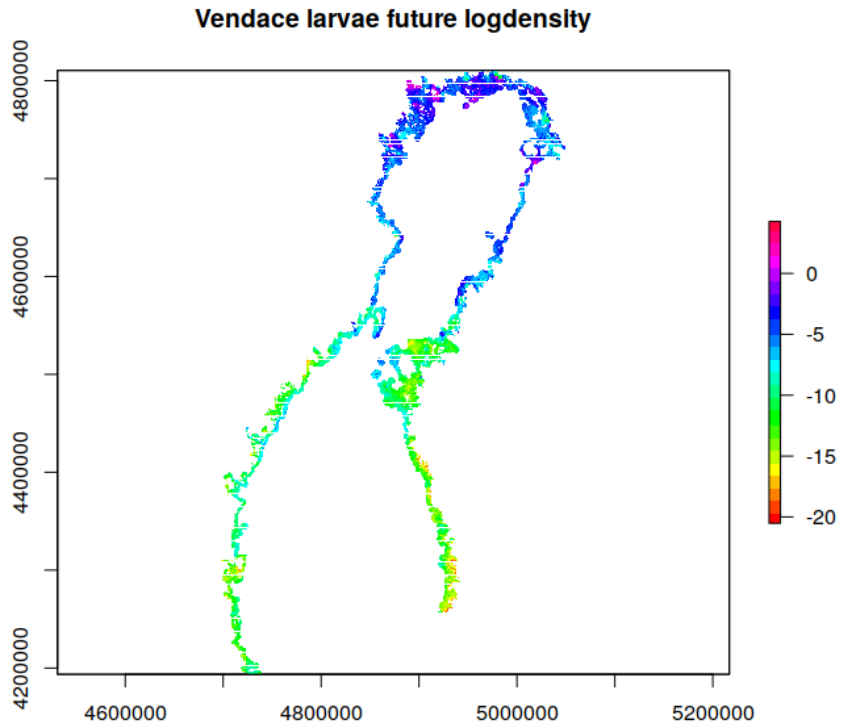


(b) Predicted whitefish larvae logdensity for years 2049-2059

Vendace predictive logdensity



(a) Predicted vendace larvae logdensity, current situation



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

Conclusion and further extension

The analyses presented in this work, are preliminary and aimed to be an initial investigation on the subject. Nevertheless, the final results are generally in line with the one obtained by previous studies. Possible extensions:

- non-parametric response curves for both species
- use zero-inflated or hurdle models
- more flexible parameterization of the Negative Binomial distribution: $\sigma^2 = \omega\mu + \theta\mu^2$
- take a causal statistics approach to account for spatiotemporal autocorrelation among the covariates
- account for species to species interaction by implementing a joint SDM

Thank you!