

# 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
  - The data
  - Target and covariates
- 2 Model
  - HSDM
  - Random effects
  - Covariance function and prior setting
- 3 Posterior analysis and conclusions
  - Posterior distributions
  - Response curve
  - Predictive logdensity

# Thesis objectives

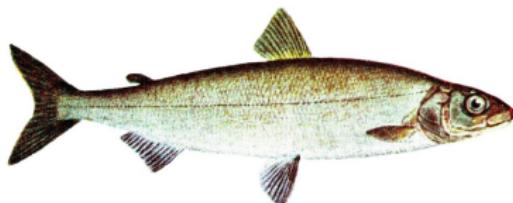
The applied goals of the thesis are twofold:

- 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.

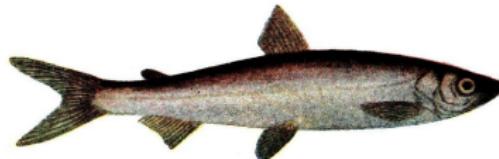
## Study area

The Gulf of Bothnia (GoB) is a brackish water basin between Sweden and Finland, in the northern part of the Baltic Sea. It 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.

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*).



(a) Whitefish

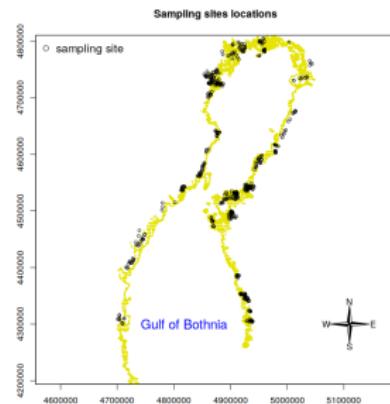


(b) Vendace

## Whitefish dataset

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

larvae abundance data were measured at 653 different sampling sites in 26 sub-areas along the Finnish and Swedish coastal region of the GoB, during 2009-2011.



Environmental 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.

## SmartSea data

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<sup>1</sup>. SmartSea Gulf of Bothnia as Resource for Sustainable Growth.

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 are used for years 2006-2059.

---

<sup>1</sup><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, salinity and ice coverage

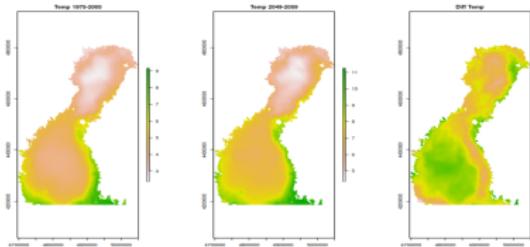


Figure: Temperature

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.

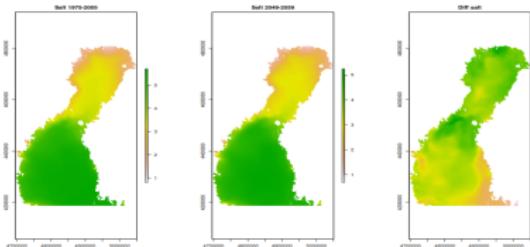


Figure: Salinity

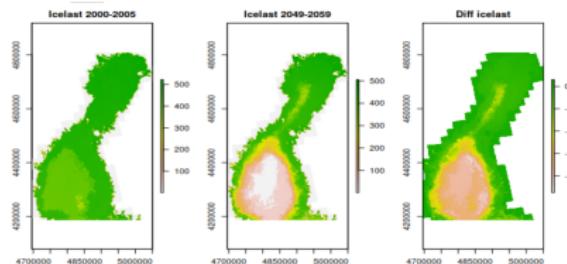


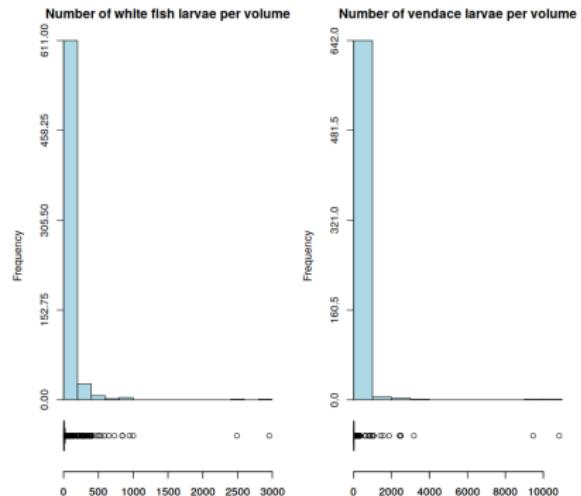
Figure: Last day of ice coverage

## 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/VOLUME 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	<p>Bottom type classification, a categorical variable with classes</p> <p>0 = not shallow,      1 = open water,      2 = other,      3 = sand,      4 = sand/mud,      5 = sand/stone</p>	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

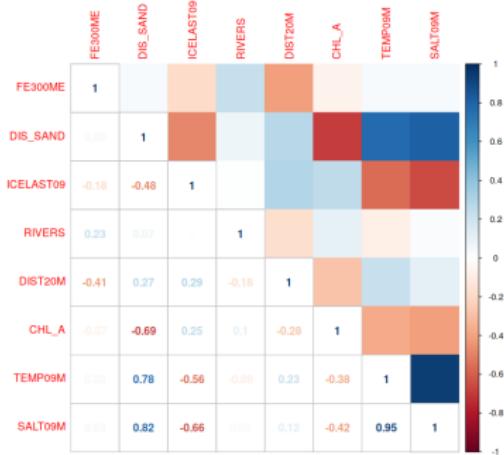


Figure: Correlation plot of the continuous covariates

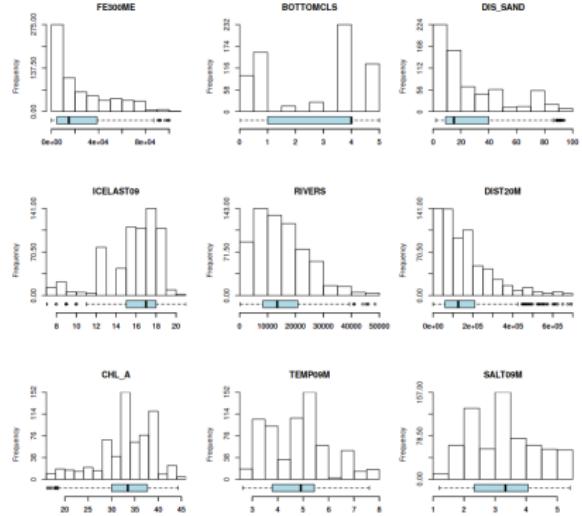


Figure: Training covariates distribution.

## 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.

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.

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.

# Hierarchical spatial models

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

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

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

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

# Larvae density SDM

Let  $\mathbf{y}$  represents the number of whitefishes or vendaces caught at the various sampling locations. Let  $i \in 1, \dots, N$

$$y_i | f_i \sim \text{Poisson}(V_i e^{f(\mathbf{x}_i, \mathbf{s}_i)})$$

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

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

## Spatial random effect

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

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

which implies the marginal distribution for any  $\mathbf{f} = [\mathbf{f}(\mathbf{s}_1), \dots, \mathbf{f}(\mathbf{s}_n)]$  is again Gaussian:

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

where  $\mathbf{X}(\mathbf{S})^\top = [\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n)]$ ,  $\Sigma_\phi = \text{Cov}(\phi, \phi)$  and  $\phi = [\phi(\mathbf{s}_1), \dots, \phi(\mathbf{s}_n)]^\top$ .

## Additional iid random effect

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

where  $\epsilon_i$  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  $\text{Var}[\epsilon_i] = 1/r$ . This yields to

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

## Covariance function

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

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

where  $l$  and  $\sigma_\phi^2$  are the range and intensity parameters of the covariance function. This 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

$$\beta_d \sim N(0, 10) \quad \forall d \in 1, \dots, 15$$

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

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

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. The code to reproduce the analysis and the results presented in this thesis, is available at the following github repository :  
<https://github.com/ilapia/thesis-pia>

To improve the efficiency of the algorithm, we 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  $\mathbf{f}$ .

```

transformed data {
  vector[N] mu;
  nmatrix[N, N] Dlst_spatial;
  nmatrix[N, N] Sigma_llin;
  ...
  // linear covariance
  Sigma_llin[i, j] = 10 * dot_product(x[i],x[j]); // off-diagonal elements
  Sigma_llin[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 {
  nmatrix[N, N] Sigma;
  nmatrix[N, N] L;
}

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

// 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;

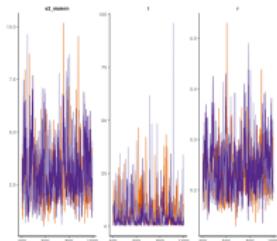
for (n in 1:N)
  y[n] ~ neg_binomial(V[n]*exp(ff[n]), r);
}

generated quantities {
  vector[N] f;
  // derived quantity (transform)
  f = L*z;
}

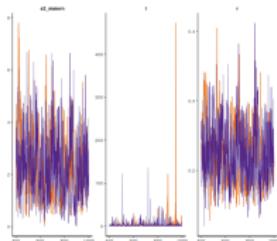
```

# Convergence

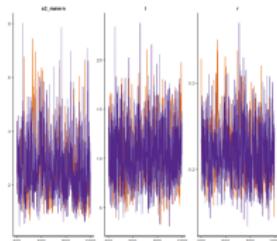
In the first stages of the analysis, we work on a subset of the data. For each of the target we consider four models: two linear models and two models with additional quadratic effect.



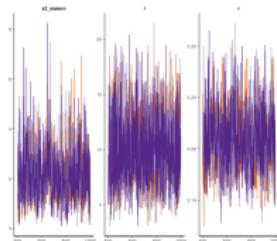
(a) whitefish,  
student-t prior,  
linear mod.



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



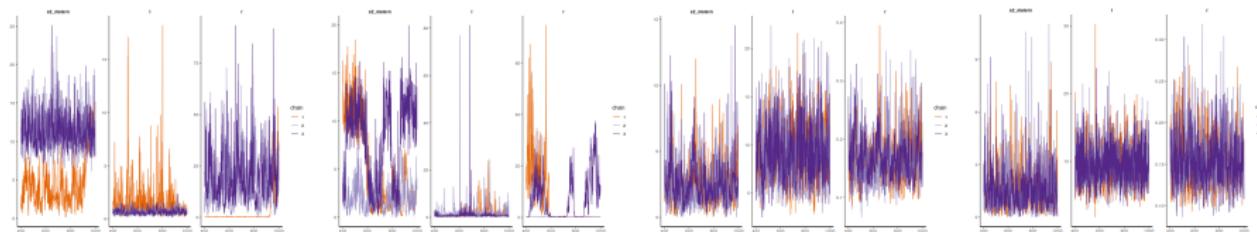
(c) whitefish,  
gamma prior, linear  
mod.



(d) whitefish,  
gamma prior, quadratic  
mod.

Figure: 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, we set as  $\ell$  prior a distribution which favors much shorter length-scale, we can observe clear problems in chain mixing. This confusion between  $\ell$  and  $r$ , and the different roles of the two random effects present in the model, lead to not uniquely identifiable results.



(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: Vendace trace plots for the four models considered.

## Model comparison

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

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

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

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

## Linear weights posterior

We examine the posterior distributions of the model parameters. We first sample  $f$  and the other model parameters from their posterior, using Stan. We then recover Monte Carlo approximations for  $\beta$  by sampling from the conditional

$$\beta | \mathbf{f}, \mathbf{X}, \sigma^2, \Sigma_\phi \sim N(\mu_p, \Sigma_p) \quad (5)$$

where

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

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

# Models' parameters posteriors

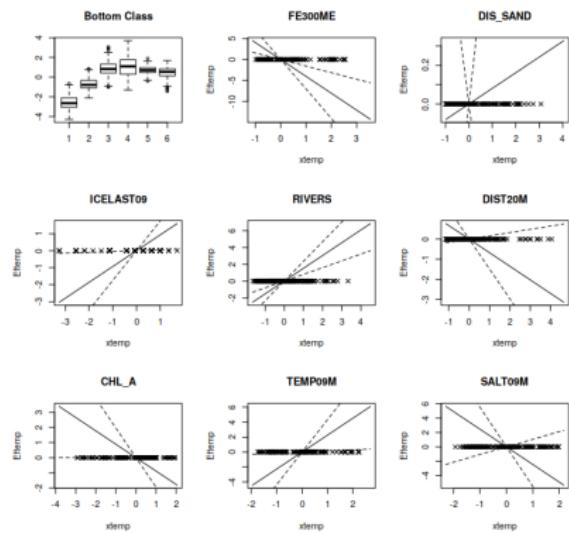
Covariate	Mean	Stand.Dev	quant2.5%	quant97.5%	Covariate	Mean	Stand.Dev	quant2.5%	quant97.5%
s2_matern	2.982	0.084	1.278	6.115	s2_matern	2.446	0.087	1.654	6.693
I	5.647	0.414	6.787	23.965	I	9.145	0.101	3.103	15.834
r	0.239	0.003	0.047	0.356	r	0.167	0.002	0.042	0.274
intercept	-0.282	3.222	-6.148	6.117	intercept	-4.142	2.908	-9.470	1.333
BOTTOMCLS_0	-2.600	0.677	-4.004	-1.117	BOTTOMCLS_0	-3.659	0.933	-5.784	-1.944
BOTTOMCLS_1	-0.680	0.549	-1.536	0.426	BOTTOMCLS_1	-2.397	0.823	-3.984	-1.004
BOTTOMCLS_2	0.854	0.725	-0.613	2.390	BOTTOMCLS_2	0.077	0.686	-1.249	1.369
BOTTOMCLS_3	1.079	1.030	-0.642	3.109	BOTTOMCLS_3	1.129	1.780	-2.105	4.385
BOTTOMCLS_4	0.708	0.413	-0.140	1.482	BOTTOMCLS_4	-0.022	0.496	-0.961	1.112
BOTTOMCLS_5	0.458	0.557	-0.850	1.366	BOTTOMCLS_5	0.744	0.569	-0.328	1.862
FE300ME	-2.335	2.091	-8.420	0.009	FE300ME	-4.015	1.772	-7.862	-1.277
DIS_SAND	-0.359	0.355	-0.998	0.291	DIS_SAND	0.081	0.604	-1.094	1.210
ICELAST09	0.304	0.375	-0.421	1.038	ICELAST09	0.932	0.521	0.004	1.960
RIVERS	0.166	0.348	-0.524	0.822	RIVERS	1.522	0.485	0.648	2.514
DIST20M	-0.155	0.394	-0.896	0.528	DIST20M	-0.677	0.533	-1.673	0.273
CHL_A	-0.438	0.594	-1.802	0.421	CHL_A	-0.885	0.585	-2.281	0.139
TEMP09M	0.644	1.153	-2.054	2.746	TEMP09M	2.266	1.421	-0.228	5.385
SALT09M	0.579	1.954	-1.700	6.279	SALT09M	-2.447	1.896	-5.791	1.461

**Table:** Posterior distribution of model parameters for Whitefish (left) and Vendace (right) SDMs

# Whitefish HSDM response curve

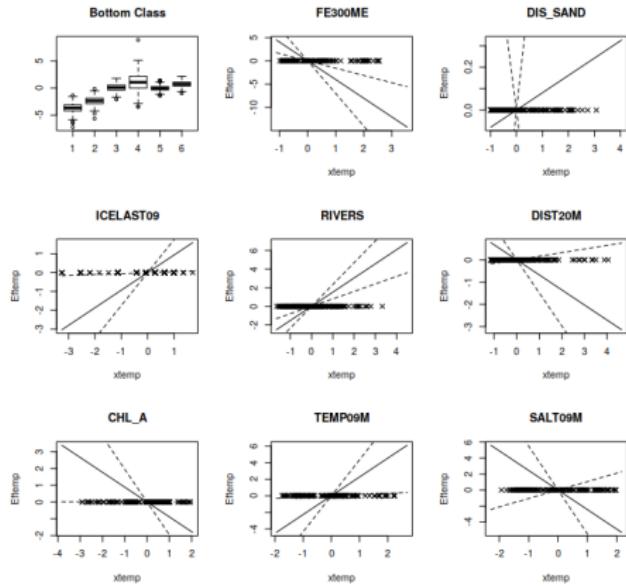
Whitefish spawning is favored by high temperature and low salinity conditions. The near presence of rivers, and longer ice coverage of the waters also benefits their reproduction, while high rate of eutrophication and the exposure to wind prevent it.

Whitefishes spawns grow better in 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 HSDM response curve

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.



## 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

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

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

where

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

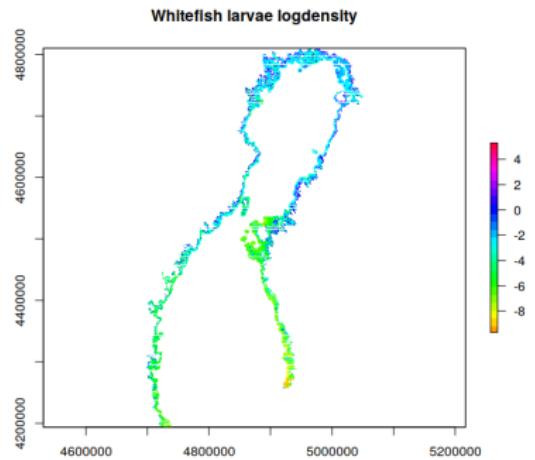
$$\mathbf{K}_{\tilde{\mathbf{f}}, \tilde{\mathbf{f}}} = \tilde{\mathbf{X}}(\tilde{\mathbf{S}}) \boldsymbol{\Sigma}_\beta \tilde{\mathbf{X}}(\tilde{\mathbf{S}}) + \mathbf{K}_{\tilde{\phi}, \tilde{\phi}}$$

and

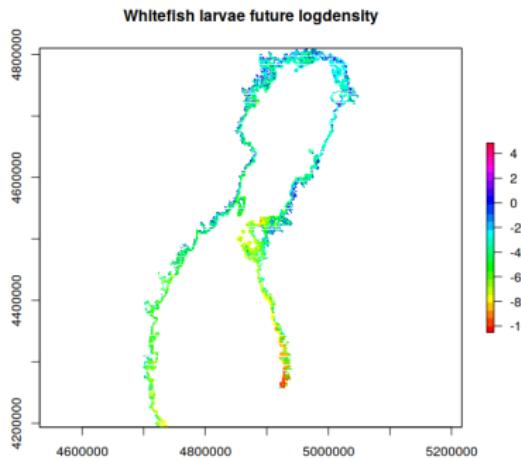
$$\mathbf{K}_{\mathbf{f}, \mathbf{f}} = \mathbf{X}(\mathbf{S}) \boldsymbol{\Sigma}_\beta \mathbf{X}(\mathbf{S}) + \mathbf{K}_{\phi, \phi}$$

# Whitefish predictive logdensity

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.



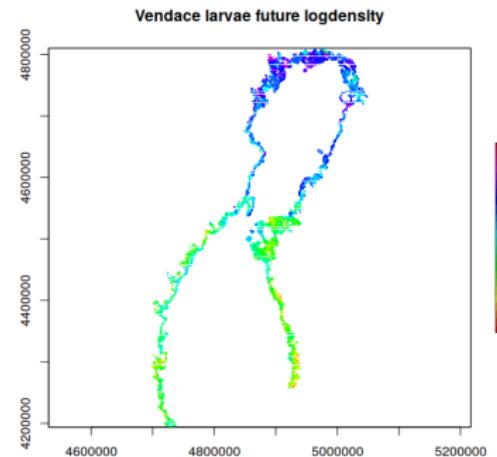
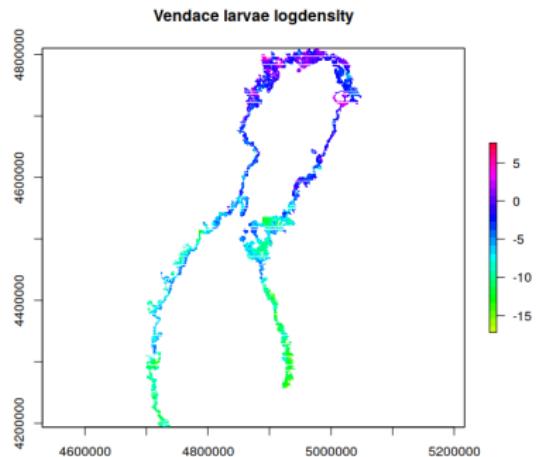
(a) Predicted whitefish larvae logdensity, current situation



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

## Vendace predictive logdensity

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.



(a) Predicted vendace larvae logdensity, current situation

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

# Thank you!

