Additive Multivariate Gaussian Processes for Joint Species Distribution Modeling with Heterogeneous Data

Jarno Vanhatalo*, Marcelo Hartmann†, and Lari Veneranta‡

Abstract. Species distribution models (SDM) are a key tool in ecology, conservation and management of natural resources. Two key components of the state-of-the-art SDMs are the description for species distribution response along environmental covariates and the spatial random effect that captures deviations from the distribution patterns explained by environmental covariates. Joint species distribution models (JSDMs) additionally include interspecific correlations which have been shown to improve their descriptive and predictive performance compared to single species models. However, current JSDMs are restricted to hierarchical generalized linear modeling framework. Their limitation is that parametric models have trouble in explaining changes in abundance due, for example, highly nonlinear physical tolerance limits which is particularly important when predicting species distribution in new areas or under scenarios of environmental change. On the other hand, semi-parametric response functions have been shown to improve the predictive performance of SDMs in these tasks in single species models.

Here, we propose JSDMs where the responses to environmental covariates are modeled with additive multivariate Gaussian processes coded as linear models of coregionalization. These allow inference for wide range of functional forms and interspecific correlations between the responses. We propose also an efficient approach for inference with Laplace approximation and parameterization of the interspecific covariance matrices on the Euclidean space. We demonstrate the benefits of our model with two small scale examples and one real world case study. We use cross-validation to compare the proposed model to analogous semi-parametric single species models and parametric single and joint species models in interpolation and extrapolation tasks. The proposed model outperforms the alternative models in all cases. We also show that the proposed model can be seen as an extension of the current state-of-the-art JSDMs to semi-parametric models.

MSC 2010 subject classifications: Primary 60G15, 60K35; secondary 62P12.

Keywords: linear model of coregionalization, hierarchical model, heterogeneous data, spatial prediction, model comparison, Laplace approximation, covariance transformation.

---

*Department of Mathematics and Statistics and Organismal and Evolutionary Biology Research Program, University of Helsinki, jarno.vanhatalo@helsinki.fi

†Department of Mathematics and Statistics and Department of Computer Science, University of Helsinki, marcelo.hartmann@helsinki.fi

‡Natural Resources Institute Finland, Finland, lari.veneranta@luke.fi

1 Introduction

Species distribution models (SDMs) are key tools in the ecologists' toolbox. They have been widely used, among other applications, to study species habitat preferences (Latimer et al., 2006), to improve identification and management of conservation areas and natural resources (Kallasvuo et al., 2017) and to evaluate species responses to environmental filtering under climate change scenarios (Clark et al., 2014; Kotta et al., 2019). The main goals of statistical inference in these contexts are to use species observations and information on the associated environment to infer the relationship between these two attributes and to predict over regions of unsampled locations to build thematic species distribution maps (Gelfand et al., 2006; Elith and Leathwich, 2009).

Species distribution modeling is, thus, directly related to inferring species' responses to environmental factors (Latimer et al., 2006). This is traditionally done using generalized linear or additive models as well as an array of machine learning approaches such as regression trees or maximum entropy modeling (Elith and Leathwich, 2009). These approaches 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 (Ovaskainen and Soininen, 2011; Hui et al., 2013; Clark et al., 2014; Thorson et al., 2015; Hartmann et al., 2017). For these reasons, joint species distribution models (JSDM) have gained increasing attention in recent years (Warton et al., 2015).

Dunstan et al. (2013) and Hui et al. (2013) introduced species archetype modeling where species' responses to the environment are clustered into few archetype models. Pollock et al. (2014) proposed to use the multivariate probit regression model (Chib and Greenberg, 1998) to describe geographical co-occurrence patterns between frogs and eucalyptus trees in Australia and Clark et al. (2014) built a JSDM to infer richness and loss of species under climate change scenarios. Thorson et al. (2015) introduced a spatial latent factor model to predict spatial distribution of breeding birds and rock fish communities and recently, Ovaskainen et al. (2017) introduced the hierarchical modeling of species communities (HMSC) framework which includes detailed description of interspecific correlations in covariate responses and spatial random effects.

Current JSDMs rely on the classical framework of hierarchical generalized linear models (GLMs). Even though this approach allows flexible modeling by describing the randomness of response variables with different probabilistic models (Nelder and Wedderburn, 1972), it is still limited by its parametric assumptions. Hence, it may fail to accurately describe a species' response to environmental conditions (Vanhatalo et al., 2012; Kotta et al., 2019). Here, we propose a semi-parametric JSDM represented with multivariate Gaussian processes (GPs). GPs are flexible semi-parametric regression models where the regression function is estimated without restrictive parametric assumptions about its form (O'Hagan, 1978; Rasmussen and Williams, 2006). Our model integrates the main strengths of semi-parametric single species models (Vanhatalo et al., 2012; Golding and Purse, 2016) and generalized linear model based JSDMs.

First, we model the species responses to environmental covariates with additive multivariate GPs. This allows us to capture wide range of nonlinear responses and share information about these responses between species. Second, due to the hierarchical model structure, the model can simultaneously accommodate several kinds of outcome variables (observations/measurements) by assigning different types of probabilistic models for them. This is important, since it allows us to exploit different types of measurements which commonly arise in real-case scenarios of multiple species surveys. Our presentation focuses in the mostly used probabilistic models in practice, the Bernoulli (Binomial) and Poisson with over-dispersion (Negative-Binomial). We also propose an efficient computational approach build around Laplace method (Vanhatalo et al., 2010; Golding and Purse, 2016). Lastly, we present a structured cross-validation scheme to validate and compare models' performance in different kinds of prediction tasks.

This paper is organized as follows. In Section 2 we introduce a motivating case study. In Section 3, we introduce the additive multivariate GPs and in Section 4 we discuss its predictive properties and introduce the computational methods for inference. In Section 5, we illustrate the basic properties of the model through two simple examples and introduce the specific case study model. In Section 6 we present the case study results and we end by discussion and conclusions in Section 7.

2 Motivating case study: coastal species distributions in the Gulf of Bothnia

As a motivating example, we study spring distribution of four fish species and three types of algae or macro-vegetation on the coastal region of the Gulf of Bothnia in the northern Baltic Sea. The Gulf of Bothnia is a brackish water basin between Sweden and Finland 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 (Veneranta et al., 2013; Kallasvuo et al., 2017). Coastal zones are also the most sensitive regions of the Baltic sea to both natural variation and anthropogenic pressures (Reusch et al., 2018). Hence, these areas are of central importance for conservation and there is need for detailed knowledge on the distribution of coastal species and predictions concerning their response to environmental changes.

2.1 Case study species and data

The studied fish species are the juvenile or adult three-spined stickleback (*Gasterosteus aculeatus*) and nine-spined stickleback (*Pungitius pungitius*) as well as larvae of whitefish (*Coregonus lavaretus*) and vendace (*Coregonus albula*). Both whitefish and vendace are commercially important species and the sticklebacks are one key fish fauna in the Gulf of Bothnia ecosystem (Bergström et al., 2015). We treat the studied vegetation and algae in functional group level comprising of diatomous algae, filamentous and macro-vegetation. These are the dominating groups of benthic vegetation in shores where larvae of Coregonids (whitefish and vendace) occur at early developmental stages. In the scale of Gulf of Bothnia, their occurrence also reflects the salinity, nutrient balance and wind exposure of studied area (Veneranta et al., 2013).
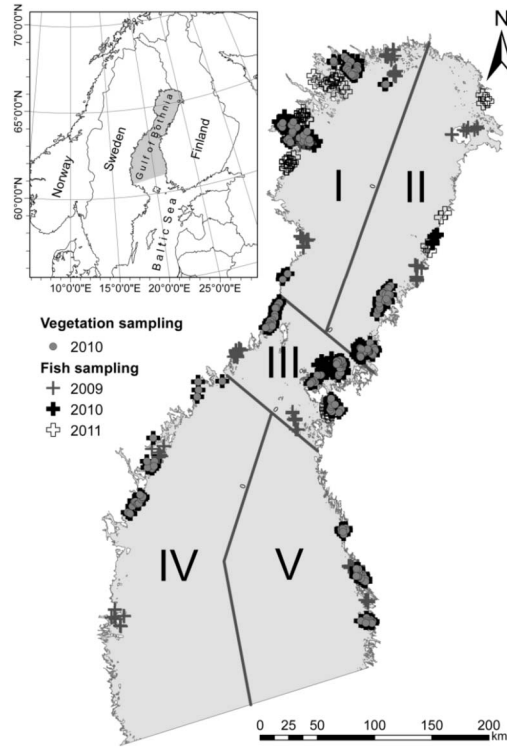
Figure 1: The study region, Gulf of Bothnia, and locations for species data. The region is divided into five subregions (I–V) to be used in cross-validation. The environmental conditions (described by the environmental covariates) are heterogeneous between these regions (Veneranta et al., 2013) which allows to test models' extrapolation performance.

The case study species reflect the large scale environmental gradients and the changes in the Baltic Sea environment in last decades. Sticklebacks in the Baltic Sea use the shallow coastal zone for reproduction (Bergström et al., 2015) and high abundances of sticklebacks in the Baltic Sea have been positively correlated with high occurrence of filamentous algae (Eriksson et al., 2009, 2012). Coregonids in coastal areas prefer more oligotrophic waters (Veneranta et al., 2013). Whitefish and vendace reproduce in stony reefs and islets of Baltic Sea in late autumn and the larvae hatch at ice break-up in spring. In sheltered areas the overwintering reeds (*Phragmites australis*) dominate the macro-vegetation in spring. Diatomous algae consist of several epiphytic species that have an early spring bloom at ice break up and settle rapidly over bottom when light and water temperature increase (Busse and Snoeijs, 2002). Filamentous algae in this study consist mostly of *Pilayella sp.* It is a fast growing annual species that dominates the exposed shores at Baltic Sea in early spring (Rönnberg and Bonsdorff, 2004).

The whole study area was divided into sampling subareas (Figure 1). Within each sub-area, data were collected at several sampling sites distributed so that they covered

Variable	Description	Resolution [m]
Openness	The average openness and exposure to winds	300
Distance to deep	Distance to 20 m depth curve	200
Sandy bottom index	Area weighted distance to the sandy shores	90
Ice breakup week	The end date (weeks) of ice cover in 2009	1852
Chlorophyll-a	Chlorophyll-a concentration	1852
River	Distance to the nearest river mouth	150
Bottom class	Bottom classification to 6 categories	200
Winter salinity	Length of ice winter	10,000

Table 1: Environmental covariates used in the case study and their original resolution. For this study, the resolution of all the environmental covariates was scaled to 300 m.

the range of most important environmental covariates in that sub-area. Sampling was conducted in 2009–2011. At each site sampling was done approximately one week after the ice break. In the scale of Gulf of Bothnia, the ice break-up happens in a period of approximately one month. The species data used in this work comprise of 160 distinct sampling sites. In 70 sites (2010 data) all species were sampled and the rest of the sites comprise samples for the fishes only (Figure 1). Fish samples comprise the number of caught fish together with information on the sampling effort at each site. The effort was measured as the volume of sampled water (m^3). A more detailed description of the data collection is provided by Veneranta et al. (2013). For plant data, the bottom was photographed at 5–13 locations at depth of 30 cm within a distance of 2 m and parallel to shore line. The occurrence of a species was recorded at 16 uniformly distributed points in all these photographs. The case study plants can grow over each others so presence of one plant at a spatial location does not exclude the presence of other plants. The whitefish and vendace data were previously analyzed by Veneranta et al. (2013) and Vanhatalo et al. (2012) but the data on other species are previously unpublished.

2.2 Environmental covariates

Gulf of Bothnia 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 Gulf of Bothnia, there is a gradient in river influence, salinity, temperature and length of ice cover period from north to south. We used seven real-valued and one categorical abiotic environmental covariates that were available throughout the study area as raster maps. Each species observation was combined with covariates from the raster cell where the sample location fell in. These are summarized in Table 1 and described and motivated in detail by Veneranta et al. (2013). Due its fresh water origin vendace is *a priori* sensitive to even small changes in salinity levels experienced in the Gulf of Bothnia whereas the other species respond to salinity only in the Baltic Sea scale. Hence, we used all covariates only for vendace but excluded the winter salinity from other species.

3   Hierarchical multivariate species distribution model

We start our model building following the generic hierarchical structure similar to the one presented by, e.g., Wikle (2003), Cressie and Wikle (2011) and Banerjee et al. (2015)

$$
\begin{aligned}
[\text{data}|\text{process}, \text{parameters}] &: \ \pi_{\boldsymbol{Y}}(\boldsymbol{y}(\mathbf{x},\mathbf{s})|\mathbf{f}(\mathbf{x},\mathbf{s}),\boldsymbol{\eta}), \\
[\text{process}|\text{parameters}] &: \ \pi_{\boldsymbol{f}}(\mathbf{f}(\mathbf{x},\mathbf{s})|\boldsymbol{\theta}), \\
[\text{parameters}] &: \ \pi(\boldsymbol{\eta},\boldsymbol{\theta}),
\end{aligned}
\tag{3.1}
$$

where the first layer of hierarchy is the probabilistic model which defines the conditional distribution for multivariate data $\boldsymbol{Y}(\mathbf{x},\mathbf{s})$, at spatial location $\mathbf{s}$ with associated covariates $\mathbf{x}$, given the model parameters $\boldsymbol{\eta}$ and the multivariate latent process $\boldsymbol{f}(\mathbf{x},\mathbf{s})$. The second layer defines the prior distribution for the latent process given the process parameters $\boldsymbol{\theta}$, and the third layer defines the prior distribution for all unknown parameters.

Let $j \in \mathcal{J} = \{1,\cdots,J\}$ be the species index set and J the total number of species in the study. Denote by $\boldsymbol{Y}(\mathbf{x},\mathbf{s})^T = [Y_1 \cdots Y_J]$ the J-variate random vector with components $Y_j = Y_j(\mathbf{x}_j,\mathbf{s})$ related to the jth species at spatial location $\mathbf{s}^T = [s_1 \ s_2]$ (coordinates) under the influence of environmental covariates $\mathbf{x}_j^T = [x_{j,1}\cdots x_{j,c_j}]$ where c_j is the number of environmental covariates for the species j. We denote by $\mathbf{x}$ the full vector of covariates and $\mathcal{X}$ the covariate space. The species specific covariates $\mathbf{x}_j$ are sub-vectors of $\mathbf{x}$ such that $\mathbf{x}_j \in \mathcal{X}_j$ where $\mathcal{X}_j \subset \mathcal{X}$ and $\mathcal{X}_j$ is c_j-dimensional. Here, we assume that given a multivariate latent process $\boldsymbol{f}(\mathbf{x},\mathbf{s}) = [f_1(\mathbf{x}_1,\mathbf{s})\cdots f_J(\mathbf{x}_J,\mathbf{s})]^T : \mathcal{X}\times\mathbb{R}^2 \to \mathbb{R}^J$, the probabilistic model for $\boldsymbol{Y} = \boldsymbol{Y}(\mathbf{x},\mathbf{s})$ factorizes as follows

$$
\pi_{\boldsymbol{Y}}(\boldsymbol{y}(\mathbf{x},\mathbf{s})|\mathbf{f}(\mathbf{x},\mathbf{s}),\boldsymbol{\eta}) = \prod_{j=1}^{J} \pi_{Y_j}(y_j|\mathrm{f}_j,\eta_j),
\tag{3.2}
$$

where $\mathrm{f}_j = \mathrm{f}_j(\mathbf{x}_j,\mathbf{s})$ is fixed but an unknown realization of the j'th process with covariates $\mathbf{x}_j$ at the spatial location $\mathbf{s}$. η_j is the vector of parameters of the probabilistic model π_{Y_j} for the species j and $\boldsymbol{\eta} = [\eta_1 \cdots \eta_J]^T$. In general, any probabilistic model for data could be used and the observation models should be chosen according to the assumed sampling process. Here, we consider in detail the Binomial and the Negative-Binomial (over-dispersed Poisson) models used in our case study (Section 2). These models can be seen as observation models resulting from inhomogeneous point process model for species abundance (Gelfand et al., 2006; Warton and Shepherd, 2010).

In practice, each sampling site consists of a small area where the observations are made. In our case study, the area covered by a sampling site is so small compared to the whole study region that the latent function is practically constant in each sampling site. The model for the count of species presences at uniformly distributed points in a sampling site (plants in our case study) is then Binomial given by (Gelfand et al., 2006)

$$
\pi_{Y_B}(y|f(\mathbf{x},\mathbf{s}),z_B(\mathbf{s})) = \binom{z_B(\mathbf{s})}{y} p(f(\mathbf{x},\mathbf{s}))^y [1-p(f(\mathbf{x},\mathbf{s}))]^{z_B(\mathbf{s})-y} I_{\{0,\ldots,z_B(\mathbf{s})\}}(y), \tag{3.3}
$$

where $z_B(\mathbf{s})$ is the total number of observation points in the sampling site at location $\mathbf{s}$ and $p(f(\mathbf{x},\mathbf{s}))$ is the success probability, which will be modeled through the logit here.

The natural model for the number of individuals in a closed area or volume in a sampling site (fish in our case study) is Poisson which we extend to over-dispersed Poisson using the Negative-Binomial given by (see Liu and Vanhatalo, 2018, for details)

$$\pi_{Y_N}(y|f(\mathbf{x},\mathbf{s}), z_N(\mathbf{s}), r) = \frac{\Gamma(r+y)}{y!\Gamma(r)} \left(\frac{r}{r+m(\mathbf{x},\mathbf{s})} \right)^r \left(\frac{m(\mathbf{x},\mathbf{s})}{r+m(\mathbf{x},\mathbf{s})} \right)^y I_{\mathbb{N}_0}(y), \quad (3.4)$$

where $m(\mathbf{x},\mathbf{s}) = z_N(\mathbf{s})\exp[f(\mathbf{x},\mathbf{s})] = E(Y_N|f(\mathbf{x},\mathbf{s}), z_N(\mathbf{x}), r)$ and r is the over-dispersion parameter. The latent process $f(\cdot)$ corresponds to the logarithm of a species (relative) density and $z_N(\mathbf{s})$ is the sampled volume of water in the sampling site at location $\mathbf{s}$. For this parameterization $V(Y_N|f(\mathbf{x},\mathbf{s}), z_N(\mathbf{x}), r) = m(\mathbf{x},\mathbf{s}) + m(\mathbf{x},\mathbf{s})^2/r$. Increasing r decreases the variance and when $r \to \infty$, the model approaches the Poisson distribution.

3.1 Univariate additive latent Gaussian process

First, we assume that marginally for any species j, the process model is given by

$$f_j(\mathbf{x}_j, \mathbf{s}) = \beta_{j,0} + h_j(\mathbf{x}_j) + \varepsilon_j(\mathbf{s}), \quad (3.5)$$

where $\beta_{j,0}$ is the offset weight with distribution $\beta_{j,0}|\sigma_{j,0}^2 \sim \mathcal{N}(0, \sigma_{j,0}^2)$ and $h_j : \mathcal{X}_j \to \mathbb{R}$ is a *predictor function* of environmental covariates. The predictor function is assumed to be additive over the covariates, $h_j(\mathbf{x}_j) = \sum_{r=1}^{c_j} h_{j,r}(x_{j,r})$, and each additive term is given an independent zero mean GP prior, so that

$$h_j(\mathbf{x}_j)|\boldsymbol{\theta}_j \sim \mathcal{GP}\left(0, \sum_{r=1}^{c_j} k_{h_{j,r}}(x_{j,r}, x'_{j,r}; \boldsymbol{\theta}_{j,r})\right), \quad (3.6)$$

where $k_{h_{j,r}}(x_{j,r}, x'_{j,r}; \boldsymbol{\theta}_{j,r}) = \mathrm{Cov}(h_{j,r}(x_{j,r}), h_{j,r}(x'_{j,r})|\boldsymbol{\theta}_{j,r})$ is a covariance function with parameter $\boldsymbol{\theta}_{j,r}$ and $\boldsymbol{\theta}_j = [\boldsymbol{\theta}_{j,1}^T \cdots \boldsymbol{\theta}_{j,c_j}^T]^T$. For example, with $k_{h_{j,r}}(x_{j,r}, x'_{j,r}; \theta_{j,r}) = x_{j,r} x'_{j,r}\sigma_{j,r}^2$ and $\theta_{j,r} = \sigma_{j,r}^2$, the predictor function corresponds to linear model $h_{j,r}(x_{j,r}) = x_{j,r}\beta_{j,r}$ where $\beta_{j,r} \sim N(0, \sigma_{j,r}^2)$ (Rasmussen and Williams, 2006). With other choices of covariance functions we can model non-linear responses in which case the model can be seen as an alternative to the traditional generalized additive models (GAMs, Hastie and Tibishirani, 1986). The general form of an additive GP prior would include also joint effects of covariates (Duvenaud et al., 2011) but this is not considered here.

The term $\varepsilon_j(\mathbf{s})$ is a spatial GP,

$$\varepsilon_j(\mathbf{s})|\sigma_j^2, \boldsymbol{\ell}_j \sim \mathcal{GP}\left(0, k_{\epsilon_j}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\ell}_j, \sigma_j^2)\right), \quad (3.7)$$

where $k_{\epsilon_j}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\ell}_j, \sigma_j^2)$ is a spatial covariance function with variance σ_j^2 and range parameter $\boldsymbol{\ell}_j$. When we consider independent models for each species, that is the traditional single species SDMs (SSDMs), the processes $h_{j,r}$ and ϵ_j are mutually independent among all species. The model outlined this far is similar to the GP based species distribution models used by, e.g., Vanhatalo et al. (2012), Kallasvuo et al. (2017) and Golding and Purse (2016). Next, we introduce models which consider dependency.

3.2   Additive multivariate Gaussian process priors

In order to account for possible species interdependence, we first include spatial species interaction into the model through the linear model of coregionalization (LMC) (Mardia and Goodall, 1993; Gelfand et al., 2004). Write $\boldsymbol{\varepsilon}(\mathbf{s})^T = [\varepsilon_1(\mathbf{s}) \cdots \varepsilon_J(\mathbf{s})]$ and assume that $\boldsymbol{\varepsilon}(\mathbf{s})$ has LMC covariance structure with species specific correlation functions $\tilde{k}_{\epsilon_j}(\cdot, \cdot\,; \boldsymbol{\ell}_j)$. The covariance structure of the LMC with interspecific spatial dependence is then,

$$\text{Cov}\left(\varepsilon_j(\mathbf{s}), \varepsilon_{j'}(\mathbf{s}') | \Sigma_\epsilon, \boldsymbol{\ell}\right) = \sum_{l=1}^{J} u_{\epsilon,l}(j,j') \tilde{k}_{\epsilon_l}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\ell}_l) \tag{3.8}$$

with $\boldsymbol{\ell}^T = [\boldsymbol{\ell}_1^T \cdots \boldsymbol{\ell}_J^T]$ and $u_{\epsilon,l}(j,j')$ the entry (j,j') of $U_{\epsilon,l} = L_{\epsilon,l} L_{\epsilon,l}^T$ where $L_{\epsilon,l}$ is the lth column of the Cholesky decomposition L_ϵ of the coregionalization matrix Σ_ϵ, that is matrix of interspecific correlations between spatial GP. Hence, the vector-valued latent process $\boldsymbol{f}(\mathbf{x}, \mathbf{s})^T = [f_1(\mathbf{x}_1, \mathbf{s}) \cdots f_J(\mathbf{x}_J, \mathbf{s})]$ unconditional on $\beta_{1,0}, \ldots, \beta_{J,0}$ follows a multivariate GP which we denote as

$$\boldsymbol{f}(\mathbf{x}, \mathbf{s}) | \Lambda_1 \sim \mathcal{MGP}\left(\mathbf{0}, \Sigma_0 + \sum_{r=1}^{c} k_{h_r}(\mathbf{x}_r, \mathbf{x}_r'; \boldsymbol{\theta}) + \sum_{j=1}^{J} \tilde{k}_{\epsilon_j}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\ell}_j) U_{\epsilon,j}\right), \tag{3.9}$$

where $\Lambda_1 = (\Sigma_0, \Sigma_\epsilon, \boldsymbol{\theta}, \boldsymbol{\ell})$, $\Sigma_0 = \text{diag}(\sigma_{1,0}^2, \ldots, \sigma_{J,0}^2)$, $\boldsymbol{\theta}^T = [\boldsymbol{\theta}_1^T \cdots \boldsymbol{\theta}_J^T]$ and $k_{h_r}(\mathbf{x}_r, \mathbf{x}_r'; \boldsymbol{\theta}) = \text{diag}\left(k_{h_{1,r}}(x_{1,r}, x_{1,r}'; \boldsymbol{\theta}_{1,r}), \cdots, k_{h_{J,r}}(x_{J,r}, x_{J,r}'; \boldsymbol{\theta}_{J,r})\right)$. If the predictor functions, $h_{j,r}$, were linear, the prior (3.9) would correspond to traditional multivariate spatial model with independent linear effects and spatial LMC (Gelfand et al., 2004).

We extend (3.9) to an *additive multivariate GP prior* where the dependence between the species specific additive predictor functions is modeled with LMC. We demonstrate this with a model where the set of covariates is equal for all species, that is, $\dim(\mathcal{X}_j) = c$ and $\mathbf{x}_j = \mathbf{x}$ for all j. Then the model is written as

$$\boldsymbol{f}(\mathbf{x}, \mathbf{s}) | \Lambda_2 \sim \mathcal{MGP}\left(\mathbf{0}, \Sigma_0 + \sum_{r=1}^{c} \sum_{j=1}^{J} \tilde{k}_{h_{j,r}}(x_r, x_r'; \boldsymbol{\theta}_{j,r}) U_{h_r,j} + \sum_{j=1}^{J} \tilde{k}_{\epsilon_j}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\ell}_j) U_{\epsilon,j}\right), \tag{3.10}$$

where $\Lambda_2 = (\Lambda_1, \Sigma_{h_1}, \ldots, \Sigma_{h_c})$ and Σ_{h_r} is the interspecific covariance matrix between the species specific predictor functions of r'th covariate, $\tilde{k}_{h_{j,r}}(\cdot, \cdot; \boldsymbol{\theta}_{j,r})$ is a correlation function related to the predictor function $h_{j,r}$, $U_{h_r,j} = L_{h_r,j} L_{h_r,j}^T$ and $L_{h_r,j}$ is the jth column of the Cholesky decomposition L_{h_r} of Σ_{h_r}. When the set of covariates is not the same for all species, covariate specific predictive functions are correlated only between species that share those same covariates.

A JSDM with multivariate GP prior (3.10) can be interpreted as extension of GP based SSDMs (Vanhatalo et al., 2012; Golding and Purse, 2016) to joint species modeling similarly as done with the generalized linear model based JSDMs (Pollock et al., 2014; Ovaskainen et al., 2017). The enhanced inferential ability of the multivariate additive GP compared to univariate GP models lies in its capability to infer similarity

(dissimilarity) in species specific responses to covariates and in the spatial random effect. The similarity/dissimilarity of responses of two species j and j' along a covariate r is indicated by a positive/negative value for correlation and hence, examining the LMC covariance matrices, Σ_{h_r}, can provide new insight to species to species associations.

3.3 Marginally uniform priors for correlation parameters

Here, we define the prior for coregionalization covariance matrices but define the covariance functions and rest of the priors in Section 5. A standard choice for prior for correlation matrices is the inverse Wishart distribution. However, if there is no prior information on interspecific correlations, we follow Hartmann et al. (2017) and suggest to use marginally uniform priors for the LMC covariance matrices Σ_ϵ and Σ_{h_r}. These can be achieved by the distribution of Barnard et al. (2000) and Tokuda et al. (2012). Let $\mathcal{P}$ be a correlation matrix with elements $\rho_{j,j'}$. A prior distribution that is marginally noninformative for the correlations, that is, the marginal distribution for every $\rho_{j,j'}$ is uniform over $(-1,1)$, is achieved with the distribution

$$\pi(\mathcal{P}\,|v) = \frac{\Gamma(\frac{v}{2})^J}{\Gamma_J(\frac{v}{2})}|\,\mathcal{P}\,|^{\frac{1}{2}(v-1)(J-1)-1} \prod_{j=1}^{J} |\,\mathcal{P}_{jj}\,|^{-\frac{v}{2}} \mathbb{1}_{(0,\infty)}(\det \mathcal{P}), \qquad (3.11)$$

when $v = J - 1$. Here, $\Gamma_J(\cdot)$ is the multivariate gamma function and matrix $\mathcal{P}_{jj}$ is obtained by removing the jth column and j^{th} row of $\mathcal{P}$. When v increases, the probability density (3.11) becomes increasingly concentrated around the origin.

4 Posterior inference and predictive properties

Given a set of species observations at locations $\mathbf{s}_{i_j}, i_j = 1,\ldots,n_j$, respectively for each species $j = 1,\ldots,J$, the likelihood can be written as

$$\pi(\mathbf{y}|\mathbf{f},\boldsymbol{\eta}) = \prod_{j=1}^{J} \prod_{i_j=1}^{n_j} \pi_j(y_{j,i_j}|\mathrm{f}_{j,i_j},\eta_j), \qquad (4.1)$$

where y_{j,i_j} is the i_j'th observation of species j at the i_j'th spatial location $\mathbf{s}_{i_j}^T$ associated with a set of covariates $\mathbf{x}_{i_j} \in \mathcal{X}_j$. The observed vector $\mathbf{y} = [\mathbf{y}_1^T \cdots \mathbf{y}_J^T]^T$ with $\mathbf{y}_j^T = [y_{j,1} \cdots y_{j,n_j}]$ collects all the species observations and $\mathbf{f} = [\mathbf{f}_1^T \ldots \mathbf{f}_J^T]^T$, where $\mathbf{f}_j^T = [\mathrm{f}_{j,1} \cdots \mathrm{f}_{j,n_j}]$, collects the corresponding latent variables. Hence, the likelihood factorizes over the latent variables and the inference can be done similarly as with univariate GP models. Markov chain Monte Carlo (MCMC) would provide exact answers in the limit of large sample size but deterministic approximations such as Laplace approximation or expectation propagation algorithm have also been shown to provide accurate approximate inference for univariate GP models with much lower computational time (Nickisch and Rasmussen, 2008; Rue and Marino, 2009; Vanhatalo et al., 2010). In order to study the properties of the model, we will examine the Laplace approximation for the posterior of latent variables conditional on the hyperparameters.

4.1 Posterior predictive inference conditional on hyperparameters

Posterior of latent variables

The Laplace's method approximates the conditional posterior of the latent function values $\mathbf{f}_* = \mathbf{f}(\mathbf{s}_*, \mathbf{x}_*)$ at the spatial location $\mathbf{s}_*$ with covariates $\mathbf{x}_*$ as[1]

$$\pi(\mathbf{f}_*|\mathbf{y},\boldsymbol{\eta},\Lambda) \approx \mathcal{N}\big(\mathbf{f}_*|C_{*,\mathbf{f}}C^{-1}\hat{\mathbf{f}}, C_* - C_{*,\mathbf{f}}(C^{-1}+\mathrm{W})^{-1}C_{\mathbf{f},*}\big), \qquad (4.2)$$

where $\hat{\mathbf{f}}$ is the (conditional) maximum a posterior (MAP) estimate of latent variables,

$$\hat{\mathbf{f}} = \underset{\mathbf{f}\in\mathbb{R}^{\sum_j n_j}}{\arg\max} \sum_{j=1}^{J}\sum_{i_j=1}^{n_j} \log\pi_j(y_{j,i_j}|\mathrm{f}_{j,i_j},\eta_j) + \log\mathcal{N}(\mathbf{f}|\mathbf{0},C) \qquad (4.3)$$

and W is the Hessian matrix of the negative log-likelihood function evaluated at $\hat{\mathbf{f}}$. Here, C is the prior covariance matrix of $\mathbf{f}$, $C_{*,\mathbf{f}}$ is the (prior) covariance matrix between elements of $\mathbf{f}_*$ and $\mathbf{f}$. C_* is the prior covariance of $\mathbf{f}_*$. In case of multivariate additive GP (3.10) the prior covariance matrix is given by

$$C = \begin{bmatrix} \sigma_0^2\mathrm{J}_{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_0^2\mathrm{J}_{n_J} \end{bmatrix}$$

$$+ \sum_{r=1}^{c}\sum_{j=1}^{J} \begin{bmatrix} u_{h_r,j}(1,1)\{\tilde{K}_{h_{j,r}}\}_{1,1} & \cdots & u_{h_r,j}(1,J)\{\tilde{K}_{h_{j,r}}\}_{1,J} \\ \vdots & \ddots & \vdots \\ u_{h_r,j}(J,1)\{\tilde{K}_{h_{j,r}}\}_{J,1} & \cdots & u_{h_r,j}(J,J)\{\tilde{K}_{h_{j,r}}\}_{J,J} \end{bmatrix}$$

$$+ \sum_{j=1}^{J} \begin{bmatrix} u_{\epsilon,j}(1,1)\{\tilde{K}_{\epsilon_j}\}_{1,1} & \cdots & u_{\epsilon,j}(1,J)\{\tilde{K}_{\epsilon_j}\}_{1,J} \\ \vdots & \ddots & \vdots \\ u_{\epsilon,j}(J,1)\{\tilde{K}_{\epsilon_j}\}_{J,1} & \cdots & u_{\epsilon,j}(J,J)\{\tilde{K}_{\epsilon_j}\}_{J,J} \end{bmatrix}, \qquad (4.4)$$

where J_n is an $n\times n$ matrix full of ones $u_{\epsilon,l}(j,j')$ and $u_{h_r,l}(j,j')$ are the entry (j,j') of $U_{\epsilon,l}$ and $U_{h_r,l}$ (see (3.8)), and $\{\tilde{K}_{\epsilon_l}\}_{j,j'}$ and $\{\tilde{K}_{h_{l,r}}\}_{j,j'}$ are correlation matrices with elements $[\{\tilde{K}_{\epsilon_l}\}_{j,j'}]_{i_j,i_{j'}} = \tilde{k}_{\epsilon,l}(\mathbf{s}_{i_j},\mathbf{s}_{i_{j'}};\boldsymbol{\ell}_l)$ and $[\{\tilde{K}_{h_{l,r}}\}_{j,j'}]_{i_j,i_{j'}} = \tilde{k}_{h_{l,r}}(x_{i_j,r},x_{i_{j'},r}|\boldsymbol{\theta}_{l,r})$. The other correlation matrices are constructed similarly. If data on all species are available at all sampling locations, the covariance matrix reduces to Kronecker product similarly as in LMC models by Mardia and Goodall (1993) and Gelfand et al. (2004), so that $C = \Sigma_0 \otimes J_n + \sum_{r=1}^{c}\sum_{j=1}^{J} U_{h_{j,r},j} \otimes \tilde{K}_{h_{j,r}} + \sum_{j=1}^{J} U_{\epsilon,j} \otimes \tilde{K}_{\epsilon_j}$ where $n_j = n$ for all $j = 1,\ldots,J$. Since, in general, the second and third term of C are full matrices, it can be seen from (4.2) and (4.3) that the posterior of $\mathbf{f}$ and the posterior predictive distribution of $\mathbf{f}_*$ are affected by observations of all species at any spatial locations.

---

[1]In case of Gaussian observation model, this equals to the true posterior density of $\mathbf{f}|\mathbf{y},\boldsymbol{\eta},\Lambda$.

The marginal posterior of additive terms

Typically, we are also interested in species' responses along covariates encoded by the additive predictor functions. Their marginal posteriors, conditional on hyperparameters, are analogous to (4.2). The matrices C_* and $C_{*,\mathbf{f}}$ are only constructed using the correlation functions corresponding to the latent function of interest. For example, to study the response of species j along covariate x_r we evaluate the posterior predictive distribution for $h_{j,r}(x_r)$ using (4.2) so that we replace $C_{*,\mathbf{f}}$

$$C_{h_{j,r},\mathbf{f}} = \sum_{l=1}^{J} \left[u_{h_r,l}(j,1)\{\tilde{K}_{h_{l,r}}\}_{j,1}, \cdots, u_{h_r,l}(j,J)\{\tilde{K}_{h_{l,r}}\}_{j,J} \right], \tag{4.5}$$

and similarly for C_*. However, in case of the traditional LMC (3.9) only the spatial random effects are correlated between species whereas the predictor functions are not. Hence, Σ_{h_r} is diagonal for all r so that $u_{h_r,l}(j,j') = \sigma_l^2$ if $l = j = j'$ and zero otherwise. The second terms of (4.4) is then block diagonal and only the j'th block column in (4.5) is non-zero for which reason there is no information exchange between species specific predictor functions. In case of univariate GP prior all the terms in (4.4) are block diagonal and (4.2) reduces to J independent Gaussian distributions with no information exchange among species at all. Hence, the predictive performance of the univariate GP and the two multivariate GP models are very different. This is illustrated in Section 5.

The (marginal) posterior expectation and variance for new observations

When predicting species occurrence probability or abundance, we need to marginalize over the posterior of the latent variables. The posterior expectation and variance for a new outcome for species j at a location $(\mathbf{x}_*, \mathbf{s}_*)$ conditional on hyperparameters are

$$\begin{aligned}
\mathbb{E}[Y_j(\mathbf{x}_*, \mathbf{s}_*)|\mathbf{y}, \boldsymbol{\eta}, \Lambda] &= \mathbb{E}[\mathbb{E}(Y_j(\mathbf{x}_*, \mathbf{s}_*)|f_j(\mathbf{x}_*, \mathbf{s}_*), \mathbf{y}, \boldsymbol{\eta}, \Lambda)] \\
\mathbb{V}[Y_j(\mathbf{x}_*, \mathbf{s}_*)|\mathbf{y}, \boldsymbol{\eta}, \Lambda] &= \mathbb{E}[\mathbb{V}(Y_j(\mathbf{x}_*, \mathbf{s}_*)|f_j(\mathbf{x}_*, \mathbf{s}_*), \mathbf{y}, \boldsymbol{\eta}, \Lambda)] \\
&\qquad + \mathbb{V}[\mathbb{E}(Y_j(\mathbf{x}_*, \mathbf{s}_*)|f_j(\mathbf{x}_*, \mathbf{s}_*), \mathbf{y}, \boldsymbol{\eta}, \Lambda)].
\end{aligned} \tag{4.6}$$

When the probabilistic model for species j is assumed to be the Negative-Binomial or Binomial model with logistic link function (3.3) we can find either an exact result or an analytical approximation for these expectations and variances. These are given in the Supplementary material (Vanhatalo et al., 2019). These solutions speed up the predictive calculations considerably compared to numerically integrating f over $\mathbb{R}$.

Conditional scenario predictions

In some applications, one might be interested in scenario predictions conditional on changes in species composition. For example, one might be interested in how removing from or introducing specific species into an area would affect other species. In our case study setting, we could be interested in, for example, effects of management actions that would clean filamentous algae from the shoreline. Such scenario predictions would

be naturally tackled with predictive causal inference (Lindley, 2002; Pearl, 2009). In predictive causal inference the parameters of the model are inferred with the available data so far and predictions made considering alternative scenarios.

To illustrate this let's first introduce a short notation $\mathbf{y}_{\mathcal{J}_1,*} = \mathbf{y}_{\mathcal{J}_1}(\mathbf{x}_*, \mathbf{s}_*)$ and $\boldsymbol{f}_{\mathcal{J}_1,*} = \boldsymbol{f}_{\mathcal{J}_1}(\mathbf{x}_*, \mathbf{s}_*)$ where $\mathcal{J}_1$ denotes the set of species to be predicted and $\mathcal{J}_2$ the *scenario species* assumed to be "managed". For brevity, we omit the conditioning on hyperparameters and data. The conditional distribution of $\boldsymbol{Y}_{\mathcal{J}_1}(\mathbf{x}_*, \mathbf{s}_*)|\mathbf{y}_{\mathcal{J}_2}$ is

$$\pi(\mathbf{y}_{\mathcal{J}_1,*}|\mathbf{y}_{\mathcal{J}_2}) = \int \pi(\mathbf{y}_{\mathcal{J}_1,*}|\mathbf{y}_{\mathcal{J}_2}, \mathbf{f}_{\mathcal{J}_1,*})\pi(\mathbf{y}_{\mathcal{J}_2}, \mathbf{f}_{\mathcal{J}_1,*})\mathrm{d}\mathbf{f}_{\mathcal{J}_1,*}/\pi(\mathbf{y}_{\mathcal{J}_2}), \tag{4.7}$$

where $Y_{\mathcal{J}_1}(\mathbf{x}_*, \mathbf{s}_*)|\mathbf{y}_{\mathcal{J}_2}, \mathbf{f}_{\mathcal{J}_1,*}$ only depends on $\mathbf{f}_{\mathcal{J}_1,*}$. The Laplace approximation for this conditional predictive distribution is shown in Supplementary material.

4.2 Parameter inference

We used Laplace approximation (Vanhatalo et al., 2010) to approximate the conditional posterior of latent variables $\pi(\mathbf{f}|\mathbf{y}, \boldsymbol{\eta}, \Lambda)$ and the marginal likelihood of the hyperparameters $\pi(\mathbf{y}|\boldsymbol{\eta}, \Lambda) = \int \pi(\mathbf{y}|\mathbf{f}, \boldsymbol{\eta})\pi(\mathbf{f}|\Lambda)d\mathbf{f}$. We searched for the (approximate) maximum a posterior (MAP) estimate for hyperparameters given by

$$(\hat{\boldsymbol{\eta}}, \hat{\Lambda}) = \underset{\boldsymbol{\eta}, \Lambda}{\arg\max} \log q(\mathbf{y}|\boldsymbol{\eta}, \Lambda) + \log \pi(\boldsymbol{\eta}, \Lambda), \tag{4.8}$$

where $\log q(\mathbf{y}|\boldsymbol{\eta}, \Lambda)$ is the Laplace approximation for the log marginal likelihood for parameters. This approach produces also good approximation for the posterior predictive distribution for latent variables (Vanhatalo et al., 2010; Vehtari et al., 2016), which are the main interest in this work. Hence, we fixed hyperparameters to their MAP estimate.

In order to avoid constrained optimization, all the parameters were transferred to unconstrained space. We used log transformation for covariance function parameters, and for the interspecific correlation matrices we used the transformation presented by Kurowicka and Cooke (2003) and Lewandowski et al. (2009), which is a bijective mapping between the space of correlation matrices and $\mathbb{R}^{\binom{J}{2}}$. This is summarized in Supplementary material. We used scaled conjugate gradient optimization for locating the MAP and checked carefully that a (local) mode had really been found by verifying that gradients along all hyperparameters were zero. The required gradients of $\log q(\mathbf{y}|\boldsymbol{\eta}, \Lambda)$ were solved analytically as described by Rasmussen and Williams (2006) and Vanhatalo et al. (2010). The Supplementary material summarizes the derivatives w.r.t. to the covariance parameters in Σ_ϵ, Σ_{h_r}, $r = 1, \ldots, c$.

5 Experiments

Here, we first illustrate the properties of the hierarchical multivariate GP models with two simple examples. These examples highlight particular properties of the proposed model. After this we introduce the model and analysis for our case study (Section 2). We implemented all the models using the GPstuff package (Vanhatalo et al., 2013).
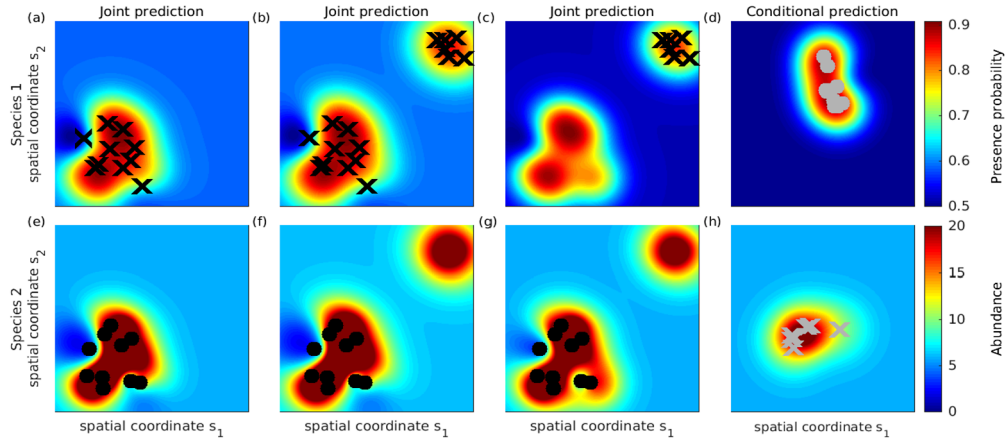
Figure 2: Illustration of multivariate GP prior for JSDM with two species and spatial only component. Crosses and dots are simulated data locations of species 1 (Binomial data) and species 2 (Negative-Binomial data). See Section 5.1 for details.

5.1 Demonstration with simulated spatial data

Figure 2 presents simulated data and posterior predictions for spatial modelling of two species. The model follows spatial LMC; that is model (3.9), where the covariate terms are dropped out. The spatial correlation function used in this demo is the Gaussian $k(\mathbf{s}, \mathbf{s}') = e^{-||\mathbf{s}-\mathbf{s}'||^2/2l^2}$, where $||\mathbf{s} - \mathbf{s}'||$ is the Euclidean distance and l is the length-scale. The first row of subplots shows the posterior predictive probability of observing species 1 ($E[Y_1]$, $Y_1(\mathbf{s}) \sim$ Binomial) and the second row shows the expected number of species 2 ($E[Y_2]$, $Y_2(\mathbf{s}) \sim$ Negative-Binomial). Plots (a) and (e) show the predictions when the species observations are from the same region but not from exactly the same locations. In this case, the prediction of multivariate GP is similar to predictions by univariate GPs. Plots (b) and (f) show the predictions when data are available on both species from the lower left corner and additionally data on species 1 is available from upper right corner. There is positive correlation between species, which has been inferred from the data in the lower left region, so the prediction for species 2 in upper right corner is informed by data on species 1 in that region. The last two columns illustrate the conditional scenario predictions (4.7). In the third column, the training data is the same as in the first column and the expected values in plots (c) and (g) show joint scenario prediction in new region of same size and shape. In this scenario, species in the new region were observed so that species 1 is known to be in the top right, and species 2 in the bottom left. The fourth column shows the conditional scenario prediction for the species separately so that the grey marks show the locations where the other species would be introduced in these scenarios.
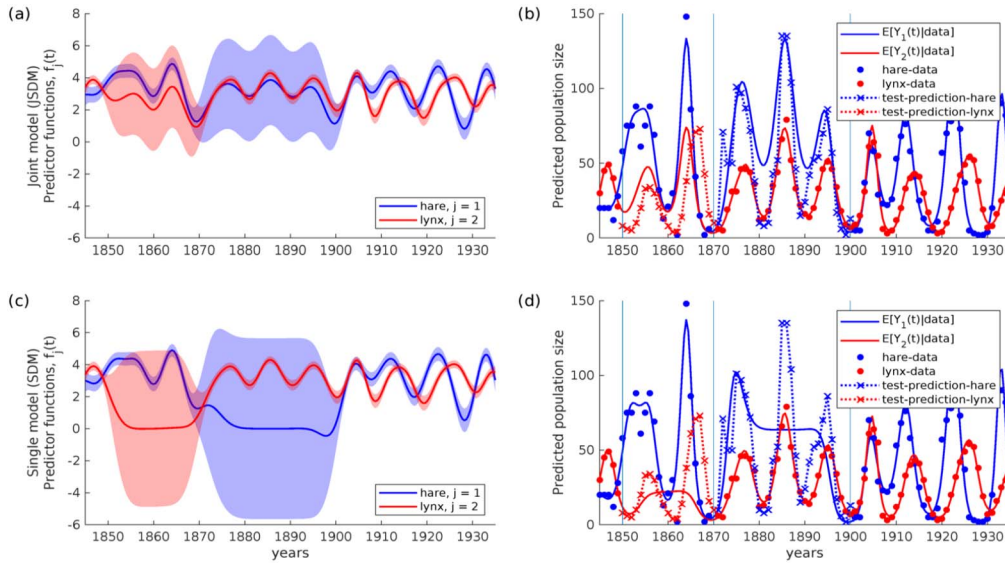
Figure 3: The results from the analysis of hare and lynx interaction (predator-prey system) analyzed with single and joint SDM. Plots a) and c) show the posterior mean and 95% credible interval of the latent functions and plots b) and d) show the training data, test data and posterior expectation of observations. See Section 5.2 for details.

5.2   Demonstration with time series of species abundances

Next, we consider the classical predator-prey system containing annual population counts of hare and lynx in the northern Canada from 1845 to 1935 (Odum, 1953). The data is not spatially explicit since the observations are total counts over the region so we model it as time series. This illustrates the behavior of individual additive covariate effect in the multivariate GP model (3.10) with time being the covariate. Let's denote by $Y_{1,t}$ and $Y_{2,t}$ the population sizes of hare and lynx at time t, and assume that $Y_{1,t}|f_1(t) \sim \text{Poisson}(\exp(f_1(t)))$ and $Y_{2,t}|f_2(t) \sim \text{Poisson}(\exp(f_2(t)))$. We compared two alternative priors for the latent functions, one with independent GP priors and another with joint bivariate GP prior with the LMC structure. In order to compare models' predictive performances when some species have not been observed, we removed parts of the original data from the training set, the period of 1870–1900 for hare and 1850–1870 for lynx. Figure 3 displays the result of the data analysis. The model with bivariate GP prior clearly outperforms the independent model since it predicts better the test data points in periods where training data was removed. Its predictions have also smaller uncertainty than the predictions of independent GPs in those periods of time.

5.3   The case study on coastal species distribution

Case study models

The plant data are modeled with the Binomial observation model (3.3) and the fish data are modeled with the Negative-Binomial observation model (3.4). In order to test the effect of different model components and the effect of semi-parametric response functions versus standard quadratic response functions we compare the following models:

1) additive univariate GP predictor functions and univ. spatial random effects (3.5),
2) additive univariate GP predictor functions and LMC spatial random effect (3.9),
3) additive multivariate GP predictor function and LMC spatial random effect (3.10),
4) additive univariate GP predictor functions only (model (3.5) without $\epsilon_j(\mathbf{s})$),
5) univariate spatial random effects only (model (3.5) without $h_j(\mathbf{x}_j)$),
6) additive univariate quadratic predictors and univariate spatial random effects (model (3.5) with $h_j(\mathbf{x}_j) = \mathbf{x}_j\beta_j$ and $\mathbf{x}_j$ includes the covariates and their squares),
7) additive univariate quadratic predictors and LMC spatial random effect (model (3.9) with $h_j(\mathbf{x}_j) = \mathbf{x}_j\beta_j$ and $\mathbf{x}_j$ includes the covariates and their squares),
8) additive multivariate quadratic predictors and LMC spatial random effects (model (3.10) with $h_j(\mathbf{x}_j) = \mathbf{x}_j\beta_j$ and $\mathbf{x}_j$ includes the covariates and their squares),
9) additive multivariate quadratic predictors only (model (3.10) without $\epsilon(\mathbf{s})$ with $h_j(\mathbf{x}_j) = \mathbf{x}_j\beta_j$ and $\mathbf{x}_j$ includes the covariates and their squares).

In each model, the spatial random effects were given the Mátern covariance function with $\nu = 3/2$ degrees of freedom $k_{\epsilon_j}(\mathbf{s}, \mathbf{s}'; \sigma_j^2, \boldsymbol{\ell}_j) = \sigma_j^2\big(1 + \sqrt{3}\ r_j(\mathbf{s}, \mathbf{s}')\big)e^{-\sqrt{3}r_j(\mathbf{s}, \mathbf{s}')}$ where σ_j^2 is the variance parameter, $\boldsymbol{\ell}_j^T = [\ell_{j,1}\ \ell_{j,2}]$ is the vector of length-scales and $r_j(\mathbf{s}, \mathbf{s}') = [(\mathbf{s} - \mathbf{s}')^T(\mathrm{diag}(\boldsymbol{\ell}_j)^2)^{-1}(\mathbf{s} - \mathbf{s}')]^{1/2}$. The continuous covariate effects in the additive GP models (models 1-4) were given the Gaussian correlation function $\tilde{k}_{h_j,r}(x_r, x_r') = e^{-||x_r - x_r'||^2/2l_{j,r}^2}$. The additive linear models (models 6–8) were coded as additive GPs with $\tilde{k}_{h_{j,r}}(x_r, x_r') = x_r x_r' \sigma_{j,r}^2$. Even though this is not a proper correlation function, when used in LMC it implies a generalized linear model with interspecific dependencies between weights, β_j, that are the key components in state-of-the-art parametric JSDMs. See Discussion for details. For the categorical covariate (Bottom class, Table 1) we used a delta function so that $\tilde{k}_{h_{j,r}}(x_r, x_r'; \boldsymbol{\theta}) = \sigma_{\delta_j}^2\delta_{x_r}(x_r')$, where $\delta_{x_r}(x_r') = 1$ if $x_r = x_r'$ and zero otherwise. This corresponds to having an own intercept for each category. We used the marginally uniform priors (Section 4.1) for the between species correlations and weakly informative priors for the rest of the parameters. The variance parameters were given Half-Student-$t(\mu = 0, \sigma^2 = 4, \nu = 4)$ priors and the length-scale parameters were given Half-Inverse-Student-$t(\mu = 0, \sigma^2 = 1, \nu = 4)$. Hence, *a priori* more weight is given for smooth functions with small variability.

Model validation and comparison

We aim to provide models that give reliable posterior predictions. Hence, it is natural to compare models with the goodness of their posterior predictive distributions. This can

be done efficiently with cross-validation (CV) using the log predictive density diagnostics (Vehtari and Ojanen, 2012). Let $\mathcal{D}$ denote the full data-set. Fix K disjoint sub-sets of $\mathcal{D}$, say $\mathcal{D}_1, \ldots, \mathcal{D}_K$, such that their union is $\mathcal{D}$. The K-fold CV log point-wise marginal predictive density is then $L_K(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \log \pi(y_i|\mathcal{D}_{\sim k(i)})$ where $\mathcal{D}_{\sim k(i)} = \{\cup_{r=1}^K \mathcal{D}_r : r \neq k(i)\}$ and $k(i)$ is such that $y_i \in \mathcal{D}_{k(i)}$. We compare the models with the leave-one-out CV (LOO-CV, $K = n$) and structured 5-fold CV. We used the MAP estimate for the hyperparameters. In 5-fold CV we found the MAP for each training set separately. The LOO-CV was conducted at the MAP found with full data and we used Laplace approximation to approximate the LOO-CV log predictive densities (Vehtari et al., 2016). Laplace approximation for LOO-CV is shown to work well in GP models and, since our data is rather large, leaving only one data point out of the training set has only negligible effect on the posterior of the hyperparameters (Vehtari et al., 2016).

The rationale for calculating both LOO-CV and 5-fold CV comparison is the following. Since multiple species were sampled at every sampling site and each of the sampling sites has other sites spatially nearby it, the LOO-CV log predictive densities are affected significantly by the spatial random effects. The LOO-CV, thus, measures models' interpolation performance which can be good even if the models were not able to represent well responses along covariates (predictor functions) (Vanhatalo et al., 2012; Veneranta et al., 2013). For this reason, we structured the 5-fold CV by dividing the data into five spatially distinct groups corresponding to regions I–V in Figure 1. The sampling sites in different groups are spatially so far from each others that the spatial random effects do not affect the posterior predictive distributions for test groups. Hence, the 5-fold CV tests mostly the extrapolation performance of a model, which is governed by the goodness of the predictor functions, whereas the LOO-CV tests the interpolation performance, which is governed also by the spatial random effects.

6 Results

6.1 Predictive performance of models

Table 2 summarizes the CV comparisons over all species and Tables 1–2 in Supplementary material show the models' predictive performance for each species separately. Tables 3–4 in Supplementary material show the pairwise comparison of the CV point wise log predictive densities between the best GP/parametric model (models 3 and 8) and the other GP/parametric models with both the covariate effects and spatial random effect (models 1–2 and 6–7). The results show that model 3, which includes multivariate GP predictors and multivariate spatial random effects (3.10), is the best in both LOO-CV and 5-fold CV with a difference of 10^{-2} or more to the other models. Since the mean log predictive density is an average of $n = 850$ point-wise predictions the difference of 10^{-2} corresponds to a difference of 8.5 in log (point-wise) joint predictive densities. Analogously to Bayes factors, which compare log prior predictive distribution, this can be considered a significant difference between two models (Kass and Raftery, 1995).

Hence, the multivariate GP (model 3) has significantly better overall predictive performance over all species than the other models. According to posterior predictive

GP models	LOO Cross Validation	5-fold Cross validation
1) Univariate GP $h(\mathbf{x})$, univ. $\epsilon(\mathbf{s})$	-2.230 (0.082)	-2.465 (0.094)
2) Univariate GP $h(\mathbf{x})$, multiv. $\epsilon(\mathbf{s})$	-2.081 (0.080)	-2.480 (0.100)
3) Multiv. GP $h(\mathbf{x})$, multiv. $\epsilon(\mathbf{s})$	**-1.669 (0.080)**	**-2.316 (0.086)**
4) Multiv. GP $h(\mathbf{x})$ only	-2.228(0.089)	-2.590(0.012)
5) Multiv. $\epsilon(\mathbf{s})$ only	-2.351(0.087)	-2.500(0.086)
6) Univ. quadr. $h(\mathbf{x})$, univ. $\epsilon(\mathbf{s})$	-2.262 (0.083)	-2.517 (0.095)
7) Univ. quadr. $h(\mathbf{x})$, multiv. $\epsilon(\mathbf{s})$	-2.117 (0.084)	-2.484 (0.109)
8) Multiv. quadr. $h(\mathbf{x})$, multiv. $\epsilon(\mathbf{s})$	-2.105 (0.084)	-2.416 (0.099)
9) Multiv. quadr. $h(\mathbf{x})$ only	-10.246(1.297)	-9.305(1.008)

Table 2: Model comparison with leave-one-out (LOO) and 5-fold cross validation using the mean point wise log marginal predictive density statistics (and its standard error) over all species. See Section 5.3 for model descriptions.

checks (Gelman et al., 2013) there are no serious discrepancies between its predictions and observed data. In extrapolation (5-fold CV), model 3 performs best also for all species separately. However, in interpolation (LOO-CV), it is not the best for all species separately (Tables 1–2 in Supplementary material). Moreover, the semi-parametric GP models (models 1–4) work better than the corresponding parametric models (quadratic environmental response, models 6–9) in both interpolation and extrapolation in general. Dropping either spatial random effect or covariate effect out from the model decreases its performance clearly. All models work better in interpolation than in extrapolation and compared to univariate models including interspecific correlations either in spatial random effects or also in environmental predictors improves models' performance in both of these tasks. Moreover, including interspecific correlations between environmental responses, $h_{j,r}(\mathbf{x}_r)$, improves the extrapolation performance relatively more than including interspecific correlations only for spatial random effects. Hence, multivariate spatial random effect improved more interpolation whereas multivariate predictors has larger effect for extrapolation as would intuitively be expected.

6.2 Effects of environmental covariates

The interspecific correlations (Figure 4) show that the responses to environment are similar among Coregonids and among sticklebacks whereas there are clear differences among these groups. These fish groups respond differently to sandy bottom and distance to deep. Moreover, there is strong positive spatial correlation among Coregonids and among sticklebacks but not between these groups. All species have negative spatial correlation with filamentous algae whereas there is either weak positive or negligible spatial correlation between fish species and macrovegetation and diatomous algae.

Figure 5 shows the marginal effects of the predictor functions for both univariate (SSDM) and multivariate (JSDM) GP models calculated as described in Section 4.1. In general, the responses in JSDM and SSDM are similar. In most of the cases the uncertainty in response function is smaller (narrower credible regions) in JSDM than in SSDM although the differences are not large. Most of the responses are smooth but
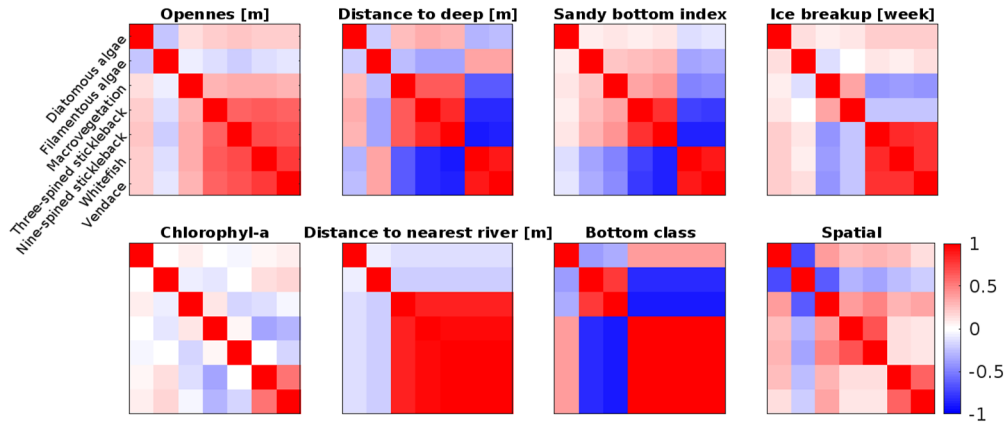
Figure 4: The maximum a posterior estimate of the correlation matrices in the multivariate additive GP model with Gaussian covariance function (model 3).
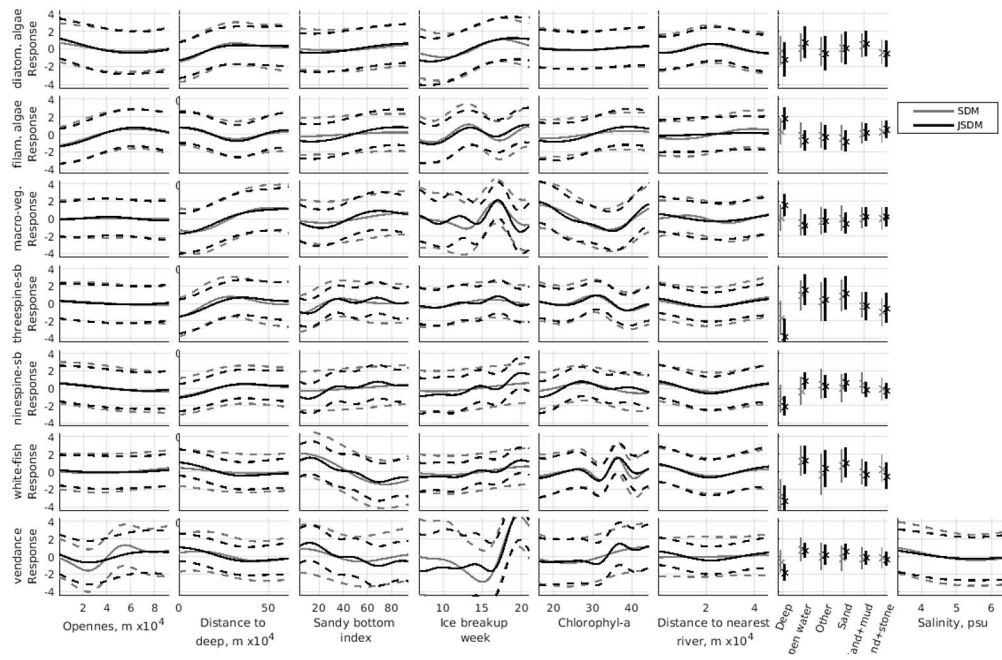

Figure 5: Marginal latent responses along covariates, $h_{j,r}$. Grey corresponds to single species GP model (model 1) and black to multivariate additive GP model (model 3).

show patterns that would be hard to capture with a quadratic function. For example, three and nine spine stickleback, macrovegetation and filamentous algae show logistic style response to either distance to deep or openness. Moreover, the response of vendace
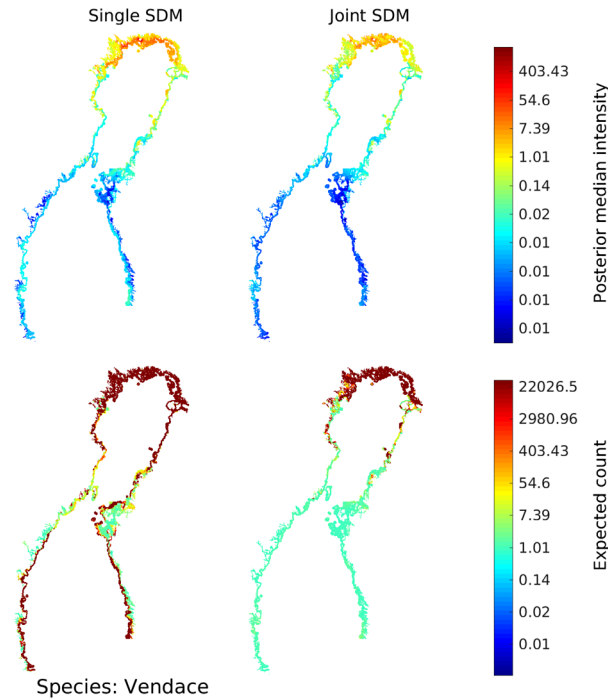
Figure 6: Posterior predictive median of intensity and the expected count of individuals in replicate sampling for vendace predicted by SSDM (model 1) and JSDM (model 3).

on ice break up week is first constant but has very steep increase after week 16. The responses to ice break up or chlorophyll-a show non-linear and non-quadratic responses also for white fish, macrovegetation and filamentous algae. The MAP estimates of other hyper-parameters are summarized in Supplementary material.

6.3   Spatial predictions

Figure 6 shows the distribution maps as posterior median of intensity and expected count of individuals in replicate sampling produced by SSDM (model 1) and JSDM (model 3) for vendace. The distribution maps for other species together with maps on posterior predictive variance are shown in Supplementary material. In broad scale, the posterior median of the intensity looks similar with SSDM and JSDM whereas SSDM predicts larger species counts than JSDM for all species throughout the study region and the uncertainty related to SSDM predictions is larger than that of JSDM predictions. Both SSDM and JSDM predict that vendace is distributed mostly in the northern Gulf of Bothnia. However, SSDM predicts somewhat higher median intensity than JSDM also in the southern Gulf of Bothnia. In relation to median intensity similar pattern that JSDM predicts more restricted distribution range than the single species model is seen also in the predictions concerning sticklebacks and whitefish. In case of diatomous and

filamentous algae SSDM and JSDM predict the distribution pattern similarly whereas for macrovegetation JSDM predicts slightly larger distribution ranges. The posterior distributions of spatial lenght-scale parameters were concentrated near one kilometer or less (see Table 5 in Supplementary material). Hence, the differences in distribution predictions cannot be explained by the spatial random effects over large areas but the spatial random effect explains local deviations from the predictions based on covariates.

7 Discussion and concluding remarks

7.1 Case study results

The GP based SDMs had better predictive performance than the parametric SDMs and JSDMs had better predictive performance than the corresponding SSDMs. The differences between SSDM and JSDM are most apparent in the distribution maps and predictor functions. The JSDM predicted in general smaller distribution ranges than SSDMs (macrovegetation being the only exception) and the posterior uncertainty in their predictions were smaller than in SSDMs.

When interpreting the results, it should be remembered that the sampling was targeted to larvae of sea-spawning Coregonids (whitefish and vendace) and planned to cover their plausible distribution range in terms of spatial (coastal areas of Gulf of Bothnia) and temporal extent (spring). Other species were sampled alongside Coregonids and the sampling area covers only limited portion of their full distribution range in the spring. Moreover, the abundance and distribution of all the studied species vary annually. Fish change their distribution areas seasonally and their larval stages last only few weeks. Vegetation and algal cover vary due changes in temperature and ice effects. For these reasons the results are most representative for larvae of whitefish and vendace. For other species the results describe their spring distribution in the shallow coastal regions only.

Our results correspond rather well to the earlier knowledge on the studied species. The responses to the environmental covariates and the interspecific correlations (Figures 5 and 4) indicate that whitefish and vendace larvae are, in general, distributed in different areas than sticklebacks. Most of the literature on Baltic Sea sticklebacks focus on three-spined sticklebacks whereas the nine-spined stickleback is not well studied. In early spring, stickleback abundances have been found to be highest in sheltered archipelago areas, where part of the population overwinter. High abundance of sticklebacks are typically thought to indicate structural complexity on the bottom; such as the presence of stones and boulders as well as reeds and other macrovegetation that function as shelter and provide food (Peltonen et al., 2004; Sieben et al., 2011). On contrary, highest densities of whitefish and vendace larvae are observed in open sandy shores near deep areas and in structurally simple shores, without macrovegetation, boulders and stones (Hudd et al., 1988; Leskelä et al., 1991; Veneranta et al., 2013). The distribution of vendace is highly influenced by ice break up week and salinity so that long ice cover period and low salinity increase their abundance. Ice winter is longer and salinity is lower in the northern than in the southern Gulf of Bothnia and thus it correlates positively to Coregonid presence (Vanhatalo et al., 2012). In general, optimal habitats for Coregonid larvae are mainly located in the northernmost Gulf of Bothnia.

Sticklebacks are known to prey mainly on mesozooplankton, but also on grazers (Peltonen et al., 2004; Sieben et al., 2011). Sticklebacks feed also on fish larvae if available (Byström et al., 2015), but there are no studies on potential predation risk to Coregonids. Based on our results, in the scale of coastal region of Gulf of Bothnia, sticklebacks are not thread to Coregonid larvae since their high density areas do not overlap. Moreover, there was no spatial correlation between sticklebacks and coregonids (Figure 4) which could indicate interspecific interaction of any kind. The presence of sticklebacks has been connected to higher eutrophication status and stickleback reproduction and growth benefit from increasing temperature and eutrophication (Candolin et al., 2008; Lefébure et al., 2011; Meier et al., 2012) On contrary, these environmental characteristics are assumed to affect negatively the Coregonid reproduction (Müller, 1992; Cingi et al., 2010; Vanhatalo et al., 2012; Veneranta et al., 2013). In our results whitefish and vendace larvae have positive response to Chlorophyl-a, which is a strong indicator for increasing eutrophication. However, in the scale of Gulf of Bothnia Chlorophyl-a concentration is typically higher near river mouths where salinity is lower and nutrient inflow high. Hence the result more likely reflects the high river influence through low salinity than preference for eutrophicated water.

The vegetation and algae distribution in our results reflect the nutrient status during winter and early spring as well as effect of ice cover and ice scraping in wind exposed shallow areas. Both the SSDMs and JSDMs indicate that filamentous algae occur in all coastal areas in high densities, except in some sheltered inner coastal areas. Macrovegetation and diatomous algae had highest presence probabilities at areas with lower filament presence probability. This pattern agrees well with the general understanding of these species groups. Filamentous algae are typically distributed in wind and wave exposed shores, that epiphytic diatoms and reeds that require shelter cannot tolerate. Filamentous algae can, however, grow over macrovegetation and it is possible that macrovegetation distribution is underestimated if filamentous algae growth over macrovegetation has hided macrovegetation from the sampling pictures. The higher nutrient levels in sheltered areas have been found to affect positively reed belt growth, and high abundances of macrovegetation species have been found from archipelago areas, lagoons, bays and river inlets (Pitkänen et al., 2013; Altartouri et al., 2014). The JSDM model reflects these smaller scale occurrences better than SSDM. In our results, diatomous algae are more common in estuaries and northern parts of the study area (see Supplementary material). This is likely explained by longer ice winter towards northern Gulf of Bothnia and longer distance to deep water.

7.2 Multivariate additive Gaussian process

Next we discuss some of the similarities and differences between our model and the existing JSDMs and highlight the novel methodological contributions in this paper.

Interspecific correlations

From the predictive point of view, the interspecific correlations in predictive functions are attractive for many reasons. In many applications of SDMs, the aim is to pre-

dict species distribution over regions that include locations spatially far from the data (Record et al., 2013) or to conduct scenario predictions related to, for example, climate change or land use (Guisan et al., 2013). In these applications, predictions are based on the responses to environmental covariates. The inclusion of interspecific dependence allows information flow between species which improves the estimates for predictive functions especially for species with only scarce data (Hui et al., 2013; Clark et al., 2014; Thorson et al., 2015; Ovaskainen et al., 2017).

We used the marginally uniform priors of (Barnard et al., 2000; Tokuda et al., 2012) for the interspecific correlations. This is justified by prior ignorance on correlations. Prior information on interspecific correlations could, however, be added into model with, e.g., informative inverse Wishart prior. With many species inferring the full covariance matrix is hard in which case we could use spatially dependent latent factors which induce for the spatial random effects a covariance structure $\text{Cov}\big(\epsilon_j(\mathbf{s}), \epsilon'_j(\mathbf{s}')\big) = \sum_{q=1}^{M} \lambda_{qj}\lambda_{qj'}k_q(\mathbf{s}, \mathbf{s}')$, where k_q is the q'th spatial covariance function and λ_{qj} the corresponding species specific factor loading. Here, $M < J$ so that this covariance structure is a low rank representation of LMC (Lopez, 2000; Lopez et al., 2008). Latent factor representation was first introduced to SDMs by Thorson et al. (2015) and it is used also in HMSC Ovaskainen et al. (2017). Recently Taylor-Rodríguez et al. (2017) proposed a clustering scheme where the species are clustered to less than J factor loading vectors.

Interspecific correlation between response functions has received less interest. Typically the response functions are assumed to be independent among species. Exception are species archetype models (Dunstan et al., 2013) and the HMSC framework of Ovaskainen et al. (2017). In the latter, the response functions are defined as $h_j(\mathbf{x}_j) = \mathbf{x}_j\theta_{j\cdot}$ where $\theta = [\theta_{1\cdot}^T, \ldots, \theta_{J\cdot}^T]^T$ is a $J \times p$ matrix of regression weights with hierarchical prior. The species specific weights are given Gaussian prior $\theta_{j\cdot} \sim N(\mu_{j\cdot}, V_j)$ where $\mu_{j\cdot}$ is the expected response of a species that can be common for all species, common within groups of species or modelled through species traits $\mu_{jr} = \tau_j^T\gamma_r$ where τ_j is a vector of trait values of species j and $\gamma_r \sim N(0, V_\gamma)$ are the effects of traits to response along covariate r. With $V = \sigma^2 I_{J \times J}$ and common prior mean $\mu_{j\cdot} = \mu \sim N(0, \sigma_\mu^2)$ for all $j \in 1, \ldots, J$, the induced covariance between species specific additive response functions is

$$\text{Cov}\left(h_{j,r}(x_r), h_{j',r}(x'_r)\right) = \mathbb{E}[\text{Cov}\left(x_r\theta_{jr}, x'_r\theta_{j'r}\right)] + \text{Cov}\left(\mathbb{E}[x_r\theta_{jr}], \mathbb{E}[x'_r\theta_{j'r}]\right)$$
$$= (\sigma_\mu^2 + \delta_j(j')\sigma^2)x_r x'_r$$

and with trait dependent prior mean $\text{Cov}(h_{j,r}(x_r), h_{j',r}(x'_r)) = (\tau_j^T V_\gamma \tau_{j'} + \delta_j(j')\sigma^2)x_r x'_r$. Hence, these alternatives have the same covariance structure as an additive multivariate GP (3.10) with $k_{h_j,r} = x_r x'_r$ for all $j = 1, \ldots, J$, and interspecific covariances $\Sigma_{h_r} = \sigma_\mu^2 + \delta_j(j')\sigma^2$ and $\Sigma_{h_r} = \tau_j^T V_\gamma \tau_{j'} + \delta_j(j')\sigma^2$. Hence, the hierarchical prior structure of HMSC could easily be extended to semiparametric models as well if we had trait information from our species. The benefit would be that these hierarchical model structures contain ecologically relevant prior information and, hence, using the induced interspecific covariance structure in the additive multivariate GP model would make it ecologically more realistic and potentially improve its predictive performance. The HMSC framework has many other features as well, such as phylogenetic relationships Ovaskainen et al. (2017), which could in principle be incorporated with our approach

as well. The structure of interspecific correlations is likely to be especially important in scenario based predictive analyses, such as climate change predictions.

Since our model does not make any mechanistic assumptions on species interdependencies the results concerning interspecific correlations have to be interpreted with care. For example, a positive correlation between spatial random effects could arize due mutualism, predator prey association or similar response to unobserved environmental covariates. Hence, interpreting these correlations should be done in light of more general ecological knowledge on the species. This is, however, a common challenge with all current JSDMs (Dunstan et al., 2013; Thorson et al., 2015; Taylor-Rodríguez et al., 2017; Ovaskainen et al., 2017). Moreover, the property of our model, as well as most other JSDM, is that if the training data shows positive or negative spatial correlation between two species this positive correlation is assumed to hold everywhere in the study domain. In many cases, this might be an unrealistic assumption for which reason one interesting future development would be to extend our models to allow species-to-species associations to depend on the environmental context (Fox and Dunson, 2015; Tikhonov et al., 2017). However, the colocalization patterns are explained by both environmental covariates and spatial random effects and regional differences in colocalization patterns can be explained by different environmental covariates in different regions.

The interspecific correlation in our approach are modelled on the latent variable level and, hence, are not directly comparable to interspecific correlations in data. Some authors have argued that this is a conceptual weakness of latent variable models since it makes interpretating their results hard and thus less transparent. An alternative would be to model the correlations in the observation error model (see, e.g., Ovaskainen et al., 2010; Pollock et al., 2014). Clark et al. (2016) proposed generalized joint attribute model (GJAM) to fix this interpretation challenge by aiming to model the correlations between species on the data scale. However, for example, in case of presence absence observations GJAM corresponds to multivariate probit model that is the marginal likelihood of our model in case of Bernoulli observation model. It is also unclear how to disentangle the process underlying the species occurrence and abundance from the observation process leading to the actual data in GJAM approach. Hence, despite the challenge of interpretating the correlations, we prefer the hierarchical latent variable modeling since it provides a coherent approach to separate these two processes.

Response functions

The response functions along environmental covariates provide basis to understand how environment affects species distribution and to predict the species distribution in new areas. Linear and other parametric models are efficient in finding overall trends in responses but have trouble in locating abrupt changes and change points due, for example, physical tolerance limits. Such limits, for example, along temperature and salinity are typical for large variety of taxa in aquatic domains (MacKenzie et al., 2007; Kotta et al., 2019). With SSDMs Vanhatalo et al. (2012), Shelton et al. (2014), Golding and Purse (2016) and Kallasvuo et al. (2017) have demonstrated the benefits of semiparametric models in such situations. The GP approach for modeling environmental responses is similar to generalized additive models (Guisan et al., 2002) but the latter have not

been implemented as JSDMs. The flexibility of semiparametric models provides also challenges compared to the more restricted parametric models. The prior for covariance function parameters has to be set with care (Golding and Purse, 2016) and we may also need to pose monotonicity constraints to the response functions Kotta et al. (2019) in order to prevent overfitting. Inferring the responses reliably requires typically more data compared to parametric models. The multivariate additive GP helps in these challenges as well since the interspecific correlations increase the effectively amount of data to be used to infer each response function and hence decreases the uncertainty in them.

Additive multivariate GP framework opens also new questions related to the choice of covariance functions and interspecific correlations. A typical choice for general purpose covariance function is a radial basis function. However, predictions using these covariance functions revert to prior predictive distributions when predicting beyond covariate range covered by data. Hence, in extrapolation tasks stationary covariance functions may not be the optimal choice (Vanhatalo et al., 2012; Kallasvuo et al., 2017) and combining the multivariate GP models with functional constraints (e.g. Kotta et al., 2019) could improve their predictive performance further.

Spatial random effects

Our model comparison results show clearly that including spatial random effects into model improves their predictive performance in both inter- and extrapolation. This result is well in line with earlier works demonstrating that environmental conditions alone may not sufficiently explain species distribution and spatial random effects improve their predictive performance (Latimer et al., 2006; Vanhatalo et al., 2012; Clark et al., 2014; Thorson et al., 2015; Kallasvuo et al., 2017). This is reasonable, since the distribution of species is shaped by the interplay of environmental covariates, stochastic processes and species interaction (Ovaskainen et al., 2017). Hence, a justified model should account for random processes and as demonstrated also by our results also to species interactions in these processes. However, adding spatial random effects into model may lead to problems in model identifibiality which we discuss in the next section.

Posterior inference and predictive performance

The JSDM built with multivariate additive GP had clearly the best overall predictive performance. Moreover, the uncertainty in the predictions by JSDMs were smaller than in the SSDMs which, together with the best log predictive density performance, indicates that the predictions were also more accurate. These findings have important implications to practical use of SDMs for management and other purposes. For example Kallasvuo et al. (2017) used SSDMs to classify coastal areas of Finland to unsuitable, suitable and highly productive spawning regions for four commercially important fish species. They based their estimates on the expected numbers of larvae in the coastal region which, as shown by Figure 6 is highly sensitive to the uncertainty of the predictions. Moreover, by using SSDMs we can overestimate the total biomass of all species (Clark et al., 2014).

An inherent challenge with the type of models considered here is the model identifibiality. Spatial random effects are known to affect the fixed effects estimates (Hodges and

Reich, 2010) and in some cases the spatial random effect can actually capture variability otherwise explainable with fixed effects (Paciorek, 2010; Bose et al., 2018). Moreover, with uniform (uninformative) priors the length-scale and variance hyperparameters of Matérn class of covariance functions are non-identifiable (Zhang, 2004) in general. In order to alleviate these identifibiality challenges we used weakly informative priors for the variance and length-scale parameters defined through the principles proposed by Gelman (2006), Simpson et al. (2017) and Fuglstad et al. (2018). Our priors for the covariance function parameters favor small variance and large length-scale parameter values. The former penalizes for large magnitudes and the latter penalizes for wiggly responses along covariates or space. In case of covariate responses this is ecologically justified since according to ecological niche theory species' responses to environmental covariates are typically either monotonic or have only one mode (Ovaskainen et al., 2016). Moreover, Fuglstad et al. (2018) show that these type of priors work well for spatial random effects with Matérn covariance functions in general. Favoring longer spatial length-scales is justified also from the model identifibiality point of view. Identifiability between fixed and spatial random effects is of less concern if spatial random effect varies with larger spatial range than the environmental covariates (Paciorek, 2010, Section 3). Since the environmental covariates in Finnish coastal region vary with relatively small spatial range, our weakly informative prior effectively favors models that explain this variation with covariate responses instead of with spatial random effects. Our model comparison suggests that the models with both covariate effects and spatial random effects lead to most reliable inference on environmental responses and spatial random effects. A model's extrapolation performance is governed mainly by the environmental covariates so better extrapolation performance indicates more reliable response functions. Similarly more accurate interpolation performance indicates more reliable joint effect of environmental covariates and spatial random effect.

An evident challenge with multivariate additive GP is the computation. The core element in the multivariate additive GP is the covariance matrix induced by the GP components. With many species and sampling sites the covariance matrix can become so large that it makes the implementation infeasible. Here we used Laplace method to speed up the computation by decreasing the number of costly posterior density calculations compared to full MCMC. However, in order to scale up the methods for large data sets involving thousands of sampling sites, the model would need to be implemented even more efficiently. One option to reduce the computational time could be to exploit the property that linear LMC can be parameterized through species specific conditional distributions (Gelfand et al., 2004). We could also implement multivariate GPs with sparse GP approximations (Vanhatalo et al., 2010; Alvarez et al., 2010) and replace the full rank LMC model with latent factor model (Thorson et al., 2015; Ovaskainen et al., 2017). However, we leave these considerations for future.

7.3 Summary

Our JSDM based on multivariate additive GP combines the key ideas from semi-parametric SSDMs and state-of-the-art parametric JSDMs. In our case study, it showed superior predictive performance compared to existing GP based SSDMs and parametric SSDMs and JSDMs in both interpolation and extrapolation. Hence, the multivariate

additive GP can be seen as the first step towards integration of the semi-parametric SSDMs and hierarchical JSDMs. We propose also an efficient approach for inference by utilizing Laplace approximation and gradient based optimization for hyperparameters. The multivariate additive GP model is not restricted only to species distribution modeling but can be used in wide variety of other applications as well.

Supplementary Material

Supplementary Material: Additive multivariate Gaussian process for joint species distribution modeling with heterogeneous data (DOI: 10.1214/19-BA1158SUPP; .pdf). The supplementary material contains additional mathematical formulation of the methodology proposed in this paper and additional figures and tables for the case study analysis.

References

Altartouri, A., Nurminen, L., and Jolma, A. (2014). "Modeling the role of the close-range effect and environmental variables in the occurrence and spread of Phragmites australis in four sites on the Finnish coast of the Gulf of Finland and the Archipelago Sea." *Ecology and Evolution*, 4: 987–1005. 435

Alvarez, M. A., Luengo, D., Titsias, M. K., and Lawrence., N. D. (2010). "Efficient multi-output Gaussian processes through variational inducing kernels." *JMLR Workshop and conference proceedings*, 9: 25–32. 439

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman Hall/CRC, second edition. MR3362184. 420

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). "Modelling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage." *Statistical Sinica*. MR1804544. 423, 436

Bergström, U., Olsson, J., Casini, M., Eriksson, B. K., Fredriksson, R., Wennhage, H., and Appelberg, M. (2015). "Stickleback increase in the Baltic Sea – A thorny issue for coastal predatory fish." *Estuarine, Coastal and Shelf Science*, 163: 134–142. 417, 418

Bose, M., Hodges, J. S., and Banerjee, S. (2018). "Toward a diagnostic toolkit for linear models with Gaussian-process distributed random effects." *Biometrics*, 74(3): 863–873. MR3860707. doi: https://doi.org/10.1111/biom.12848. 439

Busse, S. and Snoeijs, P. (2002). "Gradient responses of diatom communities in the Bothnian Bay, northern Baltic Sea." *Nova Hedwigia*, 3–4: 501–525. 418

Byström, P., Bergström, U., Hjälten, A., Ståhl, S., Jonsson, D., and Olsson, J. (2015). "Declining coastal piscivore populations in the Baltic Sea: Where and when do sticklebacks matter?" *Ambio*, 44: 462–471. 435

Candolin, U., Engström-Öst, J., and Salesto, T. (2008). "Human-induced eutrophication

enhances reproductive success through effects on parenting ability in sticklebacks."
Oikos, 117: 459–465.   435

Chib, S. and Greenberg, E. (1998). "Analysis of multivariate probit models."
Biometrika, 85(2): 347–361.   416

Clark, J. S., Gelfand, A., Woodall, C. W., and Zhu, K. (2014). "More than the sum of
the parts: forest climate response from joint species distribution models." *Ecological
Applications*, 24(5): 990–999.   416, 436, 438

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., and Zhang, S. (2016).
"Generalized joint attribute modeling for biodiversity analysis: median-zero, multi-
variate, multifarious data." *Ecological Monographs*, 87(1): 34–56.   437

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatial-Temporal Data*. Wiley Series
in Probability and Statistics. MR2848400.   420

Dunstan, P. K., Foster, S. D., Hui, F. K. C., and Warton, D. I. (2013). "Finite Mix-
ture of Regression Modeling for High-Dimensional Count and Biomass Data in Ecol-
ogy." *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3): 357–
375. MR3110898. doi: https://doi.org/10.1007/s13253-013-0146-x.   416, 436,
437

Duvenaud, D., Nickisch, H., and Rasmussen, C. (2011). "Additive Gaussian Processes."
Neural Information Processing Systems.   421

Elith, J. and Leathwich, J. R. (2009). "Species Distributions Models: Ecological Ex-
planation and Predictions Across Space and Time." *The Annual Review of Ecology,
Evolution and Systematics*, 40: 677–697.   416

Eriksson, B. K., Ljunggren, L., Sandstrom, A., Johansson, G., Mattila, J., Rubach,
A. E., Råberg, S., and Snickars, M. (2009). "Declines in predatory fish promote bloom-
forming macroalgae." *Ecological Applications*, 19: 1975–1988.   418

Eriksson, B. K., Rubach, A., Batsleer, J., and Hillebrand, H. (2012). "Cascading preda-
tor control interacts with productivity to determine the trophic level of biomass ac-
cumulation in a benthic food web." *Ecological Research*, 27: 203–210.   418

Fox, E. B. and Dunson, D. B. (2015). "Bayesian Nonparametric Covariance Regression."
Journal of Machine Learning research, 16(1): 2501–2542. MR3450515.   437

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2018). "Constructing pri-
ors that penalize the complexity of Gaussian random fields." *Journal of the Ameri-
can Statistical Association*, 1(1): 1–8. MR3941267. doi: https://doi.org/10.1080/
01621459.2017.1415907.   439

Gelfand, A. E., Silander, Jr, J. A., Wu, S., Latimer, A., Lewis, P. O., Rebelo, A. G., and
Holder, M. (2006). "Explaining Species Distribution Patterns through Hierarchical
Modelling." *Bayesian Analysis*, 1(1): 41–92. MR2227362. doi: https://doi.org/10.
1214/06-BA102.   416, 420

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004). "Nonstation-
ary multivariate process modeling through spatially varying coregionalization." *Test*,

13(2): 263–312. MR2154003. doi: https://doi.org/10.1007/BF02595775. 422, 424, 439

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 439

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC, third edition. MR3235677. 431

Golding, N. and Purse, B. V. (2016). "Fast and flexible Bayesian species distribution modelling using Gaussian processes." *Methods in Ecology and Evolution*, 7: 598–608. 416, 417, 421, 422, 437, 438

Guisan, A., Edwards, T. C., and Hastie, T. (2002). "Generalized linear and generalized additive models in studies of species distributions: setting the scene." *Ecological Modelling*, 157(2–3): 89–100. 437

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P., and Buckley, Y. M. (2013). "Predicting species distributions for conservation decisions." *Ecology Letters*, 16(12): 1424–1435. 436

Hartmann, M., Hosack, G. R., Hillary, R. M., and Vanhatalo, J. (2017). "Gaussian process framework for temporal dependence and discrepancy functions in Ricker-type population growth models." *Annals of Applied Statistics*, 11(3): 1375–1402. MR3709563. doi: https://doi.org/10.1214/17-AOAS1029. 416, 423

Hastie, T. and Tibishirani, R. (1986). "Generalize additive models." *Statistical Science*, 1(3): 297–318. MR0858512. 421

Hodges, J. S. and Reich, B. J. (2010). "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love." *The American Statistician*, 64(4): 325–334. MR2758564. doi: https://doi.org/10.1198/tast.2010.10052. 438

Hudd, R., Lehtonen, H., and Kurttila, I. (1988). "Growth and abundance of fry; factors which influence the year-class strength of whitefish (*Coregonus lavaretus widegreni*) in the southern Bothnian Bay (Baltic)." *Finnish Fisheries Research*, 9: 213–220. 434

Hui, F. K. C., Warton, D. I., Foster, S. D., and Dunstan, P. K. (2013). "To mix or not to mix: Comparing the predictive performance of mixture models vs. separate species distribution models." *Ecology*, 94(9): 1913–1919. 416, 436

Kallasvuo, M., Vanhatalo, J., and Veneranta, L. (2017). "Modeling the spatial distribution of larval fish abundance provides essential information for management." *Canadian Journal of Fisheries and Aquatic Sciences*, 74: 636–649. 416, 417, 421, 437, 438

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 430

Kotta, J., Vanhatalo, J., Jänes, H., Orav-Kotta, H., Rugiu, L., Jormalainen, V., Bobsien, I., Viitasalo, M., Virtanen, E., Sandman, A. N., Isaeus, M., Leidenberger, S., Jonsson, P., and Johannesson, K. (2019). "Integrating experimental and distribution data to predict future species patterns." *Scientific Reports*, 9(1): 1821. 416, 437, 438

Kurowicka, D. and Cooke, R. (2003). "A parameterization of positive definite matrices in terms of partial correlation vines." *Linear Algebra and its Applications*, 372(Supplement C): 225–251. MR1999149. doi: https://doi.org/10.1016/S0024-3795(03)00507-X. 426

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander, Jr., J. A. (2006). "Building Statistical Models to Analyze Species Distributions." *Ecological Applications*, 16(1): 33–50. 416, 438

Lefébure, R., Larsson, S., and Byström, P. (2011). "A temperature dependent growth model for the three-spined stickleback *Gasterosteus aculeatus*." *Journal of fish biology*, 79: 1815–1827. 435

Leskelä, A., Hudd, R., Lehtonen, H., Huhmarniemi, A., and Sandström, O. (1991). "Habitats of whitefish (*Coregonus lavaretus (L.) s.l.*) larvae in the Gulf of Bothnia." *Aqua Fennica*, 21: 145–151. 434

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). "Generating random correlation matrices based on vines and extended onion method." *Journal of Multivariate Analysis*, 100(9): 1989–2001. MR2543081. doi: https://doi.org/10.1016/j.jmva.2009.04.008. 426

Lindley, D. V. (2002). "Seeing and Doing: The Concept of Causation." *International Statistical Review*, 70(2): 191–214. 426

Liu, J. and Vanhatalo, J. (2018). "Bayesian model-based spatiotemporal survey design for log-Gaussian Cox process." *arXiv e-prints*, arXiv:1808.09200. 421

Lopez, H. F. (2000). "Bayesian Analysis in Latent Factor and Longitudinal Models." Ph.D. thesis, Institute of Statistics and Decision Sciences, Duke University. MR2700959. 436

Lopez, H. F., Salazar, E., and Gamerman, D. (2008). "Spatial Dynamic Factor Analysis." *Bayesian Statistics*, 3(4): 759–792. MR2469799. doi: https://doi.org/10.1214/08-BA329. 436

MacKenzie, B. R., Gislason, H., Möllmann, C., and Köster, F. W. (2007). "Impact of 21st century climate change on the Baltic Sea fish community and fisheries." *Global Change Biology*, 13: 1348–1367. 437

Mardia, K. V. and Goodall, C. R. (1993). "Spatio-Temporal analysis of Multivariate Environmental Monitoring Data." *Multivariate Environmental Statistics*, 347–386. MR1268443. 422, 424

Cingi, S., Keinänen, M., and Vuorinen, P. J. (2010). "Elevated water temperature impairs fertilization and embryonic development of whitefish *Coregonus lavaretus.*" *Journal of Fish Biology*, 76: 502–521. [435]

Meier, H. E. M., Hordoir, R., Andersson, H. C., Dieterich, C., Eilola, K., Gustafsson, B. G., Höglund, A., and Schimanke, S. (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 Dynamics*, 39: 2421–2441. [435]

Müller, R. (1992). "Trophic state and its implications for natural reproduction of salmonid fish." In *Dynamics and Use of Lacustrine Ecosystems*. Springer, Dordrecht. [435]

Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized Linear Models." *Journal of the Royal Statistical Association*, 135(3): 370–384. [416]

Nickisch, H. and Rasmussen, C. E. (2008). "Approximations for binary Gaussian process classification." *Journal of Machine learning*, 9: 2035–2078. MR2452620. [423]

Odum, E. P. (1953). *Fundamentals of ecology*. Wiley Subscription Services, Inc., A Wiley Company. [428]

O'Hagan, A. (1978). "Curve Fitting and Optimal Design for Prediction." *Journal of Royal Statistical Society B*, 40(1): 1–42. MR0512140. [416]

Ovaskainen, O., de Knegt, H. J., and del Mar Delgado, M. (2016). *Quantitative Ecology and Evolutionary Biology – Integrating models with data*. Oxford University Press. [439]

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). "Modeling species co-occurence by multivariate logistic regression generates new hypotheses on fungal interaction." *Ecology*, 91(9): 2514–2521. [437]

Ovaskainen, O. and Soininen, J. (2011). "Making more out of sparse data: Hierarchical modelling of species communities." *Ecology*, 92(2): 289–295. [416]

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." *Ecology Letters*, 20(5): 561–576. [416, 422, 436, 437, 438, 439]

Paciorek, C. J. (2010). "The Importance of Scale for Spatial-Confounding Bias and Precision of Spatial Regression Estimators." *Statistical Science*, 25(1): 107–125. MR2741817. doi: https://doi.org/10.1214/10-STS326. [439]

Pearl, J. (2009). "Causal inference in statistics: An overview." *Statistics Surveys*, 3: 96–146. MR2545291. doi: https://doi.org/10.1214/09-SS057. [426]

Peltonen, H., Vinni, M., Lappalainen, A., and Pönni, J. (2004). "Spatial feeding patterns of herring (*Clupea harengus* L.), sprat (*Sprattus sprattus* L.), and the three-spined stickleback (emphGasterosteus aculeatus L.) in the Gulf of Finland, Baltic Sea." *ICES Journal of Marine Science*, 61: 966–971. [434, 435]

Pitkänen, H., Peuraniemi, M., Westerbom, M., Kilpi, M., and von Numers, M. (2013). "Longterm changes in distribution and frequency of aquatic vascular plants and charophytes in an estuary in the Baltic Sea." *Annales Botanica Fennica*, 50: 1–54. 435

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., Hara, R. B. O., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). "Understanding co-occurence by modelling species simultaneously with joint species distribution model (JSDM)." *Methods in Ecology and Evolution*, 5: 397–406. 416, 422, 437

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. MR2514435. 416, 421, 426

Record, S., Fitzpatrick, M. C., Finley, A. O., Veloz, S., and Ellison, A. M. (2013). "Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change." *Global Ecology and Biogeography*, 22(6): 760–771. 436

Reusch, T. B. H., Dierking, J., Andersson, H. C., Bonsdorff, E., Carstensen, J., Casini, M., Czajkowski, M., Hasler, B., Hinsby, K., Hyytiäinen, K., Johannesson, K., Jomaa, S., Jormalainen, V., Kuosa, H., Kurland, S., Laikre, L., MacKenzie, B. R., Margonski, P., Melzner, F., Oesterwind, D., Ojaveer, H., Refsgaard, J. C., Sandström, A., Schwarz, G., Tonderski, K., Winder, M., and Zandersen, M. (2018). "The Baltic Sea as a time machine for the future coastal ocean." *Science Advances*, 4(5). 417

Rönnberg, C. and Bonsdorff, E. (2004). "Baltic Sea eutrophication: area-specific ecological consequences." *Hydrobiologia*, 514: 227–241. 418

Rue, H. and Marino, S. (2009). "Approximate Bayesian Inference for latent Gaussian models by using integrated Laplace approximantions." *Journal of the Royal Statistical Society*, 71(2): 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 423

Shelton, A. O., Thorson, J. T., Ward, E. J., and Feist, B. E. (2014). "Spatial semiparametric models improve estimates of species abundance and distribution." *Canadian Journal of Fisheries and Aquatic Sciences*, 71(July): 1655–1666. 437

Sieben, K., Ljunggren, L., Bergström, U., and Eriksson, B. K. (2011). "A meso-predator release of stickleback promotes recruitment of macroalgae in the Baltic Sea." *Journal of Experimental Marine Biology and Ecology*, 397: 79–84. 434, 435

Simpson, D. P., Rue, H. v., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. MR3634300. doi: https://doi.org/10.1214/16-STS576. 439

Taylor-Rodríguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S., and Gelfand, A. E. (2017). "Joint Species Distribution Modeling: Dimension Reduction Using Dirichlet Processes." *Bayesian Analysis*, 12(4): 939–967. MR3724974. doi: https://doi.org/10.1214/16-BA1031. 436, 437

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., and Kristensen, K. (2015). "Spatial factor analysis: a new tool for estimating joint species

distributions and correlations in species range." *Methods in Ecology and Evolution*, 6(6): 627–637.   416, 436, 437, 438, 439

Tikhonov, G., Abrego, N., Dunson, D., and Ovaskainen, O. (2017). "Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context." *Methods in Ecology and Evolution*, 8(4): 443–452. MR3889899.   437

Tokuda, T., Goodrich, B., Mechelen, I. V., and Gelman, A. (2012). "Visualizing Distributions of Covariance Matrices."   423, 436

Vanhatalo, J., Hartmann, M., and Veneranta, L. (2019). "Supplementary Material: Additive multivariate Gaussian process for joint species distribution modeling with heterogeneous data." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1158SUPP.   425

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). "Approximate inference for disease mapping with sparse Gaussian processes." *Statistics in Medicine*, 29(15): 1580–1607. MR2758849. doi: https://doi.org/10.1002/sim.3895.   417, 423, 426, 439

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). "GPstuff : Bayesian Modeling with Gaussian Processes." *Journal of Machine Learning Research*, 14: 1175–1179. MR3063621.   426

Vanhatalo, J., Veneranta, L., and Hudd, R. (2012). "Species distribution modelling with Gaussian processes: A case study with youngest stages of sea spawning whitefish (Coregonus lavatus L. s.l.) larvae." *Ecological Modelling*, (228): 49–58.   416, 419, 421, 422, 430, 434, 435, 437, 438

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016). "Bayesian Leave-One-Out Cross-Validation Approximations for Gaussian Latent Variable Models." *Journal of Machine Learning Research*, 17: 1–38. MR3543509.   426, 430

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6: 141–228. MR3011074. doi: https://doi.org/10.1214/12-SS102.   430

Veneranta, L., Hudd, R., and Vanhatalo, J. (2013). "Reproduction areas of sea-spawning coregionids reflect the environment in shallow coastal areas." *Marine Ecology Progress Series*, 477: 231–250.   417, 418, 419, 430, 434, 435

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. (2015). "So Many Variables: Joint Modeling in Community Ecology." *Trends in Ecology and Evolution*, 30(12): 766–779.   416

Warton, D. I. and Shepherd, L. C. (2010). "Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology." *Annals of Applied Statistics*, 4(3): 1383–1402. MR2758333. doi: https://doi.org/10.1214/10-AOAS331.   420

Wikle, C. K. (2003). "Hierarchical Models in Environmental Science." *International Statistical Review*, 71(2): 181–199.   420

Zhang, H. (2004). "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics." *Journal of the American Statistical Association*, 99(465): 250–261. MR2054303. doi: https://doi.org/10.1198/016214504000000241.   439

Acknowledgments