Received: 15 January 2020 | Accepted: 5 April 2021

DOI: 10.1002/env.2681

RESEARCH ARTICLE

WILEY

Spatiotemporal clustering using Gaussian processes embedded in a mixture model

Jarno Vanhatalo[1] | **Scott D. Foster**[2] | **Geoffrey R. Hosack**[2]

[1] Department of Mathematics and Statistics and Organismal and Evolutionary Biology Research Program, University of Helsinki, Helsinki, Finland

[2] Data61, CSIRO, Hobart, Tasmania, Australia

Correspondence
Jarno Vanhatalo, Department of Mathematics and Statistics and Organismal and Evolutionary Biology Research Program, University of Helsinki, P.O. Box 68, Helsinki FIN-00014, Finland.
Email: jarno.vanhatalo@helsinki.fi

Funding information
Academy of Finland, Grant/Award Number: 317255; Data Analytics Program of Data61 (CSIRO); Research Funds of the University of Helsinki, Grant/Award Number: 465/51/2014

Abstract

The categorization of multidimensional data into clusters is a common task in statistics. Many applications of clustering, including the majority of tasks in ecology, use data that is inherently spatial and is often also temporal. However, spatiotemporal dependence is typically ignored when clustering multivariate data. We present a finite mixture model for spatial and spatiotemporal clustering that incorporates spatial and spatiotemporal autocorrelation by including appropriate Gaussian processes (GP) into a model for the mixing proportions. We also allow for flexible and semiparametric dependence on environmental covariates, once again using GPs. We propose to use Bayesian inference through three tiers of approximate methods: a Laplace approximation that allows efficient analysis of large datasets, and both partial and full Markov chain Monte Carlo (MCMC) approaches that improve accuracy at the cost of increased computational time. Comparison of the methods shows that the Laplace approximation is a useful alternative to the MCMC methods. A decadal analysis of 253 species of teleost fish from 854 samples collected along the biodiverse northwestern continental shelf of Australia between 1986 and 1997 shows the added clarity provided by accounting for spatial autocorrelation. For these data, the temporal dependence is comparatively small, which is an important finding given the changing human pressures over this time.

KEYWORDS

clustering, community ecology, Gaussian process, Laplace approximation, mixture, regions of common profiles, spatial, spatiotemporal

1 | INTRODUCTION

Identifying regions of relative homogeneity in data is a common goal in most, and probably all, data-driven disciplines. This process of "clustering" or "unsupervised classification" has a long history dating back to the 1930s (Driver & Kroebe, 1932; Zubin, 1938). There are numerous textbooks devoted to the topic (Kaufman & Rousseeuw, 1990), and even societies and journals (see Murtagh & Kurtz, 2016). It is easy to understand why clustering is popular amongst applied data analysts: it gives results that humans find easy to interpret whilst capturing salient patterns of variation in the data. The ease of interpretation stems from the observation that humans are naturally predisposed to understanding categorizations (e.g., color labels, Linnaean system of taxonomy, and so forth).

Our motivation for studying clustering methods comes from ecology, and in particular the task of regionalization (biogeography)—where an analyst wants to find groups of sites that have similar assemblages of species (Pielou, 1984; Woolley et al. 2019). Here, we focus on the biogeographic patterns of bony (teleost) fish on the north-west shelf (NWS) of Australia, which is a productive and biodiverse ecosystem of long-term scientific interest (Considine, 1985; Nowara and Newman 2001). Biogeographic data, such as the fish data, are sampled in physical space and often through time. Spatial and temporal dependence may therefore arise among biological observations but nearly all cluster analyses ignores this possibility. In doing so, these studies inadvertently ignore the potential for spatiotemporal correlation to be confused with ecological groups.

In this work, we extend the model-based clustering method of Foster et al. (2013) to include spatial and spatiotemporal dependence into the model for observations. The model is a mixture-of-experts model (Jacobs et al., 1991; Jordan & Jacobs, 1994), where the probability of observing each cluster is allowed to vary with environmental covariates and we additionally allow for spatial or spatiotemporal correlation by including Gaussian processes (GPs; e.g., see Rasmussen & Williams, 2006). An appealing feature of this approach is that it allows predictions from the correlated-model to leverage off the spatial locations, and time, of the observed data as well as the inherent relationships of biology and the environment—even when predicting at locations with no direct observation.

Our approach is novel in that it is for multivariate observations (measurements on hundreds of species per site is not uncommon), and it is defined for continuous space and time using semiparametric GP response functions. Previous spatial clustering methods include, for example: (1) Spatial scan statistics (Kulldorff, 1997), which ignores environmental effects and focuses entirely on spatial properties in an algorithmic framework. (2) Two-step approaches where a hard-coded label prediction is produced algorithmically and subsequently regressed on spatial covariates and possibly spatiotemporal coordinates (Anderson et al., 2014; Bilancia & Demarinis, 2014). This approach ignores uncertainty associated with the prediction of the label. (3) Clustering areal data (Alfó et al., 2009; Green & Richardson, 2002; Lawson et al., 2017; Neelon et al., 2014; Torabi, 2016; Wall & Liu, 2009), which requires data to be gridded. The gridding is an unnatural representation of most ecological data, which is best represented in continuous spatiotemporal domain. (4) Digital image analysis (Ambroise et al., 1997; Nguyen & Wu, 2012; Woolrich et al., 2005) and specification of priors that encourage neighboring sites to share cluster labels (Corander et al., 2008; Guillot et al., 2005). Whilst close to our representation, these models do not easily allow for inclusion of covariate effects. Unlike the previously introduced models, our approach allows for covariates and for spatiotemporal autocorrelation within the data. This is achieved in a single analysis, which avoids the problems of propagating uncertainty through multiple stages of an analysis. In addition to these desirable qualities, our approach utilizes semiparametric functions for modeling the responses of cluster probabilities along covariates. We also present novel methods to summarize these effects in an intuitively clear manner.

To address the complexities introduced by the inclusion of spatial and spatiotemporal dependence, we introduce novel methods to conduct approximate Bayesian inference that scale well with both the number of samples and the dimensionality of those observations. Our fastest, and crudest inference method, is based on the Laplace approach, which is also the basis of the integrated nested Laplace approximation (Rue et al., 2009) approach that has been shown to perform well for large number of latent Gaussian variable models. Even though the Laplace method has been used for spatial clustering by, for example, Bilancia and Demarinis (2014) and Anderson et al. (2014), these earlier approaches have utilized it only for the hierarchical model conditional on the predefined cluster structure. In our approach, however, clusters are probabilistic and we use the Laplace approximation to marginalize over the spatiotemporally varying cluster probabilities and to approximate their posterior distributions. This approach is technically similar to the Laplace approximation for multiclass and Multinomial GP models (Juntunen et al., 2011; Rasmussen & Williams, 2006; Riihimäki et al., 2013). We propose also to combine Laplace approximation with partial Markov chain Monte Carlo (MCMC) to improve the accuracy of inference for the key model parameters. This combined approach is similar in nature to the approaches of Vanhatalo et al. (2010), Vanhatalo et al. (2013) and Gómez-Rubio and Rue (2018) in that the Laplace approximation is used for approximately marginalizing over a set of model parameters within an MCMC algorithm. A full MCMC procedure is also examined for comparison.

We demonstrate and test our methods with a simulation study. We then analyze 854 samples of 253 teleost fish on the NWS of Australia (see Figure 1) to test our methods in large real-world data and illustrate the effects of including spatial and spatiotemporal effects by fitting models with and without them. The temporal component may be particularly important for this region that has been subject to differing exploitation rates of fish as well as different resource management paradigms (Considine, 1985; Sainsbury et al., 1993).

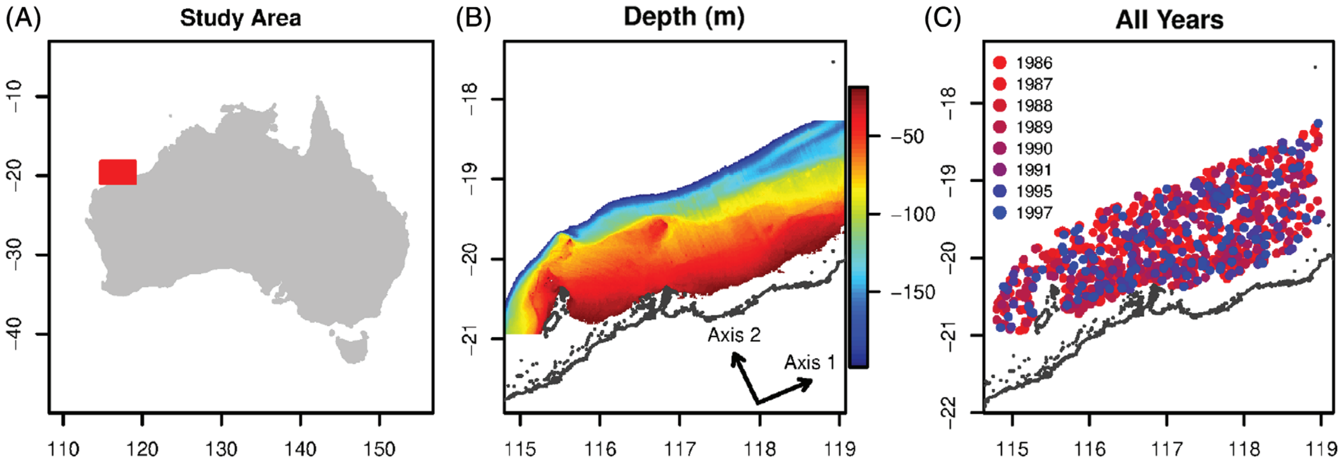

FIGURE 1 Summary of north-west shelf (NWS) data. (A) Map of Australian continent with NWS region defined. (B) Depth covariate of the NWS region (also shown are the rotated geography axes). Deeper water (maximum of 197 m) is deep blue and shallower water (minimum of 20 m) is colored red. (C) Locations of all biological samples colored by year of sampling

2 | STUDY AREA AND DATA

2.1 | NWS region

The NWS region of Australia (see Figure 1(A)) is a remote but resource rich marine area in tropical north west Australia. The continental shelf along the NWS supports a productive ecosystem influenced by both tropical and subtropical systems. Since the mid-1960s, the NWS has supported fisheries and a number of different species have been targeted (Considine, 1985; Wallner & Phillips, 1988; Sainsbury et al., 1993; Nowara and Newman, 2001). At times, the total finfish catch from this area has been much greater than that from any other waters in Australia (Considine, 1985). To understand the effect that early fishing effort had on the composition and health of the finfish stocks, and to assess the possibility of developing a domestic fishery, a number of surveys were undertaken in the early 1980s (Sainsbury, 1979; Nowara and Newman, 2001). More surveys were then sporadically conducted until 1991 with the goal of investigating management options (Sainsbury et al., 1993) with further surveys until 1997. The data contain information about how fish biodiversity varies in space and time. In particular, we wish to uncover the patterns of variation in fish assemblages.

2.2 | Biological data

The NWS data consists of 854 trawls spread from October 1986 to August 1997. The vast majority of these trawls are "community" focused with an object to describe the fish species community, where attributes of all species are recorded. For this analysis, we chose to exclude earlier data where there was ambiguity about the survey objective (e.g., see Sainsbury & Whitelaw, 1984; Thresher et al., 1986). The raw data is available from the CSIRO data trawler[1] and the data used in our analysis is available as Supplementary Material. The data spanned almost 200 m of depth, with 21 m being the shallowest (see Figure 1(B,C)). We analyze the presence-absence of the 253 species (reduced from 579) that were present in 15 or more trawls. Species with very few presences are unlikely to substantially contribute to the evidence-base for biogeographic patterns.

2.3 | Physical environment data

The physical environment was delineated using climatologies (long-term averages), which are hence time invariant. These are the same sources of physical covariates as was used by Foster et al. (2013) and Figure 1(B) gives the example

[1] https://www.cmar.csiro.au/data/trawler/

of depth. The climatological covariates used in this work are depth, intraannual standard deviation (SD) of nitrate (NO3 SD), intraannual SD of dissolved oxygen (O2 SD) and annual mean of salinity. Intraannual SD can be important to ecological systems as it measures the range of environmental conditions that a single location may encounter. A dense grid throughout the region—bounded by latitude, longitude and depth—of all these covariates is used for prediction.

We delineate space with respect to orthogonal axes rotated relative to easting and northing coordinates. The rotation was undertaken since patterns of variation in the NWS region tend to be (approximately) north-east to south-west aligned (as is the sampling region itself, see Figure 1(B)). This was achieved by aligning the spatial coordinates to the first two principal directions of variation in the sampling locations.

3 | METHODS

3.1 | Spatiotemporal clustering model

Clustering methods aim to partition the multivariate samples into K groups that are more similar to each other than they are to observations in different groups. Using mixture models (McLachlan & Peel, 2000), the clusters are found by encapsulating each multivariate observation's latent group label into the model. Formally, we define the latent label for the ith observation ($i = 1 \dots n$ indexes the sampling sites) as the $1 \times K$ vector $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \dots, z_{iK})$ with the kth element equal to "1" if the observation belongs to group k and "0" otherwise; the groups are assumed mutually exclusive such that the observation is assigned to only one of the K groups. The variable $\boldsymbol{z}_i$ is assumed to follow a categorical distribution with mean $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{iK})$.

We build upon the model of Foster et al. (2013) who assume that, conditional on the latent group label, $\boldsymbol{z}_i$, the expectation for the observed species data is constant among sampling sites. That is, the multivariate data for the J species at a sampling location $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iJ})$ has conditional elementwise expectation $\mathrm{E}\left(y_{ij}|z_{ik} = 1\right) = \mu_{kj}$ that is constant over sampling sites belonging to the same cluster. We follow the nomenclature of Foster et al. (2013) who call the vector of conditional expectations $\mathrm{E}\left(\boldsymbol{y}_i|z_{ik} = 1\right) = \boldsymbol{\mu}_k$ "a profile," and the regions of covariate space that the groups occupy are "regions of common profile" (RCP). All observations within an RCP have the same species specific conditional expectations and thus satisfy the requirement that observations within an RCP are more similar than observations in different RCPs. Define $p\left(\boldsymbol{y}_i|z_{ik} = 1, \boldsymbol{\mu}\right) = p\left(\boldsymbol{y}_i|z_{ik} = 1, \boldsymbol{\mu}_k\right) = \prod_{j=1}^{J} p\left(y_{ij}|z_{ik} = 1, \mu_{kj}\right)$ to be the conditional probability density of the ith observation, with $\boldsymbol{\mu}$ being all of the groups' profiles ($K \times J$ matrix). Note that we have partitioned $\boldsymbol{\mu}$ into its group-specific components $\boldsymbol{\mu}_k$ (a $1 \times J$ vector). The unconditional distribution of the observation is the mixture distribution

$$p\left(\boldsymbol{y}_i|\boldsymbol{\pi}_i, \boldsymbol{\mu}\right) = \sum_{k=1}^{K} \pi_{ik} p\left(\boldsymbol{y}_i|z_{ik} = 1, \boldsymbol{\mu}_k\right). \tag{1}$$

For the NWS data, the observations are binary (present/absent) and so we assume that the conditional distributions for each y_{ij} are independent Bernoulli random variables and parameterize the conditional observation models through their mean, $p\left(y_{ij}|z_{ik} = 1, \mu_{kj}\right) = \mathrm{Bernoulli}(y_{ij}|\mu_{kj})$.

We extend this mixture model for spatial and temporal dependence by allowing the expectation of the group label, $\boldsymbol{\pi}_i$, to vary with covariates *and* the spatiotemporal coordinates of the observation. That is $\boldsymbol{\pi}_i = \boldsymbol{\pi}(\boldsymbol{x}_i, \boldsymbol{s}_i, t_i)$ where $\boldsymbol{s}_i$ is the vector of spatial coordinates of site i, t_i is the sampling time and $\boldsymbol{x}_i = \boldsymbol{x}(\boldsymbol{s}_i)$ are the covariates associated with site i. Here, we choose $\boldsymbol{\pi}(\boldsymbol{x}_i, \boldsymbol{s}_i, t_i)$ to be the softmax function (Neelon et al., 2014) but other link functions (Aitchison, 1982; Daganzo, 1979) could be used as well. The softmax function gives the kth element as

$$\pi_k(\boldsymbol{x}_i, \boldsymbol{s}_i, t_i) = \frac{\exp\left(\alpha_k + h_k(\boldsymbol{x}_i) + \phi_k(\boldsymbol{s}_i, t_i)\right)}{\sum_{k'=1}^{K} \exp\left(\alpha_{k'} + h_{k'}(\boldsymbol{x}_i) + \phi_{k'}(\boldsymbol{s}_i, t_i)\right)}, \tag{2}$$

where $\alpha_k \sim N(0, \sigma_\alpha^2)$ are the groupwise constant terms, $h_k(\boldsymbol{x}_i)$ are the groups' responses to covariates, and $\phi_k(\boldsymbol{s}_i, t_i)$ are the residual spatiotemporal patterns. The groupwise constant terms, α_k, are mutually a priori independent and the responses to the covariates and the spatiotemporal patterns will be modeled using mutually a priori independent GPs (Cressie & Wikle, 2011; Gelfand et al., 2010; Rasmussen & Williams, 2006). We will denote by $f_k(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{s}, t) = \alpha_k + h_k(\boldsymbol{x}) + \phi_k(\boldsymbol{s}, t)$ a latent function combining the covariate and spatiotemporal effects for the kth RCP group.

The spatiotemporal random effects, $\{\phi_k(\boldsymbol{s}_i, t_i)\}_{k=1}^K$, are the main distinguishing feature between our model and the model introduced by Foster et al. (2013), who only considered models with low-order polynomial basis expansions. Additionally we will perform Bayesian inference, whereas Foster et al. (2013) considered a maximum likelihood approach. The spatiotemporal GPs provide a way for observations to "borrow strength" from other observations nearby by capturing spatial and temporal correlations. Such correlations could arise, for example, from missing covariates or from inherent properties of the ecosystem (e.g., fish foraging behavior and reproduction strategies in our application).

We give the spatiotemporal random effects independent zero mean GP priors

$$\phi_k(\boldsymbol{s}, t)|\boldsymbol{\theta}_{\phi,k} \sim GP\left(0, c_{\phi,k}((\boldsymbol{s}, t), (\boldsymbol{s}', t')|\boldsymbol{\theta}_{\phi,k})\right), \tag{3}$$

where $c_{\phi,k}\left((\boldsymbol{s}, t), (\boldsymbol{s}', t')|\boldsymbol{\theta}_{\phi,k}\right)$ is a separable spatiotemporal covariance function with hyperparameters $\boldsymbol{\theta}_{\phi,k}$. The spatial covariance function is chosen to be the Matérn covariance function with 3/2 degrees of freedom (Rasmussen & Williams, 2006) and we use an exponential correlation function for the temporal process so that

$$c_{\phi,k}\left((\boldsymbol{s}, t), (\boldsymbol{s}', t')|\boldsymbol{\theta}_{\phi,k}\right) = \sigma^2_{\phi,k}\left(1 + \sqrt{3}r(\boldsymbol{s}, \boldsymbol{s}')\right)e^{-\sqrt{3}r(\boldsymbol{s},\boldsymbol{s}')}e^{-|t-t'|/l_{\phi,k,3}}, \tag{4}$$

where $r(\boldsymbol{s}, \boldsymbol{s}') = \sqrt{\sum_{q=1}^2 (s_q - s'_q)^2/l_{\phi,k,q}}$ is a scaled Euclidean distance between the observation sites. The covariance function is parameterized by a variance and a "length-scale" parameter in both spatial (see Figure 1 for the spatial axes) and in time dimensions—giving hyperparameters $\boldsymbol{\theta}_{\phi,k} = \{\sigma^2_{\phi,k}, l_{\phi,k,1}, l_{\phi,k,2}, l_{\phi,k,3}\}$ for the kth spatiotemporal process. In the NWS data analysis, the spatiotemporal process is used in model M5 and it reduces to spatial process when the temporal covariance function is dropped out (corresponding to $l_{\phi,k,3} = \infty$) which is used in models M3 and M4 (see Section 4.2).

We model the functions of covariates with additive, mutually independent, GPs

$$h_k(\boldsymbol{x})|\boldsymbol{\theta}_{h,k,1}, \dots, \boldsymbol{\theta}_{h,k,D} \sim GP\left(0, \sum_{d=1}^{D} c_{h,k,d}(x_d, x'_d|\boldsymbol{\theta}_{h,k,d})\right), \tag{5}$$

where $c_{h,k,d}(x_d, x'_d|\boldsymbol{\theta}_{h,k,d})$ is the covariance function for the response along the dth covariate ($d = 1 \dots D$), in the kth RCP and $\boldsymbol{\theta}_{h,k,d}$ are the corresponding hyperparameters. The GP formulation for predictive functions allows for the linear models, $h_k(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, with Gaussian distributed weights, $\boldsymbol{\beta} \sim N(0, I\sigma^2_\beta)$, as a special case with covariance functions $c_{h,k,d}(x_d, x'_d|\boldsymbol{\theta}_{h,k,d} = \sigma^2_\beta) = x_d x'_d \sigma^2_\beta$ (Rasmussen & Williams, 2006). Incorporating polynomial regression through a linear model is straight-forward with a suitable basis-expansion of the covariates. In the NWS data analysis we test quadratic covariate responses (models M1 and M3 in Section 4.2). For more flexible models, we use the squared exponential covariance function $c_{h,k,d}(x_d, x'_d|\boldsymbol{\theta}_{h,k,d}) = \sigma^2_{h,k,d}e^{-(x_d - x'_d)^2/l^2_{h,k,d}}$ (Rasmussen & Williams, 2006) with hyperparameters $\boldsymbol{\theta}_{h,k,d} = \{\sigma^2_{h,k,d}, l_{h,k,d}\}$. In the NWS data analysis these semiparametric response functions are used in models M2, M4, and M5 (see Section 4.2). We denote all the covariance function parameters by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\phi,k}, \boldsymbol{\theta}_{h,k,1}, \dots, \boldsymbol{\theta}_{h,k,D}\}_{k=1}^K$.

3.1.1 | Priors

We consider it a priori likely that the spatial variability in the class probabilities could be well explained by the environmental responses. To this end, we prefer priors for variance parameters of the spatial and spatiotemporal covariance functions (models M3–M5 in Section 4.2) which give small effects. We use a heavy-tailed prior to allow for departures from this prior assumption where it is supported by the data. Hence, we use a weakly informative half Student-*t* prior for the SD of the spatial random effects $\sigma_{\phi,k}$ (Gelman, 2006) with four degrees of freedom and scale 0.1. Moreover, we prefer models where RCPs change slowly in space and consider it a priori more likely that the spatial random effects capture nonlocal variation that is not explained by covariates. So, we specify a half Student-*t* priors with four degrees of freedom and scale 1 for the inverse length-scales $1/l_{\phi,k,1}, 1/l_{\phi,k,2}, 1/l_{\phi,k,3}$, for all $k = 1, \dots, K$, to give more weight for spatially and temporally smooth processes. These priors for the spatial model lead to a joint prior which shrinks the model towards a null model in which there is no spatial or temporal effects.

We prefer relatively "stiff" functions for the effects of environmental covariates that do not cross their mean multiple times within the range of plausible covariate values since it is a priori reasonable that species assemblages have a unimodal relationship with each covariate. Hence, in the models with GP response along covariates (models M2, M4, and M5 in Section 4.2) we prefer length-scales that are of the same order as the range of the environmental covariates. To encode this we first scale the covariates to have a SD of 1 and then give a Student-*t* prior with scale one and four degrees of freedom for the inverse length-scales, $1/l_{h,k,d}$, of the covariate effects at the standardized scale. Moreover, we assume that it is plausible that the RCP probabilities do not respond to some of the covariates. To encode this, we give half Student-*t* priors for the SDs of the GP response functions, $\sigma_{h,k,d}$ with four degrees of freedom and scale one. The prior variances of the group specific constants, σ_α^2, and the prior variances for linear weights, σ_β^2, in the models that use linear model for covariate responses (models M1 and M3 in Section 4.2), are fixed to ten so that they correspond to fixed effects.

Independent prior information for most species is sparse or nonexistent. So, we specify mutually independent vague priors for the log odds ratio of the conditional observation probabilities, $\text{logit}(\mu_{kj}) \sim N(0, \sigma_\mu^2)$ where $\sigma_\mu^2 = 10$.

3.2 | Inferential methods

The parameter space of our model is large. For the NWS data with $K = 5$ (see Section 4.2), $J = 253$ species and $D = 4$ environmental covariates, the model includes $K \times J = 1265$ conditional observation probabilities ($\boldsymbol{\mu}$) and a minimum of $4K + 2DK = 60$ covariance function parameters ($\boldsymbol{\theta}$) for the full spatiotemporal model (55 for a spatial model). Additionally there are $n = 854$ multivariate spatial/spatiotemporal sampling sites leading to $n \times K = 4270$ random latent variables, which correspond to the values of the latent function, $f_{ik} = f_k(\boldsymbol{x}(\boldsymbol{s}_i), \boldsymbol{s}_i, t_i) = \alpha_k + h_k(\boldsymbol{x}_i) + \phi_k(\boldsymbol{s}_i, t_i)$, at these $i = 1, \ldots, n$ sampling sites. Note that n refers to the total number of unique spatiotemporal coordinates $(\boldsymbol{s}_i, t_i)$ in our data. Moreover, there are no spatial replicates from exactly same sampling sites even though sites at different times may be located close to each other. Further complexity stems from the fact that a mixture model's likelihood and posterior densities are known to be "bumpy" with many local maxima (Foster et al., 2018; McLachlan & Peel, 2000). In order to make analyses feasible, we propose to use approximate Bayesian inference methods in three increasing levels of accuracy and computation time.

First, as the fastest and crudest approach, we utilize two Laplace approximations in combination (inference method 1): one for the marginal likelihood of model parameters; and another for the (conditional) posterior distribution of latent variables given the hyperparameters (Rasmussen & Williams, 2006; Tierney & Kadane, 1986; Vanhatalo et al., 2010). The former is for estimating hyperparameters and the latter for estimating the posterior of the latent variables. The key benefits are that the dimension of the parameter space is significantly reduced due to (approximate) marginalization over latent variables, and we can use optimization instead of sampling based MCMC approach. Second, as an approach of intermediate accuracy and computational complexity, we use MCMC either to estimate the conditional posterior of the species profiles at the approximate posterior mode of the latent variables (inference method 2a) or to estimate the joint posterior of latent variables and species profiles at the (approximate) posterior mode of the covariance function parameters (inference method 2b). The posterior modes of the latent variables and covariance function parameters are estimated using the Laplace approximation of the inference method 1. Third, as the most accurate but the computationally heaviest approach, we consider full MCMC for all model parameters (inference method 3). The full MCMC would, however, be computationally infeasible for our NWS data.

We introduce the inference methods in detail below. The performance of the proposed methods is examined with simulated data in Section 4.1. All inferential methods were implemented in Matlab by utilizing parts from the GPstuff toolbox (Vanhatalo et al. 2013). The code used for this work is made available at https://github.com/jpvanhat/SpatClustMixtures.

3.2.1 | Parameter inference using Laplace approximations (inference method 1)

Denote by $\boldsymbol{Y}$ the $n \times J$ matrix of all outcome measurements and by $\boldsymbol{X}$ the respective $n \times D$ matrix of covariates. Let $\boldsymbol{f}_k = [f_{1k}, \ldots, f_{nk}]^T$ be the $n \times 1$ vector of latent variables at all observations corresponding to the kth RCP class. We stack $\boldsymbol{f}_k$ to give the full set of latent variables $\boldsymbol{f}$. The prior, conditional on hyperparameters, for the latent variables is a zero mean multivariate Gaussian $\boldsymbol{f} \sim \text{N}(\boldsymbol{0}, \boldsymbol{C})$ where $\boldsymbol{C}$ is a $nK \times nK$ block-diagonal matrix so that the kth block contains elements

$$[\boldsymbol{C}_k]_{ii'} = \sigma_\alpha^2 + \sum_{d=1}^{D} c_{h,k,d}(x_{i,d}, x_{i',d} | \boldsymbol{\theta}_{h,k,d}) + c_{\phi,k}\left((\boldsymbol{s}_i, t_i), (\boldsymbol{s}_{i'}, t_{i'}) | \boldsymbol{\theta}_{\phi,k}\right). \tag{6}$$

For notational clarity, we have omitted the conditional dependence on hyperparameters, the number of RCPs, the exact spatiotemporal coordinates and also environmental covariates.

The conditional posterior of latent variables, given the hyperparameters, is

$$p\left(\boldsymbol{f}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu}\right) \propto p\left(\boldsymbol{f}|\boldsymbol{C}\right)\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}(\boldsymbol{f}_i)p\left(\boldsymbol{y}_i|z_{ik}=1,\boldsymbol{\mu}_k\right), \tag{7}$$

where $\boldsymbol{f}_i = [f_{i1}, \dots, f_{iK}]$ collects the latent variables at the ith observation site. This posterior has no analytical form but we can use Laplace's method (Rasmussen & Williams, 2006; Rue et al., 2009; Tierney & Kadane, 1986; Vanhatalo et al., 2010) to approximate it with a normal density with mean $\hat{\boldsymbol{f}}$ and covariance given by the inverse of $-\nabla_{\boldsymbol{f}}^2 \log p\left(\boldsymbol{f}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu}\right)$ evaluated at $\boldsymbol{f}=\hat{\boldsymbol{f}}$. Here, $\hat{\boldsymbol{f}}$ is the mode of $\log p\left(\boldsymbol{f}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu}\right)$, $\nabla_{\boldsymbol{f}}$ the gradient operator, and $\nabla_{\boldsymbol{f}}^2$ the Hessian operator with respect to $\boldsymbol{f}$. We denote this approximation by $q(\boldsymbol{f}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu})$. The mode, $\hat{\boldsymbol{f}}$, is located using a Newton algorithm as described in the Web Appendix 1. After solving for $q(\boldsymbol{f}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu})$ we apply a Monte Carlo approximation for the RCP probabilities, $\pi_k\left(\boldsymbol{x}(\boldsymbol{s}),\boldsymbol{s},t\right)$, by sampling from the multivariate Gaussian approximation for $\boldsymbol{f}$, and employing the softmax transformation (2) to obtain samples of the RCP probabilities.

For posterior distribution of latent variables at unobserved locations (prediction), denote by $\tilde{\boldsymbol{f}} = f\left(\boldsymbol{x}(\tilde{\boldsymbol{s}},\tilde{t}),\tilde{\boldsymbol{s}},\tilde{t}\right)$ the K latent function values at an unobserved location $\left(\tilde{\boldsymbol{s}},\tilde{t}\right)$. Given the posterior approximation $q(\boldsymbol{f}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu})$, we can derive a Gaussian approximation for the posterior predictive distribution for $\tilde{\boldsymbol{f}}$ given by

$$p(\tilde{\boldsymbol{f}}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu}) \approx N\left(\tilde{\boldsymbol{f}}\ \middle|\ \mathrm{E}[\tilde{\boldsymbol{f}}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu}], \mathrm{Cov}\left(\tilde{\boldsymbol{f}}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu}\right)\right). \tag{8}$$

The mean and covariance of the posterior predictive distribution are given in Web Appendix 2. The posterior distribution for RCP probabilities at unobserved locations $\pi_k\left(\boldsymbol{x}(\tilde{\boldsymbol{s}}),\tilde{\boldsymbol{s}},\tilde{t}\right)$ is again approximated by Monte Carlo by using posterior predictive samples of $\tilde{\boldsymbol{f}}$.

To estimate the hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$, we first transform them so that they are supported by the entire real line. That is, we log transform the elements of $\boldsymbol{\theta}$, and logit transform the elements of $\boldsymbol{\mu}$ to generate a vector of transformed hyperparameters $\boldsymbol{\vartheta} = [\boldsymbol{\vartheta}_\theta, \boldsymbol{\vartheta}_\mu] = [\log\boldsymbol{\theta}, \mathrm{logit}(\boldsymbol{\mu})]$. Then we use the Laplace method again to approximately integrate over the Gaussian latent variables $\boldsymbol{f}$ (Vanhatalo et al., 2010). This provides us with the approximate marginal likelihood for the hyperparameters $q(\boldsymbol{Y}|\boldsymbol{\vartheta})$. We estimate the transformed hyperparameters by their approximate maximum a posteriori (aMAP) value:

$$\hat{\boldsymbol{\vartheta}} = [\hat{\boldsymbol{\vartheta}}_\theta, \hat{\boldsymbol{\vartheta}}_\mu] = \arg\max_{\boldsymbol{\vartheta}}\ \log q(\boldsymbol{Y}|\boldsymbol{\vartheta}) + \log p(\boldsymbol{\vartheta}). \tag{9}$$

Note that the prior for the transformed parameters, $\log p(\boldsymbol{\vartheta})$, is induced by the priors for the original parameters, $\boldsymbol{\theta}, \boldsymbol{\mu}$, (see Section 3.1.1) after applying the multivariate Jacobian of the transformation. The gradients of $\log q(\boldsymbol{Y}|\boldsymbol{\vartheta})$ with respect to $\boldsymbol{\vartheta}$ can be solved analytically (Rasmussen & Williams, 2006; Vanhatalo et al., 2010), which allows gradient-based optimization for the hyperparameters. See Web Appendix I for specific details.

To guard against making inference at a local (not global) mode of parameters, we employ two strategies. The first is to seek a region of reasonable starting values—there is no point in searching for local modes far from the likely position of the maximum (Foster et al., 2013; Foster et al., 2017). The second is to perform several random starts from within this region. We implement this strategy as follows. First, we hard-clustered the observation vectors $\boldsymbol{y}_i, i = 1, \dots, n$ by using the K-means clustering (see Kaufman & Rousseeuw, 1990, for example). Then we initialized logit($\boldsymbol{\mu}$) around the logit transformed observed prevalences of each species in each hard clustered group by adding random noise, $\mathrm{N}(0, \sigma^2 = 0.2^2)$, to these logit transformed prevalences. We tested alternative SDs for the Gaussian perturbation on starting values and found that a value of 0.2 worked reasonable well. To guard against forming parameter combinations that the optimization could not escape from, we set all initial values of $\boldsymbol{\mu}$ smaller than 0.2 (greater than 0.8) to be 0.2 (0.8). We also routinely include the *unperturbed* starting values as these may reflect the RCP groups well, especially if there is only moderate spatial and covariate effects. The other hyperparameters (length-scales and variances) are initialized on the log scale using independent Gaussian realizations with SD 0.2. The mean for the log length-scale parameters was 0 and for the log variances 0.1. These initializations for covariate effects correspond to GP functions of moderate flexibility (recall that the covariates are standardized).

3.2.2 | Parameter inference using Laplace method and partial MCMC (inference method 2a and 2b)

Typically we are interested in RCP probabilities, $\pi_k\,(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{s}, t)$, and species profiles, $\boldsymbol{\mu}_k$, but not on the covariance function parameters, $\boldsymbol{\theta}$, as such. Hence, for improved posterior inference we consider MCMC schemes with increasing level of accuracy for these key parameters ($\boldsymbol{\mu}_k$) and increasing computational time demands. Our first partial MCMC approach (inference method 2a) is to sample from the conditional posterior for species profiles given the aMAP estimate for covariance function parameters and latent variables. That is, we sample from

$$p(\boldsymbol{\mu}|\boldsymbol{Y},\hat{\boldsymbol{f}}) \propto p(\boldsymbol{\mu})\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}(\hat{\boldsymbol{f}})p\left(\boldsymbol{y}_i|z_{ik}=1,\boldsymbol{\mu}_k\right). \tag{10}$$

If$\hat{\boldsymbol{f}}$is a good summary for the marginal posterior of$\boldsymbol{f}$sampling from this distribution can provide good approximation for the posterior distribution for species profiles. Alternatively, we can sample conditional on $\hat{\boldsymbol{\pi}} = \mathrm{E}[\boldsymbol{\pi}|\hat{\boldsymbol{\vartheta}}]$ which is also provided by the Laplace approximation as described in Section 3.2.1. This approach is fast since for each proposal of $\boldsymbol{\mu}$ we need to recalculate only the prior density, $p(\boldsymbol{\mu})$, and the likelihood terms $p\left(\boldsymbol{y}_i|z_{ik}=1,\boldsymbol{\mu}_k\right)$ which are computationally cheap. Moreover, given the latent variables, and our independent priors $p(\boldsymbol{\mu}) = \prod_{k,j} p(\mu_{kj})$, species profile parameters $\boldsymbol{\mu}$ may be nearly independent in the conditional posterior (10), and so constructing an efficient sampler is easy. In practice, we sampled from Equation (10) so that we first sampled from the posterior distribution of the logit transformed species profiles, $\boldsymbol{\vartheta}_{\mu}$, using Hamiltonian Monte Carlo (HMC, Neal, 2011) as implemented in the `hmc2` function of GPstuff package (Vanhatalo et al., (2013)) and then retransformed these samples back to the original parameters using $\boldsymbol{\mu} = \mathrm{logit}^{-1}(\boldsymbol{\vartheta}_{\mu})$.

As a second partial MCMC option (inference method 2b), we consider sampling from the conditional posterior for latent variables and likelihood function parameters

$$p(\boldsymbol{f},\boldsymbol{\mu}|\boldsymbol{Y},\hat{\boldsymbol{\vartheta}}_{\theta}) \propto N\left(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{C}(\hat{\boldsymbol{\vartheta}}_{\theta})\right)p(\boldsymbol{\mu})\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}(\boldsymbol{f}_i)p\left(\boldsymbol{y}_i|z_{ik}=1,\boldsymbol{\mu}_k\right), \tag{11}$$

where$\boldsymbol{f}_i = [f_{i,1}, \dots, f_{i,K}]^T$ collects the groupwise latent variables at the ith sampling site, and $\hat{\boldsymbol{\vartheta}}_{\theta}$ is the aMAP estimate for $\log\boldsymbol{\theta}$ given in Equation (9). With large datasets, the posterior distribution for $\boldsymbol{\vartheta}_{\theta}$ is narrow and the posterior for latent variables is rather insensitive to changes in covariance function parameters within the highest posterior probability region around $\hat{\boldsymbol{\vartheta}}_{\theta}$ (see, e.g., Vanhatalo et al., 2010). In many cases, the conditional posterior of Equation (11) can, thus, be good surrogate for the true marginal posterior $p(\boldsymbol{f},\boldsymbol{\mu}|\boldsymbol{Y})$.

Sampling from Equation (11) is considerably faster than sampling from the full posterior since the Cholesky decomposition and inverse of $\boldsymbol{C}(\hat{\boldsymbol{\vartheta}}_{\theta})$ need to be calculated only once before the sampling. The time needed for these operations scale as $O(n^3)$ with the number of sampling locations n after which the remaining calculations in (11) scale as $O(n^2)$. Hence, this approach is feasible in many practical applications where approximating the full posterior (requiring multiple $O(n^3)$ operations) would be infeasible. In practice, we do Gibbs sampling by sequentially sampling from the conditional distributions $p(\boldsymbol{\mu}|\boldsymbol{Y},\hat{\boldsymbol{\vartheta}}_{\theta},\boldsymbol{f})$ and $p(\boldsymbol{f}|\boldsymbol{Y},\hat{\boldsymbol{\vartheta}}_{\theta},\boldsymbol{\mu})$. We again use HMC for the former conditional and an elliptical slice sampler (Murray et al., 2010) (implemented in `esls` function in the GPstuff package) to sample from the latter.

3.2.3 | Parameter inference using full MCMC (inference method 3)

Asymptotically the most accurate, but also computationally most demanding, MCMC approach is to sample from the full posterior $p(\boldsymbol{f},\boldsymbol{\mu},\boldsymbol{\theta}|\boldsymbol{Y})$. We again apply Gibbs sampling and sample the latent variables, species profiles and covariance function parameters from their full conditionals. We use an elliptical slice sampler for latent variables ($p(\boldsymbol{f}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\mu})$), HMC for species profiles ($p(\boldsymbol{\mu}|\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{f})$) and slice sampling for the covariance function parameters ($p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{\mu},\boldsymbol{f})$). The latter is implemented as described by Neal (2003) and implemented in function `sls` in GPstuff. The species profiles, $\boldsymbol{\mu}$, are sampled at the logit transformed space and the covariance function parameters, $\boldsymbol{\theta}$, are again sampled in log transformed space.

In full MCMC, the computational bottle necks are the covariance matrix operations, which have to be done for each proposal of covariance function parameters and latent variables. Moreover, strong posterior dependence can occur between the covariance function parameters and latent variables (Vanhatalo et al., 2010). As a result, the full MCMC approach was infeasible for the NWS fish data due to the number of sampling sites and the number of fish species. For that reason, we compared the Laplace approximation and the two above mentioned MCMC methods to the full MCMC using a smaller simulated dataset (Section 4.1), which showed that the Laplace approximation for the posterior of the latent variables and the conditional MCMC approaches agree well with the full MCMC approximation.

3.2.4 | Identifiability of the parameters and inferring covariate effects

Our model has several components that impose identifiability considerations. First, the components of the latent function $f_k(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{s}, t) = \alpha_k + h_k(\boldsymbol{x}) + \phi_k(\boldsymbol{s}, t)$ are unidentifiable since the overall level α_k can be absorbed by $\phi_k(\boldsymbol{s}, t)$ and the GP formulation of $h_k(\boldsymbol{x})$ (Knorr-Held, 2000). In our model, we note that this confounding is mitigated by the specification of priors for the parameters of the GP components; the effects are shrunk to zero in the absence of a spatial or temporal trend. A remedy leading to explicit identifiablity, as proposed by Knorr-Held (2000) and Hanks et al. (2015), would be to impose sum to zero constraints over the observation locations to the random effects $\phi_k(\boldsymbol{s}, t)$ and the GP based response functions $h_k(\boldsymbol{x})$. Second, due to the sum constraint of the probabilities, the softmax function, as defined in Equation (2), is not identifiable for the unnormalized latent functions $f_k(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{s}, t)$ with $k = 1, \ldots, K$. A common remedy to make the latent functions identifiable over groups (up to label-switching, see the discussion at the end of this section) is to fix the latent function of one of the groups to zero (see, e.g., Foster et al., 2013; Neelon et al., 2014; Foster et al., 2018).

Apart from prior distributions, we did not apply any of the above proposed identifiability constraints for the latent function components. The reasons are the following. Due to the nonlinear softmax link function, the covariate effects or spatiotemporal effects to RCP areas are hard to interpret by looking at the latent responses, such as $h_k(x_d)$, only. Hence, we do not want to interpret the latent functions,f_k, $k = 1, \ldots, K$ or their additive components directly. Rather, we want to interpret normalized RCP probabilities $\pi_k(\boldsymbol{x}, \boldsymbol{s}, t)$ of Equation (2). Even though the latent functions or their additive components are not identifiable, the model is identifiable for these RCP probabilities. From the point of view of interpreting RCP probabilities ($\pi_k(\boldsymbol{x}, \boldsymbol{s}, t)$) we would not gain anything from posing identifiability constraints to the latent variables. On the other hand, implementing these constraints in our inference methods (especially the Laplace approximation) would be cumbersome and lead to increase in computational demand. Hence, the implementation of our model was more straightforward by allowing the latent functions of all groups to vary.

We follow Hill et al. (2017) and Kallasvuo et al. (2017) and apply the posterior inference on the conditional responses of RCP probabilities π_k at different covariate combinations. We denote the conditional response by $\pi_k(x_d|\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t)$ as the probability of the kth RCP as a function of the dth covariate only, conditioning on the remaining covariates ($\boldsymbol{x}_{\setminus d}$) and spatial and temporal locations are fixed to ($\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t$) (note that π_k is a function of x_d, not a probability density function and $|$ is a similar conditioning statement as used, for example, in likelihood function notation). The conditional response is a random function whose posterior distribution depends on the posterior distribution of the model parameters. We study the probability changes relative to the average probability within the covariate limits at ($\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t$):

$$\Delta\pi_k(x_d|\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t) = \pi_k(x_d|\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t) - \int_{\min(x_d)}^{\max(x_d)} \pi_k(x_d|\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t) dx_d / (\min(x_d) - \max(x_d)), \tag{12}$$

Since the normalized RCP probabilities are identifiable (up to label switching) these conditional responses are identifiable as well.

In the NWS analysis, we plot the expectation of the conditional responses, $E\left[\Delta\pi_k(x_d|\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t)\right]$, with covariates from 50 randomly chosen sampling sites from the original survey data, where the expectation is taken over the posterior distribution of π_k where the posterior distribution for π_k is estimated with one of the inference methods described above. We add to this plot an average response, which is taken over the 50 conditional responses in a pointwise manner. Note that we could also report the posterior uncertainty in individual $\Delta\pi_k(x_d|\boldsymbol{x}_{\setminus d}, \boldsymbol{s}, t)$ but we suppress this information to reduce clutter in our plots.

One potential additional complication to the posterior inference is label-switching (Neelon et al., 2014), which in our application means that the posterior inference is essentially identical if the order of the RCP class memberships change. Whilst a problem for interpretation, label switching forms little issue for numerical optimization. This is because any ordering is as good as any other, and the estimation goal is to find any one of the MAPs. Hence, label switching is not a problem for Laplace approximation nor inference method 2a in Section 3.2.2 since the Laplace approximation for the posterior of latent variables will be formed around a single class membership combination. Unlike optimization, label switching can be problematic in MCMC (Stephens, 2000) as the sampler can jump between different states. However, we did not encounter obvious problems in our MCMC routines. This is reasonable since the RCP profiles, $\boldsymbol{\mu}_k$, are typically very long vectors so the posterior distributions conditional on different label combinations are naturally far from each other. If label-switching was a problem in the MCMC inference, we could follow the relabeling strategy proposed by Stephens (2000).

3.3 | Model comparison via cross-validation

Even though comparison of models with different structures and the choice of number of RCPs is not our focus in this work, we suggest that this could be performed using cross-validation with the average log posterior predictive density as the performance measure (Vehtari & Ojanen, 2012). Cross-validation approach has been utilized in mixture models previously by Wall and Liu (2009). Briefly, we divide the data into 10 randomly chosen parts of equal size, and predict each of these hold-out subsets based on the remaining nine subsets. The predictive performance was assessed using the average log posterior predictive density of the hold-out datasets based on the model estimated from the remaining data. We hold out the same partitions for all the different models under consideration, which can help guard against stochastic noise.

As with any mixture model, detecting too many groups from predictive performance is problematic. In the context of RCPs, any RCP can be split into two or more RCPs, with the same profile and essentially the same log posterior predictive density for the hold out data. Moreover, finite data and randomness in CV splitting induce randomness to cross-validation log predictive densities through which overly complex model can perform best by chance. Hence, there is risk of overfitting if we keep increasing RCPs until a model's log predictive density starts decreasing. For this reason, we calculate also the standard error of the average log predictive density of the hold out data and interpret it as an estimate of the randomness in the cross-validation result (Vehtari & Ojanen, 2012, p. 191). We then choose the largest number of RCPs that increases the average log predictive density by more than 3/2 standard errors compared with the model with one RCP less. This model choice criterion corresponds, roughly, to choosing a model that is better than the alternative models with 90% confidence level.

As an additional, qualitative check, we follow Paci and Finazzi (2018, section 3.4) who proposed that the choice of the number of clusters (K) should not be solely based on model performance metric but should also consider if adding an extra cluster significantly changes interpretation of the model. Hence, if two models have similar cross-validation log predictive densities but the RCP distribution or species profiles do not differ significantly between models, we should prefer the model with the smaller number of RCPs.

4 | EXPERIMENTS

4.1 | Tests with simulated data

In order to test the goodness of our approximate inference methods (Sections 3.2.1–3.2.3) we conducted a simulation study. We used the same simulation set up as described by Foster et al. (2013) with presence/absence observations of $J = 100$ species at $n = 150$ randomly distributed observation locations throughout a spatial rectangular area of size $[-10, 10] \times [-10, 10]$. The species data at simulated sampling sites were generated using the clustering model (1) and each sampling site was probabilistically assigned to one of $K = 3$ RCP clusters. The RCP probabilities were defined as cubic polynomial functions of spatial location. In addition to the simulation model in Foster et al. (2013) we add spatial variation using the Matérn covariance function in the GP prior. We conducted the posterior analysis using the Laplace approximation (Section 3.2.1) and the three MCMC schemes described in Sections 3.2.2 and 3.2.3.

Figure 2 summarizes the posterior distribution of the RCP cluster parameters $\pi_k(\boldsymbol{s})$ as approximated by the full MCMC and Laplace approximation. Both approximations give practically identical posterior mean estimates and very similar results for the 5% and 95% posterior quantiles. The Laplace approximation describes the center of the posterior predictive distribution well, but seems to overestimate the 5% quantile and underestimate the 95% quantile at a small number of locations. Figure 3 summarizes the posterior distribution of species profiles, $\mu_{k,j}$, as approximated by MCMC for species profiles at $\hat{\boldsymbol{f}}$ (inference method 2a), MCMC for $\boldsymbol{\mu}$ and $\boldsymbol{f}$ at $\hat{\boldsymbol{\theta}}$ (inference method 2b) and by full MCMC (inference method 3). There are no systematic differences between these approximations and all three methods provide close to identical

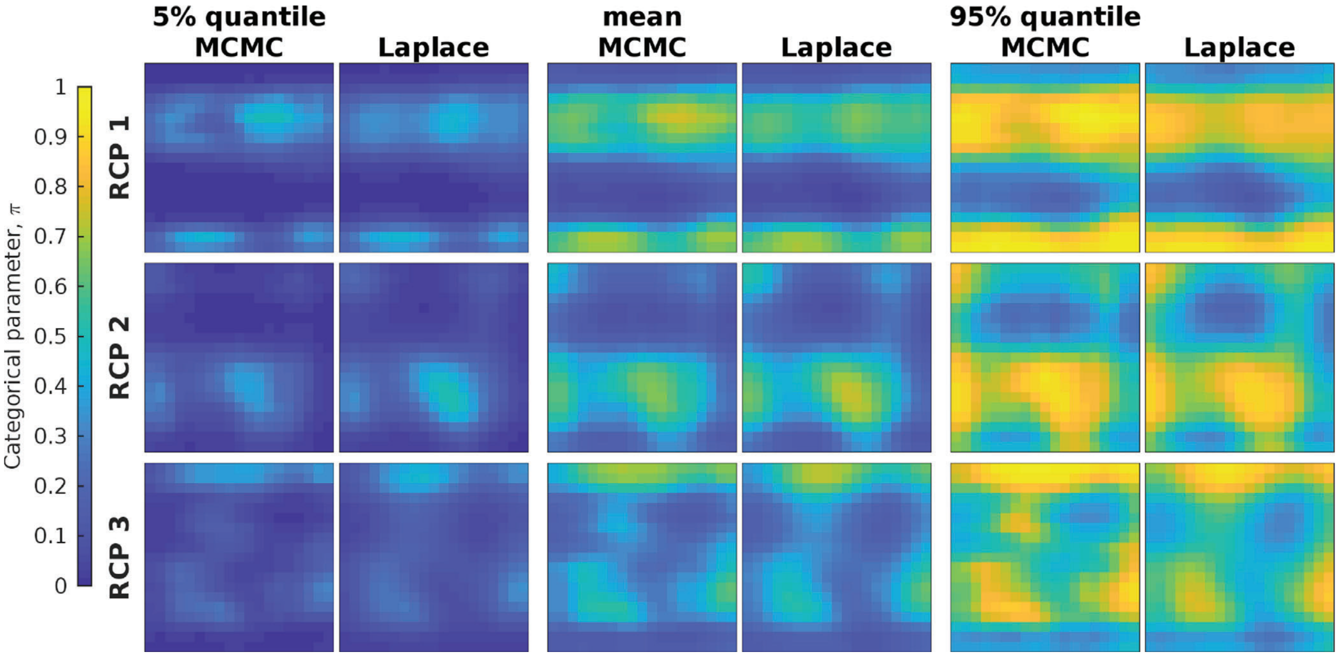

FIGURE 2 Visualization of the posterior distribution for $\pi_k, k = 1, 2, 3$ over the study region in the simulated data experiment as approximated by full MCMC (inference approach 3) and Laplace approximation (inference approach 1) for each RCP at all spatial locations. Each of the 18 subplots covers the simulated spatial study region of size $[-10, 10] \times [-10, 10]$. The rows correspond to different RCPs so that the first row contains results for π_1, the second for π_2 and the last row for π_3. The two left most columns show the 5% lower posterior quantile, the two middle columns show the posterior mean and the two right most columns show the 95% posterior quantiles for MCMC and Laplace approximation. MCMC, Markov chain Monte Carlo; RCP, regions of common profile.

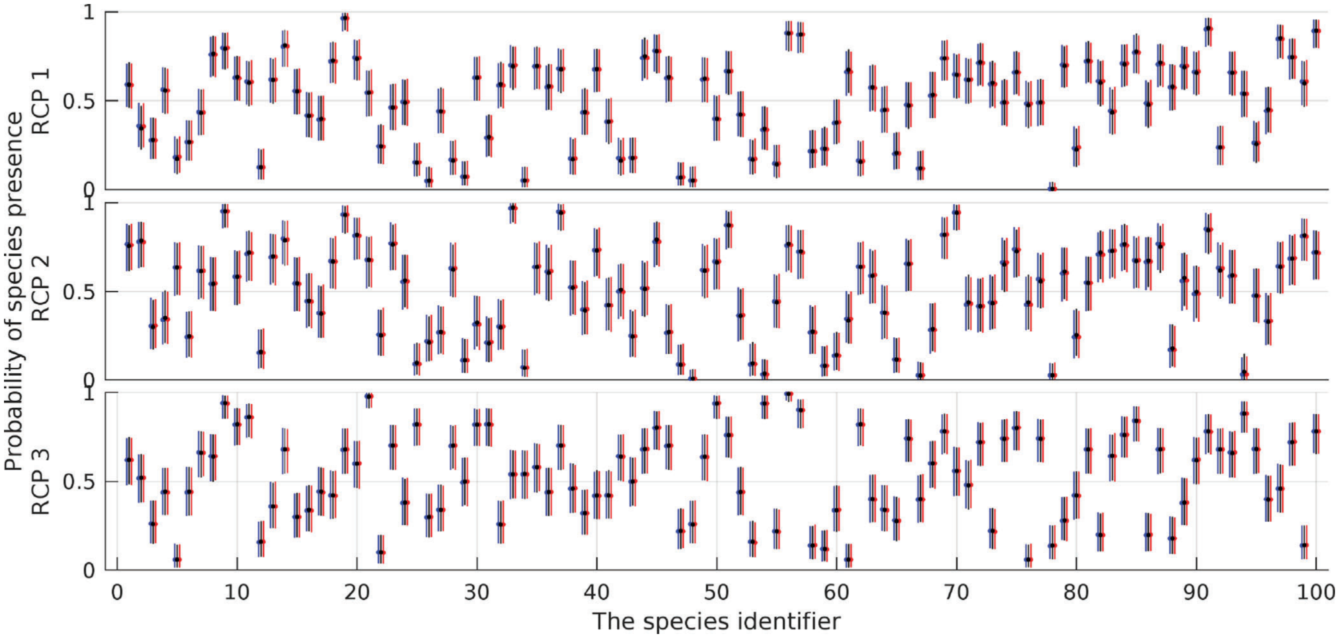

FIGURE 3 Visualization of the posterior distribution for species profiles, $\boldsymbol{\mu}_k, k = 1, 2, 3$ in simulated data as approximated by full MCMC (blue; inference method 3), MCMC for latent variables and likelihood function parameters at the aMAP estimate of covariance function parameters (red; inference method 2b) and MCMC for likelihood function parameters only at aMAP estimate of latent variables (black; inference method 2a). The lines show the 95% central credible interval and the dots show the posterior mean. MCMC, Markov chain Monte Carlo

TABLE 1 The models considered for analysis of the NWS fish data

	Correlation structure	Covariate effects
M1	None	Quadratic
M2	None	GP
M3	Spatial	Quadratic
M4	Spatial	GP
M5	spatiotemporal	GP

Note: The model defined in Foster et al. (2013) corresponds to M1.
Abbreviations: GP, Gaussian processes; NWS, north-west shelf.

uncertainty estimates for the species profiles. Table S1 in the Supplementary Material shows the 10-fold CV comparison between models with 2–5 RCPs. The model with correct number of RCPs (three) has the highest mean log predictive density. However, the difference in posterior predictive performance of models with 2–4 RCPs is small compared with the standard error of the log predictive density estimates indicating that the choice of number of RCPs should not be based only on the average log predictive density comparisons as discussed in Section 3.3.

In our experiments it took approximately 20 s on a Linux laptop with Intel(R) i7-6600U CPU @ 2.60 GHz processor to find the Laplace approximation for this simulation study. Sampling 10^4 samples from the conditional posterior of species profiles (Equation (10), inference method 2a) took another 30 s. Sampling 10^4 samples of latent variables and species profiles at the aMAP of covariance function parameters (Equation (11), inference method 2b) took approximately an hour and sampling 10^4 samples from the full posterior (inference method 3) took approximately 3 h. These performance statistics are naturally highly dependent on the sampler options and, hence, only reflect the relative performance differences between the methods. The reported values correspond to sampling after careful tuning.

4.2 | Analyses of NWS data

4.2.1 | Posterior inference and model comparison

To contrast different models and to show the real-world utility of the model, we analyze the NWS data with nonspatial, spatial and spatial-temporal models—see Table 1. For the model without spatial correlation, we use only environmental covariates to predict the RCP locations using either quadratic or GP responses (models M1 and M2). For the models that incorporate spatial correlation, we additionally employ GPs over space (models M3 and M4). Finally, we also analyze the data using a model with GP covariate effects and the spatiotemporal effects (model M5). We assessed a range of numbers of RCPs, from 2 to 6, which covers the $K = 5$ solution that Foster et al. (2013) found best for data collected in 1983 from the same region.

The predictive performance of the models increases significantly until $K = 5$ RCPs since, for the first five RCPs, each additional RCP increases the average log predictive density by more than 3/2 standard errors of its estimate (see Table S2 in the Supplementary Material). Increasing the number of RCP groups beyond $K = 5$ produced RCPs that are small (represented by few sites) and are only minor variations of existing ones. Decreasing the number of RCP groups produced models that have amalgamations of these five main RCPs. In particular, with four RCP groups the groups 3 and 4 (see Figure 4) would be merged. These two RCP groups are distributed in similar spatial locations along the depth gradient but they show opposing effects along the salinity gradient (see Figures 4 and 5) which is reflected by differences in the fine scale structure of the spatial distribution of these RCP regions. The difference in their species profiles are also significant (see Table 3 and Figure 4). Hence, even though these RCP regions show similar spatial patterns we concluded that they represent significantly different communities and used the models with $K = 5$ RCPs for the final analyses.

For the NWS data, we did all computations for this case study with a Linux desktop with Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz. Finding the aMAP estimate for hyperparameters took 4–7 h after which MCMC for conditional posterior for species profiles only was done in less than a minute and species profiles and latent variables within 3 h.

4.2.2 | The effect of covariates and the spatial term

The predicted maps of expected probability of each RCP from the models with GP (M2) and quadratic (M1) covariate effects are presented in Figures 4 and S1, respectively. In both model types (GP and quadratic), the addition of a spatial

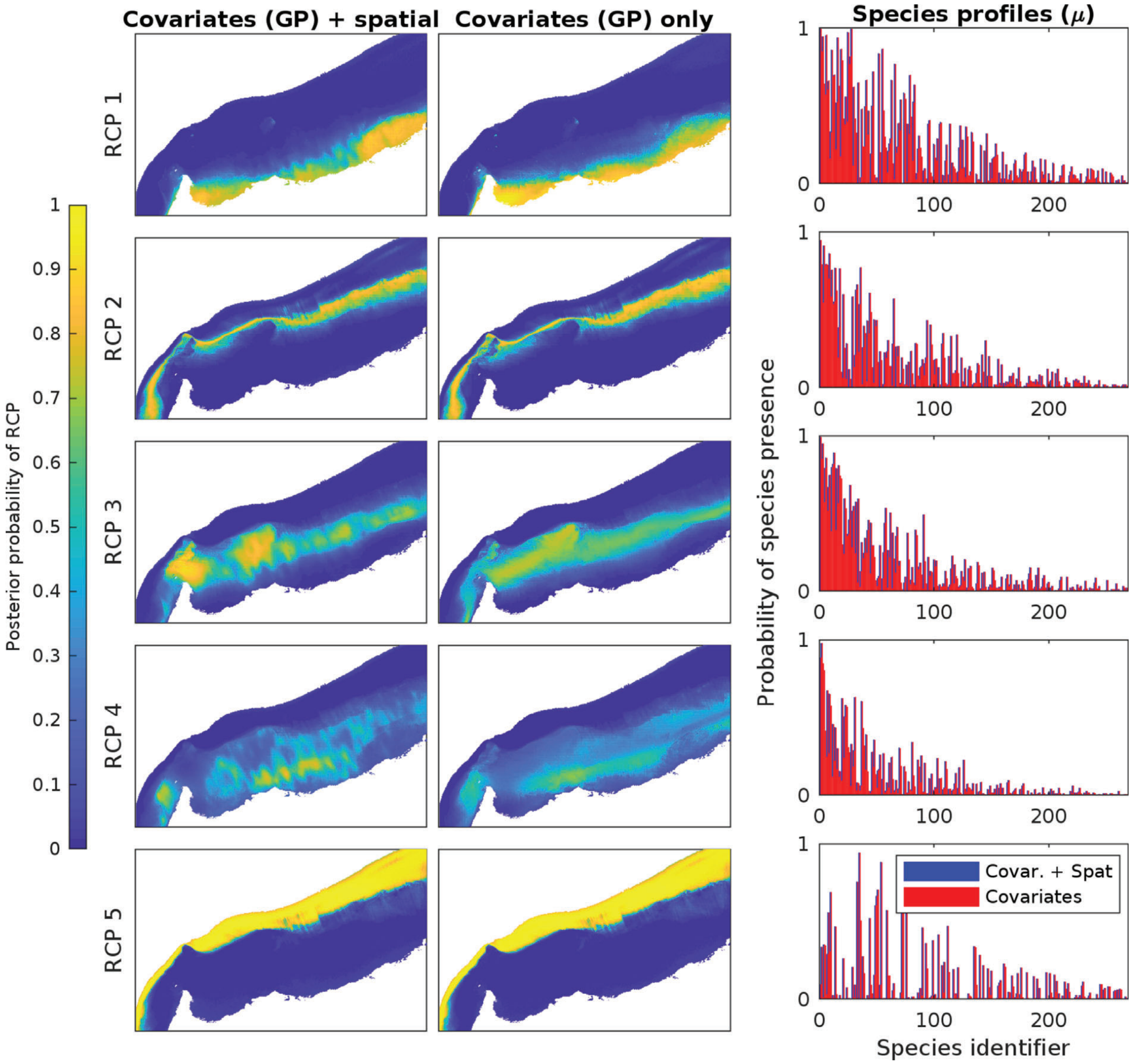

FIGURE 4 The posterior expected probability of each RCP at all spatial locations (left and middle columns). Note, the pixelwise probabilities in the images sum to one over the rows in the first two columns. The left most column is for the model with spatial and GP covariate effects (model M4). The middle column does not have the spatial effects (model M2). The column on right shows species profiles (the aMAP estimate of the probability of observing species) in each RCP. Ordering, from most prevalent species to least prevalent is for visual appeal only and it does not alter the model in any way. GP, Gaussian processes; RCP, regions of common profile

effect increases the contrast between areas of high and low probabilities of many RCPs (e.g., RCPs 3 and 4)—compare model M1 with M3, and M2 with M4. For quadratic covariate effects (M1), the spatial distribution of RCPs 3 and 4 also change noticeably with the addition of spatial effects (M3), presumably due to the effects being nonquadratic.

The maps produced by the two models with both covariates and spatial effects (left columns of Figure 4 for M4 and Figure S1 for M3) are qualitatively similar implying that the combined effect of covariates and space is similar irrespective of the form of the covariate contributions. However, if the spatial effect is removed (giving models M2 and M1) then the models do differ, and substantially so for RCPs 3 and 4 (center columns of Figures 4 and S1). RCP 2 in particular is not as sharply defined using quadratic covariate effects (model M1 in Figure S1). This implies that, for these data, the importance of using flexible covariate effects (model M2) is lessened when the spatial effect is present (model M3). The reason for this is that the functional form of the covariates is approximately quadratic when spatial effects are present (model M4, see Figure 5). The addition of spatial random effects to the model increased the uncertainty estimates of some of the model components, mainly the estimates concerning the responses along environmental covariates.

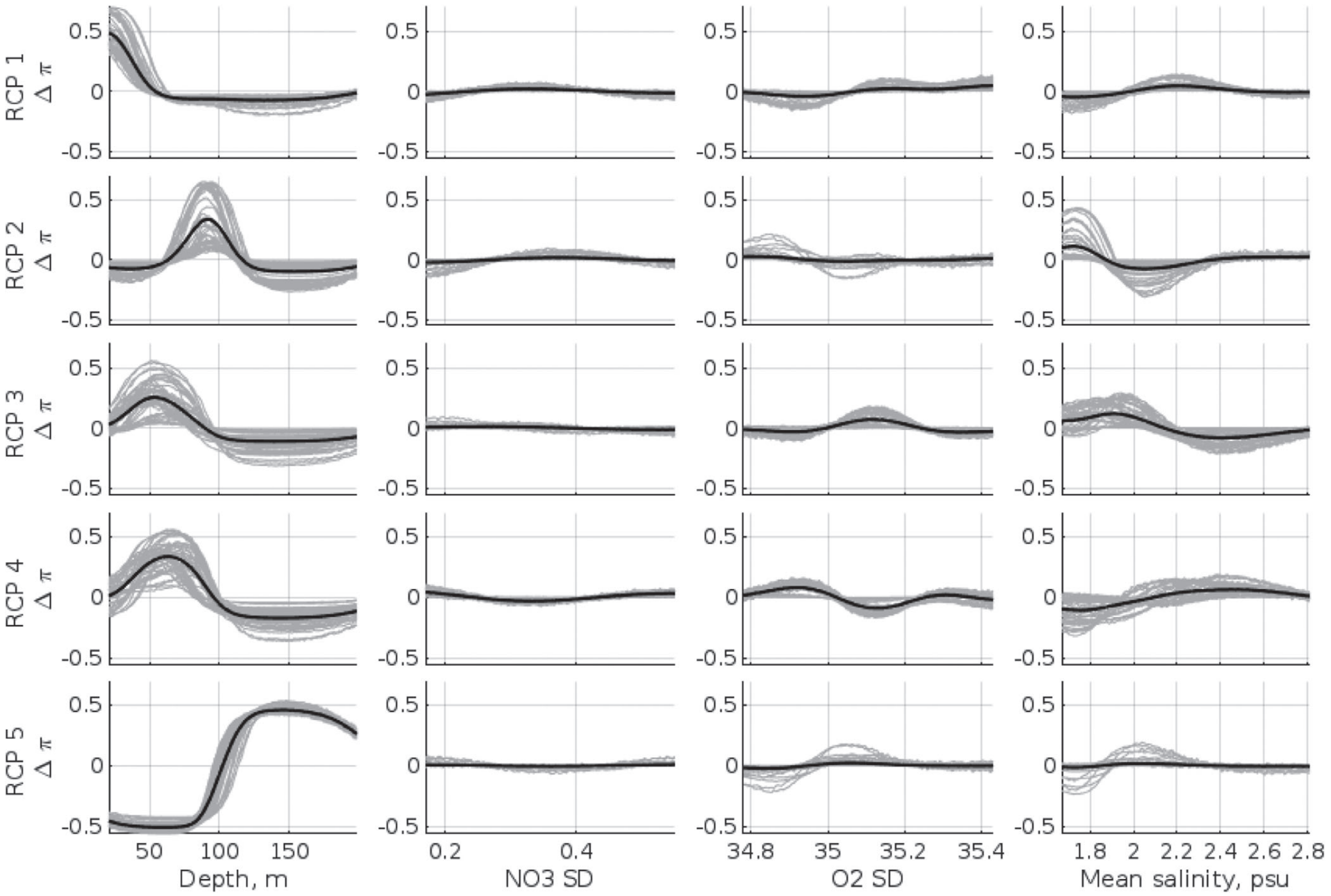

FIGURE 5 The responses of RCP probabilities along covariates in the 5RCP model with spatial random effects and covariates (GP; model M4). The rows correspond to the RCPs from $k = 1$ to $k = 5$ and the columns correspond to the environmental covariates. For details of the statistic plotted, $\Delta\pi$, see Section 3.2.4. GP, Gaussian processes; RCP, regions of common profile

TABLE 2 The posterior mean (and 95% central credible interval) for expected species-richness for each RCP for the NWS spatial model (using GP effects for the covariates, model M4)

	Mean richness	Lower CI	Upper CI
RCP 1	51.6	50.8	52.5
RCP 2	38.4	37.7	39.4
RCP 3	40.0	39.3	40.7
RCP 4	25.7	25.1	26.3
RCP 5	24.2	23.7	24.7

Note: For a map of likely location of each RCP type, see Figure 4 (left column).
Abbreviations: GP, Gaussian processes; NWS, north-west shelf; RCP, regions of common profile.

For these data, the major environmental driver is depth (Figure 5). This matches ecological understanding of the region, and of marine ecosystems in general (Hill et al., 2017; Koslow et al., 1997). In the NWS data, there appears to be: a shallow water group (RCP 1) located near the coast; two mid-depth groups (RCPs 3 and 4) that are spatially segregated but not in terms of their depth preference; one deeper water group (RCP 5), and; one group located near the start of the more rapid change in the depth gradient (RCP 2).

The species-richness patterns varied across RCPs and hence depth. Species-richness is defined as the number of different species at a sampling location, and here we quantify this as the expected richness within an RCP group: $\sum_{j=1}^{J} \mu_{kj}$ (Table 2). We also calculated the total absolute difference in species profiles between two RCPs: $\sum_{j=1}^{J} |\mu_{kj} - \mu_{k'j}|$ (Table 3).

TABLE 3 The posterior mean (and 95% central credible interval) for the total absolute difference in species profiles between RCP for the NWS spatial model (using GP effects for the covariates, model M4)

	RCP 1	RCP 2	RCP 3	RCP 4
RCP 2	48.5 (47.2, 49.7)			
RCP 3	32.0 (31.0, 33.1)	28.1 (27.0, 29.2)		
RCP 4	35.0 (34.0, 35.9)	32.6 (31.6, 33.8)	27.0 (26.1, 27.8)	
RCP 5	61.8 (60.9, 62.8)	34.7 (33.5, 35.7)	49.1 (48.2, 49.9)	35.1 (34.4, 35.9)

Note: For a map of likely location of each RCP type, see Figure 4 (left column).
Abbreviations: GP, Gaussian processes; NWS, north-west shelf; RCP, regions of common profile.

The shallow group (RCP 1) has a greater expected species-richness, while the deeper group (RCP 5) is less rich but tends to have quite a different set of species in the RCP assemblage. The RCPs with intermediate depth have richness that is in between these two (Table 3). RCP 4 arguably has less richness than one might expect given its intermediate depth. There could be a number of ecological reasons for this: (1) its location over a very particular habitat, which requires highly specialized traits; (2) the dominance of a small number of species monopolizing the resources at those sites (an uneven community), or; (3) this group is the result of the sampling process itself (e.g., the trawls were performed using slightly different protocols that affected catch diversity and/or rates. We are unable to investigate which of these three options is appropriate without extra information that is not available within the data themselves.

4.2.3 | Spatiotemporal analysis of the NWS data

To investigate possible temporal, as well as spatial, heterogeneity in the fish data, we extended the model to have spatiotemporal dependence (as per Section 3.1). This is model M5 in Table 1. In particular, we fitted a model with covariate effects given by GPs, a Matérn spatial dependence and an exponential temporal dependence as in Equation (4). Figure 6 shows the predicted maps of posterior expected probability of each RCP for years 1986–1997. The spatiotemporal model produces in general similar results as the spatial only model. There is relatively little change in the RCP clustering through time. This is important from a natural resource management perspective, as it means that many management decisions (like zonation) were likely to be enduring and so do not have the need for frequent reassessment.

Whilst remaining relatively static, there are some minor differences over the years for some of the RCP groups. For example, RCP 4 is becoming more common in a small patch near the center of the study area and RCP 1 may be retracting from its southern areas (Figure 6). Conversely, RCP 5 seems to change little over the study period and RCP 3 also experiences only minor changes.

5 | SUMMARY AND DISCUSSION

In this work we have developed a statistical model that groups samples according to their multivariate observations. The model extends the model of Foster et al. (2013) by introducing spatial, and spatiotemporal effects to deal with correlation. The methodology works by allowing the probability of any particular sample belonging to each group to depend on covariates and also their location in space and time. This is achieved by adding flexible spatial and spatiotemporal terms into the mixture-of-experts model (Foster et al., 2013; Jacobs et al., 1991; Jordan & Jacobs, 1994), see Section 3.1. An important benefit of incorporating spatial and spatiotemporal terms into the model as GP effects is that it changes the qualitative nature of posterior prediction at locations that have not previously been sampled. This is done by leveraging observed data from nearby (in space and time) locations. This enables cohesive spatial posterior prediction (even at the locations where data was sampled) and a probabilistic representation of uncertainty. In addition, we have introduced flexible Gaussian process methods to model dependence on environmental covariates.

We develop efficient techniques to estimate the models using approximate Bayesian methods for inference. We used this novel methodology to analyze a dataset of tropical fish distribution on the continental shelf of northwestern Australia to show important relationships with physical covariates and spatial dependence as well as the temporal stability of species groups (RCPs). Our analysis suggests that depth is the major delineating variable between the RCP groups (Figure 5), which agrees with ecological understanding of fish behavior elsewhere and also for other taxonomic groups

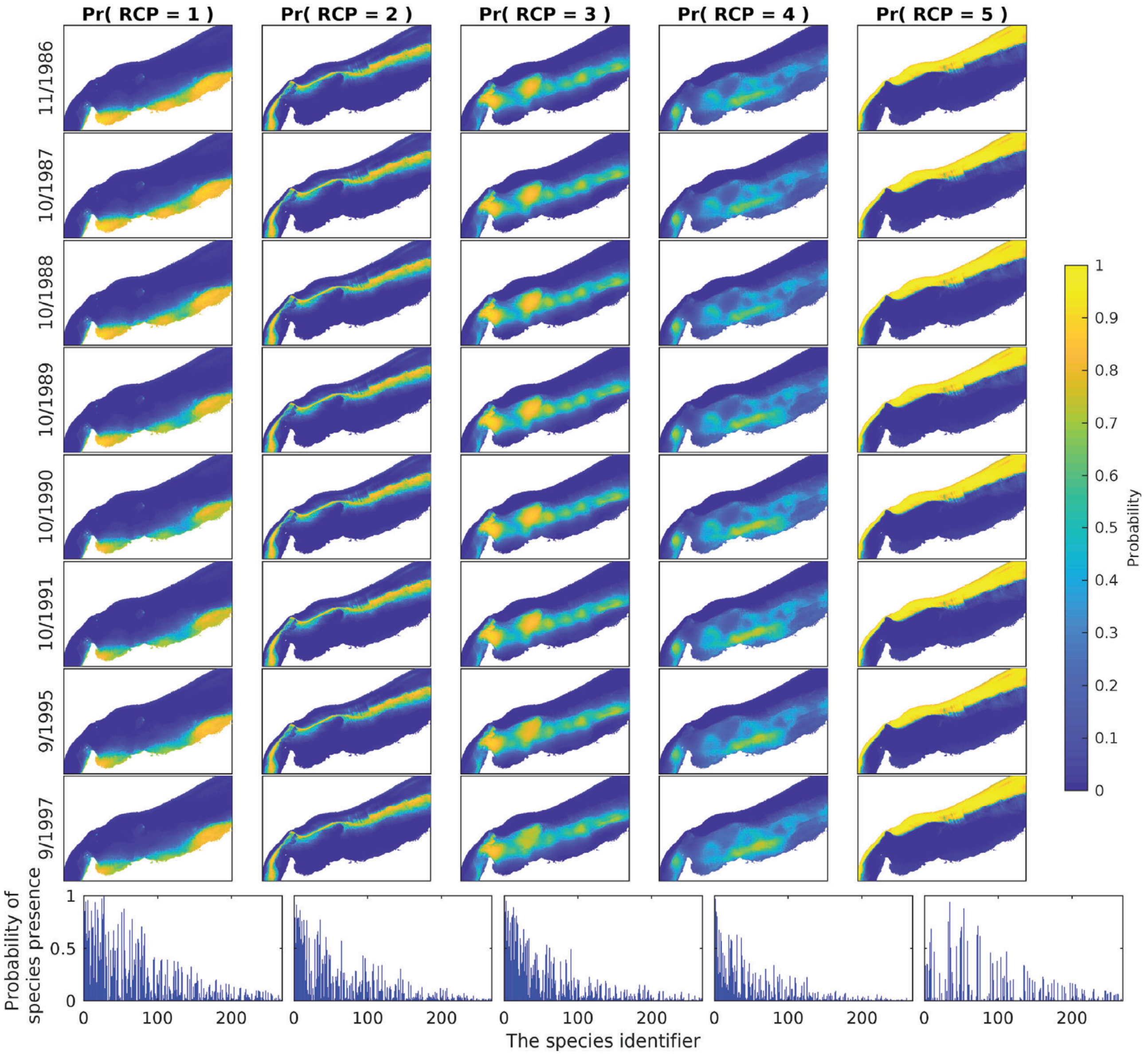

FIGURE 6 The posterior expected probability of each RCP at all spatial locations for years 1986–1997. These maps are created using the spatiotemporal model with covariates effects added using GPs (model M5). Only years when data were collected are presented. GP, Gaussian processes; RCP, regions of common profile

(Hill et al., 2017; Koslow et al., 1997). Interestingly, the analysis suggests that the temporal component of variation in the data is relatively minor. This is in spite of the changing human utilization in the area.

Our computational strategy is efficient enough to allow estimation for our example application, which has ~850 samples and ~250 measurements (species presence/absence) per sample given five RCP clusters. The size of this modeled dataset is large compared with earlier spatial clustering examples in the literature (Alfó et al., 2009; Green & Richardson, 2002; Lawson et al., 2017; Neelon et al., 2014; Torabi, 2016; Wall & Liu, 2009). To enable inference, we performed approximate Bayesian inference, as defined by a Laplace approximation and MCMC for conditional posterior of latent variables and species profiles (Sections 3.2.1 and 3.2.2). According to the simulation study, these methods provide good approximation for the posterior of latent variables, posterior probabilities of RCP regions and the posterior of species profiles. Full MCMC (Section 3.2.3), however, was infeasible for the NWS data in a reasonable time. The time requirement of Laplace approximation and the full MCMC increases as $O(n^3)$ but the constant factor is considerably smaller for Laplace approximation (only tens of optimization steps) compared with full MCMC (thousands of sample proposals). The time requirement of sampling only latent variables and species profiles increase as $O(n^2)$ whereas the time needed to sample only species profiles increases as $O(J)$. Hence, full MCMC will become increasingly hard as the number of sampling sites increases but the two partial MCMC schemes (Sections 3.2.1 and 3.2.2) will remain attainable as long as constructing the Laplace approximation is feasible.

The model in Section 3.1 further extends the earlier spatial clustering models by allowing more flexible nonparametric responses to environmental covariates through the use of GPs. In the NWS study, this added flexibility was ultimately not needed as the expected latent response tended to be near quadratic. However, this was only the case for models with a spatial effect. Without the spatial effect, the resulting maps for the GP model and the quadratic model were noticeably different. Hence, it cannot be assumed in general that the relationship between a covariate and the response will necessarily be quadratic, or will even follow a less strict form of ecological niche theory such as unimodality. The GP approach provides added value compared with parametric response functions.

ACKNOWLEDGMENTS

We would like to thank the generous financial support of Academy of Finland (grant 317255), Research Funds of the University of Helsinki (decision No. 465/51/2014) and the Data Analytics Program of Data61 (CSIRO). The authors also acknowledge CSC–IT Center for Science, Finland, for computational resources. We would like to thank K. Sainsbury, F. Althaus, R. Pitcher, P. Brodie, and M. Fuller for advice about details of the data.

ORCID

Jarno Vanhatalo https://orcid.org/0000-0002-6831-0211
Scott D. Foster https://orcid.org/0000-0002-6719-8002
Geoffrey R. Hosack https://orcid.org/0000-0002-6462-6817

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. *The Journal of the Royal Statistical Society – Series B*, *44*, 139–177.

Alfó, M., Nieddu, L., & Vicari, D. (2009). Finite mixture models for mapping spatially dependent disease counts. *Biometrical Journal*, *51*, 84–97.

Ambroise, C., Dang, M., & Govaert, G. (1997). Chapter. Clustering of spatial data by the EM algorithm. *geoENV I — Geostatistics for Environmental Applications: Proceedings of the Geostatistics for Environmental Applications Workshop, Lisbon, Portugal, 18–19 November 1996*, Springer Netherlands, Dordrecht. (pp. 493–504).

Anderson, C., Lee, D., & Dean, N. (2014). Identifying clusters in Bayesian disease mapping. *Biostatistics*, *15*, 457–469.

Bilancia, M., & Demarinis, G. (2014). Bayesian scanning of spatial disease rates with integrated nested Laplace approximation (INLA). *Statistical Methods & Applications*, *23*, 71–94.

Considine, M. L. (1985). Small fry or big fish on the North West shelf. *ECOS*, *44*, 23–28.

Corander, J., Sirén, J., & Arjas, E. (2008). Bayesian spatial modeling of genetic population structure. *Computational Statistics*, *23*, 111–129.

Cressie, N., & Wikle, C. (2011). *Statistics for spatio-temporal data*. Wiley.

Daganzo, C. F. (1979). *Multinomial probit: The theory and its application to demand forecasting*. Academic Press.

Driver, H. E., & Kroebe, A. L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology*, *31*, 211–256.

Foster, S., Givens, G., Dornan, G., Dunstan, P., & Darnell, R. (2013). Modelling biological regions from multi-species and environmental data. *Environmetrics*, *24*, 489–499.

Foster, S. D., Feutry, P., Grewe, P. M., Berry, O., Hui, F. K. C., & Davies, C. R. (2018). Reliably discriminating stock structure with genetic markers: Mixture models with robust and fast computation. *Molecular Ecology Resources*, *18*(6), 1310–1325.

Foster, S. D., Hill, N. A., & Lyons, M. (2017). Ecological grouping of survey sites when sampling artefacts are present. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*, 1031–1047.

Gelfand, A., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

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

Gómez-Rubio, V., & Rue, H. (2018). Markov chain Monte Carlo with the integrated nested laplace approximation. *Statistics and Computing*, *28*, 1033–1051.

Green, P. J., & Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, *97*, 1055–1070.

Guillot, G., Estoup, A., Mortier, F., & Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics*, *170*, 1261–1280.

Hanks, E. M., Schliep, E. M., Hooten, M. B., & Hoeting, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, *26*, 243–254.

Hill, N. A., Foster, S. D., Duhamel, G., Welsford, D., Koubbi, P., & Johnson, C. R. (2017). Model-based mapping of assemblages for ecology and conservation management: A case study of demersal fish on the Kerguelen plateau. *Diversity and Distributions*, *23*, 1216–1230.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computing*, *3*, 79–87.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computing*, *6*, 181–214.

Juntunen, T., Vanhatalo, J., Peltonen, H., & Mantyniemi, S. (2011). Bayesian spatial multispecies modelling to assess pelagic fish stocks from acoustic- and trawl-survey data. *ICES Journal of Marine Science*, *69*, 95–104.

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

Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley and Sons.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, *19*, 2555–2567.

Koslow, J. A., Williams, A., & Paxton, J. R. (1997). How many demersal fish species in the deep sea? A test of a method to extrapolate from local to global diversity. *Biodiversity & Conservation*, *6*, 1523–1532.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, *26*, 1481–1496.

Lawson, A. B., Carroll, R., Faes, C., Kirby, R. S., Aregay, M., & Watjou, K. (2017). Spatiotemporal multivariate mixture models for Bayesian model selection in disease mapping. *Environmetrics*, *28*, e2465.

McLachlan, G., & Peel, D. (2000). *Finite mixture models Wiley Series in Probability and Statistics* (1st ed.). Wiley-Interscience.

Murray, I., Adams, R., & MacKay, D. (2010). *Elliptical slice sampling*. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, volume 9 of Proceedings of Machine Learning Research* (pp. 541–548). Chia Laguna Resort, PMLR.

Murtagh, F., & Kurtz, M. J. (2016). The classification society's bibliography over four decades: History and content analysis. *Journal of Classification*, *33*, 6–29.

Neal, R. (2011). *MCMC using Hamiltonian dynamics*. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 113–162). Chapman & Hall/CRC Press.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*, 705–767.

Neelon, B., Gelfand, A. E., & Miranda, M. L. (2014). A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*, 737–761.

Nguyen, T. M., & Wu, Q. M. J. (2012). Gaussian-mixture-model-based spatial neighborhood relationships for pixel labeling problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *42*, 193–202.

Nowara, G., & Newman, S. (2001). *A history of foreign fishing activities and fishery-independent surveys of the demersal finfish resources in the Kimberley region of Western Australia*. Fisheries Research Report.

Paci, L., & Finazzi, F. (2018). Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing*, *28*, 359–374.

Pielou, E. (1984). *The interpretation of ecological data: A primer on classification and ordination*. A Wiley-Interscience publication.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Riihimäki, J., Jylänki, P., & Vehtari, A. (2013). Nested expectation propagation for Gaussian process classification. *Journal of Machine Learning Research*, *14*, 75–109.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*, 319–392.

Sainsbury, K. (1979). Defining fish stocks on NW Shelf. *Australian Fisheries*, *38*, 4–12.

Sainsbury, K., Campbell, R., & Whitelaw, A. (1993). Effects of trawling on the marine habitat on the North West Shelf of Australia and implications for sustainable fisheries management. *Sustainable fisheries through sustaining fish habitat*. *Proceedings of the Australian Society for Fish Biology Workshop held at Victor Harbor, SA, 12-13 August, 1992. BRS, AGPS, Canberra*.

Sainsbury, K., & Whitelaw, A. (1984). Biology of Peron's threadfin bream, *Nemipterus peronii* (*Valenciennes*), from the North West shelf of Australia. *Marine and Freshwater Research*, *35*, 167–185.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, *62*, 795–809.

Thresher, R. E., Sainsbury, K. J., Gunn, J. S., & Whitelaw, A. W. (1986). Life history strategies and recent changes in population structure in the lizardfish genus, *Saurida*, on the Australian Northwest Shelf. *Copeia*, *1986*, 876–885.

Tierney, L., & Kadane, J. B. (1986). Accurate approximation for posterior moments and marginal densities. *Journal of American Statistical Association*, *81*, 82–86.

Torabi, M. (2016). Hierarchical multivariate mixture generalized linear models for the analysis of spatial data: An application to disease mapping. *Biometrical Journal*, *58*, 1138–1150.

Vanhatalo, J., Pietiläinen, V., & Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, *29*, 1580–1607.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., & Vehtari, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, *14*, 1175–1179.

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.

Wall, M. M., & Liu, X. (2009). Spatial latent class analysis model for spatially distributed multivariate binary data. *Computational Statistics & Data Analysis*, *53*, 3057–3069.

Wallner, B. G., & Phillips, B. F. (1988). From scampi to deepwater prawns: developments in the North West shelf deepwater trawl fishery. *Australian Fisheries*, *47*, 34–38.

Woolley, S. N. C., Foster, S. D., Bax, N. J., Currie, J. C., Dunn, D. C., Hansen, C., Hill, N., O'Hara, T. D., Ovaskainen, O., Sayre, R., Vanhatalo, J. P., & Dunstan, P. K. (2019). Bioregions in marine environments: combining biological and environmental data for management and scientific understanding. *BioScience*, biz133. *70*(1), 48–59.

Woolrich, M., Behrens, T., Beckmann, C., & Smith, S. (2005). Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. *IEEE Transactions on Medical Imaging*, *24*, 1–11.

Zubin, J. (1938). A technique for measuring likemindedness. *Journal of Abnormal and Social Psychology*, *33*, 508–516.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Vanhatalo J, Foster SD, Hosack GR. Spatiotemporal clustering using Gaussian processes embedded in a mixture model. *Environmetrics*. 2021;32:e2681. https://doi.org/10.1002/env.2681