HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Spatial confounding in Bayesian species distribution modeling

Mäkinen, Jussi

2022-11

Mäkinen , J , Numminen , E , Niittynen , P , Luoto , M & Vanhatalo , J 2022 , ' Spatial confounding in Bayesian species distribution modeling ' , Ecography , vol. 2022 , no. 11 , e06183 . https://doi.org/10.1111/ecog.06183

http://hdl.handle.net/10138/350497
https://doi.org/10.1111/ecog.06183

cc_by
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

ECOGRAPHY

Research article

Spatial confounding in Bayesian species distribution modeling

Jussi Mäkinen, Elina Numminen, Pekka Niittynen, Miska Luoto and Jarno Vanhatalo

J. Mäkinen (https://orcid.org/0000-0001-6599-8279) ✉ (jussi.makinen@yale.edu) and J. Vanhatalo (https://orcid.org/0000-0002-6831-0211), Research Centre for Ecological Change, Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, Univ. of Helsinki, Helsinki, Finland. JM also at: Dept of Ecology and Evolutionary Biology, Yale Univ., New Haven, CT, USA. – E. Numminen and JV, Dept of Mathematics and Statistics, Faculty of Science, Univ. of Helsinki, Helsinki, Finland. – P. Niittynen (https://orcid.org/0000-0002-7290-029X) and M. Luoto, Dept of Geosciences and Geography, Faculty of Science, Univ. of Helsinki, Helsinki, Finland.

Ecography
2022: e06183
doi: 10.1111/ecog.06183

Subject Editor: Miguel Nakamura
Editor-in-Chief: Miguel Araújo
Accepted 25 July 2022

1) Species distribution models (SDMs) are currently the main tools to derive species niche estimates and spatially explicit predictions for species geographical distribution. However, unobserved environmental conditions and ecological processes may confound the model estimates if they have direct impact on the species and, at the same time, they are correlated with the observed environmental covariates. This, so-called spatial confounding, is a general property of spatial models and it has not been studied in the context of SDMs before.

2) We examine how the estimation accuracy of SDMs depends on the type of spatial confounding. We construct two simulation studies where we alter spatial structures of the observed and unobserved covariates and the level of dependence between them. We fit generalized linear models with and without spatial random effects applying Bayesian inference and recording the bias induced to model estimates by spatial confounding. After this we examine spatial confounding also with real vegetation data from northern Norway.

3) Our results show that model estimates for coarse scale covariates, such as climate covariates, are likely to be biased if a species distribution depends also on an unobserved covariate operating on a finer spatial scale. Pushing higher probability for a relatively weak and smoothly varying spatial random effect compared to the observed covariates improved the model's estimation accuracy. The improvement was independent of the actual spatial structure of the unobserved covariate.

4) Our study addresses the major factors of spatial confounding in SDMs and provides a list of recommendations for pre-inference assessment of spatial confounding and for inference-based methods to decrease the chance of biased model estimates.

Keywords: estimation bias, Gaussian process, spatial confounding, spatial random effect, species distribution model

NORDIC SOCIETY OIKOS

www.ecography.org

© 2022 The Authors. Ecography published by John Wiley & Sons Ltd on behalf of Nordic Society Oikos
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Introduction

Species distribution models (SDMs) are popular tools in ecology to study the spatial and temporal distributions of species populations, and the impact of environmental variables on species occurrence probability (Elith and Leathwick 2009, Franklin 2010). However, species distributions may depend also on unobserved covariates, such as unmeasured environmental conditions, or ecological processes such as interspecific interactions (Wisz et al. 2013), dispersal (Shurin et al. 2009) or population dynamics (Mielke et al. 2020). Owing to their impacts on species distribution, traditional 'covariates-only' models may be subject to spatially autocorrelated residual errors which create overly optimistic uncertainty estimates for the covariate effects (Legendre 1993). Moreover, a physical or an ecological dependence, or a dependence through sampling design, between the observed and unobserved covariates may confound the estimation of covariate effects in an SDM (Feng et al. 2019b). We can speculate that, in such a case, the estimated covariate effects reflect the realized distribution and habitat of a species, and do not inform about the pure physical constraints of the environment on the species population (Soberón and Nakamura 2009).

A popular and well-documented solution to explain variation in species occurrence pattern originating from an unobserved covariate is to include a spatial random effect into an SDM, applied for example by Ver Hoef et al. (2018) and Soriano-Redondo et al. (2019). Here, we refer to such models as *spatially explicit* SDMs (see Supporting information for a more thorough description of the inference with a spatial random effect). Previous studies have shown that a spatial random effect successfully explains the residual spatial autocorrelation in the model and returns higher and more realistic variance estimates for the covariate effects (Legendre 1993) and predictions (Guélat and Kéry 2018). However, there are also results showing that including a spatial random effect in a model does not necessarily account for the effect of the unobserved covariate, and may even affect the estimates of the covariate effects in a counterintuitive manner (Hodges and Reich 2010). Generally, in a spatially explicit SDM, the covariate effect estimate depends on the associations between the covariate and the response variable, and between the observed and unobserved covariates (Paciorek 2010). Moreover, correlation between the supposedly independent model components (covariates and a spatial random effect) may increase bias in covariate effect estimates, which is generally referred to as spatial confounding and has been found in many study settings (Diniz-Filho et al. 2003, Hodges and Reich 2010, Hanks et al. 2015). Spatial smoothness (i.e. large spatial autocorrelation range) of a covariate tends to increase the correlation between a covariate and a spatial random effect increasing the risk of spatial confounding (Paciorek 2010). Thus, including a spatial random effect in a model can bias the estimate for the covariate effect (Reich et al. 2006). In SDM applications, there are contradictory results about how spatial autocorrelation ranges impact the estimation accuracy of a spatially explicit model: some results support the dependence of estimation accuracy on the spatial autocorrelation range of the covariate (Lichstein et al. 2002, Hawkins et al. 2007) and others oppose that (Kissling and Carl 2008, Betts et al. 2009, Bini et al. 2009).

Here, we formulate SDMs as generalized linear models with and without spatial random effects, and study their estimation and prediction accuracy with respect to two key aspects of the studied phenomenon: 1) the amount of correlation between an observed and an unobserved covariate, and 2) the spatial autocorrelation ranges of the covariates. These issues have already been studied in the context of general spatial models (Wakefield 2007) and maximum likelihood inference (Hodges and Reich 2010, Paciorek 2010). The novelty of this work is that we combine and extend the earlier results (Bini et al. 2009, Hodges and Reich 2010, Paciorek 2010) with the general theory of species distribution modeling (Austin 2007, Franklin 2010) and study the phenomenon from a Bayesian inference point of view (Clark 2004). Taking the Bayesian standpoint naturally raises the question: how prior distributions for the hyperparameters of the spatial random effect (magnitude of variation and spatial autocorrelation range) affect the results (Soerbye et al. 2019). A typical choice is to prioritize low magnitude of variation of the spatial random effect compared to the covariate effect (Kallasvuo et al. 2017, Fuglstad et al. 2018, Mäkinen and Vanhatalo 2018) and use an informative (Kallasvuo et al. 2017, Fuglstad et al. 2018, Mäkinen and Vanhatalo 2018, Soerbye et al. 2019, Soriano-Redondo et al. 2019) or a uniform (Tikhonov et al. 2020) prior distribution for spatial autocorrelation range relative to the size of the study area. To our knowledge, there are no previous studies about comparing different priors in spatially explicit SDMs. Hence, we tested different informative priors on the variance and range parameter in experiments with both simulated and empirical data.

Our final aim is to provide suggestions on how to design a spatial study setting for an SDM and prioritize different components of the model to improve the accuracy of model estimation and spatial predictions. We do this with simulated species presence–absence data, where we vary the spatial autocorrelation ranges and cross-dependencies of the observed and unobserved covariates. We conduct the same analysis with an empirical presence–absence data set, and test how omitting a covariate from a model affects the estimates for other covariate effects. We focus on analyzing presence–absence data and studying a species distribution through the species' presence probability.

Material and methods

Data

This section reviews the method for simulating species data and presents the empirical species data.

Simulated data

We conducted two simulation studies for spatial confounding: first studying bias of the estimator of the covariate effect,

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and second studying absolute deviation of the estimators from the true values for two covariate effects. We conducted both simulation studies in a rectangular region of size (0,1) × (0,1) into which we generated species occurrence data with alternative spatial confounding scenarios. We divided the region into a grid of 50 × 50 cells and, for each set of simulated data, we first generated two spatially structured covariates in the grid cells: an observed one denoted with $x(s)$ and an unobserved one denoted with $z(s)$ (Fig. 1). Conditional on $x(s)$ and $z(s)$, we constructed a latent field $f(x(s),s)$ over the grid cells. In the first simulation study $f(x(s),s) = \beta_1 x(s) + \beta_2 z(s)$, and in the second simulation study $f(x(s),s) = \beta_1 x(s) + \beta_2 x(s)^2 + \beta_3 z(s) + \beta_4 z(s)^2$. Finally, we randomly chose n = 200 grid cells at which we simulated species' occurrences, denoted by $y(s)$, by sampling from $y(s) \sim$ Bernoulli $(\pi(f(x(s),s)))$ and saving the values of s_i, $y_i = y(s_i)$ and $x_i = x(s_i)$ for $i = 1, \ldots, n$. A probit-link function was used for π and the effects β_1, β_2, β_3, β_4 were set to one. The purpose of having two simulations is to empirically compare the estimation accuracy in models that have linear and non-linear responses to a covariate and link the simulation to the empirical study, where the non-linear responses are ecologically reasoned and an essential part of the inference.

We used Gaussian processes with Matérn 3/2 type covariance function to simulate smooth realizations of $x(s)$ and $z(s)$ (presented in the 'Species distribution modeling' section for the inference model). With different parameterizations of the covariance function, we controlled the expected spatial autocorrelation range for $x(s)$ and $z(s)$. We refer to this parameter as length-scale parameter, and higher values of it create higher spatial correlations and more smoothly varying fields, for example mimicking mesoclimatic covariates, while low values of this parameter yield locally varying and on average less correlated covariates (see Supporting information for more detailed definition of the method). Methods to simulate variables and later infer these variables rely on the same parameterization of a latent Gaussian process.

By smoothing and appropriately weighting the same white noise when generating both $x(s)$ and $z(s)$, we were able to create data with potential for spatial confounding and mimic a situation in which the observed and unobserved covariates have a common driver. Different weights for the common white noise generated different levels of dependency between the two fields. We narrowed down the possibilities of different scenarios and created only positive dependence between fields that realized as positively correlated samples of $x(s)$ and $z(s)$ on average. For each realization, we computed the sample correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$ and the spatial autocorrelation ranges in $\boldsymbol{x}$ and $\boldsymbol{z}$. The latter was defined as the length-scale of Gaussian process regression models fitted to the observed values of $\boldsymbol{x}$ and $\boldsymbol{z}$ (see Supporting information for details; Gaussian process is introduced in 'Species distribution modeling' section). We defined four categories for the realized correlation [0.0–0.2; 0.2–0.4; 0.4–0.6; 0.6–0.8] and five categories for the estimates of length-scale parameters [0.0–0.2; 0.2–0.4; 0.4–0.6; 0.6–0.8; 0.8–1.0]. Examples of covariates with different length-scale parameter estimates are shown in Fig. 1 (panel b). We sampled 50 realizations of simulated data from each of the 100 different combinations of the spatial autocorrelation ranges of the covariates and the correlation between the covariates.

Empirical data

We conducted the real data experiment on 85 vascular plant species distributions in northern Norway (70°0′N, 26°14′E) (species list in Supporting information). The study area is located around two mountain massifs, which create strong elevation and climate gradients (120–1064 m a.s.l.), where vegetation consists of Arctic, alpine and boreal species (Niittynen and Luoto 2018). The full data set consists of 1325 vegetation plots (1 × 1 m) with recorded occurrence of all vascular plant species. The study area covers 195 km^2, where the minimum distance between plots is

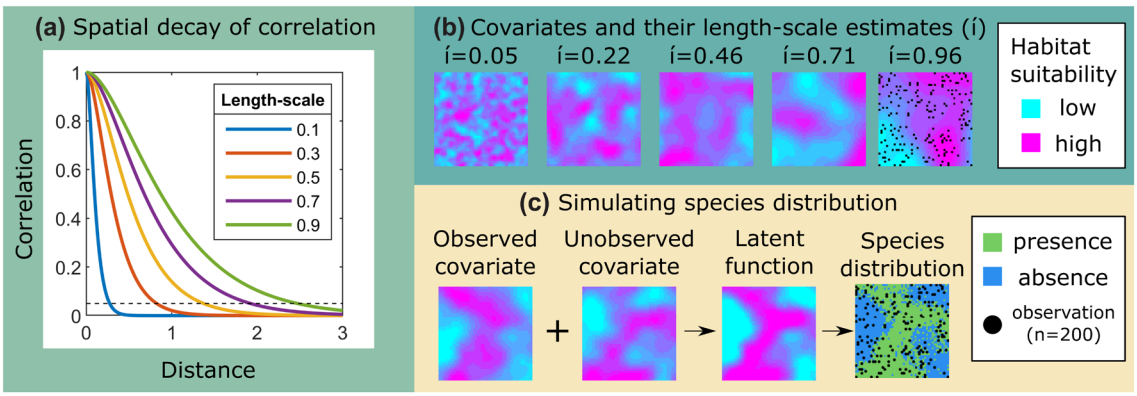


Figure 1. Panel (a) shows the expected spatial correlation as a function of distance for different values of the length-scale parameter (dashed line denotes the 5% correlation, which is used as a threshold for practical independence between two points). Panel (b) shows realizations of spatial covariate surfaces generated with a Gaussian process (going from the smallest length-scale estimate to the largest one). Sub-plot titles in panel (b) summarize the length-scale estimates of spatial Gaussian process regression models fitted to 200 observation locations shown as black points in the rightmost sub-plot. Panel (c) shows how a species distribution was simulated in the first simulation study by sampling from the latent function which is a sum of an observed covariate (x) and an unobserved covariate (z).

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

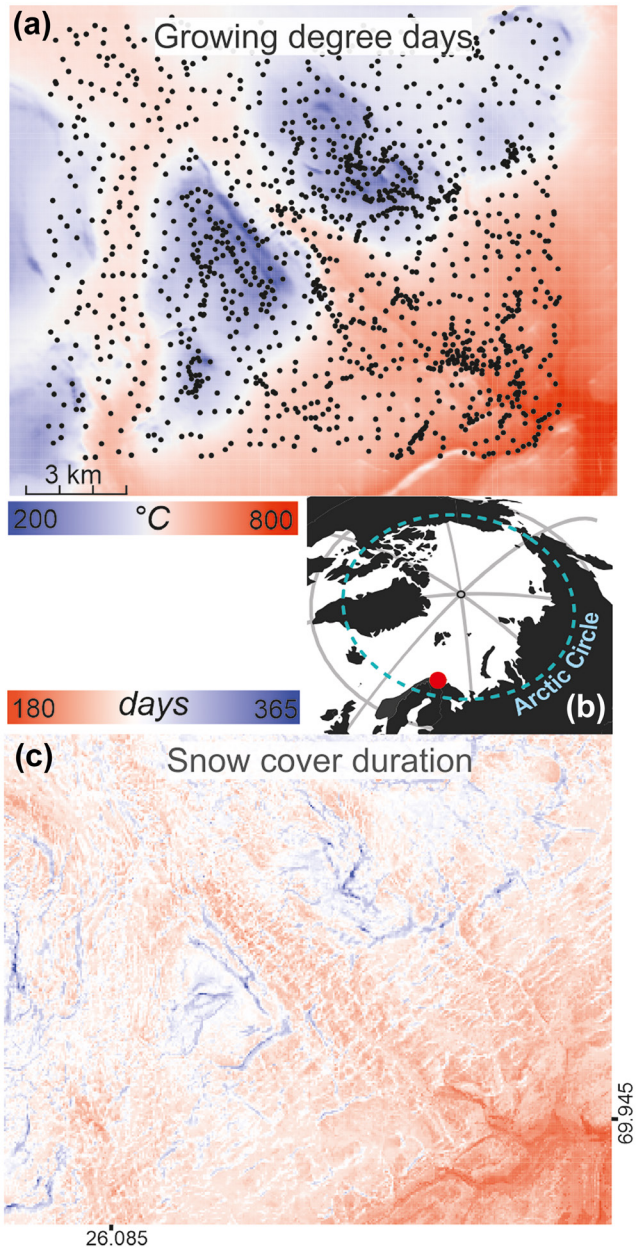


Figure 2. Panel (a) shows the spatial distributions of the vegetation sampling sites (black dots) and growing degree days (temperature sum of the growing season). Panel (b) shows the location of the study area in an Arctic context. Panel (c) shows the spatial distribution of snow cover duration (days).

19 m and the maximum distance is 18 900 m (Fig. 2). We derived six covariates describing habitat characteristics from remotely sensed data (snow cover duration, surface deposit quality), and spatial statistical models (growing degree days, potential incoming radiation, slope and topographic wetness index). Surface deposit quality is a categorical covariate and other covariates are continuous. Spatial resolution of the original environmental covariate layers varies between 2 and 30 m. The data are presented more thoroughly by Niittynen and Luoto (2018) and Niittynen et al. (2020).

For each species, we constructed six data sets that differed in their covariate composition. The first data set contained all the covariates and from the remaining five data sets we dropped in turn one of the continuous covariates. The first data set plays the role of reference data where all covariates are present whereas the six other data sets correspond to data sets from which a potentially important covariate is missing. We estimated the relationship between the omitted and retained covariates through their spatial autocorrelation ranges and mutual correlations, which were estimated separately prior the fitting of SDMs (see more thorough description of the method in Supporting information). Thus, we were able to compare how the covariate effect estimates behave with respect to the spatial autocorrelation ranges and cross-correlations of the observed and unobserved covariates. Spatial coordinates were scaled so that they are comparable to the spatial coordinates in the simulation data.

Species distribution modeling

In the analysis of simulated and empirical data sets, we formalized spatially explicit SDMs under a hierarchical Bayesian modeling framework, where the hierarchical model for species observations $y(s)$ at location s is (Eq. 1):

$$\begin{aligned} y(s) &\sim \text{Bernoulli}\left(y(s)|\pi\left(f\left(x(s),s\right)\right)\right) \\ f\left(x(s),s\right) &= \alpha + \beta x(s) + g(s) \end{aligned} \tag{1}$$

where $y(s)$ is modeled conditionally on a latent function, $f(x(s),s)$. The latent function was squeezed to vary between zero and one to represent the species presence probability with a probit-link function (π). The latent function in (Eq. 1) depends on the environmental covariates $x(s)$ and the spatial random effect, denoted by $g(s)$ (Cressie and Wikle 2015). The parameter α is a model intercept that captures the average presence probability over the observations and β is a vector of linear weights for spatially varying covariates $\boldsymbol{x}(s)$. In the first simulation study, $\boldsymbol{x}(s)$ consisted of the simulated covariate. In the second simulation and empirical study, we used both (standardized) covariates and their squares, since previous studies have provided evidence for non-linear covariate effects on the species in the vegetation data (Niittynen and Luoto 2018).

The model (Eq. 1) is a basic building block in most of the state-of-the-art single and joint species distribution models that are used for predictive analyses (Ovaskainen et al. 2015, Thorson et al. 2015, Vanhatalo et al. 2020). In these applications, the spatial random effect $g(s)$ models explicitly spatial autocorrelation in residual error and explains the effects of those processes that were not captured with the covariate data (Ovaskainen et al. 2015). The model structure is applicable to different data types used for fitting SDMs, such as presence-only data with a point process observation model, or abundance data with a log-normal observation model. It also

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

corresponds to many of the statistical models used in large SDM model comparisons (Valavi et al. 2021).

We modeled the spatial autocorrelation with a Matérn covariance function with 3/2 degrees of freedom, such that $\mathrm{Cov}[g(s), g(s')] = \sigma^2 k(s - s'/l) = \sigma^2(1+\sqrt{3}(s-s'/l))\exp(1-\sqrt{3}(s-s'/l))$ where σ^2 is a variance parameter, $k(\cdot)$ gives the correlation as a function of Euclidean distance $||\cdot||$ and l is a length-scale parameter governing the spatial autocorrelation range (a third of the distance at which the autocovariance has dropped to 5% of its maximum) of the spatial random effect. The Matérn covariance function is presented in more detail by Minasny and McBratney (2005).

The inference of a spatially explicit SDM in a hierarchical Bayesian framework is steered with the prior distributions of the parameters. Without a strong prior belief, we assigned α and β (in Eq. 1) priors with a zero mean Gaussian distribution and high variance ($\sigma^2 = 10$). These are generally considered as valid priors which allow the response to vary from strongly negative to strongly positive (Lemoine 2019). For the length-scale parameter we tested priors that prefer either large or small autocorrelation ranges relative to the size of the study region. For the variance parameter, we tested priors that prefer either high or low variance relative to anticipated covariate effects. We were able to apply these priors for simulated and empirical data sets, since covariates and coordinates were standardized on the same scales in both data sets.

We tested different default prior settings for length-scale and variance used in common ecologically oriented statistical softwares, such as HMSC (Tikhonov et al. 2020) and SPBayes (Finley et al. 2007). We compared them with manually configured priors, which represent a higher level of prior belief about the characteristics of a spatial random effect. The uniform prior for length-scale in HMSC and SPBayes was approximated with an uninformative log-Normal distribution. The priors, along with their ecological interpretation, are shown in Table 1.

For both the simulated and empirical data sets, we conducted the inference for covariance function parameters of the spatial random effect with Markov Chain Monte Carlo (MCMC) sampling. We took first the Maximum A Posteriori estimate with Laplace approximation (Vanhatalo et al. 2010) and used the point estimate as the initial value for MCMC sampling. We took 600 samples of which we discarded first 100 samples and treated the last 500 as posterior distribution. We checked that samples converged by computing potential scale reduction factor (PSRF) and kept only models that had PSRF less than 1.1. All models in simulation and empirical tests passed the evaluation. Model estimation and evaluation, and sampling evaluation, were conducted with GPstuff toolbox (Vanhatalo et al. 2013) (the development branch from <https://github.com/gpstuff-dev/gpstuff>) in Matlab (2018b).

Estimation and prediction accuracy

In the first simulation study, we compared models with respect to the bias of the posterior mean for the covariate effect and the average posterior variance of the covariate effect over the simulated data sets. We defined bias as the expected difference of the posterior mean ($\hat{\beta}$) and the true data generating value (β_{true}), $\mathrm{Bias}(\beta) = E[\hat{\beta} - \beta_{\mathrm{true}}]$, where the expectation is taken as an average over the 50 realizations of each simulation scenario. The average posterior variance of the estimate was defined as $E[\mathrm{Var}[\beta|\mathrm{Data}]]$ where Data denotes a simulated data set and the expectation is again taken over simulations.

In the second simulation study, we compared models with respect to the expected absolute deviation of the estimated first and second order effects from the true covariate effect values so that $\mathrm{Deviation}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_{\mathrm{true}}) = E[\sqrt{\sum[[\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_{\mathrm{true}}]^2]}]$, where $\hat{\boldsymbol{\beta}}$ is a vector containing the estimated first and second order effects and $\boldsymbol{\beta}_{\mathrm{true}}$ is a vector containing the true values. Expected deviation is calculated over the 50 realizations of each simulation scenario. The deviation informs only about absolute accuracy of model estimates and not about the sign (positive or negative) of the shift in estimates compared to the true values.

For empirical data sets, we did not have a data-generating value to which we would have been able to compare the estimated covariate effects. Hence, we fitted the covariate-only and the five spatially explicit models to all 6×85 empirical data sets described in the 'Empirical data' section. For each of the 85 species, we treated the covariate-only model conditional on all six covariates as the 'true' model, after which we compared it to the six missing covariate models. The omitted

Table 1. Priors for the length-scale (l) and variance (σ^2) of a spatial random effect.

Prior	$p(l)$	$p(\sigma^2)$	Interpretation
Prior 1	$\log - N(-2.2, 0.4)$ 95% credibility interval (CI) = [0.05,0.25]	$\log - N(-1.5, 0.4)$ 95% CI = [0.1,0.5]	Spatial range is small and variation low.
Prior 2	$\log - N(-2.2, 0.4)$ 95% CI = [0.05,0.25]	$\log - N(0.3, 0.2)$ 95% CI = [1,2]	Spatial range is small and variation high.
Prior 3	$\log - N(-0.5, 0.1)$ 95% CI = [0.45,0.8]	$\log - N(-1.5, 0.4)$ 95% CI = [0.1,0.5]	Spatial range is large and variation low.
Prior 4	$\log - N(-0.5, 0.1)$ 95% CI = [0.45,0.8]	$\log - N(0.3, 0.2)$ 95% CI = [1,2]	Spatial range is large and variation high.
Prior 5 (HMSC/SPBayes)	$\log - N(0.80, 0.76)$ 95% CI = [0.5,10]	Gamma(1,0.3) 95% CI = [0.01,1.12]	Spatial range varies from small to large and variation alternates from low to high.

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

covariate in empirical data sets corresponds to the unobserved covariate in the simulated data set, creating analogy between the simulated and empirical tests.

For the full analysis, we chose only the species whose distributions we were able to explain with the environmental covariates and which did not have strong spatial autocorrelation. We computed the proportion of the variation explained by a spatial random effect when all covariates were included in the model and chose the species with less than 10% of the variation explained by a spatial random effect (in total, 42 out of 85 species). We used prior 1 from Table 1 as the spatial model to measure spatial autocorrelation. To compare the true model to the missing covariate models, we used the deviation of the estimated covariate effects, as we did for the second simulation study.

In the simulation tests, we validated models with respect to their predictive accuracy in spatial interpolation and extrapolation. For spatial interpolation, we randomly chose 200 locations from the study grid. For spatial extrapolation, we simulated a test data set and placed it geographically far from the training data set so that the spatial random effect cannot impact the mean of the predictive distribution. We used the log posterior predictive density (Vehtari and Ojanen 2012) and Tjur-R^2 as the measures of predictive accuracy. The log posterior predictive density measures the goodness of the predictive probability density as a whole whereas Tjur-R^2 measures the discriminatory ability of the prediction.

To summarize, we tested estimation accuracy (bias and deviation) and precision (variance) with respect to dependence between observed and unobserved covariates, and spatial autocorrelation ranges of the observed and unobserved covariates, as well as prior specifications for the spatial random effect parameters in the inference model.

Results

Simulated data

Estimates of the covariate-only model were negatively biased when there was low correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$ (less than 0.2), regardless of the spatial autocorrelation ranges of $\boldsymbol{x}$ and $\boldsymbol{z}$ (Fig. 3: panel (a)); whereas they were unbiased in most cases when there was moderate correlation (0.2–0.4) between $\boldsymbol{x}$ and $\boldsymbol{z}$, and positively correlated when there was significant correlation (over 0.4) between $\boldsymbol{x}$ and $\boldsymbol{z}$. In the covariate-only model, bias varied only little with respect to the spatial autocorrelation ranges of $\boldsymbol{x}$ and $\boldsymbol{z}$.

In spatially explicit models, unbiased covariate effect estimates were obtained only when the correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$ was low (0–0.2) or moderate (0.2–0.4). However, also in those cases, the estimates were unbiased only for certain combinations of spatial autocorrelation ranges of $\boldsymbol{z}$ and priors for the covariance function parameters (Fig. 3). When correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$ was low (< 0.2), the bias in spatially explicit models was consistently smaller than in the covariate-only model. When the correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$ was greater than 0.4, the spatially explicit models returned covariate effect estimates that were positively biased, and bias was larger in the spatially explicit than in the covariate-only model in these cases. Moreover, the bias in spatially explicit models was systematically larger when $\boldsymbol{x}$ had a larger spatial autocorrelation range than $\boldsymbol{z}$ compared (cells below the diagonal in Fig. 3) to the simulations where $\boldsymbol{x}$ had a smaller spatial autocorrelation range than $\boldsymbol{z}$ (cells above the diagonal in Fig. 3).

Comparisons of different prior specifications for the spatially explicit model showed that the bias was the smaller the more the prior reduced the variance of the spatial covariance function towards zero (Fig. 3: compare panels n, p and r to panels o and q). This difference was most evident in cases where $\boldsymbol{x}$ had a larger spatial autocorrelation range than $\boldsymbol{z}$ (Fig. 3: in each panel cells below the diagonal). Comparisons with respect to the prior for the length-scale showed that priors favoring long length-scale had lower bias than priors favoring short length-scale (when comparing between priors 1 and 3 and between priors 2 and 4). Uniform prior for length-scale ranked between the priors 1 and 3. The differences were small and independent of the spatial autocorrelation ranges of $\boldsymbol{x}$ and $\boldsymbol{z}$.

Variance of covariate effect estimates over simulations behaved qualitatively similarly to the bias of covariate effect estimates (Supporting information). In general, variance was higher when $\boldsymbol{x}$ had a larger spatial autocorrelation range than $\boldsymbol{z}$ and when the prior preferred high variance of the spatial random effect.

In the second simulation study, deviation followed a similar pattern with respect to the spatial autocorrelation ranges and cross-correlation of $\boldsymbol{x}$ and $\boldsymbol{z}$ as in the first simulation study. Results from the second simulation were similar to those from the first, with respect to having a spatial random effect in the model or not, and with respect to different prior choices (Fig. 4). When there was strong correlation between the covariates, the deviation was less associated with the spatial autocorrelation range of the unobserved covariate and more associated with the spatial autocorrelation range of the observed covariate.

Spatially explicit models performed better in spatial interpolation with respect to log-predictive density and Tjur-R^2 in both simulation studies (Supporting information). Priors for long length-scale and small variance (prior 3) improved spatial interpolation compared to other prior settings. In spatial extrapolation, all models except the spatially explicit model with uninformative prior for length-scale and prior preference for small variance (prior 5) performed equally well (Supporting information). Prior 5 had slightly lower log-predictive density than other models across different scenarios of cross-dependence and spatial autocorrelation ranges of $\boldsymbol{x}$ and $\boldsymbol{z}$.

Empirical data

We focused on the scenarios where Saga wetness index (SWI), snow cover duration (SCD) or growing degree days (GDD)

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

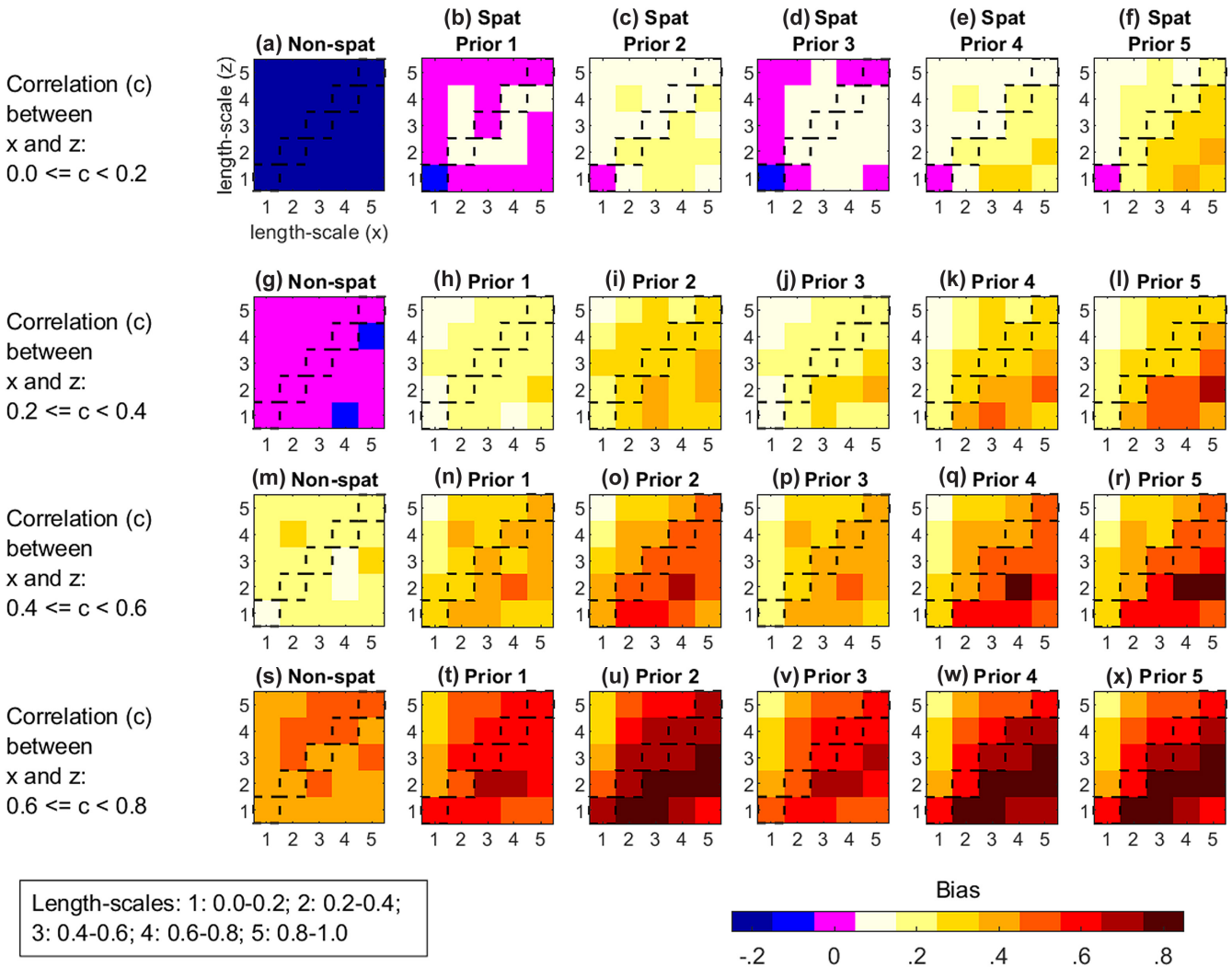


Figure 3. Comparison of estimation bias of a covariate effect (β) in a covariate-only model (column 1) and spatially explicit models (columns 2–6) in simulation experiment 1 (see Table 1 for a description of the priors). Each plot corresponds to a specific category of correlation between an observed covariate (*x*) and an unobserved covariate (*z*). Each cell in a plot corresponds to a particular combination of spatial autocorrelation ranges of *x* and *z*, measured as the estimate of the length-scale parameter in a spatial Gaussian process regression model. The dashed line around cells in the diagonal of each plot shows the cases where *x* and *z* have equal spatial autocorrelation ranges.

was omitted from a model (Fig. 5). Estimated deviation was strong for SCD and GDD when omitting either of them, which is due to their high cross-dependence. GDD and SCD were the most correlated pair of covariates followed by SWI and slope (see panel d in Fig. 5). Omitting SWI created the strongest estimate deviation for slope (see panel a in Fig. 5) which was likely due to a high correlation between them. Generally, the dependence between the omitted and observed covariates impacted the estimate deviation more strongly than the difference in their spatial autocorrelation ranges.

The estimated deviations were equally large in the covariate-only and spatially explicit models. This was slightly contrary to the simulations, where spatially explicit models returned higher bias and estimate deviations if observed and unobserved covariates were correlated. The impact of priors on model estimation was in line with simulation results. In particular, when SWI was omitted from a model, a uniform prior for length-scale and a prior preferring high variance returned higher estimate deviation for slope, SCD and GDD (see panel (a) in Fig. 5).

Discussion

The ecological interpretation of SDMs relies on an accurate estimation of the covariate effects. Despite high predictive performance, spatially explicit SDMs have had issues with unintuitive model estimates (Kissling and Carl 2008, Bini et al. 2009, Lany et al. 2019, Renner et al. 2019, Kim 2021). Our results showed that, given independent observed and unobserved covariates, the covariate-only model had negatively biased estimates (Fig. 3: panels (a)–(b)). The likely reason for this is that the unobserved covariate created overdispersion around the response along the observed covariate

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

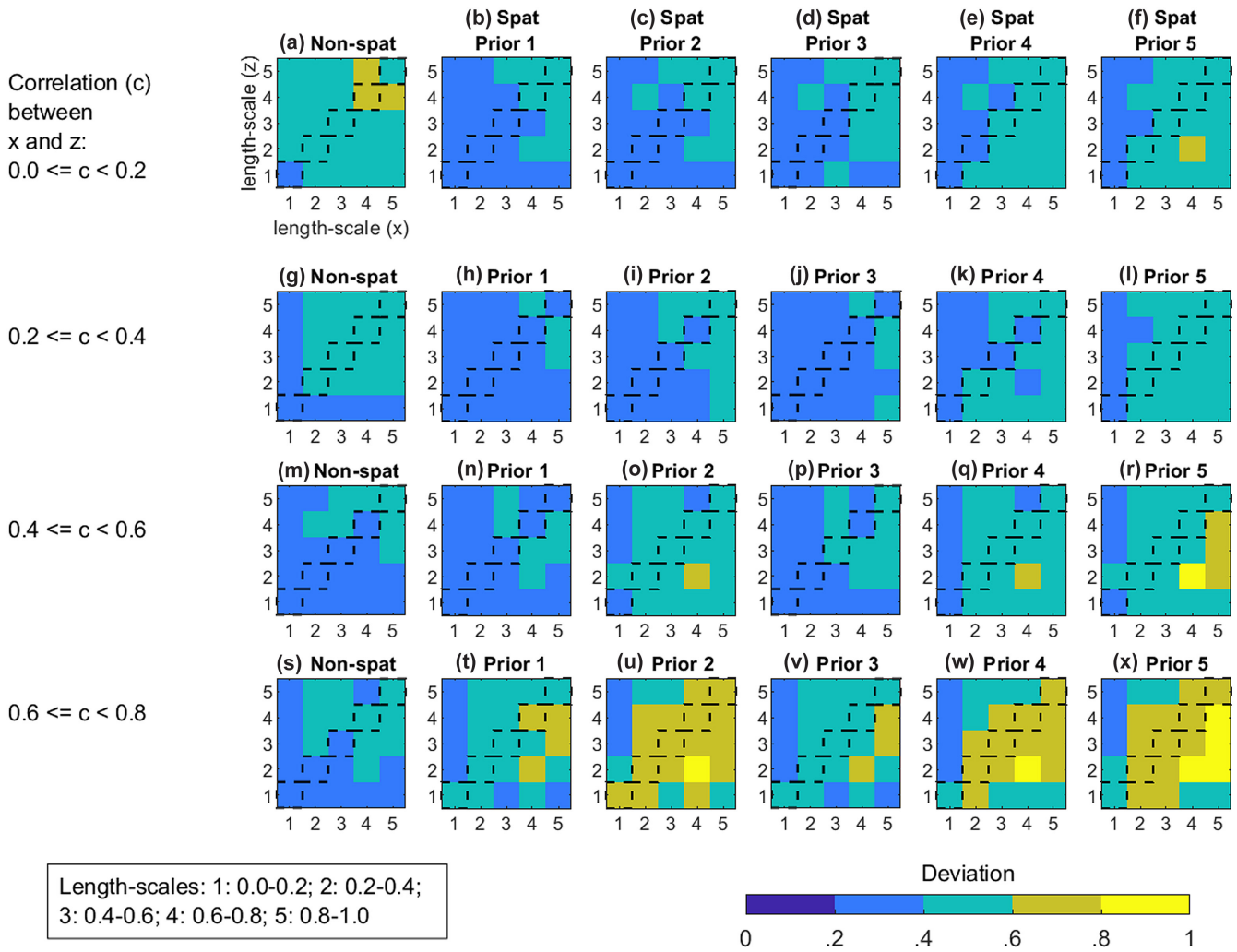


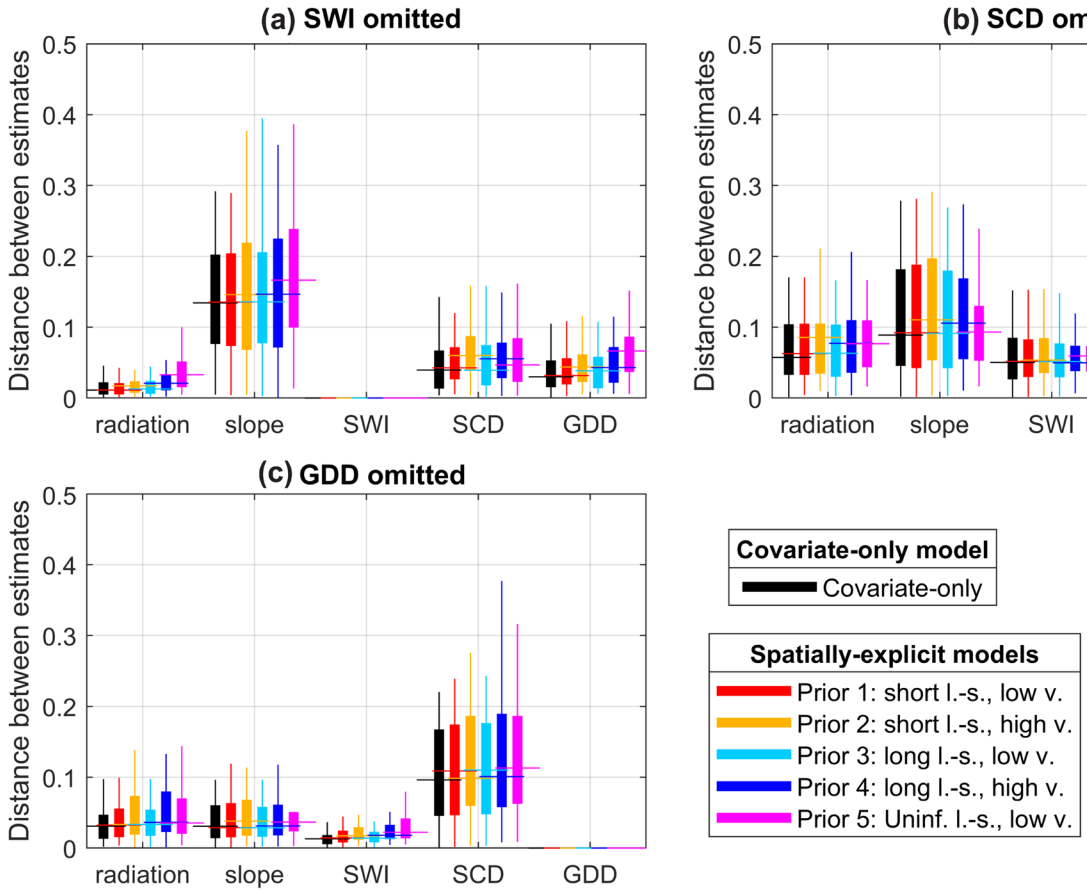
Figure 4. Comparison of deviation of covariate effects (β) in covariate-only (column 1) and spatially explicit models (columns 2–6) with different prior specifications (specified in titles) in simulation experiment 2. The subplots correspond to different simulation scenarios as in Fig. 3.

but the model was not able to account for this appropriately. Because the variance in the Bernoulli model (Eq. 1) attains its maximum at $\pi = 1/2$ and it decreases to zero when $\pi = 0$ and $\pi = 1$, the covariate-only model was able to explain the over-dispersed species observations best by pulling species presence probability towards $\pi = 1/2$ throughout the study region. This effectively means that the model must underestimate the strength of the response along the observed covariate. A spatial random effect alleviated the over-dispersion caused by the unobserved covariate and improved the estimation accuracy when correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$ was low (Fig. 3 and 4).

The dependence between the observed and unobserved covariates exposed both covariate-only and spatially explicit models to positive bias and higher deviation. Since the estimate for the covariate effect accounted for the effect of the unobserved covariate, which positively correlates with the observed covariate, this was reasonable. For SDMs, this setting is analogous to confounding the covariate effects by a species interacting with the focal species (Heikkinen et al. 2007) or by having an unobserved environmental process, like snow cover (Niittynen and Luoto 2018) or topography (Luoto and Heikkinen 2008). Given that there is a strong dependence between the covariates (correlation over 0.4), a covariate-only model had lower levels of bias and deviation than a spatially explicit model. A likely explanation for this is that the over-dispersion caused by independent variation in $\boldsymbol{z}$ compensated for the positive bias induced by the correlation between $\boldsymbol{z}$ and $\boldsymbol{x}$. The over-dispersion issue is specific to Bernoulli distributed data and, to our knowledge, has not been well addressed for spatially explicit SDMs (but see Dupont et al. 2021).

Estimates of a spatially explicit model were more biased and had higher deviation when the observed covariate had a larger spatial autocorrelation range than the unobserved covariate (Fig. 4 and 5: lower triangular), as found also by Paciorek (2010) in the context of linear models. This happens because, by default, a spatial random effect avoids overfitting

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



(d) Correlation matrix

	radiation	slope	SWI	SCD	GDD
radiation	1	-0.09	0.03	0.08	0.15
slope	-0.09	1	-0.52	0.2	-0.09
SWI	0.03	-0.52	1	-0.03	0.11
SCD	0.08	0.2	-0.03	1	-0.54
GDD	0.15	-0.09	0.11	-0.54	1

(e) Length-scale estimate

	l.-s.
radiation	0.02
slope	0.03
SWI	0.06
SCD	0.11
SDQ	0.33
GDD	0.44

Figure 5. In panels (a)–(c) boxplots correspond to the distribution of the deviation associated to the covariate effect estimates between omitted covariate models and the full model over the 42 species. The bars show the distribution of deviations over species. Panels (d)–(e) show the correlation matrix of continuous covariates and the length-scale (l.-s.) estimates of the Gaussian process regression models fitted to each covariate. SWI = saga wetness index, SCD = snow cover duration, SDQ = soil deposit quality, GDD = growing degree days.

the data and pushes the estimate of the length-scale parameter to higher values (Hodges and Reich 2010). If $\boldsymbol{z}$ had a relatively large spatial autocorrelation range, a spatial random effect more likely correctly captured its variation, and therefore more likely also explained the shared variation between $\boldsymbol{x}$ and $\boldsymbol{z}$. This alleviated spatial confounding compared to the opposite case, where $\boldsymbol{z}$ had a relatively small spatial autocorrelation range and a spatial random effect was unable to capture its structure. These results guide modelers in assessing the sensitivity of covariate effect estimates based on the spatial autocorrelation range of the covariates, which can be estimated prior to model fitting. Covariates that have large spatial autocorrelation range, like climatic covariates, may expose spatially explicit SDMs to biased covariate effect estimates, since the effects of unobserved covariates operating at a fine spatial scale are difficult to capture with a spatial random effect. Here our simulation studies tested estimation accuracy only when correlation between observed and unobserved covariates was positive. However, we can expect that creating negative correlation between unobserved and observed covariates flips bias to be negative, and does not reduce the estimate bias towards zero.

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Our study shows that, in the presence of dependence between the observed and unobserved covariates, informative priors improve the covariate effect estimates. This supports their use against uniform priors, as suggested by Keil et al. (2014) and Lemoine (2019), but especially for spatial random effects as suggested by Soerbye et al. (2019). We also found support for how to formulate informative priors. In the simulation case, increasing the variance of a spatial random effect improved the fit of the spatial random effect on the unobserved covariate. Thus, counterintuitively, higher variance of a spatial random effect decreased the impact of over-dispersion caused by the unobserved covariate and strengthened the biasing effect of the unobserved covariate. Preferring long length-scale improved estimation accuracy despite the unobserved covariate having a short spatial autocorrelation range (cells below diagonal in plots in Fig. 3 and 4). In the case of preferring low variance, pushing more prior probability to a long length-scale increased overdispersion, which corrected for the bias that originated from correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$. Optimal prior choice for magnitude of variance and length-scale parameters may conflict with our prior belief about the unobserved covariate due to the inability of the model to account for the correlation between covariates and a spatial random effect, or due to biased prior beliefs. Based on the simulation and empirical tests, in cases where the unobserved covariate impacted the focal species equally much or less than the observed covariates, we recommend prioritizing low variance of a spatial random effect.

We base this recommendation also on the estimate's uncertainty, which increased to become overly high when prioritizing high variance. A preferable prior pushes probability density of variance parameter towards zero so that 95% of the probability mass is, for example, below 0.5. This can be generalized to different study settings and models, where continuous covariates are standardized to have mean zero and standard deviation one. For the length-scale parameter, we suggest formalizing prior candidates that support different length-scales compared to the spatial autocorrelation ranges of the observed covariates. Fitting a model with prior candidates gives an idea about the sensitivity of the model to the prior choice and about the possible conflicts between a spatial random effect and covariates. The main reason for spatial confounding is that the spatially explicit SDMs do not explicitly model the correlation between covariates and the spatial random effect. To partially get around this issue, previous studies have restricted a spatial random effect by orthogonalizing it to covariates (Hughes and Haran 2013, Johnson et al. 2013, Hanks et al. 2015) and some have proposed models to explicitly model this correlation (Page et al. 2017, Dupont et al. 2021). The former approach increases posterior variance of the covariate effects appropriately but it does not remove the bias in the estimates (Hanks et al. 2015), since the covariate effect estimates from such a model correspond to covariate-only model estimates, which may be biased as well (Johnson et al. 2013, Hanks et al. 2015). Our approach conducts the restriction through the priors of the parameters of the spatial random effect and does not restrict the posterior of the spatial random effect. Along with informative priors, we consider the approach of explicitly modeling the correlation between the covariates and the spatial random effect as a more reasonable way forward in building spatially explicit SDMs.

When interpreting results, we need to keep in mind that the method for simulating data was relatively simple compared to how real species distributions are formed. In real species data, there are multiple unobserved spatially structured covariates, whose effects on focal species may be impossible to capture with a single spatial random effect, and there are different levels of cross-dependencies between the covariates. Thus, the impacts of spatial autocorrelation ranges of covariates and prior settings on model estimates were not fully identical between the tests on simulated and empirical data. Despite these uncertainties, both tests were consistent in their main conclusions about spatial confounding, which allowed us to give recommendations for future studies.

Based on this study, we suggest analyzing the spatial structures of the covariates before running SDMs (noted also by Kim (2021)). A pre-model analysis of covariates and their spatial autocorrelation ranges allows us to formulate reasonable prior candidates for the length-scale parameter and gives prior insight into which covariates are the most susceptible to produce confounded estimates. Such spatial analysis would extend the recently published suggestions on which features and processes should be considered in building SDMs (Araújo et al. 2019, Feng et al. 2019a). Furthermore, our study shows yet undiscovered benefits from including microhabitat, like microclimate variables in the model. Replacing mesoclimate variables with downscaled microclimate variables incorporates more detailed information about the environment (Lembrechts et al. 2019) but, based on our results, we expect that decreasing the spatial autocorrelation range of the climate variable through downscaling also improves the estimation accuracy as there is less conflict with a spatial random effect.

Our study showed that improperly formulated priors for a spatial random effect may decrease the predictive accuracy. Despite more biased estimates of spatially explicit models (priors 1–3) compared to covariate-only models, all models still had equally good predictive accuracy. This was likely due to more appropriate uncertainty estimation of spatially explicit models compared to covariate-only models. This shows that proper uncertainty estimation plays an important role in spatial transferability of SDMs, although it has not been generally considered in previous model transferability assessments (Sequeira et al. 2016, Yates et al. 2018). However, in ecological risk assessments, realistic uncertainty estimates have been shown to significantly alter optimal ecological risk mitigation strategies compared to approaches that ignore uncertainty (Helle et al. 2020).

Even though spatial confounding may result in a larger bias of covariate effects in spatially explicit than in covariate-only models, we still recommend using the former. The reason is that both models lead to biased inferences on covariate effects in the presence of spatial confounding, but spatially explicit models give larger uncertainty estimates compared to covariate-only models. Spatially explicit models had also

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

better predictive performance in interpolation and equally good predictive performance in extrapolation as the covariate-only models (Supporting information). However, our study highlights the importance of technical improvements in formalizing SDM structures. In study settings where we a priori assumed collinearity between covariates and an unobserved covariate (such as distribution of a strongly competing species, whose distribution partly depends on the observed covariates), the model structure can account for the correlation between covariates and a spatial random effect with a separate function. Such approaches have been tested by Page et al. (2017) and Dupont et al. (2021) with promising results, but they have not been applied to SDMs. There is a demand for more informative prior settings, such as penalized complexity priors (Fuglstad et al. 2018), which would avoid overfitting but also avoid the spatial scales of the covariates.

Given the importance of SDMs for ecological research and environmental management, and that many SDMs are used for predicting species distributions far from the sampled area, our results are topical for ecological research. Our study brings a new perspective to the role of spatial structure of environmental covariates in fitting SDMs and shows the general limitation of current SDMs. Most importantly, the spatial structures of the covariates matter with respect to how accurately the covariate effects can be estimated. Moreover, estimation accuracy can be improved by prioritizing smooth and weak spatial random effects compared to the covariates. Our results help practitioners of the current SDMs to alleviate the problems related to spatial confounding.

Acknowledgements – We thank all members of the BioGeoClimate Modelling Lab for assisting with the fieldwork.
Funding – This work was funded by the Academy of Finland (grant no. 317255) and Jane and Aatos Erkko foundation (Jussi Mäkinen, Elina Numminen and Jarno Vanhatalo); Kone foundation, Societas pro Fauna et Flora Fennica (Pekka Niittynen).

Author contributions

Jussi Mäkinen: Conceptualization (equal); Formal analysis (equal); Methodology (equal); Validation (lead); Visualization (lead); Writing – original draft (lead). **Elina Numminen**: Conceptualization (equal); Formal analysis (equal); Methodology (equal); Supervision (equal); Validation (equal); Visualization (supporting); Writing – original draft (equal). **Pekka Niittynen**: Conceptualization (supporting); Data curation (lead); Formal analysis (supporting); Resources (lead); Validation (supporting); Visualization (supporting); Writing – original draft (supporting). **Miska Luoto**: Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Resources (supporting); Supervision (equal); Validation (supporting); Visualization (supporting); Writing – original draft (supporting). **Jarno Vanhatalo**: Conceptualization (equal); Formal analysis (equal); Funding acquisition (lead); Methodology (equal); Software (lead); Supervision (equal); Validation (equal); Visualization (supporting); Writing – original draft (equal).

Transparent peer review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.06183>.

Data availability statement

Data are available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.hdr7sqvm5> (Mäkinen et al. 2022).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Araújo, M. B. et al. 2019. Standards for distribution models in biodiversity assessments. – Sci. Adv. 5: eaat4858.
Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. – Ecol. Model. 200: 1–19.
Betts, M. G. et al. 2009. Comment on 'Methods to account for spatial autocorrelation in the analysis of species distributional data: a review'. – Ecography 32: 374–378.
Bini, L. M. et al. 2009. Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. – Ecography 32: 193–204.
Clark, J. S. 2004. Why environmental scientists are becoming Bayesians. – Ecol. Lett. 8: 2–14.
Cressie, N. A. and Wikle, C. K. 2015. Statistics for spatio-temporal data. – Wiley.
Diniz-Filho, J. A. F. et al. 2003. Spatial autocorrelation and red herrings in geographical ecology. – Global Ecol. Biogeogr. 12: 53–64.
Dupont, E. et al. 2021. Spatial+: a novel approach to spatial confounding. – Biometrics.
Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.
Feng, X. et al. 2019a. A checklist for maximizing reproducibility of ecological niche models. – Nat. Ecol. Evol. 3: 1382–1395.
Feng, X. et al. 2019b. Collinearity in ecological niche modeling: confusions and challenges. – Ecol. Evol. 9: 10365–10376.
Finley, A. O. et al. 2007. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. – J. Stat. Softw. 19: 1.
Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. – Cambridge Univ. Press.
Fuglstad, G. A. et al. 2018. Constructing priors that penalize the complexity of Gaussian random fields. – J. Am. Stat. Assoc. 114: 445–452.
Guélat, J. and Kéry, M. 2018. Effects of spatial autocorrelation and imperfect detection on species distribution models. – Methods Ecol. Evol. 2018: 1614–1625.
Hanks, E. M. et al. 2015. Restricted spatial regression in practice: geostatistical models, confounding and robustness under model misspecification. – Environmetrics 26: 243–254.

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hawkins, B. A. et al. 2007. Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. – Ecography 30: 375–384.
Heikkinen, R. K. et al. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. – Global Ecol. Biogeogr. 16: 754–763.
Helle, I. et al. 2020. Impacts of oil spills on Arctic marine ecosystems: a quantitative and probabilistic risk assessment perspective. – Environ. Sci. 54: 2112–2121.
Hodges, J. S. and Reich, B. J. 2010. Adding spatially-correlated errors can mess up the fixed effect you love. – Am. Stat. 64: 325–334.
Hughes, J. and Haran, M. 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. – J. R. Stat. Soc. B 75: 139–159.
Johnson, D. S. et al. 2013. Spatial occupancy models for large data sets. – Ecology 94: 801–808.
Kallasvuo, M. et al. 2017. Modeling the spatial distribution of larval fish abundance provides essential information for management. – Can. J. Fish. Aquat. Sci. 74: 636–649.
Keil, P. et al. 2014. Uncertainty, priors, autocorrelation and disparate data in downscaling of species distributions. – Divers. Distrib. 20: 797–812.
Kim, D. 2021. Predicting the magnitude of residual spatial autocorrelation in geographical ecology. – Ecography 44: 1121–1130.
Kissling, W. D. and Carl, G. 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. – Global Ecol. Biogeogr. 17: 59–71.
Lany, N. K. et al. 2019. Complementary strengths of spatially-explicit and multi-species distribution models. – Ecography 43: 456–466.
Legendre, P. 1993. Spatial autocorrelation trouble or new paradigm. – Ecology 74: 1659–1673.
Lembrechts, J. J. et al. 2019. Incorporating microclimate into species distribution models. – Ecography 42: 1267–1279.
Lemoine, N. P. 2019. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. – Oikos 128: 912–928.
Lichstein, J. W. et al. 2002. Spatial autocorrelation and autoregressive models in ecology. – Ecol. Monogr. 72: 445–463.
Luoto, M. and Heikkinen, R. K. 2008. Disregarding topographical heterogeneity biases species turnover assessments based on bioclimatic models. – Global Change Biol. 14: 483–494.
MATLAB 2018. Version 9.5 (R2018b). – The MathWorks Inc., Natick, MA, USA.
Mäkinen, J. and Vanhatalo, J. 2018. Hierarchical Bayesian model reveals the distributional shifts of Arctic marine mammals. – Divers. Distrib. 24: 1381–1394.
Mäkinen, J. et al. 2022. Data from: Spatial confounding in Bayesian species distribution modeling. – Dryad Digital Repository, <https://doi.org/10.5061/dryad.hdr7sqvm5>.
Mielke, K. P. et al. 2020. Disentangling drivers of spatial autocorrelation in species distribution models. – Ecography 43: 1741–1751.
Minasny, B. and McBratney, A. B. 2005. The Matérn function as a general model for soil variograms. – Geoderma 128: 192–207.
Niittynen, P. and Luoto, M. 2018. The importance of snow in species distribution models of arctic vegetation. – Ecography 41: 1024–1037.
Niittynen, P. et al. 2020. Decreasing snow cover alters functional composition and diversity of Arctic tundra. – Proc. Natl Acad. Sci. USA 117: 21480–21487.
Ovaskainen, O. et al. 2015. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. – Methods Ecol. Evol. 7: 549–555.
Paciorek, C. J. 2010. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. – Stat. Sci. 25: 107–125.
Page, G. L. et al. 2017. Estimation and prediction in the presence of spatial confounding for spatial linear models. – Scand. J. Stat. 44: 780–797.
Reich, B. J. et al. 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. – Biometrics 62: 1197–206.
Renner, I. W. et al. 2019. Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization. – Methods Ecol. Evol. 10: 2118–2128.
Sequeira, A. M. et al. 2016. Transferability of predictive models of coral reef fish species richness. – J. Appl. Ecol. 53: 64–72.
Shurin, J. B. et al. 2009. Spatial autocorrelation and dispersal limitation in freshwater organisms. – Oecologia 159: 151–159.
Soberón, J. and Nakamura, M. 2009. Niches and distributional areas: concepts, methods and assumptions. Proc. Natl Acad. Sci. USA 106: 19644–19650.
Soerbye, S. H. et al. 2019. Careful prior specification avoids incautious inference for log-Gaussian Cox point processes. – J. R. Stat. Soc. C Appl. 68: 543–564.
Soriano-Redondo, A. et al. 2019. Understanding species distribution in dynamic populations: a new approach using spatio-temporal point process models. – Ecography 42: 1092–1102.
Thorson, J. T. et al. 2015. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. – Methods Ecol. Evol. 6: 627–637.
Tikhonov, G. et al. 2020. Joint species distribution modelling with the r-package Hmsc. – Methods Ecol. Evol. 11: 442–447.
Valavi, R. et al. 2021. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. – Ecol. Monogr. 1: e01486.
Vanhatalo, J. et al. 2010. Approximate inference for disease mapping with sparse Gaussian processes. – Stat. Med. 29: 1580–1607.
Vanhatalo, J. et al. 2020. Additive multivariate Gaussian processes for joint species distribution modeling with heterogeneous data. – Bayesian Anal. 15: 415–447.
Vanhatalo, J. et al. 2013. GPstuff: Bayesian modeling with Gaussian processes. – J. Mach. Learn. Res. 14: 1175−1179.
Vehtari, A. and Ojanen, J. 2012. A survey of Bayesian predictive methods for model assessment, selection and comparison. – Stat. Surv. 6: 142–228.
Ver Hoef, J. M. et al. 2018. Spatial autoregressive models for statistical inference from ecological data. – Ecol. Monogr. 88: 36–59.
Wakefield, J. 2007. Disease mapping and spatial regression with count data. – Biostatistics 8: 158–183.
Wisz, M. S. et al. 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. – Biol. Rev. 88: 15–30.
Yates, K. L. et al. 2018. Outstanding challenges in the transferability of ecological models. – Trends Ecol. Evol. 33: 790–802.

16000587, 2022, 11, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ecog.06183 by University Of Helsinki, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License