ECOGRAPHY

Research

Complementary strengths of spatially-explicit and multi-species distribution models

Nina K. Lany, Phoebe L. Zarnetske, Andrew O. Finley and Deborah G. McCullough

N. K. Lany (https://orcid.org/0000-0003-0868-266X) ✉ (lanynina@msu.edu), A. O. Finley (https://orcid.org/0000-0002-2277-2912) and D. G. McCullough (https://orcid.org/0000-0002-9765-0338), Dept of Forestry, Michigan State Univ., East Lansing, MI, USA. DGM also at: Dept of Entomology, Michigan State Univ., East Lansing, MI, USA. – P. L. Zarnetske (https://orcid.org/0000-0001-6257-6951) and NKL, Ecology, Evolutionary Biology, and Behavior Program, Michigan State Univ., East Lansing, MI, USA. PLZ also at: Dept of Integrative Biology, Michigan State Univ., East Lansing, MI, USA.

Ecography
43: 456–466, 2020
doi: 10.1111/ecog.04728

Subject Editor: Carsten Dormann
Editor-in-Chief: Miguel Araújo
Accepted 1 November 2019

Species distribution models (SDMs) project the outcome of community assembly processes – dispersal, the abiotic environment and biotic interactions – onto geographic space. Recent advances in SDMs account for these processes by simultaneously modeling the species that comprise a community in a multivariate statistical framework or by incorporating residual spatial autocorrelation in SDMs. However, the effects of combining both multivariate and spatially-explicit model structures on the ecological inferences and the predictive abilities of a model are largely unknown. We used data on eastern hemlock *Tsuga canadensis* and five additional co-occurring overstory tree species in 35 569 forest stands across Michigan, USA to evaluate how the choice of model structure, including spatial and non-spatial forms of univariate and multivariate models, affects ecological inference about the processes that shape community composition as well as model predictive ability.

Incorporating residual spatial autocorrelation via spatial random effects did not improve out-of-sample prediction for the six tree species, although in-sample model fit was higher in the spatial models. Spatial models attributed less variation in occurrence probability to environmental covariates than the non-spatial models for all six tree species, and estimated higher (more positive) residual co-occurrence values for most species pairs. The non-spatial multivariate model was better suited for evaluating habitat suitability and hypotheses about the processes that shape community composition. Environmental correlations and residual correlations among species pairs were positively related, perhaps indicating that residual correlations were due to shared responses to unmeasured environmental covariates. This work highlights the importance of choosing a non-spatial model formulation to address research questions about the species–environment relationship or residual co-occurrence patterns, and a spatial model formulation when within-sample prediction accuracy is the main goal.

Keywords: eastern hemlock, joint species distribution model, Michigan, species co-occurrence, *Tsuga canadensis*, USA

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Community composition in space is a function of both biotic and abiotic processes. A local community is a subset of species that dispersed from a regional species pool (MacArthur and Wilson 1967), and was filtered by both environmental conditions and biotic interactions (Andrewartha and Birch 1954, Rosenzweig and MacArthur 1963, MacArthur 1972). The assembly of communities emphasizes that dispersal, environmental filtering and biotic interactions are not independent nor sequential processes, but rather affect community assembly interactively (HilleRisLambers et al. 2012, Kraft et al. 2015, Cadotte and Tucker 2017). Species distribution models (SDMs) are often used to model the spatial distributions of species and assemblages, and project the outcome of these three processes onto geographic space (Guisan and Zimmermann 2000, Pulliam 2000).

Fundamentally, SDMs use regression models or machine learning to correlate abiotic conditions with occurrence or abundance of a single species to make predictions about responses to climatic change or to identify potential conservation areas solely based on the species–environment relationship (Elith and Leathwick 2009, Araújo and Peterson 2012). Additionally, SDMs can simultaneously model the species that comprise a community by incorporating residual co-occurrence patterns that reflect species interactions or residual responses to unmeasured abiotic conditions in a multivariate modeling framework (Ovaskainen et al. 2010, Kissling et al. 2012, Pollock et al. 2014, Nieto-Lugilde et al. 2018). Models can also account for spatial autocorrelation in SDMs via spatial random effects (Keitt et al. 2002, Latimer et al. 2006). Some modeling approaches combine both the multivariate and spatially-explicit approaches to unify environmental filtering, species co-occurrence patterns and spatial processes into a common framework (Finley et al. 2009a, Latimer et al. 2009, Thorson et al. 2015, 2016, Ovaskainen et al. 2016, Schliep et al. 2018). However, estimates of the regression coefficients that describe the species–environment relationship may be affected by spatial confounding when covariates have spatial structure of their own and spatial random effects are added to the model (Hanks et al. 2015). Effects of combining both multivariate and spatially-explicit model structures on the ecological inferences and the predictive abilities of a model are largely unknown.

We explicitly compare univariate, univariate–spatial, multivariate and multivariate–spatial models of tree species occurrence. We focus on eastern hemlock *Tsuga canadensis* and its associates across Michigan, USA. Eastern hemlock is a long-lived, foundational species of conservation concern because it is threatened across much of its native range by hemlock woolly adelgid *Adelges tsugae*, an invasive sap-feeding insect (Orwig et al. 2012, Havill et al. 2014). Localized infestations of HWA were first detected in in Michigan in 2015. Eastern hemlock does not occur in all locations with suitable habitat (Evans and Gregoire 2007, Doucette et al. 2009, Fitzpatrick et al. 2012, Ferrari et al. 2014), suggesting interactions with other tree species may have strong effects on eastern hemlock occurrence. Indeed, accounting for co-occurring species in the community is fundamentally important for modeling distribution and abundance across a wide variety of tree species (Meier et al. 2010, Clark et al. 2014, Taylor-Rodríguez et al. 2017). Additionally, univariate models that incorporate residual spatial autocorrelation reduce false-positive occurrence predictions for eastern hemlock (Record et al. 2013). We focus on eastern hemlock and its common associates in part because accurate prediction of eastern hemlock occurrence is essential for hemlock woolly adelgid detection and eradication efforts.

With eastern hemlock and its associated species, we quantify how the combination of both multivariate and spatially-explicit elements in a statistical modeling framework affects resulting ecological inferences and predictive ability.

We also compare the magnitude of the improvements in prediction gained by incorporating residual dependence across species versus across space. Specifically, we ask: 1) how does the improvement in prediction gained from a spatially-explicit model compare to that of a multi-species model? 2) How does the choice of model affect ecological inference about environmental filtering, niche overlap and residual co-occurrence patterns?

We expect that the improvement in prediction gained by incorporating dependence among co-occurring species is likely small compared to the improvements gained by incorporating residual spatial autocorrelation, and that the spatial models will yield lower estimates of overall importance of environmental conditions in shaping distribution and abundance (Dormann et al. 2007, Ver Hoef et al. 2018). We would expect negative residual co-occurrence patterns between species that have similar environmental tolerances if competition is structuring these tree communities (Kohli et al. 2018), as trees are limited by space and one species cannot increase unless other species decline (Clark et al. 2014, Taylor-Rodríguez et al. 2017). However, residual co-occurrence patterns often result from shared responses to unmeasured environmental covariates (Dormann et al. 2018, and references within). In this case we would expect positive residual co-occurrence values for species that respond similarly to environmental conditions and negative residual co-occurrence values for species that respond differently to environmental conditions (Kohli et al. 2018). The effect of incorporating residual spatial autocorrelation on residual co-occurrence values is unknown.

Methods

Study system

Eastern hemlock occurs across a wide range of abiotic and soil conditions (Rogers 1978). It was originally thought to be a relic of glacial eras that persists only in cool moist pockets (Clements 1934), but the current view holds that eastern

hemlocks have wide climatic tolerances similar to co-occurring shade tolerant northern hardwoods (George et al. 1974). Hemlock is commonly associated with both deciduous and evergreen species in eastern forests, including sugar maple *Acer saccharum*, eastern white pine *Pinus strobus*, northern white-cedar *Thuja occidentalis* and yellow birch *Betula alleghaniensis* (Frelich et al. 1993). We focused on these focal species in addition to eastern hemlock because they are prevalent, commonly associated with eastern hemlock in Michigan, and have not been widely planted in forested areas. We also included one common species – jack pine *Pinus banksiana* – that is not commonly associated with eastern hemlock.

Forest inventory data

We compiled data on the occurrence of tree species from 72 689 forest stand inventories performed on state and national forest land in Michigan, USA. We obtained percent cover of each overstory species that comprised at least 2% of the canopy on state land from the State of Michigan GIS Open Data Portal (Michigan Dept of Natural Resources 2013) and percent cover of overstory tree species on National Forest land from USDA Forest Service (United States Dept of Agriculture 2011). Data from each stand were spatially referenced as polygons. We included stands that ranged in area from 0.1 ha to 8.1 ha in the analyses (n = 35 569 stands). The maximum distance between pairs of stands was 649 km. Estimates of relative abundance (percent cover) were converted into presence/absence for analysis (see Supplementary material Appendix 1 Table A1 for the prevalence of each focal tree species).

Abiotic covariates

We obtained data on abiotic conditions in raster format for Michigan from online data repositories and prepared them for analysis using R ver. 3.4.4 (R Core Team), GDAL (GDAL Development Team 2017) and the 'rgdal' package (Bivand et al. 2018). Elevation data were obtained at 1-arc second (approx. 30 m) resolution from the NASA Shuttle Radar Topography Mission via USGS (NASA Jet Propulsion Laboratory 2013). Slope and aspect were calculated from the topography data for a 3 × 3 pixel kernel around each central pixel. Available soil water storage (0–100 cm depth) was obtained at 10 m resolution from the gridded soil survey geographic (gSSURGO) database (Soil Survey Staff 2017).

We acquired 30-yr climate normals for monthly mean temperature and mean monthly precipitation at 800 m resolution over the period 1981–2010 from the PRISM database (PRISM Climate Group 2017). These data layers were used to calculate bioclimatic variables commonly used in SDMs using the R package dismo (Hijmans et al. 2017). We chose covariates that were both biologically meaningful and not highly correlated with one another (max r = 0.50), including: minimum temperature (BIOCLIM 6); precipitation sum (BIOCLIM 19) during the coldest quarter of the year; maximum temperature (BIOCLIM 5); and precipitation sum (BIOCLIM 18) during the warmest quarter of the year.

We calculated two additional, biologically meaningful covariates – actual evapotranspiration and climatic water deficit – at 30 m resolution across Michigan. These variables describe how temperature and water interact to affect plants in a mechanistic way, and are often good predictors of tree distributions (Stephenson 1998, Lutz et al. 2010). Climatic water deficit describes the evaporative demand not met at a site – i.e. the additional amount of water that would have been transpired by vegetation had it been available (Thornthwaite 1948, Stephenson 1998). The two covariates were calculated following Itter et al. (2017) using monthly mean temperature and precipitation (30-yr normals), slope, aspect, latitude and soil water storage. Actual evapotranspiration and minimum winter temperature were positively correlated (r = 0.62).

We also obtained raster data on land cover class derived from Landsat imagery from the National Land Cover Database (NLCD) at 30 m resolution (Homer et al. 2015). All gridded data layers were reprojected to match the resolution of the NLCD land cover class data (30 × 30 m) using the bilinear method. Values for cells that were not classified as forested according to the NLCD cover class database were removed from all covariate layers. Covariates for each stand were extracted as the mean value of the grid cells that comprised the polygon outlining each stand. The covariates displayed spatial structure (Supplementary material Appendix 1 Fig. A1) and were mean centered and standardized for analysis.

Statistical models

We compared four generalized linear logistic regression models, including univariate, spatially-explicit univariate, multivariate and spatially-explicit multivariate. These models estimated the probability of occurrence for each focal species in each stand using minimum winter temperature (MIN), maximum summer temperature (MAX), total precipitation in the coldest quarter of the year (WIP), total precipitation in the warmest quarter of the year (SUP), annual actual evapotranspiration (AET) and annual climatic water deficit (DEF) as fixed covariates. Spatially-varying intercepts were added using either a univariate or multivariate Gaussian Process. A linear model of coregionalization was used to estimate the cross-covariance within the multivariate model (Gelfand et al. 2004, Banerjee et al. 2014). Due to computational limitations, the occurrence data were randomly split and a dataset of n = 17 784 stands was used to fit the full models.

Inference was obtained in a Bayesian framework using uninformative priors. Regression coefficients, β, described the relationship between occurrence probability and each covariate specific to each tree species and were assigned unbounded uniform priors. In spatial models, spatial decay parameters, ϕ, were assigned a uniform prior with support across the extent of the study area. For the univariate spatial model, the variance parameter Σ^2 was assigned an inverse-Gamma prior with a shape of 2 and scale of 1. Similarly, the variance–covariance matrix parameter $\mathbf{K}$ in the multivariate models, with N = 6 species, was assigned an inverse-Wishart prior with N + 1 degrees of freedom and N × N diagonal scale

matrix with diagonal elements set to 0.1. Given the computational demand induced by the size of the training dataset, the Gaussian processes were replaced with their Gaussian predictive process approximations (Banerjee et al. 2008, Finley et al. 2009b).

First, univariate and univariate spatial models were fit separately for each of the six tree species for 50 000 MCMC iterations with a target acceptance rate of 0.42 using the R package spBayes (Finley et al. 2007, 2015). spBayes uses the logistic implementation of the multivariate model as described in Wilkinson et al. (2019, section 2.3.2). The first 37 500 iterations were discarded as burn-in. Then, we ran a multivariate spatial model using the posterior mean parameter estimates from the univariate spatial models as initial values and wide proposal variances to keep the acceptance rate as close to 0.42 as possible. The multivariate spatial model was run for 5000 iterations, and the first 2500 were discarded as burn-in. The non-spatial multivariate model was fit using the open-source JAGS software (Plummer 2003) within R following the hierarchical, latent-variable probit formulation using code published in Pollock et al. (2014, Supplementary material Appendix 1). This model formulation is described in Wilkinson et al. (2019, section 2.3.3). The model was run for 50 000 iterations and the first 37 500 were discarded as burn-in. The residual variance–covariance matrix for each multivariate model was rescaled to create a correlation matrix.

Model evaluation

We performed out-of sample prediction using blocked cross-validation (Roberts et al. 2017) for each of the models. We split the 17 784 stands according to the five large hydrologic units (watersheds) that comprised the data (Fig. 1). The watershed boundary dataset (United States Geological Survey 2017) was developed as a coordinated effort between the United States Dept of Agriculture-Natural Resources Conservation Service (USDA-NRCS), the United States Geological Survey (USGS) and the Environmental Protection Agency (EPA). Spatial autocorrelation of each environmental covariate was evaluated with correlograms using the R package 'ncf' (Bjornstad 2019), and positive spatial correlation extended to 150 km or less for all covariates. Blocking by large watershed for out-of-sample prediction minimized dependence between the training and test datasets. We performed cross validation by sequentially withholding data from each watershed, fitting the model and predicting to the holdout watershed. We also evaluated in-sample model fit using the 17 785 stands randomly withheld from the full dataset. We calculated the area under the curve (AUC), a commonly-used model performance statistic (Fielding and Bell 1997). Potential values of the AUC are between 0 and 1, where 0.5 is no better than random.

We calculated the portion of the total variation in occurrence probability attributable to abiotic covariates for each species as the sum of the products of the squared regression coefficients for that species (which represents the variance due to environmental covariates because the abiotic covariates are standardized to unit variance) divided by the total variance of that species, which is the variance due to environmental covariates plus residual variance (Sæther et al. 2000, Mutshinda et al. 2011). The component of between-species correlation due to shared responses to environmental conditions was calculated for each species-pair as a function of the products of the regression coefficients and covariances of the environmental covariates obtained from the multivariate model following Pollock et al. (2014). We evaluated the absolute shifts in pairwise residual co-occurrence as the posterior mean estimate from the multivariate spatial residual correlation matrix minus the posterior mean estimate from the multivariate residual correlation matrix for each species-pair.

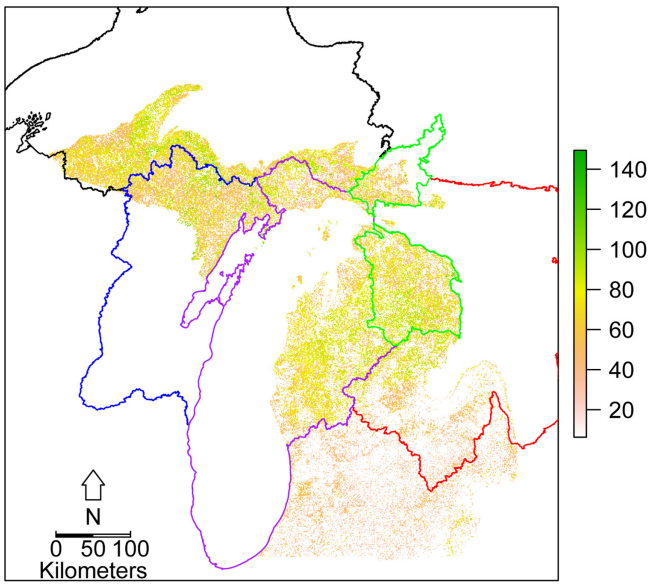

Figure 1. Climatic water deficit (mm yr^{-1}) in Michigan, USA. Colored polygons indicate the five large watersheds used for blocked cross-validation.

Results

Incorporating residual spatial autocorrelation did not consistently improve out-of-sample prediction accuracy, as measured by AUC, but did improve model fit (Fig. 2). However, incorporating residual dependence among co-occurring species did not affect model fit or out-of-sample prediction (Fig. 2). Areas with the highest predicted occurrence probability also had the greatest uncertainty associated with prediction (shown for eastern hemlock in Fig. 3). Modeled spatial random effects, predicted occurrence probability and prediction uncertainty for the remaining focal species are shown in Supplementary material Appendix 1 Fig. A2–A6.

The spatial models attributed less variation to environmental conditions than the non-spatial models (Fig. 4). We used the non-spatial multivariate model to compare posterior estimates of the β coefficients that describe the species–environment relationship across species. Responses to

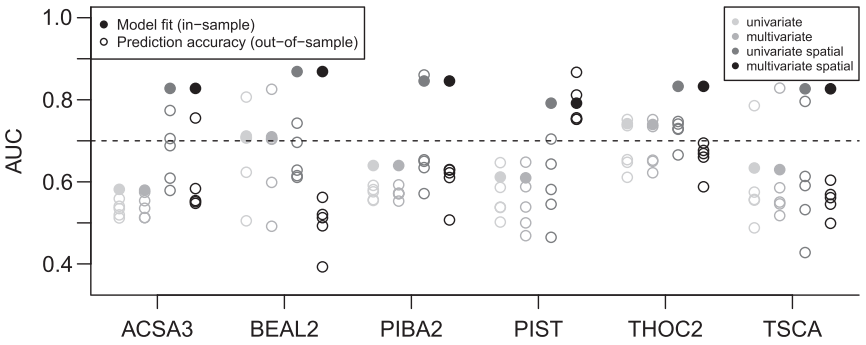

Figure 2. Comparison of predictive accuracy, as indicated by area under the curve (AUC), for six overstory tree species in Michigan, USA according to four different model specifications. The horizontal dashed line indicates the minimum AUC required for a useful model (0.7). Tree species codes are given according to the United States Dept of Agriculture standard codes: ACSA3 – *Acer saccharum*; BEAL2 – *Betula alleghaniensis*; PIBA2 – *Pinus banksiana*; PIST – *Pinus strobus*; THOC2 – *Thuja occidentalis*; TSCA – *Tsuga canadensis*.

abiotic covariates were species-specific (Fig. 5). Yellow birch and northern white cedar occurrence probabilities were most strongly explained by the included environmental covariates. Occurrence probability for yellow birch had the largest proportion of variation attributed to environmental covariates (0.76), and was related to every covariate except winter precipitation. Relationships with minimum winter temperature and summer precipitation were positive, indicating that warmer winters and wetter summers were associated with a higher probability of occurrence. Relationships with summer temperature, actual transpiration and climatic water deficit were negative, indicating that yellow birch is particularly sensitive to warm summers and drought conditions. For northern white cedar, 73% of the variation in occurrence

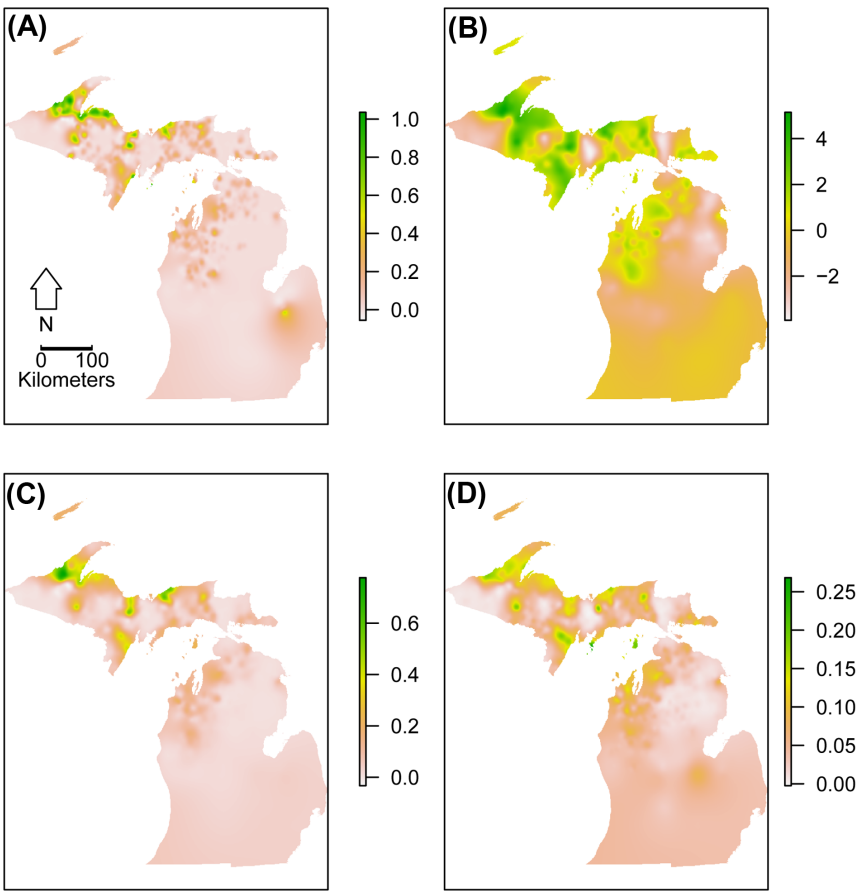

Figure 3. Interpolated occurrence data (A), spatial random effect estimates (B), predicted occurrence probability (C) and prediction uncertainty (D) for eastern hemlock *Tsuga canadensis*. Fitted values and spatial random effects made with a spatially-explicit univariate model with dimension reduction to 300 knots.

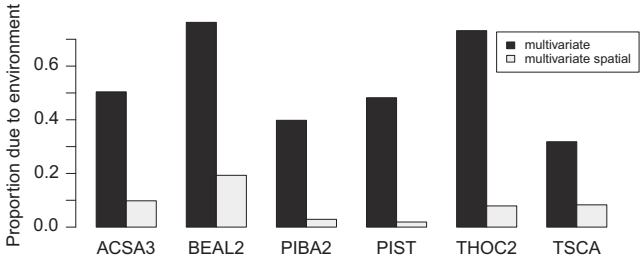

Figure 4. Proportion of variation attributed to environmental covariates in the multivariate and multivariate spatial models of tree species occurrence probability. Tree species codes are given according to the United States Dept of Agriculture standard codes: ACSA3 – *Acer saccharum*; BEAL2 – *Betula alleghaniensis*; PIBA2 – *Pinus banksiana*; PIST – *Pinus strobus*; THOC2 – *Thuja occidentalis*; TSCA – *Tsuga canadensis*.

probability was explained by environmental covariates. Occurrence probability of northern white cedar was positively related to winter temperature and actual transpiration, but negatively related to summer temperature and precipitation, and climatic water deficit.

Sugar maple and white pine occurrence probabilities were partly explained by summer environmental covariates. For sugar maple, 50% of the variation in occurrence probability was explained by environmental covariates. Sugar maple occurrence probability was positively related to winter temperature and summer precipitation, but negatively related to actual evapotranspiration and climatic water deficit. Eastern white pine occurrence probability also had a relatively large proportion of variation attributed to environmental covariates (0.48), and was related to every covariate except actual evapotranspiration. Occurrence probability was positively related to summer temperature, climatic water deficit and winter temperature and precipitation, but negatively related to summer precipitation.

Less of the variation in occurrence probability for either jack pine and eastern hemlock was explained by the included environmental covariates. For jack pine, 40% of the variation in occurrence probability was explained by environmental covariates. There was a strong positive relationship with winter precipitation and climatic water deficit, along with a strong negative relationship with minimum winter temperature. For eastern hemlock, 32% of the variation in occurrence probability was explained by environmental covariates. Eastern hemlock occurrence probability was positively related to winter temperature and summer precipitation, but negatively related to summer temperature, actual evapotranspiration and climatic water deficit.

Because the responses to each of the six environmental covariates were highly variable across species, quantifying pairwise shared responses to the covariates we included (i.e. the environmental correlations) was helpful for interpreting the overlap in response to environmental conditions among species. Shared responses to environmental conditions ranged from strongly positive to strongly negative and were positively related to residual correlation (Fig. 6). For example, eastern hemlock showed both positive environmental correlation and positive residual correlation with yellow birch, northern white cedar and sugar maple, whereas both environmental and residual correlations between eastern hemlock and jack pine were negative. However, eastern hemlock showed negative environmental but positive residual correlation with white pine. Changes in the magnitude of posterior residual co-occurrence estimates occurred in the multivariate versus multivariate spatial models (Fig. 7). Posterior mean estimates of residual correlation were higher in the multivariate spatial model in all but two instances.

Discussion

Incorporating residual spatial autocorrelation via spatial random effects did not improve out-of-sample prediction accuracy. Although superior performance of a spatial model compared with a non-spatial model has been observed in other studies (Record et al. 2013, Roberts et al. 2017), prediction accuracy may be optimistically high when blocked cross-validation methods are not used (Roberts et al. 2017). In this study, model fit was higher in the spatial models. When spatial prediction or spatial smoothing is the primary aim, and model over-fitting is less of a concern, spatial random effects models are a logical choice (Ver Hoef et al. 2018). For example, resources for surveying for the invasive hemlock woolly adelgid that attacks and kills eastern hemlocks are limited, and stakeholders are interested in accurately predicting hemlock occurrence probability at unobserved locations within the matrix of observed locations to help decide where to allocate detection efforts.

The multivariate spatial model attributed substantially less variation in occurrence probability to environmental covariates for all six tree species than the non-spatial multivariate model. This was consistent with our expectation that the spatial models in general give lower estimates of overall importance of environmental conditions in shaping distribution and abundance (Dormann et al. 2007). Some authors have argued that adding spatial random effects to account for spatially autocorrelated residuals is essential for interpreting the species–environment relationship (Keitt et al. 2002, Latimer et al. 2006). Alternatively, Bini et al. (2009) showed that coefficient shifts in spatial versus non-spatial models are idiosyncratic, and therefore ecological interpretation of beta coefficients should be performed cautiously for both spatial and non-spatial models. Residual spatial autocorrelation may not induce bias in coefficient estimates, although it does affect the standard errors of the coefficient estimates (Diniz-Filho et al. 2003). Spatial confounding, however, can affect coefficient estimates when covariates have strong spatial structure of their own (Reich et al. 2006, Hanks et al. 2015). Methods for handling spatial confounding in spatial regression are continuing to develop, but currently non-spatial models are preferred when interpreting regression coefficients (Reich et al. 2006, Hodges and Reich 2010). Therefore,

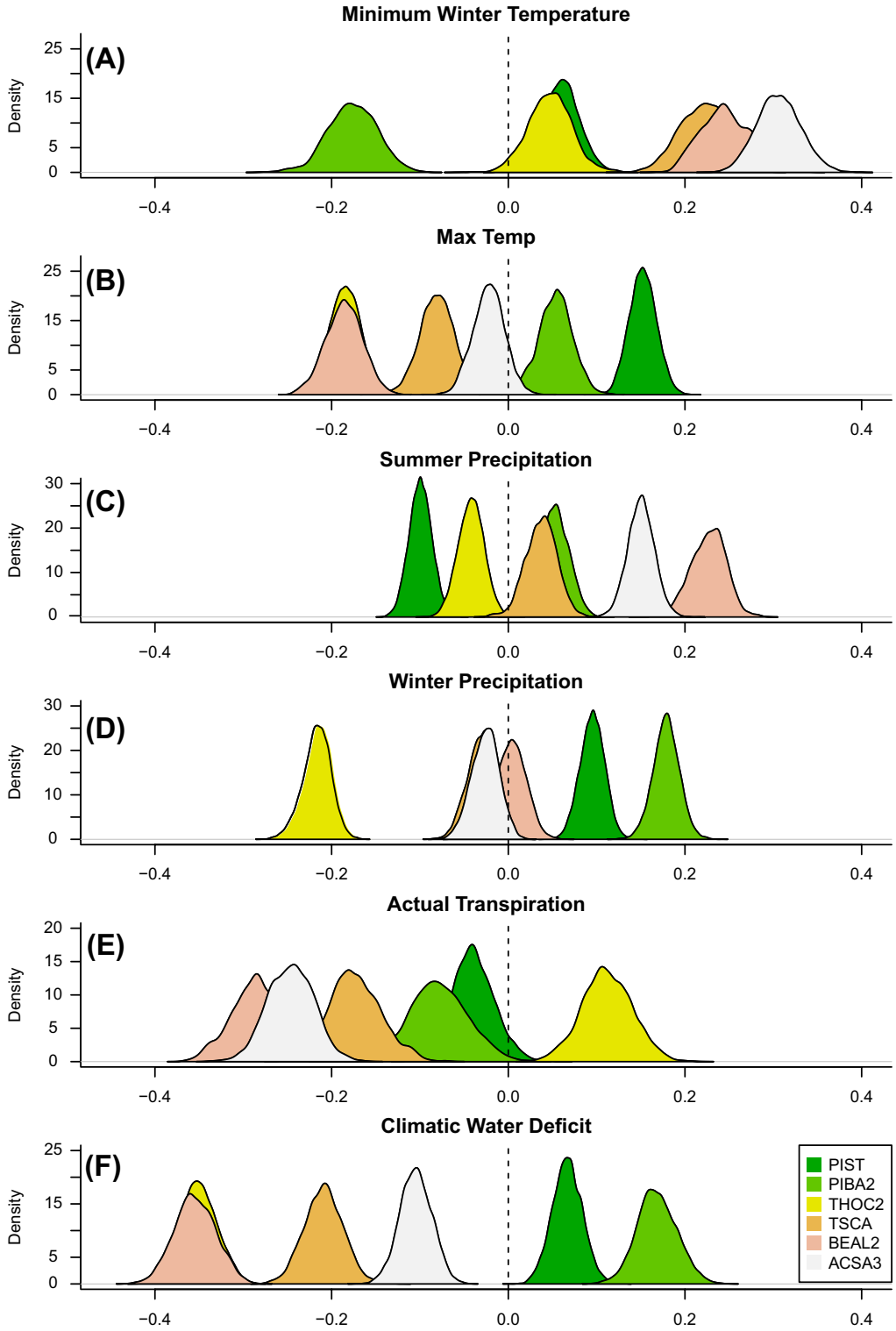

Figure 5. Comparison of posterior estimates of beta parameters describing the species–environment relationship for each abiotic covariate from the multivariate model across all six tree species. Covariates were calculated from 30-yr normals (1981–2010) and included minimum winter temperature (A), maximum summer temperature temperature (B), sum of precipitation in the warmest three months (C), sum of precipitation in the coldest three months (D), actual evapotranspiration (E) and climatic water deficit (F). Tree species codes are given according to the United States Dept of Agriculture standard codes: ACSA3 – *Acer saccharum*; BEAL2 – *Betula alleghaniensis*; PIBA2 – *Pinus banksiana*; PIST – *Pinus strobus*; THOC2 – *Thuja occidentalis*; TSCA – *Tsuga canadensis*.

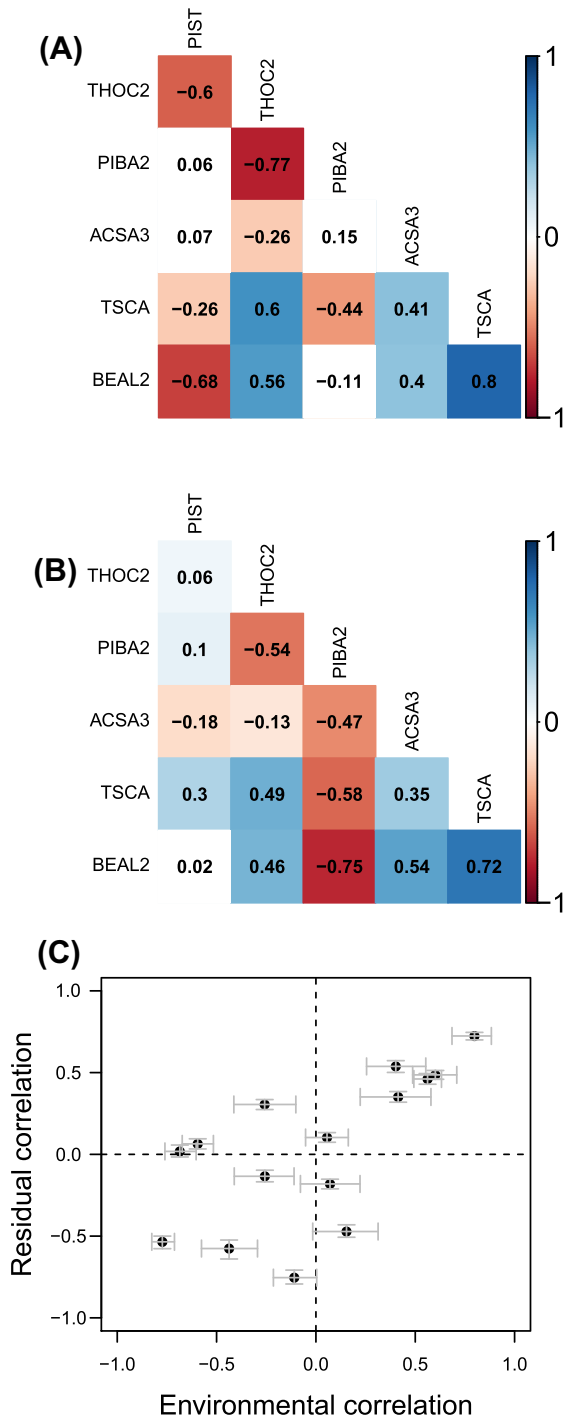

Figure 6. Interspecific correlations due to shared responses to environmental covariates (A) and residual dependence (B) for six overstory tree species in Michigan, USA. Environmental and residual correlations tended to be positively related (C). In A and B, posterior parameter estimates that were not significantly different than zero according to the 95% credible interval are not shaded. In (C), bars represent 95% posterior credible intervals. Tree species codes are given according to the United States Dept of Agriculture standard codes: ACSA3 – *Acer saccharum*; BEAL2 – *Betula alleghaniensis*; PIBA2 – *Pinus banksiana*; PIST – *Pinus strobus*; THOC2 – *Thuja occidentalis*; TSCA – *Tsuga canadensis*.

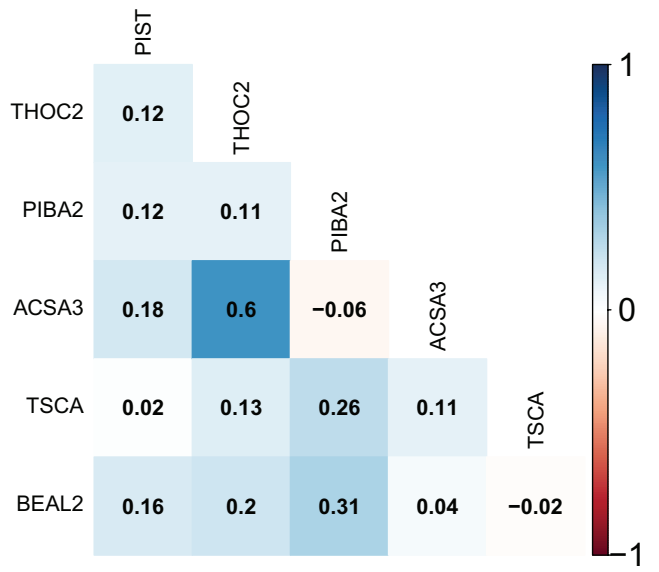

Figure 7. Magnitude of shifts in residual co-occurrence parameters for multivariate versus multivariate spatial models. Positive values (blue tones) indicate residual pairwise co-occurrence was higher in the multivariate spatial model. Tree species codes are given according to the United States Dept of Agriculture standard codes: ACSA3 – *Acer saccharum*; BEAL2 – *Betula alleghaniensis*; PIBA2 – *Pinus banksiana*; PIST – *Pinus strobus*; THOC2 – *Thuja occidentalis*; TSCA – *Tsuga canadensis*.

we used the non-spatial multivariate model for evaluating hypotheses about the species–environment relationship and the processes that shape community composition.

The posterior estimates of the parameters in the non-spatial multivariate model – the β coefficients that describe the species–environment relationship and the environmental correlations among species – were consistent with published descriptions of the range and natural history of each species (Burns and Honkala 1990) and with Michigan forest community types (Burger and Kotar 2003). Eastern hemlock, sugar maple and yellow birch had similar β species–environment relationships. The direction of the relationship between occurrence probability and each environmental covariate was the same across species, and the magnitude of the effect was often similar. This group of species was more likely to occur where winters were comparatively warm, summers were comparatively cool and moist, and both actual transpiration and climatic water deficit were low. This supports the hypothesis that eastern hemlock has broad physiological tolerances similar to co-occurring late-successional overstory species (Rogers 1978), as opposed to the relict hypothesis (Clements 1934), and that abiotic processes have a large role in structuring these communities at the stand scale. Hemlock-maple-yellow birch associations are common in upland deciduous forests Michigan (Burger and Kotar 2003). All three species are described as occurring in cool, moist areas and none of the distributions of these three species extend above the 50° North parallel (Burns and Honkala 1990). The northernmost points in the study area approach the northern distributional limits of these species.

The two *Pinus* species, jack pine and white pine, shared unique responses to several environmental conditions. No other species showed a positive relationship between occurrence probability and summer temperature, winter precipitation or climatic water deficit. Both *Pinus* species are drought tolerant (Burns and Honkala 1990). Indeed, estimates of the effect of climatic water deficit coefficients closely matched the published drought sensitivity classifications for all of the six species included in the study. According to Gustafson and Sturtevant (2013), jack pine and white pine are drought tolerant, sugar maple is somewhat drought tolerant and eastern hemlock, northern white cedar and yellow birch are somewhat intolerant. This matches the relative drought sensitivities estimated by the multivariate model shown in Fig. 5 and the relative drought sensitivities estimated by the multivariate spatial model.

Two species, jack pine and northern white cedar, each had one β estimate that was of different direction and magnitude than other species. These unique differences matched the natural history description of each species. Jack pine was the only species that showed a negative relationship between occurrence probability and winter temperature (i.e. had higher occurrence probability in places with cold winters). The native range of jack pine extends above the 65° North parallel, which is further north than those of the other focal species (Burns and Honkala 1990). Jack pine was at the lower edge of its range limits in this study. Selective management for species of higher economic value than jack pine on more productive sites in Michigan, along with management actions promoting young, dense jack pine stands for Kirtland's warbler *Setophaga kirtlandii* nesting habitat, could also contribute to this pattern. Northern white cedar is sensitive to snow and ice damage (Burns and Honkala 1990), and was the only species with a significant negative relationship between occurrence probability and winter precipitation, which typically falls as snow in the study area (Fig. 5).

Environmental and residual correlations among species-pairs were positively related and typically had the same sign (Fig. 6), indicating that residual correlations likely resulted from shared responses to unmeasured environmental covariates (Kohli et al. 2018). We incorporated meaningful, physiology-based covariates derived from soil, topograpy and climate characteristics for each stand. Climatic water deficit encompasses much about the water holding capacity of the soil ($r = -0.91$) and soil organic carbon (-0.80). These soil characteristics, and soil water holding capacity in particular, can improve models of plant distributions (Cianfrani et al. 2019). Still, unmeasured covariates such as geophysical characteristics or management history could generate the positive residual co-occurrence patterns among eastern hemlock, sugar maple and yellow birch. We are unaware of positive biotic feedbacks between these species at any life stage that could explain the positive residual co-occurrence.

Higher estimates of pairwise residual co-occurrences in the multivariate spatial model versus the non-spatial multivariate model could reflect the positive spatial autocorrelation that was observed in the measured covariates (and perhaps unmeasured environmental covariates). The residual co-occurrence matrix, similar to β coefficients, may be affected by a type of spatial confounding. It is important to demonstrate the effect of choosing a spatially-explicit model structure on estimates of community co-occurrence structure. Given that spatially-autocorrelated covariates are an unavoidable part of ecological studies, future work should explore whether these shifts are consistently positive across systems and whether these shifts can be predicted according to spatial characteristics of the environment or other variables.

In conclusion, we found that incorporating residual spatial autocorrelation via spatial random effects did not improve out-of-sample prediction but did improve model fit. Spatial random effects models were appropriate when spatial smoothing within the dependence structure is the primary aim. However, the spatial models attributed substantially less variation in occurrence probability to environmental covariates than the non-spatial models for all six tree species, and estimated higher residual co-occurrence values for most species pairs. The non-spatial multivariate model was better suited for evaluating hypotheses about environmental filtering, niche overlap and residual co-occurrence patterns. Other approaches for incorporating co-occurrence structure and missing covariates into SDMs, including latent factor analysis and spatial factor analysis (Thorson et al. 2015, 2016, Ovaskainen et al. 2016), are an area of active research that present alternatives to the methods used here, and evaluating abundance rather than binary occurrence could yield different conclusions (Van Couwenberghe et al. 2013). This work highlights the difference between ecological and statistical model selection, and the importance of choosing an appropriate model formulation for a specific research question.

Data availability statement

Data available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.fj6q573qg> (Lany et al. 2019).

Acknowledgements – We thank two anonymous reviewers for insight and helpful suggestions.
Funding – This study was funded by Michigan Dept of Natural Resources Invasive Species Grant Program #IS16-3002, United States Dept of Agriculture Hatch Project #1010055 and Michigan State Univ. NKL was supported by a fellowship from the Arnold and Mabel Beckman Foundation. AOF was supported by National Science foundation grants DMS-1916395, EF-1241874, EF-1253225 and the NASA Carbon Monitoring System program.
Author contributions – NKL, PLZ and DGM designed the study; NKL prepared the data; NKL and AOF performed the analyses; NKL wrote the first draft with input from all authors on revisions.

References

Andrewartha, H. G. and Birch, L. C. 1954. Distribution and abundance of animals. – Univ. of Chicago Press.
Araújo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.

Banerjee, S. et al. 2008. Gaussian predictive process models for large spatial data sets. – J. R. Stat. Soc. B 70: 825–848.
Banerjee, S. et al. 2014. Hierarchical modeling and analysis for spatial data, 2nd ed. – Taylor and Francis.
Bini, M. L. et al. 2009. Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. – Ecography 32: 193–204.
Bivand, R. et al. 2018. rgdal: bindings for the 'geospatial' data abstraction library. – R package ver. 1.4–6, <https://CRAN.R-project.org/package=rgdal>.
Bjornstad, O. N. 2019. ncf: spatial covariance functions. – R package ver. 1.2–8, <https://CRAN.R-project.org/package=ncf>.
Burger, T. L. and Kotar, J. 2003. A guide to forest communities and habitat types of Michigan. – Dept of Forest Ecology and Management, Univ. of Madison-Wisconsin.
Burns, R. M. and Honkala, B. H. 1990. Silvics of North America: 1. conifers; 2. hardwoods. – Agriculture Handbook 654.
Cadotte, M. W. and Tucker, C. M. 2017. Should environmental filtering be abandoned? – Trends Ecol. Evol. 32: 429–437.
Cianfrani, C. et al. 2019. Spatial modelling of soil water holding capacity improves models of plant distributions in mountain landscapes. – Plant Soil 438: 57–70.
Clark, J. S. et al. 2014. More than the sum of the parts: forest climate response from joint species distribution models. – Ecol. Appl. 24: 990–999.
Clements, F. E. 1934. The relict method in dynamic ecology. – J. Ecol. 22: 39–68.
Diniz-Filho, J. A. F. et al. 2003. Spatial autocorrelation and red herrings in geographical ecology. – Global Ecol. Biogeogr. 12: 53–64.
Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.
Dormann, C. F. et al. 2018. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. – Global Ecol. Biogeogr. 27: 1004–1016.
Doucette, J. S. et al. 2009. A rules-based approach for predicting the eastern hemlock component of forests in the northeastern United States. – Can. J. For. Res. 39: 1453–1464.
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.
Evans, A. M. and Gregoire, T. G. 2007. A geographically variable model of hemlock woolly adelgid spread. – Biol. Invas. 9: 369–382.
Ferrari, J. R. et al. 2014. Modeling the spread of invasive species using dynamic network models. – Biol. Invas. 16: 949–960.
Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. – Environ. Conserv. 24: 38–49.
Finley, A. O. et al. 2007. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. – J. Stat. Softw. 63: 1–24.
Finley, A. O. et al. 2009a. Hierarchical spatial models for predicting tree species assemblages across large domains. – Ann. Appl. Stat. 3: 1052–1079.
Finley, A. O. et al. 2009b. Improving the performance of predictive process modeling for large datasets. – Comput. Stat. Data Anal. 53: 2873–2884.
Finley, A. O. et al. 2015. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. – J. Stat. Softw. 63: 1–28.
Fitzpatrick, M. C. et al. 2012. Modeling range dynamics in heterogeneous landscapes: invasion of the hemlock woolly adelgid in eastern North America. – Ecol. Appl. 22: 472–486.
Frelich, L. E. et al. 1993. Patch formation and maintenance in an old-growth hemlock-hardwood forest. – Ecology 74: 513–527.
GDAL Development Team 2017. GDAL – geospatial data abstraction library, ver. 2.1.0. – Open Source Geospatial Foundation.
Gelfand, A. E. et al. 2004. Nonstationary multivariate process modeling through spatially varying coregionalization. – Test 13: 263–312.
George, M. F. et al. 1974. Low temperature exo-therms and woody plant distribution. – Hortic. Sci. 9: 519–522.
Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.
Gustafson, E. J. and Sturtevant, B. R. 2013. Modeling forest mortality caused by drought stress: implications for climate change. – Ecosystems 16: 60–74.
Hanks, E. M. et al. 2015. Restricted spatial regression in practice: geostatistical models, confounding and robustness under model misspecification. – Environmetrics 26: 243–254.
Havill, N. P. et al. 2014. Biology and control of hemlock woolly adelgid. – Technology transfer FHTET-2014-05, USDA Forest Service Forest Health Technology Enterprise Team.
Hijmans, R. J. et al. 2017. dismo: species distribution modeling. – R package ver. 1.1–4, <https://CRAN.R-project.org/package=dismo>.
HilleRisLambers, J. et al. 2012. Rethinking community assembly through the lens of coexistence theory. – Annu. Rev. Ecol. Evol. Syst. 43: 227–248.
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.
Homer, C. G. et al. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States – representing a decade of land cover change information. – Photogramm. Eng. Remote Sens. 81: 345–354.
Itter, M. S. et al. 2017. Variable effects of climate on forest growth in relation to climate extremes, disturbance and forest dynamics. – Ecol. Appl. 27: 1082–1095.
Keitt, T. H. et al. 2002. Accounting for spatial pattern when modeling organism–environment interactions. – Ecography 25: 616–625.
Kissling, W. D. et al. 2012. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. – J. Biogeogr. 39: 2163–2178.
Kohli, B. A. et al. 2018. A trait-based framework for discerning drivers of species co-occurrence across heterogeneous landscapes. – Ecography 41: 1921–1933.
Kraft, N. J. B. et al. 2015. Community assembly, coexistence and the environmental filtering metaphor. – Funct. Ecol. 29: 592–599.
Lany, N. K. et al. 2019. Data from: Complimentary strengths of spatially-explicit and multi-species distribution models. – Dryad Digital Repository, <https://doi.org/10.5061/dryad.fj6q573qg>.
Latimer, A. M. et al. 2006. Building statistical models to analyze species distributions. – Ecol. Appl. 16: 33–50.
Latimer, A. M. et al. 2009. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. – Ecol. Lett. 12: 144–154.
Lutz, J. A. et al. 2010. Climatic water deficit, tree species ranges and climate change in Yosemite National Park. – J. Biogeogr. 37: 936–950.

MacArthur, R. H. 1972. Geographical ecology: patterns in the distribution of species. – Princeton Univ. Press.
MacArthur, R. H. and Wilson, E. O. 1967. The theory of island biogeography. – Princeton Univ. Press.
Meier, E. S. et al. 2010. Biotic and abiotic variables show little redundancy in explaining tree species distributions. – Ecography 33: 1038–1048.
Michigan Dept of Natural Resources 2013. Michigan DNR forest canopy data. – <http://gis-michigan.opendata.arcgis.com>, accessed 18 December 2017.
Mutshinda, C. M. et al. 2011. A multispecies perspective on ecological impacts of climatic forcing. – J. Anim. Ecol. 80: 101–107.
NASA Jet Propulsion Laboratory 2013. NASA Shuttle Radar Topography Mission United States 1 arc second. – NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD, <https://lpdaac.usgs.gov>, <http://dx.doi.org/10.5067/MEaSUREs/SRTM/SRTMUS1.003>, accessed 15 April 2015.
Nieto-Lugilde, D. et al. 2018. Multiresponse algorithms for community-level modelling: review of theory, applications and comparison to species distribution models. – Methods Ecol. Evol. 9: 834–848.
Orwig, D. A. et al. 2012. A foundation tree at the precipice: *Tsuga canadensis* health after the arrival of *Adelges tsugae* in central New England. – Ecosphere 3: 16.
Ovaskainen, O. et al. 2010. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. – Ecology 91: 2514–2521.
Ovaskainen, O. et al. 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. – Methods Ecol. Evol. 7: 428–436.
Plummer, M. 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. – <www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf>.
Pollock, L. J. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). – Methods Ecol. Evol. 5: 397–406.
PRISM Climate Group 2017. Oregon State University. – <http://prism.oregonstate.edu>, created 26 March 2017.
Pulliam, H. R. 2000. On the relationship between niche and distribution. – Ecol. Lett. 3: 349–361.
Record, S. et al. 2013. Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change. – Global Ecol. Biogeogr. 22: 760–771.
Reich, B. J. et al. 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. – Biometrics 62: 1197–1206.
Roberts, D. R. et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. – Ecography 40: 913–929.
Rogers, R. S. 1978. Forests dominated by hemlock (*Tsuga canadensis*): distribution as related to site and postsettlement history. – Can. J. Bot. 56: 843–854.
Rosenzweig, M. L. and MacArthur, R. H. 1963. Graphical representation and stability conditions of predator–prey interactions. – Am. Nat. 97: 209–223.
Sæther, B. E. et al. 2000. Population dynamical consequences of climate change for a small temperate songbird. – Science 287: 854–856.
Schliep, E. M. et al. 2018. Joint species distribution modeling for spatio-temporal occurrence and ordinal abundance data. – Global Ecol. Biogeogr. 27: 142–155.
Soil Survey Staff 2017. National value added look up (Valu1) table for the gridded soil survey geographic (gSSURGO). – Database for Michigan served by the USDA-NRCS, <https://gdg.sc.egov.usda.gov/>, accessed 31 October 2017.
Stephenson, N. 1998. Actual evapotranspiration and deficit: biologically meaningful correlates of vegetation distribution across spatial scales. – J. Biogeogr. 25: 855–870.
Taylor-Rodríguez, D. et al. 2017. Joint species distribution modeling: dimension reduction using dirichlet processes. – Bayesian Anal. 12: 939–967.
Thornthwaite, C. W. 1948. An approach toward a rational classification of climate. – Geogr. Rev. 1: 55–94.
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.
Thorson, J. T. et al. 2016. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. – Global Ecol. Biogeogr. 25: 1144–1158.
United States Dept of Agriculture 2011. Field sampled vegetation stands. – United States Dept of Agriculture Forest Service (National Forest System), <www.fs.usda.gov/main/ottawa/landmanagement/gis> accessed 15 December 2017.
United States Geological Survey 2017. Watershed Boundary Dataset for HUC4. – United States Dept of Agriculture, and Natural Resources Conservation Service, <http://datagateway.nrcs.usda.gov>, accessed 22 April 2017.
Van Couwenberghe, R. et al. 2013. Can species distribution models be used to describe plant abundance patterns? – Ecography 36: 665–674.
Ver Hoef, J. M. et al. 2018. Spatial autoregressive models for statistical inference from ecological data. – Ecol. Monogr. 88: 36–59.
Wilkinson, D. P. et al. 2019. A comparison of joint species distribution models for presence–absence data. – Methods Ecol. Evol. 10: 198–211.

Supplementary material (available online as Appendix ecog-04728 at <www.ecography.org/appendix/ecog-04728>). Appendix 1.