Identifying and Separating the Processes Underlying Boreal Forest Understory Community Assembly

Malcolm Itter[1,2], Elina Kaarlejärvi[2], Anna-Liisa Laine[3,2], Leena Hamberg[4], Tiina Tonteri[4], and Jarno Vanhatalo[5,2]

[1]Department of Environmental Conservation, University of Massachusetts Amherst, Amherst, MA, USA
[2]Research Center for Ecological Change, University of Helsinki, Helsinki, Finland
[3]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[4]Natural Resources Institute Finland, Helsinki, Finland
[5]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Abstract

Identifying the ecological processes underlying community assembly remains an elusive goal in community ecology. We formalize assembly hypotheses as alternative models and apply each to predict 1,918 out-of-sample boreal forest understory communities to identify and separate the processes driving community assembly. Models are specified within a Bayesian joint species distribution framework that allows for the inclusion and separation of stochastic processes, environmental filtering, and two different species dependence structures. We found clear evidence that study communities are structured by both environmental filtering and compositional dependence highlighting the importance of selection in community assembly. The relative importance of environmental filtering was greater than compositional dependence in predicting both understory communities and the abundance of constituent species across broad successional and bioclimatic gradients. Contrary to ecological expectations, the inclusion of a flexible residual species dependence structure (accounting for more than compositional dependence) did not improve model predictions after accounting for the strong role of environmental filtering. Our results provide novel inference on the processes underlying community assembly facilitated by applying empirical approximations of alternative assembly processes to predict communities across a range of environmental conditions.

Introduction

Understanding and predicting community assembly is a foundational topic in community ecology. Despite well-developed theory, process-based models capable of making predictions of community structure across a broad range of conditions are limited to select community types (see Chauvet et al., 2017; Falster et al., 2017, for forest community examples). The importance of general process-based models of community assembly is underscored by rapid global change contributing to large-scale biodiversity loss and altered ecosystem function (Tylianakis et al., 2008; Yates et al., 2018; Blowes et al., 2019). Determining the relative importance of deterministic and stochastic processes is critical to our understanding of community assembly and a necessary step toward predicting communities under global change (Chase and Myers, 2011; Götzenberger et al., 2012; Thuiller et al., 2013; D'Amen et al., 2017). However, few methods exist to separate the effects of these processes. Here, we apply a joint species distribution model (JSDM) that allows for the expression and separation of alternative assembly processes. We apply the model to an extensive natural community dataset to test for empirical evidence of alternative processes within local communities across a range of environmental conditions.

Community assembly is hypothesized to be driven by a hierarchy of deterministic and stochastic processes (Vellend, 2016). The regional species pool is affected by speciation and dispersal, which include both deterministic and stochastic components (Mittelbach and Schemske, 2015). At local scales, communities are shaped by deterministic selection, including environmental and biotic filters that affect each species' relative fitness within a local community (Vellend, 2016). The environmental filter describes the exclusion of species lacking the ability to tolerate or persist under local conditions (Keddy, 1992; Kraft et al., 2015). The biotic filter describes interactions among species contributing to their coexistence within a community (Götzenberger et al., 2012). In addition to these deterministic processes, local communities may also be shaped by ecological drift, stochastic processes reflecting random variation in the relative abundances of species such that individual fitness is independent of species identity (Vellend, 2016).

Contemporary approaches to identify the above processes center on the theory of trait-based community ecology and attempt to infer the processes underlying community assembly from observed functional trait distributions (Weiher and Keddy, 1995; McGill et al., 2006; Götzenberger et al., 2012). Despite the prevalence of trait-based community assembly approaches, recent work highlights the challenges associated with detecting specific assembly processes from functional trait patterns alone (see Münkemüller et al., 2020, for a review). Given these challenges, Clark (2016) predicted functional trait distributions by first modeling species' abundances and then mapping predicted abundances to traits. This so-called "species-first" approach yielded better predictions of functional trait distributions than did models fit directly to trait data. This result suggests that species contain more information about the processes shaping a natural community than functional traits alone (Clark, 2016).

The current analysis builds on Clark's species-first, JSDM approach to test for evidence of alternative community assembly processes. We begin by defining a multinomial JSDM for relative abundance data applying a similar structure as several existing and commonly-

applied JSDM frameworks. The multinomial JSDM allows for the inclusion and separation of observed environmental conditions and species dependence. We apply our framework to quantify the empirical evidence for alternative models representing different hypotheses about the processes underlying community assembly to a large holdout dataset spanning 1,918 natural communities composed of 39 species. Building on recent work demonstrating that community assembly processes may vary dynamically over space and time (Jabot et al., 2020), we further estimate the relative importance of different assembly processes across successional and bioclimatic gradients. Additional insights on assembly processes are gained through predictions of natural community structure under a range of environmental conditions.

Material & methods

Natural community data

Natural community data consist of observations of Finnish boreal forest understory communities collected as part of national vegetation inventories. Understory vegetation was surveyed in 1985-1986, 1995, and 2006 across 1,700 unique plot locations established on mineral-soil in forested land. These sites were part of the systematic sampling network connected to the 8th Finnish National Forest Inventory (Reinikainen et al., 2000). This network consisted of clusters located 16 km apart in southern Finland, and 24 and 32 km apart in northern Finland along east–west and north–south axes, respectively. Each cluster consisted of four linearly located sampling sites 400 m apart in southern Finland and three sampling sites 600 m apart in northern Finland. Data includes 1,494 sites measured in 1985-1986, 1,673 sites in 1995, and 435 sites in 2006 (3,602 sites total). The survey in 2006 was part of the BioSoil project carried out under the Forest Focus scheme, a subset of the pan-European International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests plot network (Level I, Lorenz and Fischer, 2013). This survey included only one site per cluster (hence the smaller sample size), but the spatial extent of the inventory was comparable to the 1985 and 1995 surveys and spanned the entirety of Finland (Fig. S.1). The surveys included observations of a total of 380 understory plant species. To ensure that meaningful values of species-specific model parameters could be estimated, we focused only on those species which occurred in at least five percent of sites in each inventory year reducing the total number of species to 39 (a common practice in high-dimensional JSDM settings, see Clark et al., 2017).

In all three surveys, vascular plant species were identified and each species' percent cover (0.1 to 100 percent) was visually estimated within four permanent 2 m^2 quadrats located 5 m apart within each site. There was no requirement that the percent cover values of all species sum to 100 percent given that a quadrat may not be fully occupied or may have a total cover greater than 100 percent if plants over top one another (a common occurrence in understory communities). To facilitate modeling, percent cover values were treated as numeric counts (percent cover values were rounded up to the nearest integer) representing the relative abundance of each species with the total abundance given by the sum across all species. When modeling understory communities, we calculated the average percent cover

across the four sampled quadrats before converting percentages to estimates of relative and total abundance.

A number of environmental variables characterizing local conditions were measured along with understory community composition. The basal area of the trees at each site was estimated using measurements of stem diameter at breast height (Tomppo et al., 2011) and provides information on past forest management given basal area is inversely related to harvest intensity (Fig. S.2). Shrub cover at each site was quantified as the projected percent cover of shrubs and 0.5-1.5 m tall trees within a 9.8 m radius circular plot centered on the permanent vegetation survey site. Soil fertility at each site was determined in the field using six ordinal classes based on vegetation (Cajander, 1949; Tomppo et al., 2011). For the purposes of this study, we re-classified these into two groups representing "high" and "low" fertility. Finally, growing degree days were estimated as the average annual sum of daily mean temperatures exceeding +5°C per site over the decade preceding the inventory year based on 10 km^2-resolution interpolated daily temperature values modeled by the Finnish Meteorological Institute (Venäläinen et al., 2005).

Methods

Modeling approach

We defined a multinomial data model within a common hierarchical Bayesian JSDM framework (Clark et al., 2017; Ovaskainen et al., 2017; Poggiato et al., 2021) to accommodate the relative abundance understory community data. Under the multinomial model, a species' relative abundance was estimated as a function of its expected success relative to the expected success of all species given local conditions. Specifically, we defined the relative abundance of each species in a local community as $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})$, an m-dimensional vector where m is the total number of species and i indexes sites ($i = 1, 2, \ldots, n$). The relative abundance of each species in a local community was modeled as,

$$\mathbf{y}_i | \mathbf{z}_i \sim \text{Multinom}(y_{i\cdot}, \boldsymbol{\pi}_i) \tag{1}$$

where $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ defines an unobserved (latent) continuous variable, $y_{i\cdot} = \sum_{j=1}^{m} y_{ij}$ where j indexes species ($j = 1, 2, \ldots, m$), and $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{im})$ where π_{ij} is the expected probability of success of species j at site i. The expected probability of success was modeled applying a softmax function of $\mathbf{z}_i$,

$$\pi_{ij} = \frac{\exp z_{ij}}{\sum_{j=1}^{m} \exp z_{ij}} \tag{2}$$

ensuring each $\pi_{ij} \in (0, 1)$ and $\sum_{j=1}^{m} \pi_{ij} = 1$. The z_{ij} variables were modeled as,

$$z_{ij} = \mathbf{x}_i' \boldsymbol{\beta}_j + w_{ij} \tag{3}$$

where $\mathbf{x}_i$ is a vector of site-level variables characterizing the local environment, $\boldsymbol{\beta}_j$ is a vector of species-specific responses to each site-level variable, and w_{ij} is an additive species random effect accounting for variation not attributable to observed site conditions (reflecting random variation in species abundance or unobserved site conditions). A zero-centered Gaussian

distribution was applied to model species random effects either independently or dependently (jointly) among species depending on the model (see Alternative models). Additional details on the JSDM structure including differences between independent and dependent random effect terms is provided in the Supporting information.

Similar versions of the above multinomial model have been developed and applied to model microbiome data as well as plant compositional cover data (Xia et al., 2013; Damgaard, 2015). The structure of the multinomial model has a strong connection to community assembly processes. The expected probability of success of each species (π_{ij}) can be interpreted as the species' relative fitness under the conditions at site i. The total abundance among all species ($y_{i\cdot}$) can be interpreted as the community size under current growing conditions. Species dependence arises naturally under this model given the sum-to-one constraint on the species' probabilities of success ($\sum_{j=1}^{m} \pi_{ij} = 1$). This dependence is reflected in the multinomial correlation between two species defined as:

$$\rho_i(j, j') = \frac{-\pi_{ij}\pi_{ij'}}{\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ij'}(1-\pi_{ij'})}}. \tag{4}$$

Note the multionmial correlation is modeled as a function of each species probability of success at a local site as determined by Eqtn. (2). As such, $\rho_i(j, j') \in (-1, 0)$ with the maximum correlation strength (≈ -1.0) occurring only when two species have probabilities near 0.5 and all other species have probabilities near zero (i.e., richness ≈ 2.0).

The above multinomial correlation is narrow in scope accounting only for the non-independence of species' relative abundances: as the relative abundance of one species increases, the total relative abundance of all other species must decrease ($\pi_{ij} = 1 - \sum_{j' \neq j} \pi_{ij'}$) for a fixed community size. Despite this negative dependence, measurable multinomial correlation between two species does not necessarily indicate a competitive interaction. We refer to the dependence captured by the multinomial correlation as "compositional dependence" hereafter underscoring the structure of understory community observations. Additional species dependence including both positive and negative associations can be incorporated into the multinomial model through dependent species random effects (see Alternative Models and technical details in Supporting information) as is done in existing JSDM frameworks (Clark et al., 2017; Ovaskainen et al., 2017; Poggiato et al., 2021). We refer to additional species dependence arising through the random effect term as "residual dependence" as it captures non-independence among species after controlling for the effects of the environment.

We implemented our model framework within the Hierarchical Model of Species Communities (HMSC) `R` package (Ovaskainen et al., 2017; Tikhonov et al., 2020, 2021). The HMSC package has a similar model structure to the one presented above, but does not include a multinomial data model (Eqtn. 1). Instead, we applied the well-established Poisson approximation to the multinomial to induce a multinomial likelihood for the linear predictor ($\mathbf{z}_i$), which uniquely defines species-specific probabilities (McCullagh and Nelder, 1989; Baker, 1994). Details on the Poisson approximation to the multinomial and its implementation within HMSC are provided in the Supporting information.

Alternative models

We formalized hypotheses about the processes underlying community assembly as eight alternative models representing stochastic processes, environmental filtering, and species dependence in different forms (Table 1). The models applied alternative site-level variables and species random effect structures in combination with two alternative data models: the multinomial model defined in Eqtn. (1); and a log-normal Poisson model, which assumed species' relative abundances are independent conditional on $\mathbf{z}_i$,

$$\mathbf{y}_i|\mathbf{z}_i \sim \prod_{j=1}^{m} \text{Pois}(\exp z_{ij}). \tag{5}$$

As described in the data section, several site-level variables were observed along with species abundances describing local growing conditions: forest density as expressed by basal area per hectare (BA), growing degree days (GDD), soil fertility (SoilFert), and the percentage of shrub cover (ShrubCover). All models included an intercept term within $\mathbf{x}_i$ used to estimate the mean abundance of each species across all sites. Models excluding environmental filtering as a community assembly process included the intercept term alone (stochastic models, Table 1). Models designed to test the hypothesis that environmental filtering is determined by past management alone incorporated the BA variable to represent time since harvest (management models, Table 1). Models designed to test the hypothesis that environmental filtering is determined by past management and local environmental conditions included the BA variable along with GDD, SoilFert, and ShrubCover (environment models, Table 1). We applied formal Bayesian model selection to identify the combination of management and environmental variables (including interactions terms) generating the best out-of-sample prediction (as described in the following subsection). The final set of environmental filtering variables included the following terms: Intercept, SoilFert, BA, GDD, ShrubCover, SoilFert $\times$ BA, SoilFert $\times$ GDD, BA $\times$ GDD, SoilFert $\times$ BA $\times$ GDD.

Models designed to test the hypothesis that species dependence plays a role in community assembly applied the multinomial data model (Eqtn. 1; community and communityPlus models, Table 1). Models excluding species dependence as a community assembly process applied the log-normal Poisson data model (Eqtn. 5; independent models, Table 1). All but two models (communityPlus-management, communityPlus-environment) incorporating species dependence applied independent species random effects such that species dependence arises only through the multinomial data model capturing compositional dependence. In the communityPlus models (Table 1), we applied both the multinomial data model accounting for compositional dependence and dependent species random effects accounting for residual dependence.

Model selection

We applied Bayesian model selection methods to assess the eight alternative models based on their ability to predict out-of-sample data. These methods aim to identify the model best predicting out-of-sample understory communities while accounting for uncertainty in true, underlying community assembly processes, uncertainty in model parameters, and variability

in community data (Gelman et al., 2014; Hooten and Hobbs, 2015). Although not a formal hypothesis test, the predictive performance of each model provides a measure of its transferability to new communities used to assess the level of evidence for a model given observed data (Gelfand et al., 1992; Hooten and Hobbs, 2015).

Out-of-sample predictive performance was quantified using the log posterior predictive density score (hereafter log score; Gelman et al., 2014; Hooten and Hobbs, 2015). The log score reflects the statistical utility of a model to predict the complete distribution of out-of-sample data taking into account uncertainty in model parameters. We estimated the joint log score for a subset of the 1985 and 1995 datasets including 1,918 out-of-sample natural communities by fitting each alternative model twice: once using the 2006 inventory data alone to predict 435 communities within the 1995 data ($1995|2006$), and once using the 1995 and 2006 inventory data combined to predict 1,483 communities within the 1985 data ($1985|1995, 2006$). The statistical basis for our out-of-sample predictive approach is provided in the Supporting information.

We further assessed the ability of a subset of four models representing stochastic processes (independent-stochastic), environmental filtering (independent-environment), species dependence (community-stochastic), and both environmental filtering *and* species dependence (community-environment) to predict individual species and communities across successional and bioclimatic gradients. Specifically, we partitioned the 1985 and 1995 inventory data based on basal area and regional climatic conditions. The basal area variable was divided into three ranges to represent management history: 0-10 m^2/ha, 10-30 m^2/ha, and > 30 m^2/ha. The low basal area range is indicative of understory communities early in ecological succession, while the high basal area range is indicative of late-successional communities given its connection to forest management (Fig. S.2). We utilized three bioclimatic zones spanning the latitudinal gradient of Finland (south, mid, and north boreal) to represent regional climatic conditions (Table S.1, Fig. S.1; Ahti et al., 1968). This resulted in eight data partitions with a minimum of 54 communities (mid-boreal, high basal area) and a maximum of 565 communities (south-boreal, mid basal area; Table S.2). Within each partition, we estimated the mean log score (across sites) under the four models for each species and the overall community.

Results

Joint log scores for each alternative model are provided in Fig. 1. Based on the log score results, the community-environment model including both environmental filtering and species dependence had the highest joint predictive performance. The top three performing models all included environmental filtering using the complete set of environmental predictors (environment models). The next best three models included environmental filtering as measured by forest density alone (management models). The poorest performing models were those without any environmental filtering included (stochastic models). The top-performing community-environment model included compositional dependence alone, while the second-best performing independent-environment model did not include any species dependence. The communityPlus-environment model with both compositional and residual dependence

had the poorest performance of the three environment models. In-sample and out-of-sample posterior predictive model checks under the best predicting community-environment model revealed high correlation with observed data and species-level R-squared values exceeding 0.60 for the majority of species across all inventory years (Table S.4, Fig. S.3; see Supporting information for model checking details).

Mean log score values (across all out-of-sample sites) for community-level predictions under the four selected models representing different assembly processes (independent-stochastic, independent-environment, community-stochastic, community-environment) are presented for each of the eight data partitions in Fig. 2. Similar to the joint log score results in Fig. 1, the community-environment model had the highest predictive performance across all data partitions. The independent-environment model had the second-best predictive performance followed by the community-stochastic model and, finally, the independent-stochastic model in all partitions. The differences between the predictive performance of the community-environment and independent-environment models were small with 95 percent confidence intervals for the two models overlapping in all partitions. There was clear separation of the two environment models relative to the two stochastic models, but the 95 percent confidence intervals for all four models overlapped in half of the data partitions (Fig. 2).

Mean log score results for species-level predictions under the four selected models are presented for each data partition in Fig. 3. The community-environment model had the highest predictive performance for the majority of species within every data partition. The independent-environment model was the best predicting model for the second-highest number of species in each partition followed by the community-stochastic model (Fig. 3).

Posterior predictions of understory community structure under the community-environment model across a range of basal areas, growing degree days, and soil fertilities are provided in Figs. 4 and 5. Predictions of the relative abundances of species are shown in Fig. 4, while Fig. 5 provides predictions of pairwise multinomial correlations among species. The highest diversity value of 12.1 (95 percent credible interval [CI]: 3.7-19.7), calculated using true Shannon diversity (Jost, 2006), was predicted to occur under low basal area, high growing degree days, and high soil fertility (Fig. 4). The lowest diversity value of 3.5 (95 percent CI: 1.6-7.4) was predicted to occur under high basal area, low growing degree days, and low soil fertility. Independent of growing degree days and soil fertility, the community-environment model consistently predicted a decrease in community diversity moving from low to high basal areas representative of ecological succession (Fig. 4). Two dwarf shrub species, *Vaccinium myrtillus* and *V. vitis-idaea*, were predicted to have the highest relative abundances across all modeled conditions except under low basal area (Fig. 4).

The above dynamics are further reflected in the predicted pairwise multinomial correlations among species (Fig. 5). Under high basal area, the community-environment model predicted measurable correlations ($|\rho(j, j')| > 0.01$) mainly with the two dominant dwarf shrubs. The maximum correlation was predicted to occur between the two dwarf shrubs with lesser correlations predicted for secondary species. The community-environment model predicted the largest number of measurable correlations, 0.35 of all species pairings (95 percent CI: 0.13-0.52), at low basal area, high growing degree days, and high soil fertility sites consistent with these communities having the greatest diversity (Fig. 5).

Discussion

We found clear evidence that both environmental filtering and species dependence play a strong role in structuring boreal forest understory communities in Finland based on the predictive performance of alternative models across 1,918 out-of-sample communities. The importance of environmental filtering was clearly demonstrated by the fact that the top six predicting models all incorporated measures of environmental filtering with those including the full set of environmental variables (environment models) outperforming those including only forest density (management models; Fig. 1).

The role of species dependence was more nuanced in the model results. Species dependence was included in the alternative models in three ways: (1) no dependence (independent models), (2) compositional dependence alone (community models), (3) compositional and residual dependence allowing for both positive and negative associations (communityPlus models). The community-environment model, which incorporated both the full set of environmental predictors and compositional dependence alone, had the best predictive performance as measured by its joint log score (Fig. 1). The ordering of the environment models in terms of their predictive performance was community-environment > independent-environment > communityPlus-environment (Fig. 1).

Interestingly, the relative importance of alternative species dependence structures changed when forest density alone was used to represent the environmental filter (management models). Among the management models, there was statistical support for both compositional and residual species dependence as evidenced by the ordering of the models: communityPlus-management > community-management > independent-management (Fig. 1). This ordering is what we would expect to find based on ecological theory. That is, community assembly driven by a complex combination of abiotic and biotic filters with the latter including a range of potential interactions among species (Vellend, 2016).

While the ordering of the management models matches ecological expectations, when compared to the ordering among the environment models it hints at a broader and known limitation of JSDMs: the confounding of unobserved environmental effects and species dependence (Poggiato et al., 2021). Specifically, there was strong evidence for the inclusion of residual species dependence among the management models that included only a reduced set of environmental variables (Fig. 1). When the full set of environmental variables was included, however, there was near equal predictive ability of the independent-environment and communityPlus-environment models despite the former assuming no species dependence and the latter including both compositional and residual species dependence. This result suggests the variability accounted for by the residual species dependence in the communityPlus-management model was captured by the fixed environmental effects in the three environment models.

The fact that the independent-environment model yielded roughly equivalent community predictions as the communityPlus-environment model does not mean that species do not interact in complex ways in boreal forest understory communities; rather, it indicates that residual dependence among species in modeled communities is limited after controlling for the strong effects of the environmental filter. Statistically, the residual species dependence

captured by dependent species random effects in the communityPlus-environment model led to greater variation in posterior predictions of understory community composition without a strong improvement in the accuracy of the predictions. From an ecological perspective, modeling complex dynamics through the mean (the fixed effects of the models herein) rather than through residual covariance (the species random effects) is preferred given that it is often easier to estimate dependence in the mean than in the residuals (Wikle, 2003) and the ecological interpretation of such dependence follows directly from the formulation of the mean (Clark et al., 2010).

Importantly, we still found evidence that species dependence plays a role in structuring boreal forest understory communities even after including the full set of environmental variables (the community-environment model had the highest overall performance). Species dependence, however, was limited to the compositional dependence captured by the multinomial correlation structure (Eqtn. 4). This does not necessarily mean that only negative associations exist among species in boreal forest understory communities. Rather, the result suggests that after controlling for the strong role of environmental filtering, compositional dependence alone best approximates the biotic filter acting upon boreal forest understory communities among the models compared.

There are a number of reasons to expect strong compositional dependence in the modeled boreal forest understory communities. First, such dependence reflects the nature of the observations. The understory communities were measured using fixed area, 2 m^2 quadrats, with high relative abundance of one species resulting in less available growing space and other resources for other species in the community. This type of resource limitation is observed regularly in forest ecosystems and is the basis for common conceptual models of forest development driven by density-dependent competition (e.g., Reineke's self-thinning law; Reineke, 1933). Although compositional dependence does not necessarily indicate resource competition among species, model predictions combined with previous experimental results suggest such a connection may exist in the modeled boreal forest understory communities.

The community-environment model predicted that *Vaccinium myrtillus* and *V. vitis-idaea* will dominate late-successional understory communities regardless of local environmental conditions (Figs. 4, 5). Previous experimental studies have found that these dwarf shrubs compete for nitrogen and growing space in the boreal understory (Gundale et al., 2012). Greater relative abundances of other species were predicted to occur in early-successional communities (with low forest densities) and in late-successional communities under the warmest and most fertile sites (Fig. 4) consistent with the hypothesis that productive sites support more species (Heikkinen and Mäkipää, 2010). The negative multinomial correlations predicted between less abundant species and the two dominant dwarf shrubs highlight the strong compositional dependence in these communities. Specifically, high diversity communities are characterized by a large number of low-strength species' correlations while low diversity communities are characterized by few, large-strength correlations among the most dominant species (Fig. 5). While we cannot infer species interactions from compositional dependence alone, the increase in strength of the negative correlations among subordinate species and the two dominant dwarf shrubs as succession advances reflects the fact that for a lower abundance species to increase its relative abundance, there must be a corresponding

decrease in the relative abundance of the dominant species in the community.

In addition to the complete holdout data, the community-environment model also had the best performance in terms of predicting communities and individual species across data partitions reflecting successional and bioclimatic gradients (Figs. 2, 3). We expected alternative processes to play varying roles in community assembly conditional on the successional stage and relative harshness of the regional climate (Jabot et al., 2020). Contrary to our expectations, however, a consistent ordering of models emerged across all data partitions: community-environment > independent-environment > community-stochastic > independent-stochastic (Fig. 2; note the relative differences among models varies by partition). The strong predictive performance of both the community-environment and independent-environment models across a range of successional stages and bioclimatic conditions at both the community and individual species scales underscores the importance of environmental filtering in structuring modeled boreal forest understory communities. The small differences in the predictive performance of the community- and independent-environment models indicate that the compositional dependence discussed above plays only a small role in structuring the modeled communities relative to the local environmental conditions. While the relative importance of environmental filtering appears to be greater than compositional dependence in the modeled communities, there was evidence that *both* play a role in understory community assembly.

As the above discussion demonstrates, the model framework applied here allows for testing of hypotheses regarding the processes underlying community assembly. Although the alternative models developed and tested in the current analysis are phenomenological, they are process-based in the sense that they represent different combinations of ecological processes contributing to observed patterns in natural communities (Table 1). The model framework can be applied to any natural community data where the relative abundances of species are observed. Given its implementation within the more general HMSC framework, the model can deal with a large number of species and locations and allows additional information on functional traits and phylogeny to be incorporated as predictors of environmental responses (Ovaskainen et al., 2017). As such, the model framework itself, as much as the specific results described herein, constitutes an important step forward in our ability to understand and predict community assembly.

Acknowledgements

This study was supported by funding from the Jane & Aatos Erkko Foundation, Academy of Finland (grant number: 317255), and the Finnish Cultural Foundation. We thank colleagues at the Research Center for Ecological Change for their thoughtful comments and feedback on the study design and framework. We further thank colleagues at Natural Resources Institute Finland for their assistance in preparing and interpreting understory survey data.

References

Ahti, T., Hämet-Ahti, L., and Jalas, J. (1968). Vegetation zones and their sections in northwestern europe. *Annales Botanici Fennici*, 5(3):169–211.

Baker, S. G. (1994). The multinomial-poisson transformation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(4):495–504.

Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., et al. (2019). The geography of biodiversity change in marine and terrestrial assemblages. *Science*, 366(6463):339–345.

Cajander, A. K. (1949). Forest types and their significance. *Acta For. Fenn.*, 56:1–71.

Chase, J. M. and Myers, J. A. (2011). Disentangling the importance of ecological niches from stochastic processes across scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1576):2351–2363.

Chauvet, M., Kunstler, G., Roy, J., and Morin, X. (2017). Using a forest dynamics model to link community assembly processes and traits structure. *Functional Ecology*, 31(7):1452–1461.

Clark, J. S. (2016). Why species tell more about traits than traits about species: predictive analysis. *Ecology*, 97(8):1979–1993.

Clark, J. S., Bell, D., Chu, C., Courbaud, B., Dietze, M., Hersh, M., et al. (2010). High-dimensional coexistence based on individual variation: a synthesis of evidence. *Ecological Monographs*, 80(4):569–608.

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., and Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, 87(1):34–56.

D'Amen, M., Rahbek, C., Zimmermann, N. E., and Guisan, A. (2017). Spatial predictions at the community level: from current approaches to future frameworks. *Biological Reviews*, 92(1):169–187.

Damgaard, C. (2015). Modelling pin-point cover data of complementary vegetation classes. *Ecological Informatics*, 30:179–184.

Falster, D. S., Brännström, Å., Westoby, M., and Dieckmann, U. (2017). Multitrait successional forest dynamics enable diverse competitive coexistence. *Proceedings of the National Academy of Sciences*, 114(13):E2719–E2728.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Stanford Univ CA Dept of Statistics.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016.

Götzenberger, L., de Bello, F., Bråthen, K. A., Davison, J., Dubuis, A., Guisan, A., et al. (2012). Ecological assembly rules in plant communities—approaches, patterns and prospects. *Biological Reviews*, 87(1):111–127.

Gundale, M. J., Hyodo, F., Nilsson, M.-C., and Wardle, D. A. (2012). Nitrogen niches revealed through species and functional group removal in a boreal shrub community. *Ecology*, 93(7):1695–1706.

Heikkinen, J. and Mäkipää, R. (2010). Testing hypotheses on shape and distribution of ecological response curves. *Ecological Modelling*, 221(3):388–399.

Hooten, M. B. and Hobbs, N. T. (2015). A guide to bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28.

Jabot, F., Laroche, F., Massol, F., Arthaud, F., Crabot, J., Dubart, M., et al. (2020). Assessing metacommunity processes through signatures in spatiotemporal turnover of community composition. *Ecology Letters*, 23(9):1330–1339.

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

Keddy, P. A. (1992). Assembly and response rules: two goals for predictive community ecology. *Journal of Vegetation Science*, 3(2):157–164.

Kraft, N. J., Adler, P. B., Godoy, O., James, E. C., Fuller, S., and Levine, J. M. (2015). Community assembly, coexistence and the environmental filtering metaphor. *Functional Ecology*, 29(5):592–599.

Lorenz, M. and Fischer, R. (2013). Pan-European forest monitoring: an overview. *Developments in Environmental Science*, 12:19–32.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

McGill, B. J., Enquist, B. J., Weiher, E., and Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, 21(4):178–185.

Mittelbach, G. G. and Schemske, D. W. (2015). Ecological and evolutionary perspectives on community assembly. *Trends in Ecology & Evolution*, 30(5):241–247.

Münkemüller, T., Gallien, L., Pollock, L. J., Barros, C., Carboni, M., Chalmandrier, L., et al. (2020). Dos and don'ts when inferring assembly rules from diversity patterns. *Global Ecology and Biogeography*, 29(7):1212–1229.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., et al. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., and Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. *Trends in Ecology & Evolution*.

Reineke, L. H. (1933). Perfecting a stand-density index for even-aged forest. *Journal of Agricultural Research*, 46:627–638.

Reinikainen, A., Mäkipää, R., Vanha-Majamaa, I., and Hotanen, J.-P., editors (2000). *Kasvit muuttuvassa metsäluonnossa [Summary in English: changes in the frequency and abundance of forest and mire plants in Finland since 1950]*. Kustannusosakeyhtiö Tammi.

Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K., et al. (2013). A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology Letters*, 16:94–105.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J., et al. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution*, 11(3):442–447.

Tikhonov, G., Ovaskainen, O., Oksanen, J., de Jonge, M., Opedal, O., and Dallas, T. (2021). *Hmsc: Hierarchical Model of Species Communities*. R package version 3.0-11.

Tomppo, E., Heikkinen, J., Henttonen, H. M., Ihalainen, A., Katila, M., Mäkelä, H., et al. (2011). *Designing and conducting a forest inventory-case: 9th National Forest Inventory of Finland*, volume 22. Springer Science & Business Media.

Tylianakis, J. M., Didham, R. K., Bascompte, J., and Wardle, D. A. (2008). Global change and species interactions in terrestrial ecosystems. *Ecology Letters*, 11(12):1351–1363.

Vellend, M. (2016). *The Theory of Ecological Communities (MPB-57)*. Princeton University Press.

Venäläinen, A., Tuomenvirta, H., Pirinen, P., and Drebs, A. (2005). A basic Finnish climate data set 1961–2000–description and illustrations. *Finnish Meteorological Institute, Reports*, 5:1–27.

Weiher, E. and Keddy, P. A. (1995). Assembly rules, null models, and trait dispersion: new questions from old patterns. *Oikos*, pages 159–164.

Wikle, C. K. (2003). Hierarchical bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., et al. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, 33(10):790–802.

Table 1: Summary of alternative models representing different processes underlying community assembly. Variable abbreviations are defined as follows: BA = basal area per hectare, SoilFert = soil fertility, GDD = growing degree days, ShrubCover = percent shrub cover.

Model	Assembly processes represented	Environmental variables ($\mathbf{x}_i$)	Species dependence	Data model
Independent-Stochastic	Stochastic	None	None	Log-normal Poisson
Independent-Management	Environmental filtering	BA	None	Log-normal Poisson
Independent Environment	Environmental filtering	BA, SoilFert, GDD, ShrubCover	None	Log-normal Poisson
Community-Stochastic	Species dependence	None	Compositional	Multinomial
Community-Management	Environmental filtering and species dependence	BA	Compositional	Multinomial
Community-Environment	Environmental filtering and species dependence	BA, SoilFert, GDD, ShrubCover	Compositional	Multinomial
CommunityPlus-Management	Environmental filtering and species dependence	BA	Compositional + residual	Multinomial
CommunityPlus-Environment	Environmental filtering and species dependence	BA, SoilFert, GDD, ShrubCover	Compositional + residual	Multinomial

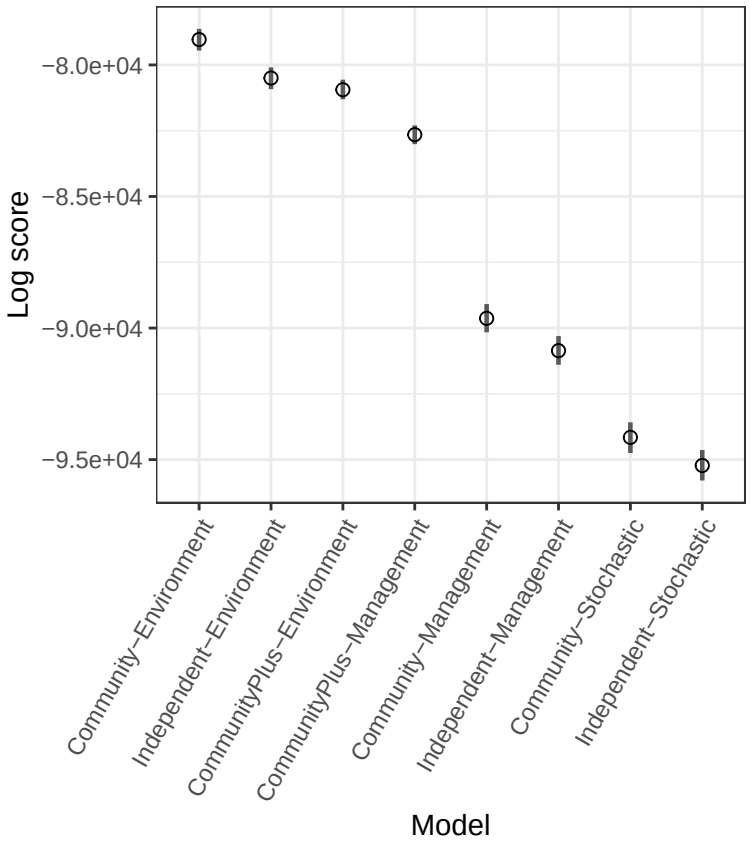

Figure 1: Joint log scores for alternative models (as calculated by Eqtns. S.4 or S.5). Points represent bootstrapped median values, while lines indicate bootstrapped 95 percent confidence intervals. Models are ordered from highest to lowest log score with higher values indicating better out-of-sample prediction. Models names are defined in Table 1.

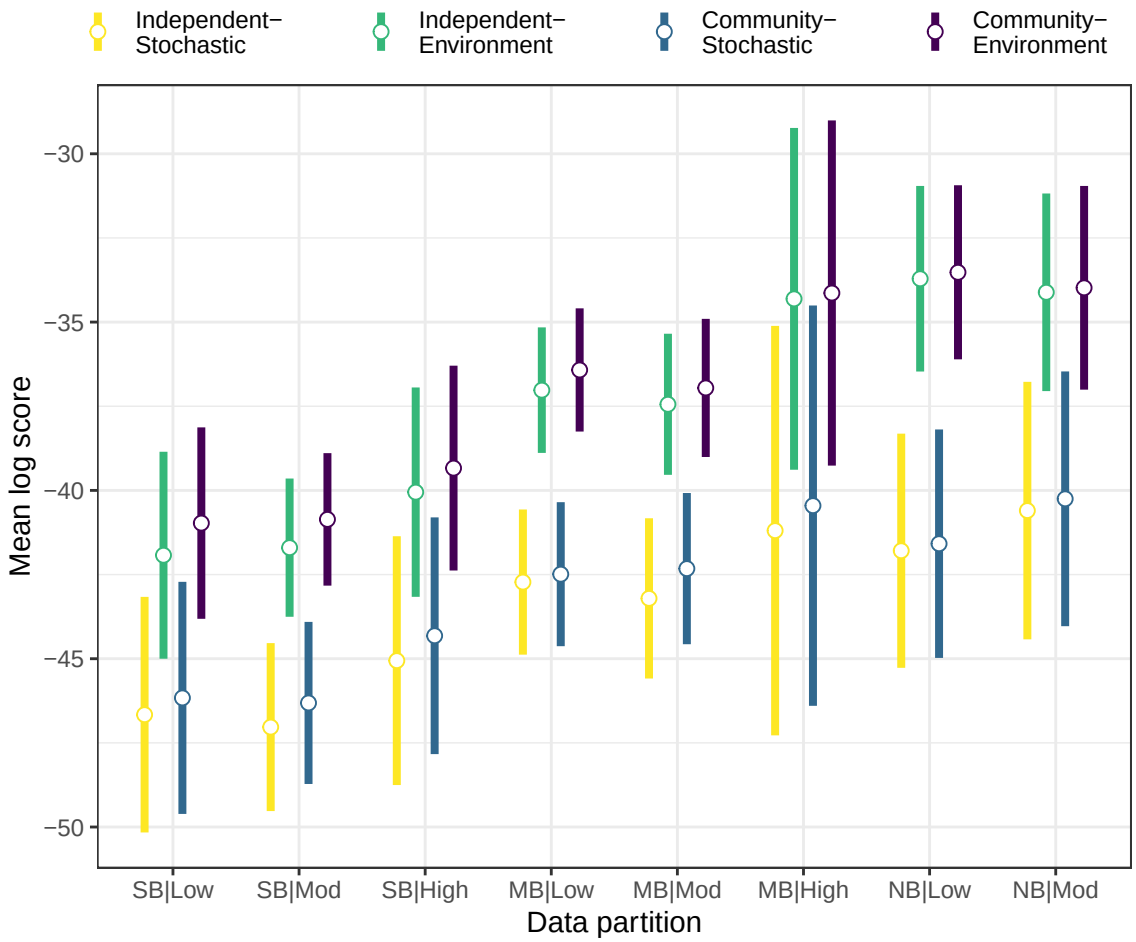

Figure 2: Mean, per-site log scores for community-level prediction under the four alternative models applied within each data partition (as calculated by Eqtn. S.6). Points represent the mean across all sites within a partition, while lines indicate 95 percent confidence intervals. Data partition labels are defined as: bioclimatic zone|basal area (BA). Bioclimatic zones: SB = south boreal, MB = mid boreal, NB = north boreal; BA levels ($m^2 \cdot ha^{-1}$): Low = [0-10), Mod = [10-30), High = $\geqslant$ 30.

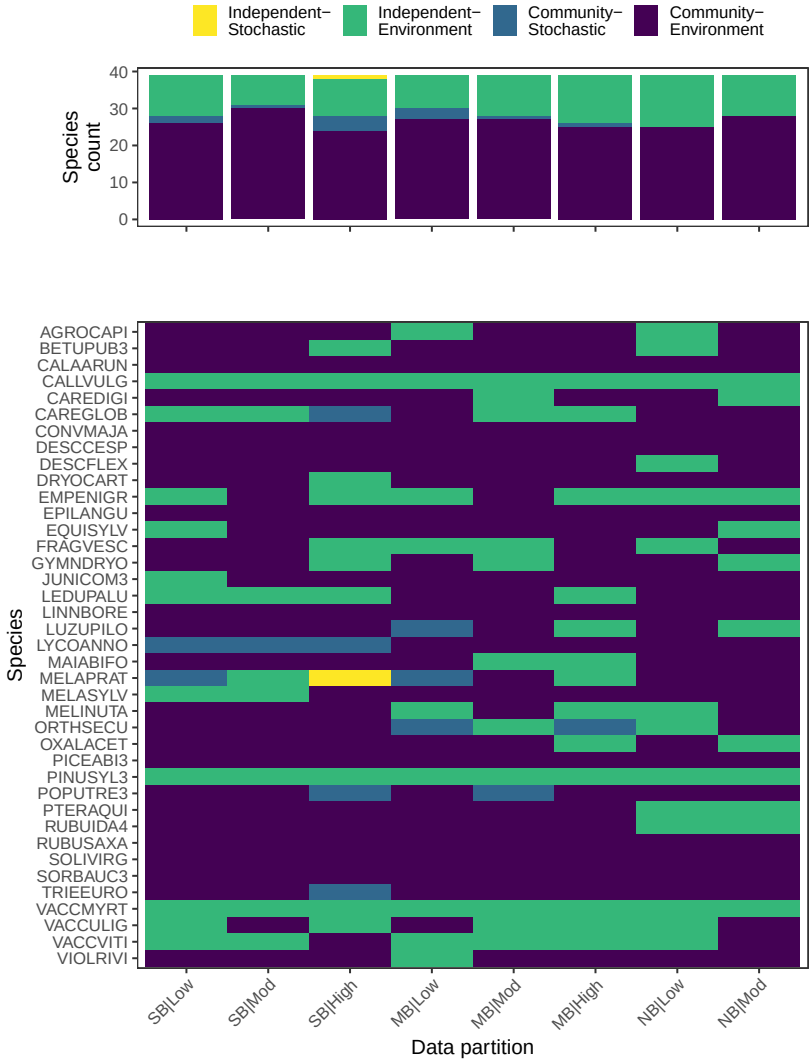

Figure 3: Summary of mean, per-site log scores for species-level prediction under the four alternative models applied within each data partition (as calculated by Eqtn. S.7) including: the number of species for which each model had the highest mean log score (upper panel) and the model with the highest mean log score for each species-by-partition combination (lower panel). Data partition labels are defined as: bioclimatic zone|basal area (BA). Bioclimatic zones: SB = south boreal, MB = mid boreal, NB = north boreal; BA levels ($m^2 \cdot ha^{-1}$): Low = [0-10), Mod = [10-30), High = $\geqslant$ 30. Species codes are defined in Table S.3.

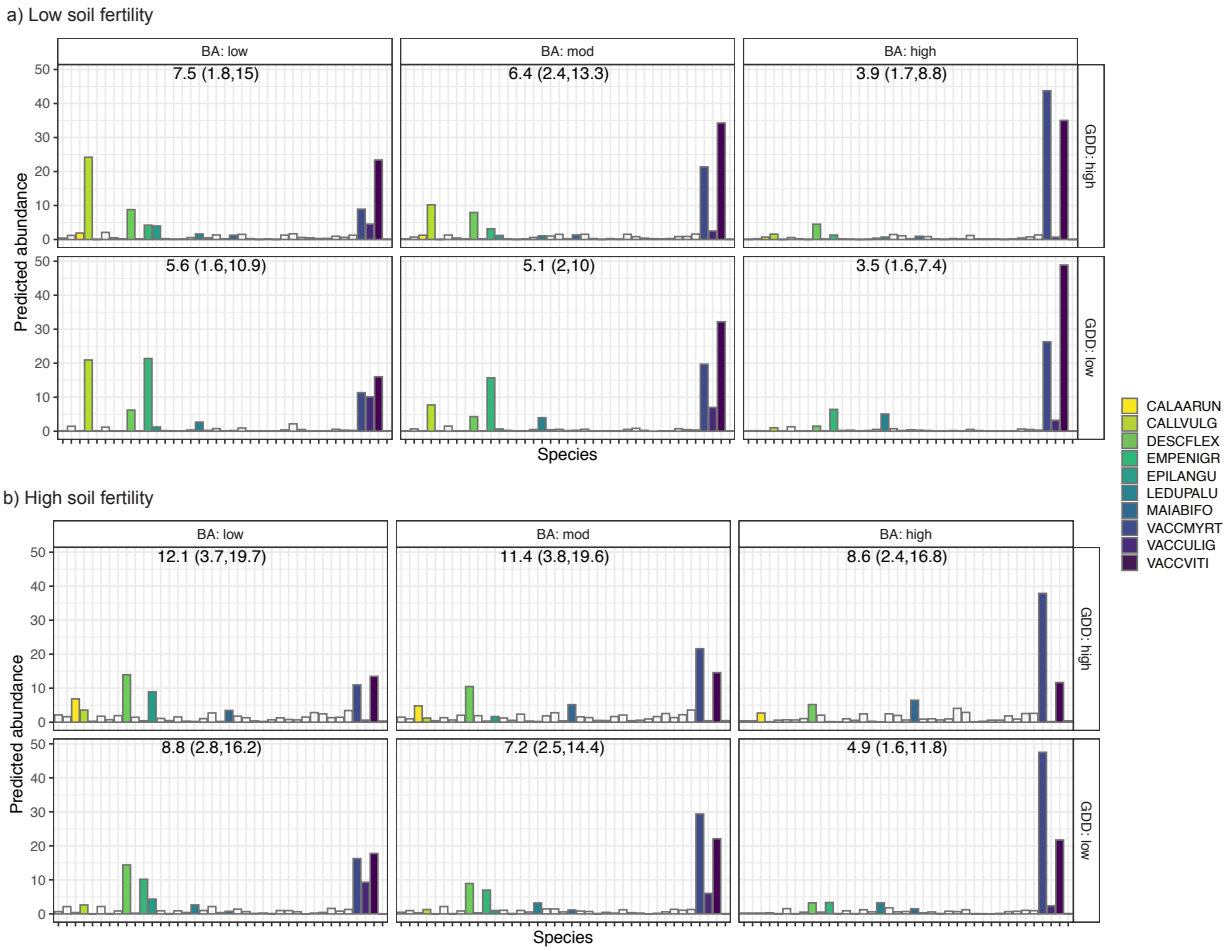

Figure 4: Predicted natural communities applying the community-environment model across a range of basal areas (BA), growing degree days (GDD), and low (a) and high (b) soil fertility. Bars indicate the posterior median predicted abundance of each species within a community of 100 individuals. The ten species with the largest overall abundance are color-coded for reference (all other species are shown in white). Text within panels indicates the posterior median predicted true Shannon diversity (Jost, 2006) and its corresponding 95 percent credible interval (in parentheses). BA ($m^2 \cdot ha^{-1}$): low = 0, mod = 15, high = 35; GDD: low = 900, high = 1200. Species codes are defined in Table S.3.

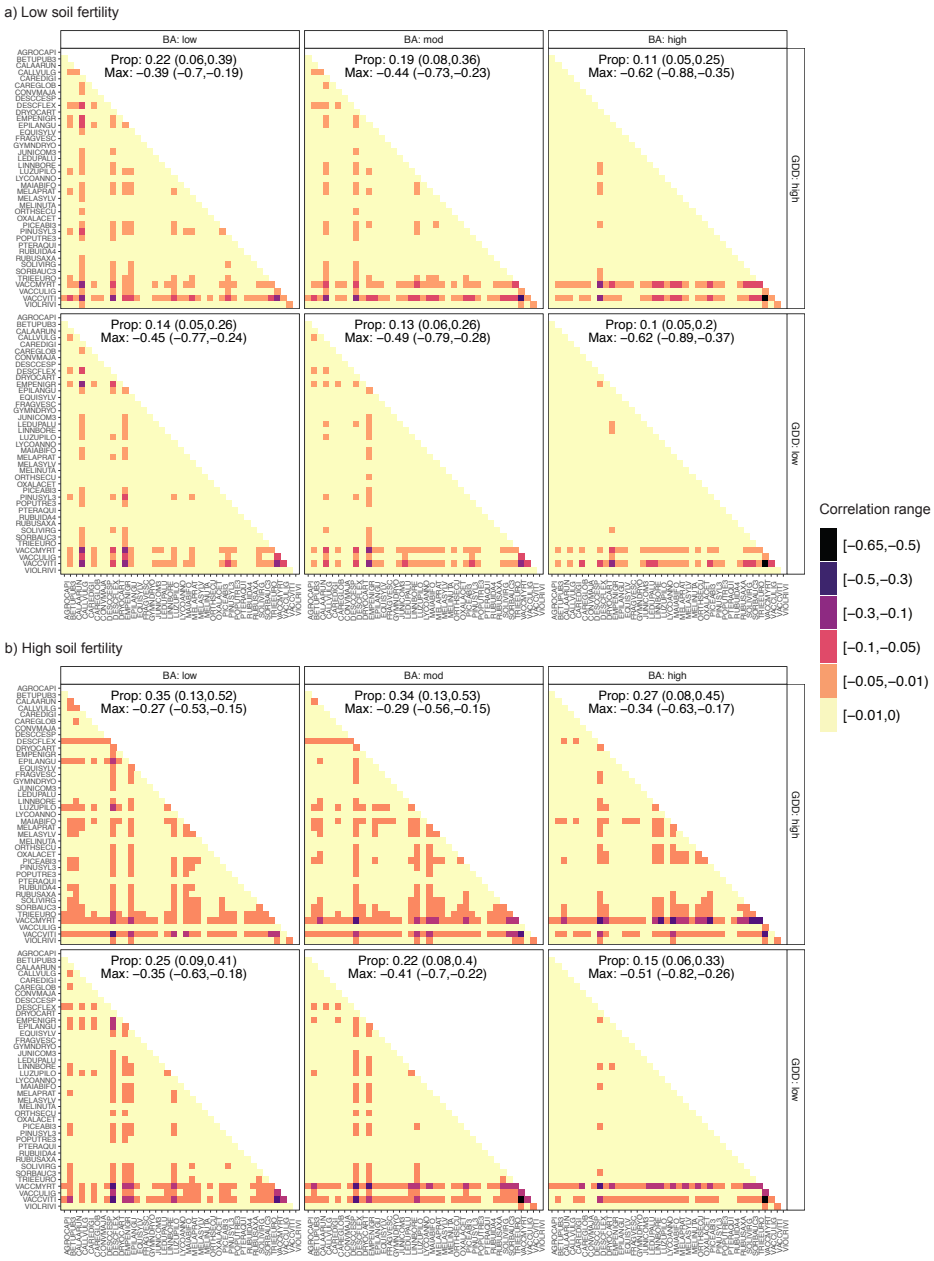

Figure 5: Predicted species correlations applying the community-environment model across a range of basal areas (BA), growing degree days (GDD), and low (a) and high (b) soil fertility. Values indicate the posterior median predicted correlation of each species pair. Text indicates the posterior median proportion of all species combinations (741 total) that had non-zero correlation: $|\rho(j, j')| > 0.01$ (Prop.), and the posterior median maximum strength of correlation (Max), with corresponding 95 percent credible intervals (in parentheses). BA ($m^2 \cdot ha^{-1}$): low = 0, mod = 15, high = 35; GDD: low = 900, high = 1200. Species codes are defined in Table S.3.