The Internal Structure of Metacommunities

Authors and Affiliations:

Mathew A. Leibold, Department of Biology, University of Florida. 32611

mleibold@ufl.edu

Javiera Rudolph, Department of Biology, University of Florida.

javiera.av@ufl.edu

F. Guillaume Blanchet, Département de Biologie, Université de Sherbrooke, 2500 Boulevard Université, Sherbrooke, Québec, J1K 2R1, Canada.

guillaume.blanchet@usherbrooke.ca

Luc De Meester, Laboratory of Aquatic Ecology, Evolution and Conservation, University of Leuven, 3000 Leuven, Belgium; Leibniz Institut für Gewässerökologie und Binnenfischerei (IGB), 12587 Berlin, Germany; Institute of Biology, Freie Universität Berlin, Königin-Luise-Strasse 1-3, 14195 Berlin, Germany; Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), Königin-Luise-Str. 2-4, 14195 Berlin, Germany

Luc.DeMeester@kuleuven.be

Dominique Gravel, Département de Biologie, Université de Sherbrooke, 2500 Boulevard Université, Sherbrooke, Québec, J1K 2R1, Canada.

Dominique.Gravel@USherbrooke.ca

Florian Hartig, Theoretical Ecology, University of Regensburg, Germany,
florian.hartig@ur.de

Pedro Peres-Neto, Department of Biology, Concordia University.
pedro.peres-neto@concordia.ca

Lauren Shoemaker, Botany Department, University of Wyoming.
lshoema1@uwyo.edu

Jonathan M. Chase, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig; Department of Computer Sciences, Martin Luther University Halle-Wittenberg;
jonathan.chase@idiv.de

Running title: Internal Metacommunity Structure

Keywords: Interspecific competition, site dynamic theory, hierarchical community models, metacommunity paradigms, variation partitioning.

Type of Article: Concepts and Synthesis

Number of words in Abstract: 261
Number of words in text (excluding abstract, acknowledgements, tables and figures, and references): 4697
Number of references: 37

Number of figures: 6, Number of tables: 1, Number of boxes: 1

Corresponding author:

Mathew A. Leibold

Department of Biology

University of Florida

220 Bartram Hall

P.O. Box 118525

Gainesville, FL 32611-8525

mleibold@ufl.edu

+1 (512) 636-9824

Statement of authorship: MAL and JMC conceived the overall idea for the paper. JMC obtained funding for the workshop that produced this publication. DG developed the model and JR, DG, LS and FGB produced the simulations and the analysis of the results. All authors developed the ideas and findings of the paper and wrote the paper jointly.

Data accessibility statement: Should this manuscript be accepted, the data and code supporting the results will be archived in an appropriate public repository such as Github.

ABSTRACT:

Metacommunity ecology has become an important subdiscipline of ecology, but it is increasingly evident that its foundational theoretical and analytical frameworks do not adequately incorporate a realistic continuum of environmental and biotic process at play. We propose an approach that develops stronger links between theoretical and statistical frameworks to shift the focus towards the study of the 'internal structure' of metacommunities by dissecting how different species and different sites contribute to overall metacommunity structure. To illustrate this, we simulate data from a model that includes environmental variation, dispersal, biotic interactions, and stochasticity as the basic ecological processes that influence species' (co)-distributions. We analyze the simulated data with hierarchical community models and propose a new method to visualize and analyze the simultaneous role of species co-distribution, environment, space, and stochasticity in this emerging statistical approach. We focus in particular on quantifying how species affect the overall structure of the metacommunity via differences in their dispersal and niche traits, and how environmental filtering, dispersal and species interactions varies from site to site in relation to environmental conditions and connectivity. Although there are still challenges ahead, this framework provides a roadmap for a more comprehensive approach by jointly developing more mechanistic theory based on community assembly processes and the analytical tools needed to map these concepts onto data in diverse landscapes and systems.

Introduction

The assembly of communities and their dynamics are driven by processes that act at both local and regional scales (e.g. Levin 1974, Ricklefs 1987, Lawton 1999, Leibold et al. 2004). The local-regional connection embodied in the metacommunity framework, which examines the role of local (i.e. patch) dynamics and their connectivity at the regional (i.e. metacommunity) scale and, by doing so, has provided important advances to our understanding of the coexistence, diversity and distribution of species in space and time (see Leibold et al. 2004, Logue et al 2011, Brown et al. 2107, Leibold and Chase 2017, Ovaskainen et al. 2019). Nevertheless, our understanding of mechanisms governing these dynamics in diverse systems and in complex landscapes, and our ability to infer them from field surveys, is still poorly developed. In particular, it remains difficult to connect abundance patterns in patches and metacommunities to underlying assembly mechanisms (Leibold and Chase 2018, Ovaskainen et al. 2019).

The classical conceptual framework of metacommunity ecology is based on four idealized models (the so-called 'archetypes') of metacommunity dynamics (e.g., species sorting, neutral, patch dynamics and mass effects; Leibold et al. 2004). Until now, these have mostly been studied by quantifying how the community composition varies, for example in relation to spatial or environmental factors (Cottenie 2005, Peres-Neto et al. 2006), and attributing a given system to one of the archetypes based on this analysis. Although these idealizations served as useful benchmarks to classifying metacommunities according to major drivers such as environmental heterogeneity and dispersal (e.g. Leibold 2011, Logue et al. 2011, Lindstrom and Langenheder 2011), it is increasingly recognized that this

overlooks many important features of metacommunities (e.g. Brown et al. 2017, Leibold and Chase 2018, Ovaskainen et al. 2019, Thompson et al. 2020). Consequently, it has been difficult to expand these models and the associated theoretical framework to highly diverse communities, while still maintaining their general insights and intuitive appeal (e.g. Fournier et al. 2017, Ovaskainen et al. 2019).

Drawing from parallels with evolutionary biology, Vellend (2010, 2016) proposed an alternative, more process-oriented theoretical framework for community ecology. In this view, the assembly of species is the result of the interaction between four main processes: a) dispersal (analogous to gene flow in evolution), b) speciation (and other ways species arrive into the regional species pool; analogous to mutations), c) demographic stochasticity (ecological drift, analogous to genetic drift), and d) species interactions and responses to environmental factors (analogous to selection, including both frequency-dependent and frequency independent components). If we adopt this view, and consider that speciation is beyond the scope of regional metacommunity dynamics, our aim is thus to explain the distribution of species across sites in a metacommunity by ecological drift, dispersal, environmental filtering (i.e. responses to 'abiotic' factors) and selection linked to species interactions (Fig. 1). This links the basic processes identified by Vellend (2010, 2016) to the classical concepts of community assembly via accessibility (dispersal), abiotic and biotic ‘filters’ (Soberon et al. 2005). The additional advantage is that all these components can be estimated with joint species distribution models (JSDMs), a relatively new and promising class of statistical community models.

140 The categorical approach based on archetypes makes the implicit assumption that there is
141 a dominant assembly process that applies to most species across most local sites in a
142 metacommunity (Brown et al. 2017, Leibold and Chase 2018). However, it ignores the way
143 that different species may experience a given metacommunity in different ways, for
144 example because they differ in dispersal abilities or by having distinct responses to habitat
145 heterogeneity. It also ignores that different sites within a metacommunity may have
146 species compositions that are determined by distinct processes. For example, sites that are
147 very isolated may be occupied by species that occupy neighboring sites due to dispersal
148 limitation, regardless of how well suited they are to local environmental conditions
149 whereas sites that are very closely connected in space may be more susceptible to source-
150 sink relations with neighboring sites.
151
152 Furthermore, previous work has not adequately accounted for the role of species
153 interactions that can act to influence the distribution of species in metacommunities. For
154 example, variation partitioning (Borcard et al. 1992, Cottenie 2005) only addresses effects
155 of environmental and spatial factors but does not account for any additional possible
156 effects of species interactions (Logue et al. 2011). The roles of species interactions other
157 than interspecific competition has also been largely neglected (Guzman et al. 2019, Gravel
158 and Massol 2020).
159
160 In this article, we explain how these shortcomings might be addressed in a more
161 comprehensive framework that builds on Vellends (2010) approach and can better account
162 for species interactions (see also Ovaskainen et al. 2019). We especially focus on

accounting for heterogeneities among species and localities by dissecting overall metacommunity structure into species-specific and patch-specific components which we call the 'internal structure' of the metacommunity. We also focus more directly on key processes that drive community assembly dynamics in metacommunities (e.g., stochastic drift, dispersal, selection, heterogeneity, Vellend 2010, 2016), recognizing that a continuum of outcomes can emerge depending on the relative strengths of those processes (see also Gravel et al. 2006, Leibold and McPeek 2006, Vellend 2010, 2016, Boulangeat et al. 2012, Munkenmuller et al. 2012, Fournier et al. 2017, Wiegand et al. 2017, Ovaskainen et al. 2019, Rybicki et al. 2018, Thompson et al. 2019, Shoemaker et al. 2020, Gravel and Massol 2020). We illustrate this with a flexible theoretical model that captures the key elements and processes of metacommunity dynamics, including species and patch-level variation. Although we limit ourselves to interspecific interactions via competition, we recognize that ongoing and future work will address the full scope of species interactions, including facilitation and cross-trophic level interactions.

Quantifying the link between process and pattern using a simple metacommunity simulation and refined statistical approach:

To test and exemplify the analytical power of our framework, we simulate data from a process-based model, which allows us to create observations with complete knowledge about the underlying mechanisms. Our process-based model is based on a spatial implementation of early spatially implicit site occupancy models (e.g., Levins and Culver 1971, Horn and MacArthur 1972, Levin 1974, Hastings 1980, Hanski 1991) to describe

dynamics in heterogeneous, multispecies metacommunities. For each species in each patch, we model occupancy using two key equations (details in the SI). The first of these describes the colonization of patch z by species i during a discrete time interval, Δt:

$$P\left(X_{i,z,t+\Delta t} = 1 | X_{i,z,t} = 0\right) = I_{i,z,t} S_{i,z,t} C_{i,z,t} \quad (1)$$

where $X_{i,z,t}$ is a stochastic variable representing the occurrence of species i at location z at time t, $I_{i,z,t}$ is the number of immigrants of species, $S_{i,z,t}$ is the effect of environmental filtering on the probability of establishing a viable local population, and $C_{i,z,t}$ is the effect of ecological interactions on the establishment probability. Second, we consider the alternative possibility: the extinction of species i in patch z during the time interval Δt:

$$P\left(X_{i,z,t+\Delta t} = 1 | X_{i,z,t} = 0\right) = M_{i,z,t} E_{i,z,t} \quad (2)$$

where $M_{i,z,t}$ and $E_{i,z,t}$ are the responses of the extinction probability to the local environment and to ecological interactions, respectively.

At steady state the solution to this model is:

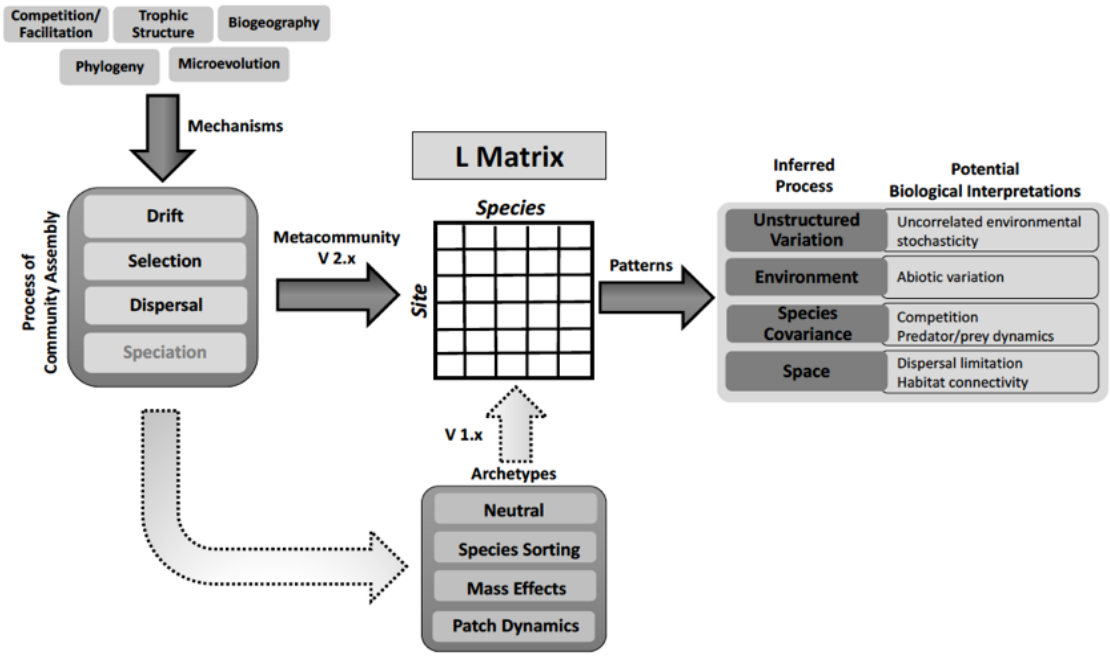
$$\log \frac{\hat{p}^{i,z}}{1-\hat{p}^{i,z}} = \log I^{i,z} + \log \frac{S^{i,z}}{M^{i,z}} + \log \frac{C^{i,z}}{E^{i,z}} \quad (3)$$

where $\hat{P}^{iz}$ is the expected probability that site *z* is occupied by species *i.*

This formulation supposes that immigration($I^{i,z}$), 'abiotic selection' ($S^{i,z}$ *and* $M^{i,z}$) and interactions ($C^{i,z}$ *and* $E^{i,z}$) can be separated into distinct effects that do not interact. This may not apply in all cases in nature if, for instance the strength of species interactions or dispersal could vary with the environment. Equation 3 naturally follows a log-linear structure, suggesting that an appropriately structured JSDM can be used to separate the contributions of these effects into spatial effects (driven by immigration), environmental filtering (driven by abiotic selection) and species co-distribution unrelated to either space or environment. This very general analytical model's probabilistic nature suggests that in a case with a finite number of patches there should be an additional fraction of the variation that is independent of these factors and thus stochastic within this context (Shoemaker et al. 2020). Furthermore, the likelihood of every observation can be marginalized over each species (by summing the likelihoods for a given species across all patches) to describe the variation among species. Alternatively, the likelihood can be marginalized by patch (by summing the likelihoods for a patch across all species) to describe the variation across the region. Doing so, we can quantify the importance of environment, species co-distribution, and space for predicting metacommunity structure as a whole, as well as their importance for predicting the abundance of individual species or community composition at individual patches. In this way, we can study the 'internal structure' of the metacommunity. This can

226 clarify how these key processes work together in a diverse and complex metacommunity to

227 explain species distributions within the metacommunity (Figure 1).



228

229 Figure 1: A summary of the metacommunity problem. Species distributions, denoted by the

230 species-by-sites **L** matrix , are the outcome of drift, selection, dispersal, and speciation.

231 These basic processes can be influenced by species interactions, food web structure,

232 biogeography, phylogeny and micro-evolution. Metacommunity theory mainly focuses on

233 drift, selection and dispersal. We view previous approaches based on the four archetypes

234 of Leibold et al. (2004) as being much more indirect and idealized. Instead, we call for a

235 more direct evaluation of how the basic processes affect the **L** matrix, and how to dissect

236 the consequences to the distributions of different species and the occupancy of different

237 sites, for example by using a JSDM to identify main effects and interspecific variability in

238 the importance of unstructured, biotic, environmental, and spatial effects on **L**. This

239 approach allows us to recognize and address the effects of heterogeneities among species
240 and among patches on the overall structure of the metacommunity.
241
242 To illustrate the utility of this approach in a more realistic framework, we implemented the
243 key processes of drift, environmental filtering, dispersal, and species interactions (Vellend
244 2010, 2016) in a flexible simulation version of the model (described in more detail in the
245 SI). The simulation model allows us to vary each process separately for each species in a
246 heterogeneous spatially explicit landscape. It simulates the dynamics of a metacommunity
247 across a set of patches and generates a spatial network that specifies the connectivity
248 among patches. The state variables of the simulation are the occupancy of each species in
249 every patch (i.e. presence/absence, though future implementations could also address
250 abundance data, e.g. Rybicki et al. 2018, Ovaskainen et al. 2019, Thompson et al. 2020).
251 Each patch can be colonized from nearby patches depending on their location in the
252 landscape, dispersal rate of the species and proximity of extant populations in neighboring
253 patches. Each species in each patch is subject to extinctions that reflect demographic
254 and/or environmental stochasticity. Patches can differ in local environmental conditions
255 that differentially influence baseline colonization and extinction probabilities. Species
256 interactions are modeled in two ways. First, the presence of other species in a patch can
257 modify baseline colonization probability (a reduction in the case of competition). Second,
258 co-occurring species can modify baseline extinction probability (an increase in the case of
259 competition).
260

261 We next apply a JSDM to the resulting distribution of species among patches. Specifically,
262 we use the package 'Hierarchical Modeling of Species Communities' (HMSC, Ovaskainen et
263 al. 2017, Ovaskainen and Nerea 2020) that accounts for the effect of the environment,
264 spatial autocorrelation and species co-distribution. After fitting the model using HMSC, we
265 partition variation between the model processes into four components (or fractions) using
266 an approach akin to classic variation partitioning (Borcard et al. 1992; Peres-Neto et al.
267 2006 – see Box 1 for a historical retrospective of the development of this methodology).
268 Details about how this type of variation partitioning is computed through HMSC is given in
269 the Supporting Information. Specifically, the variation was partitioned to quantify the
270 effects of environment (labeled [E]), spatial patterning (labeled [S]), co-distribution among
271 species (labeled [C]) and residual (unexplained) variation that cannot be attributed to any
272 of the three previously mentioned fractions. This is expressed as $1-R^2$, where R^2 is the
273 proportion of variation explained by the model that includes fractions [E], [S], and [C].
274
275 Compared to the classical method, this variation partitioning approach provides two major
276 advances. The first is that we can infer a third fraction that quantifies co-variation (or co-
277 distribution) among species using latent variables (fraction [C]), which is distinct from
278 either the environment effect (fraction [E|) or the spatial effects (fraction [S]). In our model
279 this fraction reflects the consequences of species association, and we emphasize that this
280 interpretation must be done with caution (see Dormann et al. 2018 and Blanchet et al.
281 2020 for critical reviews of this issue). The second advance is that all fractions are
282 calculated at the species level, where species-specific variation detail the internal structure
283 of the metacommunity from the perspective of each individual species. Fractions of the

variation at the metacommunity level are then calculated by averaging across all species to obtain metacommunity level descriptors akin to those obtained from variation partitioning. This allows us to subsequently examine potential predictors quantifying the variation in these different effects among species. In addition, we developed a way to partition each fraction on a site-by-site basis as described in the SI. The technical details for these calculations are a bit different than those used for the species-level analyses and in particular they do not allow for averaging the R^2 of the model across patches; rather we instead use the additive fraction of the models to the overall variation (see the SI for more details). This way to partition variation among sites informs us about the individual contribution each site has to variation in the data. The ability to resolve species-specific and site-specific components means that we can study the 'internal structure' of the metacommunity in ways that were absent in classic variation partitioning.

Although the analytical model described in equations 1-3 suggests that we can make a link between process and pattern using JSDMs, we wished to evaluate if this was also likely in less idealized situations such as those used in our metacommunity model. Our goal here is to illustrate how this link might be made and we leave a more extensive evaluation of these methods for future work (see also Ovaskainen and Nerea 2020). To do this, we simulated a number of scenarios that vary the strength of environmental selection, dispersal and competition. Comparing scenarios that vary niche breadth and competition (scenarios A-D, Fig. 2 and Fig 3), and a more complex case where species compete and vary in both dispersal and response to the environment (scenario G, Figure 4, scenarios E and F in the

306 SI) highlight how HMSC provides an avenue for distinguishing between underlying
307 processes based on abundances of species across metacommunity patches.
308
309 **Simulation Results**
310 In the first set of simulations, we considered a situation where species had distinct
311 environmental optima along an environmental gradient and had limited dispersal (Figure
312 2, Table 1). We contrasted the case where the environmental niches were narrow (steep
313 changes in baseline colonization success and extinction rates with small deviations in
314 environment) with the case with identical optima, but with wide environmental niches
315 (much flatter changes in colonization and extinction with environmental value). As
316 expected, we find that these differences in environmental niche breadth had strong effects
317 on the relative importance of environmental filtering (fraction [E]) versus spatial
318 patterning (fraction [S]). Specifically, we find stronger spatial patterning when niche
319 breadths were broad and stronger environmental filtering when niches were narrower.
320 We also found that the R^2 values were higher for the case with narrow niches than with
321 wide niches. Finally, we also found non-zero (though relatively weak) variance
322 components for co-distributions (fraction [C]) in both cases, especially when niches were
323 broad even though our analytical model would predict the absence of such variation
324 components since colonization and extinctions were not affected by species interactions in
325 these simulations. We discuss possible explanations for this observation below.
326
327 Table 1: Summary of metacommunity level variation components for the seven different
328 scenarios modeled in this study.

		Fractions			
Scenario	Corresponding Figure	E (SD)	S (SD)	C (SD)	Residuals $1-R^2$ (SD)
A	2, upper panels	0.75 (0.11)	0.026 (0.019)	0.044 (0.062)	0.18 (0.11)
B	2, lower panels	0.019 (0.015)	0.15 (0.037)	0.019 (0.036)	0.81 (0.04)
C	3, upper panels	0.42 (0.31)	0.21 (0.2)	0.14 (0.16)	0.23 (0.12)
D	3, lower panels	0.018 (0.026)	0.35 (0.3)	0.28 (0.33)	0.35 (0.33)
E	Supplement	0.63 (0.21)	0.067 (0.069)	0.13 (0.15)	0.18 (0.089)
F	Supplement	0.74 (0.097)	0.035 (0.025)	0.047 (0.057)	0.18 (0.097)
G	4	0.5 (0.17)	0.08 (0.072)	0.2 (0.16)	0.22 (0.12)

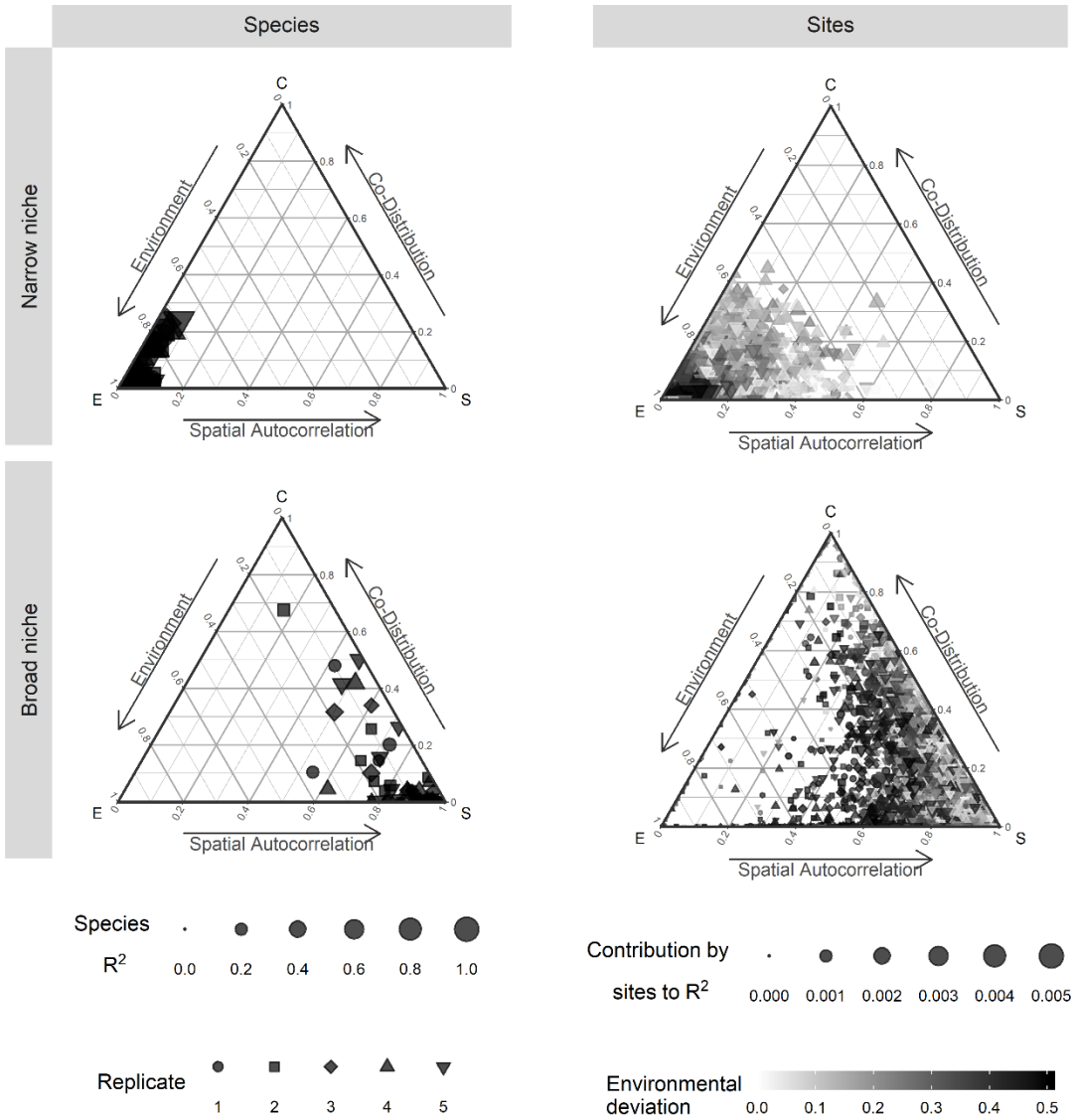


Figure 2: Ternary plots describing the three components of metacommunity internal structure for two different simulation scenarios with no species interactions (independent metapopulations): The upper panels correspond to narrow environmental niches whereas lower panels correspond to wide environmental niches. Each dot represents a species (left panels) or a site (right panels). The size of the symbol is proportional to the R^2 of the model (note the different scales used for species and sites) and the location indicates the proportion of explained variation attributed to environmental factors (*E* - lower left),

spatial effects (*S* - lower right) and remaining co-distributions (*C* - upper apex) (see SI for details). In the species panels (left side) different symbols indicate different replicate simulations; generally, these indicate that the distribution of species responses are variable within replicates but that the overall variation among replicates are repeatable. In the site panels (right side), the shading indicate how central (lighter) or extreme (darker) the local environmental conditions are on the gradient; these also show substantial variation but indicate that more extreme environmental conditions increase the effects of local environment on occupancy patterns than more central conditions.

We next simulated metacommunities with identical parameters as above, except with added interspecific competition effects (Figure 3, Table 1). As in the case without species interactions (compare with Figure 2), narrow niches enhanced the relative strength of environmental filtering (fraction [E]) and reduced spatial patterning (fraction [S]) when compared to wide niches. In these simulations, however, the co-distribution components (fraction [C]) were much more substantial than without species interactions. We also found that adding interspecific competition substantially increased the amount of variation explained (R^2) by the model.

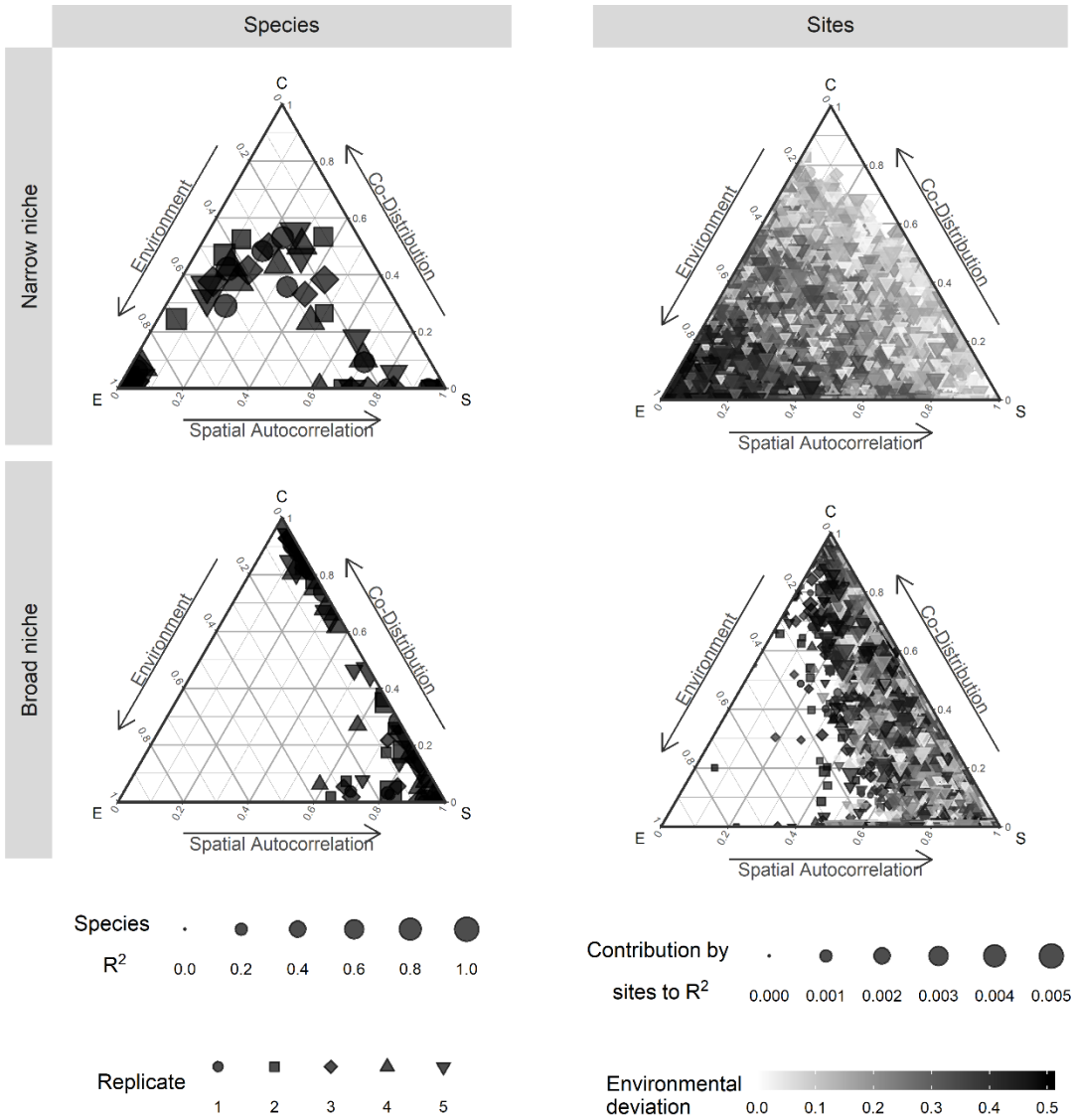


Figure 3: Ternary plots describing the three components of metacommunity internal structure for different simulation scenarios with competition among the species. Notation is the same as in Figure 2. The upper panels correspond to narrow environmental niches whereas the lower panels correspond to wide environmental niches. Left-hand panels show variation components for different species whereas panels on the right hand side of the figure correspond to variation components for different sites.

365 We conducted a number of other simulations to explore if these ternary plots (see Box 1)
366 that describe interspecific variation in the effects of dispersal, niches breadth, and
367 interactions on environment (fraction [E]), space (fraction [S]), and co-distributions
368 (fraction [C]). Illustrative examples are shown in the supplemental information and
369 summary statistics are shown in Table 1. We illustrated one of these that includes
370 heterogenous dispersal to show how the internal structure can reveal how dispersal
371 variation affects distributions (Figure 4). We found that we could distinguish species by
372 the degree to which their distributions are related to environment (fraction [E]), space
373 (fraction [S]) and co-distributions (fraction [C]) (Figure 5a). Species with higher dispersal
374 ability and more specialized environmental niche positions had distributions better
375 predicted by the environment than those that were dispersal limited and had distributions
376 that were more spatially autocorrelated (fraction [S]). Species with optima closer to the
377 middle of the environmental gradient also had a larger fraction [C] than those with more
378 extreme optima.

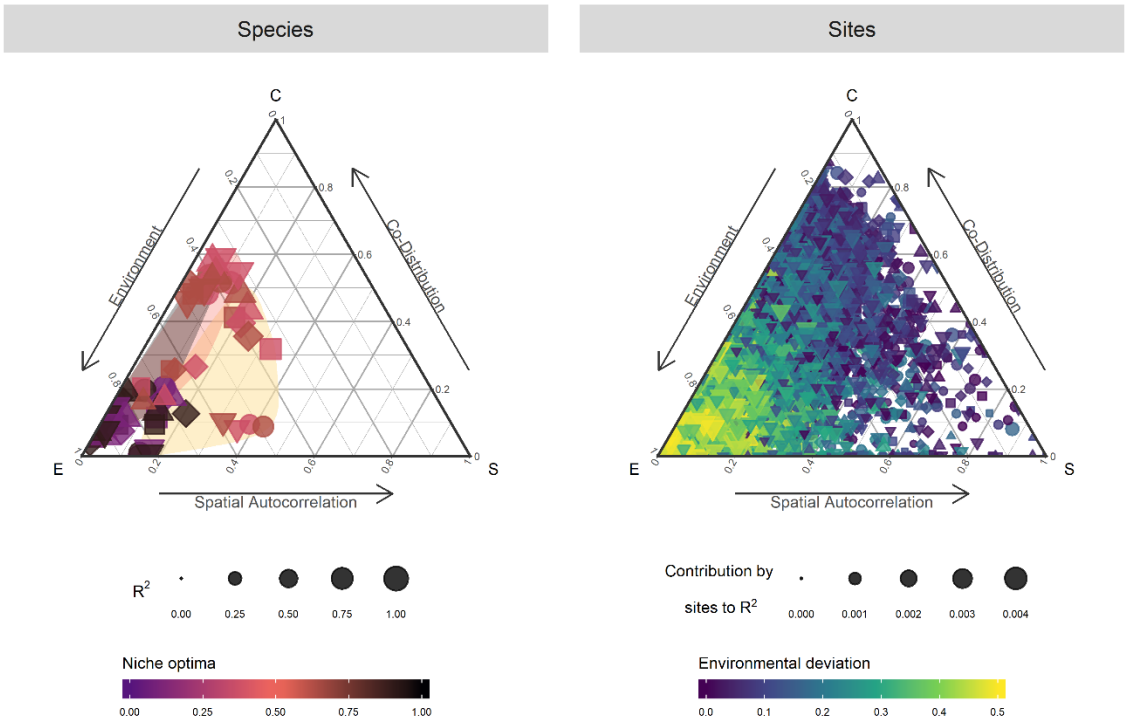


Figure 4: Ternary plots for species (left panel) and sites (right panel) for simulations with species that differ in environmental position along the gradient and dispersal ability. The size of the symbol indicates the R^2 of the model for each species or site). In the left panel (species) the color indicates the preferred local environmental conditions for species (yellow for species that prefer centrally located environmental conditions, purple or magenta for species with more extreme environmental optima). The symbol indicates the dispersal rate of the species (circles are more dispersal limited, squares are least dispersal limited and triangles are intermediate). In the right hand panel the color indicates the degree of deviation from centrality along the environmental gradient (as in Figure 2).

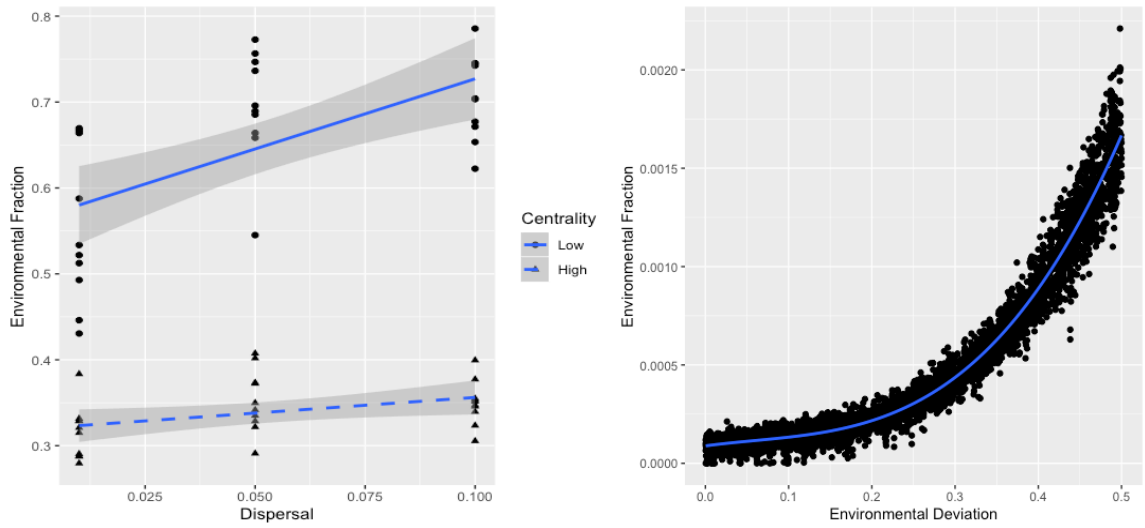


Figure 5: Effects of species traits (left panel) and site attributes (right panel) on the environmental fraction of variation in species distributions and site occupancy. A) Higher dispersal ability and lower niche centrality (i.e. greater deviation from mean niche value) enhance the degree to which different species (individual symbols) have distributions that correlate with environmental variation. B) Sites that differ more from the mean environmental value (environmental deviation) are more likely to be occupied by species with niche traits that are locally favored.

Sites also differed in how their species composition was related to environmental effects (fraction [E]), spatial effects (fraction [S]) and co-distributions (fraction [C] - Figure 4b). Some sites tended to be occupied by locally dominant species (in the lower left of the ternary plot, nearer to fraction [E]), while others were occupied by species found in nearby sites (lower right of the ternary plot, nearer to fraction [S]). Some sites were also occupied by combinations of species that were differentially associated with each other regardless of environment or dispersal (upper apex of the ternary plot, nearer to fraction [C]). As can

407 understood from figure 4b, there were also a wide range of intermediate conditions. A
408 major driver of this variation is the local environmental condition, especially in relation to
409 how distinct the local environment is from the overall mean environment of the
410 metacommunity (Figure 5b).
411
412 We further investigated the structure of the species co-distribution (fraction [C]). In our
413 analyses, this covariation can be directly attributable to species interactions due to our
414 ability to closely model the underlying processes. However, even in our model, species co-
415 distribution may not directly link to pairwise interaction coefficients, but rather may
416 emerge as a complex relationship between species interactions and environmental
417 conditions (Cazelles et al. 2015, Blanchet et al. 2020). To illustrate this, we show the co-
418 distribution among species as a heat map separately for each of the five individual
419 simulations presented in Figure 4 and compared them to the actual interaction matrix that
420 describes interspecific competition in our model (Figure 6; similar heat maps obtained
421 with the other scenarios are shown in the Supplemental Information). Despite the fact that
422 the same interaction matrix was used for all five of these simulations, the resulting co-
423 distribution patterns are inconsistent in their details. However, these matrices show that
424 there is consistency in several features of the co-distribution pattern. For example, they all
425 share the predominance of strong negative correlations along the main diagonal that match
426 the interaction matrix we used. They also share a strong 'checkerboard' pattern with
427 alternating negative and positive co-distributions between species when these are ranked
428 against their environmental optima. Given the simple scheme of species interactions we
429 used (Fig. 5 and SI), these results are consistent with the predictions that direct

interactions are stronger than indirect ones and tend to weaken with the number of links in indirect chains even if the details of these effects are less predictable.

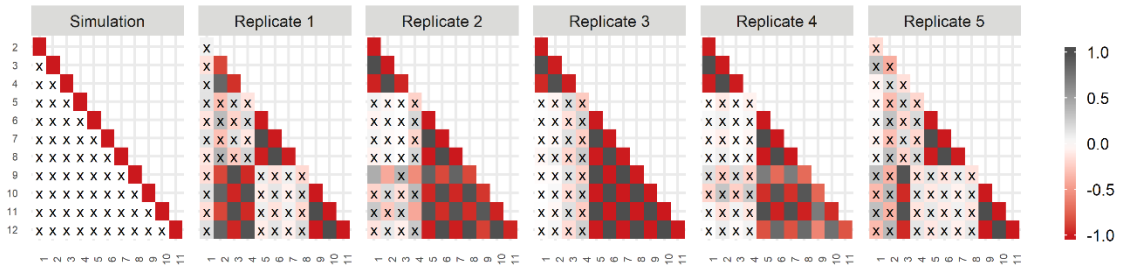


Figure 6: Comparisons of the interaction matrix (Simulation) with the co-distribution of species in five replicate runs (Iteration 1-5) of the scenario with interspecific variation in dispersal and competition among species. In each figure species are ranked by the position of their environmental optima along the environmental gradient. The co-distributions are shown as heat maps with the strength of the covariation proportional to the intensity of color and the color indicating negative (green) or positive (gray) covariation among pairs of species. These can be compared to the pattern of direct species interactions (left panel called Simulations). The Xs denote no significant association although the color indicates the trend.

Discussion

While metacommunity ecology has made great progress in the past decades, the assumptions that species and sites were relatively homogeneous in their underlying processes (reviewed in Leibold and Chase 2017) have limited the applicability of metacommunity theory to more realistic species pools within complex landscapes. Here,

450 by combining a tractable process-based model with emerging analytical methods, we have
451 provided a general approach that accounts for multiple interacting processes that may
452 operate differently among species or in different parts of landscapes.
453
454 Although there are some important challenges to consider, our study illustrates important
455 insights about the 'internal structure' of metacommunities , including:
456 - Variation partitioning using JSDMs (here implemented using HMSC) can be
457 linked to basic processes of community assembly (e.g. environmental
458 selection, dispersal, biotic selection, and drift).
459 - Quantifying co-distributions of species in metacommunities can improve
460 predictive ability even when the processes that generate these distributions
461 are complex (stochasticity, complex spatial landscapes, and species
462 interactions).
463 - Species can have distributions that are determined by combinations of the
464 basic community assembly processes depending on features of their ecology
465 (dispersal and environmental preferences)
466 - Predominant assembly processes can differ among adjacent sites in a
467 metacommunity (e.g. sites that are occupied by species most fit for
468 environmental conditions vs sites occupied by species in nearby sites that
469 may differ in environmental conditions)
470
471 Despite these advances, it is important to emphasize that there remain some substantial
472 challenges in moving forward with the overall approach we have advocated. This includes

473 some technical issues (e.g. estimation of parameters, interpretation of results in more
474 complex models in JSDMs and with real data) as well as conceptual ones (e.g. accounting
475 for other processes such as speciation, local adaptation, historical biogeography).
476 Nevertheless, we believe such an endeavor will be fruitful, allowing a deeper
477 understanding of ecological dynamics in more realistic, but necessarily complex, spatial
478 landscapes.
479
480 Our model simplifies several potentially complex processes (e.g. non-linearities and
481 interactive effects) into an approximation involving colonization-extinction dynamics. It is
482 possible that more realistic and complex mechanisms driving these processes will weaken
483 associations between pattern and process or create biases in the partitioning of the
484 variation revealed by JSDMs. However, the development of JSDMs are still progressing, and
485 we anticipate that future developments could solve some of these problems.
486
487 The co-distribution component of the JSDMs (fraction [C]) is particularly concerning. We
488 find that this component can be biased, especially when species have broad environmental
489 niches (Figure 2). It is widely recognized that there are particular challenges in quantifying
490 and interpreting species co-distributions (reviewed by Blanchet et al. 2020). Perhaps the
491 most obvious challenge is that this component can reflect <u>any</u> process that is inadequately
492 quantified by the environmental (fraction [E]) and spatial components (fraction [S]). In
493 addition to species interactions, this would include, for example, unmeasured
494 environmental factors (see Blanchet et al. 2020). Teasing apart the effects of species
495 interactions from these confounding factors should thus be a major focus for future work.

496 Nevertheless, it is important to understand that including the co-distribution component in
497 our analyses allows us to account for them, rather than lumping them with residual
498 variation where they give an exaggerated impression of stochasticity.
499
500 The basic framework we used to address the process to pattern link is most transparent for
501 mechanisms that focus on interspecific competition because of the analogies with genetic
502 change in evolutionary genetics (Vellend 2010, 2015). However, species interactions in
503 metacommunities are much more variable and include consumer-resource interactions,
504 mutualisms and facilitative interactions. Although such interactions can easily be
505 incorporated in simulations, the interpretation that might link process to pattern in such
506 cases becomes more complex. Likewise, future work could include local (co-)evolutionary
507 dynamics (see Urban et al. 2020) and historical effects of biogeography and speciation (e.g.
508 Leibold and Chase 2018, Overcast et al. 2020).
509
510 In this paper, we have also retained a simple two-level perspective on spatial scale (local
511 discrete sites in a broader regional landscape). It is increasingly apparent that
512 metacommunity dynamics occur over multiple nested scales and that habitats can be
513 continuous and/or nested, rather than discretely patchy (e.g. Munkenmuller et al. 2012,
514 Rybicki et al 2018, Ovaskainen et al. 2019, Viana and Chase 2019). Refining our approach
515 to address multiple spatial scales is a logical next step.
516
517 Here, we have argued that a focus on the internal structure of metacommunities, by
518 examining site-specific and species-specific variation components, can enhance our study

519 of the links between process and pattern in the distributions of species and the occupancies
520 of sites. We also argue that such developments are essential if metacommunity ecology is
521 to address the dynamics and structure of realistic metacommunities that have typically
522 high biodiversity and are in complex landscapes. This focus on internal structure
523 represents a shift from traditional approaches that used descriptors of overall variation
524 components at the metacommunity scale that generalizes previous ad hoc approaches to
525 similar internal variation in metacommunity patterns (e.g. Pandit et al. 2009). The
526 dynamics and structure of distributions of realistically diverse species in a realistically
527 structured landscape of sites likely involves the interaction of community assembly
528 processes including environmental filtering, dispersal, and drift, and these are unlikely to
529 be adequately described by simple metacommunity level metrics (see Ovaskainen et al.
530 2019). Consequently, dissecting the internal structure of metacommunities on the basis of
531 species and site contributions could provide key insights into the processes underlying
532 metacommunity assembly.
533

534 Box 1: Contrasting Analytical Approaches to Metacommunity Ecology:
535 a) Variation partitioning: The key elements of this approach were developed by Borcard et
536 al. (1992) and recognized that sets of communities could often show spatial patterning that
537 was independent of environmental factors in determining local species composition. Since
538 then the approach has been refined in numerous ways, but the overall idea remains the
539 same. It models overall community composition (ignoring the individual effects on
540 different species or sites) as being related to environmental predictors (labeled E in Figure
541 B1, often considered as fixed and generally suitable for abiotic factors) or to spatial

542 patterning (labeled S in the figure, often because communities that are close to each other
543 are likely to have similar species even if environments differ). There results four partitions
544 of the overall variation in composition: one that is solely predicted by the environmental
545 conditions, one that is solely predicted by spatial effect, one that is predictive but cannot be
546 resolved into either environment or spatial factors because they are confounded, and
547 residual variation that is often interpreted as 'stochastic'.
548 b) Joint Species Distribution Models used to study overall variation in community structure
549 and related methods. Here we recognize that there is an additional possible set of
550 predictors that influence co-distributions of species besides environment and space
551 (labeled I in the figure). This method recognizes that the 'residuals' of the variation
552 partitioning method in a) are not necessarily random but may show cross-correlations
553 among species. The mechanisms are vague but possibly include the influence of
554 unmeasured environmental variables or the influence of species. They thus add a third
555 major variation component to the 'variation partitioning' method described in a). To date,
556 results using these methods have mostly focused on the patterns at the metacommunity
557 scale to assess metacommunity structure (but see Ovaskainen et al. 2017).
558 c) Species-specific and Site-specific elaborations on jDSMs: Ovaskainen et al. (2017)
559 suggested that trait or phylogenetic relations among species could be used as predictors for
560 the various overall fractions described by b). Here we take this suggestion further by
561 refining the details of the method to partitioning the overall pattern in b) into contributions
562 by individual species and/or individual sites. The results can be visualized using ternary
563 plots (with terminal axes that correspond to E, S and I) that display the relative
564 contributions of environmental, spatial, and residual covariation among species and among

sites. We further visualize the overall joint predictive effects of all three of these predictors as proportional to the size of symbol (i.e. small symbols show species who's distribution if more stochastic than large symbols). We then examine how these components of variation are correlated with species traits (in relation to species-specific variation) or with site attributes (in relation to site-specific variation). Only this last approach reveals the internal structure of the metacommunity and tries to identify the key species- and site-attributes that might drive these effects.

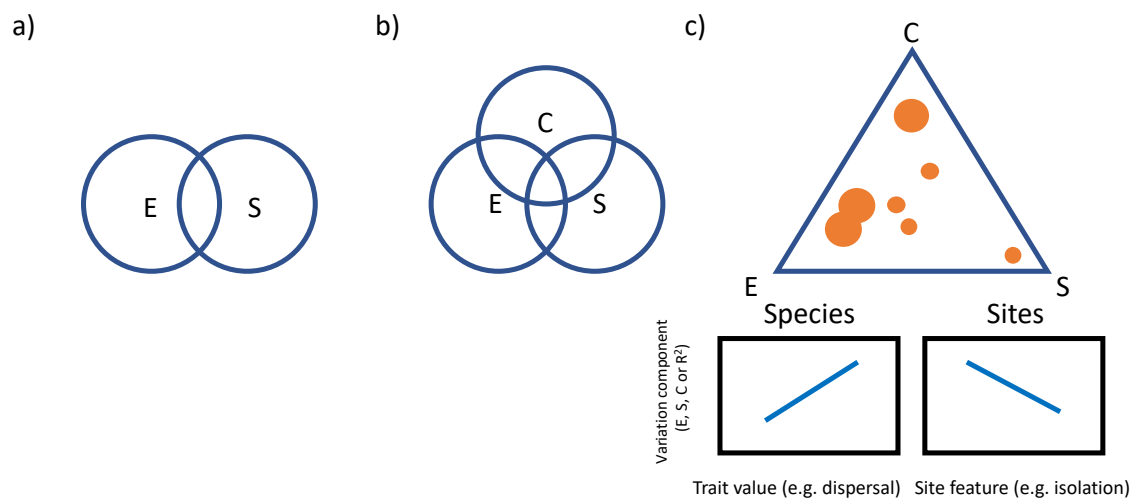


Figure B1: Highly simplified diagramatic representation of variation partitioning methods in metacommunity ecology: a) variation partititioning (Borcard et al. 1992), total variation in community variability is partitioned in relation to environment (E) and space (S). b) joint-species distribution models (Ovaskainen et al. 2017), latents are used to further quantify variation due to co-distribution (C) patterns unrelated to E or S. c) internal structure of metacommunities (current contribution), the elements of E, S and C as well as the overall explanatory ability to predict distributions of individual species or sites is dissected (upper panel) and can be related to features of species and sites (lower panels).

Acknowledgments: This paper emerged from workshops funded with the support (to JMC) from the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig funded by the German Research Foundation (FZT 118). MAL would like to acknowledge the Alexander von Humboldt Foundation for a Research Award that helped fund this work. LDM acknowledges KU Leuven Research Fund project C/2017/002 and FWO projects G0B9818 and G0C3818. We also thanks F Engel, C. Rakowski, K. Taylor, R. Pelinson, X. Zhao, L. Juen, T. Michelan and A. Rudolf for comments on earlier versions of the manuscript.

References:

Blanchet, F.G., Cazelles, K., Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23, 1050–1063

Borcard, D., Legendre, P. & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, 73, 1045–1055.

Boulangeat, I., Gravel, D., Thuiller, W. (2012). Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters,* 15, 584-593.

Brown, B.L., Sokol, E.R., Skelton, J. & Tornwall, B. (2017). Making sense of metacommunities: dispelling the mythology of a metacommunity typology. *Oecologia*, 183, 643–652.

Cottenie, K. (2005). Integrating environmental and spatial processes in ecological community dynamics: Meta-analysis of metacommunities. *Ecology Letters*, 8, 1175–1182.

Dormann, C.F., Bobrowski, M., Dehling, M. et al. (2018) Biotic interactions in species distribution modelling: ten questions to guide interpretation and avoid false conclusions. *Global Ecology and Biogeography*. 27.9:1004-1016

Fournier, B., Mouquet, N., Leibold, M.A. & Gravel, D. (2017). An integrative framework of coexistence mechanisms in competitive metacommunities. *Ecography*, 40, 630–641.

Gravel, D., Canham, C.D., Beaudet, M. & Messier, C. (2006). Reconciling niche and neutrality: the continuum hypothesis. *Ecology Letters*, 9, 399–409.

Gravel, D., & Massol, F. (2020). Toward a general theory of metacommunity ecology. *In*: McCann, K.S. & Gellner, G. (eds.). Theoretical Ecology. Oxford University Press. *In press.*

Guzman, L.M., Germain, R.M., Forbes, C., Straus, S., O'Connor, M. I., Gravel, D. Srivastava, D.S., and Thompson, P.L. (2019). Towards a multi-trophic extension of metacommunity ecology. *Ecology Letters,* 22, 19-33.

Hastings, A. (1980). Disturbance, coexistence, history, and competition for space. *Theoretical Population Biology*, 18, 363–373.

Horn, H.S. & MacArthur, R.H. (1972). Competition among fugitive species in a harlequin environment. *Ecology*, 53, 749–752.

Lawton, J.H. (1999). Are there general laws in ecology? *Oikos*, 84, 177.

Leibold, M.A. & Chase J.M. (2018). *Metacommunity ecology*. Princeton University Press, Princeton.

Leibold, M.A., Economo, E.P. & Peres-Neto, P. (2010). Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. *Ecology Letters*, 13, 1290–1299.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., *et al.* (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, 7, 601–613.

Leibold, M.A., and McPeek, M.A., (2006). Coexistence of the niche and neutral perspectives in community ecology. *Ecology*, 87, 1399-1410.

Levin, S.A. (1974). Dispersion and population interactions. *American Naturalist*, 108, 207-228.

Levins, R., and Culver, D., (1971). Regional coexistence of species and competition between rare species. *Proceedings of the National Academy of Sciences, U.S.A.*, 68, 1246-1248.

Lindstrom, E.S. and Langenheder, S., (2012). Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports, 4, 1-9.*

Logue, J.B., Mouquet, N., Peter, H. & Hillebrand, H. (2011). Empirical approaches to metacommunities: a review and comparison with theory. *Trends in Ecology & Evolution*, 26, 482–491.

T. Münkemüller, T., de Bello, F., Meynard, C. N., Gravel, D,, Lavergne, S., Mouillot, D., Mouquet, N., and Thuiller, W. (2012). From diversity indices to community assembly processes: a test with simulated data. *Ecography*, 35, 468–480.

Ovaskainen, O. and Abrego, N., (2020). Joint Species Distribution Modelling With Applications in R. Cambridge University Press.

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, 561–576.

Ovaskainen, O., Rybicki, J., and Abrego, N. (2019). What can observational data reveal about metacommunity processes? *Ecography*, 42, 177-1886.

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation partitioning of species data matrices: Estimation and comparison of fractions. *Ecology*, 87, 2614–2625.

Peres-Neto, P.R., Leibold, M.A. & Dray, S. (2012). Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. *Ecology*, 93, S14–S30.

Ricklefs, R.E. (1987). Community diversity: Relative roles of local and regional processes. *Science*, 235, 167–171.

Rybicki, J., Abrego, N. and Ovaskainen, O., (2019). Habitat fragmentation and species diversity in competitive communities. *Ecology Letters*, 23, 506-517.

Shoemaker, L.G., Sullivan, L.L., Donohue, I., Cabral, J.S., Williams, R.J. Mayfield M.M., Chase, J.M., Hcu, C. Harpole, W.S., Huth, A., HilleRisLambers, J., James, A.R.M., Kraft, N.J.B., May, F., Muthukrishnan, R., Satterlee, S., Taubert, F., Wang, X., Wiegand, T., Yang,Q., and Abbott, K.S. (2020). Integrating the underlying structure of stochasticity into community ecology. *Ecology*, 101, e02922.

Soberon, J. and Peterson, A.T. (2005). Interpretation of models of fundamental ecological niches and species distributional areas. *Biodiversity Informatics* 2, 1-10.

669 Thompson, P.L., Guzman, L.M., De Meester, L., Horvath, Ptacnik, R., Vanschoenwinkle, B.,
670 Viana, D.S., and Chase, J.M. (2019). A process based framework for metacommunity
671 ecology. bioRxiv doi.org/10.1101/832170
672 Urban, M.C., Strauss, S.Y., Pelletier, F., Palkovacs, E.P., Leibold M.A., Hendry A.P., De
673 Meester, L., Carlson, S.M., Angert, A.L. and Giery, S.T. (2020). Evolutionary origins for
674 ecological patterns. *Proceedings of the National Academy of Sciences, U.S.A.*, in press.
675 Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly Review of*
676 *Biology*, 85, 183–206.
677 Vellend, M. (2016). The theory of ecological communities. Princeton University Press.
678 Viana, D.S. and Chase, J.M. (2019). Spatial scale modulates the inference of metacommunity
679 assembly processes. *Ecology,* 100, e02576.
680 Wiegand, T., Uriarte, M., Kraft, N.J.B., Shen, G., Wang, X., and He, F. (2017). Spatially explicit
681 metrics of species diversity, functional diversity, and phylogenetic diversity:
682 Insights into plant community assembly processes. Annual Review of Ecology,
683 Evolution, and Systematics 48, 329–351.
684
685
686

The Internal Structure of Metacommunities: Supplementary Information

Mathew A. Leibold, Javiera Rudolph, Pedro Peres-Neto, Dominique Gravel, Luc De Meester, Lauren Showmaker, Florian Hartig, F. Guillaume Blanchet, Jonathan M. Chase

Table of Contents

Description of Model 1

Description of Statistical Framework 4

Parameterization and Simulation Scenarios 8

Example of metacommunity simulation functions and analyses 10

Description of Model

Patch locations and environmental variables

In our model, the metacommunity consists of N patches distributed over a spatially heterogeneous landscape, with multiple environmental variables(although, the current simulations only have one environmental variable D and 1000 patches) that could either be randomly distributed or spatially autocorrelated. Each patch has a set of coordinates in a two-dimensional space, and all possible coordinates are feasible such that this is a continuous space model that is not restricted to a lattice or some other kind of regular spatial arrangement of spatial units.

A patch may be empty, or be occupied by a single or by several species. We define $X_{i,z,t}$ as a stochastic variable representing the occurrence of species i at location z and time t. Occurrence, $X_{i,z,t}$, takes a value of 1 when species i is present and a value of 0 when it is absent. Similarly, we define $\mathbf{Y}_{z,t} = X_{1,z,t}, X_{2,z,t}, \dots, X_{R,z,t}$ as a vector containing the presence-absence of each species from the regional pool R.
The model only tracks the occupancy of the patches (not population densities). Spatial dynamics occurs because of colonization events, in both empty patches and patches that

are occupied by other species, and because of extinction events. The emerging species co-distributions are a result of a dynamic balance between these events. Ecological interactions can impact either or both the colonization and the extinction probabilities. For instance, the presence of a competitor pre-empting a patch can reduce the colonization probability by another competitor. Alternatively, the presence of a competitor in a patch could increase the extinction probability of another species. Similarly, the environment could influence both the colonization and the extinction probabilities.

Patch Colonization

We consider a discrete-time Markovian process to represent the dynamics of presence-absence of all species and we incorporate the effect of dispersal, environmental filtering and ecological interactions in such a way that we could cover all possible scenarios wherein species differ in any combination of ways. We include interspecific competition along with other types of spatial dynamics such as predator-prey interactions (Gravel et al. 2011), priority effects (Shurin et al. 2004), or mutualistic interactions (e.g. Gilarranz et al. 2015). Following a colonization event from time t to $t + \Delta$ corresponds to:

$$P(X_{i,z,t+\Delta t} = 1 \,|\, X_{i,z,t} = 0) = I_{i,z,t} S_{i,z,t} C_{i,z,t}$$

where $I_{i,z,t}$ is the number of immigrants of species i reaching patch z at time t, $S_{i,z,t}$ is the effect of environmental filtering on the probability of establishing a viable local population and $C_{i,z,t}$ is the effect of ecological interactions on the establishment probability. We note that because we represent a stochastic process, the product of these three functions has to be bounded between 0 and 1. We consequently define these quantities:

The effect of immigration is given by:

$$I_{i,z,t} = \frac{\sum k(z,\omega) X_{i,\omega,t}}{\sum k(z,\omega)}$$

which is a weighted average of the occurrence probability of species i in the neighborhood of z. The function $k(z,\omega)$ is a dispersal kernel that depends on the location of patch z and the neighborhood ω. For convenience, we consider an exponential function of the Euclidean distance between localities. We add to the kernel a low distance and neighborhood-independent constant m in order to account from immigration from outside the simulated metacommunity. This assumption is required to prevent total extinction by drift under perfectly neutral dynamics.

The effect of the environment is given by a product of the establishment performance over all environmental variables E_n:

$$S_{i,z,t} = \prod f(E_{n,z}\, \mu_{i,n}\, \sigma_{i,n})$$

In our simulations, for convenience, we consider that the function f has a quadratic form for all species and all environmental variables, though the model is flexible enough to consider other responses that also differ among species.

Ecological interactions on establishment probability

To incorporate all possible ecological interactions, we start by representing the interaction network by a community matrix $\mathbf{A}$ of R species. The elements α_{ij} of $\mathbf{A}$ quantify the effect of species j on the dynamics of species i. When α_{ij} is negative, the colonization probability of species i decreases and/or its extinction probability increases when j is found locally. Inversely, when α_{ij} is positive, the colonization probability increases and/or the extinction probability decreases. To account for the cumulative effects of local interactions on transition probabilities, we make colonization and extinction probabilities community dependent. As explained above, at a time t, the $\mathbf{Y}_{z,t}$ vector gives the local assemblages. We calculate the sum of interactions at any time and for each species as $\nu = \mathbf{A}_{z,t}\mathbf{Y}_{z,t}$. Our approach can be interpreted as a spatial analogue to the generalized Lotka–Volterra model because it takes into account the impact of the whole network of interactions on each species dynamics and can deal with any type of interaction. We now define the function

$$C_{i,z,t} = g(\nu_i, z, t)$$

representing the total effect of ecological interactions on the colonization probability. For convenience, we will use a sigmoid function, with g ranging between c_{min} at high negative interactions and c_{max} at high positive interactions, where c_{max} should be interpreted as the maximal colonization probability when the environmental conditions are optimal and there are no dispersal limitations.

Patch Extinction

The definition of the extinction probability follows exactly the same rules as for colonization, except that extinction is independent of the neighborhood composition. We follow the same logic to define the effect of ecological interactions and of variation in the environment. Consequently, we get the Markovian process:

$$P\left(X_{i,z,t+\Delta t} = 1 \,\middle|\, X_{i,z,t} = 0\right) = M_{i,z,t}E_{i,z,t}$$

where $M_{i,z,t}$ and $E_{i,z,t}$ are the responses of the extinction probability to the local environment and to ecological interactions, respectively. The difference with the colonization functions defined in the previous section is that the extinction probability must be larger when interactions are negative and smaller when they are positive. In addition, the extinction rate should be minimal (instead of maximal) at environmental optimum.

Interpretation

To interpret the model, note that, at steady state, for each species, we obtain the expected occurrence probability ($\hat{P}$) at each site as:

$$\frac{\hat{P}^{iz}}{1-\hat{P}^{iz}} = \frac{I^{iz} \cdot S^{iz} \cdot C^{iz}}{M^{iz} \cdot E^{iz}}$$

After a log transformation, this yields:

$$\log\left(\frac{\hat{P}^{iz}}{1-\hat{P}^{iz}}\right) = \log\left(I^{iz}\right) + \log\left(\frac{S^{iz}}{M^{iz}}\right) + \log\left(\frac{C^{iz}}{E^{iz}}\right)$$

The last equation can be interpreted as a macroscopic description of the expected species distribution pattern (Thuiller et al. 2013). In this formulation, $\log(I)$ describes the tendency of a patch to resemble other nearby patches due to the spatial contagion by dispersal, $\log\left(\frac{S}{M}\right)$ describes the tendency of sites to be occupied by species with similar fitness responses to environmental gradients, and $\log\left(\frac{C}{E}\right)$ describes the remaining influence of other species on co-occurrence due to interactions among species. The values for these indices will depend on what choices are made for the components of eq. 1 (see Supporting Information for details on how we implemented this simulations model).

This modeling framework can represent the classical archetypes (Box 1), but also permits more intricate (and likely far more realistic) metacommunity scenarios and predictions. For example, we could use the model to examine how species traits (and environmental context) link to metacommunity dynamics. Moreover, continuous mixtures of different metacommunity extremes (archetypes) can be represented by appropriate choices for dispersal, competitive abilities, and environmental preferences. For instance, species sorting would require a relatively large colonization to extinction ratio along with species-specific environmental requirements and regional similarity (sensu Mouquet and Loreau, 2002). Alternatively, coexistence within competition-colonization trade-offs requires

species to have similar responses to the environment and appropriate heterogeneities in the I, C and E functions, but no environmental preferences.

The implemented mechanisms in the simulation model can be partly mapped onto variation partitioning components. For instance, at equilibrium, we could expect dispersal limitation (the $\log(I)$ term in equation 3) to create positive spatial autocorrelation at the dispersal scale (the [S/E] fraction in variation partitioning). Environmental filtering (the $\log\left(\frac{S}{M}\right)$ term in equation 3) should lead to a correlation between composition and environment (the [E/S] fraction in variation partitioning). The last term in equation 3, however, describing the effect of interactions on distribution (the $\log\left(\frac{C}{E}\right)$), is novel and has no equivalent in the context of classical variation partitioning.

Description of the Statistical Framework

Hierarchical Community Models

In their simplest form, Hierarchical Community Models (HCMs) resemble standard species distribution models that regress species presences/absences against environmental predictors. However, to reduce model complexity, HCMs assume that all species in a metacommunity will react to environmental heterogeneity following a similar response function (e.g., linear vs quadratic or Gaussian). To account for spatial autocorrelation, either spatial variables such as Moran's eigenvectors maps (MEM, Dray et al. 2006) or spatially auto-correlated latent variables (Ovaskainen et al. 2016b) can be incorporated to the model. To account for biotic interactions, non-spatially auto-correlated latent variables are used. If we use a linear specification approach (here, this can also include quadratic terms that capture Gaussian responses to environment as imposed in our model), we can write:

$$\mathbf{L}^{zi} = \mathbf{X}^{zk}\mathbf{B}^{ki} + \boldsymbol{\epsilon}^{zi}$$

with

$$\mathbf{B}^{ki} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\mathbf{L}^{zi}$ is the presence (or absence) of species i (out of m species) at patch z (out of n patches), $\mathbf{X}^{zk}$ is the value of the environmental variable k (out of p variables) at site z, $\mathbf{B}$ is a matrix of regression parameters, $\boldsymbol{\mu}$ is a vector of length p that describes the mean environmental response of all species, Σ is a $p \times p$ covariance matrix that describes how species vary (diagonal) and co-vary (off-diagonal) around the mean environmental

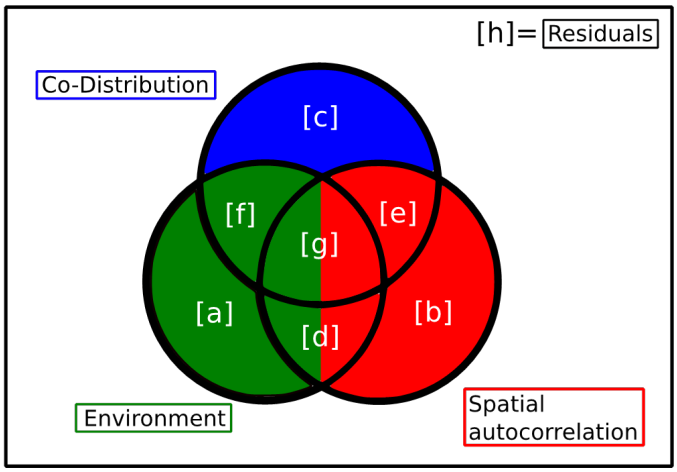
response (Ovaskainen and Soininen 2011), and $\boldsymbol{\epsilon}^{zi}$ is a residual value. Estimating species parameters hierarchically around a community mean reduces the degrees of freedom and makes the model easier to fit with limited data. Both $\boldsymbol{\mu}$ and Σ can be further informed or constrained by species traits or phylogeny if desired (Ovaskainen et al. 2017). To account for biotic interactions, we consider latent variables $\mathbf{H}^{zl}$ (where l refers to a latent variable measured at site z) and their associated parameters $\boldsymbol{\Lambda}^{li}$ (Ovaskainen et al. 2016a). This yields:

$$\mathbf{L}^{zi} = \mathbf{X}^{zk}\mathbf{B}^{ki} + \mathbf{H}^{zl}\boldsymbol{\Lambda}^{li} + \boldsymbol{\epsilon}^{zi}$$

Note that it is not necessary to always include all of these components in one model; they can be considered in any combination deemed relevant for a particular question. In this paper, we used Moran's Eigenvector Maps (MEMs; Dray et al. 2006), a powerful and commonly used method to account for space in statistical models involving species distributions. Note that we could have also used autocorrelated latent variables (Ovaskainen et al. 2016b) to account for the space into the model.

Calculating Variation Partitioning for the HCM As in any generalized linear mixed effect model, we can now partition the explained variation into different components, notably environmental heterogeneity, co-distribution (biotic interactions), space and unexplained variation (Figure S1). To estimate the unique contributions of each of these four fractions for each species, we calculated semi-partial coefficients of determination (i.e., based on Type III sum-of-squares as specified in Peres-Neto et al. 2006) using the implementation suggested by Tjur (2009) as being more appropriate for presence-absence data (i.e., logit link) than the traditional variation partitioning based on an identity link. To adjust for the number of variables used to quantify each fraction of the variation partitioning analysis, we applied the adjustment to the coefficient of determination proposed by Gelman and Pardoe (2006) in the variation partitioning analysis, which is designed for hierarchical models. As shown in Figure A1, the different partitions were combined so that a unique value was associated to environment (fractions [a], [d]/2, [f] and [g]/2), co-distribution (fraction [c]), space (fractions [b], [d]/2, [e] and [g]/2) and the unexplained portion of the variation (fraction [h]). Latent variables are quite powerful to isolate structure in the data. As such, in the calculation of the variation partitioning, latent variables will capture almost all variation associated to the environment and space, giving an artificial inflation of the overlapping partitions between co-distribution and environment and co-distribution and space. For this reason, all partitions overlapping with co-distribution (fractions [e], [f] and [g]) were associated to either environment (fractions [f] and [g]) or space (factions [e] and [g]). In this calculation, a unique measure of explained variation (in adjusted R2) is associated to co-distribution (fraction [c]) but this is not the case for environment and

space. To associate a unique value to environment and space, and represent the results as we did in Figure 2 and 3 we divided the fractions overlapping environment and space between these two components. As such, the sum of fractions [a], [f], half of fraction [d] and half of fraction [g] were used to measure the effect of the environment while fractions was considered [b], [e], half of fraction [d] and half of fraction [g] were used to measure the effect of space.



Supplementary Figure 1.

Calculation of the coefficient of determination

Classic coefficient of determination

The coefficient of determination, R^2, that was partitioned in the variation partitioning analysis is calculated for species j as

$$R_j^2 = 1 - \frac{\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^{n}(y_{ij} - \overline{y}_{ij})^2)}$$

where $y_i j$ is the data associated with species j (out of p species) at site i (out of n sites), $\hat{y}_{ij}$ is the model associated to species j at site i and $\overline{y}_j$ is the average of the data for species j across all sites.

Community-level coefficient of determination

Although having an R_j^2 for each species j can be highly informative, it is sometimes useful to also have a single R^2 for the entire community. This is obtained by averaging all R_j^2

$${}^{C}R^2 = \frac{\sum_{j=1}^{p} R_j^2}{p}$$

where the ${}^{C}R^2$ is the community-level R^2.

Site contribution to the coefficient of determination

In the paper, we use the contribution of each site to R_j^2 to present how each site contributes differently to the environment, space and co-distribution for the community. The calculation of the site i contribution to the R_j^2, is calculated as

$$R_{ij}^2 = \frac{1}{n} - \frac{(y_{ij} - \hat{y}_{ij})^2}{(y_{ij} - \overline{y}_{ij})^2}$$

When studying equation A12, a few considerations can be made. The first part of the equation where the 1 of the classic R_j^2 is divided by n is included to make sure that if we sum all R_{ij}^2 across all sites for species j, the resulting value is R_j^2 and not some other value.

More importantly, what can be noticed is that by calculating R_{ij}^2, the contribution of sites to each species R_j^2, we obtain a matrix that has the same dimension as the site by species matrix (an $n \times p$ matrix). Using this matrix, if we sum across all sites, we obtain the R_j^2. However, if we average across the species for site i we obtain the site's contribution to the community-level ${}^{C}R^2$ or ${}^{C}R_i^2$.

Pseudo coefficient of determination

There exist different variants of $\text{pseudo} - R^2$ that can be applied to specific types of data (e.g. Tjur's D (Tjur 2009) is designed to presence-absence data) or more broadly (e.g. Efron's $\text{pseudo} - R^2$ (Efron 1978)). Because all the different variants of $\text{pseudo} - R^2$ can be applied to a model with a single response variable (i.e. a single species), community-level $\text{pseudo} - R^2$ can be obtained following the approach mentioned above. However, this is not always possible when the interest is to calculate the sites' contribution to a $\text{pseudo} - R^2$, especially for variants of $\text{pseudo} - R^2$ that do not have a mathematical structure similar to the classical R^2.

Parameterization and Simulation Scenarios

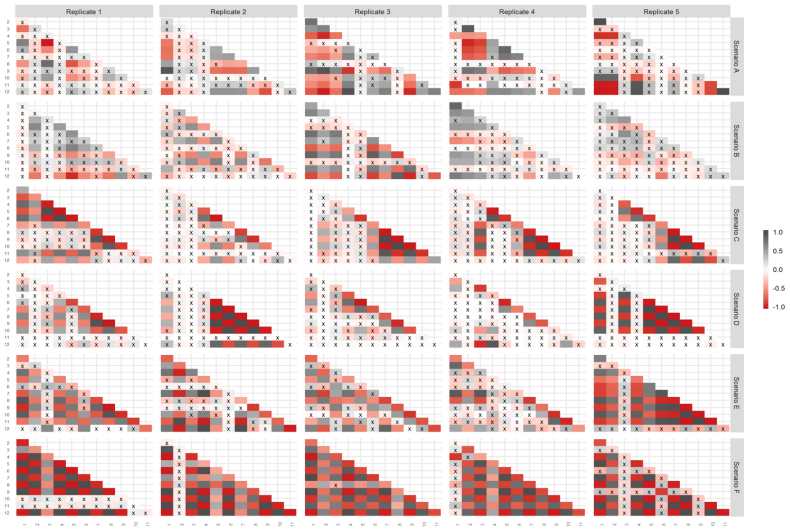
We simulated metacommunity dynamics with a landscape of 1000 patches over 200 time steps and an initial occupancy of 0.8. Patches were placed randomly in a two-dimensional plane with coordinates drawn from a uniform distribution with a minimum of 0 and a maximum of 1. The environment varied spatially, with values drawn from a random distribution between 0 and 1. In the specific simulations we studied in the paper, colonization was the only component of the species that were affected by the environment (i.e. $E_{i,z,t} = 1$). Specifically, colonization reacted to the environment following a quadratic curve.

For all scenarios considered, we simulated 12 species. Niche optimums for the species were evenly distributed between 0.1 and 0.9 while niche breadth was set to 0.8 for simulations with narrow niches (scenarios A, B, E, F, G), and to 2 for simulations where niche was assumed to be broad (scenarios C and D). For dispersal, we considered an exponential

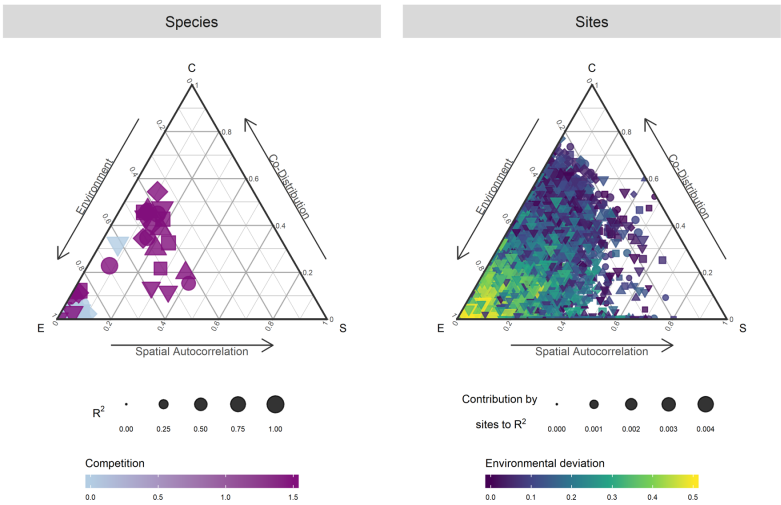
dispersal kernel, with a distance-independent immigration probability of 0.001 and an α parameter of 0.05. For scenario G, where we have variable dispersal kernels (Figure 3), α was 0.01 for 1/3 of species, 0.05 for 1/3 of species, and 0.1 for the other 1/3 of species. We used a sigmoid function to relate the total number of interactions with colonization and extinction coefficients following the implementation by Cazelles et al. (2016). Colonization probability in the absence of interactions was set at 0.4, which tends to zero as negative interactions tend to infinity, while it asymptotes at a 1 with infinite positive interactions. All other aspects of the colonization-interaction curve were the same for all scenarios. Similarly, extinction in the absence of interactions was set at 0.025, and tended to 1 with infinite negative interactions, while its asymptote tended to 0 with infinite positive interactions. In both cases, the parameter setting the shape of the sigmoid function was set to 0 for the scenarios without competition (scenarios A, B, and F), and 1.5 in the presence of competition (scenarios C, D, E, G) . If there were interactions, then a focal species only interacted with the two species that had the closest niches. For all scenarios, five sets of metacommunities were simulated and analyzed to obtain the results found in Figures 2, 3, and 4 in the main text, and supplementary figures.

All scenarios have been implemented in R and the project's repository can be found here: github.com/javirudolph/testingHMSC

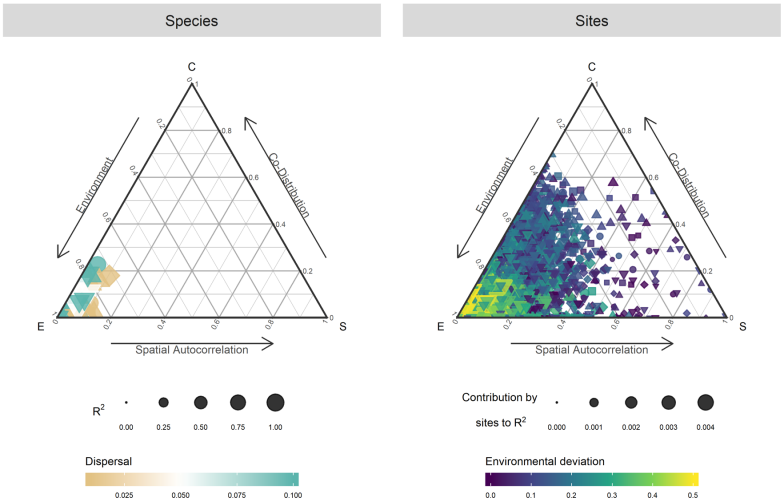
Supplementary Figure 2. Species interactions for all scenarios.



Supplementary Figure 3. In which we have half of the species with interactions and the other without.



Supplementary Figure 4. In which we change dispersal only, α was 0.01 for 1/3 of species, 0.05 for 1/3 of species, and 0.1 for the other 1/3 of species.



Example of metacommunity simulation functions and analyses

The following example of code shows the overall processes involved in the metacommunity simulation for our model. This example shows the scenario for 20 patches and one environmental variable. The model gives the option for a random or spatially aggregated structure for the patches. In the aggregated case, we determined four clusters, denoted by `Nclusters` in the code below. The value of the environmental variable for each patch is shown with the color hue. In this case, the environmental variable is randomly distributed.

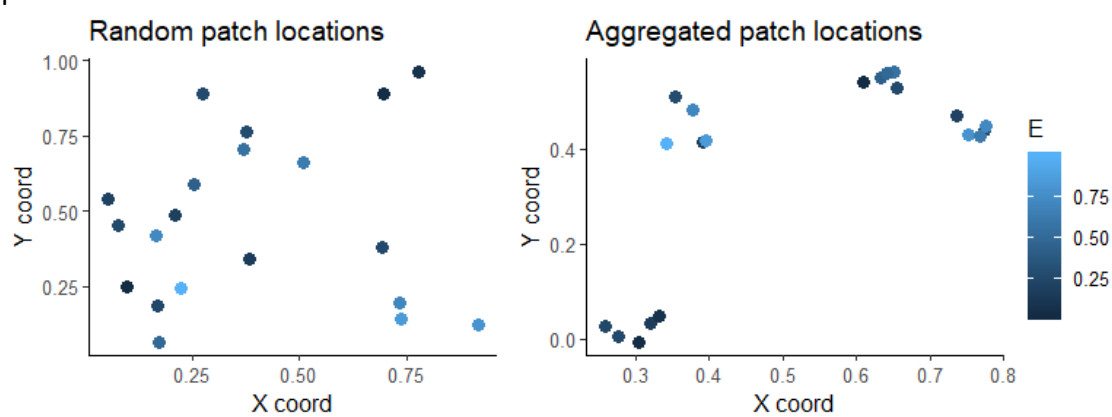
```
set.seed(227)
# Random XY coordinates
```

```r
# Each coordinate is drawn from a random uniform distribution
get_XY = function(N) cbind(runif(N),runif(N))

# Aggregation of XY coordinates
get_XY_agg = function(N, Nclusters, sd_xy) {

  Xclust = runif(Nclusters)
  Yclust = runif(Nclusters)

  X = rnorm(N, rep(Xclust,N/Nclusters), sd_xy)
  Y = rnorm(N, rep(Yclust,N/Nclusters), sd_xy)

  cbind(X,Y)
}

# Random uniform environmental values
get_E = function(D, N) matrix(runif(D*N), nr = N, nc = D)
```

Here, we set N, the number of patches to 20, and D, the number of environmental variables to one.

```r
N <- 20
D <- 1

rXY <- get_XY(N)
agXY <- get_XY_agg(N, 4, 0.02)
E <- get_E(D = D, N = N)
```

The following figure shows a side by side comparisson between random and aggregated patches obtained from our functions.

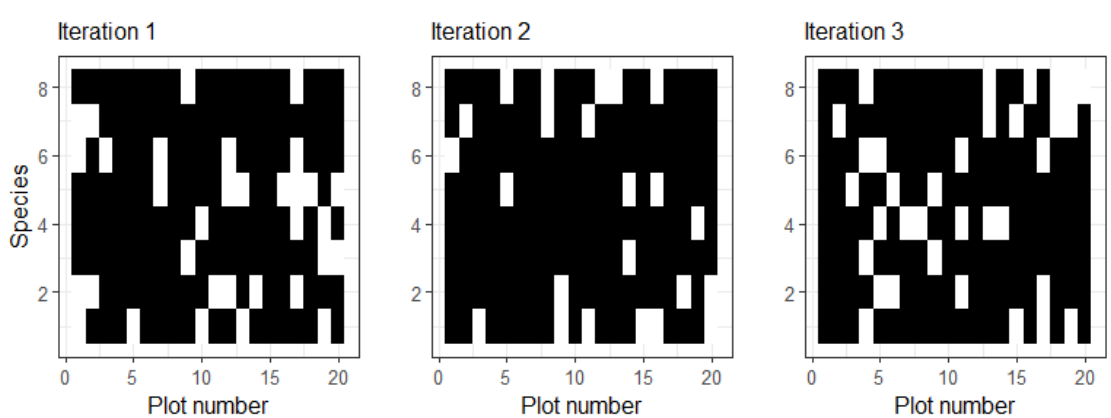


Initial Occupancy

The toy model allows for setitng the initial occupancy in the metacommunity. For example, to create the initial conditions, $t = 0$, of presence absence, species occupancy is drawn from a random uniform distribution, and values smaller than 0.8 are considered as species presence. Patches or locations z are represented by rows in the matrix, whereas each species is a column. Each cell in the matrix is $X_{i,z,t}$ and each row is $\mathbf{Y}_{z,t}$ for $t = 0$. The following figure shows three

different iterations of this process, where areas in black represent occupancy = 1, and white denotes an absence.

```
#Get your initial conditions:
R <- 8
Y0 = matrix(0, nrow = N, ncol = R)
rand = matrix(runif(N*R), nr = N, nc = R)
Y0[rand < 0.8] = 1

ggplot(melt(Y0), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "black")+
  labs(x = "Plot number", y = "Species", subtitle = "Iteration 1") +
  theme_bw() -> A

# Second iteration for the same initial conditions
Y0 = matrix(0, nrow = N, ncol = R)
rand = matrix(runif(N*R), nr = N, nc = R)
Y0[rand < 0.8] = 1

ggplot(melt(Y0), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "black")+
  labs(x = "Plot number", y = "" , subtitle = "Iteration 2") +
  theme_bw() -> B

# Third iteration
Y0 = matrix(0, nrow = N, ncol = R)
rand = matrix(runif(N*R), nr = N, nc = R)
Y0[rand < 0.8] = 1

ggplot(melt(Y0), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "black")+
  labs(x = "Plot number", y = "", subtitle = "Iteration 3") +
  theme_bw() -> C

# Set the three iterations to be displayed side by side
ggarrange(A, B, C, ncol = 3, common.legend = TRUE, legend = "none")
```



When we consider the immigration component, we need to also consider the connectivity matrix. In the code below, `I_f` calculates the probabilty of immigration for each species,

based on the occupancy matrix and the dispersal kernel, K. The argument `K` is the connectivity matrix. The argument `XY` corresponds to the patch coordinates, whereas `alpha` is the dispersal parameter associated to the exponential distribution used for dispersal. It can be computed with the following function:

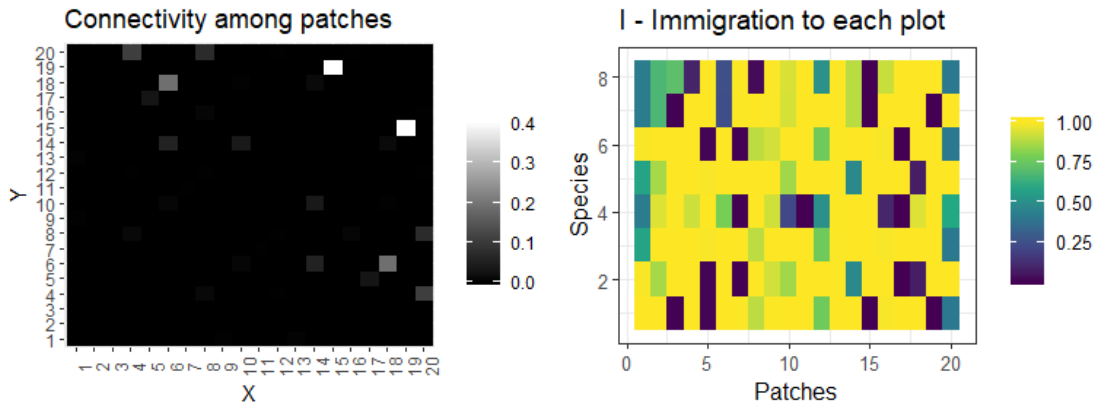
```
# Compute the propagule pressure
I_f = function(Y, K, m) I = (1-m)*(K%*%Y)/(K%*%matrix(1,nr=N,nc=R)) + m
```

The arguments for this function are: `Y, K, m`. We calculated $\mathbb{Y}$, species presence or absence, in the previous section with the case for initial conditions. Argument `m` is set in the parameters as a value `m = 0.001` and the connectivity matrix `K` is calculated below.

```
# Compute the connectivity matrix
get_K = function(XY, alpha) {
    N = nrow(XY)
    distMat = as.matrix(dist(XY, method = "euclidean", upper = T, diag = T))
    ConMat = exp(-1/alpha*distMat)
    diag(ConMat) = 0
    return(ConMat)
}
```

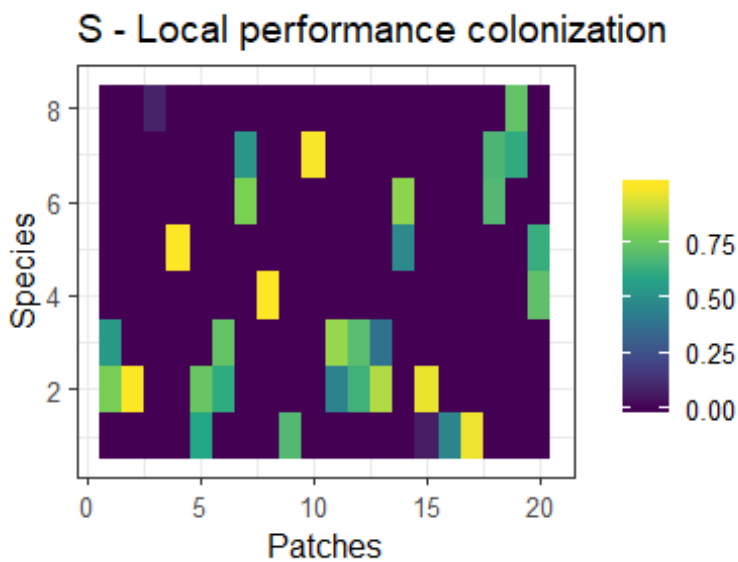
As an example, using the aggregated XY coordinates for 20 patches and our initial occupancy matrix with 8 species ,we can see the connectivity between patches, and can calculate the contribution of immigration from each species to each patch.

```
# We can use the aggregated XY coordinates for this example:
XY <- agXY

# Connectivity matrix
alpha <- 0.005
K <- get_K(XY, alpha)
plot_K <- cbind(rownames(K), stack(as.data.frame(K)))
names(plot_K) <- c("X", "Fill", "Y")
plot_K %>%
  as_data_frame() %>%
  mutate(X = factor(X, levels = c(1:20)),
         Y = factor(Y, levels = c(1:20))) -> plot_K
ggplot(plot_K, aes(x = X, y = Y, fill = Fill)) +
  geom_raster() +
  scale_fill_gradient(low = "black", high = "white") +
  #scale_fill_viridis_c() +
  labs(title = "Connectivity among patches") +
  theme(legend.title = element_blank(),
        axis.text.x = element_text(angle = 90)) -> A
# Immigration
m <- 0.001
Y <- Y0
I <- I_f(Y, K, m)
rastPlot(I, title = "I - Immigration to each plot", x = "Patches", y = "Species") +
  theme(legend.title = element_blank()) -> plot_I
ggarrange(A, plot_I, ncol = 2)
```



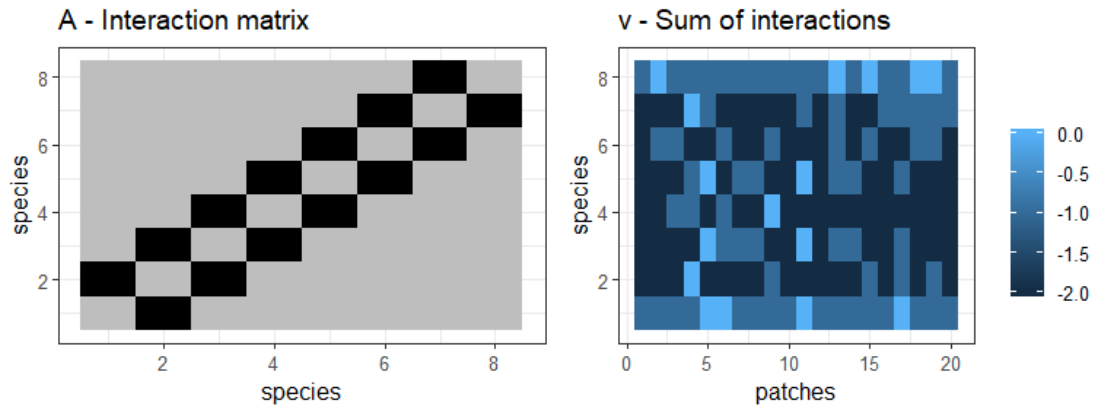
The effect of the environment on each species, depending on each species niche optima, is computed in the following code section. The argument `E` corresponds to the vector of values for the environmental variable in each patch. The other two arguments in this function are the niche optima (`u_s`) for each species and niche breadth(`s_c`).

```
# Compute the local performance of propagules
S_f_quadratic <- function(E, u_c, s_c) {
  R <- ncol(u_c)
  N <- nrow(E)
  D <- ncol(E)
  S <- matrix(1, nr = N, nc = R)
  for(i in 1:D){
    optima <- matrix(u_c[i,],nrow = N,ncol = R,byrow = TRUE)
    breadth <- matrix(s_c[i,],nrow = N,ncol = R,byrow = TRUE)
    S <- S * ((-1 / (breadth/2)^2) * (E[,i] - optima)^2 + 1)
    S <- ifelse(S < 0, 0 , S)

  }
  return(S)
}

# Understood as niche optima for each species, for each environmental variable
  u_c = matrix(nr = D, nc = R)
  u_c[1,] = seq(0.1,0.9, length=R)
# Understood as niche breadth
  s_c = matrix(0.2, nr = D, nc = R)

# Local performance, colonization
S <- S_f_quadratic(E, u_c, s_c)
plot_S <- rastPlot(S, title = "S - Local performance colonization", x = "Patches", y = "Specie
s") +
  guides(fill = guide_colorbar(title = ""))
plot_S
```



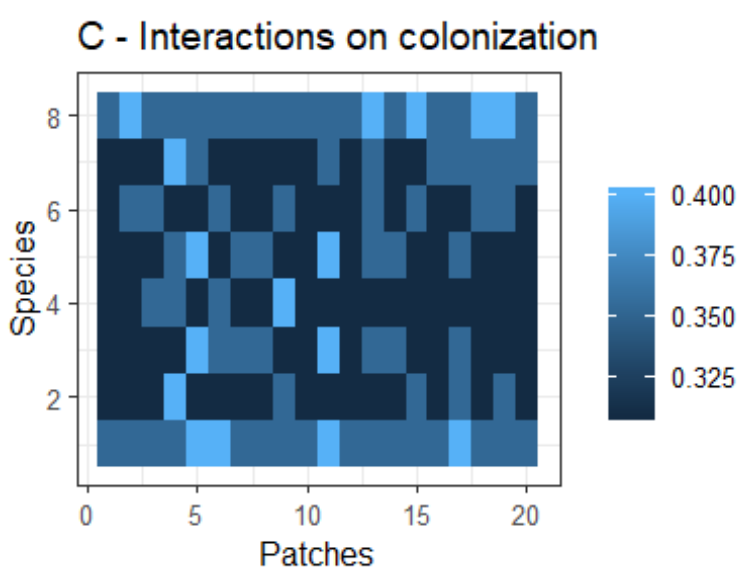
When incorporating species interactions into the toy model, we use the following interaction matrix **A**, where the colored black sections show species with potential of interacting:

```
  # # Interaction matrix
  A = matrix(0,nr=R,nc=R)
  d = as.matrix(dist(c(1:R),upper=TRUE,diag=T))
  A[d<=1] = -1
  diag(A) = 0

  plot_A <- rastPlot(A, title = "A - Interaction matrix", x = "species", y = "species") +
    scale_fill_gradient(low = "black", high = "grey") +
  theme(legend.position = "none")

# Compute the sum of ecological interactions for every location and every species
sum_interactions = function (A, Y) t(A%*%t(Y))
# this is considered to be "v"
v <- sum_interactions(A, Y)

plot_v <- rastPlot(v, title = "v - Sum of interactions", x = "patches", y = "species") +
  guides(fill = guide_colorbar(title = "")) +
  scale_fill_gradient()

ggarrange(plot_A, plot_v, ncol = 2)
```

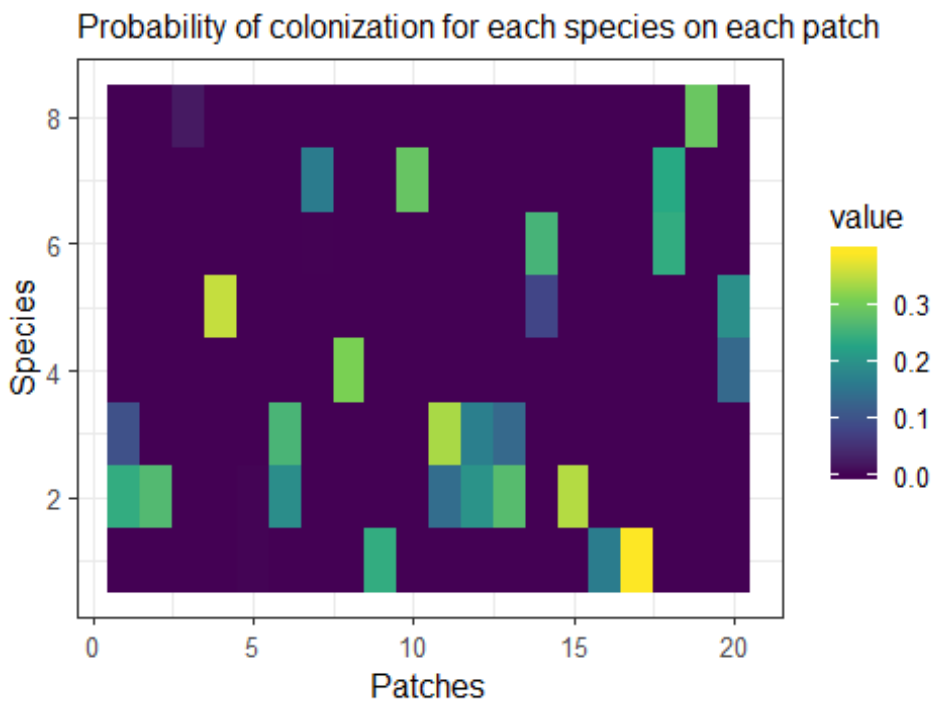


We compute the effect of ecological interactions on colonization probability with the function below.The arguments for this function are `v` as the resulting matrix from the sum of interactions, `d_c` as the sensitivity to interactions, `c_0` and `c_max` as the colonization parameters:

```
C_f = function(v, d_c, c_0, c_max) c_max*(1 +(1/c_0 - 1)*exp(-v*d_c))^-1

  # # Colonization function
  c_0 = rep(0.4, R) # Colonization at 0 interactions
  c_max = rep(1, R) # Colonization at max interactions

  # # Sensitivity to interactions
  d_c = 0.2

C <- C_f(v, d_c, c_0, c_max)
plot_C <- rastPlot(C, title = "C - Interactions on colonization", x = "Patches", y = "Species
") +
  guides(fill = guide_colorbar(title = "")) +
  scale_fill_gradient()
plot_C
```



With all the components calculated, we can now compute the colonization probability $P\left(X_{i,z,t+\Delta t} = 1 \,\middle|\, X_{i,z,t} = 0\right) = I_{i,z,t} S_{i,z,t} C_{i,z,t}$



The following function calculates the effect of the environment on the extinction, with `E` being the environmental variable, and `u_e` and `u_s` being species level effect and the assymptote.

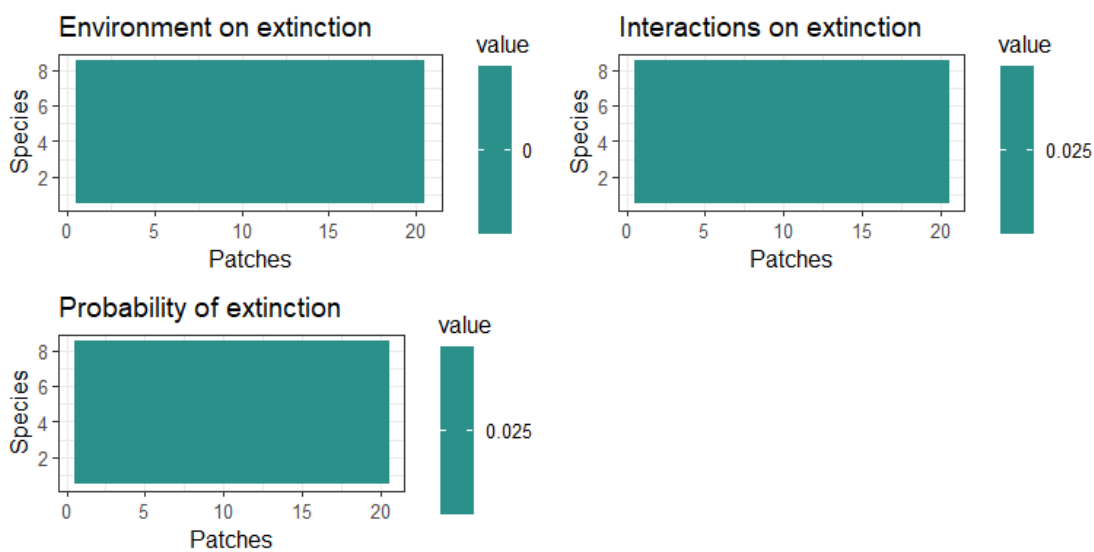
```
M_f = function(E, u_e, s_e) {
    R = ncol(u_e)
    N = nrow(E)
    D = ncol(E)
    M = matrix(1, nr = N, nc = R)
    for(i in 1:D){
      M = M*(1-exp(-(E[,i]-matrix(u_e[i,],nr=N,nc=R,byrow=TRUE))^2 / matrix(s_e[i,],nr=N,nc=R,
byrow=TRUE)^2))
      }
    return(M)
}

# Set the function arguments
  # # Effect of the environment on extinction
  u_e = matrix(nr = D, nc = R)
  u_e[1,] = c(rep(0.5, R-1), 0.05) # One species having a lower level of extinction from envir
onmental effect
  s_e = matrix(Inf, nr = D, nc = R)

  #u_e
  #head(s_e)
```

The following shows the effect of ecological interactions on extinction, using the same v matrix calculated above.

```
E_f = function(v, d_e, e_0, e_min) {

    e_min_mat = matrix(e_min, nr = N, nc = R, byrow=TRUE)

    e_min_mat+(1/(1-e_min_mat)+(1/(e_0-e_min_mat)-1/(1-e_min_mat))*exp(d_e*v))^-1

}


#With the arguments computed as:


# # Extinction function
  #e_0 = c(rep(0.025, R-1), 0.5) # Extinction at 0 interactions, with one species having a hig
her value.
e_0 = rep(0.025, R)
e_min = rep(0, R) # Exinction at max interactions

  # # Sensitivity to interactions

  d_e = 0
```

We can now compute the probability of extinction $P\left(X_{i,z,t+\Delta t} = 1 \,\middle|\, X_{i,z,t} = 0\right) = M_{i,z,t}E_{i,z,t}$

The figure shows these as identical graphs, since we have made all species have the same probability of extinction and the environment not having an effect on extinction.

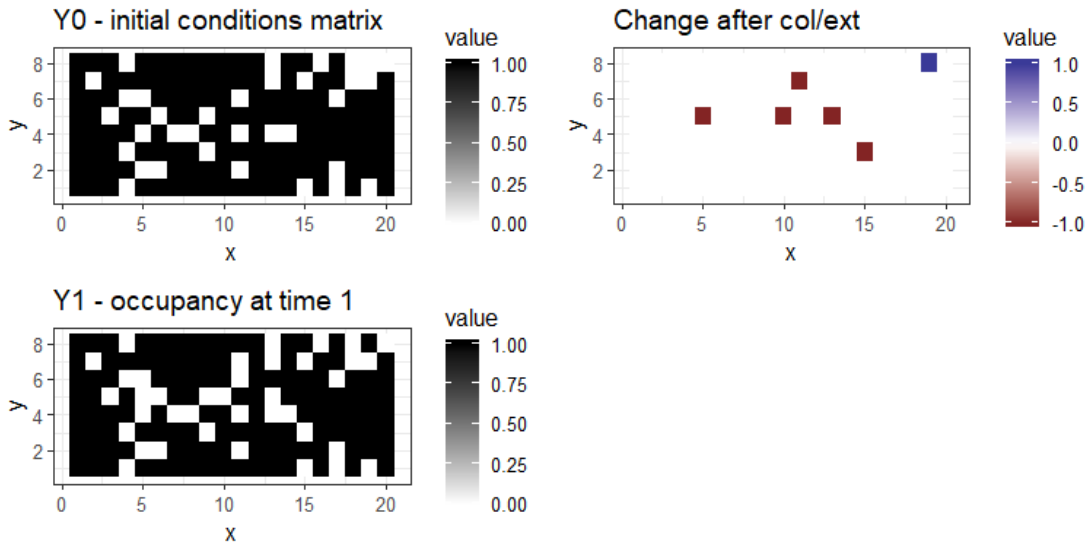


The way parameters are set, the extinction component is the same for all species in every patch.

Testing and changes

```
# Perform the test
delta <- matrix(0, nr = N, nc = R)
```

```
rand <- matrix(runif(N*R), nr = N, nc = R)
delta[Y == 0 & rand < P_col] <- 1

# Perform the test
rand = matrix(runif(N*R), nr = N, nc = R)
delta[Y == 1 & rand < P_ext] = - 1
```



References:

Dray, S., P. Legendre, and P.R. Peres-Neto. (2006). Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecological Modelling, 196, 483-493.

Efron, B. 1978. Regression and ANOVA with zero-one data: Measures of residual Variation. Journal of the American Statistical Association 73:113–121.

Gelman, A., and I. Pardoe. (2006). Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics 48, 241–251.

Gilarranz, L. J., Sabatino, M., Aizen, M. A., & Bascompte, J. (2015). Hot spots of mutualistic networks. Journal of Animal Ecology, 84(2), 407-413.

Gravel, D., Massol, F., Canard, E., Mouillot, D. & Mouquet, N. (2011). Trophic theory of island biogeography. Ecology Letters, 14, 1010–1016.

Gelman, A., and I. Pardoe. 2006. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics 48:241–251.

Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016a). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. Methods in Ecology and Evolution, 7, 549–555.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology, 91, 2514–2521.

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016b). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods in Ecology and Evolution, 7, 428–436.

Ovaskainen, O. & Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. Ecology, 92, 289–295.

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, 561–576.

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation partitioning of species data matrices: Estimation and comparison of fractions. Ecology, 87, 2614–2625.

Shurin, J.B., Amarasekare, P., Chase, J.M., Holt, R.D., Hoopes, M.F. & Leibold, M.A. (2004). Alternative stable states and regional community structure. Journal of Theoretical Biology, 227, 359–368.

Tjur, T. (2009). Coefficients of determination in logistic regression models— A new proposal: the coefficient of discrimination. American Naturalist, 63, 366–372.