Community composition exceeds area as a predictor of long-term conservation value

Running title: Predictors of conservation value

Authors: Jacob D. O'Sullivan[1], J. Christopher D. Terry[1], Ramesh Wilson[1], Axel G. Rossberg[1]

Affiliations: [1]School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, United Kingdom

Corresponding author: Jacob Dinner O'Sullivan (j.osullivan@qmul.ac.uk), School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom

Document statistics: Abstract - 225 words; Main Text - 6576 words, 11 figures; Appendix 1058 words, 2 figures.

Pre-print URL: `https://www.biorxiv.org/content/10.1101/2021.07.09.451808v1`

Keywords: biodiversity — macroecology — spatial ecology — metacommunity — conservation — land-use change

Competing interests: The authors declare no competing interests.

Data/code availability: Simulation data required to reproduce the results summarised in this study will be made available upon publication. Previous versions of the software used to generate results are already available at `https://github.com/jacobosullivan/LVMCM_src/`. Study specific updates will be made available upon publication.

Funding: This work forms part of the project 'Mechanisms and prediction of large-scale ecological responses to environmental change' funded by the Natural Environment Research Council (NE/T003510/1).

Abstract

Conserving biodiversity often requires deciding which sites to prioritise for protection. Predicting the impact of habitat loss is a major challenge, however, since impacts can be distant from the perturbation in both space and time. Here we study the long-term impacts of habitat loss in a mechanistic metacommunity model in terms of both immediate extinctions and secondary species losses. We find that biomass-at-site, closely related to site area, is a poor predictor of long-term regional species losses following site removal. Knowledge of the compositional distinctness (average between-site Bray-Curtis dissimilarity) of the removed community can markedly improve the prediction of impacts at the regional scale, even when biotic responses play out at substantial spatial or temporal distance from the removed site. Fitting our model to empirical species-by-site tables describing Andean diatoms and Brazilian lichen-fungi, we show that compositional distinctness surpasses area as a predictor of long-term species losses in the empirically relevant parameter range. Our results robustly demonstrate that site area alone is not sufficient to gauge conservation priorities; analysis of compositional distinctness permits improved prioritisation at low cost.

1 Introduction

Habitat loss due to conversion of natural landscapes is the leading cause of biodiversity loss today (Millenium Ecosystem assessment, Reid et al. 2005; IPBES, Díaz et al. 2019). Immediate species losses that result when the habitats of endemic species are destroyed only represent part of the impact of land conversion. Additional losses may occur as 'extinction debts' are paid (Tilman et al., 1994). These additional extinctions can arise due to a suite of complex processes that ripple through the wider landscape, often involving multiple species (Swift and Hannon, 2010; Mouquet et al., 2011; Nee and May, 1992; Livingston et al., 2012; Terborgh et al., 2001; Taheri et al., 2021). The complexity of these ecological responses to habitat loss makes predicting long-term impacts a major challenge. Essential to meeting this challenge is an understanding of how changes in the abiotic and biotic structure of the landscape are likely to interact (Chase et al., 2020).

Decisions in conservation ecology often require identification of the least worst outcomes of landscape conversion (Wilson et al., 2009). For such assessments, sufficient mechanistic understanding of the biodiversity impacts of habitat loss is required (Figueiredo et al., 2019). Predictions of such impacts often rely on phenomenological models of species-area relationships (SAR) (Kuussaari et al., 2009; Figueiredo et al., 2019; Thompson et al., 2019), which assume that impacts follow simple scaling relationships (e.g. Thomas et al. 2004; Foster et al. 2013). However, the scaling of diversity with area arises due the fact that ecological assemblages are internally heterogeneous: it is usually not area *per se* that determines species richness but the diversity of ecological associations that a landscape can support. As such, it is plausible that metrics that directly quantify the internal diversity structure of a assemblages may outperform area alone as predictors of a site's contribution to regional diversity.

Here, we explore the long-term outcomes of habitat destruction using a spatially explicit simulation model called the Lotka-Volterra Metacommunity Model (LVMCM), which has been shown to reproduce a large variety of well-documented patterns of spatio-temporal macroecology (O'Sullivan et al., 2019, 2021a). We model habitat conversion as the complete removal of sites from a metacommunity, and find that indeed the biotic impacts following site removal can be complex. Secondary extinctions including extinction cascades are common. These cascades can cause extinctions of populations distant from the removed site. The area of the removed site only weakly correlates with *conservation value*, which we defined as the long-term, landscape-level species loss following site removal. A stronger predictor of regional extinctions is often the compositional distinctness of the removed community, despite the cascading, far-reaching impacts removal can have. To test whether empirical systems fall into the parameter space where compositional distinctness is a stronger predictor than area, we developed a method

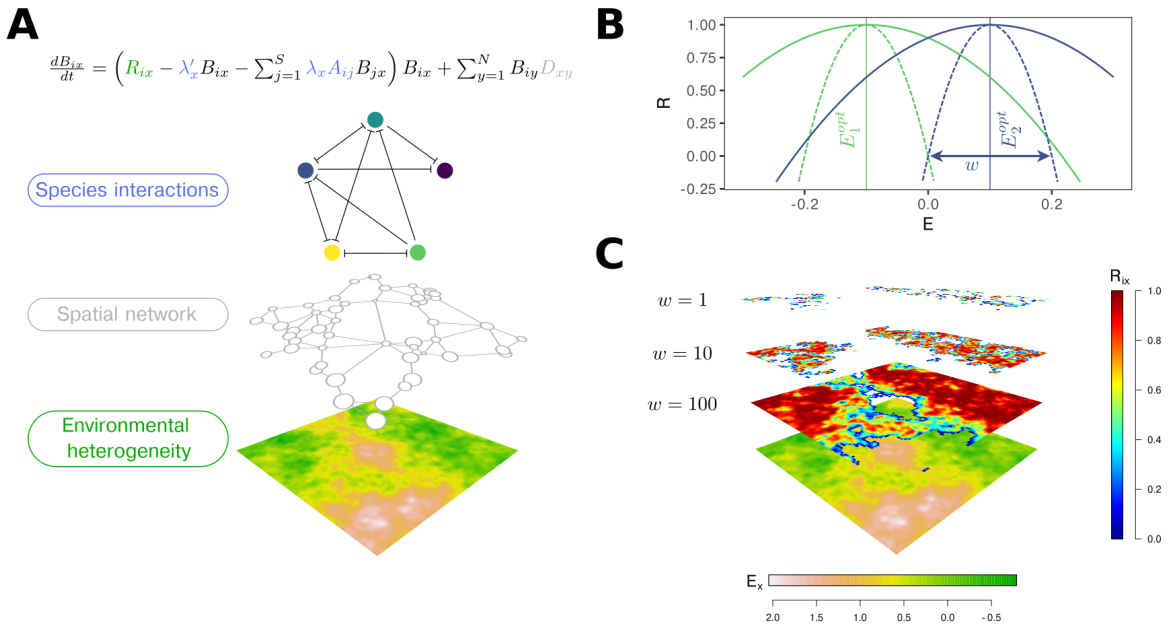

Figure 1: **Elements of the metacommunity model. A**: Model metacommunities include a spatially autocorrelated environment of at least one abiotic variable and a spatial network, in which local sites of unequal area are modelled by scaling the local interaction coefficients by λ_x and λ'_x. **B**: Intrinsic growth rates $R = R_{ix}$, representing species' Hutchinsonian niches, are modelled as quadratic functions of the environmental variables. Niche width is controlled by a parameter w and each species is assigned a randomly sampled environmental optimum E_i^{opt} (solid lines exemplify large w, dashed lines small w, colours represent different species). **C**: Three examples illustrating of how niche width w affects the distribution of areas of positive R over the landscape, highlighting the relationship between niche width and effective heterogeneity.

to fit the LVMCM to biodiversity patterns derived from empirical species-by-site tables. Using this method, we demonstrate not only the higher predictive power of compositional distinctness for empirical systems, but also the wider potential of mechanistic metacommunity models as decision support tools in applied conservation (Chase et al., 2020).

1.1 Model description - Overview

The LVMCM extends the conventional Lotka-Volterra competition model to a spatially explicit system of connected sites (Fig. 1). It models the dynamics of a metacommunity formed by a guild of competing species. Sites differ in their suitability for each species and their total area, and nearby communities are connected by dispersal. The underlying environment is modelled as a spatially autocorrelated random field with mean zero and unit standard deviation and species responses to the environment are modelled by quadratic environmental response functions,

with each species allocated a random environmental optimum. Species disperse between sites according to an exponential dispersal function. We kept the width of the environmental response function, w, and the characteristic dispersal length ℓ fixed for all species in a given simulation, but varied them systematically between simulations. For this study we kept the number of sites fixed at 20 while, under variation of these key parameters, the regional γ-diversity ranged from 250 to 2500 species.

We allowed metacommunity models to self-organise via a regional assembly process (Post and Pimm, 1983) through which species invade the metacommunity and distribute across the landscape, responding to the abiotic and biotic conditions in the sites. We then systematically perturbed assembled metacommunities by removing each site in turn, simulating the biotic response and measuring long-term impacts of single site removal at the regional scale (for technical details see methods).

1.2 Predicting long-term species losses

Process based metacommunity models like the LVMCM permit direct comparison between the immediate effects and long-term outcomes of perturbations. Immediate species losses can be predicted by asking which species have a global range limited to the removed site. We denote the predicted immediate species loss $\Delta\gamma_0$. In contrast, we denote by $\Delta\gamma$ the actual long-term species loss determined by simulating metacommunity dynamics. We define the conservation value of the removed site as the proportional long-term species loss, relative to the pre-disturbance regional species richness γ^*.

We first we asked how $\Delta\gamma/\gamma^*$ depends on model parameters. Determining these parameters empirically, however, can be costly. Therefore we also tested how well $\Delta\gamma/\gamma^*$ can be predicted by the proportion of biomass immediately removed, B_x^*/B^*, where B_x^* and B^* represent the pre-disturbance biomass of the removed site x and of the entire metacommunity, respectively. We use B_x^* as a proxy for the area of site x, with the advantage that B_x^* is directly represented in LVMCM model communities. As a second predictor we use a measure of compositional distinctness, $\bar{\beta}_x$, defined as the mean Bray-Curtis Bray and Curtis (1957) dissimilarity of the focal site x to all other sites.

We assessed the effects of B_x^*/B^*, $\bar{\beta}_x$ and various spatial network properties (centrality measures) on $\Delta\gamma/\gamma^*$ using simple multivariate linear regression, with the best model being selected by comparison of AIC (**?**). Goodness-of-fit of predictive models was assessed using the adjusted R^2. The proportions of variance explained by area and compositional distinctness were then partitioned using partial regression redundancy analysis.

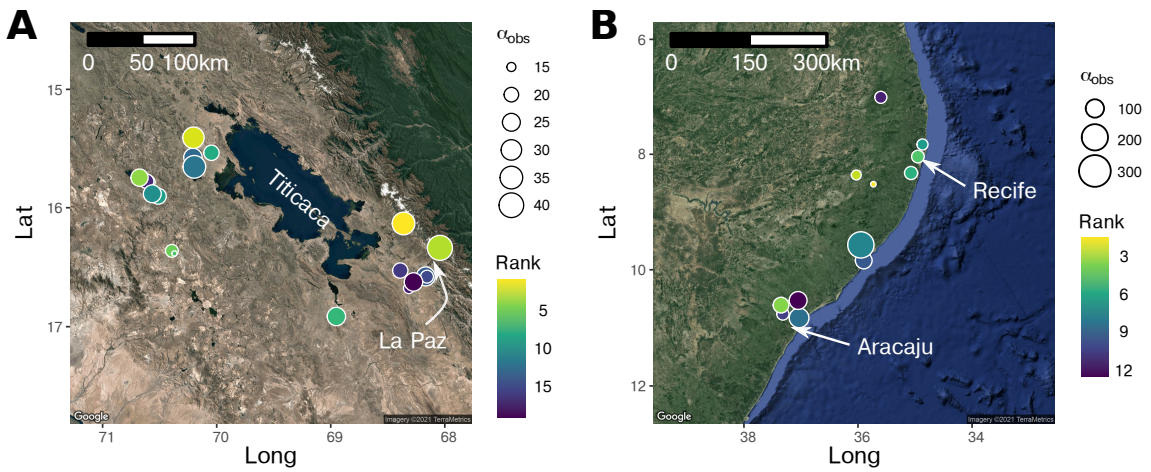

Figure 2: **Datasets used to fit the LVMCM. A**: The subset of the Andean diatom database (Benito et al., 2018) straddling Lake Titicaca and the Peruvian-Bolivian border. **B**: The Brazilian lichen-fungi database (da Silva Cáceres et al., 2017) with nearby samples pooled together. The colour of the points represents the ranked impact of site removal on regional diversity in *in silico* removal experiments, with the lightest colour corresponding to the site with the greatest impact. The size of the points represents the observed species richness.

1.3 Fitting the LVMCM to empirical data

We demonstrate how the LVMCM can be fit to data using two datasets, one describing the distributions of diatoms (D) in lakes straddling the Peruvian-Bolivian border (Benito et al., 2018), the other the distribution of Lichen-Fungi (L) in Eastern Brazil (da Silva Cáceres et al., 2017). These datasets were chosen for their high species richness and because we can assume that the ecological interaction networks of these guilds are reasonably represented by competition within a single trophic level. For dataset D, which covers a large region, we selected a subset of the full database for which the first two principal components of the key environmental variables formed a distinct cluster, including 19 sites and 221 taxa. For dataset L we reduced the total number of sites (and the computational effort) by pooling observations separated by less than 20km. This left 12 sites and 784 taxa.

In the Appendix we give a detailed summary of the model fitting procedure that we outline here. The spatial structure (spatial distance matrix) and key environmental variables from the empirical dataset were input directly into the model, defining the distance between sites and the underlying abiotic environment. We then used a battery of simulations to estimate the abiotic niche width parameter w that best fit the empirical species by site table. This was done by quantitatively comparing the occupancy frequency distribution (OFD) (McGeoch and Gaston, 2002) for the empirical observation to that of the simulation under systematic variation of w.

The LVMCM is a parameter rich model, making direct fitting to data impossible. Therefore we developed a

simplified patch occupancy model that predicts high-level metacommunity properties with a single fitting parameter, which we termed the 'mixing rate' m. Elsewhere we have shown that this simple framework can explain the spatial and temporal properties of riverine communities (O'Sullivan et al., 2021b). While unsuitable for in-depth analyses of biotic responses to perturbation, the simplicity of the patch occupancy model means it can be directly fitted to both empirical observations and LVMCM model communities, thus serving as a bridge between the complex LVMCM simulations and empirical ecosystems. The theory behind this simple model and the procedure for model fitting are summarised in the Appendix.

The parameter m can be estimated from the OFD for a given dataset. For LVMCM models with spatial and environmental distributions matching the observation, the abiotic niche width parameter space of the model was scanned. The value of w that best reproduces the mixing rate of the empirical observation was then located. This value gives a meaningful estimate of the typical niche widths relative to the total environmental variation in the assemblage which includes explicit consideration of the role of biotic interactions in determining species ranges.

2 Results and discussion

Even though none of the sites removed in simulation experiments comprised more than 12% of the total regional biomass/area, we detected regional extinction of at least one species in over 75% of cases. The highest impact of a single site removal was a proportional loss of $\Delta\gamma/\gamma^* = 0.23$. In a supplementary movie we show examples of extinction cascades following site removal, demonstrating species loss from sites both near and distant from the removed site.

2.1 The dynamics of extinction following habitat removal

Proportional long-term species loss $\Delta\gamma/\gamma^*$ decreased with increasing abiotic niche width (Fig. 3A). This is plausible, since with wider niches single-site removal tends to remove a smaller proportion of the area representing a species' Hutchinsonian niche. By contrast we found that dispersal length had surprisingly little effect on $\Delta\gamma/\gamma^*$ (Fig. 3A).

The process by which species are lost following site removals was complex. In Fig. 3B we show a random sample of these complex extinction dynamics for various dispersal lengths with abiotic niche widths fixed at $w = 0.63$, the value for which impacts are largest. We find that site removals can either lead to the loss of endemics only (cummulative extinctions effectively independent of time lag), or they can lead to extinction cascades which can

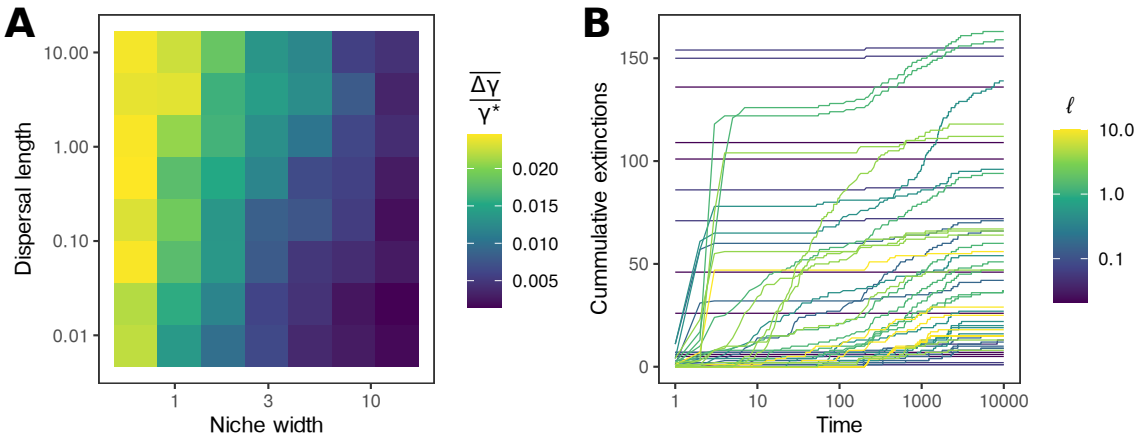

Figure 3: **Species extinctions following site removal. A**: The average proportional species losses for all combinations of the abiotic niche width and dispersal length. **B**: Secondary extinctions occur due to the interuption of source-sink dynamics and extinction cascades. Here we show the outcome of removing sites for the smallest niche width $w = 0.63$.

play out over long times. These extinction cascades, particularly prevalent in higher dispersing metacommunities, demonstrate the complexity of potential metacommunity responses to site removal.

The secondary extinctions of non-endemics can be categorised into two distinct types. Those occurring due the disruption of mass effects (sink populations which are lost following the removal of source sites), and those occurring due to a complex restructuring of the metacommunity as species ranges shift in space. Extinctions of the first type typically occur within around 50 unit times and in sites adjacent to the removed site. The average distance between the site in which a species was last detected and the original perturbation was $0.05L$, where L is the side length of the model landscape, in the first 50 unit times (smallest niche width, across dispersal lengths). The second type can occur much later and at almost any site in the metacommunity. For extinctions occurring after more than 50 unit times, the average distance of the final local extinction in a given species' decline to extinction was $0.45L$. For species lost after 120 unit times, the distance from the initial perturbation had a mean ($0.5004L$) that was statistically indistinguishable from the average inter-site distance ($0.5002L$), suggesting the location of subsequent species losses was essentially random with respect to the initial impact.

2.2 Predicting long-term impacts based on empirically measurable quantities

Predicting the outcome of single site removals, given the complexity of the biotic response (Fig.3C) is non-trivial. It is not the case that species losses in the model can be generically explained by reference to immediate losses of

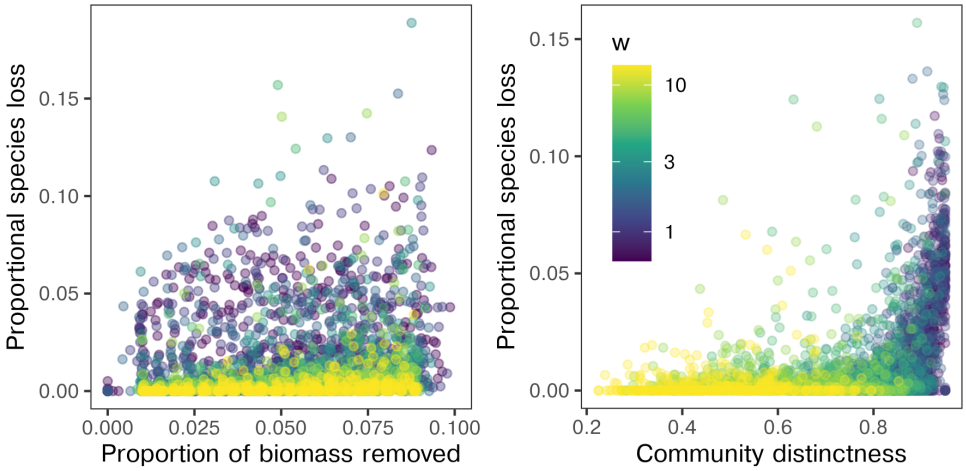

Figure 4: **Predicting long-term species losses.** The proportional long-term species losses following site removal plotted against the proportion of biomass removed initially, a proxy for site area, and the compositional distinctness of the removed site. Here we show a random subset of the simulated removal experiments for visual clarity.

endemics. Instead, we test whether area and compositional distinctness, both empirically accessible properties of communities, correlate with the long-term outcomes of the often complex structural redistribution precipitated by single site removal.

We find a clear positive association between the conservation value of sites and B_x^*/B^*, but with substantial spread (Fig. 4, Spearman's $\rho = 0.33$, all parameter combinations pooled). In contrast, long-term species losses were more strongly associated with compositional distinctness of the removed site $\bar{\beta}_x^s$ (Fig. 4, $\rho = 0.52$, all parameter combinations pooled).

The best model selected by AIC included both proportional area removed and compositional distinctness of removed site as predictors of long-term species losses. Remarkably, none of the standard centrality measures available (degree-, closeness-, betweenness-, eigenvector centrality) were selected as predictors.

Decomposing the simulation results by w, we find that the strength of the association between long-term impacts and the properties of the removed site varies considerably with nice width. In Fig. 5 we show how the R^2 and the regression coefficients of the multivariate linear models vary over the parameter space.

For the smallest niche widths studied here, the R^2 (bars in Fig. 5) of the full model (both predictors) increased up to 0.55 for the smallest niche width. For large niche widths R^2 stabilised at 0.13. Thus, predicting $\Delta\gamma/\gamma^*$ when niche widths are large relative to environmental variation is particularly challenging. This is partly due to the fact that long-term species losses are rare when environments are benign, but may also reflect the fact that

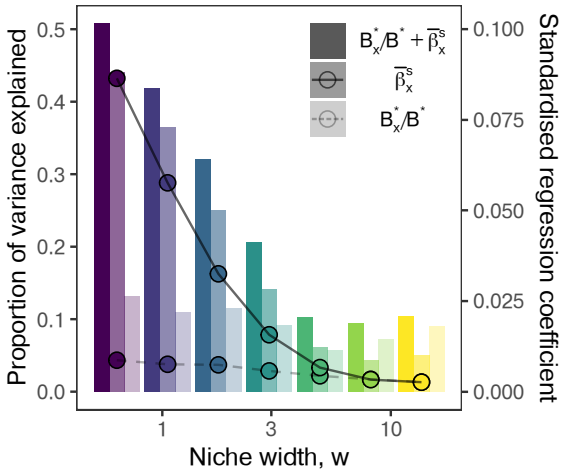

Figure 5: **Dependence of predictive power on niche width.** We find a clear decay with increasing niche width w in the variance explained (bars) by both the full model (combining both predictors) and model based on distinctness. In contrast, the proportion of variance explained by area was largely independent of w. Coincident with the decay in variance explained, the standardised regression coefficients (points) also decayed with niche width. The standardised effects of area (dashed line) was consistently lower than that of compositional distinctness (solid line).

the biotic response to perturbation depends much more on dispersal and local competitive effects–absent from the multivariate regression–where abiotic filtering is largely neutral.

The proportion of variance explained by area following partial regression redundancy analysis was largely consistent across niche width parameterisations. In contrast, the proportion of variance explained by compositional distinctness decayed with increasing niche width. As a result, for the largest niche widths studied, the variance explained by area actually exceeds that by composition, though only once total R^2 had dropped to its minimum. The standardised regression coefficients (points in Fig. 5) show that for all but the widest abiotic niches, the effect of compositional distinctness exceeded that of area on long-term losses (solid and dashed lines respectively).

By repeating this analysis for random sub-sets of the simulated metacommunities containing a fixed number of species, we verified that these differences are not due to differences in overall species richness of low- and high-w metacommunities.

Thus, we conclude that compositional distinctness typically outperforms biomass as a predictor of long-term losses, despite the fact that the population scale impacts are not uniquely felt in or near the impacted site. In order to estimate the long-term effects of habitat destruction on biodiversity, it is critically important to take the community composition of the affected areas into account.

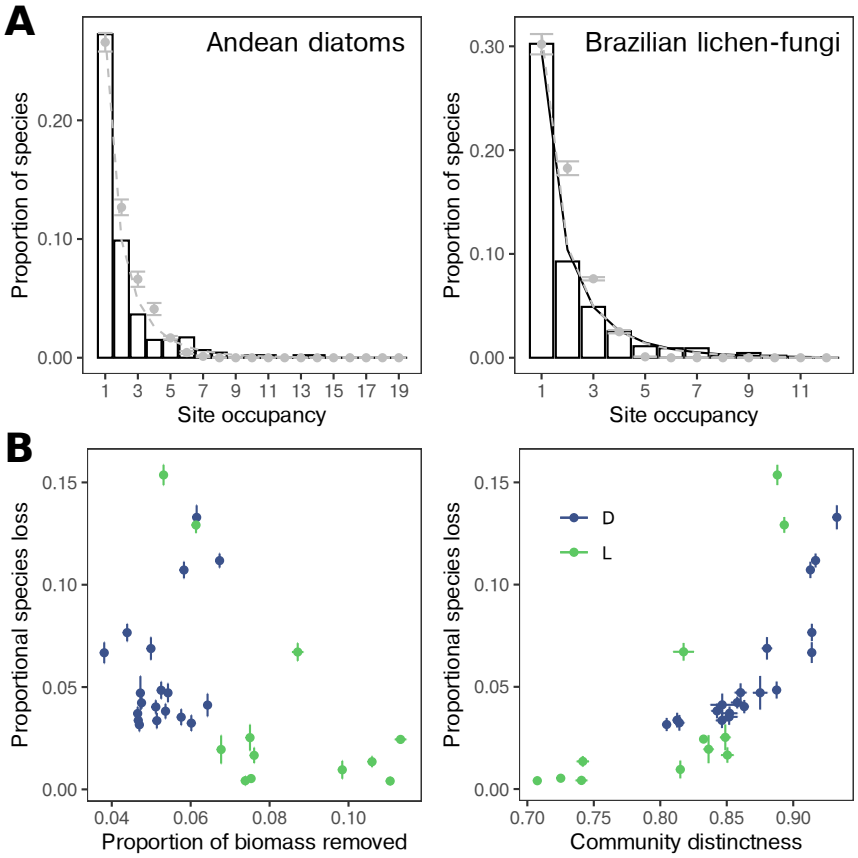

Figure 6: **Fitting the LVMCM to empirical OFD and simulated habitat loss.** **A**: Bars represent the empirical OFDs The black curves are the OFD of the patch dynamics model with the mixing rate fitted to the observation. Grey points are the mean occupancies of 10 replicate LVMCM simulations with parameters that best fit the empirical observations. The grey dashed lines are the OFDs of the patch dynamics model fitted to the simulation. **B**: The outcomes of the simulated removal experiments for the best fitting parameter combination. Each point represents a site, error bars are the standard deviation in the outcome over 10 replicate assemblies.

2.3 Fitting the LVMCM to empirical observation

In view of the strong parameter dependence of the predictive power of compositional distinctness, it is critical to ascertain which region of the parameter space is most representative of natural ecosystems at large scales. We constructed a set of LVMCM metacommunities with spatial and environmental distributions taken from the two empirical datasets. By comparing the mixing rate m estimated from the empirical species by site table to that for LVMCM models we found that dataset D was best fit when $w = 3.46$. In the case of dataset L $w = 1.08$ gave the closest match to observations.

The simulations deviate from the empirical data in two of important respects (Fig. 6A). The empirical OFDs

tend to have a sharper peaks at single site occupancy than the simulations and slightly fatter right tails. This is most likely due to the model's simplifying assumption that all species have the same abiotic niche width. Inspection of the OFD for various niche widths (Fig. S4) suggests that a better fit to the empirical distribution could be achieved by relaxing this assumption. We did not attempt to incorporate a distribution in environmental specialism in our model to avoid over-parameterization and because the shape of such distributions is not well understood empirically (O'Sullivan et al., 2021b).

Despite these differences, our simple fitting procedure generated model metacommunities with macroecological spatial structure (OFD) approximately matching that of the empirical observation. The key finding is that both datasets are fit by abiotic niche widths of order $w \approx 1$—representing around one standard deviation of the landscape scale heterogeneity—suggesting the parameter space in which compositional dissimilarity substantially exceeds biomass as a predictor of conservation value is the most biologically relevant.

Using these fitted models, we also performed a direct mechanistic assessment of the rank conservation priority of each site (Fig. 2), performing systematic removal experiments as described above. In these experiments the proportional drop in diversity $\Delta\gamma/\gamma^*$ after removal of a single site ranged from 0.03 to 0.13 for dataset D and 0.004 to 0.15% for dataset L (Fig. 6B). In both cases, B_x^*/B^* was a rather poor predictor of simulated species losses. On the other hand, a strong non-linear relationship between long-term impact and $\bar{\beta}_x^s$ was found, consistent with the results shown in Fig. 5. For dataset D, conservation priority typically increased with site altitude as spatial isolation and deviation from environmental averages increased. Surprisingly, for dataset L the greatest impacts occurred when either of the two smallest sites were removed. This apparently incongruous result is explained by the fact that for this dataset observed species richness was negatively associated with environmental rarity, measured as the mean Euclidean distance of focal site from all other sites in environment space (Spearman's $\rho = -0.56$, $p = 0.057$). In dataset L, the fact that small sites occupy the most distinct environments, and therefore host the most distinct communities, means the relationship between long-term impact and site area was reversed compared to SAR-based expectation.

We acknowledge that testing the accuracy of these prediction is a challenge. The present study demonstrates the possibility of a bespoke, simulation-based assessment of conservation value, but leaves further refinements and validation of the procedure to future work.

3 Conclusions

The biotic response of metacommunities to localised perturbation can be complex and far reaching. While endemic species are by definition lost immediately, secondary extinctions can be substantial as metacommunities restructure. Predicting the long-term impacts of this restructuring is a major challenge, and one that currently can only be studied using process-based metacommunity models like the LVMCM. With this study we have shown that compositional distinctness can substantially exceed area alone as a predictor of long-term impacts, including secondary extinctions, on biodiversity following habitat destruction.

Compositional distinctness, measured in terms of average dissimilarity, is likely to be empirically accessible because Bray-Curtis dissimilarity between sites is numerically dominated by the locally dominatant populations, the sizes of which are readily quantified. However, the usefulness of the predictor is based not only to what it tells us about the dominating species, but also what the dominating set of species tells us about the abiotic characteristics of a site. The compositional distinctness of the dominant taxa is likely to be informative as to the distinctness of the non-dominant taxa, which is harder to measure.

While at first sight it might appear that application of this predictor requires a well-defined metacommunity to average over (which in practice may be ambiguous), this is not actually the case. Addition or omission of sites from a dataset affects ordering of two sites in terms of mean Bray-Curtis dissimilarity only for those other sites in the dataset that have compositions similar to either of the two sites.

To illustrate the relevance of our results, we note that the recently published first draft of the Post-2020 Global Biodiversity Framework[1] sets as a its primary goal the enhancement of ecosystem integrity, including increasing the area and connectivity of natural ecosystems, ensuring the robustness of populations and the reduction in extinction rates. Accompanying this document is a set of indicators proposed to help monitor progress toward these goals[2]. We can roughly categorise the 56 indicators applicable to the primary goal of the Framework into those relating to (i) area, extent or coverage, (ii) species-level assessments (of extinction risk, habitat integretity etc.) (iii) intactness (describing the proportion of historical assemblages that persists today) and (iv) spatial beta diversity intrinsic to an ecological unit (region, nation). The number of indicators in each category in the current draft are (i) 19, (ii) 12, (iii) 3 and (iv) 2. Thus, while our demonstration that compositional rarity is a key biotic quantity that must be conserved in order to protect regional biotas may be intuitive in hindsight, the weighting of area and composition in key management tools currently remains strongly skewed toward area-based measures. We therefore advocate for

[1] CBD, https://www.cbd.int/doc/c/abb5/591f/2e46096d3f0330b08ce87a45/wg2020-03-03-en.pdf
[2] https://www.cbd.int/sbstta/sbstta-24/post2020-indicators-en.pdf

increased prioritisation of descriptors of the internal structure of biodiversity when predicting impacts and setting conservation targets.

4 Materials and methods

4.1 Model description - Environmental heterogeneity and spatial structure

Environmental heterogeneity (Fig. 1A) was modelled by assigning to each each site x a set of K independent random variables E_{kx} ($1 \leq k \leq K$), representing, e.g. the principal component(s) of abiotic environmental variation. The E_{kx} were sampled from spatially correlated Gaussian random fields (Adler, 2010) with mean 0, standard deviation 1, and correlation length 1.

Each species i in the model was allocated environmental optima E_{ik}^{opt} for each environmental variable k, sampled from uniform distributions in the range $[1.25 \cdot \min_x E_{xk}; 1.25 \cdot \max_x E_{xk}]$. The effective heterogeneity of the environment was controlled by varying the niche width parameter w. The intrinsic growth rate of the ith species in site x was computed as

$$R_{ix} = 1 - \left(\frac{2}{w}\right)^2 \sum_{k=1}^{K} (E_{kx} - E_{ik}^{opt})^2, \tag{1}$$

that is, to simplify parameterisation, we assumed identical niche widths for all species and all n independent components of environmental variation. For the random metacommunities we set n to either 1 or 2 and observed no major change in outcomes. For the fitted metacommunites, the first two principal components of the observed environmental variation were used.

The spatial network of sites (Fig. 1A) was modelled by randomly sampling Cartesian coordinates from a uniform distribution in the range $[0; \sqrt{N}]$, where N is the number of sites. As in previous studies (O'Sullivan et al., 2019, 2021a), we linked nearby sites using a Gabriel graph (Gabriel and Sokal, 1969).

4.2 Model description - Scaling local site area

An essential technical innovation that made this study possible is a technique we developed to model local differences in site area by scaling the intensity of local ecological interactions. The precise functional form of the area dependence of intra- and inter-specific ecological interactions is not currently known. In order to overcome this knowledge gap we assembled metacommunities of different total area a and measured the decay in effective

interaction strengths. Previous work has shown that regional scale interaction coefficients C_{ij} describing the interaction between species i and j can be estimated for models by perturbing the biomass of each species in turn and summing the impacts on all other species over the metacommunity (O'Sullivan et al. 2019 and S1 of the appendix for details). The resulting interaction matrix $\mathbf{C}$ captures the combined effects of differences in environmental preference, limited dispersal, indirect and direct interactions.

The decay in average interaction strength with area, which we refer to as the *competition area relation* (CAR), can be modelled by two power laws, one for inter-specific interactions, $\overline{C_{ij}} \propto a^{v}$ $(i \neq j)$, and one for intra-specific interactions, $\overline{C_{ii}} \propto a^{v'}$ (Fig.S1A). The exponents v and v' depend on model parameters, in particular the abiotic niche width w, which strongly affects species' range sizes (we found the effect of dispersal length to be weak by comparison). The CAR can be incorporated into the model dynamics in order to model variation in the local biomass (area) between sites by scaling the site-specific interaction matrix $\mathbf{A}_x$. Thus we model each site as an implicit sub-network of various nodes and scale the interactions to those expected for the corresponding (sub-)metacommunity.

Formally, if $\mathbf{A}_0$ is a hollow matrix with zeros on the diagonal, the scaled matrix is given by

$$\mathbf{A}_x = \lambda'_x \mathbf{I} + \lambda_x \mathbf{A}_0. \tag{2}$$

where $\mathbf{I}$ is the identity matrix, and

$$\lambda_x = \left(\frac{a_x}{a_0}\right)^{v}, \tag{3a}$$

$$\lambda'_x = \left(\frac{a_x}{a_0}\right)^{v'}. \tag{3b}$$

Here a_x is the area of the xth site and a_0 the area of a reference site with (unscaled) interaction matrix $\mathbf{A} = \mathbf{I} + \mathbf{A}_0$. We measure site area in units such that $a_0 = 1$. For random metacommunities, the a_x values were randomly assigned to sites x such that they covered the range from 1 to 30 biomass units in equal intervals. For fitted metacommunities, the a_x were extracted from the empirical species by site tables (see below). The exponents v and v' were set based on the relationships between the CAR and SAR found in simulations, assuming that within-site SARs are typically linear (see S1 of the appendix).

4.3 Model description - Metacommunity dynamics

Metacommunity dynamics were modelled using a spatial extension to the classic Lotka-Volterra community model (O'Sullivan et al., 2019, 2021a)

$$\frac{dB_{ix}}{dt} = \left(R_{ix} - \lambda'_x B_{ix} - \sum_{j=1}^{S} \lambda_x A_{ij} B_{jx} \right) B_{ix} + \sum_{y=1}^{N} B_{iy} D_{xy}, \tag{4}$$

where B_{ix} represents the biomass of the ith species in the xth site. R_{ix} are the intrinsic growth rates, which vary across the landscape.

For simplicity, the unscaled local interspecific interaction coefficients A_{ij} were sampled from a discrete distribution with $P(A_{ij} = 0.3) = 0.3$, and $A_{ij} = 0$ otherwise. D_{xy} are the elements of the spatial connectivity matrix describing the inter-site dispersal. Emigration and immigration rates are given by $D_{xx} = -e$, $D_{xy} = (e/k_y)\exp(-d_{xy}/\ell)$ for sites x, y connected by the spatial network, and $D_{xy} = 0$ otherwise; e is an emigration rate, d_{xy} the distance between sites x, y, and ℓ the characteristic dispersal length, which was systematically varied. The normalisation k_y represents the unweighted degree of the yth site.

4.4 Model description - Metacommunity assembly and removal experiments

Model metacommunities were assembled by iteratively generating species i with randomly sampled E_{ik}^{opt} and A_{ij}, and introducing them as invaders to the model at low biomass. Dynamics were then simulated over 500 unit times using Eq. 4. Before each invasion, the metacommunity was scanned for any species whose biomass had fallen below the extinction threshold of 10^{-4} biomass units in all sites. These species were considered regionally extinct and removed from the model. Metacommunity models assembled in this fashion eventually saturated with respect to both average local and regional species richness (O'Sullivan et al., 2019) due to the onset of ecological structural instability (Rossberg, 2013). After saturation is reached, each invasion generates, on average, a single extinction.

Following pilot studies metacommunities were assembled with w assigned 10 values logarithmically spaced in the range $0.5 \leq w \leq 15$. The parameter ℓ was logarithmically spaced in the range $2 \cdot 10^{-2} \leq w \leq 10$, again with 10 distinct values included. In both cases, a couple extreme values were excluded from the analysis either because beyond a threshold, no further change in outcomes was observed or, in the case of ℓ, because very small values lead to numerical errors in the simulation.

After assembling model metacommunities of 20 sites to regional diversity limits, each site in turn, and all

associated edges were systematically removed and the simulation advanced over 10^4 unit times. For completeness, the dispersal matrix $\mathbf{D}$ was updated to reflect the new degree distributions.

References

Robert J. Adler. *The Geometry of Random Fields*. SIAM, 2010.

Andrew P Allen and Ethan P White. Effects of range size on species–area relationships. *Evolutionary Ecology Research*, 5(4):493–499, 2003.

Xavier Benito, Sherilyn C Fritz, Miriam Steinitz-Kannan, Maria I Vélez, and Michael M McGlue. Lake regionalization and diatom metacommunity structuring in tropical south america. *Ecology and evolution*, 8(16): 7865–7878, 2018.

J. Roger Bray and John T. Curtis. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

Jonathan M Chase, Alienor Jeliazkov, Emma Ladouceur, and Duarte S Viana. Biodiversity conservation through the lens of metacommunity ecology. *Annals of the New York Academy of Sciences*, 1469(1):86–104, 2020.

Marcela Eugenia da Silva Cáceres, André Aptroot, and Robert Lücking. Lichen fungi in the atlantic rain forest of northeast brazil: the relationship of species richness with habitat diversity and conservation status. *Brazilian Journal of Botany*, 40(1):145–156, 2017.

Jeffrey J Danielson and Dean B Gesch. *Global multi-resolution terrain elevation data 2010 (GMTED2010)*. US Department of the Interior, US Geological Survey, 2011.

Sandra Myrna Díaz, Josef Settele, Eduardo Brondízio, Hien Ngo, Maximilien Guèze, John Agard, Almut Arneth, Patricia Balvanera, Kate Brauman, Stuart Butchart, et al. The global assessment report on biodiversity and ecosystem services: Summary for policy makers. 2019.

Ludmilla Figueiredo, Jochen Krauss, Ingolf Steffan-Dewenter, and Juliano Sarmento Cabral. Understanding extinction debts: spatio–temporal scales, mechanisms and a roadmap for future research. *Ecography*, 42(12): 1973–1990, 2019.

Nicola L Foster, Andrew Foggo, and Kerry L Howell. Using species-area relationships to inform baseline conservation targets for the deep north east atlantic. *PLoS one*, 8(3):e58941, 2013.

K. Ruben Gabriel and Robert R. Sokal. A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18(3):259–278, 1969.

Benjamin Gilbert, Tyler D. Tunney, Kevin S. McCann, John P. DeLong, David A. Vasseur, Van Savage, Jonathan B. Shurin, Anthony I. Dell, Brandon T. Barton, and Christopher DG Harley. A Bioenergetic Framework for the Temperature Dependence of Trophic Interactions. *Ecology Letters*, 17(8):902–914, 2014.

G Huffman, D Bolvin, D Braithwaite, K Hsu, R Joyce, and P Xie. Integrated multi-satellite retrievals for gpm (imerg), version 4.4. *NASA's Precipitation Processing Center*, 2014. URL `ftp://arthurhou.pps.eosdis.nasa.gov/gpmdata/`.

G Hulley and S Hook. Mod21a2 modis/terra land surface temperature/3-band emissivity 8-day l3 global 1km sin grid v006. *NASA EOSDIS Land Processes DAAC*, 2017. doi: https://doi.org/10.5067/MODIS/MOD21A2.006.

Carles Ibáñez, Carles Alcaraz, Nuno Caiola, Albert Rovira, Rosa Trobajo, Miguel Alonso, Concha Duran, Pere J Jiménez, Antoni Munné, and Narcís Prat. Regime shift from phytoplankton to macrophyte dominance in a large river: top-down versus bottom-up effects. *Science of the Total Environment*, 416:314–322, 2012.

Mikko Kuussaari, Riccardo Bommarco, Risto K Heikkinen, Aveliina Helm, Jochen Krauss, Regina Lindborg, Erik Öckinger, Meelis Pärtel, Joan Pino, Ferran Rodà, et al. Extinction debt: a challenge for biodiversity conservation. *Trends in ecology & evolution*, 24(10):564–571, 2009.

George Livingston, Miguel Matias, Vincent Calcagno, Claire Barbera, Marine Combe, Mathew A Leibold, and Nicolas Mouquet. Competition–colonization dynamics in experimental bacterial metacommunities. *Nature communications*, 3(1):1–8, 2012.

Melodie McGeoch and Kevin J Gaston. Occupancy frequency distributions: patterns, artefacts and mechanisms. *Biological Reviews*, 77(3):311–331, 2002.

Nicolas Mouquet, Birte Matthiessen, Tom Miller, and Andrew Gonzalez. Extinction debt in source-sink metacommunities. *PLoS One*, 6(3):e17567, 2011.

Sean Nee and Robert M May. Dynamics of metapopulations: habitat destruction and competitive coexistence. *Journal of Animal Ecology*, pages 37–40, 1992.

Jacob D. O'Sullivan, Robert J. Knell, and Axel G. Rossberg. Metacommunity-scale biodiversity regulation and the self-organised emergence of macroecological patterns. *Ecology Letters*, 22(9):1428–1438, 2019. ISSN 1461-0248. doi: 10.1111/ele.13294.

Jacob D. O'Sullivan, J. Christopher D. Terry, and Axel G. Rossberg. Intrinsic ecological dynamics drive biodiversity turnover in model metacommunities. *Nature Communications*, 2021a. doi: 10.1101/2020.05.22.110262.

Jacob D. O'Sullivan, J. Christopher D. Terry, and Axel G. Rossberg. Temporally robust spatial structure in ecosystems explained by local biodiversity regulation. *In preparation*, 2021b.

Wilfred M. Post and Stuart L. Pimm. Community Assembly and Food Web Stability. *Mathematical Biosciences*, 64(2):169–192, 1983.

Walter V Reid et al. *Millennium ecosystem assessment*. 2005.

James Rosindell and Stephen J Cornell. Species–area relationships from a spatially explicit neutral model in an infinite landscape. *Ecology letters*, 10(7):586–595, 2007.

Axel G. Rossberg. *Food Webs and Biodiversity: Foundations, Models, Data*. John Wiley & Sons, 2013.

David Storch. The theory of the nested species–area relationship: geometric foundations of biodiversity scaling. *Journal of Vegetation Science*, 27(5):880–891, 2016.

Trisha L Swift and Susan J Hannon. Critical thresholds associated with habitat loss: a review of the concepts, evidence, and applications. *Biological reviews*, 85(1):35–53, 2010.

Shirin Taheri, David García-Callejas, and Miguel B Araújo. Discriminating climate, land-cover and random effects on species range dynamics. *Global Change Biology*, 27(6):1309–1317, 2021.

John Terborgh, Lawrence Lopez, Percy Nuñez, Madhu Rao, Ghazala Shahabuddin, Gabriela Orihuela, Mailen Riveros, Rafael Ascanio, Greg H Adler, Thomas D Lambert, et al. Ecological meltdown in predator-free forest fragments. *Science*, 294(5548):1923–1926, 2001.

Chris D Thomas, Alison Cameron, Rhys E Green, Michel Bakkenes, Linda J Beaumont, Yvonne C Collingham, Barend FN Erasmus, Marinez Ferreira De Siqueira, Alan Grainger, Lee Hannah, et al. Extinction risk from climate change. *Nature*, 427(6970):145–148, 2004.

Samuel ED Thompson, Ryan A Chisholm, and James Rosindell. Characterising extinction debt following habitat fragmentation using neutral theory. *Ecology letters*, 22(12):2087–2096, 2019.

David Tilman, Robert M May, Clarence L Lehman, and Martin A Nowak. Habitat destruction and the extinction debt. *Nature*, 371(6492):65–66, 1994.

Kerrie A Wilson, Josie Carwardine, and Hugh P Possingham. Setting conservation priorities. *Annals of the New York Academy of Sciences*, 1162(1):237–264, 2009.

Appendix

S1 Harvesting experiment and the CAR

Following O'Sullivan et al. (2019) we computed an $S \times S$ spatially unresolved competition matrix $\mathbf{C}$ that summarised biotic interactions at the regional scale. For this we constructed a spatially unresolved Lotka-Volterra system,

$$\frac{dB_i}{dt} = \left(\rho_i - \sum_{j=1}^{S} C_{ij} B_i \right) \cdot B_i, \tag{S1}$$

where B_i represents total biomass of species i, as an approximation of the spatially resolved model. The regional scale interaction coefficients C_{ij} were then estimated using a computational harvesting experiment, which assumes that interaction strengths can be inferred from the changes in regional abundances that result from controlled changes in the regional abundances of harvested species (Gilbert et al., 2014).

For the unresolved model, Eq. S1, the harvesting of species i at a rate h, represented by substituting $\rho_i - h$ for ρ_i in Eq. S1, produces a shift in the equilibrium biomasses given by

$$\sum_y \left(B_{jy} - B^*_{jy} \right) = \Delta B_j = -C^{-1}_{ij} h. \tag{S2}$$

where $\mathbf{B}^*$ represents the unharvested equilibrium. Near an equilibrium B^*_{ix} of the full metacommunity model,

$$\frac{dB_{ix}}{dt} = \left(\sum_{jy} J_{ixjy} (B_{jy} - B^*_{jy}) \right) - h B^*_{ix}, \quad \text{and} \tag{S3a}$$

$$\frac{dB_{kx}}{dt} = \left(\sum_{jy} J_{kxjy} (B_{jy} - B^*_{jy}) \right) \quad \text{for } k \neq i\ . \tag{S3b}$$

where h is the harvesting rate, and $\mathbf{J}$ is the numerically estimated $SN \times SN$ Jacobian of the full metacommunity model. We vectorize the matrix $\mathbf{B}$ (denoted $\vec{\mathbf{B}}$) to match the dimensionality of the spatially resolved Jacobian, and the write the equilibrium condition for Eq. S3a as

$$\mathbf{J}(\vec{\mathbf{B}} - \vec{\mathbf{B}}^*) - \vec{\mathbf{H}} = 0, \tag{S4}$$

where the elements of vector $\vec{\mathbf{H}}$ are hb^*_{jy} for $j = i$ and 0 otherwise. From this, we obtain

$$h^{-1}\left(\vec{\mathbf{B}} - \vec{\mathbf{B}}^*\right) = \mathbf{J}^{-1}h^{-1}\vec{\mathbf{H}}. \tag{S5}$$

The left hand side of Eq. S5 represents the local change in biomasses due to the harvesting of the focal species i per unit h. From Eq. S5 we compute the change in total biomass $\Delta B_j = \sum_x B_{jx} - B^*_{jx}$ for each species j, which gives row i of $\mathbf{C}^{-1}$ (Eq. S2). Iterating over all species $i = 1 \ldots S$, we computed $\mathbf{C}^{-1}$ and from this the spatially unresolved interaction matrix $\mathbf{C}$. This method requires that the metacommunity is at equilibrium with respect to biomass. In practice, this assumption is violated by the onset of autonomous compositional turnover for large N. For this reason we restricted our analysis to a range of N for which autonomous turnover is absent or low.

The approximate functional form of the CAR and the scaling relationships between v, v', z and w can be studied by assembling metacommunities of various numbers of sites (total area) and computing the mean regional-scale interaction coefficients $\overline{C_{ij}}$ using the harvesting approach (Figs. S1A). The exponents v and v' depend on model parameters, in particular the abiotic niche width w, which strongly affects species' range sizes (we found the effect of dispersal length to be weak by comparison). Future developments in the theory of LV metacommunities are likely to reveal close relationships between the CAR, the SAR ($S \propto a^z$) and the effective heterogeneity of the landscape (here determined mainly by w). In lieu of analytic results, we report the statistical relationships between these key quantities in Figs. S1B-D. The exponent z decays with abiotic niche width w, suggesting greater abiotic heterogeneity (smaller niche width) results in a steeper SAR. This is consistent with previous results (O'Sullivan et al., 2019) and qualitatively matches the observation of a steep phase in the empirical SAR at the highest spatial scales (greatest effective heterogeneity) (Allen and White, 2003; Rosindell and Cornell, 2007; Storch, 2016). The ratio of the exponents of the CAR $v'/v \geq 1$ holds over the examined parameter space and further that $v'/v \approx 1 - z$. This suggests a pivotal role for the ratio between the area dependence of inter- and intra-specific interaction coefficients in determining the accumulation of species with area, and thereby the exponents of the SAR. Finally, the rate of decay of the intra-specific interaction coefficient, determined by v', and abiotic niche width w are negatively correlated. When abiotic niche widths are very small, intra-specific interaction coefficients are largely insensitive to total area (v' approaches 0), since species ranges are primarily limited by the environment. In contrast, when niche widths are large and species distributions can encompass much of the landscape, effective intra-specific interactions decay more rapidly with total area (v' approaches v, Fig. S1A). The linear associations revealed in Figs. S1B-D may hold only for a subset of the parameter space of the model, however this subset

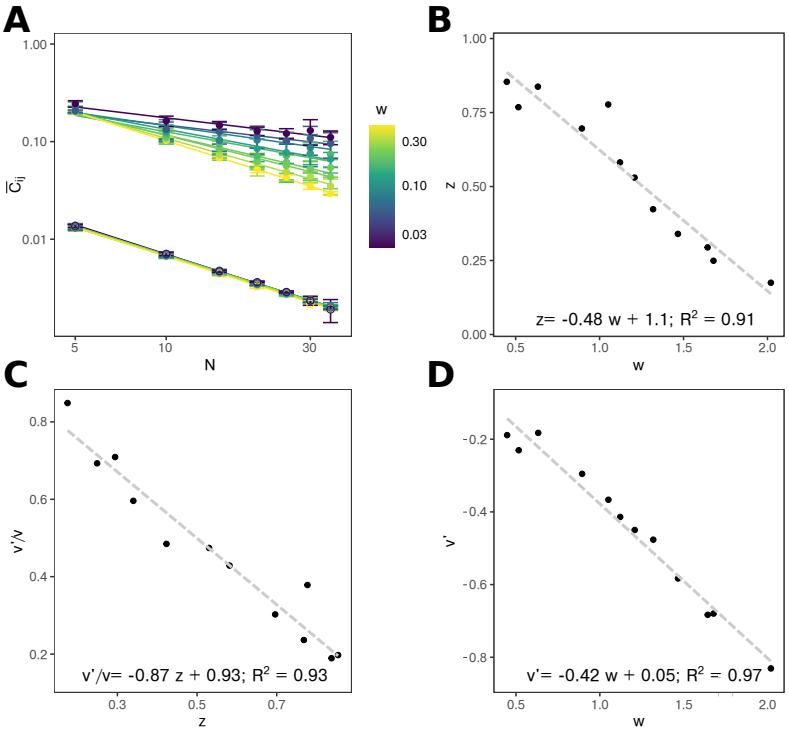

Figure S1: **Competition area relations in the LVMCM.** Assembling metacommunities with different total area reveals a power-law decay in average regional scale competitive interaction coefficients (A). We show the CAR for random metacommunities of $5 \leq N \leq 35$ sites for a range of niche widths w. The area dependence of mean inter-specific interaction strengths are largely unimpacted by niche width. In contrast there is a substantial effect of w on the exponent v' for intra-specific interactions. Linearly regressing the exponent z of the species-area relation (SAR) against niche width w (B), v'/v against z (C) and v' against w (D) offers some indication of the mechanisms through which effective heterogeneity impacts the SAR and inform the scaling of interactions in our model. In B-D, the exponents z, v and v' were computed by assembling metacommunities for given w and $N = 5, 10, 15, 20, 25, 30, 35$ sites in 10-fold replication, such that each point in B-D represents an regression across 70 independent simulations.

includes the parameter combinations that we find to best fit empirical data.

We assume that accumulation of diversity with area within sites is linear, that is, that $z = 1$. This simplification is biologically reasonable given the tendency for the initial phase of empirical SARs to be linear. From the relationships between the key quantities summarised in Fig. S1 we can therefore set the exponents of the CAR in this linear phase as $v = -0.59$, $v' = -0.04$.

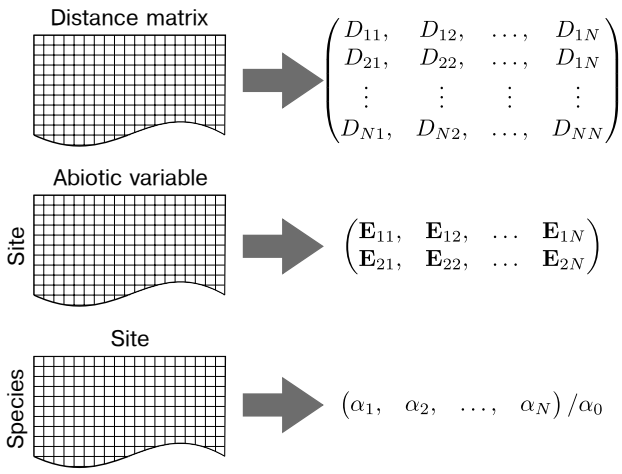

Figure S2: **Extracting data for fitting to empirical observation.** To fit the LVMCM to empirical observation we extract spatial coordinates, scaled to a square of longest side length $\sqrt{20}$ so that parameter values correspond to those used in the random metacommunity simulations. From the species by site presence-absence table, we extract the local site richness, presumed to scale with site area (quality). This is then used to generate λ, and λ' which scale the interaction coefficients $\mathbf{A}_{ij}$ so that the simulated metacommunity reproduces the observed distribution in α-diversity. From the available environmental metadata we extract the first two principle components and input these as the vectors $\mathbf{E}_k$ in the simulation.

S2 Detailed procedure for fitting to empirical data

Our approach to fitting the LVMCM to empirical species-by-site tables is as follows: We determined for each data set the Universal Transverse Mercator coordinates of each site, the two principal components of a multivariate distribution of environmental variables scaled to mean zero and unit variance, and the species richness recorded at each site (Fig. S2). The first two are input directly into the model. For better comparison to the parameters used above, we scaled the spatial network such that it occupies an arena of which the longest extension was $\sqrt{20}$. We again use a Gabriel graph to model the landscape connectivity, but note that alternative networks based on topographical data could be readily implemented.

To characterise environmental variability in dataset D, we followed the approach of the authors of the original study and used the NIPALS (nonlinear estimation by iterative partial least squares) algorithm, which allows ordination of datasets including missing values (Ibáñez et al., 2012), to extract the principal components of 14 environmental variables describing water chemistry, average climate, climate seasonality and variables describing the topography of landscape (see Benito et al. 2018 for details). Dataset L did not include quantitative environmental variables. We therefore extracted elevation, mean annual temperature, and mean annual precipitation data from publicly available remote sensing databases (further details given in the S4).

Because the ecological interactions between the observed species are insufficiently known, we randomly sampled the interaction coefficients $\mathbf{A}_{ij}$ from the same discrete distribution used in the random metacommunities. For a community model with no spatial extent for which $P(A_{ij} = 0.3) = 0.3$, and otherwise $A_{ij} = 0$, the onset of ecological structural instability occurs when $S_0 \approx 40$. From the SAR the effective area of a given site can be estimated using the equation

$$a_x = a_0 \left(\frac{S_x}{S_0} \right)^{\frac{1}{z}}, \tag{S6}$$

where we assume $z = 1$ as explained above. Substituting Eq. S6 into Eq. 2, the scaling of the competitive overlap matrix for the fitted metacommunity model can be written as

$$\mathbf{A}_x = \left(\frac{S_x}{S_0} \right)^{\frac{v}{z}} \mathbf{A}_0 + \left(\frac{S_x}{S_0} \right)^{\frac{v'}{z}} \mathbf{I}, \tag{S7}$$

where S_x, the observed richness at a given site, is extracted from the empirical data.

When assembling fitted metacommunity models, we again made the simplifying assumption that abiotic niche widths (w) are identical for both environmental variables and all species. Greater accuracy in model fitting could be achieved by estimating the typical abiotic niche widths of each variable independently, and by incorporting inter-specific variation in niche width.

For simulated LVMCM metacommunities with environmental and topographic distributions matching empirical observations we fit the parameter w based on the match between the observed and simulated occupancy frequency distribution (OFD) (McGeoch and Gaston, 2002), the proportion of species occupying a given number of sites (ranging from 1 for the site with the highest estimated conservation value to N the number of sites in the dataset). Since we find the parameter ℓ to have very little impact on the response to site removal, this was kept fixed at an intermediate value when fitting to observations. The OFD of the observation was quantitatively compared to the simulation by fitting to a pseudo-mechanistic patch occupancy approximation of the LVMCM, which is outlined in S3 of the appendix and will be explored in more detail in future work.

S3 Patch dynamics fitting model

We use a simple locally saturated patch occupancy model (LSPOM) as an intermediary for comparison of the occupancy frequency distributions (OFD) of LVMCM and empirical datasets. This is advantageous because the

LSPOM has only a single fitting parameter, called the 'mixing rate', and can be solved analytically. Thus, comparison between OFDs reduced to comparison of this single fitting parameter. We discuss motivation and validity of the LSPOM in detail elsewhere (O'Sullivan et al., 2021b). Here, we only provide a bare-bones statement of the model and its analytic solution.

The LSPOM conceptualizes a metacommunity as a spatially implicit species-by-site occupancy matrix describing an assemblage of S species occupying N patches, each capable of hosting n populations. This constraint on local richness reflects the fact that in each patch in the full LVMCM the number of co-existing species is limited by ecological structural stability constraints. Thus, each site can be thought to contain n 'slots', each of which can host a single local population. Each species present in the metacommunity occupies one or more slots, and the distribution of the numbers of slots occupied by species is the OFD. If most species occupy much fewer slots than there are patches (as we find here), then one can disregard, in a first instance, the constraint that a species cannot occupy more than one slot in the same patch.

At each time step of the model a new species invades the metacommunity from outside. Its initial site occupancy is set to 1, i.e. each new species colonises a single site in the metacommunity. Simultaneously, each extant local population colonises an additional site with a probability $m/(nN)$, where m is the mixing rate. Each colonisation event, either from within or from outside the metacommunity, produces a corresponding extirpation event of a randomly selected occupant of the colonised site in accordance with the assumption of local ecological saturation.

The steady state OFD for this process is such that, in a large metacommunity with a total of Z slots, the mean number of species occupying n slots is Zs_n, where for $m = 0$

$$s_n = \frac{1 - Q(n-1, \lambda - 1)}{n\lambda} \tag{S8}$$

and for $m > 0$

$$s_n = \frac{1}{n\lambda}\left[1 - \left(\frac{m}{\lambda + m}\right)^n - Q(n, \lambda - 1) + \exp\left(\frac{(\lambda-1)\lambda}{m}\right)\left(\frac{m}{\lambda+m}\right)^{n-1} Q\left(n, \frac{(\lambda-1)(\lambda+m)}{m}\right)\right], \tag{S9}$$

with $Q(a, x) = \Gamma(a, z)/\Gamma(z)$ denoting the regularised incomplete gamma function. Noting that the total number of species in this metacommunity is $Z\sum_n s_n$, the definition of λ translates to

$$\lambda - 1 = \frac{Z}{Z\sum_n s_n} = \frac{1}{\sum_n s_n}. \tag{S10}$$

This is solved numerically by adjusting λ. In Fig. S3 we show the average occupancy frequency of the steady state of 1000 spatially implicit patch dynamics models assembled according to the basic principles described above. The dashed line represents the predicted OFD after numerically fitting the mixing rate.

To find the best fitting abiotic niche width w for an empirical dataset, we estimated the mixing rate m for the empirical OFD, then, for model metacommunities with abiotic variables extracted from observation data as described above, we scanned assembled metacommunities with various combinations of the focal parameters (Fig. S4). The best fitting w was selected as that with emergent OFD which closest matched the mixing rate m of the observation. Note that, for simplicity we define the site occupancy with reference to source populations only. These are the local populations which persist following a long relaxation (10^4 unit times) after dispersal is switched off. In effect this assumes that only source populations are detected in the empirical data due to type II sampling errors. Alternative approaches to defining local site occupancy in simulations are available but typically require additional model assumptions or setting an arbitrary numerical threshold.

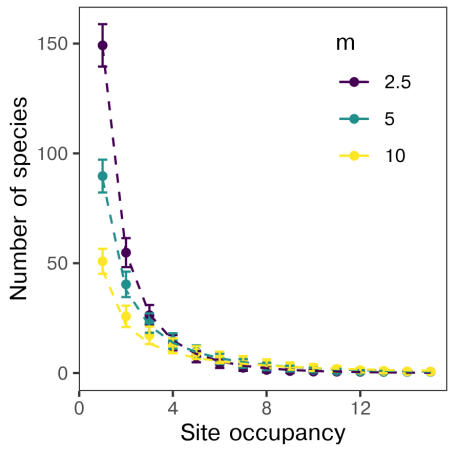

Figure S3: **The steady state OFD of the locally saturated patch dynamics model.** Points represent the average number of species in each site occupancy for 1000 realisations of the LSPOM described in the text for three different values of the mixing rate m. Dashed lines are the predicted OFDs computed by numerically solving Eq. (S10). Figure taken from O'Sullivan et al. (2021b).

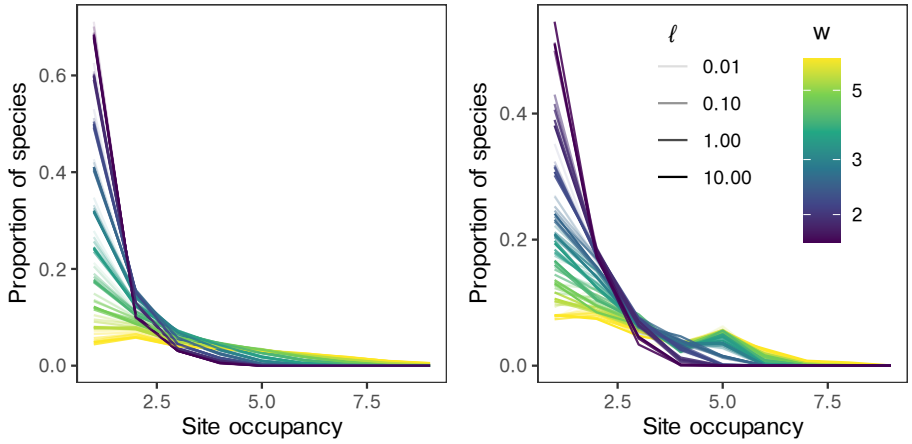

Figure S4: **The impact of parameter variation on the OFD for fitted models.** The average proportional occupancy frequency distributions for various values of w with landscape properties extracted from empirical datasets L and D. Colour represents abiotic niche width w.

S4 Remote sensing data collection

We extracted temperature, precipitation and elevation data for the 23 sites of the Brazilian Lichen-Fungi survey (da Silva Cáceres et al., 2017). The first two abiotic variable were time averaged over the 20 years from January 2000 to January 2020. We pooled nearby sites in order to reduce the dimensionality and therefore the computational load of the simulations. For these sites we took abiotic values, including spatial coordinates, as the between site averages.

Temperature data were collected from the MODIS satellite database (Hulley and Hook, 2017). For each site, sensor products were filtered by land surface temperature and emissivity to retrieve data from the MODIS thermal infrared bands. Precipitation data were collected from the IMERG satellite database (Huffman et al., 2014). Elevation data were taken from the GTOPO30 digital elevation model database (Danielson and Gesch, 2011) which provides data from collective satellite inputs, compiled by the U. S. Geological Survey.