



Beyond neutrality: disentangling the effects of species sorting and spurious correlations in community analysis

SYLVIE CLAPPE,^{1,2,4} STÉPHANE DRAY ^{1,2} AND PEDRO R. PERES-NETO³

¹Université de Lyon, F-69000 Lyon, France

²CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France

³Department of Biology, Concordia University, Montreal, Quebec H4B 1R6 Canada

Abstract. The methods of direct gradient analysis and variation partitioning are the most widely used frameworks to evaluate the contributions of species sorting to metacommunity structure. In many cases, however, species are also driven by spatial processes that are independent of environmental heterogeneity (e.g., neutral dynamics). As such, spatial autocorrelation can occur independently in both species (due to limited dispersal) and the environmental data, leading to spurious correlations between species distributions and the spatialized (i.e., spatially autocorrelated) environment. In these cases, the method of variation partitioning may present high Type I error rates (i.e., reject the null hypothesis more often than the pre-established critical level) and inflated estimates regarding the environmental component that is used to estimate the importance of species sorting. In this paper, we (1) demonstrate that metacommunities driven by neutral dynamics (via limited dispersal) alone or in combination with species sorting leads to inflated estimates and Type I error rates when testing for the importance of species sorting; and (2) propose a general and flexible new variation partitioning procedure to adjust for spurious contributions due to spatial autocorrelation from the environmental fraction. We used simulated metacommunity data driven by pure neutral, pure species sorting, and mixed (i.e., neutral + species sorting dynamics) processes to evaluate the performances of our new methodological framework. We also demonstrate the utility of the proposed framework with an empirical plant dataset in which we show that half of the variation initially due to the environment by the standard variation partitioning framework was due to spurious correlations.

Key words: direct gradient analysis; environmental effect; limited dispersal; metacommunity ecology; Moran spectral randomization; neutral dynamics; spatial autocorrelation; spatially constrained null model; variation partitioning.

INTRODUCTION

Determining which factors and processes underlie the variation in species compositions across local communities within large landscapes is a focal question in ecology (Cottenie 2005, Jones et al. 2008, Sharma et al. 2011, Wan et al. 2015). Among different frameworks for investigating the mechanisms governing species distributions in space, metacommunity ecology focuses on how dispersal interacts with local community assembly processes (e.g., niche preferences and biotic interactions) in determining how and which species are selected to compose local communities (Leibold et al. 2004). Species sorting via environmental filtering has been long recognized as a key mechanism underlying metacommunity structure in which variation in species compositions is driven by differential species responses to environmental heterogeneity among patches (Chase and Leibold 2003, Leibold and Chase 2017). Notwithstanding, it is also widely accepted that the spatial structure of metacommunities can be also influenced by other processes such as patch dynamics, mass-effect dynamics, neutral dynamics (Leibold et al. 2004), among other processes (e.g., historical biogeography; Leibold et al. 2010). Notably, neutral dynamics caused by stochastic events such as death, birth and limited dispersal can lead to spatial arrangements in the distribution of species within a

metacommunity that are as complex as the ones generated by species sorting (Bell 2001, Hubbell 2001). Because both mechanisms are likely to be important in many metacommunities, a major challenge in contemporary ecology has been to understand the relevance of species sorting to metacommunity structure that are beyond those expected under neutrality (Gravel et al. 2006, Paknia and Pfeiffer 2014).

Among the quantitative tools currently available to study metacommunities, variation partitioning (hereafter referred as to VP; Borcard et al. 1992, Peres-Neto et al. 2006) is the only one often assumed to be able to distinguish between species sorting and neutral dynamics (Gilbert and Lechowicz 2004, Cottenie 2005, Chang et al. 2013, Paknia and Pfeiffer 2014, Arellano et al. 2016). VP is a widely used quantitative framework in community ecology (i.e., more than 3,000 WoS citations for Borcard et al. 1992 and Peres-Neto et al. 2006) that allows decomposing the variation in species composition (or abundance) across local communities (i.e., beta-diversity) into the amount of variation explained by environmental (fraction [ab]) and spatial components (fraction [bc]). Independent contributions are also estimated and represent: non-spatially structured environmental variation (fraction [a]), pure spatial variation (fraction [c]), spatialized (i.e., spatially autocorrelated) variation of the environment (fraction [b]), and non-explained and non-spatialized residual variation (fraction [d]) (Fig. 1).

Pure spatial variation (fraction [c]) cannot be directly linked to a unique process as it may be due to stochastic neutral dynamics via dispersal limitation, missing spatialized

Manuscript received 14 September 2017; revised 12 March 2018; accepted 21 March 2018. Corresponding Editor: Helmut Hillebrand.

⁴E-mail: sylvie.clappe@gmail.com

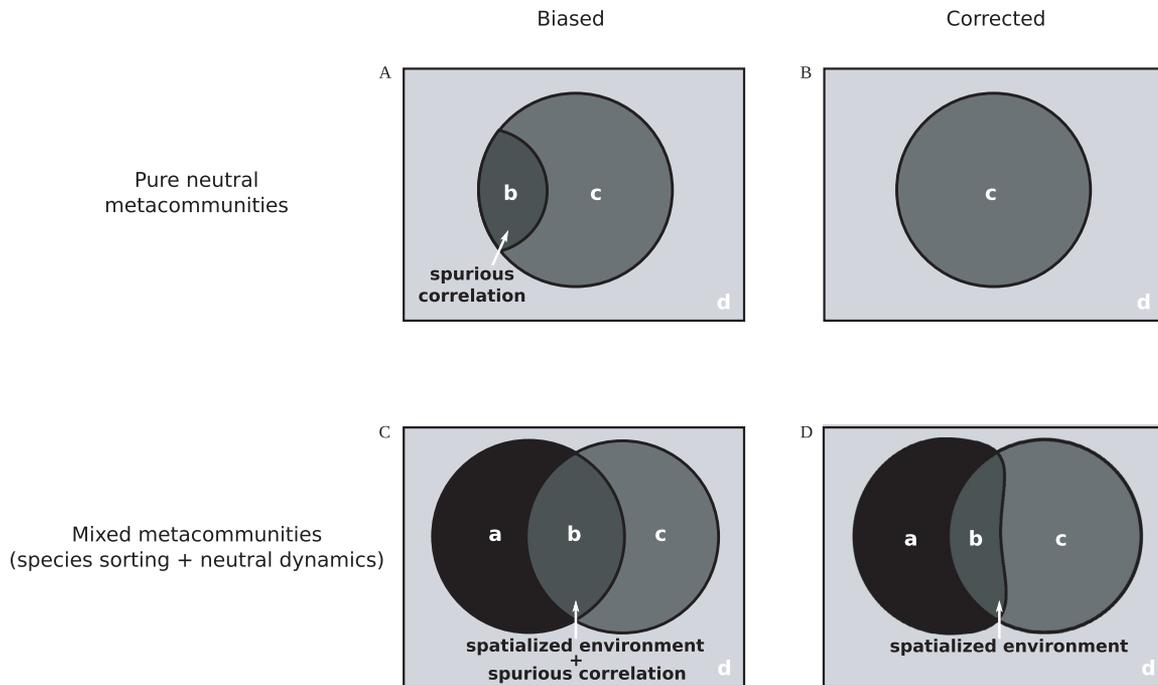


FIG. 1. Objective of the study: adjusting for spurious correlation in variation partitioning due to neutral dynamics. (A) The effects of spurious correlation on variation partitioning on metacommunities driven by neutral dynamics, leading to inflated fraction [b] (and [ab] as a consequence) but not [a]; (B) The effects of our procedure on variation partitioning on metacommunities driven by neutral dynamics in which only a spatial effect (fraction [c]) should be expected; (C) The effects of spurious correlation on variation partitioning on metacommunities driven by neutral dynamics and species sorting, leading to inflated fraction [b] (and [ab] as a consequence) but not [a]; (D) The effects of our procedure on variation partitioning on metacommunities driven by neutral dynamics and species sorting, leading to a reduced fraction [b] (and [ab] as a consequence). Note that the spatial fraction [bc] does not change as a function of our procedure, but fraction [c] increases given that fraction [b] decreases. (A–D) [a] non-spatially structured environmental variation; [b] spatialized environmental variation; [c] pure spatial variation; and [d] non-spatialized residual variation.

environmental predictors influencing species sorting and/or other types of spatial dynamics (e.g., biotic interactions). Additionally, spatialized-environmental variation (fraction [b]) may not necessarily represent the contribution of spatialized environment to species sorting given that fraction [b] can be inflated by spurious correlations between spatialized (measured) environmental variables and spatial variation in species distributions due to spatial processes such as neutral dynamics (Smith and Lundholm 2010, Fig. 1A, C). As such, VP can in principle only estimate the importance of species sorting by analyzing and testing its non-spatialized environmental contribution (fraction [a]; Peres-Neto and Legendre 2010, Logue et al. 2011). Ideally, however, one should estimate the contribution of the total environmental variation (i.e., fraction [ab]) containing both its non-spatial (fraction [a]) and spatial (fraction [b]) components (Fig. 1). This would allow to adjust the estimates of the environmental contribution and, in many cases, increase the power of detection of species sorting given that fraction [ab] is greater than fraction [a] alone. To do that, however, one needs to separate the variation in fraction [b] that is due to the spatialized measured environment and the variation due to neutral dynamics (Fig. 1A, C).

Although not often acknowledged, estimating the contribution of species sorting over neutral dynamics parallels the issues of statistical estimation and testing for species-environmental associations (Peres-Neto and Legendre 2010, Dray et al. 2012). Spatial autocorrelation in species distributions and their environments is extremely common (Houlihan

et al. 2007) and poses a challenge in statistical testing and estimation because it violates the assumption of data independence (Legendre 1993). VP and canonical analyses such as Redundancy Analysis and Canonical Correspondence Analysis (see Legendre and Legendre 2012 for a review) are regression-based models (Peres-Neto et al. 2006) whose estimates and tests are known to be affected by spatial autocorrelation (Dormann et al. 2007). Direct gradient analyses are a commonly used quantitative framework in ecology that are based on canonical analyses but do not consider spatialized variation. Statistical testing in direct gradient analyses is based on testing fraction [ab] and, as such, they need appropriate solutions to separate the environmental from neutral contributions. This latter issue has escaped the attention of ecologists who use direct gradient analysis to describe and test for the importance of environmental variation on species distributions. Finally, performing VP with both environmental and spatial predictors to estimate and test for species sorting (i.e., environmental contribution to species distributions, fraction [ab]) fails to control for biases due to spatial autocorrelation (Gilbert and Bennett 2010, Peres-Neto and Legendre 2010, Smith and Lundholm 2010). Given that VP have become the de facto framework to describe and assess the importance of species sorting, robust solutions to the potential biases generated by the spatial structure in species distributions and their environments need to be developed.

The goal of this paper is to provide a synthesis of the issues and implications of spatial autocorrelation in

community and metacommunity analyses as well as a solution to address current challenges when distinguishing between neutral dynamics and species sorting. More specifically we: (1) demonstrate via simulations the conditions in which spatial autocorrelation biases direct gradient analyses (i.e., canonical analyses) and VP in estimating and testing for the importance of species sorting in metacommunities; and (2) derive new estimates and statistical tests for assessing the role of species sorting in structuring species distributions and beta-diversity patterns. Finally, we demonstrate the utility of the proposed framework in an empirical plant data set in which we show that half of the variation initially due to the environment using the standard variation partitioning framework was due to spurious correlations.

In this study, we chose to use neutral dynamics as the source of spatial autocorrelation in species distributions and bias in canonical analysis given that neutrality has been the focus of much research regarding the mechanisms underlying the structure of metacommunities. Note, however, that our framework is able to deal with any mechanism generating spatial autocorrelation and thus tackle other sources of biases while testing and estimating the importance of species sorting in ecological communities.

METHODOLOGICAL FRAMEWORK

Variation partitioning

Variation partitioning (Borcard et al. 1992) is used in metacommunity studies to decompose the total variation of a species distribution matrix \mathbf{Y} (communities by species; abundance or presence-absence data) into four fractions of variation: [a] unique variation in \mathbf{Y} explained by the environment, [c] unique variation due to space, [b] shared variation between environment and space, and [d] non-explained (residual) variation. The approach is based on three separate sets of multi-response multivariate linear regressions with the response matrix \mathbf{Y} (i.e., sites by species matrix) (Peres-Neto et al. 2006): (1) a regression of \mathbf{Y} on a matrix \mathbf{X} containing environmental predictors to estimate [a]; (2) a regression of \mathbf{Y} on a matrix \mathbf{W} of spatial predictors to estimate [c]; (3) a regression of \mathbf{Y} on both \mathbf{X} and \mathbf{W} (all columns in \mathbf{X} and \mathbf{W} are used together as predictors) to estimate the total amount of variation explained by space and environment together (fraction [abc]). All fractions are then obtained via subtractions ([a] = [abc] - [bc]; [c] = [abc] - [ab]; [b] = [abc] - [a] - [c]; [d] = 1 - [abc]). Further computational details can be found in Peres-Neto et al. (2006). Note that decomposing a species distribution matrix \mathbf{Y} is equivalent to decomposing a beta-diversity matrix (community by community) into its environmental and spatial components (Legendre et al. 2005) and, as such, the issues discussed here also apply to the analysis of beta-diversity. The matrix \mathbf{W} (sites by spatial predictors) can, in principle, contain the geographical coordinates of sites, trend surface polynomials performed on these coordinates or spatial eigenvectors (Dray et al. 2012). However, geographical coordinates (simple or their polynomials) are only able to deal with simple large-scale spatial patterns whereas spatial eigenvectors are capable to model complex spatial structures occurring at both large and very fine scales

(Griffith and Peres-Neto 2006). As such, in this study we use Moran's Eigenvector Maps (MEM; Dray et al. 2006) to compose the spatial matrix \mathbf{W} which are routinely used in VP and known to be an efficient set of predictors to model simple and complex spatial patterns at multiple scales.

Fractions of variation are coefficients of determination (i.e., R^2) and need to be adjusted to account for the number of predictors as shown in Peres-Neto et al. (2006). Adjustments to correct for inflated variation explained in \mathbf{Y} due to irrelevant predictors is standard in multiple regression analysis (Ezekiel 1929) and as such also apply to canonical analysis and VP (Peres-Neto et al. 2006) which are simple extensions of multiple regressions. For instance, in the case of fraction [ab], the adjusted canonical coefficient of determination $\tilde{R}_{[ab]}^2$ is defined as:

$$\tilde{R}_{[ab]}^2 = 1 - \frac{n-1}{n-p-1} (1 - R_{[ab]}^2) \quad (1)$$

where n is the number of sites (i.e., communities), p is the number of environmental predictors in \mathbf{X} , and $R_{[ab]}^2$ is the coefficient of determination calculated on the basis of a multi-response multiple regression of species table \mathbf{Y} on the environmental matrix \mathbf{X} only. The same adjustments are performed on fractions [bc] (spatial) and [abc] (spatial and environmental combined) using the appropriate degrees of freedom involved prior to calculating unique and shared fractions via subtraction as detailed above.

The significance of fractions is estimated by a randomization test where the rows (i.e., sites) of the predictors (\mathbf{X} , \mathbf{W} or \mathbf{XW}) are permuted and appropriate F -statistics related to each fraction are calculated (Legendre and Legendre 2012). Note that in the case of a simple canonical analysis without a spatial matrix (i.e., direct gradient analysis), only fraction [ab] is tested. In the case of the environmental fraction $R_{[ab]}^2$ the null hypothesis H_0 is "the absence of species-environment relationships" (i.e., species sorting is not important). Under this null hypothesis, the expected value of the unadjusted coefficient of determination is $E(R_{[ab]}^2|H_0) = p/n - 1$ (Ezekiel 1929, Peres-Neto et al. 2006) so that, after some mathematical manipulations, the adjusted value (Eq. 1) becomes (Peres-Neto et al. 2006):

$$\tilde{R}_{[ab]}^2 = 1 - \frac{1}{1 - \frac{p}{n-1}} (1 - R_{[ab]}^2) \quad (2)$$

and therefore:

$$\tilde{R}_{[ab]}^2 = 1 - \frac{1}{1 - E(R_{[ab]}^2|H_0)} (1 - R_{[ab]}^2) \quad (3)$$

As such, the expected value under the null hypothesis of no relationship between species distributions and the (measured) environment $E(\tilde{R}_{[ab]}^2|H_0) = 0$. However, when species and environmental factors are independent but both spatially autocorrelated, $R_{[ab]}^2$ becomes inflated under the null hypotheses so that $E(\tilde{R}_{[ab]}^2|H_0) > 0$. This was shown by Smith and Lundholm (2010) using neutral dynamics to impose spatial autocorrelation on species distributions and

estimating $\tilde{R}_{[ab]}^2$ based on autocorrelated environments. Moreover, when \mathbf{X} and \mathbf{Y} are autocorrelated but not associated, permutation tests used in canonical analysis (i.e., direct gradient analysis) to test for $\tilde{R}_{[ab]}^2$ are also biased, having a rejection rate that is larger than the pre-established nominal significance level α (i.e., inflated Type I error rates; Peres-Neto and Legendre 2010). Here, we propose a new framework to estimate $E(R_{[ab]}^2|H_0)$, leading to an adjusted estimator and a valid statistical test for assessing the importance of species sorting on metacommunity dynamics that would be otherwise biased by spatial autocorrelation.

A spatially-constrained null model to estimate and remove biases due to spurious environmental effects

Similarly to Chase and Myers (2011), our approach is based on the development of a constrained null model to study the environmental contribution (i.e., fraction [ab]) in the face of the stochasticity induced by neutral dynamics (via limited dispersal). To do so, here we adapt a method recently proposed by Wagner and Dray (2015), namely the Moran Spectral Randomization (MSR) that allows generating random replicates of a data matrix while keeping its original inter-correlations and spatial structures (i.e., spatially-constrained randomization) to produce unbiased model fit estimates and statistical tests in the context of direct gradient analysis (i.e., canonical analysis) and VP. As such, MSR was used to produce a spatially constrained null model in which H_{0-MSR} is best described as “the absence of species-environment relationships given the spatial structure of the environmental variables”. In addition to conserving the original spatial autocorrelation in the environmental predictors, MSR also preserves the original correlation structure of the environment, a property important to assure that randomization tests with multiple predictors have correct Type I error rates even when predictors are non-spatially autocorrelated (Peres-Neto et al. 2006). Note that, if environmental predictors are not spatially structured, the resulting replicate is also random in respect to space. The MSR was used to generate replicates of randomized environmental matrices \mathbf{X} that were independent of the species distribution matrix \mathbf{Y} (in this study we used $n_{MSR} = 1,000$ replicates).

For every randomized environmental matrix (\mathbf{X}_{MSR} replicate), a VP for \mathbf{Y} on \mathbf{X}_{MSR} and \mathbf{W} (spatial predictors based on MEMs) was produced and corresponding fractions calculated. These computations were repeated for each \mathbf{X}_{MSR} replicate so that the expected distribution of $R_{[ab]}^2$ under H_{0-MSR} was produced. As such, our spatially-constrained randomization procedure allows estimating the statistical distribution of the $R_{[ab]}^2$ values expected under pure spurious species-environment relationships for a given species data matrix \mathbf{Y} . The estimated expected value under the null hypothesis $E(R_{[ab]}^2|H_{0-MSR})$ is then simply the average of the n_{MSR} random replicates of $R_{[ab]}^2$. Consequently, the estimated expected value of the coefficient of determination adjusted for spurious correlations is simply:

$$\tilde{R}_{[ab]MSR}^2 = 1 - \frac{1}{1 - E(R_{[ab]}^2|H_{0-MSR})} (1 - R_{[ab]}^2) \quad (4)$$

Likewise, fractions [a] and [b] are adjusted as follows:

$$\tilde{R}_{[b]MSR}^2 = 1 - \frac{1}{1 - E(R_{[b]}^2|H_{0-MSR})} (1 - R_{[b]}^2) \quad (5)$$

$$\tilde{R}_{[a]MSR}^2 = \tilde{R}_{[ab]MSR}^2 - \tilde{R}_{[b]MSR}^2 \quad (6)$$

where $E(R_{[b]}^2|H_{0-MSR})$ is the mean of the n_{MSR} random replicates of $R_{[b]}^2$. Detailed calculations to obtain fractions [bc], [abc], [c] and [d] using the MSR-based new VP procedure follow similar implementations and are shown in Appendix S1.

Fig. 1 represents a summary of the issues related to spurious autocorrelation and how our framework adjusts for them. In metacommunities driven solely by neutral dynamics, the total environmental contribution (fraction [ab]) is inflated because spurious correlations generated by independent spatial autocorrelation in environmental factors and species distributions inflate fraction [b] (Fig. 1A; see *Results* section for simulations demonstrating this issue). In these cases, our proposed MSR-based VP is expected to generate an unbiased estimation of fraction [b] (and thus [ab]) as zero (Fig. 1B). In metacommunities driven by both species sorting and neutral dynamics, fraction [b] and, as consequence, fraction [ab] are inflated due to the species dynamics driven by neutral dynamics (Fig. 1C). In these cases, our MSR-based spatially constrained null model adjusts for this spurious contribution in the total environmental fraction [ab] by reducing fraction [b] to the contribution of the spatially structured environment to species distributions independent on spatial autocorrelation (Fig. 1D).

Simulation study

We designed a simulation study to (1) describe how spatial autocorrelation affects the ability of direct gradient analysis (i.e., canonical analysis) and VP to disentangle the effects of species sorting and neutrality; (2) demonstrate the performance of our environmental spatially constrained approach (i.e., MSR-based VP) to generate an unbiased test for assessing the importance of species sorting over neutral dynamics. Traditionally, spatial predictors are selected via a forward model selection to compose the spatial matrix \mathbf{W} in VP (Blanchet et al. 2008, Smith and Lundholm 2010). Peres-Neto and Legendre (2010), however, showed that selected MEM subsets do not successfully remove spatial autocorrelation from model residuals, leading to inflated estimation and type I error rates of fraction [a] (i.e., species sorting). That is why, here, we used all positively spatially autocorrelated MEMs.

Metacommunities simulations

We used a dynamic spatially explicit individual-based model originally developed by Bell (2000, 2003, 2005) and implemented by Smith and Lundholm (2010). We modified the initial model to simulate the spatial distributions of $S = 50$ species on a $M \times M$ torus ($M = 15$) where each cell (patch) represented a local community ($n = 225$ local communities). We first simulated metacommunities solely composed of 50

generalist species (i.e., a form of neutrality in which all individuals have the same fitness) where the survival probabilities of all individuals were equal and independent of the environmental conditions in a particular patch. We then simulated metacommunities composed of 50 specialist species (i.e., here driven by species sorting dynamics) in which the survival probability of an individual was a function of the difference between its species niche optima and the local environmental conditions. A third type of metacommunity was composed by a mix of 25 species structured by neutral dynamics and 25 species structured by environmental sorting (mixed metacommunity).

Simulations to generate a single metacommunity were run for 5,000 generations to meet population equilibrium without incurring major species extinctions. For each type of dynamics (i.e., neutral, species sorting, and mixed), 1,000 metacommunities were generated. Simulations details are provided in Appendix S2.

Environment

In all simulations, we used 5 spatialized (i.e., spatially-autocorrelated) environmental variables that were generated following Dray (2011) in which an autoregressive model with parameter ρ varying between 0 (no spatial structure) and 0.8 (strong positive autocorrelation) was used to generate each variable independently. In the case of neutral dynamics only, we also considered an environmental matrix X containing 10 and 20 variables generated with a strong autocorrelation ($\rho = 0.8$).

Values of all environmental variables were kept constant through time within any given local community (i.e., cell).

Statistical analysis

Spatial predictors in table W were generated using Moran's Eigenvector Maps (MEM; Dray et al. 2006) of a spatial weighting defined by a "Queen's neighborhood" (i.e., 8 closest patches). Species matrices were Hellinger transformed prior to VP as it generates unbiased estimates of fractions for abundance data (Peres-Neto et al. 2006).

As noted in the *Introduction*, VP can, in principle, only estimate the importance of species sorting. As such, here we reported results only regarding the environmental fraction: (1) estimations of fractions [ab] (environmental + spatial), [a] (environmental variation independent of space) and [b] (spatialized environmental fraction); and (2) rejection rates for significance tests of fractions [ab] and [a]; note that fraction [b] is not testable under current statistical implementations. Type I error rates corresponded to the number of cases where the null hypothesis (i.e., no association between species distributions and the environment) was detected as false although true (i.e., under neutral dynamics only) divided by the total number of tests (i.e., 1,000 metacommunities). For the two other scenarios (species sorting metacommunities and mixed metacommunities), rejection of the null hypothesis corresponded to the estimation of statistical power as, in these cases, environment acted as a real driver of species distributions (i.e., species sorting was important). Significance tests were based on 999 randomizations and a significance level (α) of 0.05 was used in all tests.

All statistical analyses and simulations were performed with R 3.2.2 (R Core Team 2016). R function for the application of the new VP procedure using MSR as a spatially-constrained null model is available in packages *ade4* (Dray and Dufour 2007) and *adespatial*. An example of the new variation partitioning procedure (and related R code) on an empirical vegetation dataset is provided in Appendix S6.

RESULTS

The results based on metacommunities driven by neutral dynamics highlight the issue of how spatial autocorrelation impacts the estimation of fractions and associated hypotheses testing (Fig. 2 left panel). The total contribution of environment (fraction [ab]) becomes greater than the expected (zero) and it increases with the number of autocorrelated environmental variables used to assess the importance of species sorting (Fig. 2A). Type I error rates for fraction [ab] are extremely high leading to non-valid tests, implying strong species sorting for neutrally generated metacommunities. Notably, although the spatial predictors generate appropriate estimates and Type I error rates for the pure environmental contribution (fraction [a]; Fig. 2B), VP based on these predictors lead to inflated estimates of the spatial component of the environment (fraction [b]; Fig. 2C), explaining why fraction [ab] (but not [a]) is biased under neutral dynamics. Our spatially-constrained null model framework completely removes the biases in estimation due to spatial autocorrelation in fraction [b] and [ab] (and associated statistical testing of fraction [ab]), leaving fraction [a] as expected, i.e., zero (Fig. 2D–F). The results presented so far were based on high levels of spatial autocorrelation in the environment ($\rho = 0.8$). Using a lower level of autocorrelation ($\rho = 0.4$), estimates of fractions [ab] and [b] were also biased (Fig. 3A–C; Appendix S3: Tables S1–S3) though (as expected) relatively closer to zero in contrast to estimates based on high levels of spatial autocorrelation ($\rho = 0.8$). Note, however, that Type I error rates for fraction [ab] were equally high for both levels of spatial autocorrelation. As for high levels of spatial autocorrelation (Fig. 2), our spatially-constrained null model also led to adjusted estimates and Type I error rates for lower levels of autocorrelation for all fractions (Fig. 3D–F; Appendix S3: Tables S1–S3).

The simulations and results based on metacommunities solely driven by species sorting demonstrated that our spatially constrained null model tends to only slightly reduce fraction estimates by about less than 0.1% in most cases (Fig. 3J–L; Appendix S3) when compared to unadjusted results (Fig. 3G–I; Appendix S3: Tables S1–S3). Note, however, that statistical power is the same between unadjusted and adjusted versions.

For the case of mixed metacommunities, where some species are driven by neutral dynamics and others by species sorting, our spatially constrained null model produces estimates for fractions [ab] and [b] that are smaller (Fig. 4D–F; Appendix S3: Tables S1–S3) in contrast to the non-adjusted (biased) estimates (Fig. 4A–C; Appendix S3: Tables S1–S3), particularly in the cases of high spatial autocorrelation ($\rho = 0.8$). This reduction is obviously a result of the fact that the relationship between some species and their environments are spurious. As such, our procedure is also relevant

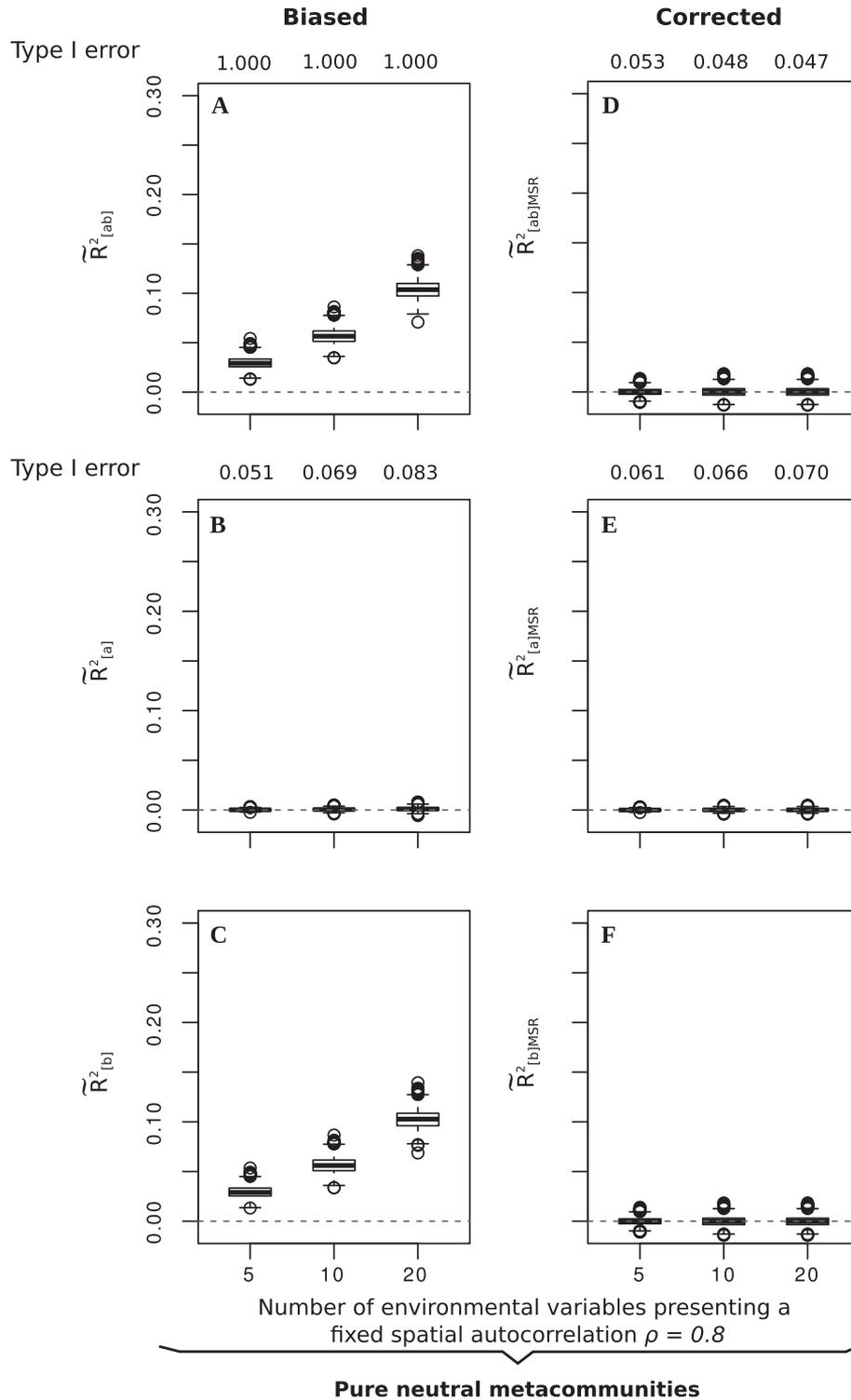


FIG. 2. Results of variation partitioning fractions [ab], [a] and [b] for neutral metacommunities as a function of the number of environmental variables. (A–C) observed \tilde{R}^2 of fractions [ab], [a], and [b] ($\tilde{R}^2_{[ab]}$; $\tilde{R}^2_{[a]}$; $\tilde{R}^2_{[b]}$ respectively) and their associated Type I error rates. (D–F) adjusted \tilde{R}^2 of fractions [ab], [a], and [b], respectively, ($\tilde{R}^2_{[ab]MSR}$; $\tilde{R}^2_{[a]MSR}$; $\tilde{R}^2_{[b]MSR}$) and their respective Type I error rates; (A–F) Note that fraction [b] cannot be tested and as a result Type I errors are not estimated. All fractions \tilde{R}^2 are reported with boxes representing the 25, 50 and 75 quartiles, the lower whisker is the first quartile minus 1.5 times the interquartile range, and the upper whisker is the third quartile plus 1.5 times the interquartile range. Extreme values (i.e., exceeding the whiskers) are plotted. Environment presents successively 5, 10 and 20 environmental variables whose spatial autocorrelation is fixed to $\rho = 0.8$. Dashed grey line (at zero) corresponds to the expected \tilde{R}^2 . Ordinates are scales from 0 to 0.3 to facilitate comparisons with Figs. 3 (left panel), 4.

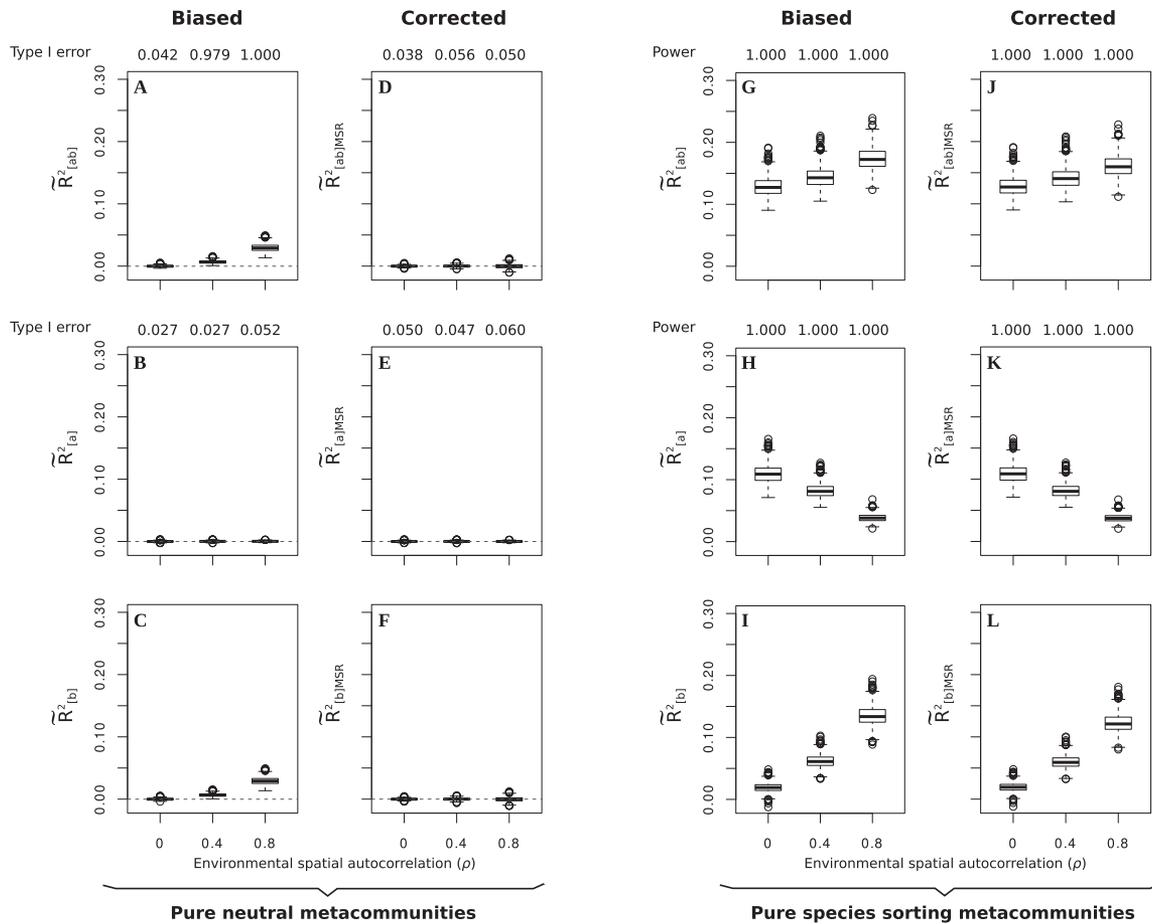


FIG. 3. Results of variation partitioning fractions [ab], [a] and [b] for metacommunities structured by neutral dynamics and species sorting. Left panel: results for neutral metacommunities. (A–C) estimation of total environmental contribution (fraction [ab]; $\tilde{R}^2_{[ab]}$), non-spatialized (fraction [a]; $\tilde{R}^2_{[a]}$), spatialized (fraction [b]; $\tilde{R}^2_{[b]}$), and their respective Type I error rates; (D–F) Estimates for adjusted fractions [ab] ($\tilde{R}^2_{[ab]MSR}$), [a] ($\tilde{R}^2_{[a]MSR}$), and [b] ($\tilde{R}^2_{[b]MSR}$) and their respective Type I error rates. Note that fraction [b] cannot be tested and as a result Type I errors are not estimated. Dashed grey line (at zero) corresponds to the expected variation explained for neutral metacommunities. Right panel: results for metacommunities structured solely by species sorting; order of figures are the same as for the left panel. (A–L) Boxplot specifics are explained in the legend of Fig. 2. Ordinates are scaled from 0 to 0.3 to facilitate comparisons with Figs. 3 (left panel), 2, 4.

when metacommunities are structured by mixes of processes (i.e., neutral and species sorting).

Applied to an empirical dataset, our new procedure allowed to substantially reduce estimates of fractions [ab] and [b] as [b] was reduced from 0.12 to 0.060 (Appendix S6: Table S1), thus adjusting for the effect of spatial autocorrelation on both species distributions and environmental data. This result demonstrates that the original total environmental fraction [ab], which contains fraction [b], was heavily driven by spurious spatial autocorrelations in these data.

DISCUSSION

The goal of this paper was twofold: (1) identify and summarize the issues regarding how the importance of species sorting is artificially inflated by neutral dynamics if spatial autocorrelation is not accounted for; and (2) propose a general and flexible procedure based on a spatially constrained null model to tackle these issues regarding parameter estimation (fractions) and inference (hypothesis testing). We centered on direct gradient analysis and VP because they are the

most used quantitative frameworks to infer species sorting in community ecology (Cottenie 2005, Soininen 2016, Leibold and Chase 2017), though our framework is general and relatively simple enough to be imbedded in any type of analysis (e.g., null models of species distributions, trait and phylogenetic community analysis, analyses of beta-diversity patterns).

VP applied to neutral metacommunities showed increasing inflations of fraction [ab] ($\tilde{R}^2_{[ab]}$) and associated Type I error rates as environmental spatial autocorrelation increases. In particular, Type I error rates became extremely high (equal to one) for a strong spatial autocorrelation in environmental features, thus systematically detecting a significant species sorting contribution when in fact there was none. Bias in estimation and inflation of Type I error rates rapidly increased with the number of environmental variables and can be easily explained by the fact that increasing the number of predictors provides greater chance to obtain spurious correlations between environment and species distribution data. It is now well recognized in ecology that when model residuals are spatially autocorrelated (i.e., when the model does not fully account for the spatial variation in the response \mathbf{Y} ; Dormann

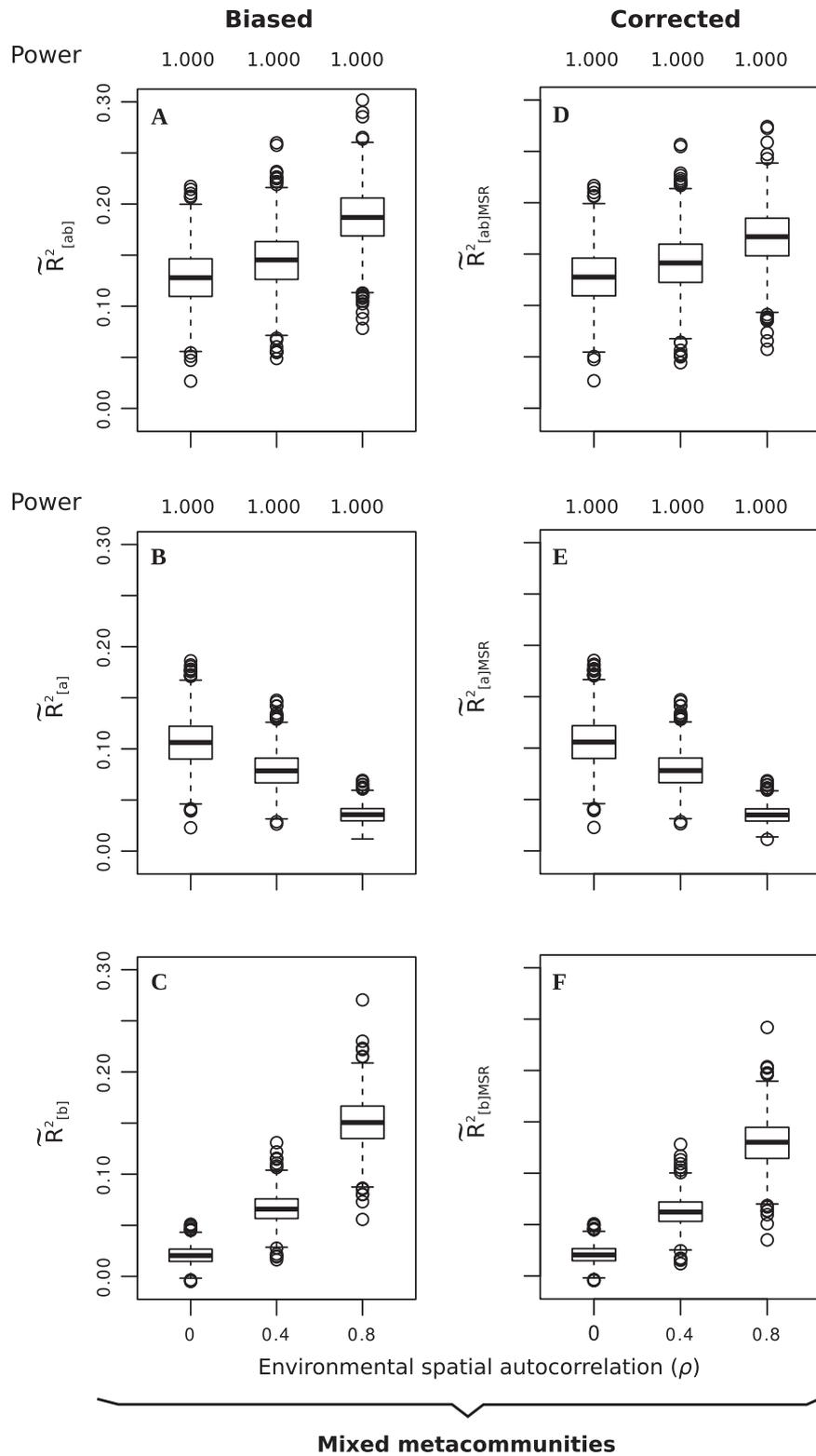


FIG. 4. Results of variation partitioning fractions [ab], [a] and [b] for mixed metacommunities. (A–C) observed estimates of fractions [ab] ($\tilde{R}_{[ab]}^2$), [a] ($\tilde{R}_{[a]}^2$) and [b] ($\tilde{R}_{[b]}^2$), and their respective power as a function of different levels of spatial autocorrelation in environmental variables. (D–F) Estimates for adjusted fractions [ab] ($\tilde{R}_{[ab]MSR}^2$), [a] ($\tilde{R}_{[a]MSR}^2$), and [b] ($\tilde{R}_{[b]MSR}^2$) to account for spatial autocorrelation. (A–F) Boxplot specifics are explained in the legend of Fig. 2. Ordinates are scaled from 0 to 0.3 to facilitate comparisons with Figs. 3 (left panel), 2.

et al. 2007, Dray et al. 2012) model slope estimates have much greater uncertainty (large confidence intervals) than expected under the assumption of residual independence. Although slopes are not biased per se (i.e., their average across sampling values equals the true population value; Dormann et al. 2007), their distribution is wider so that large slope values (positive and negative) will contribute to increase R^2 (i.e., fraction) values (adjusted or not), leading to inflated fraction [ab]. This is clearly the reason why under neutral dynamics, species sorting is detected when using only fraction [ab]. Note that studies only carrying out direct gradient analysis (i.e., canonical analysis) without considering VP need to be also aware of this issue and adjust estimates and tests accordingly using an appropriate spatially constrained such as the one developed here.

The main contribution of the present study is that we determined a way to consider spatial autocorrelation when testing for species sorting. As such, we propose new estimates of the environmental contribution (species sorting) via an adjustment to fraction [ab] instead as previous studies that only considered fraction [a] as a test of species sorting given that the assumption was that the entire fraction [b] had to be foregone because one could not separate the effects of spurious correlation from the spatial structure of the environment. Moreover, we expect, in many cases, our procedure will increase the statistical power for detecting species sorting given that fraction [ab] is greater than fraction [a] alone. Note that for communities structured by neutral dynamics only, even though the expected value of fraction [ab] ($\bar{R}_{[ab]}^2$) was properly estimated (i.e., average value equal to zero), its variance increased slightly with the level of environmental spatial autocorrelation. Given that this increase of variance was also present for the initial value of fraction [ab] ($\bar{R}_{[ab]}^2$) for communities generated solely by neutral dynamics and was absent from the others scenarios considered in this study (i.e., species sorting and mixed), it indicates that this issue is not linked to our new procedure but rather to the fact that increased spatial autocorrelation leads to great uncertainty (variance) in parameter estimates (model slopes; Dormann et al. 2007). This issue deserves further attention in future studies.

In this study, we developed a method and articulated a rationale in which fraction [ab] (and not fraction [a] alone) should be interpreted as the total (spatial and non-spatial) environmental contribution to species distributions (and related patterns such as beta-diversity). This is because our method separates the spatialized (i.e., spatially-autocorrelated) contribution of the environment (fraction [b]) that is due to spurious correlations from the contributions of the environment to species sorting. We did not consider the issues of estimating the spatial fractions ([bc] and [c]) given that spatial variation cannot be directly linked to a unique process and may be due to stochastic neutral dynamics via dispersal limitation, missing spatialized environmental predictors influencing species sorting and/or other types of spatial dynamics (e.g., biotic interactions). Note, however, that our framework needs to consider a spatial component (matrix **W** containing MEMs). Because most studies that uses VP apply a model selection (usually forward selection; e.g., Gilbert and Bennett 2010, Smith and Lundholm 2010) to reduce the number of spatial predictors and increase the

statistical power of detecting a significant fraction [a] (i.e., non-spatialized environmental contribution) and now fraction [ab] as per our study, we felt important to further discuss this issue here. In the original VP procedure, although estimates of fraction [a] are not biased under neutral dynamics (i.e., average values are zero; Figs. 2B, 3B) with (Appendix S4: Fig. S1B; Smith and Lundholm 2010) or without selection of spatial predictors (our study; Figs. 2B, 3B; Appendix S4: Fig. S1E), tests of fraction [a] have inflated Type I errors under neutral dynamics when subsets (i.e., selected) of spatial predictors are used (Appendix S4: Fig. S1B; Peres-Neto and Legendre 2010). This is related to the fact that variance of fraction [a] (as for fraction [ab]) is inflated under neutral dynamics leading to inflated Type I errors and, as such, a number of fraction [a] estimates are higher (due to the high variance issue as discussed above) than fraction [a] estimates from communities that are driven by species sorting, leading to inflated Type I error rates. This issue, however, is corrected by our framework in which both estimates and Type I error rates of fractions [a] and [b] (and [ab] as a consequence) are adjusted with and without forward selection for the spatial predictors (Appendix S4: Fig. S1G–I).

A novel finding in the present study is that estimates of fraction [b] under neutral dynamics are the ones inflated (Figs. 2C, 3C) even when all spatial predictors are used, leading to an inflation of fraction [ab]. Note that our spatially-constrained null model fully accounts for this bias (Figs. 2F, 3F). Given these results, we can now establish that fraction [b] has four origins:

1. Spatial structure of the measured environment;
2. Spatial structure of unmeasured spatialized environmental variables that influence species distributions and that are partially correlated with the measured environment;
3. Dispersal dynamics that track measured and/or unmeasured environmental variables (i.e., non-neutral). In other words, species can track suitable environments through dispersal (i.e., environmental tracking; Grönroos et al. 2013, Gianuca et al. 2017, Hill et al. 2017). As species are not considered equivalent in species sorting dynamics, even if they can reach many different environments through dispersal, they will only persist in those environments that are well suited. In our simulation model, this corresponds to the birth and death rate (weighted by survival probability and environmental selection) which determine if species establishment succeeds (i.e., species survive and reproduce in the community) or fails (i.e., species do not reproduce and die just after arriving in the community);
4. Dispersal dynamics via neutral dynamics (i.e., independent of environmental variation).

Therefore, our spatially constrained null model allows separating the contributions of spatially structured species sorting via environment (origins 1–3) in contrast to what is expected under neutral dynamics (4th origin; i.e., fraction [b] due to spurious relationships between autocorrelated species and environmental features, Appendix S3). Note, however, that our model cannot separate between origins 1 and 3. As discussed above, this separation allows to use

fraction [ab] and not only [a] to estimate and test for species sorting.

As species-environment relationships were defined as constant in our simulations, the increase of fraction [ab] ($\tilde{R}_{[ab]}^2$) as a function of increased levels of environmental spatial autocorrelation (Fig. 3G, J) in the case of species-sorting metacommunities can appear somewhat counter-intuitive. One could indeed expect fraction [ab] to be also constant across different levels of spatial autocorrelation of the environment. This result can be easily explained by the interaction between the spatial structure of the environment and dispersal (3rd origin of fraction [b]; i.e., dispersal dynamics tracking environmental features). Given that dispersal limitation was kept fixed across all simulation scenarios, as the spatial autocorrelation of the environment increases, the probability of finding suitable environments increases as well, leading to greater variation explained by the environment (i.e., fraction [ab]). Note that our spatially-constrained null model does not affect this pattern because this increase is truly related to species sorting (i.e., not induced via spurious correlation due to spatial autocorrelation).

The environmental fraction [ab] ($\tilde{R}_{[ab]}^2$) for mixed metacommunities was reduced by our framework (Appendix S3: Tables S2, S3). This can be explained by species undergoing neutral dynamics that inflates fraction [b] by adding the unwanted 4th origin (i.e., dispersal without species sorting; Appendix S5: Fig. S1). Our spatially constrained null model was able to estimate the spurious contributions in fraction [b] (Appendix S3) and produce adjusted estimates of the environmental contribution.

In this study, we demonstrated that the environmental contribution (fraction [ab]) in direct gradient analysis and VP could be strongly inflated in the case of spatially autocorrelated environment. As spatial autocorrelation is inherent to environmental predictors and given that both latter frameworks are extremely used in ecology to estimate environmental effect (Soininen 2014), it is most likely that the role of species sorting has been overemphasized in recent ecological studies. Our MSR-based VP produces adjusted estimates of the VP fractions that would be otherwise inflated by spatial autocorrelation. As such, it allows for a better evaluation of the importance of species sorting shaping beta diversity. More importantly, our new framework allows ecologists to revise their previous results without collecting new data.

Our study highlights the need for more sophisticated null models than those furnished by standard permutation procedures. Constrained null models offer a nice framework to consider the properties of the ecological data, allowing testing for more complex ecological hypotheses than the usual full random hypothesis. Consequently, they are promising tools to disentangle processes underlying metacommunity and community structure (Buschke et al. 2015, Mori et al. 2015, Brown et al. 2016).

ACKNOWLEDGMENTS

We thank the computing facilities of the CC LBBE/PRABI with which all simulations and statistical tests were performed. This work was supported by an NSERC (Natural Sciences and Engineering Research Council of Canada) Discovery Grant and the Canada Research Chair in Spatial Ecology and Biodiversity to P. R. Peres-

Neto. We would like to thank Dave Jenkins and an anonymous reviewer for their comments which greatly contributed to improve this manuscript.

LITERATURE CITED

- Arellano, G., J. S. Tello, P. M. Jørgensen, A. F. Fuentes, M. I. Loza, V. Torrez, and M. J. Macía. 2016. Disentangling environmental and spatial processes of community assembly in tropical forests from local to regional scales. *Oikos* 125:326–335.
- Bell, G. 2000. The distribution of abundance in neutral communities. *American Naturalist* 155:606–617.
- Bell, G. 2001. Neutral macroecology. *Science* 293:2413–2418.
- Bell, G. 2003. The interpretation of biological surveys. *Proceedings of the Royal Society of London B: Biological Sciences* 270:2531–2542.
- Bell, G. 2005. The co-distribution of species in relation to the neutral theory of community ecology. *Ecology* 86:1757–1770.
- Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Forward selection of explanatory variables. *Ecology* 89:2623–2632.
- Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. *Ecology* 73:1045–1055.
- Brown, C., J. B. Illian, and D. F. R. P. Burslem. 2016. Success of spatial statistics in determining underlying process in simulated plant communities. *Journal of Ecology* 104:160–172.
- Buschke, F. T., L. De Meester, L. Brendonck, and B. Vanschoenwinkel. 2015. Partitioning the variation in African vertebrate distributions into environmental and spatial components – exploring the link between ecology and biogeography. *Ecography* 38:450–461.
- Chang, L.-W., D. Zelený, C.-F. Li, S.-T. Chiu, and C.-F. Hsieh. 2013. Better environmental data may reverse conclusions about niche- and dispersal-based processes in community assembly. *Ecology* 94:2145–2151.
- Chase, J. M., and M. A. Leibold. 2003. *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press, Chicago, Illinois, USA.
- Chase, J. M., and J. A. Myers. 2011. Disentangling the importance of ecological niches from stochastic processes across scales. *Philosophical Transactions of the Royal Society B* 366:2351–2363.
- Cottenie, K. 2005. Integrating environmental and spatial processes in ecological community dynamics. *Ecology Letters* 8:1175–1182.
- Dormann, F. C., et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30:609–628.
- Dray, S. 2011. A new perspective about moran's coefficient: spatial autocorrelation as a linear regression problem. *Geographical Analysis* 43:127–141.
- Dray, S., and A.-B. Dufour. 2007. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* 22:1–20.
- Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling* 196:483–493.
- Dray, S., et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs* 82:257–275.
- Ezekiel, M. 1929. The application of theory of error to multiple and curvilinear correlation. *Proceeding Supplement. Journal of the American Statistical Association* 24:99–104.
- Gianuca, A. T., S. A. J. Declerck, P. Lemmens, and L. De Meester. 2017. Effects of dispersal and environmental heterogeneity on the replacement and nestedness components of β -diversity. *Ecology* 98:525–533.
- Gilbert, B., and J. R. Bennett. 2010. Partitioning variation in ecological communities: Do the numbers add up? *Journal of Applied Ecology* 47:1071–1082.
- Gilbert, B., and M. J. Lechowicz. 2004. Neutrality, niches, and dispersal in a temperate forest understory. *Proceedings of the National Academy of Sciences USA* 101:7651–7656.

- Gravel, D., C. D. Canham, M. Beaudet, and C. Messier. 2006. Reconciling niche and neutrality: the continuum hypothesis. *Ecology Letters* 9:399–409.
- Griffith, D. A., and P. R. Peres-Neto. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* 87:2603–2613.
- Grönroos, M., J. Heino, T. Siqueira, V. L. Landeiro, J. Kotanen, and L. M. Bini. 2013. Metacommunity structuring in stream networks: roles of dispersal mode, distance type, and regional environmental context. *Ecology and Evolution* 3:4473–4487.
- Hill, M. J., J. Heino, I. Thornhill, D. B. Ryves, and P. J. Wood. 2017. Effects of dispersal mode on the environmental and spatial correlates of nestedness and species turnover in pond communities. *Oikos* 126:1575–1585.
- Houlahan, J. E., et al. 2007. Compensatory dynamics are rare in natural ecological communities. *Proceedings of the National Academy of Sciences USA* 104:3273–3277.
- Hubbell, S. P. 2001. *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton, New Jersey, USA.
- Jones, M. M., H. Tuomisto, D. Borcard, P. Legendre, D. B. Clark, and P. C. Olivás. 2008. Explaining variation in tropical plant community composition: influence of environmental and spatial data quality. *Oecologia* 155:593–604.
- Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? *Ecology* 74:1659–1673.
- Legendre, P., D. Borcard, and P. R. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs* 75:435–450.
- Legendre, P., and L. F. J. Legendre. 2012. *Numerical ecology*. Elsevier, Amsterdam, The Netherlands.
- Leibold, M. A., and J. M. Chase. 2017. *Metacommunity ecology*. Princeton University Press, Princeton, New Jersey, USA.
- Leibold, M. A., E. P. Economo, and P. R. Peres-Neto. 2010. Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. *Ecology Letters* 13:1290–1299.
- Leibold, M. A., et al. 2004. The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters* 7:601–613.
- Logue, J. B., N. Mouquet, H. Peter, and H. Hillebrand. 2011. Empirical approaches to metacommunities: a review and comparison with theory. *Trends in Ecology & Evolution* 26:482–491.
- Mori, A. S., S. Fujii, R. Kitagawa, and D. Koide. 2015. Null model approaches to evaluating the relative role of different assembly processes in shaping ecological communities. *Oecologia* 178:261–273.
- Paknia, O., and M. Pfeiffer. 2014. Niche-based processes and temporal variation of environment drive beta diversity of ants (Hymenoptera: Formicidae) in dryland ecosystems of Iran. *Myrmecol News* 20:15–23.
- Peres-Neto, P. R., and P. Legendre. 2010. Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography* 19:174–184.
- Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* 87:2614–2625.
- R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Sharma, S., P. Legendre, M. De Cáceres, and D. Boisclair. 2011. The role of environmental and spatial processes in structuring native and non-native fish communities across thousands of lakes. *Ecography* 34:762–771.
- Smith, T. W., and J. T. Lundholm. 2010. Variation partitioning as a tool to distinguish between niche and neutral processes. *Ecography* 33:648–655.
- Soininen, J. 2014. A quantitative analysis of species sorting across organisms and ecosystems. *Ecology* 95:3284–3292.
- Soininen, J. 2016. Spatial structure in ecological communities – a quantitative analysis. *Oikos* 125:160–166.
- Wagner, H. H., and S. Dray. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. *Methods in Ecology and Evolution* 6:1169–1178.
- Wan, Y., L. Xu, J. Hu, C. Xu, A. Wan, S. An, and Y. Chen. 2015. The role of environmental and spatial processes in structuring stream macroinvertebrates communities in a large river basin. *Clean–Soil, Air, Water* 43:1633–1639.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at <http://onlinelibrary.wiley.com/doi/10.1002/ecy.2376/supinfo>