OIKOS

Research article

Integrating several analytical methods to assess strength of ecological processes behind metacommunity assembly

Ching-Lin Huang[1], **David Zelený**[1] and **Chia-Hao Chang-Yang**[2]

[1]Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan
[2]Department of Biological Sciences, National Sun Yat-Sen University, Kaohsiung, Taiwan

Correspondence: David Zelený (zeleny@ntu.edu.tw)

Oikos
2023: e10166
doi: 10.1111/oik.10166
Subject Editor: François Munoz
Editor-in-Chief: Pedro Peres-Neto
Accepted 23 October 2023

Understanding processes and mechanisms of how species assemble in a metacommunity is crucial for illuminating the factors that contribute to the maintenance of biodiversity and developing management decisions. Ecologists have proposed a number of analytical methods for identifying the effects of various ecological processes, but there is no consensus on the best approach. Our study extends the existing framework which synthesizes multiple analytical methods and incorporates community data across space and time to understand the underlying ecological processes. We extended this framework by 1) including null-model-based analytical methods; 2) defining metacommunity archetypes that illustrate extreme cases of metacommunities, to observe how well they can be distinguished by different summary statistics, 3) applying the extended framework to real-world vegetation data from a subtropical forest and interpreting the results, and 4) discussing the potential advantages, limitations, and future directions of applying this framework. We used a process-based metacommunity simulation model to generate a simulated community dataset and applied random forest (RF) approach to estimate the strength of ecological processes in the process-based model by considering the summary statistics calculated by the analytical methods as predictors. We also quantified the performance of the trained RF and applied it to estimate the strength of ecological processes in Fushan Forest Dynamics Plot. Our results demonstrate the framework's flexibility in incorporating different analytical methods and its generality to be applied to different community systems. We highlight its theoretical values in evaluating the performance of different statistics or indices in identifying ecological processes and its practical values in assessing the strength of ecological processes underlying real-world metacommunities. Future improvements should focus on synthesizing statistics that capture specific signals of ecological processes and evaluating the robustness of estimation against dataset complexity and incompleteness.

Keywords: community assembly, empirical data, Fushan Forest Dynamics Plot, metacommunity archetypes, process-based model, random forest, simulation

NORDIC SOCIETY OIKOS

www.oikosjournal.org

Introduction

Metacommunity theory is a comprehensive framework that integrates multiple ecological processes that affect species turnover among ecological communities (Leibold et al. 2004). Four high-level ecological processes simultaneously affect the species composition dynamics across space and time. At a local scale, competition and ecological drift determine the local community structure. Different species can interact with each other in a variety of ways, such as predation (Volterra 1928) or mutualism (Boucher et al. 1982). Particularly, competition is one of the species interactions that describes the negative impact among species caused by similarity in resource consumption or hierarchy in the fitness of the species (MacArthur and Levins 1967, Chesson 2000). In contrast, ecological drift is a stochastic process that describes population dynamics caused by demographical or environmental stochasticity in the death of an individual or the local extinction of a species within a community (Hubbell 2011). At the regional scale, environmental filtering and dispersal play the role in regulating species turnover (Leibold et al. 2004). The species that cannot tolerate given abiotic environmental conditions will be excluded from the local community (Grinnell 1917, Kraft et al. 2015). Different dispersal abilities of the species may alleviate or facilitate the migration of propagules between local communities (MacArthur and Wilson 1967, Tilman 1997, Mouquet and Loreau 2003). These ecological processes may vary in strength, interact with each other, and simultaneously drive the species turnover across spatial and temporal scales (Thompson et al. 2020). 'Metacommunity archetypes' are sometimes used to describe typical or extreme cases of metacommunities, each focusing on the effects of different ecological processes (Leibold et al. 2004, Brown et al. 2017, Leibold and Chase 2017) (Table 1).

Understanding processes and mechanisms of how species assemble in a metacommunity is of great interest to both theoretical and applied ecologists. By identifying the strength of processes underlying community assembly, we can describe the factors driving species turnover and the mechanisms responsible for the maintenance of biodiversity. Furthermore, predicting the future dynamics of the metacommunity structure under the pressure of anthropogenic activities and climate change has already become an essential topic in conservation (Clark et al. 2001, Evans 2012, Chase et al. 2020, McFadden et al. 2023). For example, studies focused on the relationship between anthropogenic factors (or climatic conditions) and the strength of ecological processes across metacommunities help to identify the mechanisms underlying the loss of ecosystem services, which in turn can improve management decisions to optimize these ecosystem services (Hodgson and Halpern 2019, Chase et al. 2020, McFadden et al. 2023). How to effectively estimate the strength of ecological processes underlying the observed

Table 1. Definition of the terms used in our study.

Term	Definition
Species turnover	The changes in species composition across space and time
Metacommunity	A metacommunity is a set of local communities linked by the dispersal of multiple interacting species (Wilson 1992, Leibold et al. 2004). In our study, the observed metacommunity is defined as the metacommunity in the field, and the simulated metacommunity is the metacommunity generated by a simulation model
Ecological process	The force that drives the species turnover across space and time. In the metacommunity framework, four high-level processes are proposed to drive species turnover: competition, ecological drift, environmental filtering and dispersal (Leibold and Chase 2017)
Metacommunity archetypes	Four metacommunity archetypes are the different perspectives to study the species turnover within the metacommunity (Leibold et al. 2004, Leibold and Chase 2017)
Patch dynamics (PD)	Competition hierarchy and tradeoff in competition and colonization ability among species determine the local extinction and colonization of the species and further cause the species turnover within the metacommunity
Species sorting (SS)	Species interactions and demographic differences among species, which are associated with the environmental heterogeneity across space and time, are the main drivers of the species turnover within the metacommunity
Neutral dynamics (ND)	Species in the metacommunity are demographically similar. Dispersal limitation among local communities and ecological drift caused by the demographic stochasticity within the local patches are the main drivers of species turnover
Mass effect (ME)	The species composition is influenced not only by species interactions and demographic differences resulting from environmental heterogeneity, but also by the strong dispersal flow of propagules. The strong dispersal ability of species may prevent the local population of less competitive species from being out-competed and maintain their presence in the community
Analytical methods	The analytical methods discussed in our study are the statistical approaches that summarize the ecological information based on the ecological community data by multiple summary statistics. The summary statistics may identify the influence of certain ecological processes to the species turnover. For example, beta-diversity variation partitioning is one of the analytical methods that uses constrained ordination to relate species composition with environmental and geographical variables. The variation partitions explained by environmental and geographical variables are the summary statistics of this method, which aim to quantify the relative importance of environmental filtering and dispersal
Process-based simulation models	The process-based approach explicitly mimics the effect of ecological processes on the state of the system, e.g. population size. The parameters in the process-based model may regulate the strength of the ecological processes. By modifying the model parameters, we may generate different patterns of population dynamics in a simulated metacommunity

metacommunities thereby becomes an urgent question with direct relevance to the coexistence of humans and nature.

Since the mid-20th century, ecologists have proposed a number of analytical methods that aimed to identify the presence (or quantify the relative importance) of various ecological processes based on species composition data (MacArthur 1958, Diamond 1975). More recently, ecologists started to integrate multiple types of data, e.g. information on environmental conditions, functional traits, and phylogeny, and proposed analytical methods that can summarize ecological information from the observed metacommunity by multiple summary statistics. These summary statistics serve the purpose of quantifying the effect of certain ecological processes on the species turnover or variation in functional traits and phylogeny structure (Fig. 1A).

Null model is one of the approaches that compares the observed community data to the theoretical expectation under the null hypothesis, created by data randomization (Gotelli and McGill 2006, Gotelli and Ulrich 2012). Data randomization is aimed to remove the effect of certain assembly processes within the metacommunity. The presence of certain ecological processes is then confirmed if the observed data is considerably different from the randomized one, i.e. the null hypothesis is rejected. Without information on the habitat conditions, studying the clustering or over-dispersion of the species composition may disentangle the effect of environmental filtering and competition on species turnover among local communities (Diamond 1975, Connor and Simberloff 1979, Chase and Myers 2011). On the other hand, testing the convergence or the divergence of the functional traits and phylogenetic composition between communities may identify whether the observed metacommunity is mainly driven by environmental filtering or competition (Mayfield and Levine 2010, Borics et al. 2020). The null-model approach may also quantify the relative importance of niche and dispersal processes underlying the metacommunity based on species composition and differences in species phylogeny or functional traits (Stegen et al. 2013, Ford and Roberts 2020), or only based on species composition (Gibert and Escarguel 2019, Vilmi et al. 2021). Environmental and geographical data are not necessary for these null-model-based analytical methods. Ecologists may identify the assembly processes without collecting such data. However, it is difficult to evaluate whether the algorithm of randomization correctly diminishes the effect of certain ecological processes. The mismatch between the algorithm and the null hypothesis may result in the wrong type I error rate of the hypothesis testing (Molina and Stone 2020).

Other analytical methods are based on the summary statistics derived from the relationship between different types of data, e.g. canonical analysis and hierarchical joint species distribution approach. In the canonical analysis, species composition is regressed on environmental and geographical attributes (Borcard et al. 1992), or even functional traits and phylogeny structure within the local communities (Sîrbu et al. 2021). We will refer to this approach as beta-diversity variation partitioning in the following text. The relative importance of environmental filtering and dispersal is then estimated by the amount of compositional variation explained by environmental and geographical data (Cottenie 2005). The hierarchical joint species distribution approach uses a hierarchical generalized linear model to relate species composition with environment, species traits, and phylogeny. In the hierarchical modeling of species community (HMSC) proposed by Ovaskainen et al. (2017), the random effects of biotic interaction, and spatial and temporal autocorrelation are also considered in the model. The effect of environmental filtering, biotic interactions, and random processes on the species turnover may be quantified by the variation partitioning among the explanatory variables in the hierarchical generalized linear model (Ovaskainen et al. 2017). Additionally, the structural equation model can be used to identify the

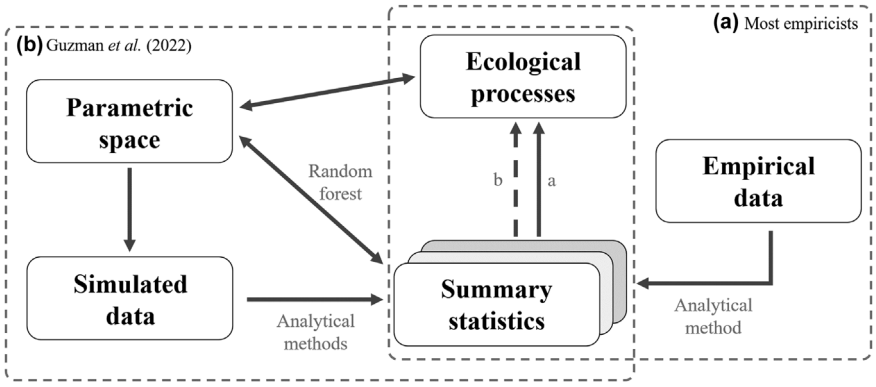

Figure 1. Flow diagram, showing the difference between the path most empiricists would take (box (a) at the right side), compared to the path proposed by Guzman et al. (2022) (box (b) at the left side). (a) Most empiricists infer the ecological processes underlying the observed metacommunities based on the summary statistics calculated by a single analytical method. For example, beta-diversity variation partitioning is one of the analytical methods to quantify the relative importance of environmental filtering and dispersal on species turnover based on the explained variations in constrained ordination. (b) Guzman et al. (2022) extended this process by linking multiple summary statistics derived from different analytical methods to the parametric space of the process-based simulation model. The model parameters that regulate the strength of the ecological processes may be estimated based on the summary statistics derived from the empirical data.

causal relationship between ecological processes and spatio-temporal species turnover (Jabot et al. 2020).

Although these analytical methods are widely used in ecological studies, ecologists have not reached an agreement on which of them can best disentangle the ecological processes underlying the observed metacommunity. Since, in natural communities, the real impact of these processes on the species turnover is unknown, the efficiency of these analytical methods is difficult to evaluate. Additionally, even though the effectiveness of these methods has been assessed independently through simulation approach (Smith and Lundholm 2010, Tucker et al. 2016, Ning et al. 2019), and criticized (Tuomisto et al. 2012, Vellend et al. 2014, Brown et al. 2017, Molina and Stone 2020), they have rarely been systematically compared. One possible solution is to use a process-based modeling approach to simulate metacommunity data structured by known strengths of assembly processes, to enable a systematic evaluation of the analytical methods. A process-based model is defined as 'a model that characterizes changes in a system's state as explicit functions of the events that drive those state changes' (Connolly et al. 2017). The model parameters in the process-based model can control the strength of specific ecological processes, and contribute to the variation in population sizes of different species over space and time (thereby affecting the system's states). By defining a comprehensive parametric space, we may study whether the summary statistics derived from these analytical methods can successfully reflect the variation in the strength of ecological processes.

Guzman et al. (2022) is one of the few studies that attempted to synthesize multiple analytical methods and incorporate community data across space and time, with the aim of understanding the underlying ecological processes (Fig. 1B; Ovaskainen et al. 2019). Guzman et al. (2022) evaluated the performance of different sets of analytical methods in estimating the underlying ecological processes by employing simulated metacommunity data with controlled strength of assembly processes and concluded that no single analytical method had outstanding performance. Furthermore, these authors developed a framework that integrates summary statistics obtained from several different analytical methods, and showed that enhancing the sampling completeness of species turnover across space and time also improves the accuracy in estimating the underlying ecological processes. However, Guzman et al. (2022) did not discuss how and whether their framework could be applied to empirical data.

In our study, we extended Guzman et al. (2022) by 1) including null-model-based analytical methods, demonstrating its flexibility in incorporating different types of methods; 2) excluding certain summary statistics considered in Guzman et al. (2022), which highlights issues in the choice of statistics when applying this framework; 3) defining metacommunity archetypes which illustrate extreme cases of metacommunities and visualizing the dynamics of different summary statistics under these extreme metacommunities across time to assess their ability to distinguish between archetypes, 4) applying the extended framework to real-world vegetation data from Fushan Forest Dynamics Plot (FFDP), a subtropical rainforest in Taiwan, and interpreting the results; and 5) discussing the potential advantages, limitations, and future directions of using this framework for practical applications.

Material and methods

In our study, we followed the framework introduced by Guzman et al. (2022) (Fig. 1B), and created the simulated metacommunity data by the process-based metacommunity framework of Thompson et al. (2020), with several model parameters regulating the strength of different ecological processes. A simulated metacommunity contained multiple patches, representing multiple local communities or plots. The three model parameters in Thompson et al.'s framework, namely niche width, competition type, and dispersal of the species, are related respectively to the relative importance of environmental filtering and stochasticity, different density-dependent biotic interactions and strength of dispersal limitation underlying the metacommunity. We retained beta-diversity variation partitioning and replaced the hierarchical modeling of species communities (HMSC) used by Guzman et al. (2022) by incorporating two alternative analytical methods: Stegen's framework (Stegen et al. 2013) and the dispersal-niche continuum index (Vilmi et al. 2021). We used the random forest (RF) approach, as Guzman et al. (2022) did in their framework, to estimate the model parameters in Thompson et al.'s model by using the summary statistics derived from the three analytical methods as predictors. The performance and the robustness of the trained RF were evaluated by calculating the correctness of the parameter estimation and the sensitivity test on sampling effort and choice of iteration time steps. To demonstrate the application of Guzman et al.'s framework, we applied it to the empirical data from the repeated census of woody plant species in FFDP, a subtropical rainforest in Taiwan. In our study, the symbols in the formulas were mostly consistent with the original papers and we have not attempted to harmonize them. All the simulations and calculations were done in `Julia` (Bezanson et al. 2017) or in R (www.r-project.org).

Process-based metacommunity framework

In the process-based metacommunity framework proposed by Thompson et al. (2020), metacommunity dynamics is assumed to be influenced by three main mechanisms: 1) density-independent abiotic responses, 2) density-dependent biotic interactions and 3) dispersal. In addition, demographic stochasticity is also considered in this model. Niche width, competition type, and dispersal ability of the species are the model parameters that modify the strength of these mechanisms and further regulate the strength of the ecological processes. Rigorously, the narrow species niche width indicates large differences in ecological preferences among species, which increases the strength of environmental filtering in

forming the species composition within the local community; the wide species niche width results in ecological preferences shared with other species, which conversely increases the stochasticity within a local community. Different species competition types regulate the level of similarity in resource usage, priority effect, or competitive hierarchy. The strength of the species dispersal ability indicates the levels of species dispersal limitation within the metacommunity. For simplicity, these three model parameters are assumed to be independent of each other.

The population size of the species i in patch x at time t (t-th iteration) is denoted by $N_{ix}(t)$. Thompson et al. (2020) considered the discrete-time model

$$N_{ix}(t+1) = N_{ix}(t)\frac{r_{ix}(t)}{1+\sum_{j=1}^{S}\alpha_{ij}N_{jx}(t)} + I_{ix}(t) - E_{ix}(t)$$

where $r_{ix}(t)$ is the density-independent growth rate of species i in patch x at time t, α_{ij} is the per capita effect of species j on species i, so-called competition coefficients, S is the total number of species, $I_{ix}(t)$ is the number of individuals of species i arrive at patch x from elsewhere in the metacommunity via dispersal at time t, and $E_{ix}(t)$ is the number of individuals of species i leave from patch x at time t via dispersal.

This metacommunity framework assumes that the discrete patches within the metacommunity are linked by the dispersal of the species. The patches are located on the torus with equal height and width to avoid edge effects, and the x- and y-coordinates of the patches are randomly generated by uniform distribution between 1 and 100. Multiple individuals and different species may co-occur within a patch.

The density-independent growth rate of species $r_{ix}(t)$ is determined by the species niche and the environment value of the patch (Supporting information). The species niche optimum is independently generated by a random value from a uniform distribution between 0 to 1. All species in the given simulation are assumed to have the same niche width σ. Moreover, the spatially autocorrelated environmental condition of the patches is embedded in the simulated metacommunity. The value describing the state of environmental conditions in each patch (hereafter called environment value) is generated by the stationary isotropic covariance model (by *RMexp* function in the 'RandomFields' R package, www.r-project.org, Schlather et al. 2015). For each patch, only one environment value ranging between 0 and 1 is generated. Contrary to Thompson et al. (2020), in our study, the environment value was set to be constant across time but varied across space with positive spatial autocorrelation. We also assumed a linear relationship between functional traits and the environment; therefore, the species niche optima were considered analogous to their functional traits in further analysis.

The density-dependent biotic interactions are categorized into five competition types with different scenarios of competition coefficients (α_{ij}): 1) no competition ($\alpha_{ii}=1$ and $\alpha_{ij}=0$), 2) stabilizing competition ($\alpha_{ii}=1$ and $\alpha_{ij} \sim$ Unif[0, 0.5]), 3) equal competition ($\alpha_{ii}=1$ and $\alpha_{ij}=\alpha_{ii}$), 4) mixed competition ($\alpha_{ii}=1$ and $\alpha_{ij} \sim$ Unif[0, 1.5]) and 5) competition–colonization tradeoff. In case of competition–colonization tradeoff, one-third of the species are assigned to be the dominant species, which have stronger competitive ability compared to the other species, which are called inferior species. Rigorously, dominant species are assumed to impose more impact on inferior species ($\alpha_{ij} \sim$ Unif[1, 1.5] and $\alpha_{ji} \sim$ Unif[0, 1] if species j is a dominant species and species i is an inferior species). The impact from the same types of species, e.g. one dominant species impacts on another dominant species, is lower ($\alpha_{ji} \sim$ Unif[0, 1]).

The migration of the individuals is determined by the dispersal ability of the species and the distances between patches (calculated by the Euclidean distance on the torus; Supporting information). The emigration rate and immigration rate are related to the distance matrix of the patches. For the first four competition types, the species are assumed to have the same dispersal ability a. In the case of competition–colonization tradeoff, the inferior species have a stronger dispersal ability (a) than the dominant species ($0.1a$).

For each replicate of the simulation, the landscape configuration, including the coordinates of the patches and the environment value of each patch, and species trait were first generated before the iteration starts (Supporting information). The environment value in each patch was fixed across all the iterations. The initial species abundance of each species in each patch was generated independently by Poisson distribution with mean 0.5. The simulation ran 2200 iterations, with 200 iterations for the burn-in stage. Within the burn-in stage, the recruitment event, which adds individuals with numbers independently generated by Poisson distribution with mean 0.5 for each species in each patch, was implemented every 20 iterations. The number of individuals of species i in patch x in time t (t-th iteration) is denoted by $N_{ix}(t)$. Then, the expected individual number of species i in patch x in time $t+1$ is defined as

$$N_{\exp,ix}(t+1) = N_{ix}(t)\frac{r_{ix}(t)}{1+\sum_{j=1}^{S}\alpha_{ij}N_{jx}(t)} + I_{ix}(t) - E_{ix}(t)$$

The number of individuals of species i in patch x at time $t+1$ was generated by Poisson distribution with the mean equal to the expected individual number at time t, i.e. $N_{ix}(t+1) \sim \text{Poisson}(N_{\exp,ix}(t+1))$. After 1800 iterations, the abundance for each species in each patch was recorded for every 20 iterations, until the simulation ended. We eventually recorded 20 snapshots of the species composition and abundance of a simulated metacommunity ($t=1, 2, \ldots, 19, 20$) (Supporting information).

We considered the same parametric space proposed by Thompson et al. (2020) (Supporting information), including 13 values for species niche width σ (0.001–10), five competition types, and 15 values for species dispersal ability a (0.00001–1). The species niche widths and dispersal abilities are transformed into ordinal variables (level 1–13 and level

1–15, respectively; Supporting information). Eighteen replicates were independently run for each parametric setting.

For further analysis, the samples with low species occurrence (one occurrence is defined by the emergence of a species in a single patch), abundance, or diversity, i.e. total occurrence ≤ 200 patches, total abundance ≤ 1000 individuals, or regional diversity ≤ 3 species, were excluded. Four metacommunity archetypes were defined subjectively by a specific combination of parameters in the parametric space to represent the extreme scenarios of the simulated metacommunity (Fig. 2F). Patch dynamics (PD) was defined as the scenario with competition–colonization tradeoff and arbitrarily determined niche width and dispersal ability; species sorting (SS) as the scenario with relatively narrow species niche width, intermediate dispersal ability, and stable competition; neutral dynamics (ND) as the scenario with relatively wide species niche width, intermediate dispersal ability, and equal competition; mass effect (ME) as the scenario with relatively strong dispersal ability, narrow species niche width, and stable competition. These archetypes were used for visually evaluating the efficiency of the summary statistics in discerning variation of ecological processes. Furthermore, they were also treated as the testing dataset for evaluating how sampling effort would influence the assessment of the ecological processes.

Beta-diversity variation partitioning

Beta-diversity variation partitioning is a popular method for quantifying the relative importance of environmental filtering and dispersal underlying species turnover. This method was proposed by Borcard et al. (1992) and has been widely used in ecological studies (Cottenie 2005, Peres-Neto et al.

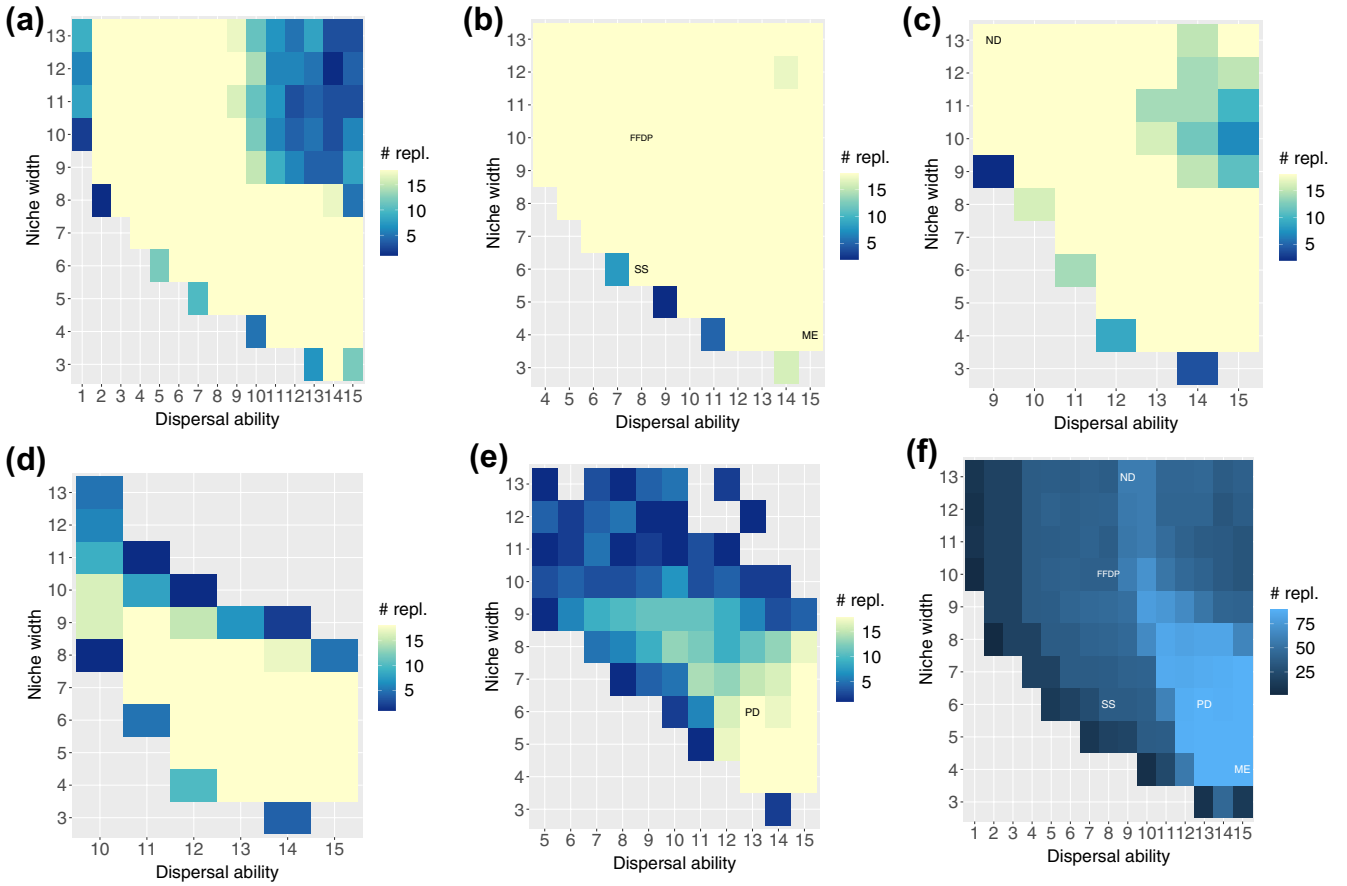

Figure 2. Parametric space defined by niche width, competition type, and dispersal ability of the species in the simulated metacommunity. The number on the two axes represents the level of species niche width and dispersal ability (Supplementary information). On the x-axis, the numbers from 1 to 15 are the dispersal ability from weak to strong. On the y-axis, the numbers from 1 to 13 are the niche widths from narrow to wide. The values in the tile plots represent the number of replicates that remain after excluding those with too low species occurrence, abundance, or diversity, or those too sparse to calculate DNCI value. If the scenario has no replicates after filtering, the values are shown as missing in the tile plots. The axes are truncated since no replicates remained after data filtering for those scenarios. (a)–(e) show the parameter space defined by niche width and dispersal ability with different competition types. (f) is the summation of the five tile plots from (a) to (e). The labels in the tile plot show the subjective definition of the four metacommunity archetypes in the parametric space: species sorting (SS), neutral dynamics (ND), mass effect (ME) and patch dynamics (PD). The species in Fushan Forest Dynamics Plot (FFDP) are predicted to interact with each other with stabilizing competition and have relatively wide niche widths and relatively weak dispersal ability.

2006, Smith and Lundholm 2010). It requires data about species composition, environment, and geographical information for each plot in the observed metacommunity, and uses a constrained ordination technique to quantify the variation in species composition explained by either environment or by the coordinates of the plots (Supporting information). The variations explained by environment and space are then used to compare the relative importance of environmental filtering and dispersal in species turnover across metacommunities (Cottenie 2005) (several issues of this method in Kraft et al. 2011, Chang et al. 2013).

In our study, we conducted canonical correspondence analysis (CCA; ter Braak 1986, Legendre and Legendre 2012) to the simulated data we generated. A snapshot of the species abundance matrix was assigned as the response variable, while environment values generated at the beginning of the simulation and spatial attributes derived from the coordinates of the patches were used as the explanatory variables. The spatial attributes were derived from applying distance-based Moran's eigenvector maps (dbMEM) on the distance matrix of the patches (Borcard and Legendre 2002). To quantify the pattern of positive autocorrelation in species abundances, only the eigenfunctions with positive eigenvalues were considered as the explanatory variables. The fraction explained by only environment [a], by only space [c], by both environment and space [b], and unexplained fraction [d], were treated as the summary statistics of beta-diversity variation partitioning (Supporting information).

Stegen's framework

Stegen et al. (2013) is a null-model-based method proposed to quantify the relative importance of selection, dispersal limitation, homogenizing dispersal, and drift, underlying the microbial metacommunity, and in original implementation requires phylogenetic and species composition data. Compared to most of the null-model-based methods that can only disentangle the presence of some ecological processes (Ulrich and Gotelli 2010, Chase and Myers 2011), Stegen et al. (2013) applied two-step null model to every pair of plots within the metacommunity. The significance of the divergence or the convergence of the phylogenetic structure and species composition between the two plots is tested. The relative importance of the four ecological processes is summarized by the significances derived from all pairs of plots (Supporting information).

In our study, we used Stegen's framework with the modification proposed by Ford and Roberts (2020) to replace phylogenetic data with functional traits. The niche optima of the virtual species in the simulation model were treated as functional traits. We applied it to all the possible pairs of patches in a simulated metacommunity to calculate the relative importance (fraction) of selection, dispersal limitation with drift, homogenizing dispersal, and pure drift based on the species composition and species traits (Supporting information). These relative importances were treated as the summary statistics of Stegen's framework.

Dispersal-niche continuum index

Dispersal-niche continuum index (DNCI) is another null-model-based method that aims to quantify the relative importance of niche and dispersal assembly processes, originally used on the paleontological community data (Vilmi et al. 2021, Supporting information). DNCI value is calculated by comparing the observed and the randomized SIMPER (similarity percentage) profile (which was proposed by Clarke (1993) and aims to summarize the relative contributions of the species in the community to the overall composition dissimilarity; Supporting information). Interestingly, only species composition data, grouped by any cluster or ordination analysis, is required to calculate the DNCI value. Environmental and functional trait data which are difficult to collect are not necessary. The value of DNCI and its SD were used to infer the relative importance of dispersal and niche assembly processes; if DNCI value is significantly larger than 0, then niche assembly processes are supposed to be the dominant processes driving the species turnover, while if DNCI value is significantly lower than 0, then dispersal assembly processes are the most influential. If DNCI value is not significantly different from 0, then both processes may be similarly important for the community assembly. The significance is determined by whether the confidence interval of DNCI value encompasses 0.

In our study, we treated DNCI value and its SD as the summary statistics of the metacommunity. Additionally, since in some cases, the species composition was too sparse (contained too many zeros) and failed to find the column permutation with nonzero row sums, we added the time limitation that if the permutation cannot be found in 30 s then we dropped that whole sample.

Linking summary statistics with parametric space of simulation model

To construct the link between the summary statistics, which are derived from beta-diversity variation partitioning, Stegen's framework and DNCI, and the model parameters in Thompson et al.'s process-based metacommunity framework, we followed a similar approach as in Guzman et al. (2022) by applying the algorithm of RF. RF is a statistical classifier that constructs a mapping between variables by the training dataset. The trained RFs were also applied to the empirical data to disentangle underlying ecological processes in the real community.

To construct a RF, the 12 replicates of the simulated data (two-thirds of 18 replicates) were used as the training data and the remaining six replicates were used as the testing data. The response variables were the model parameters that influenced the relative importance of ecological processes underlying the metacommunity: niche width, competition type, and dispersal ability of the species. Niche width and dispersal ability were treated as ordinal variables (Hornung 2020), and competition type was treated as a categorical variable. We constructed 12 RFs for each model parameter. The explanatory

variables of these RFs consisted of four sets of summary statistics (i.e. those calculated by only beta-diversity variation partitioning, only Stegen's framework, only DNCI, or all three analytical methods) which were derived from three sets of iteration time steps (i.e. one snapshot ($t=20$), four snapshots ($t=20$, 16, 12, 8) or all 20 snapshots of the species composition), resulting in a total of 12 unique combinations of summary statistics. To assess the performance of the trained RFs in estimating the model parameters, we applied the statistics calculated from the testing data to the trained RFs. We then measured the accuracy of the parameter estimations by comparing the proportion of correct links between the estimated parameters and the actual ones. The importance of each summary statistic in estimating the model parameters was also quantified. For estimating competition type, the importance of each summary statistic was defined by the mean decrease Gini (Liaw and Wiener 2002, Han et al. 2016). For estimating niche width and dispersal ability, the importance of the explanatory variables was defined by the permutation variable importance measure (Janitza et al. 2016, Hornung 2021).

Without interpolating, complete data (with no missing values) is required for constructing and evaluating the RFs. Only the samples without missing summary statistics at all 20 iteration time steps were used to construct and evaluate the RFs. The missing statistics in some samples may have been caused by the strong stochasticity or maladaptation of the species in local environmental condition, which produced low occurrences, abundance, or diversity in metacommunities, and the sparseness of the species composition, which made it not possible to calculate DNCI value.

Sensitivity analysis to sampling effort and choice of iteration time steps

We assessed the robustness of the RFs to the sampling effort by specifically examining those RFs that used the summary statistics calculated by all three analytical methods derived from four snapshots as the explanatory variables. To reduce the calculation time, we only considered the defined four archetypes (PD, SS, ND and ME) as the testing data. We calculated the summary statistics based on the three analytical methods for the four snapshots ($t=20$, 16, 12, 8) of the simulated species composition generated by the parameters setting of these four archetypes. Then, we subsampled the species composition data by randomly choosing 10, 20, ... and 90 of the patches 15 times independently. Under different sampling efforts, the summary statistics were recalculated based on the subsampled species composition and inserted in the trained RF to calculate the accuracy in estimating the model parameters. How the performance of the RFs in estimating the model parameters would be affected by the completeness of the species composition data was shown in a line graph.

We also assessed the robustness of the same RFs to the choice of iteration time steps to be the input data. Without determining the unit of the iteration time in the simulation model, we could not match the iteration time in the simulated model with the periods in the real metacommunities. Therefore, it is necessary to test whether the choice of the iteration time steps would affect the estimation of model parameters. For such assessment, the explanatory variables of the RF were the summary statistics calculated by the three analytical methods from the snapshots at $t=20$, 16, 12, 8. For each model parameter (niche width, competition type, and dispersal ability), we reassigned the summary statistics derived from randomly chosen four time steps from $t=1$, 2, …, 19, 20 in increasing order as the inputs of the trained RF. The correctness of the estimation based on the mismatched summary statistics was calculated. We randomly chose the time steps 99 times. The distribution of the accuracy for estimating three model parameters was shown in a boxplot.

Application to vegetation data from Fushan Forest Dynamics Plot

The FFDP is a 25 ha forest dynamics plot established in northern Taiwan in 2003, with the first survey finished in 2004. FFDP is located in the subtropical zone at 24°45′40″N, 121°33′28″E with elevation from 600 to 733 m a.s.l. The 500 × 500 m plot is divided into 625 20 × 20 m quadrates. The delineation of the plot was conducted in 2002–2003. Precise topography was measured by the electronic total-station theodolites. FFDP consists of multiple topographic components, such as hills, ridges, slopes, valleys, flats, and creeks (Su et al. 2007). As the environmental data, we used mean elevation, convexity, slope, and northernness (aspect folded along the N–S axis) of each quadrat, and standardized each to Z-scores before further analysis.

The species composition of FFDP was surveyed in 2003–2004, 2008–2009, 2013–2014 and 2018–2019, following the method developed by the Center for Tropical Forest Science (CTFS, now ForestGEO) (Condit 1998). These four censuses can be imagined as the four snapshots of metacommunity in FFDP. All the woody species and the tree ferns with a diameter at breast height (DBH) of ≥ 1 cm were identified, measured, mapped, and tagged. Within four censuses of FFDP, between 111 683 to 133 558 individuals of 111 species in 40 families and 68 genera were recorded. Details of the information about environmental conditions and the field survey can be found in Su et al. (2007).

Specific leaf area (SLA), leaf thickness, and leaf dry matter content (LDMC) for 103 species were measured on randomly selected 6–12 individuals for each species found in FFDP (Lasky et al. 2013).Total organic nitrogen mass per unit leaf mass (N mass) and total organic phosphorus mass per unit leaf mass (P mass) were obtained for 99 of the 103 species based on two microplate methods. The maximum tree height of the 99 species was measured according to Cornelissen et al. (2003). The wood density of 74 species was measured in FFDP, of 21 species were obtained from the literature. The wood density of the remaining species was derived from Chave et al. (2009).

The four snapshots of species composition data, together with data about topography, geographical coordinates of

each quadrat, and species leaf traits were used to calculate the summary statistics of beta-diversity variation partitioning, Stegen's framework, and DNCI. Since the incompleteness in trait data, when calculating the dissimilarity of traits between two plots in Stegen's framework (βMNTD), the species with missing traits were ignored. We obtained a total of four sets of summary statistics based on the four snapshots of the species composition, which we inputted into the trained RFs (incorporating three analytical methods and four snapshots) to estimate the niche width, competition type, and dispersal ability of the species in FFDP. Additionally, we quantified the robustness of the estimation based on empirical data against sampling effort, using the following procedure. First, we subsampled 90% of the quadrates for every census, and used the subsampled species composition data to calculate the summary statistics (by beta-diversity variation partitioning, Stegen's framework, and DNCI). Second, we inputted these summary statistics into the trained random forest to derive newly estimated parameters within the parametric space. By doing this 100 times, we got the distribution of the new estimated results and quantified how sensitive the estimation is against a 10% loss in the number of quadrates.

Results

Summary of the simulated data

Overall, we had 17 550 samples from 18 replicates of metacommunity with 975 scenarios (five competition types, 15 dispersal abilities, and 13 niche widths) generated by Thompson et al.'s framework. Each sample contained 20 snapshots of the species composition at different iteration time steps. After filtering out the incomplete samples, which missed some summary statistics at some iteration time, we retained 5150 samples and 349 scenarios and used them in further analysis.

The parametric space for the retained samples was shown by a heat map (Fig. 2). The scenarios with weak dispersal ability a and narrow niche width λ were mostly filtered out due to the sparseness of the metacommunity (i.e. relatively few individuals or species present). In addition, more samples were retained in scenarios of no competition, stabilizing competition, and equal competition (Fig. 2A–C), compared to mixed competition and competition–colonization tradeoff (Fig. 2D–E). We defined the four metacommunity archetypes, each by one of the scenarios in the parametric space (Fig. 2F).

Summary statistics of the simulated metacommunity

Statistics of beta-diversity variation partitioning, Stegen's framework and DNCI were calculated for the remaining 5150 samples. All the statistics fluctuated across time and differed between replicates (Fig. 3). The summary statistics derived from Stegen's framework could successfully separate the four archetypes. The relative importance of selection could separate SS from others; the relative importance of dispersal limitation could separate ME from others; the relative importance of homogenizing dispersal could separate PD and ND. For beta-diversity variation partitioning and DNCI, they could not identify any archetypes (except for the variation explained only by environment [a], which inclines to separate the archetypes). The ranges of the summary statistics of three analytical methods are in the Supporting information.

Performance in estimating model parameters

We assessed the accuracy of estimating three model parameters (i.e. niche width, competition type and dispersal ability of the species) for all 36 RFs (Table 2). For the RFs that only considered one analytical method, Stegen's framework had the best performance in estimating all three model parameters compared to beta-diversity variation partitioning and DNCI, if considering the same number of snapshots. By incorporating the statistics derived from multiple snapshots, the performance was improved. However, among the 12 different sets of explanatory variables, the one that incorporated all summary statistics from three analytical methods and 20 snapshots had the highest accuracy in estimating dispersal ability (69.15%) and competition type (83.55%) of the species in the simulated metacommunity. This combination also had the second-highest accuracy in estimating niche width (71.08%), which was only 0.88% less than the RF with the highest accuracy. The second best combination of the explanatory variables was the one that incorporated all summary statistics from four snapshots, which had 67.15, 71.96 and 83.02% correction rates in estimating the dispersal ability, niche width, and competition type, respectively. We showed that the RFs which integrated statistics derived from multiple analytical methods had better performance in estimating the model parameters than those which only considered the statistics derived from single analytical methods. The importance of the explanatory variables of the RFs that incorporate the summary statistics derived from three different analytical methods based on the four snapshots was quantified (Supporting information). The fraction of homogenizing dispersal and selection derived from Stegen's framework, and the value of DNCI, were the three most important statistics for estimating the competition type. The fraction of selection, homogenizing dispersal, and dispersal limitation derived from Stegen's framework and the variation explained by space derived from beta-diversity variation partitioning were the most important statistics for estimating the dispersal ability. Variation explained only by environment, the fraction of selection, and drift derived from Stegen's framework were the most important statistics for predicting the niche width.

Robustness to the sampling effort and choice of iteration time steps

For the trained RFs with summary statistics derived by the three analytical methods from four snapshots, the

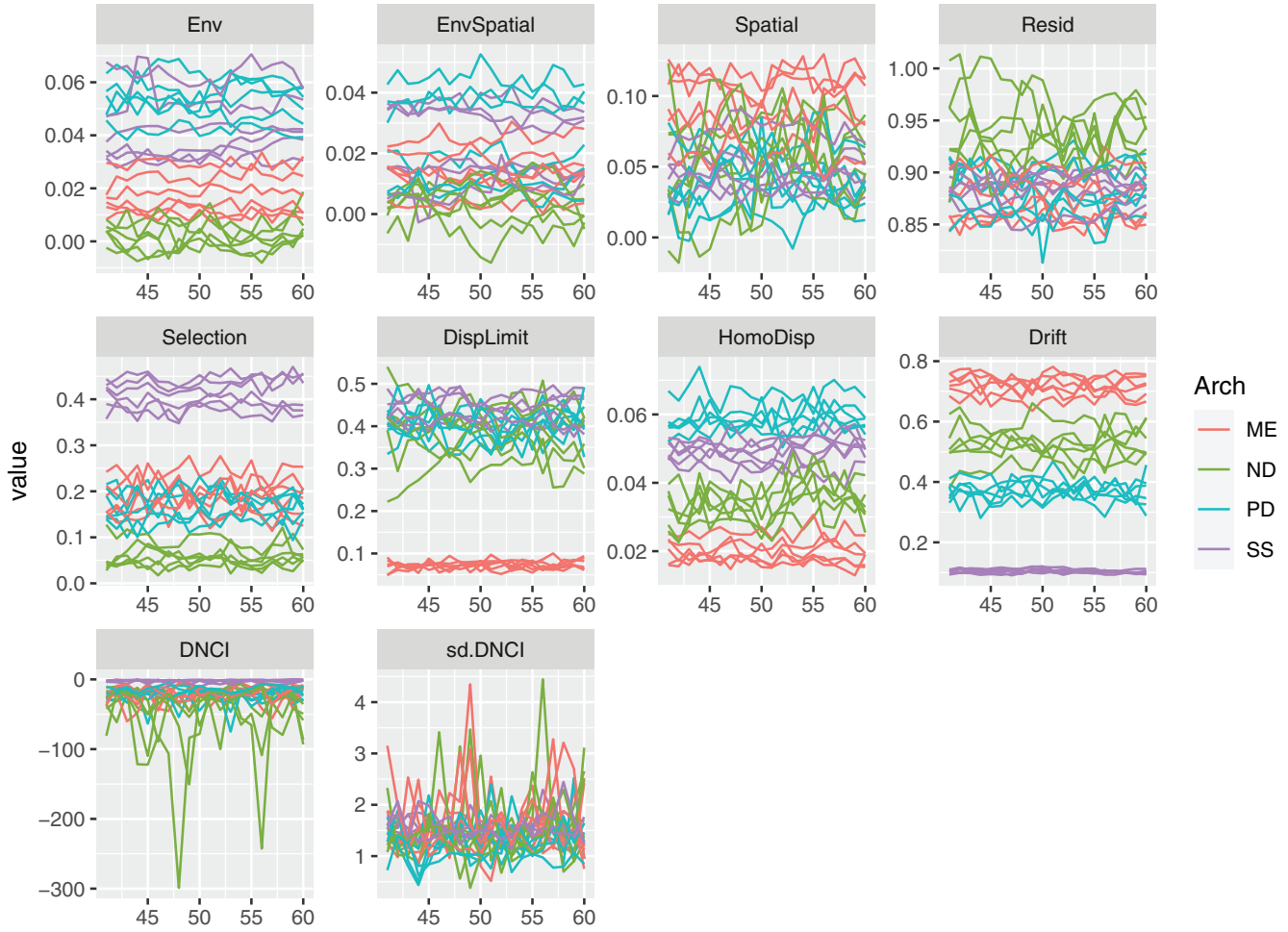

Figure 3. Comparing the dynamics of the summary statistics derived by beta-diversity variation partitioning, Stegen's framework and DNCI under four metacommunity archetypes. Within each panel, the curve shows the dynamics of the values of each summary statistic across time. The first row shows the dynamics of the variation explained only by the environment (denoted by Env), by both environment and space (EnvSpatial), only by space (Spatial), and unexplained variation (Resid) derived from beta diversity variation partitioning. The second row shows the dynamics of the fraction of selection (Selection), dispersal limitation (DispLimit), homogenizing dispersal (HomoDisp), and drift (Drift) derived from Stegen's framework. The third row shows the dynamics of the value of DNCI and its SD (sd.DNCI). Different colors represent different metacommunity archetypes. A maximum of six replicates are shown for each archetype.

performance in estimating the niche width, competition type, and dispersal ability decreased when the number of subsampled patches decreased (Fig. 4A). However, the performance in estimating the niche width, competition type, and dispersal ability had no difference even though the summary statistics were derived from the species composition at four randomly selected iteration time steps (Fig. 4B).

Application to Fushan Forest Dynamics Plot

By inputting the summary statistics derived from four snapshots of FFDP based on three analytical methods (Supporting information), we estimated that the species in FFDP act with stabilizing competition ($\alpha_{ii}=1$ and $\alpha_{ij} \sim$ Unif [0, 0.5]) and have relatively weak dispersal ability and relatively wide niche width (Fig. 2B) when compared to the extremes in the parametric space. Additionally, the estimation for niche width and competition type underlying FFDP is also robust against 10% loss of the number of quadrates. The results showed that for estimating species niche width, 99% of the new estimations remained unchanged (at species niche width level 10; only one showed different results, at level 11). For estimating competition type, 100% of the new estimations stayed identical to the original estimation (stabilizing competition). However, for estimating species dispersal ability, only 69% of the new estimations remained unchanged (at species dispersal ability level 8; however, 14 replicates were estimated by level 10, and 17 replicates were estimated by level 15). The insufficient robustness occurred when estimating the species dispersal ability, which may be caused by the relatively weaker performance of the random forest in estimating dispersal ability (67.15% correctness).

Table 2. Accuracy of 12 RFs with different explanatory variables in prediction model parameters. The first four columns show the explanatory variables in the RFs. The symbol 'O' represents the summary statistics derived from analytical methods considered in the RF, while the symbol 'X' represents those not considered in the RF. The fourth column represents how many snapshots of the species composition are used to calculate the summary statistics and considered as the explanatory variables in the RF. The accuracy for predicting niche width, competition type, and dispersal ability is shown in the last three columns. VP: beta-diversity variation partitioning. Stegen: Stegen's framework. DNCI: the value of DNCI and its SD.

Explanatory variables						
Summary statistics				Performance of prediction (%)		
VP	Stegen	DNCI	Snapshots	Niche width	Competition type	Dispersal ability
O	X	X	1	41.28	43.74	19.20
O	X	X	4	48.77	54.92	23.01
O	X	X	20	51.23	57.32	25.23
X	O	X	1	46.96	68.50	46.55
X	O	X	4	54.10	74.77	54.74
X	O	X	20	54.04	77.28	57.96
X	X	O	1	16.98	46.49	13.82
X	X	O	4	22.37	55.56	22.78
X	X	O	20	26.05	57.03	25.41
O	O	O	1	69.56	81.09	62.35
O	O	O	4	71.96	83.02	67.15
O	O	O	20	71.08	83.55	69.15

Discussion

We confirmed the observation done by Guzman et al. (2022) that to understand ecological processes forming metacommunity, applying a single analytical method on a single snapshot of community data mostly resulted in unsatisfactory estimation. Instead, applying multiple analytical methods on multiple snapshots of community data leads to considerable improvement. This may be because each analytical method extracts a different subset of the available empirical data and

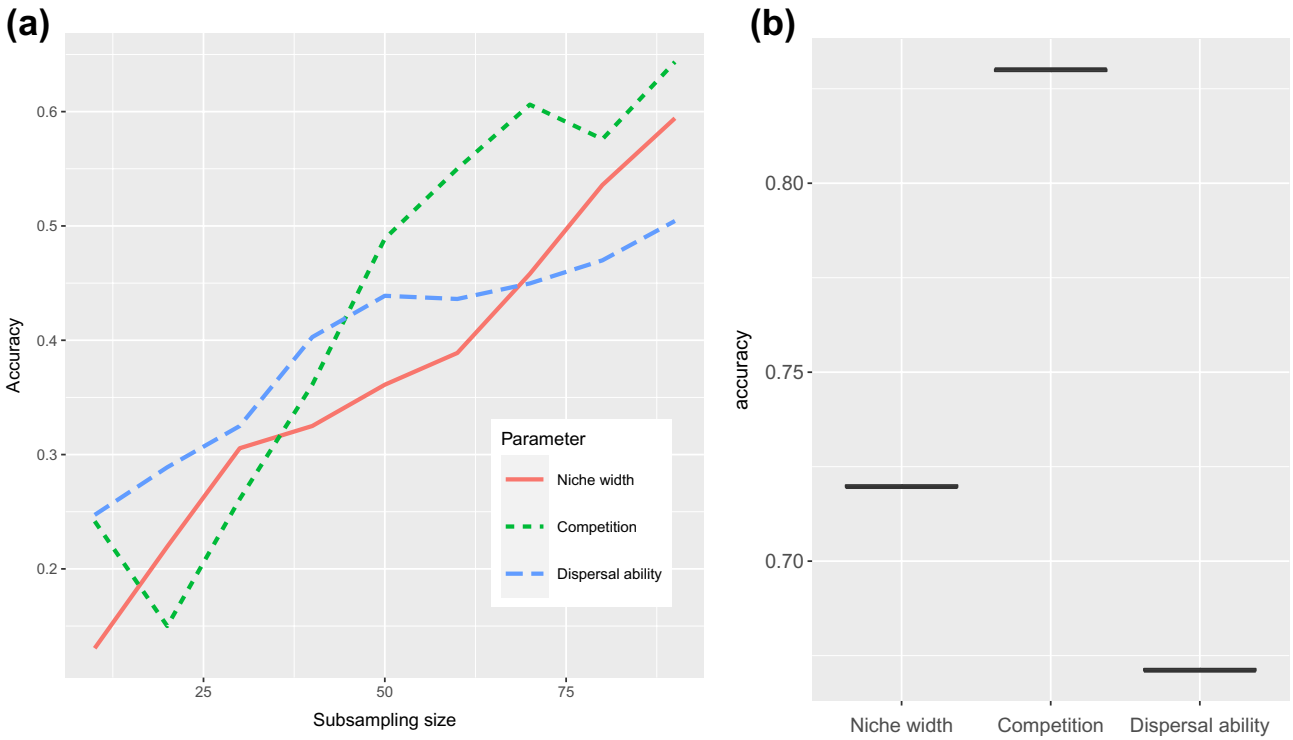

Figure 4. Robustness of random forest (RF) to the sampling effort and choice of the time steps. (a) Different curves represent the relationship between the performance of the random forest in estimating the niche width, competition type, and dispersal ability of the species, and the number of subsampled patches from a complete simulated metacommunity. (b) The distribution of the accuracy of the RFs for estimating the niche width, competition type, and dispersal ability of the species when using the summary statistics derived from the snapshots at randomly chosen iteration time steps (creating mismatch between the iteration times in training and testing dataset). The SDs of the accuracies are considerably small (< 0.03%) and not visible in the boxplots displayed.

specializes in different types of ecological processes while ignoring the others. Combining their results is like observing metacommunities from different perspectives (Overcast et al. 2021). Moreover, multiple snapshots data allows us to study the temporal variation in summary statistics. These can reveal additional ecological information and uncover a more complete pattern in community structure.

The comparison of the performance between analytical methods was made possible by applying these methods to the simulated community data generated by a process-based model with the known strength of assembly processes. While previous studies have tested some of these methods using simulated data with known properties (Smith and Lundholm 2010, Tucker et al. 2016, Ning et al. 2019), we believe that the integrated framework proposed by Guzman et al. (2022) and further elaborated by us provides a more comprehensive solution for systematically studying and comparing different analytical methods. We also found the use of metacommunity archetypes helpful for evaluating the behavior of individual summary statistics. These archetypes not only represent the extreme scenarios of metacommunity but also summarize the typical scenarios that have been studied by theoreticians. We used these archetypes to visually evaluate how summary statistics of different analytical methods distinguish between individual archetypes and how this ability changes with the advancing time of metacommunity composition development. It may be also a good idea to evaluate the effectiveness of existing statistics or indices in identifying individual archetypes, and investigate the properties of these statistics under temporal variations of ecological processes (e.g. a sudden habitat destruction that may intensify the influence of ecological drift). In addition, note that except for archetypes, there is still a large proportion of unexplored areas in the parametric space worth exploring (e.g. drift with strong dispersal limitation). We also encourage researchers to move beyond the concept of archetypes and study the effectiveness of summary statistics across continuous parametric space.

Furthermore, we demonstrated the application of Guzman et al.'s framework on real-world metacommunity data (Fig. 1), which is not considered in Guzman et al. (2022) (even though some similar approaches have been proposed, Munoz et al. (2018) and Overcast et al. (2021)). We identified a relatively strong influence of stochasticity within FFDP (associated with relatively weak environmental filtering), possibly due to the frequent disturbance caused by typhoons. While the dominant species within FFDP have shown to have low mortality rates to persist against typhoons (Su et al. 2020), further investigation is necessary to understand whether the frequent disturbed typhoons do not increase the mortality rate for the rest of the species within FFDP. If the influence is significant, the species composition may be considerably influenced by the stochastic disturbance from typhoons, which consequently aligns with our findings. Additionally, it has been reported that typhoons can induce defoliation and reduce the difference in light levels between gap and non-gap areas (Yao et al. 2015). This neutralizing effect may homogenize the seedling composition and further contribute to the relatively weaker environmental filtering observed within FFDP. Moreover, we showed that conspecific density has more impact on species mortality than heterospecific density in FFDP. However, a previous study did not find a consistent negative density dependence on dominant tree mortality in FFDP (Su et al. 2020). We also identified a relatively strong dispersal limitation within FFDP, possibly due to the recruitment limitation of the species within the plot (Chang-Yang et al. 2013, 2021). A previous study also proposed the presence of dispersal limitation in FFDP by individual-level analysis (Shen et al. 2013).

To simulate community data, our framework incorporated Thompson et al.'s model, which is principally based on high-level ecological processes and is sufficiently general to be applied to different community systems. By further relaxing the assumptions of model parameters (e.g. by making niche widths and dispersal ability of individual species unequal), this simulation model may generate an even more comprehensive spectrum of metacommunities. Theoretically, Thompson et al.'s model can also be replaced by any more specific model, allowing to study more detailed ecological processes, such as demographic change in growth, mortality, and reproduction rate in different life stages (Jops and O'Dwyer 2023). In fact, the entire framework is flexible in the sense that different sets (and different numbers of) analytical methods can be integrated. The decision on which analytical methods to include can be driven by either theoretical considerations (e.g. to make sure that these methods together evenly cover different ecological processes) or practical considerations (depending on what kind of data are available for analysis; for example, if trait data are missing, Stegen's framework modified for the use of trait matrix cannot be included). Furthermore, the researchers can use different model selection approaches to identify the random forest with the most effective combination of summary statistics to estimate the ecological processes (Guzman et al. 2022).

By applying our framework in different metacommunities, it is possible to study the differences in the strength of ecological processes across space and time. In our study, since we considered only one empirical dataset, we could only compare the estimated strength of ecological processes with the extreme cases in the parametric space, such as archetypes or the space boundary. A better use of our framework would be to incorporate multiple empirical datasets, and study the variation in the strength of ecological processes among different metacommunities (Munoz et al. 2018, Overcast et al. 2021); we encourage other researchers to follow this approach. Particularly, one may compare the strength of ecological processes underlying the metacommunities across gradients of anthropogenic disturbance (caused by climate change, invasion of alien species, or habitat destruction). Such disturbance is known to alter the strength of ecological processes and to influence the stability of biodiversity and ecosystem function (Chase et al. 2020, McFadden et al. 2023). Since the degree to which various anthropogenic activities affect the strength of ecological processes is still poorly known (McFadden et al. 2023), we suggest that our framework can help understand

the underlying ecological processes and provide an opportunity for forecasting the metacommunity dynamics and studying the species coexistence and diversity (Adler et al. 2007, 2010) under human impacts. Moreover, the strategies for conserving biodiversity, e.g. optimizing habitat connectance or size, may be altered based on the underlying processes. For example, if the metacommunity is strongly influenced by ecological drift, then protecting the area of the target habitat would be essential to avoid the decline in rare species caused by drift. If, on the other side, the strength of dispersal is an important factor in reducing the diversity, then the habitat connectance or protection from alien species invasion is more essential (McFadden et al. 2023).

In addition, how to quantify the confidence interval of the estimated ecological process strength may be a crucial question when comparing the estimated values among different communities. In our study, we attempted to achieve it by a subsampling approach. By removing a small amount of the empirical data, we may retain the similar completeness of the dataset. Then, we may reestimate the ecological processes by the subsampled empirical data and compare the results with those obtained by using the complete dataset. By doing this subsampling several times, we may quantify how the estimated results are sensitive to the data completeness and evaluate the practical usage of trained random forests.

Limitations and future directions

Before further utilization of the framework explored in this study, it is important to acknowledge that there are various limitations and potential issues that require consideration.

First, the assumptions of the process-based simulation model should align with the observed metacommunity (e.g. variability of the environmental conditions, and dispersal strategies of species). Otherwise, the resulting pattern generated by the simulation model may not be relevant to the observed metacommunity. For example, we may allow the environment value in each patch to fluctuate over time to simulate the intensively fluctuating environmental conditions in marine or tidal ecosystems. In addition, instead of competition, we may consider other types of species interactions, such as predation and mutualism (Pilosof et al. 2017). However, the tradeoff between generality, realism, and precision cannot be ignored (Levins 1966). By constructing an overly comprehensive and complex process-based model, we may lose its generality in application to different ecosystems and reduce the precision in estimating the model parameters. In this case, to improve the precision, we should create a larger simulated dataset or incorporate alternative techniques for parameter estimation (e.g. approximate Bayesian computation; van der Vaart et al. 2015, Munoz et al. 2018, Guzman et al. 2022).

Second, the process-based simulation model should generate simulated data that results in a relatively complete range of summary statistics, which are calculated using analytical methods. If the statistics calculated by the observed data are not encompassed within the extent that the simulation model can generate, the real-world metacommunity will be out of the domain of the simulated dataset. This may lead to inaccurate estimation of the strength of ecological processes. In our study, the summary statistics calculated by the empirical data from FFDP were encompassed within the range of summary statistics that resulted from the simulated dataset. Identifying the positions of the empirical data fits in the distribution of summary statistics calculated by the simulated data may be another approach to check the relevance of the estimation (Supporting information). Additionally, the usage of summary statistics should be concerned since some of the statistics (e.g. the descriptive statistics in Guzman et al. 2022) are intensively varied across systems and may not be easily controlled in the process-based simulation model. For example, gamma diversity is one of the descriptive statistics, which counts the total number of species within the metacommunity. It may not only be regulated by the ecological processes, but also by other factors, such as the size of the species pool and the number of patches of the simulated metacommunity. Without modifying these two parameters in our study, the range of the gamma diversity derived from the simulated data would be limited and may not encompass the gamma diversity in the observed metacommunity.

Third, the incompleteness of the observed metacommunity data may reduce the efficiency in identifying the strength of ecological processes. We showed that the incompleteness of the species composition reduced the performance of the trained RF in predicting the model parameters, as shown in Guzman et al. (2022). Also, unmeasured environmental variables and incomplete functional trait data are shown to considerably influence the summary statistics of the beta-diversity variation partitioning (Chang et al. 2013) and functional trait metrics (Pakeman 2014), respectively. We expected that the deviation in summary statistics caused by the incompleteness of the data may lead to the performance of the trained RF being overestimated. The magnitude of the deviation would be determined by the robustness of the analytical methods themselves. Thus, the robustness of analytical methods should be investigated and compared systematically in the future, and the summary statistics sensitive to the data incompleteness should be excluded from the set of explanatory variables when estimating the strength of ecological processes.

Fourth, without determining the unit of the time steps in the simulation model (e.g. a time step represents a week or five years), we do not know which time steps represent which census of the observed metacommunity. The choice of iteration time steps to calculate the summary statistics as the explanatory variables of the RF may influence the parameter estimation. In our study, we showed that even though the time steps to calculate the statistics as the inputs of the RF were randomly chosen, the performance of the RF was maintained. This robustness may be led by the temporal stability of the species composition and environmental conditions of the simulated communities. Additionally, none

of the statistics we used directly summarizes the temporal information of the community assembly (although in some studies time is explicitly included into the algorithm, e.g. in Jabot et al. 2020). However, in a community influenced by strong stochasticity (e.g. as a result of frequent disturbance or fluctuating environmental conditions), the species composition may keep changing over time and the iteration time to determine the inputted data may significantly influence the estimation of the RFs. For such communities, the variance of parametric estimation should be of concern for application.

We acknowledge that the complexity of the simulated data does not fully encompass the complexity of species interactions, resource usage, and dispersal strategies in real-world metacommunities. Also, the performance of the integrated framework in estimating the underlying ecological processes still needs improvement. To address this, we promote the research in constructing a more comprehensive process-based model that can incorporate the complexity of biotic and abiotic response and dispersal strategies of the species. Moreover, while developing new statistics to summarize the spatio-temporal community turnover, it is also essential to systematically evaluate the effectiveness of current or novel summary statistics in detecting signals of ecological processes. This will facilitate the synthesis of these metrics and obtain more precise estimations of ecological processes' strength.

Conclusions

Guzman et al. (2022) introduced a framework that integrates multiple analytical methods and applied them to multiple snapshots of simulated metacommunity to quantify the strength of underlying assembly processes. We show its theoretical values in studying the performance of different statistics and indices in identifying ecological processes. We propose using visualization and importance quantification as approaches to systematically compare their effectiveness. Furthermore, we highlight the practical values of this framework in assessing the strength of ecological processes underlying real-world metacommunities. Our study demonstrates the flexibility of this framework in incorporating different analytical methods and its generality to be applied to different community systems. However, there is a need to improve the effectiveness of this framework in understanding the underlying assembly processes. Further synthesis of statistics and indices that capture specific ecological processes is required. Additionally, it is necessary to evaluate their robustness against the complexity of the baseline simulated datasets and the incompleteness of community data.

Acknowledgements – The authors thank Dr Po-Ju Ke for providing valuable theoretical insights, Dr Corentin Gibert for offering practical suggestions on DNCI calculation, Dr Patrick L. Thompson for insightful explanations of the process-based model and Editor François Munoz, Dr Laura Melissa Guzman for their insightful comments on the manuscript.

Funding – Funding was provided by the Ministry of Science and Technology, Taiwan (MOST, 109-2621-B-002-002-MY3).

Author contributions

Ching-Lin Huang: Conceptualization (lead); Data curation (equal); Formal analysis (lead); Methodology (lead); Visualization (lead); Writing – original draft (lead). **David Zelený:** Conceptualization (equal); Funding acquisition (lead); Methodology (supporting); Supervision (lead); Writing – original draft (supporting). **Chia-Hao Chang-Yang:** Data curation (equal); Investigation (equal); Writing – original draft (supporting).

Data availability statement

Data are available from the Zenodo Digital Repository: https://zenodo.org/records/10153872 (Huang et al. 2023).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Adler, P. B., HilleRisLambers, J. and Levine, J. M. 2007. A niche for neutrality. – Ecol. Lett. 10: 95–104.

Adler, P. B., Ellner, S. P. and Levine, J. M. 2010. Coexistence of perennial plants: an embarrassment of niches. – Ecol. Lett. 13: 1019–1029.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. 2017. Julia: a fresh approach to numerical computing. – SIAM Rev. 59: 65–98.

Borcard, D. and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. – Ecol. Modell. 153: 51–68.

Borcard, D., Legendre, P. and Drapeau, P. 1992. Partialling out the spatial component of ecological variation. – Ecology 73: 1045–1055.

Borics, G., B-Béres, V., Bácsi, I., Lukács, B. A., T-Krasznai, E., Botta-Dukát, Z. and Várbíró, G. 2020. Trait convergence and trait divergence in lake phytoplankton reflect community assembly rules. – Sci. Rep. 10: 19599.

Boucher, D. H., James, S. and Keeler, K. H. 1982. The ecology of mutualism. – Annu. Rev. Ecol. Syst. 13: 315–347.

ter Braak, C. J. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. – Ecology 67: 1167–1179.

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

Chang, L.-W., Zelený, D., Li, C.-F., Chiu, S.-T. and Hsieh, C.-F. 2013. Better environmental data may reverse conclusions about niche-and dispersal-based processes in community assembly. – Ecology 94: 2145–2151.

Chang-Yang, C.-H., Lu, C.-L., Sun, I.-F., Hsieh, C.-F. 2013. Long-term seedling dynamics of tree species in a subtropical rain forest, Taiwan. – Taiwania 58: 35–43.

Chang-Yang, C.-H., Needham, J., Lu, C.-L., Hsieh, C.-F., Sun, I.-F. and McMahon, S. M. 2021. Closing the life cycle of forest trees: the difficult dynamics of seedling-to-sapling transitions in a subtropical rainforest. – J. Ecol. 109: 2705–2716.
Chase, J. M. and Myers, J. A. 2011. Disentangling the importance of ecological niches from stochastic processes across scales. – Phil. Trans. R. Soc. B 366: 2351–2363.
Chase, J. M., Jeliazkov, A., Ladouceur, E. and Viana, D. S. 2020. Biodiversity conservation through the lens of metacommunity ecology. – Ann. N. Y. Acad. Sci. 1469: 86–104.
Chave, J., Coomes, D., Jansen, S., Lewis, S. L., Swenson, N. G. and Zanne, A. E. 2009. Towards a worldwide wood economics spectrum. – Ecol. Lett. 12: 351–366.
Chesson, P. 2000. Mechanisms of maintenance of species diversity. – Annu. Rev. Ecol. Syst. 31: 343–366.
Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J. A., Lodge, D. M., Pascual, M., Pielke, R. Jr., Pizer, W., Pringle, C., Reid, W. V., Rose, K. A., Sala, O., Schlesinger, W. H., Wall, D. H. and Wear, D. 2001. Ecological forecasts: an emerging imperative. – Science 293: 657–660.
Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. – Aust. J. Ecol. 18: 117–143.
Condit, R. 1998. Tropical forest census plots: methods and results from Barro Colorado Island, Panama and a comparison with other plots. – Springer Science & Business Media.
Connolly, S. R., Keith, S. A., Colwell, R. K. and Rahbek, C. 2017. Process, mechanism, and modeling in macroecology. – Trends Ecol. Evol. 32: 835–844.
Connor, E. F. and Simberloff, D. 1979. The assembly of species communities: chance or competition? – Ecology 60: 1132–1140.
Cornelissen, J. H. C., Lavorel, S., Garnier, E., Díaz, S., Buchmann, N., Gurvich, D. E., Reich, P. B., Ter Steege, H., Morgan, H. D., Van Der Heijden, M., Pausas, J. G. and Poorter, H. 2003. A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. – Aust. J. Bot. 51: 335–380.
Cottenie, K. 2005. Integrating environmental and spatial processes in ecological community dynamics. – Ecol. Lett. 8: 1175–1182.
Diamond, J. M. 1975. The island dilemma: lessons of modern biogeographic studies for the design of natural reserves. – Biol. Conserv. 7: 129–146.
Evans, M. R. 2012. Modelling ecological systems in a changing world. – Phil. Trans. R. Soc. B 367: 181–190.
Ford, B. M. and Roberts, J. D. 2020. Functional traits reveal the presence and nature of multiple processes in the assembly of marine fish communities. – Oecologia 192: 143–154.
Gibert, C. and Escarguel, G. 2019. Per-simper – a new tool for inferring community assembly processes from taxon occurrences. – Global Ecol. Biogeogr. 28: 374–385.
Gotelli, N. J. and McGill, B. J. 2006. Null versus neutral models: what's the difference? – Ecography 29: 793–800.
Gotelli, N. J. and Ulrich, W. 2012. Statistical challenges in null model analysis. – Oikos 121: 171–180.
Grinnell, J. 1917. The niche-relationships of the California thrasher. – Auk 34: 427–433.
Guzman, L. M., Thompson, P. L., Viana, D. S., Vanschoenwinkel, B., Horváth, Z., Ptacnik, R., Jeliazkov, A., Gascón, S., Lemmens, P., Anton-Pardo, M., Langenheder, S., De Meester, L. and Chase, J. M. 2022. Accounting for temporal change in multiple biodiversity patterns improves the inference of metacommunity processes. – Ecology 103: e3683.
Han, H., Guo, X. and Yu, H. 2016. Variable selection using mean decrease accuracy and mean decrease Gini based on random forest. – In: 2016 7th IEEE international conference on software engineering and service science (ICSESS). IEEE Publications, pp. 219–224.
Hodgson, E. E. and Halpern, B. S. 2019. Investigating cumulative effects across ecological scales. – Conserv. Biol. 33: 22–32.
Hornung, R. 2020. Ordinal forests. – J. Classif. 37: 4–17.
Hornung, R. 2021. ordinalForest: ordinal forests: prediction and variable ranking with ordinal target variables. – R package ver. 2.4-2. https://cran.r-project.org/web/packages/ordinalForest/
Huang, C.-L., Zelený, D. and Chang-Yang, C.-H. 2023. Data from: Integrating several analytical methods to assess strength of ecological processes behind metacommunity assembly. – Zenodo Digital Repository, https://zenodo.org/records/10153872.
Hubbell, S. P. 2011. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.
Jabot, F., Laroche, F., Massol, F., Arthaud, F., Crabot, J., Dubart, M., Blanchet, S., Munoz, F., David, P. and Datry, T. 2020. Assessing metacommunity processes through signatures in spatiotemporal turnover of community composition. – Ecol. Lett. 23: 1330–1339.
Janitza, S., Tutz, G. and Boulesteix, A. L. 2016. Random forest for ordinal responses: prediction and variable selection. – Comp. Stat. Data Anal. 96: 57–73.
Jops, K. and O'Dwyer, J. P. 2023. Life history complementarity and the maintenance of biodiversity. – Nature 618: 986–991.
Kraft, N. J., Comita, L. S., Chase, J. M., Sanders, N. J., Swenson, N. G., Crist, T. O., Stegen, J. C., Vellend, M., Boyle, B., Anderson, M. J., Cornell, H. V., Davies, K. F., Freestone, A. L., Inouye, B. D., Harrison, S. P. and Myers, J. A. 2011. Disentangling the drivers of β diversity along latitudinal and elevational gradients. – Science 333: 1755–1758.
Kraft, N. J., Adler, P. B., Godoy, O., James, E. C., Fuller, S. and Levine, J. M. 2015. Community assembly, coexistence and the environmental filtering metaphor. – Funct. Ecol. 29: 592–599.
Lasky, J. R., Sun, I.-F., Su, S.-H., Chen, Z.-S. and Keitt, T. H. 2013. Trait-mediated effects of environmental filtering on tree community dynamics. – J. Ecol. 101: 722–733.
Legendre, P. and Legendre, L. 2012. Numerical ecology. – Elsevier.
Leibold, M. A. and Chase, J. M. 2017. Metacommunity ecology. – Princeton Univ. Press.
Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., Holt, R. D., Shurin, J. B., Law, R., Tilman, D., Loreau, M. and Gonzalez, A. 2004. The metacommunity concept: a framework for multi-scale community ecology. – Ecol. Lett. 7: 601–613.
Levins, R. 1966. The strategy of model building in population biology. – Am. Sci. 54: 421–431.
Liaw, A. and Wiener, M. 2002. Classification and regression by randomforest. – R news 2: 18–22.
MacArthur, R. H. 1958. Population ecology of some warblers of northeastern coniferous forests. – Ecology 39: 599–619.
MacArthur, R. and Levins, R. 1967. The limiting similarity, convergence, and divergence of coexisting species. – Am. Nat. 101: 377–385.
MacArthur, R. H. and Wilson, E. O. 1967. The theory of island biogeography. – Princeton Univ. Press.
Mayfield, M. M. and Levine, J. M. 2010. Opposing effects of competitive exclusion on the phylogenetic structure of communities. – Ecol. Lett. 13: 1085–1093.

McFadden, I. R. et al. 2023. Linking human impacts to community processes in terrestrial and freshwater ecosystems. – Ecol. Lett. 26: 203–218.
Molina, C. and Stone, L. 2020. Difficulties in benchmarking ecological null models: an assessment of current methods. – Ecology 101: e02945.
Mouquet, N. and Loreau, M. 2003. Community patterns in source-sink metacommunities. – Am. Nat. 162: 544–557.
Munoz, F., Grenié, M., Denelle, P., Taudière, A., Laroche, F., Tucker, C. and Violle, C. 2018. ecolottery: simulating and assessing community assembly with environmental filtering and neutral dynamics in R. – Methods Ecol. Evol. 9: 693–703.
Ning, D., Deng, Y., Tiedje, J. M. and Zhou, J. 2019. A general framework for quantitatively assessing ecological stochasticity. – Proc. Natl Acad. Sci. USA 116: 16892–16898.
Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. and Abrego, N. 2017. How to make more out of community data? a conceptual framework and its implementation as models and software. – Ecol. Lett. 20: 561–576.
Ovaskainen, O., Rybicki, J. and Abrego, N. 2019. What can observational data reveal about metacommunity processes? – Ecography 42: 1877–1886.
Overcast, I., Ruffley, M., Rosindell, J., Harmon, L., Borges, P. A. V., Emerson, B. C., Etienne, R. S., Gillespie, R., Krehenwinkel, H., Mahler, D. L., Massol, F., Parent, C. E., Patiño, J., Peter, B., Week, B., Wagner, C., Hickerson, M. J. and Rominger, A. 2021. A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities. – Mol. Ecol. Resour. 21: 2782–2800.
Pakeman, R. J. 2014. Functional trait metrics are sensitive to the completeness of the species' trait data? – Methods Ecol. Evol. 5: 9–15.
Peres-Neto, P. R., Legendre, P., Dray, S. and Borcard, D. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. – Ecology 87: 2614–2625.
Pilosof, S., Porter, M. A., Pascual, M. and Kéfi, S. 2017. The multilayer nature of ecological networks. – Nat. Ecol. Evol. 1: 0101.
Schlather, M., Malinowski, A., Menck, P. J., Oesting, M. and Strokorb, K. 2015. Analysis, simulation and prediction of multivariate random fields with package random fields. – J. Stat. Softw. 63: 1–25.
Shen, G., He, F., Waagepetersen, R., Sun, I. F., Hao, Z., Chen, Z. S. and Yu, M. 2013. Quantifying effects of habitat heterogeneity and other clustering processes on spatial distributions of tree species. – Ecology 94: 2436–2443.
Sîrbu, I., Benedek, A. M. and Sîrbu, M. 2021. Variation partitioning in double-constrained multivariate analyses: linking communities, environment, space, functional traits, and ecological niches. – Oecologia 197: 43–59.
Smith, T. W. and Lundholm, J. T. 2010. Variation partitioning as a tool to distinguish between niche and neutral processes. – Ecography 33: 648–655.
Stegen, J. C., Lin, X., Fredrickson, J. K., Chen, X., Kennedy, D. W., Murray, C. J., Rockhold, M. L. and Konopka, A. 2013. Quantifying community assembly processes and identifying features that impose them. – ISME J. 7: 2069–2079.
Su, S.-H., Chang-Yang, C.-H., Lu, C.-L., Tsui, C.-C., Lin, T.-T., Lin, C.-L., Chiou, W.-L., Kuan, L.-H., Chen, Z.-S. and Hsieh, C.-F. 2007. Fushan subtropical forest dynamics plot: tree species characteristics and distribution patterns. – Taiwan For. Res. Inst.
Su, S., Guan, B. T., Chang-Yang, C., Sun, I., Wang, H. and Hsieh, C. 2020. Multi-stemming and size enhance survival of dominant tree species in a frequently typhoon-disturbed forest. – J. Veg. Sci. 31: 429–439.
Thompson, P. L., Guzman, L. M., De Meester, L., Horváth, Z. and Ptacnik, R., Van-schoenwinkel, B., Viana, D. S. and Chase, J. M. 2020. A process-based metacommunity framework linking local and regional scale community ecology. – Ecol. Lett. 23: 1314–1329.
Tilman, D. 1997. Community invasibility, recruitment limitation, and grassland biodiversity. – Ecology 78: 81–92.
Tucker, C. M., Shoemaker, L. G., Davies, K. F., Nemergut, D. R. and Melbourne, B. A. 2016. Differentiating between niche and neutral assembly in metacommunities using null models of β-diversity. – Oikos 125: 778–789.
Tuomisto, H., Ruokolainen, L. and Ruokolainen, K. 2012. Modelling niche and neutral dynamics: on the ecological interpretation of variation partitioning results. – Ecography 35: 961–971.
Ulrich, W. and Gotelli, N. J. 2010. Null model analysis of species associations using abundance data. – Ecology 91: 3384–3397.
van der Vaart, E., Beaumont, M. A., Johnston, A. S. A. and Sibly, R. M. 2015. Calibration and evaluation of individual-based models using approximate bayesian computation. – Ecol. Modell. 312: 182–190.
Vellend, M., Srivastava, D. S., Anderson, K. M., Brown, C. D., Jankowski, J. E., Kleynhans, E. J., Kraft, N. J. B., Letaw, A. D., Macdonald, A. A. M., Maclean, J. E., Myers-Smith, I. H., Norris, A. R. and Xue, X. 2014. Assessing the relative importance of neutral stochasticity in ecological communities. – Oikos 123: 1420–1430.
Vilmi, A., Gibert, C., Escarguel, G., Happonen, K., Heino, J., Jamoneau, A., Passy, S. I., Picazo, F., Soininen, J., Tison-Rosebery, J. and Wang, J. 2021. Dispersal–niche continuum index: a new quantitative metric for assessing the relative importance of dispersal versus niche processes in community assembly. – Ecography 44: 370–379.
Volterra, V. 1928. Variations and fluctuations of the number of individuals in animal species living together. – ICES J. Mar. Sci. 3: 3–51.
Wilson, D. S. 1992. Complex interactions in metacommunities, with implications for biodiversity and higher levels of selection. – Ecology 73: 1984–2000.
Yao, A.-W., Chiang, J.-M., McEwan, R. and Lin, T.-C. 2015. The effect of typhoon-related defoliation on the ecology of gap dynamics in a subtropical rain forest of Taiwan. – J. Veg. Sci. 26: 145–154.