PNAS
www.pnas.org

Supplementary Information for

A General Framework for Quantitatively Assessing Ecological Stochasticity

Daliang Ning, Ye Deng, James M. Tiedje, and Jizhong Zhou

Jizhong Zhou
Email: jzhou@ou.edu

This PDF file includes:

Supplementary Text
Figs. S1 to S6
Tables S1 to S4
Supplementary References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1904623116

Supplementary Text

A. Community similarity/dissimilarity metrics

Various similarity/dissimilarity metrics have been applied in ecological research to measure β diversity, either by taxonomic or phylogenetic β-diversity metrics. Many metrics have both incidence-based (qualitative) and abundance-based (quantitative) formats. Table S3 summarizes commonly used taxonomic (1-14) and phylogenetic (15-20) similarity/dissimilarity metrics. Please see Parks and Beiko's paper (17) for more phylogenetic β-diversity metrics (39 different indexes). The taxonomic diversity metrics used in our study were calculated using function "vegdist" and/or "designdist" in R package "vegan" (21).

In principle, our method requires the value of the metric ranging from 0 to 1, and the complementarity of similarity (C) and dissimilarity (D) indexes ($C = 1 - D$). Many taxonomic and phylogenetic metrics satisfy this requirement, such as Jaccard, Ružička, Sørensen, Bray-Curtis, Kulczynski, Gower, Canberra, Morisita-Horn, unweighted Unifrac, and Phylosor, which can be directly implemented with our method. However, quite a few metrics do not have fixed upper limit and/or not have clear defined similarity measure, such as Euclidean, Manhattan, Binomial, Cao, modified Gower, βMNTD, βMPD, and most of weighted phylogenetic dissimilarity metrics. These metrics need to be standardized to meet the requirement before being applied to our method. Inspired by what Cao et al (14) did to define similarity from Cao dissimilarity index, we proposed a general method as follows.

$$D_{ij} = \frac{D'_{ij} - min\{D'\}}{max\{D'\} - min\{D'\}} = \frac{D'_{ij}}{D'_{max}} \qquad \textbf{Eq. S1}$$

where $\min\{D'\} = 0$, considering the lower limit of dissimilarity metrics is always zero.

$$D'_{max} = \begin{cases} D'_{up} & if\ upper\ limit\ is\ fixed \\ \max\limits_{ij}\{D'_{ij}, G'_{\max ij}\} & if\ upper\ limit\ is\ not\ fixed \end{cases} \qquad \textbf{Eq. S2}$$

$$C_{ij} = 1 - D_{ij} = 1 - \frac{D'_{ij}}{D'_{max}} \qquad \textbf{Eq. S3}$$

$$G_{ij} = \frac{G'_{ij}}{D'_{max}} \qquad \textbf{Eq. S4}$$

$$E_{ij} = 1 - G_{ij} = 1 - \frac{G'_{ij}}{D'_{max}} \qquad \textbf{Eq. S5}$$

D_{ij} Standardized dissimilarity between community *i* and *j*.
D'_{ij} Unstandardized (original) dissimilarity between community *i* and *j*.
C_{ij} Standardized similarity between community *i* and *j*.
D'_{max} Probable maximum unstandardized dissimilarity.
D'_{up} Defined upper limit of original dissimilarity, e.g. for Bray-Curtis, $D'_{up} = 1$.
G_{ij} Standardized null (randomly expected) dissimilarity between community *i* and *j*.
G'_{ij} Unstandardized null dissimilarity calculated between community *i* and *j*.

E_{ij} Standardized null similarity calculated between community i and j.

E'_{ij} Unstandardized null similarity calculated between community i and j.

$G'_{max\,ij}$ The estimated maximum value of null dissimilarity between community i and j. It is calculated as 3 times of smoothing bandwidth beyond the maximum simulated value of the dissimilarity D'_{ij}, using function “density” with Gaussian model in R package “stats”.

Our method described in the main text is based on similarity and dissimilarity metrics ranging from 0 to 1 or standardized according to above equations. If the upper limit of dissimilarity is not fixed, D'_{max} is still not a fixed value but depends on community data matrix and null model algorithms, which leads to the uncertainty of metrics standardization. Therefore, the original metrics ranging from 0 to 1 (i.e. there is no need for standardization) are preferred if they have the same performance in terms of accuracy and precision as standardized metrics.

We tested thirteen incidence-based metrics, which can be classified into three major categories:

i. Unique-ratio metrics: the dissimilarity is measured by the ratio of unique taxa (i.e. the taxa only observed in one of the two samples). Jaccard is the unique taxa number divided by total observed taxa number in two samples, which is exactly the same as incidence-based Canberra, modified Gower (mGower), and modified Manhattan (mManhattan), and will be incidence-based Cao if multiplied with a constant. Sørensen is the unique taxa number divided by the sum of observed taxa number in two samples, which is exactly the same as incidence-based Morisita-Horn. Kulczynski is the mean of unique taxa percentage in each sample, and Gower is the unique taxa number divided by observed taxa number across all samples. Thus, these nine metrics can be classified into the same type named unique-ratio metrics.

ii. Unique-number metrics: the dissimilarity is directly measured by the number of unique taxa in two samples. Incidence-based Manhattan is defined as this, and incidence-based Binomial is Manhattan multiplied with a constant.

iii. Squared-root metrics: the metrics are calculated from squared root of unique taxa number. This type includes incidence-based Euclidean and modified Euclidean (mEuclidean) which is Euclidean divided by total observed taxa number in the two samples.

We also tested fifteen abundance-based metrics, which can be divided into four major groups as below:

i. Relative-difference metrics: the abundance difference (or the relatively smaller abundance) of each taxon between two samples is divided by the abundances of the taxon in the samples before or after summed up. Ružička, Bray-Curtis, and Kulczynski definitely belong to this type. In Chao’s formula, the total number of individuals in the taxa shared by the two samples (Ci) is divided by the total number of individuals in a sample (i.e. abundance sum in a sample), thus Chao is also classified into this type.

ii. Average-relative-difference metrics: the sum of relative difference between two samples (or other value represent relative difference) is further divided by total taxa number in the samples. Canberra is a typical metric defined as average relative difference of taxon abundance between two samples. In the equation of mGower, the numerator is calculated from the difference of logarithmic transformed abundances which is equal to the ratio between larger and smaller abundance before logarithmic transformation, thus can be regarded as relative difference. And the denominator is total taxa number in the two samples, thus mGower can be classified as average relative difference. Cao also has a numerator related to the ratio of each taxon's abundance between two samples and the total taxa number as the denominator, thus belongs to this type.

iii. Absolute-difference metrics: the abundance difference between two samples is not divided by the taxa abundances in the two samples. Manhattan is defined as the sum of absolute abundance difference. mManhattan is Manhattan divided by total taxa number in the two samples rather than any abundance-related value, thus classified into this type. Gower and Binomial appear like relative-abundance metrics, but usually show stronger correlation with Manhattan or mManhattan than other relative-abundance metrics. For example, in the empirical data used in this study, Gower and Binomial showed obviously higher correlation coefficients with Manhattan (r=0.964 and 0.897) than with Bray-Curtis (r=0.359 and 0.558). Thus, they are classified into this type.

iv. Squared-sum metrics: the metrics are calculated from the sum of squared abundance difference or product of abundances in two samples. Euclidean is squared root of the squared difference sum and mEuclidean is Euclidean divided by total taxa number in the two samples. Morisita and Morisita-horn are calculated from the product of abundances of each taxon in two samples. By some mathematical deviation, Morisita-horn is actually the squared sum of each taxon's proportion difference divided by the sum of squared proportions, i.e. $\left[\sum_k (p_{ik} - p_{jk})^2\right] / (\sum_k p_{ik}^2 + \sum_k p_{jk}^2)$, where p_{ik} and p_{jk} are the proportions of taxon k in sample i and j, respectively. and Morisita has some minor difference. Thus, these four metrics are classified as a type of squared-sum.

B. Normalization of stochasticity ratio

Intuitively, the indexes measuring stochasticity and determinism are expected to range from 0% to 100%, and reach the extreme values when community assembly is completely deterministic or stochastic. We defined stochasticity ratio (ST_{ij}) as the ratio of average null expectation ($\overline{E_{ij}}$ or $\overline{G_{ij}}$) to observed similarity (C_{ij}) or dissimilarity (D_{ij}). Because null expectation is calculated from null model which simulates stochastic assembly, when community assembly is highly stochastic, the average observed similarity or dissimilarity can be very close to the average null expectation, and *ST* can approach the accurate value of stochasticity (i.e. 100%). However, also because null model simulates stochastic assembly, the average null expectation always has substantial deviations from 0, no matter how deterministic the observed similarity or dissimilarity is. Therefore, when the community assembly is highly deterministic, the expected stochasticity approaches 0%, but the values of ST_{ij} always has substantial deviations from 0%. It means that *ST* would obviously overestimate stochasticity when expected stochasticity is very low, although it could be relatively accurate when expected stochasticity is high. Thus, we applied the following formula to obtain normalized selection strength (*NSS*) and normalized stochasticity ratio (*NST*).

$$NSS = \frac{SS - {}^{T}SS}{{}^{D}SS - {}^{T}SS} \quad \textbf{Eq. S6}$$

$$NST = 1 - NSS = \frac{{}^{D}SS - SS}{{}^{D}SS - {}^{T}SS} \quad \textbf{Eq. S7}$$

where SS is the observed selection strength, ${}^{D}SS$ and ${}^{T}SS$ are the theoretical extreme values of SS under completely deterministic and stochastic assembly, respectively. After normalization, when community assembly is completely deterministic, SS is equal to ${}^{D}SS$, NSS will be 100%, and NST will be 0%. When community assembly is completely stochastic, SS will be equal to ${}^{T}SS$, NSS will be 0% and NST will be 100%. Thus, NSS and NST are theoretically better than SS and ST for measuring determinism and stochasticity in community assembly.

Before further derivation, we introduce a generalized function ξ to make equations simpler.

$$\xi(x, y) = \frac{x - y}{x - \delta} \quad \delta = \begin{cases} 0 & x \geq y \\ 1 & x < y \end{cases} \quad \textbf{Eq. S8}$$

If we set x as the observed similarity between community *i* and *j* (C_{ij}), and set y as the average null expectation of the similarity between community *i* and *j* ($\overline{E_{ij}}$), $x \geq y$ means type A situation, $x < y$ means type B situation, and $\xi\left(C_{ij}, \overline{E_{ij}}\right)$ is the same as our definition of *SS* between community *i* and *j*. Thus, we can simplify Eq. 1 and Eq. 3 in the main text into one equation as below.

$$SS_{ij} = \xi\left(C_{ij}, \overline{E_{ij}}\right) \quad \textbf{Eq. S9}$$

and

$$\overline{E_{ij}} = \frac{\sum_{k=1}^{N_r} E_{ij}^{(k)}}{N_r} \quad \textbf{Eq. S10}$$

where $E_{ij}^{(k)}$ is the null similarity between community *i* and *j* at the *k*th randomization time of null model analysis, and N_r is the randomization time of null model, which is usually set as 1000 times.

To estimate ${}^{D}SS$, we consider two extreme situations. If the deterministic factors lead to more similar community structure ($C_{ij} \geq \overline{E_{ij}}$, type A situation), the extremely deterministic assembly should have the similarity, ${}^{D}C_{ij}$, approaching to the maximum value of 1. In contrast, if the deterministic factors lead to more dissimilar community structure ($C_{ij} < \overline{E_{ij}}$, type B situation), the extremely deterministic assembly should have the dissimilarity close to the maximum and the similarity, ${}^{D}C_{ij}$, close to 0. Thus, ${}^{D}SS_{ij}$ and ${}^{D}SS$ can be estimated by following equations.

$${}^{D}C_{ij} = \begin{cases} 1 & C_{ij} \geq \overline{E_{ij}} \\ 0 & C_{ij} < \overline{E_{ij}} \end{cases} \quad \textbf{Eq. S11}$$

$${}^{D}SS_{ij} = \xi\left({}^{D}C_{ij}, \overline{E_{ij}}\right) \quad \textbf{Eq. S12}$$

$${}^{D}SS = \frac{\sum_{ij} {}^{D}SS_{ij}}{n} = \frac{\sum_{ij} \xi\left({}^{D}C_{ij}, \overline{E_{ij}}\right)}{n} \quad \textbf{Eq. S13}$$

where *n* is the number of pairwise comparisons.

Before estimating ${}^{T}SS$, we would like to explain why ${}^{T}SS$ cannot be simply set as zero. We need to consider the "uncertainty" of similarity/dissimilarity of communities when they are under completely stochastic assembly. Here "uncertainty" means that, similarity of each pairwise comparison under completely stochastic assembly (${}^{T}C_{ij}$) has probability to be any value within the range of null expectation, because of the randomness of stochastic assembly. The completely stochastic assembly can be simulated by null model. ${}^{T}C_{ij}$ can be estimated as null similarity E_{ij} which is not a certain value but a distribution $\left\{E_{ij}^{(k)}\right\}_k$ with highest probability usually at the average null expectation $\overline{E_{ij}}$. The estimated *SS* under complete stochastic assembly ${}^{T}SS$ is the average relative deviation of E_{ij} from the mean $\overline{E_{ij}}$. Similar to standard deviation, ${}^{T}SS$ value depends on variance of E_{ij} and could be equal to zero only if the variances of E_{ij} in every pairwise comparison are all equal to zero. Due to the randomness of stochastic assembly, E_{ij} always has variance larger than zero, thus ${}^{T}SS$ can never be zero.

To estimate ${}^{T}SS$, we simulate stochastic assembly by randomizing the observed community structure with a null model algorithm for as many times as necessary (usually N_r=1000 times). At each time of randomization, the *SS* value of each null pairwise comparison ${}^{T}SS_{ij}^{(k)}$ can be calculated from the null similarity $E_{ij}^{(k)}$. Then, we can obtain the average *SS* value of the null communities at each randomization time ${}^{T}SS^{(k)}$. To ensure the index *NSS* will not exceed 100%, ${}^{T}SS$ is calculated as the minimum value of $\left\{{}^{T}SS^{(k)}\right\}_k$. Altogether, ${}^{T}SS$ can be estimated as following equations.

$${}^{T}SS_{ij}^{(k)} = \xi\left(E_{ij}^{(k)}, \overline{E_{ij}}\right) \qquad \textbf{Eq. S14}$$

$${}^{T}SS = \min_k\left\{{}^{T}SS^{(k)}\right\} \qquad \textbf{Eq. S15}$$

Altogether, *NSS* and *NST* are calculated as below.

$$NSS = \frac{SS - {}^{T}SS}{{}^{D}SS - {}^{T}SS} = \frac{\sum_{ij}\xi\left(C_{ij}, \overline{E_{ij}}\right) - \min_k\left\{\sum_{ij}\xi\left(E_{ij}^{(k)}, \overline{E_{ij}}\right)\right\}}{\sum_{ij}\xi\left({}^{D}C_{ij}, \overline{E_{ij}}\right) - \min_k\left\{\sum_{ij}\xi\left(E_{ij}^{(k)}, \overline{E_{ij}}\right)\right\}} \qquad \textbf{Eq. S16}$$

$$NST = \frac{{}^{D}SS - SS}{{}^{D}SS - {}^{T}SS} = \frac{\sum_{ij}\xi\left({}^{D}C_{ij}, \overline{E_{ij}}\right) - \sum_{ij}\xi\left(C_{ij}, \overline{E_{ij}}\right)}{\sum_{ij}\xi\left({}^{D}C_{ij}, \overline{E_{ij}}\right) - \min_k\left\{\sum_{ij}\xi\left(E_{ij}^{(k)}, \overline{E_{ij}}\right)\right\}} \qquad \textbf{Eq. S17}$$

Because such indexes are originally derived from every pairwise comparison, they are not independent. The distribution of *NSS* or *NST* is unknown and probably not normal. Therefore, the nonparametric permutation test, permutational multivariate analysis of variance (PERMANOVA), is used to examine whether the communities under different conditions differ in their *NSS* and *NST*. The ST and NST calculation and PERMANOVA test can be performed using the function "NST" on a web-based pipeline (http://ieg3.rccc.ou.edu:8080/) built on Galaxy platform (22) or a R package "NST".

C. Estimating stochasticity in simulated communities

C1. Simulation models

(a) Spatially implicit simulation model

We built a spatially implicit simulation model to obtain a total of **21 datasets** with the expected abundance-based stochasticity ranging from 0% to 100% (5% interval, scenario A in Table S1 and Fig. S1a). Each dataset has two groups of local communities from **2 plots** under distinct environments (e.g. very hot and cold environments). The two plots share the same **metacommunity**. Each plot has **12 local communities** as biological replicates. In each local community, the total richness and total abundances of deterministic and stochastic species are set according to a certain expected stochasticity. The total abundance of microorganisms in each local community is set as 20,000, which is a normal sequencing depth of 16S rRNA gene in many microbial community studies.

The **metacommunity structure** (i.e. abundance of each species in the metacommunity) is generated according to metacommunity zero-sum multinomial distribution (mZSM) (23) derived from Hubbell's Unified Neutral Theory Model (24), using R package "sads" (25) with $J=10^8$, $\theta=5000$, and 10,000 species sampled. In each **local community**, the **stochastic species** are simulated as a random draw of 100 species (i.e. the assigned richness of stochastic species in the local community) from metacommunity, with probabilities proportional to their regional frequencies. The regional frequency of a stochastic species is calculated from its regional relative abundance according to Sloan's Neutral Model (26) which was also derived from Hubbell's neutral theory and particularly developed for microbial communities. The dispersal rate (m) is set as 0.1. The abundances of stochastic species in a local community are simulated as a random draw of a certain number of individuals (i.e. the assigned total abundance of stochastic species in the local community) from metacommunity into the stochastic species in this community, with probabilities proportional to the regional relative abundances of the species. We set only two types of **deterministic species**: one is thermophilic, and the other is psychrophilic. The local communities from hot environment have equal abundances of the thermophilic species, but no psychrophilic species. The communities from cold environment are under exactly opposite situation, such that the similarity of deterministic species is 100% within group and 0% between groups.

The expected stochasticity in a simulated community can be defined as incidence-based or abundance-based measures as below.

$$ST_{exp.in} = \frac{S_t}{S_t + S_d} \qquad \textbf{Eq. S18}$$

$$ST_{exp.ab} = \frac{J_t}{J_t + J_d} \qquad \textbf{Eq. S19}$$

$ST_{exp.in}$ Incidence-based expected stochasticity in a simulated local community.
$ST_{exp.ab}$ Abundance-based expected stochasticity in a simulated local community.
S_d Richness of deterministic species in a simulated local community.
J_d Total abundance of deterministic species in a simulated local community.
S_t Richness of stochastic species in a simulated local community.
J_t Total abundance of stochastic species in a simulated local community.

(b) Spatially explicit simulation model

To examine scale dependence of stochasticity estimation, we built a spatially explicit simulation model (Scenario B-F in Table S1, Fig. 2a, and Fig. S1b). The model has four-level metacommunities, including local (for each site), regional, continental, and global metacommunities. In the model, an area of 16,384 (128×128) cells are divided into 4 (2×2) continents, each continent is divided into 4 (2×2) regions, and each region is divided into 4 (2 × 2) sites. Each site has 4 (2×2) plots, sharing the same local metacommunity. Each plot has 64 (8×8) cells, and each cell represents a local community with 20,000 individuals (Fig. 2a). We take all individuals from a single cell as a sample, and a certain number of samples from each plot (6 samples/plot unless specified) to get a simulated dataset. Ecological stochasticity was estimated with different indexes based on the pairwise comparisons of all samples within each unit at different spatial scales, i.e. plot, site, region, continent, or global.

To investigate more complicated deterministic forces, **deterministic species** were simulated under three types of scenarios. The **first** scenario (Scenario B in Table S1 and Fig. S1b) is simple abiotic filtering without environmental noise. Plots at the same row (like latitude) have the same temperature, while temperature increases by 2°C per plot along each column (like longitude), from 0°C at the top (northmost) plot to 30°C at the bottom (southmost) plot (Fig. 2a). The temperature is homogeneous within each plot. All local communities (cells) under each temperature have equal abundance of the only deterministic species which prefer this temperature.

The **second** type of scenarios (Scenario C-E in Table S1) is abiotic filtering with environmental noise. The mean temperature of each plot is the same as that in the first scenario, but the temperature in each cell is a random value from a normal distribution with a certain standard deviation (temperature deviation, σ_t). Temperature within each cell is still set homogenous. The abundances of deterministic species in each cell are determined by a Gaussian function as below (Eq. S20). The temperature deviation in each plot is set at different level comparing to fitness deviation (σ_f, defined in Eq. S20), to simulate low (σ_t=5%σ_f, Table S1 scenario C), medium (σ_t=25%σ_f, Table S1 scenario D), and high (σ_t=200%σ_f, Table S1 scenario E) environmental noise.

$$A_{ij} = J_{d0}\exp\left[-\frac{\left(T_j - T_i\right)^2}{2{\sigma_f}^2}\right] \qquad \textbf{Eq. S20}$$

A_{ij}	Abundance of species *i* in local community *j*.
J_{d0}	Expected maximum abundance of deterministic species *i* in a local community.
T_j	Temperature of local community *j*.
T_i	Optimum temperature of species *i*.
σ_f	Fitness deviation, set as 0.4 in this study.

The **third** type of scenario is to consider biotic competition (Table S1 scenario F). Each of the 256 competitors randomly occupies one cell at the very beginning. Then, the competitors randomly disperse to an adjacent cell at each time step, with equal probabilities to all four directions, until all cells are occupied by competitors. In each cell (i.e. each local community), the first-arrived competitor excludes other competitor(s) and stops them passing through the cell.

The **fourth** type of scenario is to investigate community under complex deterministic forces (Table S1 scenario G). In each simulated community with deterministic part, deterministic species controlled by abiotic filtering without environmental noise are simulated as in the first scenario, and then combined at a certain abundance ratio with species controlled by competition which are simulated as in the third types of scenarios.

In different scenarios, **stochastic species** were simulated in the same way as below. First, a global metacommunity was generated in the same way as that in spatially implicit model, with $J=10^9$, $\theta=5000$, and 10,000 species sampled. Second, we developed a two-step random assembly model to simulate stochastic assembly in the spatially explicit model. At the first step, a certain number of species (S_t) are randomly drawn from the higher-level metacommunity to a lower-level (meta)community, according to the expected occurrence frequencies and relative abundance of all species in the higher-level metacommunity as described in the spatially implicit model. The expected occurrence frequencies are calculated according to Sloan's neutral model with a certain dispersal rate (m_1). At the second step, the species (S_t) for each lower-level (meta)community are randomly drawn from three sources, the higher-level metacommunity (with a dispersal rate of m_1), first-step pool of this lower-level (meta)community (m_2), and all first-step pools of adjacent (meta)communities (m_3), to simulate dispersal from higher-level species pool and adjacent communities, respectively. Third, we applied this two-step random assembly model to simulate the (meta)communities at each level. Each continental metacommunity (5,200 species and 8×10^7 individuals) is simulated as two-step random draw from global metacommunity with the dispersal rates of m_1=0.001, m_1=0.997, and m_2=0.002. In the same way, we simulated each regional metacommunity (2,700 species, 2×10^7 individuals, m_1=0.05, m_1=0.8, m_2=0.15), local metacommunity (each site, 1,400 species, 5×10^6 individuals, m_1=0.1, m_1=0.5, m_2=0.4), and local community (each cell, 100 species, m_1=0.2, m_1=0.2, m_2=0.6). In each local community (cell), the total individual number of stochastic species depends on the expected abundance-based stochasticity ($ST_{exp.ab}$). The dispersal rates from the adjacent (meta)community pool (m_2) are higher at lower spatial scales because dispersal is easier at smaller spatial scales.

For each scenario, we combined deterministic and stochastic species at different abundance ratios to generate 11 datasets with expected abundance-based stochasticity ranging from 0% to 100% (10% interval). For each dataset, we estimated stochasticity at different spatial scales with the three indexes and evaluated their accuracy and precision as described below.

C2. Stochasticity indexes

In each dataset, the stochasticity within each group of simulated communities was estimated by *ST* and *NST*, and the neutral species percentage (*NP*). *NST* and *ST* in simulated communities were calculated based on the null model algorithm "PF" (described in part D and Table S4) and various similarity metrics (Table S2).

Sloan et al. (26) developed a neutral model about the relationship between occurrence frequency and relative abundance in source community for microbial communities. We applied Sloan's neutral model to fit the occurrence frequency of each species in a group of communities and the relative abundance in the whole dataset. The species within the 95% confidence interval of Sloan neutral model are defined as neutral species (27). The abundance-weighted and unweighted percentage of neutral species (*NP*) were used to estimate abundance-based and incidence-based stochasticity, respectively. *NP* was calculated using the R codes reported by Burns et al. (27).

Modified Roup-Crick metrics (RC) and standardized effect size (SES) were also applied to communities simulated by spatial implicit model, calculated as previously reported (28-30). The percentage of turnovers with |SES|<2 and that with |RC|<0.95 were counted as stochastic turnover ratio (SR) based on SES (SR_{SES}) and SR based on RC (SR_{RC}), respectively (29). These indexes showed obviously worse quantitative performance than NST and ST for data of spatial implicit model, when calculated based on Bray-Curtis and Ružička (Fig. S6). Thus, we did not further apply them to other simulated data or test various metrics.

C3. Evaluating the accuracy and precision of different stochasticity indexes

We evaluated the performance of each stochasticity index quantitatively by accuracy and precision coefficients. Concordance Correlation Coefficient (CCC) was developed as a measure of agreement between two methods (31). It has meaningful components of accuracy (χ_a, Eq. S21) and precision (ρ, Eq. S22) (32), in which the precision coefficient is the same as Pearson correlation coefficient. Thus, we applied these two coefficients to evaluate the accuracy and precision of stochasticity values estimated by different methods (Table S2). Based on the equation, high accuracy coefficient value means the estimated values have very similar mean and variance as true values. In contrast, high precision coefficient means the variation of estimated values have very similar trend as true values, thus can precisely reflect the relative change of true values. Therefore, a qualified stochasticity index should have high scores in both accuracy and precision coefficients. For example, we assume the true values are 20%, 40%, 60%, 80%, 90% in sequence. When the estimated values are 90%, 60%, 80%, 40%, 20%, the accuracy is very high (χ_a=1) but the precision is very low (ρ=-0.86, negative), thus the index is useless. When the estimated values are 2%, 4%, 6%, 8%, 9%, the accuracy is very low (χ_a=0.04) but the precision is very high (ρ=1), thus the index cannot reflect the magnitude of true value but can be used to estimate the relative changes of true values. When the estimated values are 19%, 41%, 60%, 79%, 91%, the accuracy and precision are both very high (>0.99), thus the index can be used to estimate the true values.

$$\chi_a = \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + \left(\mu_x - \mu_y\right)^2} \qquad \textbf{Eq. S21}$$

$$\rho = \frac{\sigma_{yx}}{\sigma_x\sigma_y} \qquad \textbf{Eq. S22}$$

σ_{yx} Covariance of *x* and *y*. In our study, x is expected stochasticity, and y is estimated stochasticity.

σ_x^2 Variance of *x*.

σ_y^2 Variance of *y*.

μ_x Mean of *x*.

μ_y Mean of *y*.

D. Null model algorithms

In general, there are nine major types of null model algorithms for species co-occurrence analysis, previously elucidated by Gotelli (33) (Table S4). When randomizing the observed communities, different null model algorithms use different ways to constrain the occurrence frequency of each taxon and taxon richness in each sample. We listed the abbreviation and formula to calculate the probability of a taxon present in a sample in each algorithm in Table S4. If abundance weighted metrics are used, after getting occurrence data matrix, abundance can be assigned as random draw of individuals with probabilities proportional to the regional relative abundances of the taxa as previously described by Stegen et al (29). All samples of the empirical dataset were considered as from the same regional species pool, thus randomization was performed across all samples. ST and NST can be calculated based on different null model algorithms and different metrics using the function “NST” on the pipeline (http://ieg3.rccc.ou.edu:8080), or using the function “tNST” in a R package “NST”.

Supplementary Figures

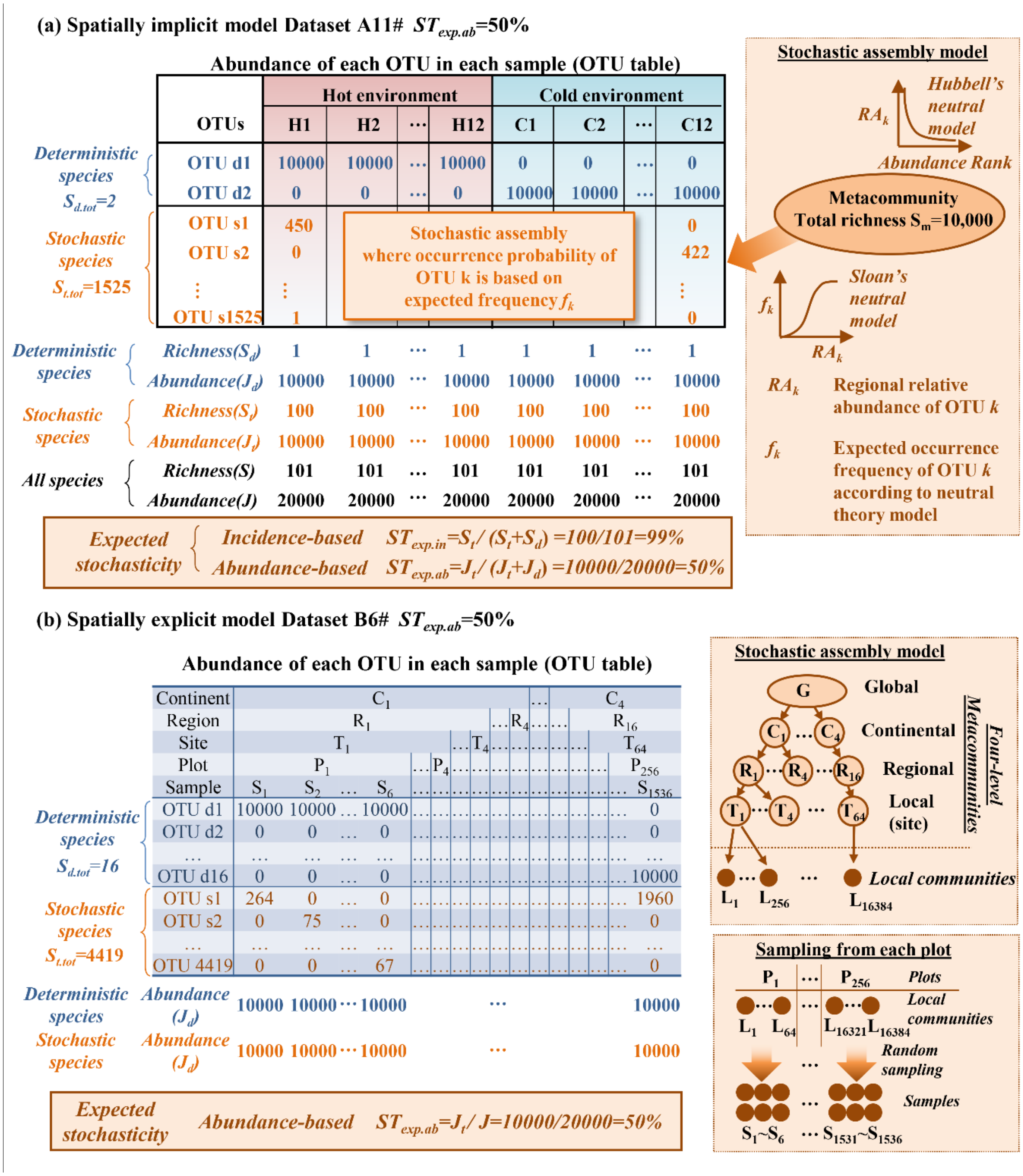

Fig. S1. Community composition, stochastic assembly model and expected stochasticity in the example datasets from (a) the spatially implicit model and (b) the spatially explicit model. The OTU tables and associated annotation (left part of each panel) show the abundances, richness, and abundance sum of deterministic (blue) and stochastic (orange) species in each sample, while $S_{d.tot}$ and $S_{t.tot}$ represent the total richness of deterministic and stochastic species across all samples in a dataset. Each column of the OTU tables represents a sample, and each OTU represents a species. The bottom box (brown) in each panel shows how the expected stochasticity is calculated. The box of stochastic assembly model on the right panel shows how the stochastic species are simulated. In both spatially implicit (a) and explicit (b) models, the top-level metacommunity is simulated according to Hubbell's neutral theory model, and each local community is generated as random draw from local metacommunity based on Sloan's neutral model. In the spatially explicit model with four-level metacommunities, lower-level metacommunities are simulated as random draw of species from higher-level metacommunities using a two-step random assembly method based on Sloan's neutral model. In the spatially implicit model, all simulated local communities (12/plot) are taken as samples. In the spatially explicit model, a certain number (6 in this example) of local communities are taken as samples from each plot (the box about sampling). See supplementary text part C and Table S1 for details.

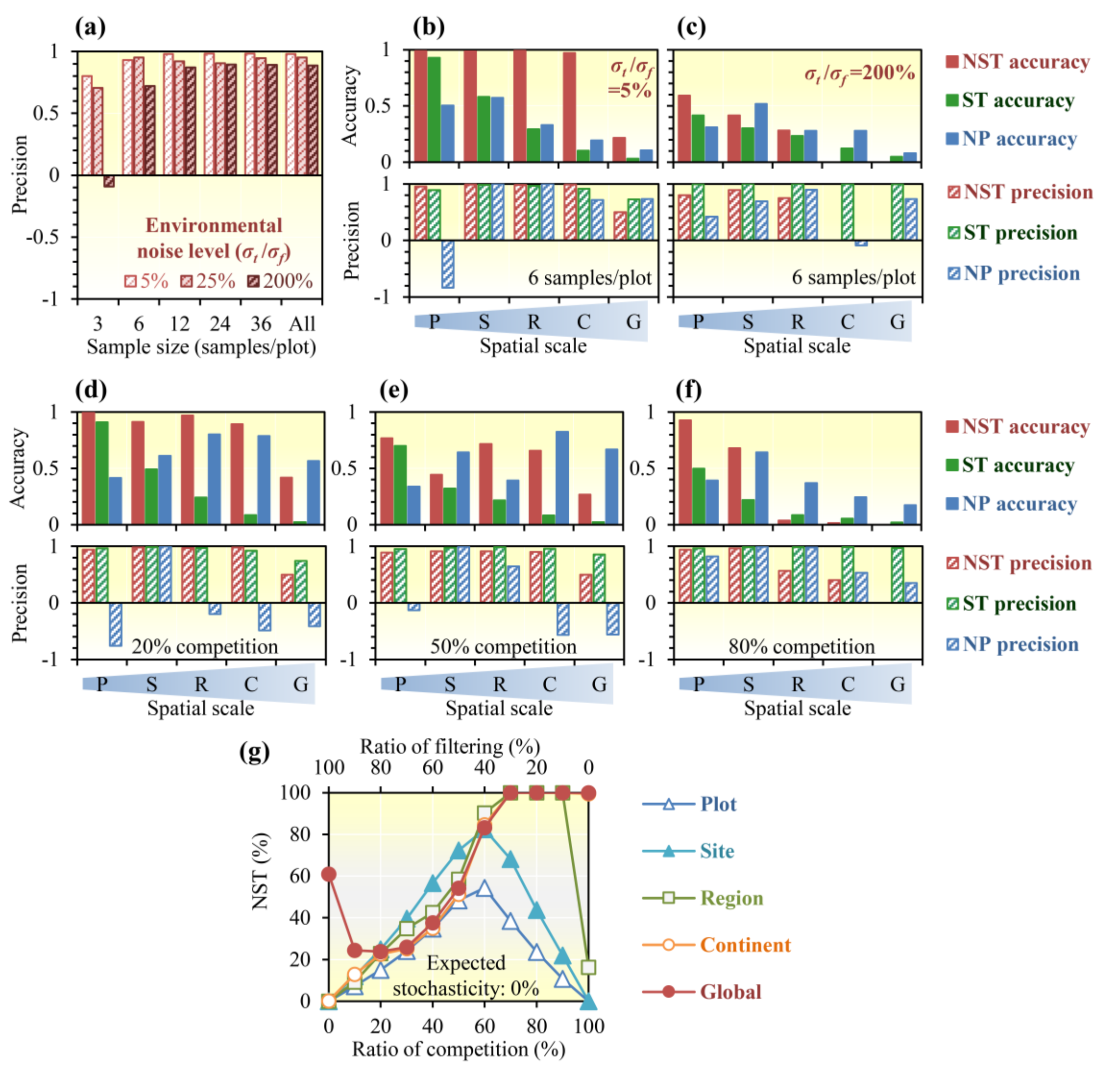

Fig. S2. Accuracy and precision of stochasticity estimation in simulated communities under abiotic filtering with noise from deterministic environmental factors (i.e. environmental noise) or with competition. (a) Influence of sample size on precision of normalized stochasticity ratio (*NST*) at plot scale with different degrees of environmental noise. The precision obviously decreased when sample size was not large enough (≤6 samples/plot), which was more obvious when environmental noise was higher. The accuracy did not show obvious trend, thus not showed here. **(b)** Accuracy and precision of stochasticity estimation under low environmental noise (σ_t/σ_f =5%, Table S1 scenario C) and **(c)** high environmental noise (σ_t/σ_f =200%, Table S1 scenario E) across different spatial scales (P, plot; S, site; R, region; C, continent; G, global). *NST* can have high accuracy and precision when environmental noise was not too high. All indexes had very low accuracy especially at large scales, although stochasticity ratio (*ST*) still showed high precision across different spatial scales. The community was simulated by spatially explicit model (see Supplementary text C and Table S1 for details). σ_t, the standard deviation of temperature in each plot; σ_f, the fitness deviation defined in Eq. S20. **(d, e, f)** Accuracy and precision of stochasticity estimation when deterministic species include some controlled by abiotic filtering and the others controlled by competition with the ratio of **(d)** 20%, **(e)** 50%, and **(f)** 80%. *NST* (red bars), normalized stochasticity ratio; *ST* (green bars), stochasticity ratio; *NP* (blue bars), neutral species percentage. *NST* and *ST* were based on

Ružička similarity index (Table S3) and the null model “PF” (Table S4). Accuracy (solid color bars) and precision (diagonal strip bars) were evaluated by the coefficients derived from concordance correlation coefficient (Eq. S21-22). **(g)** *NST* of simulated communities controlled by abiotic filtering and competition without stochastic assembly, estimated across different scales. Although the expected stochasticity is zero, *NST* still overestimated stochasticity. The overestimation is more obvious when filtering and competition are comparable (e.g. *NST*>50% when ratio of competition is 50~60%). The overestimation is the lowest at plot level. In contrast, *NST* became up to 100% at regional to global scales when the ratio of competition is 70~90%.

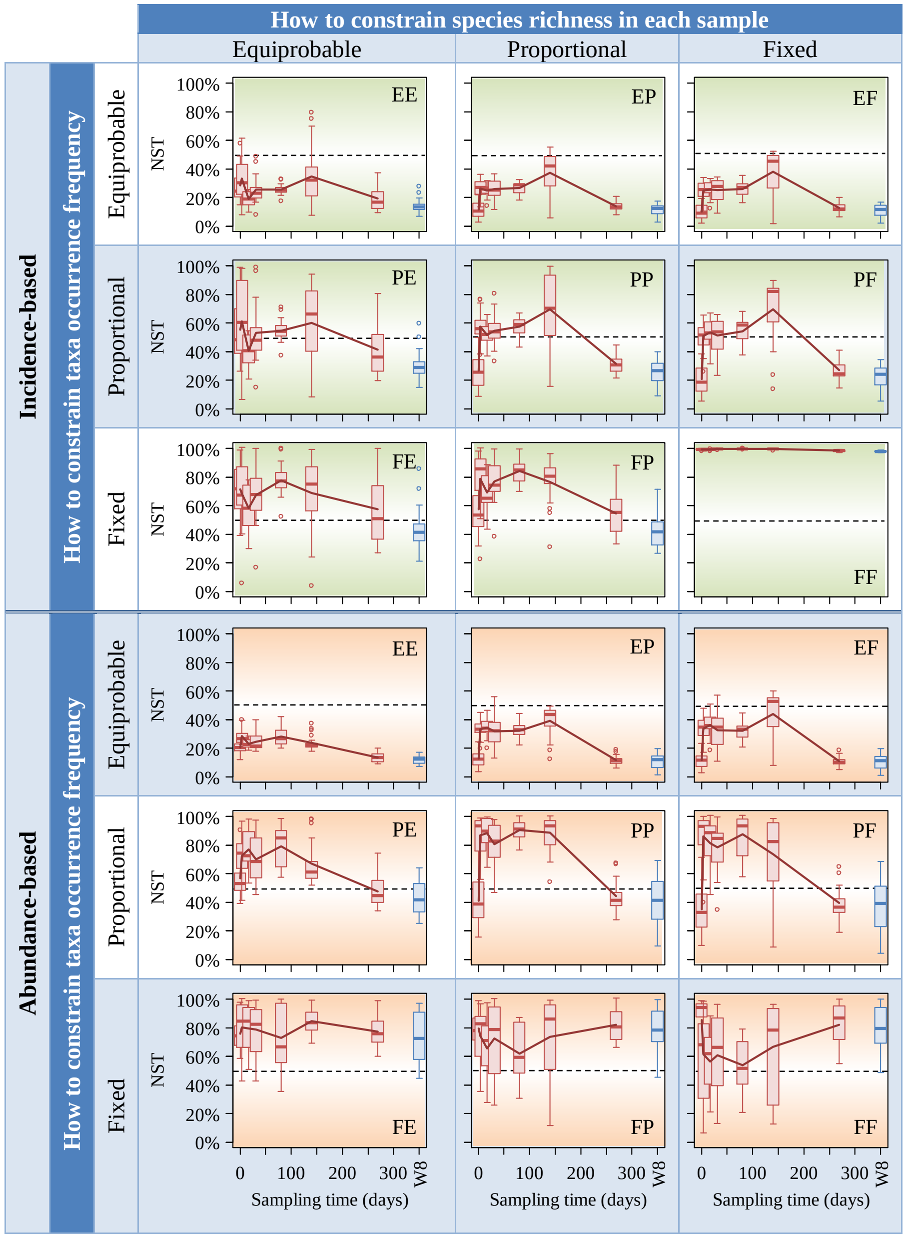

Fig. S3. Effects of null model algorithms on *NST* estimation with incidence-based Jaccard (upper) and abundance-based Ružička (lower) similarity metric. The emulsified vegetable oil was injected at Day 1 and almost exhausted at Day 269, and had minimal impact on the control well (W8). Therefore, at Day 0, Day 269, and W8, the microbial communities were under very high selection pressure caused by high concentrations of pollutants (e.g. heavy metals, nitrate) and carbon poor (34, 35), thus they should be under more deterministic assembly with low stochasticity. The vegetable oil injection significantly increased carbon resources (electron donors) and decreased some pollutants (34), thus should reduce the impact of selection and increase stochasticity. The null model PP and PF showed more significant and expected variations of stochasticity along time. Null model EP and EF showed similar trend but much less estimated stochasticity than expected. Other null models did not show consistent or clear trend. The *NST* values based on Ružička were obviously higher than those based on Jaccard although the trend is very similar. See Table S4 for the detailed algorithms of null models.

Fig. S4. Effects of similarity metrics on *NST* estimation. Most metrics showed very similar trend of stochasticity variation, but the magnitude of *NST* obviously varied among some metrics. Most abundance-based metrics showed similar trend of stochasticity variation but obviously higher magnitude of *NST* comparing to their corresponding incidence-based metrics. Null model algorithm used was “PF” (Table S4). See Table S3 for detailed definition of each similarity metric, and supplementary text part A for metrics standardization method.

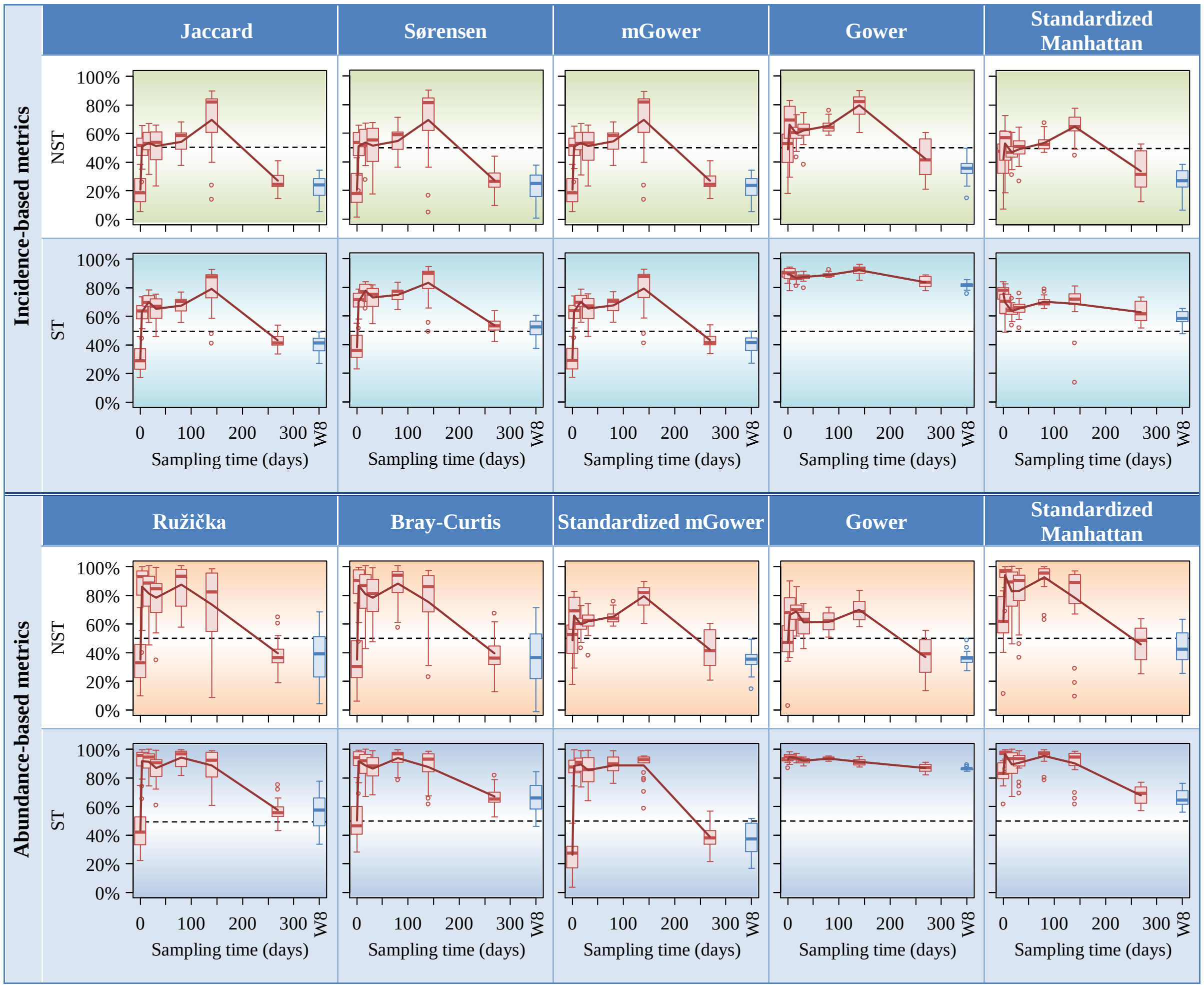

Fig. S5. Comparison between *NST* and *ST* estimated with different similarity metrics. Although *NST* and *ST* basically showed consistent trend, *NST* exhibited much less variation when based on different similarity metrics, i.e. *NST* is less sensitive to metric selection than *ST*.

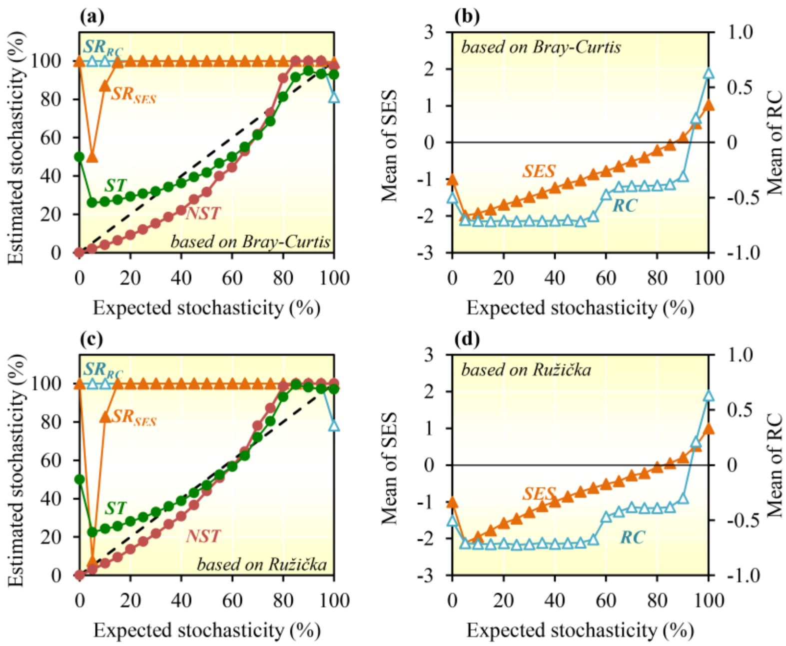

Fig. S6. Comparison of different null-model-based indexes applied to the simulated communities with various levels of expected stochasticity. (a) Estimated stochasticity with different indexes based on Bray-Curtis; **(b)** Mean of standardized effect size (SES) and modified Raup-Crick metrics (RC) based on Bray-Curtis; **(c)** Estimated stochasticity with different indexes based on Ružička; **(d)** Mean of SES and RC based on Ružička. The simulation model was spatially implicit. *NST* (red), normalized stochasticity ratio; *ST* (green), stochasticity ratio; SES (orange, panel b and d), standardized effect size; SR_{SES} (orange, panel a and c), stochastic turnover ratio based on SES, i.e. percentage of turnovers with |SES|<2; RC (aqua, panel b and d), modified Raup-Crick metrics; SR_{RC} (aqua, panel a and c), stochastic turnover ratio based on RC, i.e. percentage of turnovers with |RC|<0.95.

Supplementary Tables

Table S1. The richness (*S*) and abundance (*J*) of species under stochastic and deterministic assembly in simulated communities. Seven scenarios (A-G) are considered. Scenario A has 21 datasets, while the others have 11 datasets.

Dataset #	Expected stochasticity		In each local community						In all samples		
	Incidence -based	Abundance -based	Stochastic species		Deterministic species		All species		Stochastic species	Deterministic species	All species
	$ST_{exp.in}$	$ST_{exp.ab}$	S_t[1]	J_t	S_d	J_d	S	J	$S_{t.tot}$	$S_{d.tot}$	S_{tot}
Scenario A: Spatially implicit model, abiotic filtering without environmental noise **(12 samples/plot × 2 plots = 24 samples)**											
A1	0%	0%	0	0	1	20000	1	20000	0	2	2
A2	99%	5%	100	1000	1	19000	101	20000	1353	2	1355
A3	99%	10%	100	2000	1	18000	101	20000	1392	2	1394
A4	99%	15%	100	3000	1	17000	101	20000	1433	2	1435
A5	99%	20%	100	4000	1	16000	101	20000	1462	2	1464
A6	99%	25%	100	5000	1	15000	101	20000	1484	2	1486
A7	99%	30%	100	6000	1	14000	101	20000	1498	2	1500
A8	99%	35%	100	7000	1	13000	101	20000	1528	2	1530
A9	99%	40%	100	8000	1	12000	101	20000	1532	2	1534
A10	99%	45%	100	9000	1	11000	101	20000	1505	2	1507
A11	99%	50%	100	10000	1	10000	101	20000	1525	2	1527
A12	99%	55%	100	11000	1	9000	101	20000	1577	2	1579
A13	99%	60%	100	12000	1	8000	101	20000	1560	2	1562
A14	99%	65%	100	13000	1	7000	101	20000	1573	2	1575
A15	99%	70%	100	14000	1	6000	101	20000	1566	2	1568
A16	99%	75%	100	15000	1	5000	101	20000	1572	2	1574
A17	99%	80%	100	16000	1	4000	101	20000	1591	2	1593
A18	99%	85%	100	17000	1	3000	101	20000	1593	2	1595
A19	99%	90%	100	18000	1	2000	101	20000	1591	2	1593
A20	99%	95%	100	19000	1	1000	101	20000	1599	2	1601
A21	100%	100%	100	20000	0	0	100	20000	1559	0	1559
Scenario B: Spatially explicit model, abiotic filtering without environmental noise **(6 samples/plot × 256 plots = 1536 samples)**											
B1	0%	0%	0	0	1	20000	1	20000	0	16	16
B2	99%	10%	100	2000	1	18000	101	20000	3720	16	3736
B3	99%	20%	100	4000	1	16000	101	20000	3965	16	3981
B4	99%	30%	100	6000	1	14000	101	20000	4199	16	4215
B5	99%	40%	100	8000	1	12000	101	20000	4299	16	4315
B6	99%	50%	100	10000	1	10000	101	20000	4419	16	4435
B7	99%	60%	100	12000	1	8000	101	20000	4516	16	4532
B8	99%	70%	100	14000	1	6000	101	20000	4670	16	4686
B9	99%	80%	100	16000	1	4000	101	20000	4688	16	4704
B10	99%	90%	100	18000	1	2000	101	20000	4832	16	4848
B11	100%	100%	100	20000	0	0	100	20000	4807	0	4807

Table S1. Continued

Dataset #	Expected stochasticity		In each local community						In all samples		
	$ST_{exp.in}$	$ST_{exp.ab}$	S_t	J_t	S_d	J_d	S	J	$S_{t.tot}$	$S_{d.tot}$	S_{tot}
Scenario C: Spatially explicit model, abiotic filtering with low environmental noise (**σ_i/σ_f =5%[2], 6 samples/plot × 256 plots = 1536 samples)**											
C1	0%	0%	0	0	1	19628~ 19999	1	19628~ 19999	0	16	16
C2	99%	10.0%~ 10.2%	100	2000	1	17682~ 17999	101	19682~ 19999	3701	16	3717
C3	99%	20.0%~ 20.2%	100	4000	1	15773~ 15999	101	19773~ 19999	3980	16	3996
C4	99%	30.0%~ 30.4%	100	6000	1	13752~ 13999	101	19752~ 19999	4161	16	4177
C5	99%	40.0%~ 40.4%	100	8000	1	11814~ 11999	101	19814~ 19999	4349	16	4365
C6	99%	50.0%~ 50.4%	100	10000	1	9847~ 9999	101	19847~ 19999	4454	16	4470
C7	99%	60.0%~ 60.5%	100	12000	1	7851~ 7999	101	19851~ 19999	4549	16	4565
C8	99%	70.0%~ 70.4%	100	14000	1	5888~ 5999	101	19888~ 19999	4592	16	4608
C9	99%	80.0%~ 80.2%	100	16000	1	3945~ 3999	101	19945~ 19999	4699	16	4715
C10	99%	90.0%~ 90.1%	100	18000	1	1969~ 1999	101	19969~ 19999	4804	16	4820
C11	100%	100%	100	20000	0	0	100	20000	4798	0	4798
Scenario D: Spatially explicit model, abiotic filtering with medium environmental noise (**σ_i/σ_f =25%, 6 samples/plot × 256 plots = 1536 samples)**											
D1	0%	0%	0	0	1~2	15066~ 19999	1~2	15066~ 19999	0	16	16
D2	99%	10.0%~ 13.0%	100	2000	1~2	13358~ 17999	101~ 102	15358~ 19999	3735	16	3751
D3	99%	20.0%~ 25.4%	100	4000	1~2	11721~ 15999	101~ 102	15721~ 19999	3956	16	3972
D4	99%	30.0%~ 40.3%	100	6000	1~2	8870~ 13999	101~ 102	14870~ 19999	4147	16	4163
D5	99%	40.0%~ 49.3%	100	8000	1~2	8228~ 11999	101~ 102	16228~ 19999	4308	16	4324
D6	99%	50.0%~ 58.3%	100	10000	1~2	7145~ 9999	101~ 102	17145~ 19999	4436	16	4452
D7	99%	60.0%~ 71.0%	100	12000	1~2	4890~ 7999	101~ 102	16890~ 19999	4523	16	4539
D8	99%	70.0%~ 76.6%	100	14000	1	4286~ 5999	101	18286~ 19999	4598	16	4614
D9	99%	80.0%~ 86.3%	100	16000	1~2	2534~ 3999	101~ 102	18534~ 19999	4725	16	4741
D10	99%	90.0%~ 93.3%	100	18000	1	1289~ 1999	101	19289~ 19999	4745	16	4761
D11	100%	100%	100	20000	0	0	100	20000	4849	0	4849

Table S1. Continued

Dataset #	Expected stochasticity		In each local community						In all samples		
	$ST_{exp.in}$	$ST_{exp.ab}$	S_t	J_t	S_d	J_d	S	J	$S_{t.tot}$	$S_{d.tot}$	S_{tot}
Scenario E: Spatially explicit model, abiotic filtering with high environmental noise **(σ_t/σ_f=200%, 6 samples/plot × 256 plots = 1536 samples)**											
E1	0%	0%	0	0	1~2	3~19999	1~2	3~19999	0	16	16
E2	99%	10.0%~100.0%	100	2000	0~2	0~17999	100~102	2000~19999	3747	16	3763
E3	99%	20.0%~100.0%	100	4000	0~2	0~15999	100~102	4000~19999	3958	16	3974
E4	99%	30.0%~100.0%	100	6000	1~2	2~13999	101~102	6002~19999	4153	16	4169
E5	99%	40.0%~100.0%	100	8000	0~2	0~11999	100~102	8000~19999	4300	16	4316
E6	99%	50.0%~100.0%	100	10000	0~2	0~9999	100~102	10000~19999	4455	16	4471
E7	99%	60.0%~100.0%	100	12000	0~2	0~7999	100~102	12000~19999	4522	16	4538
E8	99%	70.0%~100.0%	100	14000	1~2	1~5999	101~102	14001~19999	4643	16	4659
E9	99%	80.0%~100.0%	100	16000	0~2	0~3999	100~102	16000~19999	4720	16	4736
E10	99%	90.0%~100.0%	100	18000	0~2	0~1999	100~102	18000~19999	4779	16	4795
E11	100%	100%	100	20000	0	0	100	20000	4887	0	4887
Scenario F: Spatially explicit model, biotic interspecies competition **(6 samples/plot × 256 plots = 1536 samples)**											
F1	0%	0%	0	0	1	20000	1	20000	0	249	249
F2	99%	10%	100	2000	1	18000	101	20000	3721	251	3972
F3	99%	20%	100	4000	1	16000	101	20000	3964	247	4211
F4	99%	30%	100	6000	1	14000	101	20000	4203	248	4451
F5	99%	40%	100	8000	1	12000	101	20000	4351	249	4600
F6	99%	50%	100	10000	1	10000	101	20000	4438	244	4682
F7	99%	60%	100	12000	1	8000	101	20000	4533	253	4786
F8	99%	70%	100	14000	1	6000	101	20000	4629	248	4877
F9	99%	80%	100	16000	1	4000	101	20000	4714	250	4964
F10	99%	90%	100	18000	1	2000	101	20000	4787	251	5038
F11	100%	100%	100	20000	0	0	100	20000	4780	0	4780
Scenario G: Spatially explicit model, abiotic filtering and competition **(6 samples/plot × 256 plots = 1536 samples)**											
G1	0%	0%	0	0	2	20000	2	20000	0	272	272
G2	99%	10%	100	2000	2	18000	102	20000	4702	272	4974
G3	99%	20%	100	4000	2	16000	102	20000	4923	272	5195
G4	99%	30%	100	6000	2	14000	102	20000	5107	272	5379
G5	99%	40%	100	8000	2	12000	102	20000	5234	272	5506
G6	99%	50%	100	10000	2	10000	102	20000	5393	272	5665
G7	99%	60%	100	12000	2	8000	102	20000	5447	272	5719

Dataset #	Expected stochasticity		In each local community						In all samples		
	$ST_{exp.in}$	$ST_{exp.ab}$	S_t	J_t	S_d	J_d	S	J	$S_{t.tot}$	$S_{d.tot}$	S_{tot}
G8	99%	70%	100	14000	2	6000	102	20000	5557	272	5829
G9	99%	80%	100	16000	2	4000	102	20000	5588	272	5860
G10	99%	90%	100	18000	2	2000	102	20000	5715	272	5987
G11	100%	100%	100	20000	0	0	100	20000	5723	0	5723

[1] S_t, S_d, and S are the richness of stochastic, deterministic, and all species in each local community; J_t, J_d, and J are the abundance of stochastic, deterministic, and all species in each local community; $S_{t.tot}$, $S_{d.tot}$, and S_{tot} are the overall richness of stochastic, deterministic, and all species in all samples.

[2] σ_t is the standard deviation of temperature in each plot, σ_f is fitness deviation defined in Eq. S20.

Table S2. Accuracy and precision of stochasticity in simulated communities estimated by different indexes based on various similarity metrics.

Types	Similarity metrics	Accuracy coefficient[1]			Precision coefficient[1]		
		NST[2]	*ST*[2]	*NP*[2]	*NST*	*ST*	*NP*
Incidence-based	Jaccard[3]	0.999	0.779		1.000	0.999	
	Sørensen[3]	0.999	0.764		1.000	0.999	
	Kulczynski	0.999	0.762		1.000	0.999	
	Gower[3]	0.994	0.773	0.744[6]	1.000	0.999	0.118
	Manhattan	0.999	0.607		1.000	0.999	
	Euclidean (S[4])	0.999	0.625		1.000	0.999	
	mEuclidean[5] (S)	0.999	0.639		1.000	0.999	
Abundance-based	Ružička	0.985	0.968		0.985	0.925	
	Bray-Curtis	0.966	0.969		0.969	0.897	
	Kulczynski	0.965	0.969		0.968	0.896	
	Canberra	0.416	0.255		0.532	0.811	
	Gower	0.616	0.308		0.660	0.865	
	mGower[5] (S)	0.989	0.716		0.989	0.920	
	Morisita	0.629	0.809		0.724	0.606	
	Morisita-Horn	0.631	0.810	0.462	0.723	0.609	0.275
	Manhattan (S)	0.986	0.754		0.987	0.917	
	mManhattan[5] (S)	0.989	0.714		0.988	0.920	
	Euclidean (S)	0.652	0.323		0.928	0.822	
	mEuclidean[5] (S)	0.637	0.275		0.936	0.829	
	Binomial (S)	0.380	0.148		0.370	0.411	
	Chao	0.967	0.919		0.981	0.940	
	Cao (S)	0.753	0.164		-0.484	0.304	

[1] Communities are simulated by the spatially implicit model described in supplementary text C. Accuracy and precision coefficients are derived from concordance correlation coefficient according to Lin et al. (31, 32).

[2] Stochasticity indexes: *NST*, normalized stochasticity ratio; *ST*, stochasticity ratio; *NP*, abundance-based or incidence-based percentage of species fitting neutral model.

[3] The incidence-based similarity metrics Canberra, modified Gower (mGower), Cao, and modified Manhattan showed exactly the same results as Jaccard metric. The incidence-based Morista-Horn metric showed the same results as Sørensen metric. The incidence-based Binomial metrics showed the same results as Gower metric. See Table S3 for the detailed definition of each similarity metric.

[4] "(S)" means the metrics need to be standardized as described in Supplementary text A before applied to *ST* and *NST*.

[5] mEuclidean, mGower, and mManhattan indicate modified Euclidean, Gower, and Manhattan indexes, respectively.

[6] *NP* does not depend on similarity metrics at all, thus only has one value here.

Table S3. List of similarity and dissimilarity metrics.

No.	Methods	Refs	Dissimilarity (D): Qualitative (D_{uw})		Dissimilarity (D): Quantitative (D_w)		Similarity (C)	
			Formula	Upper limit	Formula	Upper limit	Formula	Upper limit
Taxonomic measures								
1	Jaccard & Ružička	(1, 2)	$\frac{A+B-2J}{A+B-J}$	1	$\frac{\sum_k \lvert x_{ik}-x_{jk}\rvert}{\sum_k max\{x_{ik},x_{jk}\}}$	1	$1-D$	1
2	Sørensen & Bray-Curtis	(3, 4)	$\frac{A+B-2J}{A+B}$	1	$\frac{\sum_k \lvert x_{ik}-x_{jk}\rvert}{\sum_k (x_{ik}+x_{jk})}$	1	$1-D$	1
3	Kulczynski	(5)	$1-\frac{1}{2}\cdot\left(\frac{J}{A}+\frac{J}{B}\right)$	1	$1-\frac{1}{2}\cdot\left(\frac{\sum_k \min\{x_{ik},x_{jk}\}}{\sum_k x_{ik}}+\frac{\sum_k \min\{x_{ik},x_{jk}\}}{\sum_k x_{jk}}\right)$	1	$1-D$	1
4	Canberra	(6)	$\frac{A+B-2J}{A+B-J}$	1	$\frac{1}{A+B-J}\cdot\sum_k \frac{\lvert x_{ik}-x_{jk}\rvert}{x_{ik}+x_{jk}}$	1	$1-D$	1
5	Gower	(7)	$\frac{A+B-2J}{M}$	1	$\frac{1}{M}\cdot\sum_k \frac{\lvert x_{ik}-x_{jk}\rvert}{max\{x_k\}-min\{x_k\}}$	1	$1-D$	1
6	Modified Gower	(8)	$\frac{A+B-2J}{A+B-J}$	1	$\frac{\sum_k \lvert x'_{ik}-x'_{jk}\rvert}{A+B-J}$ where $x'_{ik}=log_{10}(x_{ik})+1$, unless $x_{ik}=0$, in which $x'_{ik}=0$	Unfix.	$C_{uw}=1-D_{uw}$ $C_w=\frac{\sum_i min\{x'_{ij},x'_{ik}\}}{A+B-J}$	1 Unfix.
7	Morisita	(9)	NA.	-	$1-\frac{2\sum_k x_{ik}x_{jk}}{(\lambda_i+\lambda_j)\sum_k x_{ik}\sum_k x_{jk}}$ where $\lambda_i=\frac{\sum_k[x_{ik}(x_{ik}-1)]}{\sum_k x_{ik}[(\sum_k x_{ik})-1]}$	1	$1-D$	1
8	Morisita-Horn	(10)	$\frac{A+B-2J}{A+B}$	1	$1-\frac{2\sum_k x_{ik}x_{jk}}{(\lambda'_i+\lambda'_j)\sum_k x_{ik}\sum_k x_{jk}}$ where $\lambda'_i=\frac{\sum_k x_{ik}^2}{(\sum_k x_{ik})^2}$	1	$1-D$	1

Table S3. Continued

No.	Methods	Refs	Dissimilarity (D)				Similarity (C)	
			Qualitative (D_{uw})		Quantitative (D_w)			
			Formula	Upper limit	Formula	Upper limit	Formula	Upper limit
9	Manhattan	(11)	$A+B-2J$	Unfix.	$\sum_k \lvert x_{ik}-x_{jk}\rvert$	Unfix.	NA.	-
10	Modified Manhattan	(8)	$\frac{A+B-2J}{A+B-J}$	1	$\frac{\sum_k \lvert x_{ik}-x_{jk}\rvert}{A+B-J}$	Unfix.	NA.	-
11	Euclidean	(11)	$\sqrt{A+B-2J}$	Unfix.	$\sqrt{\sum_k (x_{ik}-x_{jk})^2}$	Unfix.	NA.	-
12	Modified Euclidean	(8)	$\frac{\sqrt{A+B-2J}}{A+B-J}$	Unfix.	$\frac{\sqrt{\sum_k (x_{ik}-x_{jk})^2}}{A+B-J}$	Unfix.	NA.	-
13	Binomial	(12)	$(A+B-2J)log2$	Unfix.	$\sum_k \left[\frac{x_{ik}}{x_{ik}+x_{jk}} log\left(\frac{x_{ik}}{x_{ik}+x_{jk}}\right) + \frac{x_{jk}}{x_{ik}+x_{jk}} log\left(\frac{x_{jk}}{x_{ik}+x_{jk}}\right) - log\frac{1}{2}\right]$	Unfix.	NA.	-
14	Chao	(13)	NA.	-	$1-\frac{U_iU_j}{U_i+U_j-U_iU_j}$ where $U_i = \frac{C_i}{N_i} + \frac{N_j-1}{N_j}\cdot\frac{q1}{2q2}\cdot\frac{s1_i}{N_i}$, similar for U_j	1	$1-D$	1
15	Cao	(14)	$\frac{A+B-2J}{A+B-J}\rho$ where ρ=1.4954 if using natural logarithms	ρ	$\frac{\sum_k \left(log\left(\frac{x_{ik}+x_{jk}}{2}\right) - \frac{x_{ik}logx_{jk}+x_{jk}logx_{ik}}{x_{ik}+x_{jk}}\right)}{A+B-J}$ where if x_{ik}=0 or x_{jk}=0, 0.1 is assigned	Unfix.	$1-\frac{D}{\max\{D\}}$	1

Table S3. Continued

No.	Methods	Refs	Dissimilarity (D): Qualitative (D_{uw}) Formula	Upper limit	Dissimilarity (D): Quantitative (D_w) Formula	Upper limit	Similarity (C) Formula	Upper limit
Phylogenetic measures								
16	Phylogenetic analogue of Jaccard & Ružička	(15-17)	$\frac{a+b-2c}{a+b-c}$ called Unifrac	1	$\sum_n \lvert p_{in}-p_{jn}\rvert W_n$ called weighted Unifrac	Unfix.	NA.	-
					$\frac{\sum_n \lvert p_{in}-p_{jn}\rvert W_n}{\sum_n max(p_{in},p_{jn})W_n}$ Unnamed	1	$1-D$	1
17	Phylogenetic analogue of Sørensen & Bray-Curtis	(16, 18)	$\frac{a+b-2c}{a+b}$	1	$\frac{\sum_n \lvert p_{in}-p_{jn}\rvert W_n}{\sum_n (p_{in}+p_{jn})W_n}$ called normalized weighted Unifrac	1	$1-D$	1
18	βMPD	(19)	$\frac{1}{AB}\sum_{k=1}^{A}\sum_{m=1}^{B}\delta_{km}$	Unfix.	$\frac{\sum_{k=1}^{A}\sum_{m=1}^{B} p_{ik}p_{jm}\delta_{km}}{\sum_{k=1}^{A}\sum_{m=1}^{B} p_{ik}p_{jm}}$	Unfix.	NA.	-
19	βMNTD	(19, 20)	$\frac{1}{2}\left[\frac{\sum_{k=1}^{A}\min_m(\delta_{km})}{A}+\frac{\sum_{m=1}^{B}\min_k(\delta_{km})}{B}\right]$	Unfix.	$\frac{1}{2}\left[\sum_{k=1}^{A} p_{ik}\min_m(\delta_{km})+\sum_{m=1}^{B} p_{jm}\min_k(\delta_{km})\right]$	Unfix.	NA.	-

[1] A is the richness (number of taxa) in community i, while B is the richness in sample j, and J is the number of taxa that occur on both sample i and j.

[2] x_{ik} is the abundance of taxon k in sample i, while x_{jk} is the abundance of taxon k in sample j.

[3] p_{ik} is the proportion of taxon k in sample i, while p_{jm} is the proportion of taxon m in sample j.

[4] $\{x_k\}$ is the set of abundances of taxon k in all samples.

[5] M is the number of taxa in all samples.

[6] For Chao index, C_i is the total number of individuals in the taxa of sample i that are shared with sample j; N_i is the total number of individuals in sample i, N_j is the total number of individuals in sample j; $q1$ (and $q2$) are the number of species occurring in sample i that have only one (or two) individuals in sample j; $s1_i$ is the total number of individuals in the species present in sample i that occur with only one individual in sample j.

[7] a is the amount of phylogenetic tree branch length in community i, b is the amount of branch length in community j, and c is the amount of branch length shared between community i and j.

[8] p_{in} is the proportion of sequences (taxa) from community i descendant from branch n; W_n is the weight or length of branch n.

[9] δ_{km} is the phylogenetic distance from sequence (taxon) k to sequence (taxon) m.

[10] X is the set of taxa in community i, while Y is the set of taxa in community k.

Table S4. Summary of null model algorithms for species co-occurrence analysis.

No.	Abbreviation in this paper	Abbreviation in Gotelli (33)	Ways to constrain taxa occurrence frequency[1]	Ways to constrain richness in each sample[2]	Probability of taxon *i* present in sample *j*[3]
1	EE	SIM1	Equiprobable	Equiprobable	$P_{ij} = \frac{1}{N} \cdot \frac{1}{M}$
2	EP	SIM6	Equiprobable	Proportional	$P_{ij} = \frac{1}{N} \cdot \frac{S_j}{F}$
3	EF	SIM3	Equiprobable	Fixed	$P_{ij} = \frac{1}{N}$
4	PE	SIM7	Proportional	Equiprobable	$P_{ij} = \frac{f_i}{F} \cdot \frac{1}{M}$
5	PP	SIM8	Proportional	Proportional	$P_{ij} = \frac{f_i}{F} \cdot \frac{S_j}{F}$
6	PF	SIM5	Proportional	Fixed	$P_{ij} = \frac{f_i}{F}$
7	FE	SIM2	Fixed	Equiprobable	$P_{ij} = \frac{1}{M}$
8	FP	SIM4	Fixed	Proportional	$P_{ij} = \frac{S_j}{F}$
9	FF	SIM9	Fixed	Fixed	Not applicable

[1] As to occurrence frequency, "Equiprobable" means that all taxa have equal probability to occur; "Proportional" means that the occurrence probability of a taxon is proportional to its observed occurrence frequency; "Fixed" means that the occurrence frequency of a taxon is fixed as observed.

[2] As to species richness in each sample, "Equiprobable" means that all samples have equal probability to contain a taxon; "Proportional" means the occurrence probability in a sample is proportional to the observed richness in this sample; "Fixed" means the occurrence frequency of a taxon is fixed as observed.

[3] P_{ij} is the probability of taxon i present in sample j in a null model.
S_j is the observed richness in sample j, N is the total number of taxa, M is the total number of samples.
f_i is the observed occurrence frequency of taxon i, F is the total number of occurrences.
A_i is the regional abundance of taxon i, J is the total abundance of all taxa in all samples.

Supplementary References

1. Jaccard P (1912) The distribution of the flora in the alpine zone. *New Phytol* 11(2):37-50.
2. Ružička M (1958) Anwendung mathematisch-statisticher methoden in der geobotanik (Synthetische bearbeitung von aufnahmen). *Biológia, Bratislava* 13:647–661.
3. Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab. Biologiske Skrifter*. 4(1-34).
4. Bray JR & Curtis JT (1957) An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol Monogr* 27(4):326-349.
5. Kulczynski S (1928) Die Pflanzenassoziationen der Pieninen. *Bulletin de l'Académie polonaise des sciences. Série des sciences biologiques*.
6. Lance GN & Williams WT (1966) Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses"). *Comput J* 9(1):60-64.
7. Gower JC (1971) A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27(4):857-871.
8. Anderson MJ, Ellingsen KE, & McArdle BH (2006) Multivariate dispersion as a measure of beta diversity. *Ecol Lett* 9(6):683-693.
9. Morisita M (1959) Measuring of the dispersion and analysis of distribution patterns. *Memoires of the Faculty of Science, Kyushu University, Series E. Biology*. 2:215-235.
10. Horn HS (1966) Measurement of "Overlap" in Comparative Ecological Studies. *Am Nat* 100(914):419-424.
11. Krebs CJ (1999) *Ecological Methodology* (A. Wesley Longman, NY, USA) 2 Ed.
12. Anderson MJ & Millar RB (2004) Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand. *J Exp Mar Bio Ecol* 305(2):191-221.
13. Chao A, Chazdon RL, Colwell RK, & Shen T-J (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* 8(2):148-159.
14. Cao Y, Williams WP, & Bark AW (1997) Similarity measure bias in river benthic aufwuchs community analysis. *Water Environ Res* 69(1):95-106.
15. Lozupone C & Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12):8228-8235.
16. Lozupone CA, Hamady M, Kelley ST, & Knight R (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 73(5):1576-1585.
17. Parks DH & Beiko RG (2013) Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. *ISME J* 7(1):173-183.
18. Bryant JA, *et al.* (2008) Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. *Proc Natl Acad Sci U S A* 105:11505-11511.
19. Webb CO, Ackerly DD, & Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics* 24(18):2098-2100.
20. Stegen JC, Lin X, Konopka AE, & Fredrickson JK (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J* 6:1653–1664.
21. Oksanen J, *et al.* (2017) vegan: Community Ecology Package).
22. Afgan E, *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46(W1):W537-W544.

23. Alonso D & McKane AJ (2004) Sampling Hubbell's neutral theory of biodiversity. *Ecol Lett* 7(10):901-910.
24. Hubbell SP (2001) *The unified neutral theory of biodiversity and biogeography* (Princeton University Press, Princeton, NJ, USA) p 375.
25. Prado PI, Miranda MD, & Chalom A (2017) sads: Maximum Likelihood Models for Species Abundance Distributions. *R package version 0.4.1*.
26. Sloan WT, *et al.* (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ Microbiol* 8(4):732-740.
27. Burns AR, *et al.* (2016) Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J* 10(3):655-664.
28. Chase JM, Kraft NJB, Smith KG, Vellend M, & Inouye BD (2011) Using null models to disentangle variation in community dissimilarity from variation in alpha-diversity. *Ecosphere* 2(2).
29. Stegen JC, *et al.* (2013) Quantifying community assembly processes and identifying features that impose them. *ISME J* 7:2069-2079.
30. Kraft NJB, *et al.* (2011) Disentangling the drivers of beta diversity along latitudinal and elevational gradients. *Science* 333(6050):1755-1758.
31. Lin LI (1989) A concordance correlation-coefficient to evaluate reproducibility. *Biometrics* 45(1):255-268.
32. Lin L, Hedayat AS, Sinha B, & Yang M (2002) Statistical methods in assessing agreement: Models, issues, and tools. *J Am Stat Assoc* 97(457):257-270.
33. Gotelli NJ (2000) Null model analysis of species co-occurrence patterns. *Ecology* 81(9):2606-2621.
34. Zhang P, *et al.* (2017) Dynamic succession of groundwater sulfate-reducing communities during prolonged reduction of uranium in a contaminated aquifer. *Environ Sci Technol* 51(7):3609-3620.
35. He Z, *et al.* (2018) Microbial functional gene diversity predicts groundwater contamination and ecosystem functioning. *MBio* 9(1):e02435-02417.