Research

EDITOR'S CHOICE

Ecography 40: 806–816, 2017
doi: 10.1111/ecog.02302

Subject Editor: Miguel Nakamura. Editor-in-Chief: Miguel Araújo. Accepted 5 July 2016

Linking trait variation to the environment: critical issues with community-weighted mean correlation resolved by the fourth-corner approach

Pedro R. Peres-Neto, Stéphane Dray and Cajo J. F. ter Braak

P. R. Peres-Neto (peres-neto.pedro@uqam.ca), Canada Research Chair in Spatial Modelling and Biodiversity, Dépt des sciences biologiques, Univ. du Québec à Montréal, Succursale Centre-Ville Montréal, Canada. – S. Dray, Laboratoire de Biométrie et Biologie Evolutive, Univ. de Lyon, Univ. Lyon 1, CNRS, UMR 5558, Villeurbanne, France. – C. J. F. ter Braak, Biometris, Wageningen Univ. and Research Centre, Wageningen, the Netherlands.

Establishing trait–environment relationships has become routine in community ecology. Here, we demonstrate that the community weighted means correlation (CWM) and its parallel approach in linking trait variation to the environment, the species niche centroid correlation (SNC), have important shortcomings, arguing against their continuing application. Using mathematical derivations and simulations, we show that the two major issues are inconsistent parameter estimation and unacceptable significance rates when only the environment or only traits are structuring species distributions, but they themselves are not linked. We show how both CWM and SNC are related to the fourth-corner correlation and propose to replace all by the Chessel fourth-corner correlation, which is the fourth-corner correlation divided by its maximum attainable value. We propose an appropriate hypothesis testing procedure that is not only unbiased but also has much greater statistical power in detecting trait–environmental relationships. We derive an additive framework in which trait variation is partitioned among and within communities, which can be then modeled against the environment. We finish by presenting a contrast between methods and an application of our proposed framework across 85 lake-fish metacommunities.

Assembly rules have been long considered as a unifying theme in community ecology as a way to uncover general processes underlying the complex patterns found in ecological assemblages (MacArthur 1972, Keddy 1992). In the search for these processes, ecologists have been studying which traits are important in allowing species to occupy particular environments and which traits allow or hinder species to coexist (McGill et al. 2006). In particular, analyses that link environment and trait to explain community structure have given rise to a plethora of studies (Cavender-Bares et al. 2004, Garnier et al. 2004, Lavorel et al. 2008, Kühn et al. 2009, Terribile et al. 2009, Messier et al. 2010, Webb et al. 2010, Ricotta and Moretti 2011, Coyle et al. 2014).

To date, seemingly different methods have been proposed to link trait variation to environmental features (Dolédec et al. 1996, Legendre et al. 1997, Dray and Legendre 2008, Webb et al. 2010, Pavoine et al. 2011, Ricotta and Moretti 2011, Kleyer et al. 2012, Peres-Neto et al. 2012, Jamil et al. 2013, Brown et al. 2014, Dray et al. 2014). In our experience, the community-weighted trait means (CWM) and fourth-corner approaches are the most commonly used among these methods. In the CWM case, mean trait values, weighted by species abundance, are calculated for each of a set of communities and then correlated with (or regressed on) environmental factors across communities (see Lavorel et al. 2008 for a review). As we will shown in this paper, the fourth corner correlation (Legendre et al. 1997) is a weighted correlation between weighted standardized environment (by row sums, i.e. species richness per community) and the community-weighted trait means; traits are weighted standardized (by columns, i.e. total abundance or number of occurrences per species) prior to calculations. Another related approach, though less common, is the species niche centroid (SNC) method (Cavender-Bares et al. 2004), in which mean environmental values (weighted by species abundance) are calculated for each of a set of species and then correlated with (or regressed on) trait values measured across species.

CWM, SNC and fourth-corner correlations have become quite standard approaches in trait-based ecology. We will show that they are algebraically closely related, but we note already three differences: 1) the way in which traits and environmental features are standardized (non-weighted in CWM and SNC but weighted in the fourth-corner), 2) the way correlations are calculated (non-weighted for CWM and SNC, but weighted in the fourth-corner), and 3) the test procedures used to evaluate the statistical significance of trait–environment relationships (usually parametric for CWM and SNC but nonparametric using permutations for the fourth-corner). These seemingly minor differences result in dramatically different statistical behaviors. One major issue is demonstrated in Fig. 1 in which we simulated an example where the abundance distributions of 50 species

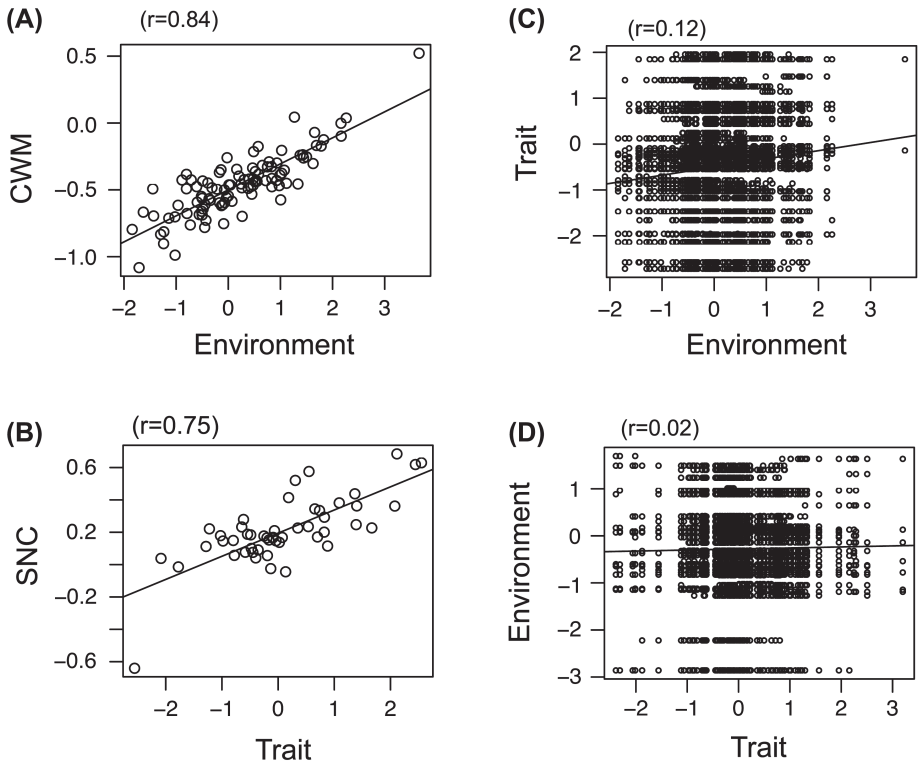

Figure 1. Simulated example (100 communities and 50 species) to show the properties of three main approaches linking trait–environmental variation. (A) Correlation between community weighted trait means and environment for a case in which species distributions were structured by environment but trait values across species were randomly generated; (B) correlation between species weighted environmental mean (SNC) and trait for a case in which species distributions were structured by trait variation but environmental values were randomly generated across communities. (C–D) Fourth-corner correlations in which appropriate weights (hence the amount of points in the graphs) are given to species and communities, leading to appropriate estimates of the absence of correlation between species distributions, environment and traits. Note that the data used in (A–B) are different from (C–D).

across 100 communities were structured according to a single environmental gradient but trait values of a single trait were randomly generated across species (Fig. 1A); and another example in which the distributions of 50 species were structured according to a single trait but environmental values were randomly generated across communities (Fig. 1B). In these examples, one should expect no correlation between trait and environment, but the correlations were quite high and statistically significant by the CWM- and the SNC-based correlations: $r = 0.84$ between CWM and the environment (Fig. 1A) and $r = 0.75$ between SNC and the trait (Fig. 1B). These are typical examples in simulated data. By contrast, the fourth-corner correlations are close to zero, as they should: $r = 0.12$ and $r = 0.02$ in the first and second examples, respectively (Fig. 1C and D). As shown later, the fourth-corner approach is the only one that yields consistent estimates of the true correlation value.

Our goal is to present a critical review of the features, shortcomings and associations among these commonly used approaches. While reviewing them, we present appropriate solutions and produce a robust framework for linking trait variation to the environment. The layout of the paper is as follows. We start by showing the properties and links among these methods. We then show that a) the estimation of trait–environment correlations by using CWM and SNC have very little precision in contrast to the fourth-corner correlation (i.e. large sampling variation around the true correlation value), particularly in the case where correlations are truly zero; b) commonly used statistical test procedures using CWM and SNC (either parametric or permutation-based) do not provide valid tests, i.e. type I error rates are high, meaning that they judge too many correlations to be significant when there is in fact no trait–environment association. We then present solutions, based on the fourth-corner approach, to these shortcomings and introduce a novel and comprehensive multivariate framework in which trait variation is partitioned among and within communities, which can then be modeled against environment. The three frameworks are contrasted to one another using simulations and a large lake-fish data set across multiple metacommunities composed of about 10 000 lakes.

Material and methods

All mathematical equations presented below are illustrated using R code (Supplementary material Appendix 1).

Weighted descriptive statistics – definitions and notation

An important distinction among approaches concerns the weighting options used to standardize variables and compute correlations. To show their links and differences, we start by

recalling some basic algebraic definitions and introducing the mathematical notation of the paper.

Consider a column vector $\mathbf{x} = [x_i]$ containing the values of a variable for n observations (communities or species, depending on the approach). The mean and the variance are given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

$$s^2(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \quad (2)$$

The standardized values for $\mathbf{x}$ are stored in $\tilde{\mathbf{x}} = \left[\tilde{x}_i\right] = \left[\left(x_i - \bar{x}\right)/s(x)\right]$. Note that variance was calculated on the basis of n degrees of freedom instead of the conventional $(n - 1)$; this, however, does not change the calculation of the correlation-based statistics (CWM, SNC and fourth-corner) described here. If we consider a second variable ($\boldsymbol{y} = [y_i]$) with standardized values in $\tilde{\boldsymbol{y}}$, the correlation equals:

$$cor(x, y) = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\,\tilde{y}_i \quad (3)$$

We can consider weights (e.g. proportional to the sum of abundances across species or communities) $(w_1, \ldots, w_n)$, with $\sum_{i=1}^{n} w_i = 1$, stored in an n-by-n diagonal matrix $\mathbf{W} = \text{diag}\ (w_1, \ldots, w_n)$. Weighted mean and variance are calculated as:

$$\bar{x}_{\mathbf{w}} = \sum_{i=1}^{n} w_i x_i \quad (4)$$

$$s^2_{\mathbf{W}}(x) = \sum_{i=1}^{n} w_i\left(x_i - \bar{x}_{\mathbf{W}}\right)^2 \quad (5)$$

Because the weights sum to 1, there is no need to divide the two sums above by their summed weights. For computational convenience, the previous equations can be rewritten as:

$$\bar{x}_{\mathbf{W}} = \mathbf{1}_n^{\mathrm{T}}\mathbf{W}\mathbf{x} \quad (6)$$

$$s^2_{\mathbf{W}}(x) = \left(\boldsymbol{x} - \bar{x}_{\mathbf{W}}\mathbf{1}_n\right)^{\mathrm{T}}\mathbf{W}\left(\mathbf{x} - \bar{x}_{\mathbf{W}}\mathbf{1}_n\right) \quad (7)$$

where $\mathbf{1}_n$ is a column vector of n ones. The weighted correlation $cor_{\mathbf{W}}(x,y)$ can then be computed from $\tilde{\mathbf{x}}_{\mathbf{w}} = \left[\left(x_i - \bar{x}_{\mathbf{W}}\right)/s_{\mathbf{W}}(x)\right]$ and $\tilde{\mathbf{y}}_{\mathbf{W}} = \left[\left(y_i - \bar{y}_{\mathbf{W}}\right)/s_{\mathbf{W}}(y)\right]$ by the equation $cor_{\mathbf{w}}(x, y) = \tilde{\mathbf{x}}_{\mathbf{W}}^{\mathrm{T}}\mathbf{W}\tilde{\mathbf{y}}_{\mathbf{w}}$, and is related to the weighted covariance: $cov(x,y) = cor_{\mathbf{w}}(x,y)s_{\mathbf{w}}(x)s_{\mathbf{w}}(y)$.

Weights are based on the species distribution matrix $\mathbf{L} = [l_{ij}]$, an n-by-S matrix containing the abundances (or presences/absences or biomasses) of S species across n communities. We can define a matrix of relative frequencies $\mathbf{P} = \left[p_{ij}\right] = \left[l_{ij}\Big/\sum_{i=1}^{n}\sum_{j=1}^{S} l_{ij}\right]$ and derive diagonal matrices of weights for species (columns) and communities (rows), respectively: $\mathbf{W}_S = diag\left(\sum_{i=1}^{n} p_{i1}, \cdots, \sum_{i=1}^{n} p_{iS}\right)$ and $\mathbf{W}_n = diag\left(\sum_{j=1}^{S} p_{1j}, \cdots, \sum_{j=1}^{S} p_{nj}\right)$.

While presenting the methods, our notation below substitutes $\mathbf{y}$ and $\mathbf{x}$ by an S-by-1 vector $\mathbf{t} = [t_j]$ containing the values of a trait measured for the S species, and an n-by-1 vector $\mathbf{e} = [e_i]$ with the values of an environmental variable measured for the n communities.

Community-weighted mean correlation

The community weighted means (CWM; Lavorel et al. 2008) of a trait for the ith community is calculated as:

$$c_i = \sum_{j=1}^{S} l_{ij}t_j \Big/ \sum_{j=1}^{S} l_{ij} \quad (8)$$

The n-by-1 vector $\mathbf{c} = [c_i]$ of CWM values for an entire set of communities is calculated by:

$$\mathbf{c} = \mathbf{W}_n^{-1}\mathbf{P}\mathbf{t} \quad (9)$$

The link between $\mathbf{c}$ and an environmental feature is measured by computing the standard correlation: $cor(c,e) = \frac{1}{n}\tilde{\mathbf{c}}^{\mathrm{T}}\tilde{\mathbf{e}}$ where $\tilde{\mathbf{e}}$ is the standardized (mean = 0 and variance = 1) environmental variable. Note, however, that environmental values are not standardized according to site weights (i.e. community total abundances) and correlations are calculated without these site weights. We demonstrate later on that this correlation has such bad statistical properties, that it cannot be reliably used to infer trait–environmental associations.

Species niche centroid correlation

The species niche centroid (SNC; Gause 1930, Tingley et al. 2009; see also ter Braak and Barendregt 1986, ter Braak and Looman 1986) is calculated for the jth species as:

$$u_j = \sum_{i=1}^{n} l_{ij}e_i \Big/ \sum_{i=1}^{n} l_{ij} \quad (10)$$

The SNC values are calculated for all species at once as:

$$\mathbf{u} = \mathbf{W}_S^{-1}\mathbf{P}^{\mathbf{T}}\mathbf{e} \quad (11)$$

The link between $\mathbf{u}$ and a species trait is measured by a standard correlation: $cor(u,t) = \frac{1}{n}\tilde{\mathbf{u}}^{\mathrm{T}}\tilde{\mathbf{t}}$. We will show that $cor(u,t)$ also has the same bad properties as $cor(c,e)$ above.

Equations (8) and (10) form the basis of Whittaker's (1960, 1967) method of gradient analysis to find ordinations of communities on the basis of a quantitative trait and of species on the basis of a quantitative environmental variable, respectively. Hill (1973) used these equations in his reciprocal averaging method, which was a rediscovery of correspondence analysis. In this method, the vectors $\mathbf{t}$ and $\mathbf{e}$ are not measured or given, but optimized in an iterative fashion so as to maximize their weighted correlation. This aspect is used in the next section.

Fourth-corner correlation and links with CWM and SNC

The basic definition of the fourth-corner correlation (Legendre et al. 1997) is a weighted correlation between weighted standardized traits ($\tilde{\mathbf{t}}_{\mathbf{w}_S}$) and weighted standardized environmental ($\tilde{\mathbf{e}}_{\mathbf{w}_n}$) values:

$$cor_{\mathbf{P}}(t,e) = \tilde{\mathbf{e}}_{\mathbf{W}_n}^{\mathrm{T}}\mathbf{P}\tilde{\mathbf{t}}_{\mathbf{W}_S} = \tilde{\mathbf{t}}_{\mathbf{W}_S}^{\mathrm{T}}\mathbf{P}^{\mathbf{T}}\tilde{\mathbf{e}}_{\mathbf{W}_n} \quad (12)$$

The fourth-corner correlation can be rewritten as:

$$cor_{\mathbf{P}}(t,e) = \tilde{\mathbf{e}}_{\mathbf{w}_n}^{\mathrm{T}} \mathbf{W}_n (\mathbf{W}_n^{-1} \mathbf{P} \tilde{\mathbf{t}}_{\mathbf{w}_S}) = \tilde{\mathbf{e}}_{\mathbf{w}_n}^{\mathrm{T}} \mathbf{W}_n \mathbf{c}_{\mathbf{w}_n} \quad (13)$$

which is the $\mathbf{W}_n$-weighted covariance between weighted standardized environmental ($\tilde{\mathbf{e}}_{\mathbf{w}_n}$) values and CWM values $\mathbf{c}_{\mathbf{w}_n} = \mathbf{W}_n^{-1} \mathbf{P} \tilde{\mathbf{t}}_{\mathbf{W}_S}$. This rewrite shows that the fourth-corner correlation is related to the weighted correlation between CWM and environment **e**, where the weights are community total abundances. Moreover, the fourth-corner is also related to the weighted correlation between SNC and trait **t** (with weights the species totals):

$$cor_{\mathbf{P}}(t,e) = \tilde{\mathbf{t}}_{\mathbf{w}_S}^{\mathrm{T}} \mathbf{W}_S (\mathbf{W}_S^{-1} \mathbf{P}^{\mathbf{T}} \tilde{\mathbf{e}}_{\mathbf{w}_n}) = \tilde{\mathbf{t}}_{\mathbf{w}_S}^{\mathrm{T}} \mathbf{W}_S \mathbf{u}_{\mathbf{w}_S} \quad (14)$$

which is the $\mathbf{W}_S$-weighted covariance between weighted standardized traits ($\tilde{\mathbf{t}}_{\mathbf{W}_S}$) and the SNC values $\mathbf{u}_{\mathbf{W}_S} = \mathbf{W}_S^{-1} \mathbf{P}^{\mathbf{T}} \tilde{\mathbf{e}}_{\mathbf{W}_n}$. Because the fourth-corner applies weights to both environment and trait, the direction in which it is calculated leads to precisely the same correlation.

The only difference between these correlations lies in the standardization of the variables involved. As CWM correlates *c* and *e*, its implicit divisor is $s(c).s(e)$ and, in its weighted form, $s_{\mathbf{W}_n}(c).s_{\mathbf{W}_n}(e)$. Similarly, SNC and fourth-corner correlations have implicit divisors (Supplementary material Appendix 2). The relation between these correlations is thus:

$$cor_{\mathbf{P}}(t,e) = \frac{s_{\mathbf{W}_n}(c)}{s_{\mathbf{W}_S}(t)} cor_{\mathbf{W}_n}(c,e) = \frac{s_{\mathbf{W}_S}(u)}{s_{\mathbf{W}_n}(e)} cor_{\mathbf{W}_S}(u,t) \quad (15)$$

As CWM values (**c**) are computed by averaging trait values (**t**), the spread in **c** is smaller than that in **t**, so that $s_{\mathbf{W}_n}(c) < s_{\mathbf{W}_S}(t)$ and thus $cor_{\mathbf{P}}(t,e) < cor_{\mathbf{W}_n}(e,c)$. Symmetrically, we have $s_{\mathbf{W}_S}(u) < s_{\mathbf{W}_n}(e)$ and thus $cor_{\mathbf{P}}(t,e) < cor_{\mathbf{w}_S}(t,u)$. Thus, correlation values obtained by the fourth-corner approach are smaller than those obtained by either weighted CWM or SNC or their non-weighted versions. This is because the CWM and SNC correlations divide by mean variation, that is, the variability in the mean trait among communities in CWM and the variability in the average environment among species in SNC, whereas the fourth-corner divides by the total trait variation (i.e. $s_{\mathbf{w}_n}(e)$ and $s_{\mathbf{w}_S}(t)$). Division by mean variation gives arbitrary results, when its true value is zero, which happens in CWM when the trait is unrelated to the species distribution (matrix **L**) and, in SNC, when the environmental variable is unrelated to **L**, as we demonstrate by simulation later on.

In practice, the fourth-corner correlation cannot reach –1 or 1 unless species distributions are perfectly ordered across columns and rows in **L**; its maximum value is equal to the square-root of the first eigenvalue of the correspondence analysis of **L**, as one of the many derivations of correspondence analysis is in fact finding the latent community and species scores that maximize their correlation (Williams 1952). We therefore propose to divide the fourth-corner correlation by this maximum, and to call this ratio the Chessel fourth-corner correlation, in honor of Daniel Chessel, who introduced this type of correlation in RLQ analysis (Dolédec et al. 1996).

In the next section, we show how the total variation can be decomposed into among- (mean) and within- (variance) components, providing a much broader framework for the analysis of trait variation than just considering the mean component.

Decomposing trait variation into among and within variation components

Because the CWM and SNC approaches are based on weighted averages, they ignore trait variation among species within communities, though this question has gained much interest (Weiher et al. 1998, Swenson and Enquist 2007, Paquette and Messier 2011, Ricotta and Moretti 2011, Coyle et al. 2014). It is now routine to use community-weighted mean trait values to assess the among-community component and average trait distances across species to assess the within-community component (e.g. trait diversity metrics such as mean pairwise trait distances across species – MPD; Ricotta and Moretti 2011, Coyle et al. 2014). However, these approaches are not additive, i.e. they do not produce values that would add up to the total trait variation. An additive partitioning would allow for an appropriate contrast between the relative importance of the among- and within-components of trait variation across different metacommunities (akin in concept to the alpha and beta components in the decomposition scheme proposed by Ackerly and Cornwell 2007). Using an ANOVA-like decomposition of variance, we can divide the total variation in a trait (or series of traits) into its within- and among-components as follows:

$$\underbrace{s^2_{\mathbf{W}_S}(t)}_{\text{Total variation}} = \underbrace{\sum_{i=1}^{n} \alpha_i s^2_{\mathbf{W}_{S_i}}(t)}_{\text{Within-communities variation}} + \underbrace{s^2_{\mathbf{W}_n}(c)}_{\text{Among-communities variation}} \quad (16)$$

where $s^2_{\mathbf{W}_S}(t)$ represents the total trait variation. The variation of traits observed in the *i*th community is given by $s^2_{\mathbf{W}_{S_i}}(t) = \sum_{j=1}^{S} \frac{l_{ij}}{\sum_{j=1}^{S} l_{ij}} (t_j - c_i)^2$. The coefficients α_i are equal to $\sum_{j=1}^{S} p_{ij}$ and are the abundance weights for the communities. As described above, these weights are stored in $\mathbf{W}_n = diag(\alpha_i, ..., \alpha_i)$ and it follows that the within-communities variation is the weighted average of traits variances computed for each community.

From single fourth-corner correlations to a multivariate regression approach

Another way to describe the fourth-corner correlation is as a $\mathbf{W}_n$-weighted regression slope between $\mathbf{W}_n$-weighted standardized environment and the community means of a weighted standardized trait:

$$cor_{\mathbf{P}}(t,e) = b_{W_n(\mathbf{c}_{\mathbf{W}_n} \sim \tilde{\mathbf{e}}_{\mathbf{W}_n})} = \left(\tilde{\mathbf{e}}_{\mathbf{W}_n}^{\mathrm{T}} \mathbf{W}_n \tilde{\mathbf{e}}_{\mathbf{W}_n}\right)^{-1} \tilde{\mathbf{e}}_{\mathbf{W}_n}^{\mathrm{T}} \mathbf{W}_n \mathbf{c}_{\mathbf{W}_n} \quad (17)$$

Note that the integration of multiple environmental predictors is straightforward by replacing the vector containing a single weighted standardized environmental predictor $\tilde{\mathbf{e}}_{\mathbf{w}_n}$ by a matrix $\tilde{\mathbf{E}}_{\mathbf{w}_n}$ in which each column is weighted standardized. Note that this approach differs from that in RLQ

in considering the correlations among environmental variables; such correlations are neglected in RLQ. In the present paper, however, we will focus only on single response trait, given that the approaches reviewed here are widely used in this context (Lavorel et al. 2008, Kleyer et al. 2012).

Statistical test procedures

Here we will show that parametric tests of CWM- and SNC-based correlations all lead to inflated type I errors. P-values estimated by a parametric test (i.e. which assumes normality of residuals) for the CWM and SNC approaches are closely related to the randomization test based on the permutation of the values in the environment **e** and trait **t** vectors, respectively. Because of these equivalences, only the results for the permutation tests are presented.

To date, only type I error and power of the fourth-corner correlation have been evaluated (Dray and Legendre 2008, ter Braak et al. 2012) but no assessment for CWM-based and SNC-based correlations have been produced. This is likely in part because no difficulties were expected and no links had been established among these seemingly different methods. In the case of the fourth-corner, two permutation test strategies must be applied (ter Braak et al. 2012): permutation of rows of either **L** or **e** (communities) and permutation of columns of either **L** or **t** (species). For each permutation procedure separately (i.e. row and column based), a p-value is estimated as the number (plus one) of random squared correlations equal to or greater than the observed squared value divided by the number of permutations plus one. By adding 1, the observed values are included as one possible value under the null distribution. In earlier work (Dray and Legendre 2008, Peres-Neto et al. 2012), we found that either permutation procedure when applied alone produces elevated type I errors for the fourth-corner correlation when only the environment or only traits structured species distributions, but they themselves were not linked. As shown later (ter Braak et al. 2012), we can only conclude that species sharing greater levels of trait similarity have also more similar habitat affinities (i.e. a link between trait, species distributions and environment), when both the row-based and the column-based permutation tests are significant (here referred as to row–column permutation scheme). By picking the largest of the two p-values, the type I error is controlled, i.e. the type I error is equal to or below the nominal level (ter Braak et al. 2012). In the next section, we use simulations to study the sampling properties of the different frameworks (i.e. sample variation around the true population value) as well as the properties of different statistical testing procedures. Code for the different procedures are shown in the Supplementary material Appendix 3.

Assessing the performance of the different frameworks via simulations

To assess the statistical properties of the different frameworks (type I error rates, statistical power, and sampling accuracy), we used a simulation protocol (after Dray and Legendre 2008) that conditions a species distribution matrix on the basis of the joint distribution of traits and environment. Power estimates were based on community matrices of 50 sites by 30 species and 100 sites by 50 species, whereas type I error and sampling variability in parameter estimation were based on a community matrix of 100 sites by 50 species (results for type I error were consistent regardless of matrix sizes). The steps involved were the following (we considered 100 sites and 50 species in this description): 1) generate an environmental vector **e** containing 100 independent random values drawn from the standard normal distribution: $\mathbf{e} \sim N(0,1)$; 2) generate a species trait vector **t** containing 50 random values: $\mathbf{t} \sim N(0,1)$; 3) generate a vector **h** (50 species) containing 50 independent uniformly distributed random values between 0.3 and 1. These values represent the abundance of any given species at its optimum; 4) generate a vector $\boldsymbol{\sigma}$ containing 50 independent uniformly distributed random values between 0 and 1 that were multiplied by a constant s (referred as to niche breadth) which took different values in order to generate different correlation strengths between trait and environment; and 5 generate a unimodal response for the *j*th species at the *i*th site as follows:

$$\mu_{ij} = 30\mathbf{h}_j exp\left[\frac{-\left(e_i - t_j\right)^2}{2s\sigma_j^2}\right] \quad (18)$$

6) The abundance values $\mathbf{L}_{ij}$ were drawn from a Poisson distribution with mean μ_{ij}.

We started by estimating the sampling variation under three scenarios in which we removed the links between environment and trait (i.e. trait–environmental correlation equals zero). Here, species distributions were first structured by both environment and trait (based on $s = 1.5$), but the environmental and trait vectors were then replaced by an independent set of normally distributed random values according to the following scenarios: 1) both trait and environmental variation were replaced by random values in relation to the species distributions, 2) environment was related to species distributions (i.e. original vector **e** was used in the analysis) but traits not (i.e. original **t** was replaced by a random vector in the analysis) and 3) trait was related to species distributions (i.e. original **t** was used in the analysis) but environment **e** was replaced by a random vector. In all simulations, 1000 samples were generated (i.e. 1000 different **L**, **t** and **e**), permutation tests were based on 9999 random permutations and a significance level (alpha) of 0.05.

Note that expanding the square of the difference between e_i and t_j in the simulation model (Eq. 18) yields two squares, e_i^2 and t_j^2, and a product $e_i t_j$. In terms of a log-linear model, this product represents the interaction between trait and environment (e.g. species tend to select environments with lower temperatures the larger their body size, a relation known as Bergman's rule). In the Supplementary material Appendix 4, we simulate simple log-linear models to verify that the row–column testing procedure applied to the CWM and SNC and fourth-corner based correlations controls the type I error when there are main effects but no interaction. In addition we replace the Poisson distribution by a zero-inflated negative binomial distribution for generating the type of overdispersion that is common in ecological data (Supplementary material Appendix 4).

Real data application: Ontario lake-fish metacommmunities

The data contain presence–absence records distributed across 9900 lakes (see Henriques-Silva et al. 2013 for details) for 134 fish species and was obtained from the Ontario (Canada) Fish Distribution Database (OFDD). Lakes within the same tertiary watershed (85 in total) were considered to form a metacommunity. Environmental data for each lake consisted of lake elevation, total dissolved solids, pH, Secchi depth, oxygen concentration, morphoedaphic index, surface area, maximum depth, mean depth, perimeter, island perimeter within lakes and growing-degree days. The only trait used in this analysis was fish (adult) body length obtained in Scott and Crossman (1973). The choice of data was due to the fact that we were able to contrast variation in the strength of these relationships across a large number of metacommunities (85 watersheds) at the same time. In contrast, previously used data to study trait–environment relationships (mostly plants) do not contain information a large number of species data matrices.

Our goal here was not to provide an example of how to make ecological inferences based on trait–environmental links, as they are abundant in the literature, but rather to show the properties of the different approaches on real data sets.

Results

Simulations

Figure 2A–F shows the distribution of correlation values for the three scenarios for the standard CWM correlation ($cor(c,e)$) and the CWM weighted correlations ($cor_{w_n}(c,e)$; Fig. 2A and B), the standard SNC correlation ($cor(t,u)$) and the SNC weighted correlations ($cor_{w_s}(t,u)$; Fig. 2C and D), the fourth-corner correlation ($cor_{\mathbf{P}}(t,e)$ Fig. 2E) and the Chessel fourth-corner correlation (Fig. 2F; see text) under which the expected correlation between trait and environment is zero. Except for the fourth-corner (and Chessel's standardized form), all statistics had very low precision around the true value of zero.

Next, we explored variation around correlations that are non-zero while modulating the strength of the correlations using values for s (niche breadth) at 1, 5 and 10 (Fig. 2G to L). As expected, for any given s (niche breadth), the standard and weighted correlations (Fig. 2G to J) produced greater values on average than the fourth-corner correlation (Fig. 2K). Note also that the values differed for the CWM and SNC approaches and that the Chessel fourth-corner correlation (Fig. 2L) is higher than four-corner, but still lower than the other correlations.

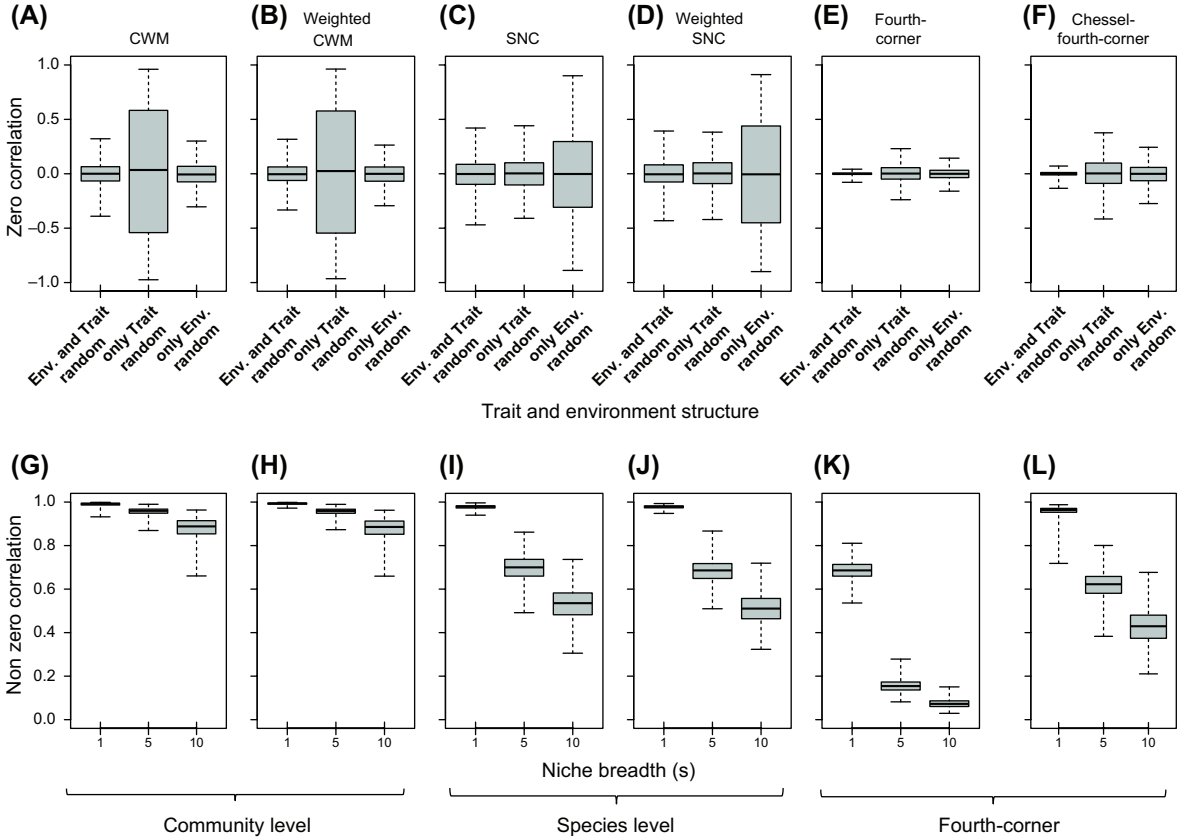

Figure 2. Sampling variation (median, lower and upper quartiles and extreme values for 1000 sample correlations) for the five different methods for correlating trait to environmental variation. Note that the median and the mean were equal. Panels (A) to (F) represent the case in which the true correlations were zero for three different scenarios relating trait to environment when simulating species distributions: both trait and environment were random, only trait was random but not environment, and only environment was random but not trait in determining species distributions. Panels (G) to (L) represent the case in which the correlations between trait and environment in determining species distributions were set to non-zero (true link). In each case, three different levels of niche breadth (s) were considered.

Regarding type I error, when both traits and environment were random (i.e. unrelated to **L**), the observed type I error rate is correct (i.e. giving values close to or less than the nominal level of 0.05) in all permutation schemes (Fig. 3). However, when either only trait or only environment are important in structuring species distributions, only the row–column scheme presented correct type I errors, whereas the row permutation or column permutation presented extremely high values (Fig. 3). Regarding power, given that our simulation approach generates very strong correlations between trait and environment, we considered larger niche breadths (5, 10 and 20) and once the matrix **L** and vectors **e** and **t** were generated, we added noise to the environment by adding to **e** a vector of normally N(0,1) randomly distributed values multiplied by a constant (either 1 or 2). Note also that only power for the row–column permutation scheme is shown because it is the only one that provides correct type I error rates (Fig. 3). The general conclusion is that the fourth-corner correlation provided the greatest power in contrast to the standard and weighed correlations based on CWN and SNC (Fig. 4). Note that power is not reported for Chessel's correlations as they have exactly the same p-value as fourth-corner correlations.

Lake-fish metacommunities

For simplicity, we only present results for the CWM and fourth-corner correlations, thus omitting SNC and Chessel's fourth-corner. The contrast between CWM and fourth-corner was made on the basis of 1) a table of watersheds by environmental variables (85 by 14 = 1190 correlations) indicating which correlations were significant under the three permutation schemes (Fig. 5); 2) the distribution of significant correlation values; and 3) a histogram showing the frequency distribution for p-values smaller than 0.10 for row–column permutation. Given that only the row–column permutation scheme provides correct type I errors for all three statistics, we can clearly conclude that the fourth-corner correlation is more powerful than the CWM and weighted CWM correlations. Note, however, that additional information can be extracted from the other two permutation schemes, when the row–column permutation scheme is not significant. When correlations are only significant based on row permutation, but not on column permutations, it indicates that only environment was important in structuring species distributions (Fig. 5); conversely, when only the permutation of columns led to significant correlations, then only body size was important in structuring species distributions (Fig. 5).

Discussion

Our goal was to show the relationships and statistical properties of some of the most commonly used methods to assess the links between trait variation and the environment. The methods reviewed here are based on simple correlations between trait and environmental features, although we showed how the fourth-corner can be directly expanded to a multivariate regression approach. Their algebra is quite similar, but they vary in the way traits and environments are standardized and whether correlations are weighted or unweighted.

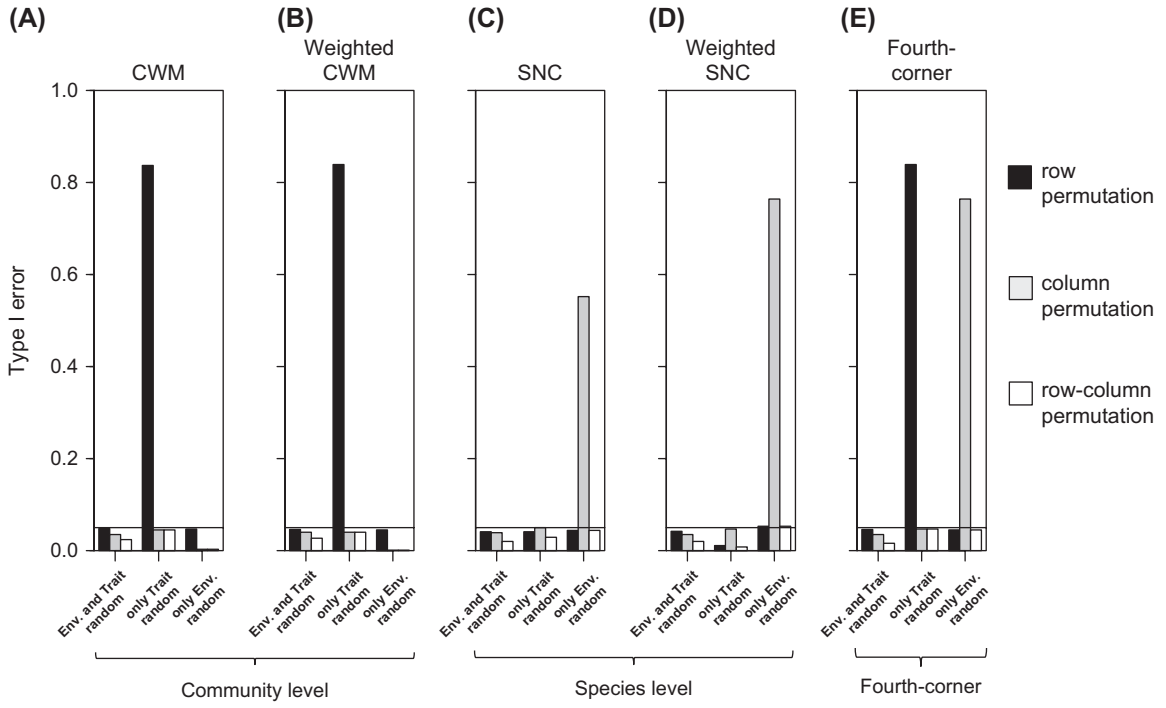

Figure 3. Type I error rates based on 1000 sample tests (9999 permutations, alpha = 0.05) for the three different permutation strategies (row permutation – across communities, column permutation – across species, and row–column permutation – across species and communities) under different ways in which trait and environment were not linked to community structure (see legend of Fig. 2 for details).

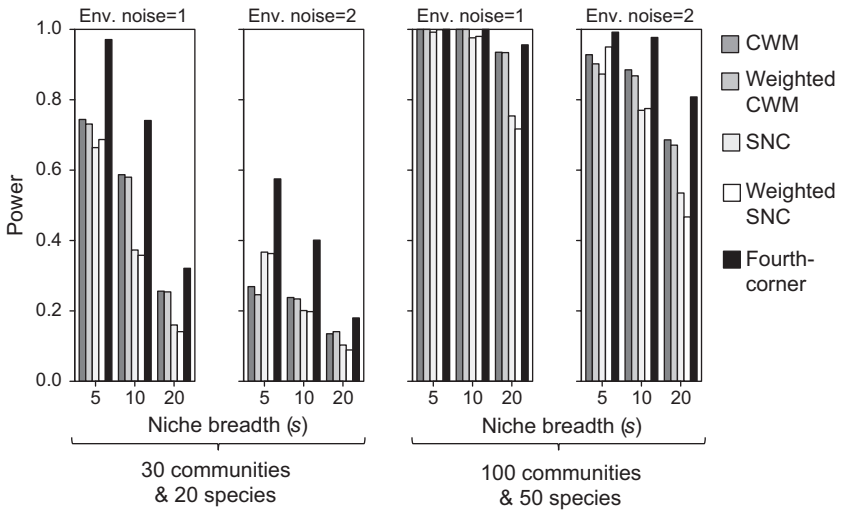

Figure 4. Rejections rates (statistical power) based on 1000 sample tests (9999 permutations, alpha = 0.05) for the row/column permutation strategy for different levels of species niche breadth (*s*): 5, 10 and 20. Only power for the row–column permutation scheme is shown because it is the only one that provides correct type I error rates.

Moreover, the units being modeled are either communities (trait averages across communities in CWM-based correlations) or species (environmental averages across species in SNC-based correlations), whereas the fourth-corner approach integrates both column (trait) and row (environment) marginals. Our results clearly indicate that the fourth-corner approach outperforms the CWM- and the SNC-based approaches in terms of statistics' sampling accuracy and statistical power. We also show that the standard statistical tests to judge significance of CWM- and SNC-based correlations are inconclusive about the existence of trait–environment association as they yield highly inflated type I error when there is in fact no trait–environment association. In these cases, only the row–column permutation scheme allows for a correct type I error for all approaches. Finally, the fourth-corner approach allowed us to derive a general framework of trait decomposition into components among (mean) and within- (variance) communities. Although we only treated the modelling of the mean component, the variance component could be equally considered, for instance, using a log-linear model in which the log of trait variance is regressed against linear predictors. Log-linear regressions has been used in the past to detect heteroscedasticity in residual variation (Cook and Weisberg 1983) but can be equally used to model variance as a response variable.

Because most studies have been using one parametric test (akin to either row-or column permutation), they have most likely being reporting extremely significant and high correlations, biasing our knowledge about trait–environment associations (Fig. 3). As shown by our type I error results, a significant CWM-based correlation based on row-permutation or parametric testing may simply indicate that environment drives species distributions but not that there is a link to trait variation. The advantage of the row–column scheme is that two p-values are produced and we can draw conclusions from each of them separately as well as from their combination (Fig. 3): 1) significant under the row–column permutation – link between trait and environment; 2) only significant under the row permutation – link of species distributions with environment but not trait; 3) only significant under the column permutation – link of species distributions with trait but not environment. By reporting the significant correlation for the three testing schemes, as in Fig. 5, we can separate three types of patterns, leading to more insight about the true nature of these relationships. We conjecture that current testing procedures of trait–environment associations in GLM-based species distribution models (Jamil et al. 2013, Brown et al. 2014) are also too tolerant as they are based upon a single test only (akin to row permutations).

The row–column permutation test, when significant, guarantees (up to the significance level) that species abundances are linked to both the trait and the environment (ter Braak et al. 2012). In a log-linear model, traits and environment may each have an important main effect, so that both influence species distributions. Without interaction, trait and environment act separately (on the logscale) and not many ecologists will say that there is trait–environment association in this case. Fortunately, the fourth-corner approach is not sensitive to detecting main effects in a log-linear model; it does not detect trait–environment association when there are only main effects, as we verified by simulation (Supplementary material Appendix 4). This result does not come unexpected as the fourth-corner approach is closely linked to correspondence analysis, which decomposes the chi-square statistic for testing interaction in contingency tables along several ordination axes. Main effects are removed in the calculation of the chi-square statistic and methods based on it are thus not sensitive to main effects. One often overlooked characteristic of the fourth-corner correlation is thus that it assesses the strength of an interaction between environment and traits in the underlying species distributions (e.g. species inhabit sites with lower temperature the smaller their body size, implying that species differ in habitat selection depending on their body size) rather than additive effects (e.g. body sizes and

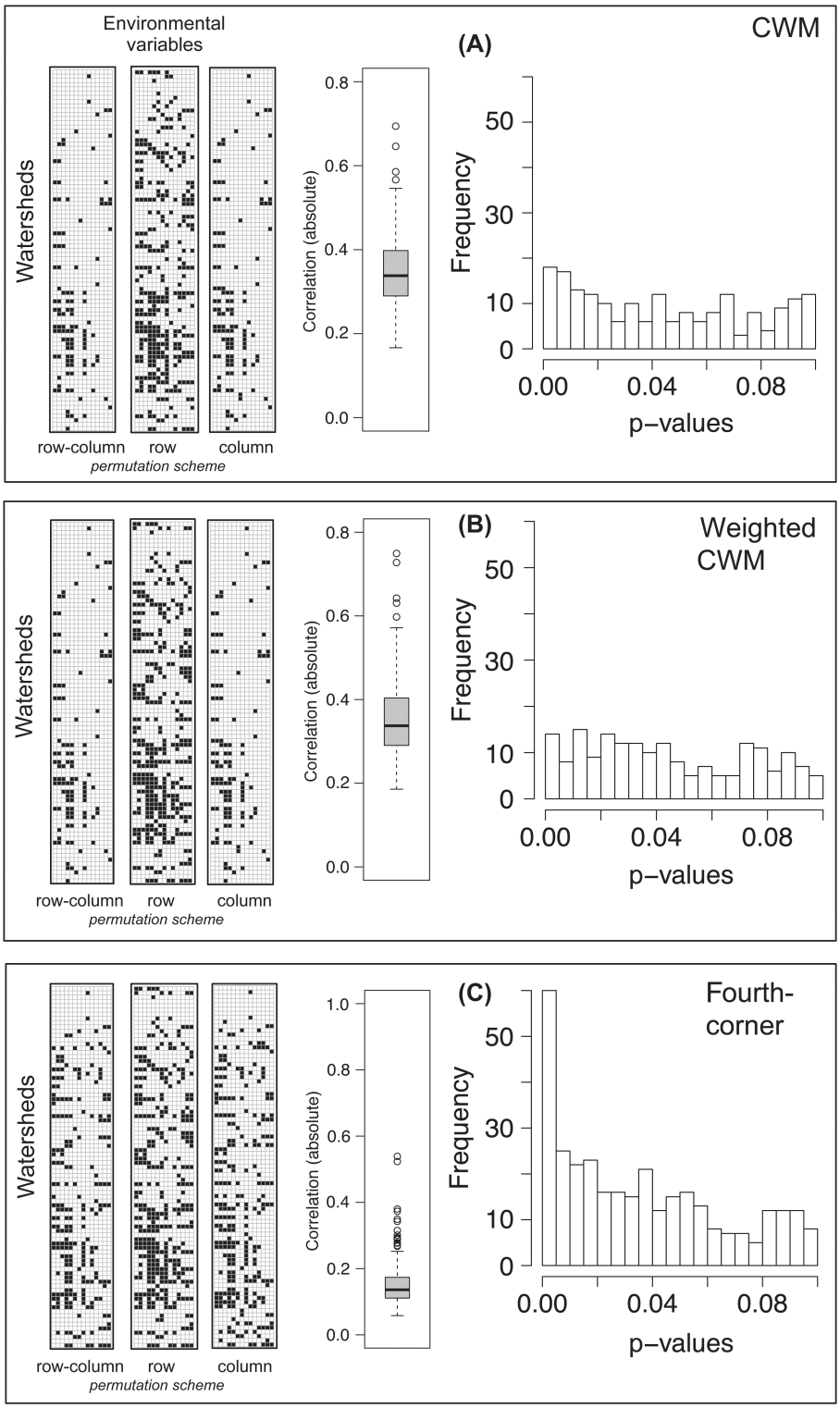

Figure 5. Trait–environment analyses for the lake-fish metacommunities (85 watersheds across 14 environmental variables) for three statistics: community weighted mean correlations (CWM), community weighted mean weighted correlations and the fourth-corner. The left inset tables show which correlations were significant (9999 permutations, alpha = 0.05) for each permutation test scheme. Environmental variables are in the following order: lake elevation (column 1), total dissolved solids, pH, Secchi depth, oxygen concentration, morphoedaphic index, surface area, maximum depth, mean depth, perimeter, island perimeter within lakes and growing-degree days (column 14). The boxplot (median, lower and upper quartiles and extreme values) represents the distribution of the significant correlations. The right inset represents frequency distributions of probabilities of rejection (smaller than 0.10) based on the row–column scheme for each of the three types of correlation.

temperature both influence abundance, but temperature-based habitat selection is independent of body size).

CWM-based correlations are likely the most widely used approach to link trait to environmental variation (Lavorel et al. 2008), though it only considers the component of variation among communities. Unfortunately, this component is typically small. In our real data sets (and others not shown here), we found that the variation due to this component is in general much less than 50%. Therefore, by using the CWM-based approach, only a small fraction of the total variation has currently been estimated and, as a result, over-optimistic correlations have been reported. For instance, one may find a correlation of 0.80 quite high at first, but low if this correlation is found to explain only, say, 5% of the total variation in trait composition; if so, one may not be as hopeful to have captured much of the drivers of a particular set of communities.

Whereas the CWM-based correlation overestimates the strength of the trait–environment relationship, the fourth-corner correlation is perhaps too low in absolute value, as it cannot reach the usual extremes of –1 and 1. Therefore we proposed to rescale the fourth-corner correlation to its full range, yielding the Chessel fourth-corner correlation. A similar rescaling is reported in the output of the rlq command of the R package ade4. The rescaling is based on the first eigenvalue of correspondence analysis (CA), a widely used ordination technique in ecology (ter Braak 1985). Note that we are not advocating replacing the fourth-corner correlation by the Chessel's form. Instead, we suggest that both correlation values should be reported. The original fourth-corner correlation provides high sampling precision (Fig. 2) whereas Chessel's form provides a relative measure of how well a single environmental–trait correlation does in explaining species distributions.

Because we know the maximum possible absolute value of the (Chessel) fourth-corner correlation based solely on the species distribution matrix, we can select in advance which species data matrices are the best candidates to uncover links between environment and traits. Moreover, based on this property, we can infer on latent environmental and trait features (Hui et al. 2015). For instance, we can consider the species (column) scores from the CA axis as latent traits and calculate fourth-corner correlations using the original environmental variables. Conversely, we can consider the community (row) scores as environmental features and calculate fourth-corner statistics with the original trait. If the correlations are greater for when the observed traits (instead of CA columns scores) were used than when the observed environmental features were used instead, we can infer that the measured traits contain less uncertainty in determining the community structure than environment. Moreover, by mapping community CA scores independent of observed environment (residual variation), one can infer on the nature of missing environmental predictors (Leathwick and Whitehead 2001, Pollock et al. 2014) to explain trait–environmental relationships, particularly in the case in which the maximum fourth-corner correlation (obtained via the first CA eigenvalue) is high but the observed fourth-corner correlation is low.

In Brown's et al. (2014) paper on GLM-based approaches for modelling species distributions based on environment, trait and their interactions, they suggested that the fourth-corner approach is a hypothesis testing only approach and that the strength of the association cannot be assessed. Here, we have shown that the fourth-corner when embedded in a weighted regression model provides a general modelling approach can estimate parameters and accommodate for different types of predictors. Finally, the fourth-corner approach allows for a full additive decomposition of trait variation. Future work is needed to contrast the differences between these two types of frameworks (e.g. species distributions versus community traits as response variables) not only in terms of statistical performance but also in terms of underlying community assembly processes.

In conclusion, the fourth-corner approach is a simple, extendable and powerful method that should replace the community weighted means and species niche centroid approaches for linking environmental and trait variation.

Acknowledgements – We would like to thank David Warton, and Pierre Legendre for their comments on early versions of the manuscript. We acknowledge the funding from Government of Canada, Natural Sciences and Engineering Research Council of Canada.

References

Ackerly, D. D. and Cornwell, W. L. 2007. A trait-based approach to community assembly: partitioning of species trait values into within- and among-community components. – Ecol. Lett. 10: 135–145.

Brown, A. M. et al. 2014. The fourth-corner solution – using predictive models to understand how species traits interact with the environment. – Methods Ecol. Evol. 5: 344–352.

Cavender-Bares, J. et al. 2004. Multiple trait associations in relation to habitat differentiation among 17 Floridian oak species. – Ecol. Monogr. 74: 635–662.

Cook, R. D. and Weisberg, S. 1983. Diagnostics for heteroscedasticity in regression. – Biometrika 76: 1–10.

Coyle, J. R. et al. 2014. Using trait and phylogenetic diversity to evaluate the generality of the stress-dominance hypothesis in eastern North American tree communities. – Ecography 37: 814–826.

Dolédec, S. et al. 1996. Matching species traits to environmental variables: a new three-table ordination method. – Environ. Ecol. Stat. 3: 143–166.

Dray, S. and Legendre, P. 2008. Testing the species traits–environment relationships: the fourth-corner problem revisited. – Ecology 89: 3400–3412.

Dray, S. et al. 2014. Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. – Ecology 95: 14–21.

Garnier, E. et al. 2004. Plant functional markers capture ecosystem properties during secondary succession. – Ecology 85: 2630–2637.

Gause, G. F. 1930. Studies on the ecology of the Orthoptera. – Ecology 11: 307–325.

Henriques-Silva, R. et al. 2013. exploring patterns in species distributions across large geographic areas. – Ecology 94: 627–639.

Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. – Ecology 54: 427–432.

Hui, F. K. C. et al. 2015. Model-based approaches to unconstrained ordination. – Methods Ecol. Evol. 6: 399–411.
Jamil, T. et al. 2013. Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. – J. Veg. Sci. 24: 988–1000.
Keddy, P. A. 1992. Assembly and response rules: two goals for predictive community ecology. – J. Veg. Sci. 3: 157–164.
Kleyer, M. et al. 2012. Assessing species and community functional responses to environmental gradients: which multivariate methods? – J. Veg. Sci 23: 805–821.
Kühn, I. et al. 2009. Combining spatial and phylogenetic filtering in a trait analysis of plant flowering phenology in Switzerland. – Global Ecol. Biogeogr. 8: 745–758.
Lavorel, S. et al. 2008. Assessing functional diversity in the field – methodology matters! – Funct. Ecol. 22: 134–147.
Leathwick, J. R. and Whitehead, D. 2001. Soil and atmospheric water deficits and the distributions of New Zealandís indigenous tree species. – Funct. Ecol. 15: 233–242.
Legendre, P. et al. 1997. Relating behavior to habitat: solutions to the fourth-corner problem. – Ecology 78: 547–562.
MacArthur, R. H. 1972. Geographical ecology. – Harper and Row.
McGill, B. J. et al. 2006. Rebuilding community ecology from functional traits. – Trends Ecol. Evol. 21: 178–185.
Messier, J. et al. 2010. How do traits vary across ecological scales? A case for trait-based ecology. – Ecol. Lett. 7: 838–848.
Paquette, A. and Messier, C. 2011. The effect of biodiversity on tree productivity: from temperate to boreal forests. – Global Ecol. Biogeogr. 20: 170–180.
Pavoine, S. et al. 2011. Linking patterns in phylogeny, traits, abiotic variables and space: a novel approach to linking environmental filtering and plant community assembly. – J. Ecol. 99: 165–175.
Peres-Neto, P. R. et al. 2012. Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. – Ecology 93: S14–S30.
Pollock, L. J. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). – Methods Ecol. Evol. 5: 397–406.

Supplementary material (Appedix ECOG-02302 at <www.ecography.org/appendix/ecog-02302>). Appendix 1–4.

Ricotta, C. and Moretti, M. 2011. CWM and Rao's quadratic diversity: a unified framework for functional ecology. – Oecologia 167: 181–188.
Scott, W. B. and Crossman, E. J. 1973. Freshwater fishes of Canada. – Bulletin 184, Fisheries Research Board of Canada.
Swenson, N. G. and Enquist, B. J. 2007. Ecological and evolutionary determinants of a key plant functional trait: wood density and its community-wide variation across latitude and elevation. – Am. J. Bot. 94: 451–459.
ter Braak, C. J. F. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. – Biometrics 41: 859–873.
ter Braak, C. J. F. and Barendregt, L. G. 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. – Math. Biosc. 78: 57–72.
ter Braak, C. J. F. and Looman, C. W. N. 1986. Weighted averaging, logistic regression and the Gaussian response model. – Plant Ecol. 65: 3–11.
ter Braak, C. J. F. et al. 2012. Improved testing of species traits–environment relationships in the fourth corner problem. – Ecology 93: 1525–1526.
Terribile, L. C. et al. 2009. Ecological and evolutionary components of body size: geographic variation of venomous snakes at the global scale. – Biol. J. Linn. Soc. 98: 94–109.
Tingley, M. W. et al. 2009. Birds track their Grinnellian niche through a century of climate change. – Proc. Natl Acad. Sci. USA 106: 19637–19643.
Webb, C. T. et al. 2010. A structured and dynamic framework to advance traits-based theory and prediction in ecology. – Ecol. Lett. 13: 267–283.
Weiher, E. et al.1998. Community assembly rules, morphological dispersion, and the coexistence of plant species. – Oikos 81: 309–322.
Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. – Ecol. Monogr. 30: 279–338.
Whittaker, R. H. 1967. Gradient analysis of vegetation. – Biol. Rev. 42: 207–264.
Williams, E. J. 1952. Use of scores for the analysis of association in contingency tables. – Biometrika 39: 274–289.